for Journal of Proteome Research

PeptidePicker: a Scientific Workflow with Web Interface for Determining Appropriate Peptides for Targeted Proteomics Experiments

Yassene Mohammed\*1,2, Dominik Domański3, Angela M. Jackson1, Derek S. Smith1,
 André M. Deelder2, Magnus N. Palmblad\*2,and Christoph H. Borchers\*1

1University of Victoria - Genome British Columbia Proteomics Centre, University of Victoria, Canada

2Center for Proteomics and Metabolomics, Leiden University Medical Center, the Netherlands

3Mass Spectrometry Laboratory, Institute of Biochemistry and Biophysics PAS, Warsaw, Poland

#### ABSTRACT

One challenge in Multiple Reaction Monitoring Proteomics analysis is to select the most appropriate surrogate peptides to represent a target protein. Several features are required for a good surrogate peptide -- it should be unique within the target proteome, efficiently liberated during enzymatic digestion, and free of post translational modifications or variants. We present here a software package to automatically generate the most appropriate surrogate peptides to represent target proteins in an LC/MRM-MS analysis. Our method integrates ~~most of the~~ information about the proteins, their tryptic peptides and suitability for MRM that is available online in UniProtKB, NCBI’s dbSNP, ExPASy, PeptideAtlas, and GPM. We introduce a new scoring algorithm that reflects our knowledge in choosing best candidate peptides for an MRM experiment based on the uniqueness of the peptide in the targeted proteome, the chemophysio properties of each peptide, and whether it was previously observed or not. The scoring mechanism can be modified to allow different weighting of peptides selecting criteria if desired. The scientific workflow allows linking to a user's database to determine whether already-used peptides are indeed the most appropriate ones or not by matching them with the selection criteria, other possible peptides, and scoring them. The modularity of the developed workflow, i.e. different processing unites each dedicated to one specific task, allows it to be easily extended and allows additional selection criteria to be incorporated. In order to hide the complexity of the underlying system, and to make it user friendly, we have developed a simple Web interface where the researcher provides the protein accession number, the subject organism, and peptide-specific options (for example, if only the mature form of the protein should be considered). Currently the software is designed for human and mouse proteomes, but additional species can be easily be added.. In comparison with manually selected peptides, our methods enriched the peptide selection by eliminating human error, considering all of a protein isoforms, and allowed faster peptide selection. Regarding speed our tool managed more than 50 proteins in 12 minutes compared to 8 proteins a day by expert. In all tests all peptides chosen by expert were a subset of the selected peptides by PeptidePicker, except when there was a human error.

*Keywords: MRM, SRM,* targeted proteomics, peptide selection, data integration, scientific workflow**INTRODUCTION**

Multiple Reaction Monitoring (MRM), also known as Selected Reaction Monitoring (SRM), is a powerful method for performing targeted proteomics. The approach is based on the generation of specific quantitative assays for each protein of interest and can be used to accurately quantitate large sets of proteins at high throughput.[1](#_ENREF_1) MRM studies are typically performed on triple quadrupole mass spectrometers. The first quadrupole is used to isolate a specific precursor peptide ion, the second quadrupole is used for collision-induced dissociation, and the third is used to isolate a characteristic fragment ion. In multiplexed analysis, specific precursor/product ion pairs are monitored while the peptide is eluting from the liquid chromatography system, which allows quantitation of specific peptides and, by inference, the corresponding proteins.

Designing an MRM/SRM assay for a specific protein starts with the selection of the most appropriate peptides to represent the target protein. Depending on the study design, these peptides are selected according to multiple criteria so that they are unique to the target protein. In addition, they must follow other rules in order to facilitate their accurate and precise measurement in the mass spectrometer. The selection process has, thus far, been cumbersome, and is usually performed by scientists with proficiency in selecting peptides for MRM analyses. We present here a scientific workflow to automatically select the most appropriate surrogate peptides to represent a target protein in an LC/MRM-MS analysis. The method integrates the data and information available online in UniProt[2](#_ENREF_2" \o "UniProt_Consortium, 2009 #14892), NCBI’s dbSNP[3](#_ENREF_3" \o "Sherry, 1999 #16341), ExPASy[4](#_ENREF_4" \o "ExPASy_Bioinformatics_Portal, 2005 #15519), PeptideAtlas[5](#_ENREF_5" \o "PeptideAtlas, 2010 #14082), and the Global Proteome Machine -- GPM.[6](#_ENREF_6) By linking one's own database of MRM peptides, the scientific workflow can also verify whether an already selected peptide is indeed the most appropriate surrogate peptide for the target protein. The workflow also integrates our knowledge and practice in using available information for choosing peptides. It also gives the researcher the freedom to switch some of the selection criteria on and off, according to the analysis. The method can also be extended or modified to include (or exclude) additional peptide-selection criteria. Compared to the manual process of choosing peptides by a scientist with proficiency in peptide selection for MRM experiment, our method is faster and more accurate. Due to the automatic enforcement of the selection criteria, the tool can also be used to find and eliminate human error in previously selected lists of peptides. The software can also be used for periodic improvement of precursor/product ion lists, for example when the discovery of new post translational modifications or SNPs might result in some peptides no longer meeting the selection criteria.

**MATERIALS AND METHODS**

***Data formats***

Our implementation uses different formats corresponding to the specific data repositories from which the information is retrieved. The protein Data downloaded from UniProt[2](#_ENREF_2" \o "UniProt_Consortium, 2009 #14892) is in XML. The online search results from GPM[6](#_ENREF_6) are in HTML. Data from PeptideAtlas[5](#_ENREF_5" \o "PeptideAtlas, 2010 #14082) is in fasta and XML formats. For the purpose of our workflow, using markup languages was essential due to the convenience of parsing them to extract different values of interest according to specific tags. Data in fasta format (*i.e.*, lists of protein sequences) are generally easier and faster to search, and do not reflect the complexity and standardization of XML formats. Our workflow takes advantage of and uses both formats whenever appropriate.

The intermediate results from each processing module of the scientific workflow is transferred to the next module using data lists that can be saved as text files or as comma-separated values lists whenever needed. The ability to see the intermediate stages of the selection process allows the researcher to see the origins and plausibility of the results at some later time point, which is useful if an unexpected result is generated. This is especially important when the content of the data repositories we are using is changing over time as new discoveries are made and as scientists upload new data.

***Scientific Workflow Engine***

We used the Taverna workflow engine to build one complete data processing workflow from the different data processing steps.[7](#_ENREF_7) Scientific workflows allow automatic transition between processing steps, and workflow engines such as Galaxy[8](#_ENREF_8), Moteur[9](#_ENREF_9" \o "Maheshwari, 2010 #48), Kepler[10](#_ENREF_10" \o "Altintas, 2004 #53), and Taverna[7](#_ENREF_7" \o "Oinn, 2004 #45) have been introduced to facilitate interfacing modular processing steps, automating analysis pipelines, scaling them up to Big Data, and making analyses reproducible and shareable. Here we used Taverna 2.4 which offers various types of processors, such as WSDL/REST Web services, Tools, and XPath processors.[7](#_ENREF_7), [11](#_ENREF_11) For our implementation, we used Beanshell processors, which enable execution of small Java code snippets as part of a workflow. Rshell processors were used to execute the R scripts for peptide selection and comparison as well as for advanced XML parsing. Finally XPath processors are used for simple XML files parsing and extracting needed information.

***Evaluation Data***

In order to evaluate the accuracy of the method in selecting the right peptides, the method was tested using multiple lists of proteins. In order to test the correctness of the results and the processing speed on a desktop computer, we used a list of 50 proteins for which peptides had been previously manually selected. In order to test the robustness of the software, to simulate future use, and to test the software for Big Data scenarios, we also used this software to select peptides for all of the proteins from human chromosome 21.

***Related Work***

Many current software packages select the “best” proteotypic peptides from a given protein based on the sequence information alone. In this context, the best peptides for MRM are considered to be those whose physicochemical properties are likely to result in high sensitivity in electrospray ionization (ESI)-MS experiments, and which will generate high fragment ion signal intensities for a given transition. Software packages using this approach include the ESP predictor [12](#_ENREF_12), Peptide Sieve [13](#_ENREF_13), and PepFly[14](#_ENREF_14), [15](#_ENREF_15) [Boja and Rodriguez 2012]. These software packages generally depend on “training sets” considering hundreds of physiochemical peptide/amino acid properties with peptides classified according to their precursor responses as derived from extracted ion chromatograms (XICs). Compared to manual empirical selection of MRM transitions which are composed of high-responding precursor/product ion pairs instead of an XIC signal, the ESP predictor, however, correctly selected on average only two out of the top five experimentally validated MRM peptides per protein (#9).[12](#_ENREF_12)

PeptideAtlas[5](#_ENREF_5), [16](#_ENREF_16) has combined its database of observed peptides with PABST (Peptide Atlas Best SRM Transition)[17](#_ENREF_17) to select the best peptides. Its score is based on *in silico* prediction and experimental MS observation, with the addition of a number of weighting factors such as reactive amino acids and peptide length criteria. However, the PABST tool can still, for example, rank peptides highly if they have multiple gene locations or contain cysteine and methionine residues if they are predicted to yield high ion signals, even though peptides that contain cysteine or methionine are undesirable in an MRM assay because these amino acids can easily undergo chemical modification such as oxidation. This makes them "non-unique" because multiple peptide forms can be produced, all of which would need to be monitored for accurate quantitation.

Brusniak et al introduced Automated and Targeted Analysis with Quantitative SRM (ATAQS),[18](#_ENREF_18) an analysis pipeline that organizes, generates, and verifies transition lists, and performs post-acquisition analysis. Although ATAQS integrates information from online repositories such as protein-protein interactions or cellular functions, this software does not take biological constraints into account or make use of available data on post-translational modifications, natural variants, biological protein processing and isoform variants during the peptide selection.

Other approaches include MRMaid,[19](#_ENREF_19) the GPMDB MRM Worksheet,[20](#_ENREF_20) SRM/MRMAtlas,[21](#_ENREF_21) TIQAM[22](#_ENREF_22) (used by the previously mentioned ATAQS workflow), Skyline,[23](#_ENREF_23) and TPP-MaRiMba[24](#_ENREF_24), [25](#_ENREF_25). These, however, are computational tools that focus primarily on the design, optimization, and selection of appropriate MRM transitions, rather than the selection of the precursor peptide. MRMaid does suggest peptide targets in addition to transitions for the submitted protein.[26](#_ENREF_26) The suggestion of ions to be monitored is based on EBI’s PRIDE spectral database[27](#_ENREF_27) with scores given for the likelihood of observing a given peptide.[28](#_ENREF_28) The score also takes into consideration the suitability of an MRM transition, based on the frequency and intensity of the observed fragment ions of the specific peptide. The software checks selected peptides for uniqueness in a given proteome, and the user can specify peptide lengths, the specific MS instrument used, and can indicate LC conditions so that predicted retention times can be taken into account. The software can also filter peptides based on additional criteria such as unwanted amino acids, but it does not go any further in using criteria that would filter out peptides based on biological context information available from UniProt, such as post-translational modifications or the presence of natural variants.[19](#_ENREF_19)

Similarly, MRM Worksheet in gpmDB and MRMAtlas in PeptideAtlas suggests MRM transitions based on spectral information contained in their respective databases but actually does less in terms of suggesting suitable peptides for MRM assay design.[25](#_ENREF_25), [29](#_ENREF_29) Skyline, MaRiMba, and TIQAM/ATAQS focus more on the design of MRM transitions for selected sets of peptides and the empirical validation of this pool of peptides from user-generated experimental MRM data,[18](#_ENREF_18), [24](#_ENREF_24), [30](#_ENREF_30). Although these tools can access information stored in spectral databases and can point out frequently observed fragment ions and peptides, the selection of the “best” peptide based on biological criteria is limited. A search of PeptideAtlas displays the occurrences of SNPs and post-translationally modified residues in peptides, but their presence is not taken into consideration when calculating the different peptide suitability scores. This should be considered as an indicator of a peptide's suitability for an MRM assay, but it is not considered at this time by these software packages.

Clearly there is a need for an automated software package that takes the biological criteria into consideration during the peptide-selection stage of the MRM assay development. Our workflow integrates information from different data sources and builds on the information stored in UniProt, NCBI’s dbSNP, GPM, and PeptideAtlas. Our method considers the post-translational modifications, natural variants, and post-translational peptide chain processing as well as allowing or excluding specific amino acids. Our software also takes into account other scores, including the PeptideAtlas Empirical and Predicted Suitability Scores.[16](#_ENREF_16) Because it is not limited to these sources, our software can be easily extended to consider additional data repositories, including the user's own peptide libraries. The ability to extend or reduce the workflow allows our method to match the demands of a specific laboratory or experiment. The reproducibility of the results is also an important part of method development and is facilitated by the use of Tavern as a workflow engine, as well as by the possibility of archiving all of the data retrieved from online repositories for later inspection.

**RESULTS**

***Scientific workflows***

Figure 1 shows the Taverna modular workflow that implements the peptide selection logic. The workflow requires a list of protein accession numbers as input, along with the target organism and specific selection criteria if they are different from the default values. Currently the tool supports human and mouse proteomes, but can be easily extended to include additional organisms. The software incorporates different tools to perform different steps: finding protein isoforms, *in silico* tryptic digestion, peptide filtering, peptide and protein information retrieval from online repositories, peptide selection criteria enforcement, and report generation. The workflow compares the peptides passing the selection criteria with the online contents of Global Proteome Machine and PeptideAtlas and indicates whether or not the peptides have been previously observed.

All processing modules run locally, and all of the data used to generate the final list is stored for later manual or automated inspection if needed. The software mainly uses three technologies: Java,[31](#_ENREF_31) R statistical language,[32](#_ENREF_32) and XML Path Language (XPath).[33](#_ENREF_33) Java Beanshell consists mainly of command-line calls for the needed software with the desired parameters, which also include connecting and downloading data from online repositories. Statistical analysis modules are written using R and are run locally. XPath parsing modules extract information snippets from XML files and report them as lists.

***Peptide Acceptance Criteria***

Surrogate proteotypic peptides should be efficiently liberated during trypsin digestion, free of PTMs and variants, and unique within the target proteome. The peptide selection can, however, be based on more stringent criteria, which are summarized below:

* Peptide must be unique in the proteome of the studied organism. To ensure correct quantitation of proteins, the representative surrogate peptides must appear only once in the proteome. We consider a peptide unique to a protein if it appears only once in the UniProt database of reviewed[34](#_ENREF_34) proteins for the organism under study.
* Previously observed peptides. PeptideAtlas and gpmDB collect experimental MS data and are therefore very good sources of information regarding peptides that can have sufficient sensitivity to be detected in a mass spectrometer. Peptides in these repositories that match all of the other obligatory criteria are therefore favored, as it is most likely that they will be observed in an MS experiment. PeptideAtlas ranks observed peptides using the Empirical Suitability Score (ESS), which is derived from the peptide identification probability, Empirical Observability Score (EOS), which is a measure of in how many samples a particular peptide is seen, relative to other peptides from the same protein and the number of times the peptide is observed, which is the total number of observations in all modified forms and charge states. The ESS is also adjusted for unfavorable sequence characteristics such as semi-tryptic or missed cleavage sites, or multiple genome locations. For peptides with no or a low number of observations in its database, PeptideAtlas reports a Predicted Suitability Score (PSS), which is derived by combining publicly available algorithms (Peptide Sieve,[35](#_ENREF_35) STEPP,[36](#_ENREF_36) ESPP,[12](#_ENREF_12) APEX,[37](#_ENREF_37) Detectability Predictor,[38](#_ENREF_38)) to define the likelihood of observing a peptide for a protein. Our workflow considers whether a peptide is in gpmDB or PeptideAtlas, and whether a peptide has ESS or PSS scores or not, and integrates this information into the final selection of peptides.
* *in silico* predicted efficiency of trypsin digestion. We use the ExPASy PeptideCutter with the “sophisticated” model for trypsin in order to estimate the digestion efficiency. The cleavage probabilities in this model are based on a statistical model derived from the charts published by Keil.[39](#_ENREF_39) Although there are different studies and systems that outperform the Kiel model, this study has been the main reference model for comparing with newer methods. Due to the modularity of our software, different digestion rules or models can be included. At the implementation level, in order to consider possible peptides with missed cleavages, our tool runs the ExPASy PeptideCutter with increasing digestion efficiencies set at a minimum of 10% to a maximum of 98% , and includes all of the resulting peptides if they obey the other obligatory criteria.
* Peptide length. For specificity and for the efficient synthesis of the standard peptides used for accurate quantitation, the software considers peptides between 7 and 20 residues long, by default. Short peptides (<6) are less likely to be unique and might have protonated regions masked by the solvent in MS analysis. The upper limit of 20 to 25 residues, is the practical limit of efficient peptide synthesis, as well, larger peptides ionize and fragment less efficiently and their large fragment ions will fall outside the range of most triplequadrupole analyzers. In our software, the upper and lower limits can be specified by the user as inputs.
* Existence of the peptides in all of the isoforms of a protein. Depending on the context of the experiment, peptides can exist in all isoforms, in the canonical sequence only, or only in one single isoform. Our workflow supports all three cases, in that the user chooses the corresponding UniProt accession number. For example in case of Nuclear receptor coactivator 3 (UniProt accession number Q9Y6Q9) which has 5 isoforms, the software accepts Q9Y6Q9 as an input if the user requires a peptide which can measure all 5 isoforms, or Q9Y6Q9-1, Q9Y6Q9-2, Q9Y6Q9-3, Q9Y6Q9-4, or Q9Y6Q9-5 if only the canonical sequence (Q9Y6Q9-1) or any of the other isoforms (-2 to -5) is the particular isoform that needs to be measured. If the canonical sequence or only one isoform is selected, only that specific sequence is considered and any peptides shared with other isoforms are considered not unique. The user can control the behavior of this selection criterion by switching off the corresponding input. In this way peptides which are not in all isoforms of the protein can also be reported. This allows all of the observed peptides in MS/MS databases to be considered in the selection process, and therefore suggest the most sensitive MRM peptides, but the top selected peptides might not cover all existing isoforms of the protein.
* Protein biological annotation. When selecting and scoring the peptides, the software takes into account the biological information regarding protein annotation present in the UniProt knowledge database. Ideally, peptides should not be present in the signal peptide or propeptide sequences which will be cleaved off in the mature form of the protein. Considerations also need to be taken with regard to the processing of the mature form of the protein into multiple chains in order to avoid selecting a peptide which might be split between two separate processed chains. Our software extracts all annotations about a protein available in UniProt and labels any peptide which is annotated, for example if it is included in or overlaps with a signal or a propeptide. The user can also choose to include only those peptides which are part of the mature form of the protein in the final report.
* Absence of methionine (M) and cysteine (C). Both M and C are highly susceptible to oxidation. Oxidation of M to sulfoxide (+16 Da mass shift) and sulfone derivatives (+32 Da mass shift) will split the total peptide MRM signal over multiple peptide forms therefore reducing the assay sensitivity and requiring MRM quantitation of all of the different species. Oxidation of C can also form insoluble aggregates through the formation of disulfide bonds.
* Preference given to those peptides which do not contain tryptophan (W), which like M, is prone to oxidation but to a much lesser extent.
* Absence of N-terminal glutamine (Q). Under acidic conditions, Q on the N-terminal cyclizes to form Pyroglutamate.[40](#_ENREF_40)
* Absence of aspartic acid paired with proline or glycine (DP and DG). These combinations can undergo hydrolysis and cause peptide cleavage under acidic conditions.
* Absence of asparagine-glycine (NG) and glutamine-glycine (QG). These are prone to deamidation at neutral and alkaline pH.
* Absence of sequential proline (P) and serine (S) residues. P or S strings can cause significant deletions during peptide synthesis resulting in low yield. This is especially important for proline residues which can undergo cis/trans isomerization and reduce peptide purity.
* Absence of significant natural variants. These are non-homologous single nucleotide polymorphisms (SNPs) that occur in the studied population at a high enough frequency to affect the biological question of interest. Amino acid substitutions that occur at a frequency higher than 1% are considered as polymorphisms. Peptides containing such substitutions, or flanking sites that will have affected lysine (K) or arginine (R) residues (for tryptic cleavage) are considered during peptide selection. Natural variants such as disease mutations are usually a less frequently observed issue, however, depending on the population being studied the significance of such modifications can be considered when choosing peptides for MRM analysis. The software checks the frequency of these allele changes for each SNP in the NCBI database of SNPs and considers it in scoring the peptides.
* Absence of significant post-translationally modified residues (PTMs). PTMs, such as phosphorylation, glycosylation, and acetylation, need to be avoided on the peptide of interest as these, like with oxidation, would mean that different forms of the same peptide would need to be monitored for an accurate quantitation. Such modifications need also to be avoided on the flanking K/R residues (next to the N-terminal of peptide of interest) because they will affect tryptic cleavage of that peptide. Depending on their frequency, modifications may affect the production of the tryptic peptide and therefore affect its MRM measurement. In general, a possible PTM site should be avoided, but if the choice of peptide is limited, the frequency of the occurrence of the PTM may need to be considered. The probability of a PTM’s occurrence can be roughly estimated by the number of literature references supporting the modification (as indicated in the UniProtKB database, for instance). PTMs with supporting experimental evidence are more significant than "potential" modification sites suggested solely on the basis of sequence motifs such as NXT/S for N-glycosylation, for example. Our tool can be set to avoid any referenced and potential modifications, or to penalize peptides containing these during the score calculation. The tool penalizes peptides containing modifications with referenced evidence more than potential sites based on non-experimental data as indicated in UniProt.

These criteria can be modified and readjusted according to the needs of the experiment. The workflow contains two filtering processors and these can be modified to include, exclude, or perform additional advanced selection criteria testing

***Scoring***

For the purpose of ranking the peptides, we have introduced a simple scoring mechanism that combines all of the selection criteria – the v-score. This scoring system is not meant to be an absolute measure of how suitable the peptide is for an MRM experiment -- it is a subjective scoring mechanism that assists the researcher by allowing him to see the effect of the selection criteria on the choice of peptides, thereby enabling him to adjust his choices by putting higher emphasis on desirable criteria, and reducing unnecessary criteria. The current v-score is specific to the protein -- it is not meant to compare peptides across proteins, but only within a single protein. It is a reproducible value and is computed as the sum of the matrix entry times the corresponding weight:

$$vscore\_{P}=\sum\_{i}^{}M\_{P,i}×w\_{i}$$

where *vscoreP* is the score of the peptide *P*, *M* is the matrix of all peptides and criteria fulfillment (see Table 1) and *w* is the weighting (see Table 2).

***Testing and validation***

We asked a researcher with expertise in selecting peptides to evaluate 55 proteins. The average speed of the peptide selection by the expert was 8 proteins per day. In comparison, our software was able to compile the results for the same 50 proteins within around 1 hour , running on a local workstation with Xeon E5-1630 processor and 4 GB of memory. The manually selected peptides were almost always a subset of the peptides chosen by the software. Differences can be related to human error, disregarding some selection criteria (for instance information about natural variants from dbSNP data), intended toleration in choosing peptides (for instance not considering the existence of peptides in all isoforms), or the availability of new data that alter a choice made in the past (for instance adding new reviewed proteins in UniProt). The most important difference is however the order of the peptides. The software would order the peptide reflecting the weightings specified in Table 2 with most appropriate peptide on the top. The manually chosen peptides did not have any ordering in regard to which peptide is more/less favorable.

A key difference is inherent in our method because of its design as a scientific workflow -- it can be scaled up and run in parallel on multiple levels. Thus, not only can different inputs (proteins) be run in parallel, but also different tests and data/information retrieval steps of a single input. To test the performance of the tool we ran a list of the 254 proteins encoded by human chromosome 21 from UniProt build June 2013. Our tool reported peptides for 248 proteins out of 254 in 28 hours. 6 proteins failed because of errors in the online repositories we are dependent on (mainly ExPASy PeptideCutter failed in case of very short proteins and incomplete protein annotations in UniProt). Approximately 40 proteins were reported as having no matching peptides. For these proteins, a second iteration could have be performed with loosened selection criteria -- for instance, peptides containing 6-25 amino acids could have been selected instead of 7-20 amino acids. We believe that this 28 hours could be reduced by using a faster server and by running more than 20 requests in parallel. However, most of the processing time is spent in retrieving the results of the different queries sent to the external repositories.

A list of proteins of a specific chromosome, disease, or pathway can be done likewise in a similar way. To test the robustness of the tool and the scale-up ability, we ran a list of 1141 proteins encoded by human chromosome 6,

**CONCLUSIONS**

Here we present a method and a scientific workflow for generating a list of the most appropriate surrogate peptides for target proteins to be analyzed by LC/MRM-MS. Before the development of this software, generating these lists was a cumbersome process, in which experts retrieved information from different online repositories and used their own reasoning to find the most appropriate peptides. Our scientific workflow integrates information from different data sources including UniProt, Global Proteome Machine, NCBI’s dbSNP, and PeptideAtlas, but is not limited to these sources. Our software can be extended to consider additional data repositories and the user's own peptide libraries. The workflow can easily be adjusted to specific laboratory demands, such as choosing shorter peptides, putting emphasis on different selection criteria, or including more (or own instrument specific) data repositories, which makes the method general and useful for a wide audience. All data retrieved from online repositories are stored locally for a later verification or inspection of the reported results if needed. In order to make the software accessible to a wide audience, we have implemented a Web interface where the user can run the software from the browser.

The results of the scientific workflow are only as good as the data available in the repositories. Our tool in the current version does not handle malformed XML files generated by online repositories or erroneous HTML results generated by the Web tools used. In such situations, the tool can still be useful as an indicator of these cases, where advanced expertise is needed to manually select appropriate peptides for the proteins of interest. The scientific workflow and all code used in this work are available under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA) on www.proteincentre.com/bioinfomratics/MRMPeptidePicker/. Access to the online implementation of the tool is available at www.proteincentre.com/MRMPeptidePicker/.

**ASSOCIATED CONTENT**

**AUTHOR INFORMATION**

***Corresponding Authors***

\*Yassene Mohammed: Phone: +31-71-5268745. Fax: +31-71-5266907. E-mail: yassene@proteincentre.com or y.mohammed@lumc.nl.

\*Magnus Palmblad: Phone: +31-71-5269526. Fax: +31-71-5266907. E-mail: n.m.palmblad@lumc.nl

\*Christoph Borchers: Phone: +1-250-4833221. Fax: +1-250-483-3238 E-mail: christoph@proteincentre.com

***Competing Financial Interests***

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We are grateful to Genome Canada and Genome BC for providing Science and Technology Innovation Centre funding and support for the UVic-Genome BC Proteomics Centre. This work was also supported by the Dutch Organization for Scientific Research (De Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) grants NRG-2010.06 and VI-917.11.398.

 **REFERENCES**

1. Percy, A. J.; Chambers, A. G.; Yang, J.; Hardie, D.; Borchers, C. H., Advances in Multiplexed MRM-based Protein Biomarker Quantitation Toward Clinical Utility. *Biochimica et Biophysica Acta* **2013,** pii: S1570-9639(13)00239-2. .

2. UniProt\_Consortium, The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research* **2009,** 37, D169–D174.

3. Sherry, S. T.; Ward, M.; Sirotkin, K., dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research* **1999,** 9, (8), 677-9.

4. ExPASy\_Bioinformatics\_Portal Peptide Cutter. (July 2012),

5. PeptideAtlas, [www.peptideatlas.org](http://www.peptideatlas.org). **2010**.

6. The\_Global\_Proteome\_Machine\_Organization The Global Proteome Machine. (November 19, 2013),

7. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P., Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **2004,** 20, (17), 3045-3054.

8. Goecks, J.; Nekrutenko, A.; Taylor, J., Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **2010,** 11, (8), R86.

9. Maheshwari, K.; Montagnat, J. In *Scientific Workflow Development Using Both Visual and Script-Based Representation*, Services (SERVICES-1), 2010 6th World Congress on, 5-10 July 2010, 2010; 2010; pp 328-335.

10. Altintas, I.; Berkley, C.; Jaeger, E.; Jones, M.; Ludascher, B.; Mock, S., Kepler: An Extensible System for Design and Execution of Scientific Workflows. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, IEEE Computer Society: 2004; p 423.

11. Langley, G. J.; Herniman, J. M.; Davies, N. L.; Brown, R., Simplified sample preparation for the analysis of oligonucleotides by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **1999,** 13, 1717-1723.

12. Fusaro, V. A.; Mani, D. R.; Mesirov, J. P.; Carr, S. A., Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology* **2009,** 27, (2), 190-8.

13. Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R., Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology* **2007,** 12, (1), 125-31.

14. Sanders, W. S.; Bridges, S. M.; McCarthy, F. M.; Nanduri, B.; Burgess, S. C., Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics* **2007,** 1, (8 Suppl 7), S23.

15. Boja, E. S.; Rodriguez, H., Mass spectrometry-based targeted quantitative proteomics: Achieving sensitive and reproducible

detection of proteins. *Proteomics* **2012,** 12, 1093–1110.

16. Deutsch, E. W.; Lam, H.; Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Reports* **2008,** 9, 429–434.

17. PABST Peptide Atlas Best SRM Transition tool. (November 19, 2013),

18. Brusniak, M.-Y. K.; Kwok, S.-T.; Christiansen, M.; Campbell, D.; Reiter, L.; Picotti, P.; Kusebauch, U.; Ramos, H.; Deutsch, E. W.; Chen, J.; Moritz, R. L.; Aebersold, R., ATAQS: A computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. *BMC Bioinformatics* **2011,** 12, (78).

19. Mead, J. A.; Bianco, L.; Ottone, V.; Barton, C.; Kay, R. G.; Lilley, K. S.; Bond, N. J.; Bessant, C., MRMaid, the web-based tool for designing multiple reaction monitoring ( MRM ) transitions. *Molecular and Cellular Proteomics* **2009,** 8 (4), 696-705.

20. Walsh, G. M.; Lin, S.; Evans, D. M.; Khosrovi-Eghbal, A.; Beavis, R. C.; Kast, J., Implementation of a data repository-driven approach for targeted proteomics experiments by multiple reaction monitoring. *Journal of Proteomics* **2009,** 72, 838-852.

21. Institute\_for\_Systems\_Biology SRMAtlas.

22. TIQAM TIQAM (Targeted Identification for Quantitative Analysis by MRM). (November 19, 2013),

23. Skyline\_SRM/MRM\_Builder Skyline Targeted Proteomics Environment. (2011),

24. Sherwood, C.; Eastham, A.; Peterson, A.; Eng, J. K.; Shteynberg, D.; Mendoza, L.; Deutsch, E.; Risler, J.; Lee, L. W.; Tasman, N.; Aebersold, R. L., Henry ; Martin, D. B., MaRiMba: a Software Application for Spectral Library-Based MRM Transition List Assembly. *J. Proteome Res.* **2009,** 8, (10), 4396–4405.

25. Cham (Mead), J. A.; Bianco, L.; Bessant, C., Free computational resources for designing selected reaction monitoring transitions. *Proteomics* **2010,** 10, (6), 1106–1126.

26. Fan, J.; Mohareb, F.; Jones, A. M.; Bessant, C., MRMaid: The SRM Assay Design Tool for Arabidopsis and Other Species. *Frontiers in Plant Science* **2012,** 3, (164).

27. PRIDE PRIDE (Proteomics Identifications Database).

28. Fan, J.; Mohareb, F.; Bond, N. J.; Lilley, K. S.; Bessant, C., MRMaid 2.0: Mining PRIDE for Evidence-Based SRM Transitions. *OMICS: A Journal of Integrative Biology* **2012,** 16, (9), 483-488.

29. Vizcaíno, J. A.; Foster, J. M.; Martens, L., Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. *Journal of Proteomics* **2010,** 73, (11), 2136-46.

30. Bereman, M. S.; MacLean, B.; Tomazela, D. M.; Liebler, D. C.; MacCoss, M. J., The development of selected reaction monitoring methods for targeted proteomics via empirical refinement. *Proteomics* **2012,** 12, (8), 1134-41.

31. Arnold, K.; Gosling, J.; Holmes, D., *The Java Programming Language*. 4 ed.; Addison-Wesley Professional: 2005.

32. Urbanek, S. Rserve - Binary R server, [www.rforge.net/Rserve/index.html](http://www.rforge.net/Rserve/index.html). [www.rforge.net/Rserve/index.html](http://www.rforge.net/Rserve/index.html) (January 21, 2013),

33. Clark, J.; DeRose, S. XML Path Language (XPath). <http://www.w3.org/TR/xpath/>

34. Magrane, M.; Consortium, U., UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation* **2011,** 2011, bar009.

35. PeptideSieve PeptideSieve. (November 19, 2013),

36. STEPP SVM Technique for Evaluating Proteotypic Peptides (STEPP) (November 19, 2013),

37. Braisted, J. C.; Kuntumalla, S.; Vogel, C.; Marcotte, E. M.; Rodrigues, A. R.; Wang, R.; Huang, S.-T.; Ferlanti, E. S.; Saeed, A. I.; Fleischmann, R. D.; Peterson, S. N.; Pieper, R., The APEX Quantitative Proteomics Tool: Generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* **2008,** 9, 529.

38. Wedge, D. C.; Gaskell, S. J.; Hubbard, S. J.; Kell, D. B.; Lau, K. W.; Eyers, C. In *Peptide detectability following ESI mass spectrometry: prediction using genetic programming.*, 9th annual conference on Genetic and evolutionary computation (GECCO), London, 2007; Thierens, D.; Beyer, H.-G.; Bongard, J.; Branke, J.; Clark, J. A.; Cliff, D.; Congdon, C. B.; Deb, K.; Doerr, B.; Tim, K.; Kumar, S.; Miller, J. F.; Moore, J.; Neumann, F.; Pelikan, M.; Poli, R.; Sastry, K.; Stanley, K. O.; Stutzle, T.; Watson, R. A.; Wegener, I., Eds. ACM Press: London, 2007; pp 2219-2225.

39. Keil, B., *Specificity of proteolysis*. Springer-Verlag: Berlin-Heidelberg-NewYork, 1992; p 335.

40. Lai, M. C.; Topp, E. M., Solid-state chemical stability of proteins and peptides. *Journal of Pharmaceutical Science* **1999,** 88, (5), 489-500.

 **Table 1**: An example of the output of the scientific workflow shown in Figure 1.

****

**Figure Legends**

**Figure 1**: A logical workflow of peptide selection for an MRM experiment as implemented in PeptidePicker. The first step, after retrieving all proteins information and *in silico* digestion, is to filter all peptides according to the obligatory criteria. At the next level, all information about the peptides that is available from PeptideAtlas and Global Proteome Machine gpmDB, and the proteome of the organism as it appears in UniProt and the NCBI’s dbSNP are retrieved. The protein annotations available from UniProt are also reflected onto the peptides at this level, *i.e.,* determining which peptide contains a modification site, whether the peptide is in the signal/chain part of the protein or if it overlaps two sections that will be cleaved, etc.. Finally, all of the information is integrated into one table/spreadsheet and the final scoring is generated by penalizing unwanted criteria and rewarding desirable criteria.

**Figure 2**: A scientific workflow for the automatic generation of the most appropriate surrogate peptides to represent target proteins in LC/MRM-MS experiments. The user provides the protein accession number (or a list thereof) as input. Other inputs are also used to control the selection if the desired criteria are different from the default (see table 2). Some parameters need to be set once, such as the output directory or the path to the html2xml executable. *Uniprot\_RetrieveXML* retrieves the full description of the protein as an XML file from Uniprot; *extract\_name* reports the protein's recommended name, *find\_isoforms* indicates whether the protein has isoforms, and, if so, reports their accession numbers. Both of the latter processors are XPath parsers. The *find\_mod\_pos* algorithm is an R script that extracts the positions of any modifications present in the protein. *Peptidecutter* uses ExPASy online tool PeptideCutter to *in silico* digest the protein (and its isoforms, if required). The next four processors *html2xml, extract\_peptides, extract\_lpmm, extract\_lpmm2* extract the information reported by the PeptideCutter into a table that includes the peptides and their lengths, masses, positions within the proteins, and calculated digestion efficiencies. The *find\_peptides* algorithm filters the reported peptides according to the different acceptance criteria (see Peptide Acceptance Criteria section); *mark\_mod* indicates any possible modifications in the reported peptides from the step before; *intersect\_peptides* finds those peptides which are present in all isoforms when applicable. The two processors, *test\_observed\_GPM*, and *test\_observed\_PeptideAtlas*, test whether the reported peptides were previously observed and whether they have entries in the Global Proteome Machine or in PeptideAtlas. In addition, *test\_LabPeptideDataset* finds whether the peptides are in the laboratory's own dataset. The *test\_unique* algorithm reports whether the peptides is unique in the (human) proteome. The final step, *produce\_report*, summarizes all findings in a Comma Separated Values (CSV) report that can easily be imported or further processed by other software.

**Figure 3:** A screen shot of the Web interface of the tool. (**A)** shows the input page where the user needs only to insert the accession number of the protein of interest, and choose the organism. The user also controls the selection criteria by switching these on and off as indicated in the figure and in table 2. (**B)** shows the default results page where the peptides are listed in descending order, according to their v-score which integrates all of the retrieved information from the different repositories with the objective of choosing the appropriate peptides. The result page also includes links to the full extensive table of all information as html page for viewing in the browser or for downloading as CSV.

Figure 1



Figure 2



|  |  |
| --- | --- |
|  |  |

Figure 3



**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criterion** | **Type** | **Explanation** | **Weight**  |
| length\_”n”\_”m”\_AA | Mandatory with user input | The length of the peptides in the number of amino acids, where n is the lower and m in the higher limits. Both are integer and normally useful values are n=(6,7) and m=(20,25), but any values where n<m are allowed. | 1 |
| dig\_efficiency | Mandatory | The digestion efficiency is considered to be the product of both side digestion probability as obtained from ExPASy digestion tool Peptide Cutter. The tool can be replaced with other online or local tools  | 1/100 |
| no\_M | Mandatory | Is true, i.e. 1 if M is not present in the peptide | 1 |
| no\_C | Mandatory | Is true, i.e. 1 if C is not present in the peptide | 1 |
| no\_PP | User input | Is true, i.e. 1 if no P strings (2 or more) are present in the peptide | 1 |
| no\_N.term.Q | Optional | Is true, i.e. 1 if no Q is present on the N-terminal | 3 |
| no\_DG | Optional | Is true, i.e. 1 if DG is not present in the peptide | 2 |
| no\_DP | Optional | Is true, i.e. 1 if DP is not present in the peptide | 2 |
| no\_NG | Optional | Is true, i.e. 1 if NG is not present in the peptide | 2 |
| no\_QG | Optional | Is true, i.e. 1 if QG is not present in the peptide | 2 |
| no\_SSS | User input | Is true, i.e. 1 if no SSS strings (3 or more) are present in the peptide | 1 |
| no\_W | User input | Is true, i.e. 1 if W is not present in the peptide | 1 |
| no\_internal\_R\_K | User input | Is true if the peptide does not include internal R or K | 1 |
| no\_mc\_only\_P | User input | Is true, i.e. 1 if the peptide does not include R or K except when it is followed by P  | 1 |
| No\_mc\_Keil | User input | Is true, i.e. 1 if the peptide does not include missed cleavages according to Keil rules | 1 |
| no\_mod | Optional | Is true if the peptide does not include post -translational modifications | 4 |
| no\_near\_NCterm\_mod | Optional | Is true if the peptide does not include post translational modifications near the N or C terminal | 1 |
| in\_all\_iso | User input  | Is true of the peptide is present in all isoforms of the protein | 3 |
| unique | Mandatory | Is true if the peptide is observed only once in the proteome  | 5 |
| gpm | Optional | Is true if the peptide is previously observed in MS/MS experiments according to gpmDB  | 10 |
| peptideatlas | Optional | Is true if the peptide is previously observed in MS/MS experiments according to PeptideAtlas | 10 |
| prot\_in\_peptideatlas | Optional | Is true if the protein in in PeptideAtlas | 1 |
| PSS | Optional | Is the value of the PeptideAtlas PSS score if available, otherwise is “NA” aka 0 | 5 |
| ESS | Optional | Is the value of the PeptideAtlas ESS score if available, otherwise is “NA” aka 0 | 10 |
| PA\_n\_obs | Optional | The number of observations of the peptides as reported in PeptideAtlas | 10/ max(PA\_n\_obs) |
| het\_score | Optional | The value of the heterozygosity score as reported by dbSNP. Only peptides with heterozygosity value of less than 0.02 are considered. | -1 |
| het\_std\_err | Optional | The value of the standard error in measuring the heterozygosity score as reported by dbSNP | -1 |
| mod\_by\_similarity\_potential | Optional | Is true if the reported post translational modification has no experimental evidence | -1 |
| mod\_with\_evidence |  | Is true if the reported post translational modification has reported experimental evidence  | -4 |