PepExplorer aids in the biological interpretation of de novo sequencing results; this is accomplished by assembling a list of homolog proteins obtained by aligning results from widely adopted de novo sequencing tools against a target-decoy sequence database. Our tool relies on pattern recognition to ensure that the results satisfy a user-given false-discovery rate (FDR). For this, it employs a radial basis function neural network that considers the precursor charge states, de novo sequencing scores, the peptide lengths, and alignment scores. PepExplorer is recommended for studies addressing organisms with no genomic sequence available. PepExplorer is integrated into the PatternLab for proteomics environment, which makes available various tools for downstream data analysis, including the resources for quantitative and differential proteomics. © 2015 by John Wiley & Sons, Inc.

Keywords: de novo sequencing • mass spectrometry • proteomics

How to cite this article:

INTRODUCTION

Shotgun proteomics data analysis relies on several distinct techniques for interpreting data generated by mass spectrometers to ultimately provide a better understanding of biological systems. The gold-standard strategy, called peptide spectrum matching (PSM), relies on comparing experimental spectra against those theoretically generated from a sequence database (Eng et al., 1994). Yet, the absence of curated databases limits the application of PSM to unsequenced organisms. De novo is a technique for interpreting mass spectra that has roots in a different paradigm; it aims at predicting peptides by interpreting mass spectra without requiring a database. Therefore, de novo sequencing is commonly applied when there is a lack of genomic or proteomic databases and thus is complementary to PSM; it can also be used, say, for the discovery of new proteoforms. The output of de novo tools is limited to a list of peptides with confidence scores (Seidler et al., 2010) and therefore does not provide a direct interpretation at the protein level.

To overcome the aforementioned limitation, a class of software that performs BLAST-like analyses using sequences from closely related organisms and substitution matrices tailored to the task at hand emerged (Shevchenko et al., 2001; Junqueira et al., 2008). PepExplorer refines the latter by introducing a target-decoy similarity search followed by filtering results with a radial basis function neural network (RBF-NN) to satisfy a given false-discovery rate (FDR). The results are presented as a user-friendly and dynamic report that includes commonly used features found in proteomic studies such as protein
coverage and number of matching peptides. This qualifies PepExplorer as a software for statistically dealing with protein inference from peptide de novo sequence data and aiding in de novo interpretation at the protein level (Leprevost et al., 2014).

The following protocols describe the process of analyzing de novo sequencing results (post de novo analysis) with PepExplorer. We provide protocols for creating a database tailored to the study at hand, as well as to set the appropriate parameters and to perform data analysis and interpretation using the dynamic report.

PepExplorer is part of the software PatternLab for proteomics (Carvalho et al., 2012, UNIT 3.19), which is freely available at http://patternlabforproteomics.org.

**CREATING AND FORMATTING A DATABASE FOR PepExplorer ANALYSIS**

Creating a database for PepExplorer is the most critical step when doing post de novo analysis. Attention is important at this step regardless of whether the aim is to obtain identifications complementary to those reported by PSM or to unravel the proteome of a recently discovered organism. The key aspect is to select sequences from organisms phylogenetically related to the organism at hand. If the study addresses a specific class of proteins, such as, say, kinases, the database should contain different members of that family, from different organisms.

Protein sequences can be obtained from curated or non-curated databases. Curated databases are smaller and populated only with curated or reference sequences that usually have some functional information associated with them, which aids in determining the localization and function of proteins; NCBI RefSeq and SwissProt are examples of curated databases. Curated databases are the product of intensive labor, every registry being reviewed individually. Non-curated databases, on the other hand, are larger because not all registries go through a review process; this causes redundancy, as well as sequences originating from experimental errors (e.g., being sequenced together with the vector), and even sequences having a low sequence quality, which means they can be wrong. The positive aspect of using a non-curated database lies in the possibility of validating new variants and uncharacterized sequences. However, one should also bear in mind that searching against larger databases requires more computational resources and may result in sensitivity loss (Borges et al., 2013).

For more on dealing with sequence databases we refer the reader to Chapter 1 in this manual.

**Necessary Resources**

**Hardware**

A PC running an English version of Windows 7 or higher, with at least 4 GB of RAM and at least 100 GB of storage space

**Software**

Microsoft .NET framework 4.5 or later

PatternLab for proteomics, which can be freely downloaded from the project’s Web site (http://patternlabforproteomics.org)

1. The instructions to prepare a database for analysis are available in UNIT 3.19 (Carvalho et al., 2012; “PatternLab: From Mass Spectra to Label-Free Differential Shotgun Proteomics”), Basic Protocol 1, “Preparing a Sequence Database to be Searched by ProLuCID or the Academic SEQUEST.”
The aforementioned protocol details how to generate a target-decoy database containing reversed sequences plus 127 common contaminants for mass spectrometry.

ADJUSTING THE PARAMETERS

PepExplorer provides the user with a set of parameters that can be adjusted to optimize the analysis process (Fig. 13.27.1). These parameters are reviewed here.

Necessary Resources

Hardware
A PC running an English version of Windows 7 or higher, with at least 4 GB of RAM and at least 100 GB of storage space

Software
PatternLab for proteomics, which can be freely downloaded from the project’s Web site (http://patternlabforproteomics.org)

Files
A results file from one of the compatible de novo sequencing tools available: PepNovo (Frank and Pevzner, 2005), PNovo+ (Chi et al., 2010), or PEAKS (Ma et al., 2003).
A database generated as specified in Basic Protocol 1

1. To access PepExplorer, execute PatternLab, then click on the Filter drop-down menu and then on PepExplorer.
2. Click the Browse button on the Database field and then select the sequence database.

Figure 13.27.1 PepExplorer user interface. This is the main window where all analysis parameters are set.
3. Write the label used for tagging decoy sequences as in the database. If Basic Protocol 1 was followed, all reverse sequences will have a Reverse tag in their headers.

4. The Alignment section allows the setting of parameters related to sequence alignment. Their values determine how the alignment algorithm will penalize gaps, gap extensions, and non-matching amino acids. The default values (shown in Fig. 13.27.1) were obtained by extensively testing all combinations on a P. furiosus dataset and verifying the combination that yielded the optimal result. Nevertheless, the default combination is not necessarily optimal for the experiment at hand. We suggest also experimenting with more stringent values.

5. Parameters:
   a. Open Gap: The penalization for opening a gap in the alignment.
   b. Extend Gap: The penalization for extending a gap that was already open during sequence alignment.
   c. TopHits/Peptide size: How many predicted peptides from each spectrum are considered to be aligned and the peptide size itself, respectively.
   d. De novo score cutoff: Every de novo tool that is compatible with PepExplorer generates a list of peptide predictions, each having a de novo score reflecting a confidence. This parameter restricts the analysis only to de novo results with a score higher than the value set here.
   e. Matrix: The sequence alignment is scored according to the substitution matrix specified in this field. PepExplorer makes available the PAM30MS matrix, tailored to mass spectrometry data analysis (Shevchenko et al., 2001).

6. Database format: Select from which online sequence database the sequences were obtained: Generic FASTA, NCBI, UniProt, NextProt. The Generic FASTA format accepts headers that are composed by the Protein ID, followed by a blank space, and then by a description.

7. Click the Browse button in the de novo output section and specify the directory containing the de novo results files. Alternatively, paste the directory path in the text box.

   If the user only has a list of peptides not scored by any of the aforementioned de novo tools, this list can be pasted into the lower text box contained in this section.

8. Lastly, select the de novo sequencing software employed by using the corresponding pull-down menu.


   The time necessary to conclude the analysis depends on database size and on the number of de novo results to be analyzed. Once PepExplorer has finished, the Data Analysis window will pop open and the results should be saved.

GUIDELINES FOR UNDERSTANDING RESULTS

After the software completes analyzing the data, it will open a results window showing the analysis report (Fig. 13.27.2). That figure shows the dynamic report generated after the analysis is finished. The first tab, called Report, lists the identified proteins and their respective matching peptides. There are two groups of elements on this window, Filter and FDR. The former makes available parameters that affect the protein selection stringency and how the protein list will be organized. For example, if the alnCount counter is set to 2, only the proteins with more than 2 alignments will be shown. The MinFullIdent parameter specifies the minimum similarity that a de novo result must have with a sequence in the database to be included in the results.
Figure 13.27.2  The dynamic report shows inferred proteins based on the sequence alignment and the RBF-NN classification. By adjusting the parameters at the top of the window, it is possible to rearrange results to better fit the research.

Figure 13.27.3  Dynamic Report Window. The report shows the list of identified proteins and the matching peptides. When the inferred proteins are selected from the upper panel, the bottom list shows all the peptides that aligned against the protein. It also shows further information on the analysis, like the alignment size, the number of peptides, how many gaps it has, and so on.

The actual FDR value is shown in the second element. The values change every time the list is filtered with one of the controls on the left. It informs both the FDR for peptide analysis and that for protein analysis.

The results are presented in both the upper and the lower panel. The upper list provides information on the identifications: database protein length, the peptide coverage, the number of matched sequences, the number of spectra, and description. Clicking on any row will update the content of the bottom panel to reflect the peptide information of the selected protein (Fig. 13.27.3). Double-clicking on a protein in the upper list will
open the Protein Coverage and Group Analysis window (Fig. 13.27.4), which shows a graphical representation of the coverage and depicts, using text, which amino acids could be explained through de novo sequencing.

The coverage window (Fig. 13.27.5) also provides access to a cloud service that enables the dynamic prediction of the domains of the identified protein (Leprevost et al., 2013). This is accomplished by sending the sequence to our servers; on our end, the HMMR algorithm is executed (Punta et al., 2012). After completing the analysis, the window will reflect the domains pertaining to that protein.

The second tab, called Plotting (Fig. 13.27.6), was conceived as a visual way to view the topography and decision surface of the RBF-NN over the results. As the RBF-NN considers the de novo peptide sequence length, de novo score, and alignment scores, not all features are represented in the two-dimensional plot. The upper controls allow selecting which features should be considered as x axis and y axis in the plot.

The plotting area located in the upper part of the window represents the distribution of the alignments based on two features chosen out of three. We note that the analysis itself and the classification process take into account, simultaneously, all the features (i.e., de novo score, alignment score, and peptide length) of the alignments between the proteins and the peptide.

The plotting area and the points are composed of different tones of red or blue. A point is red if its corresponding de novo result aligned against a decoy sequence; it is blue otherwise. The background color tone is correlated to the RBF-NN output; regions of higher confidence are dark blue; they become red otherwise. The area where both colored points are found in approximately the same proportions generally becomes a “gray zone,” as it becomes less probable that the points will be correctly classified.

A histogram is displayed right below the plotting area; it is a complementary way of representing the results and helps in understanding how the RBF-NN converged. In
Figure 13.27.5  Domain Inference Window. This window allows the user to use an on-line service to predict functional domains in proteins. Also, allows to visualize the protein coverage regarding the matching peptides. The blue bar represents the identified protein. The orange bars represent regions of the protein that can be confirmed with identified peptides.

Figure 13.27.6  RBF-NN classification viewer. This window provides the user with a visual form of representing the classification process executed by the neural network. The plotting area shows a two-dimensional representation of the RBF-NN decision surface. Blue dots represent de novo results that aligned against target proteins; likewise, red represent those that aligned against decoys. The background color intensity is correlated with the RBF-NN confidence; the stronger the blue, the higher the confidence. The lower panel displays a histogram of the distribution of scores for the "red" and "blue" classifications.
brief, the x axis represents the RBF-NN output value and the y axis the number of de novo results satisfying that score. Likewise, blue bars represent results that aligned with forward sequences, and red bars represent decoy alignments. If an FDR of, say, 1% is specified by the user, the algorithm will establish an RBF-NN cutoff score in such a way that the summed area of the red bars represent 1% of the total area.

**COMMENTARY**

**Background Information**

Most of what we know of biology has been derived from the study of cell lines and model organisms. Even so, there are many examples of groundbreaking discoveries resulting from the study of species that, until then, were given little scientific importance. For example, who would have imagined that a study of organisms living in thermal waters could become a precursor to a revolution in molecular biology? Indeed, the Taq polymerase was discovered while studying *Thermus aquaticus* (Saiki et al., 1985). This thermostable DNA polymerase enzyme paved the way for Kary Mullis to later introduce the Polymerase Chain Reaction (PCR) method, capable of amplifying short segments of DNA (Bartlett and Stirling, 2003), a discovery that would lead to his being awarded the 1993 Nobel Prize in Chemistry.

The mainstream in proteomics has been to compare experimental spectra against theoretical spectra generated from a sequence database (Eng et al., 1994). Nevertheless, only a small fraction of the species have Reference Sequences deposited in the NCBI database (http://www.ncbi.nlm.nih.gov/refseq/). There is a consensus that natural biological resources must be better assessed: the exploratory space for groundbreaking discoveries beyond existing model organisms and cell lines is virtually unlimited, but the proper tools to mine this proteomosphere are in their infancy (Junqueira and Carvalho, 2012). Continual advances in mass spectrometry, with yearly releases of more sensitive, faster, and higher-resolution equipment, make de novo sequencing more and more of a mainstream approach. PepExplorer aims to shorten the gap in efficiency between the proteomic interpretation of unsequenced organisms and the widely adopted PSM approach, notwithstanding the many limitations that still exist. For example, in our view, there is still much room for improvement on how sequence alignment is done; while current methods excel for genomic data, they do not take into account the shortcomings of mass-spectrometry data, such as the possibility of de novo sequencing swapping neighboring amino acids when spectral peaks are missing.

Other applications in which PepExplorer can be particularly useful are those addressing samples with many mutations or polymorphisms (e.g., venoms), as well as those in metaproteomics (Muth et al., 2015). Currently, there is no one-fits-all approach for handling such problems, and there are approaches that are complementary to PepExplorer. For example, blind modification searches, in our hands, have provided complementary results to PepExplorer (Na et al., 2012); this type of approach stems from a different paradigm where a sequence database is used for matching mass spectra, but in a much “looser” way. Spectral networking is also a powerful approach for treating unsequenced organisms; this strategy capitalizes on overlapping spectra to improve on de novo sequencing and then perform alignment against a sequence database. Blind-PTM approaches and spectral networks allow the discovery of unanticipated PTMs (a limitation of PepExplorer) and mutations/polymorphisms (Bandeira, 2007). However, in our hands, PepExplorer has proven to be more sensitive, but still complementary, to blind-PTM approaches when not considering PTMs. The proper, combined use of this arsenal of tools, applied to the vast biological diversity, holds the keys to unlocking a vast treasure trove of knowledge in biology.

**Troubleshooting**

**My analysis is taking too long**

The current version of PepExplorer places a high burden on the computer. Smaller databases significantly speed up the search. PepExplorer can use multiple computing cores, so we strongly recommend using computers with as many cores as possible. Depending on the available hardware, the process could take hours or even days. We are currently improving on the code to soon disclose a faster PepExplorer.

**I’m getting very few proteins in my results**

Peptide de novo sequencing is a technique that strongly depends on high data quality. For example, high-resolution tandem mass spectra should significantly improve results. Also,
double check if the organism selected for generating the sequence database is “close” to the one being analyzed.

The domain prediction service (FiCloud) is not working or it is taking too long

The domain prediction service relies on a Web server that must be contacted via the Internet. It is possible for the communication with that server to be compromised by different factors, like no connection to the Internet or some firewall blockade. If there are problems using this feature, we recommend trying it later or from another Internet connection.

Acknowledgement

The authors acknowledge CNPq, FAPERJ, CAPES (Grant 063/2010–Edital Toxinologia), and PDTIS and PAPES for financial support. We acknowledge Prof. John R. Yates, III, for participating in PepExplorer project.

Literature Cited


