Using PepExplorer to Filter and Organize *De Novo* Peptide Sequencing Results

Felipe da Veiga Leprevost,[1,2] Valmir C. Barbosa,[3] and Paulo Costa Carvalho[1]

[1]Computational Mass Spectrometry Group, Carlos Chagas Institute–Fiocruz. Curitiba, Paraná, Brazil
[2]Department of Pathology, University of Michigan, Ann Arbor, Michigan
[3]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

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:
da Veiga Leprevost, F., Barbosa, V.C., and Carvalho, P.C. 2015. Using PepExplorer to filter and organize *de novo* peptide sequencing results. *Curr. Protoc. Bioinform.* 51:13.27.1-13.27.9.
doi: 10.1002/0471250953.bi1327s51

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

Current Protocols in Bioinformatics 13.27.1-13.27.9, September 2015
Published online September 2015 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/0471250953.bi1327s51

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*.

BASIC PROTOCOL 1

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

BASIC PROTOCOL 2

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

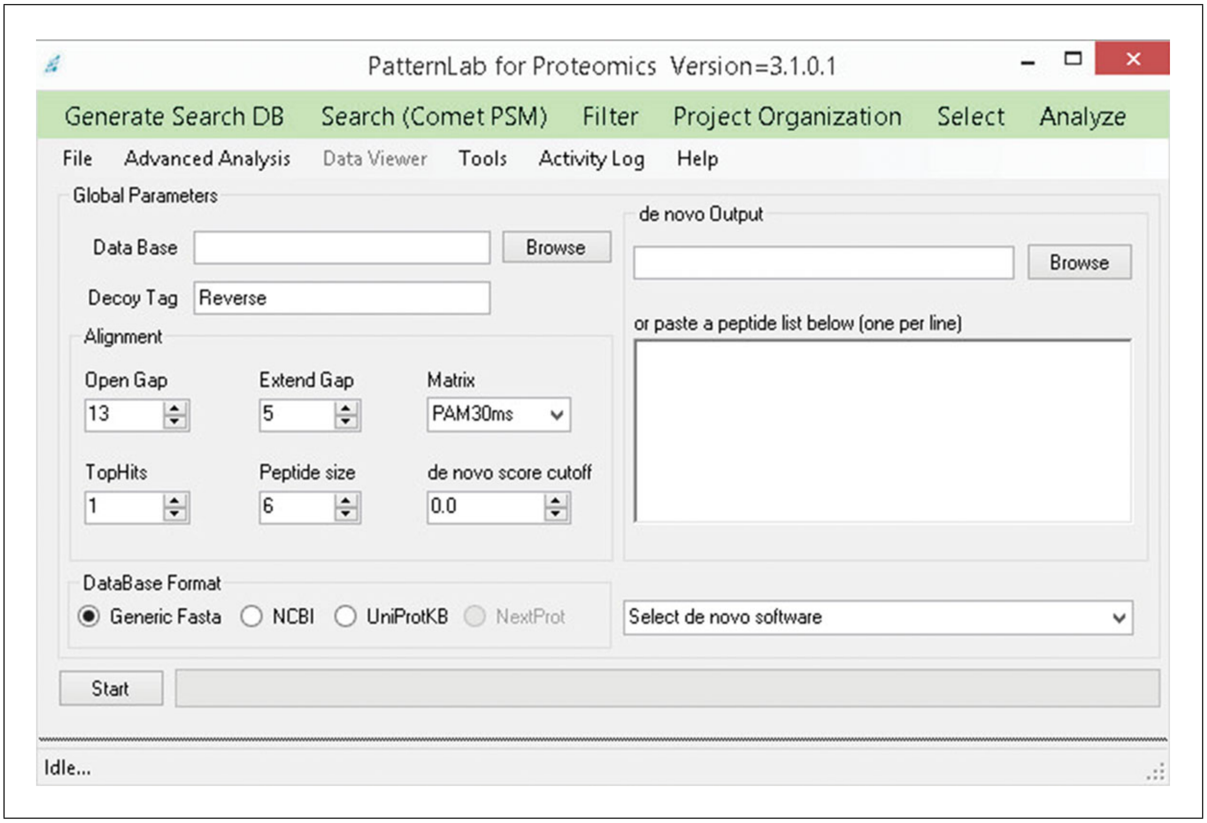
PatternLab for proteomics, which can be freely downloaded from the project's Web site (*http://patternlabforproteomics.org*)

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.



Data Base

Decoy Tag Reverse

Open Gap 13

Extend Gap 5

Matrix PAM30ms

TopHits 1

Peptide size 6

de novo score cutoff 0.0

● Generic Fasta ○ NCBI ○ UniProtKB ○ NextProt

or paste a peptide list below (one per line)

Select de novo software

Idle...

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.

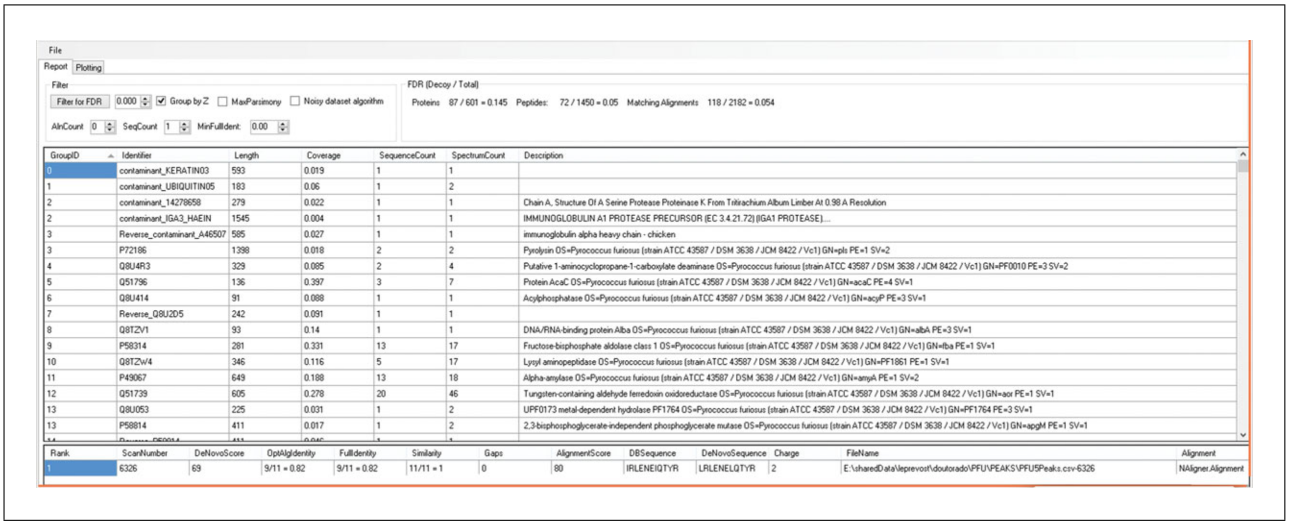
8. Lastly, select the *de novo* sequencing software employed by using the corresponding pull-down menu.

9. Press Start.

   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.



0.000 ☑ Group by Z ☐ MaxParsimony ☐ Noisy dataset algorithm

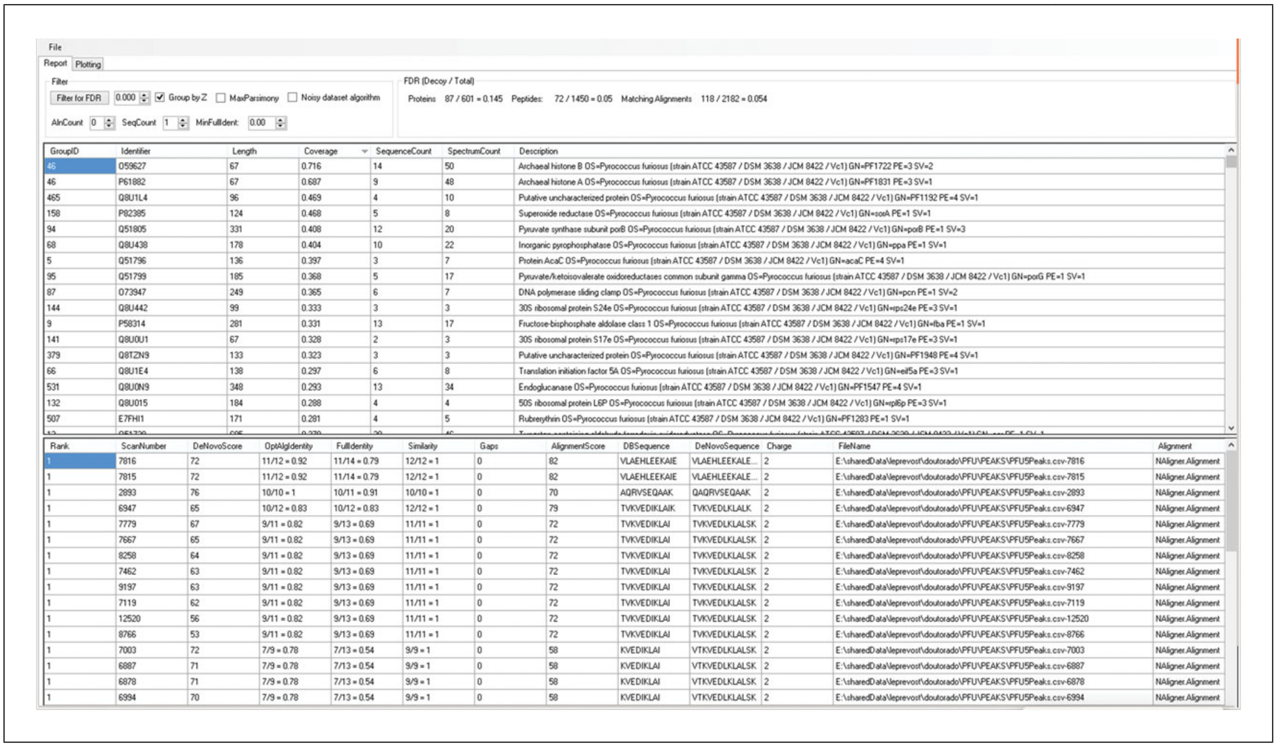
AlnCount 0 SeqCount 1 MinFullIdent: 0.00

Proteins 87 / 601 = 0.145 Peptides: 72 / 1450 = 0.05 Matching Alignments 118 / 2182 = 0.054

GroupID	Identifier	Length	Coverage	SequenceCount	SpectrumCount	Description
0	contaminant_KERATIN03	593	0.019	1	1	
1	contaminant_UBIQUITIN05	183	0.06	1	2	
2	contaminant_14278658	279	0.022	1	1	Chain A, Structure Of A Serine Protease Proteinase K From Tritirachium Album Limber At 0.98 A Resolution
2	contaminant_IGA3_HAEIN	1545	0.004	1	1	IMMUNOGLOBULIN A1 PROTEASE PRECURSOR (EC 3.4.21.72) (IGA1 PROTEASE)....
3	Reverse_contaminant_A46507	585	0.027	1	1	immunoglobulin alpha heavy chain - chicken
3	P72186	1398	0.018	2	2	Pyrolysin OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=pls PE=1 SV=2
4	Q8U4R3	329	0.085	2	4	Putative 1-aminocyclopropane-1-carboxylate deaminase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF0010 PE=3 SV=2
5	Q51796	136	0.397	3	7	Protein AcaC OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=acaC PE=4 SV=1
6	Q8U414	91	0.088	1	1	Acylphosphatase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=acyP PE=3 SV=1
7	Reverse_Q8U2D5	242	0.091	1	1	
8	Q8TZV1	93	0.14	1	1	DNA/RNA-binding protein Alba OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=albA PE=3 SV=1
9	P58314	281	0.331	13	17	Fructose-bisphosphate aldolase class 1 OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=fba PE=1 SV=1
10	Q8TZW4	346	0.116	5	17	Lysyl aminopeptidase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1861 PE=1 SV=1
11	P49067	649	0.188	13	18	Alpha-amylase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=amyA PE=1 SV=2
12	Q51739	605	0.278	20	46	Tungsten-containing aldehyde ferredoxin oxidoreductase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=aor PE=1 SV=1
13	Q8U053	225	0.031	1	2	UPF0173 metal-dependent hydrolase PF1764 OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1764 PE=3 SV=1
13	P58814	411	0.017	1	2	2,3-bisphosphoglycerate-independent phosphoglycerate mutase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=apgM PE=1 SV=1

Rank	ScanNumber	DeNovoScore	OptAlgIdentity	FullIdentity	Similarity	Gaps	AlignmentScore	DBSequence	DeNovoSequence	Charge	FileName	Alignment
1	6326	69	9/11 = 0.82	9/11 = 0.82	11/11 = 1	0	80	IRLENEIQTYR	LRLENELQTYR	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-6326	NAligner.Alignment

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.



0.000 ☑ Group by Z ☐ MaxParsimony ☐ Noisy dataset algorithm

AlnCount 0 SeqCount 1 MinFullIdent: 0.00

Proteins 87 / 601 = 0.145 Peptides: 72 / 1450 = 0.05 Matching Alignments 118 / 2182 = 0.054

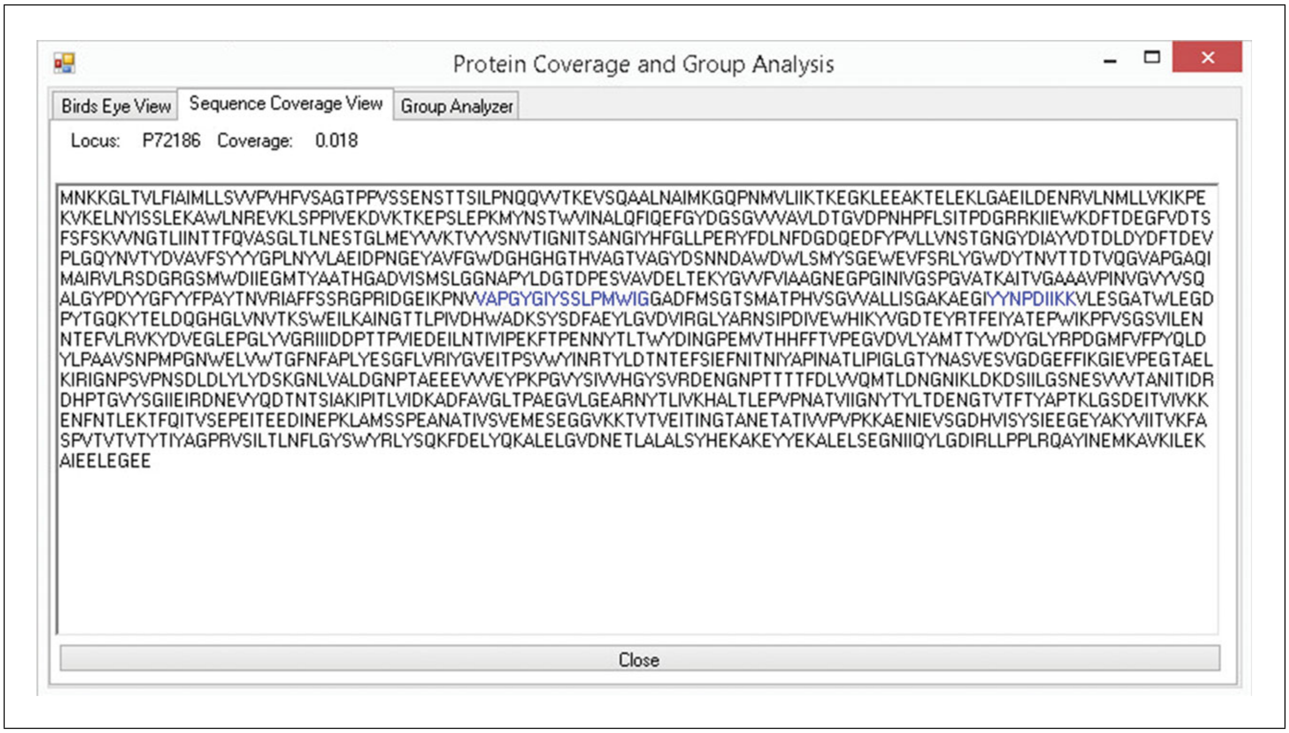
GroupID	Identifier	Length	Coverage	SequenceCount	SpectrumCount	Description
46	O59627	67	0.716	14	50	Archaeal histone B OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1722 PE=3 SV=2
46	P61882	67	0.687	9	48	Archaeal histone A OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1831 PE=3 SV=1
465	Q8U1L4	96	0.469	4	10	Putative uncharacterized protein OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1192 PE=4 SV=1
158	P82385	124	0.468	5	8	Superoxide reductase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=sorA PE=1 SV=1
94	Q51805	331	0.408	12	20	Pyruvate synthase subunit porB OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=porB PE=1 SV=3
68	Q8U438	178	0.404	10	22	Inorganic pyrophosphatase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=ppa PE=1 SV=1
5	Q51796	136	0.397	3	7	Protein AcaC OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=acaC PE=4 SV=1
95	Q51799	185	0.368	5	17	Pyruvate/ketoisovalerate oxidoreductases common subunit gamma OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=porG PE=1 SV=1
87	O73947	249	0.365	6	7	DNA polymerase sliding clamp OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=pcn PE=1 SV=2
144	Q8U442	99	0.333	3	3	30S ribosomal protein S24e OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=rps24e PE=3 SV=1
9	P58314	281	0.331	13	17	Fructose-bisphosphate aldolase class 1 OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=fba PE=1 SV=1
141	Q8U0U1	67	0.328	2	3	30S ribosomal protein S17e OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=rps17e PE=3 SV=1
379	Q8TZN9	133	0.323	3	3	Putative uncharacterized protein OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1948 PE=4 SV=1
66	Q8U1E4	138	0.297	6	8	Translation initiation factor 5A OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=eif5a PE=3 SV=1
531	Q8U0N9	348	0.293	13	34	Endoglucanase OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1547 PE=4 SV=1
132	Q8U015	184	0.288	4	4	50S ribosomal protein L6P OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=rpl6p PE=3 SV=1
507	E7FHI1	171	0.281	4	5	Rubrerythrin OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF1283 PE=1 SV=1

Rank	ScanNumber	DeNovoScore	OptAlgIdentity	FullIdentity	Similarity	Gaps	AlignmentScore	DBSequence	DeNovoSequence	Charge	FileName	Alignment
1	7816	72	11/12 = 0.92	11/14 = 0.79	12/12 = 1	0	82	VLAEHLEEKAIE	VLAEHLEEKALE...	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7816	NAligner.Alignment
1	7815	72	11/12 = 0.92	11/14 = 0.79	12/12 = 1	0	82	VLAEHLEEKAIE	VLAEHLEEKALE...	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7815	NAligner.Alignment
1	2893	76	10/10 = 1	10/11 = 0.91	10/10 = 1	0	70	AQRVSEQAAK	QAQRVSEQAAK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-2893	NAligner.Alignment
1	6947	65	10/12 = 0.83	10/12 = 0.83	12/12 = 1	0	79	TVKVEDIKLAIK	TVKVEDLKLALK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-6947	NAligner.Alignment
1	7779	67	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7779	NAligner.Alignment
1	7667	65	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7667	NAligner.Alignment
1	8258	64	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-8258	NAligner.Alignment
1	7462	63	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7462	NAligner.Alignment
1	9197	63	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-9197	NAligner.Alignment
1	7119	62	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7119	NAligner.Alignment
1	12520	56	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-12520	NAligner.Alignment
1	8766	53	9/11 = 0.82	9/13 = 0.69	11/11 = 1	0	72	TVKVEDIKLAI	TVKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-8766	NAligner.Alignment
1	7003	72	7/9 = 0.78	7/13 = 0.54	9/9 = 1	0	58	KVEDIKLAI	VTKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-7003	NAligner.Alignment
1	6887	71	7/9 = 0.78	7/13 = 0.54	9/9 = 1	0	58	KVEDIKLAI	VTKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-6887	NAligner.Alignment
1	6878	71	7/9 = 0.78	7/13 = 0.54	9/9 = 1	0	58	KVEDIKLAI	VTKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-6878	NAligner.Alignment
1	6994	70	7/9 = 0.78	7/13 = 0.54	9/9 = 1	0	58	KVEDIKLAI	VTKVEDLKLALSK	2	E:\sharedData\leprevost\doutorado\PFU\PEAKS\PFU5Peaks.csv-6994	NAligner.Alignment

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



Locus: P72186 Coverage: 0.018

```
MNKKGLTVLFIAIMLLSVVPVHFVSAGTPPVSSENSTTSILPNQQVVTKEVSQAALNAIMKGQPNMVLIIKTKEGKLEEAKTELEKLGAEILDENRVLNMLLVKIKPE
KVKELNYISSLEKAWLNREVKLSPPIVEKDVKTKEPSLEPKMYNSTWVINALQFIQEFGYDGSGVVVAVLDTGVDPNHPFLSITPDGRRKIIEWKDFTDEGFVDTS
FSFSKVVNGTLIINTTFQVASGLTLNESTGLMEYVVKTVYVSNVTIGNITSANGIYHFGLLPERYFDLNFDGDQEDFYPVLLVNSTGNGYDIAYVDTDLDYDFTDEV
PLGQYNVTYDVAVFSYYYGPLNYVLAEIDPNGEYAVFGWDGHGHGTHVAGTVAGYDSNNDAWDWLSMYSGEWEVFSRLYGWDYTNVTTDTVQGVAPGAQI
MAIRVLRSDGRGSMWDIIEGMTYAATHGADVISMSLGGNAPYLDGTDPESVAVDELTEKYGVVFVIAAGNEGPGINIVGSPGVATKAITVGAAAVPINVGVYVSQ
ALGYPDYYGFYYFPAYTNVRIAFFSSRGPRIDGEIKPNVVAPGYGIYSSLPMWIGGADFMSGTSMATPHVSGVVALLISGAKAEGIYYNPDIIKKVLESGATWLEGD
PYTGQKYTELDQGHGLVNVTKSWEILKAINGTTLPIVDHWADKSYSDFAEYLGVDVIRGLYARNSIPDIVEWHIKYVGDTEYRTFEIYATEPWIKPFVSGSVILEN
NTEFVLRVKYDVEGLEPGLYVGRIIIDDPTTPVIEDEILNTIVIPEKFTPENNYTLTWYDINGPEMVTHHFFTVPEGVDVLYAMTTYWDYGLYRPDGMFVFPYQLD
YLPAAVSNPMPGNWELVWTGFNFAPLYESGFLVRIYGVEITPSVWYINRTYLDTNTEFSIEFNITNIYAPINATLIPIGLGTYNASVESVGDGEFFIKGIEVPEGTAEL
KIRIGNPSVPNSDLDLYLYDSKGNLVALDGNPTAEEEVVVEYPKPGVYSIVVHGYSVRDENGNPTTTTFDLVVQMTLDNGNIKLDKDSIILGSNESVVVTANITIDR
DHPTGVYSGIIEIRDNEVYQDTNTSIAKIPITLVIDKADFAVGLTPAEGVLGEARNYTLIVKHALTLEPVPNATVIIGNYTYLTDENGTVTFTYAPTKLGSDEITVIVKK
ENFNTLEKTFQITVSEPEITEEDINEPKLAMSSPEANATIVSVEMESEGGVKKTVTVEITINGTANETATIVVPVPKKAENIEVSGDHVISYSIEEGEYAKYVIITVKFA
SPVTVTVTYTIYAGPRVSILTLNFLGYSWYRLYSQKFDELYQKALELGVDNETLALALSYHEKAKEYYEKALELSEGNIIQYLGDIRLLPPLRQAYINEMKAVKILEK
AIEELEGEE
```

Figure 13.27.4 Protein Coverage and Group Analysis window. Text-based representation of the protein coverage according to the identified peptides.

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.

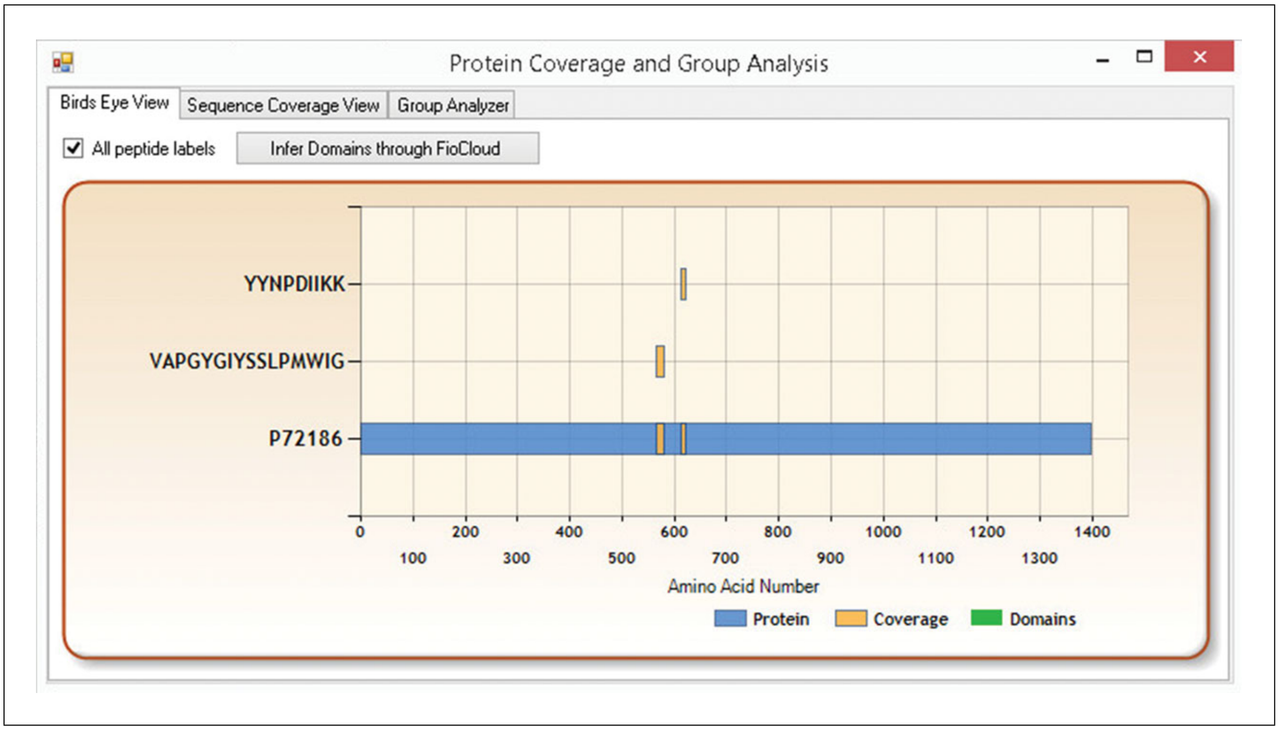
The coverage window (Fig. 13.27.5) also provides access to a cloud service that enables the dynamic prediction of the domains of the identified protein (Leprevost et al., 2013). This is accomplished by sending the sequence to our servers; on our end, the HMMR algorithm is executed (Punta et al., 2012).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

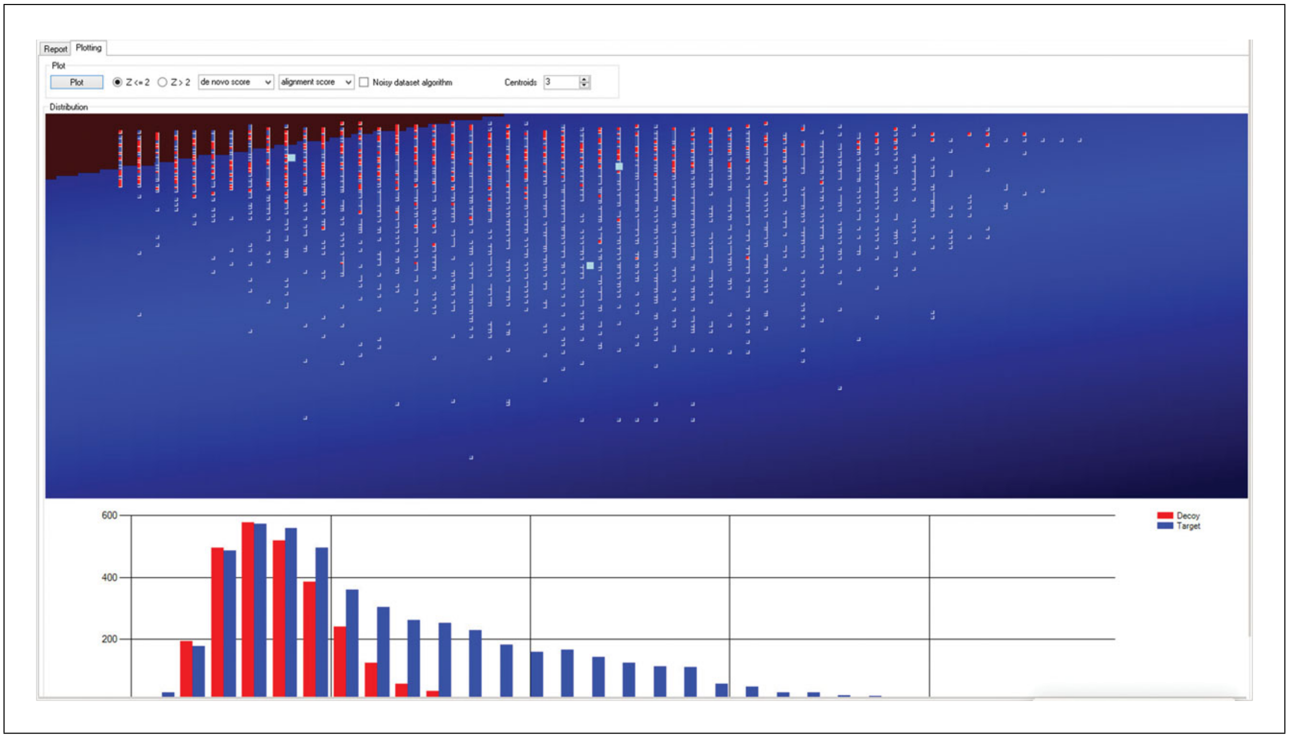
The plotting area and the points are composed of different tones of red or blue. A point is red if its corresponding *de novo* result aligned against a decoy sequence; it is blue otherwise. The background color tone is correlated to the RBF-NN output; regions of higher confidence are dark blue; they 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



☑ All peptide labels

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.



● Z <= 2 ○ Z > 2 de novo score alignment score □ Noisy dataset algorithm Centroids 3

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 (Fi°Cloud) 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

Bandeira, N. 2007. Spectral networks: A new approach to *de novo* discovery of protein sequences and posttranslational modifications. *BioTechniques* 42:687, 689, 691 passim.

Bartlett, J.M.S. and Stirling, D. 2003. A Short history of the polymerase chain reaction. *In* PCR Protocols pp. 3-6. Humana Press, Totowa, N.J. Available at *http://link.springer.com/10.1385/1-59259-384-4:3* [Accessed April 16, 2015].

Borges, D., Perez-Riverol, Y., Nogueira, F.C.S., Domont, G.B., Noda, J., da Veiga Leprevost, F., Besada, V., França, F.M.G., Barbosa, V.C., Sánchez, A., and Carvalho, P.C. 2013. Effectively addressing complex proteomic search spaces with peptide spectrum matching. *Bioinformatics* 29:1343-1344.

Carvalho, P. C., Fischer, J. S. G., Xu, T., Yates, J. R. and Barbosa, V. C. 2012. PatternLab: From mass spectra to label-free differential shotgun proteomics. *Curr. Protoc. Bioinform.* 40:13.19.1-13.19.18.

Chi, H., Sun, R.-X., Yang, B., Song, C.-Q., Wang, L.-H., Liu, C., Fu, Y., Yuan, Z.-F., Wang, H.-P., He, S.-M., and Dong, M.-Q. 2010. pNovo: *De novo* peptide sequencing and identification using HCD spectra. *J. Proteome Res.* 9:2713-2724.

Eng, J.K., McCormack, A.L., and Yates, J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*5:976-989.

Frank, A. and Pevzner, P. 2005. PepNovo: *De novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* 77:964-973.

Junqueira, M. and Carvalho, P.C. 2012. Tools and challenges for diversity-driven proteomics in Brazil. *Proteomics* 12:2601-2606.

Junqueira, M., Spirin, V., Balbuena, T.S., Thomas, H., Adzhubei, I., Sunyaev, S., and Shevchenko, A. 2008. Protein identification pipeline for the homology-driven proteomics. *J. Proteomics* 71:346-356.

Leprevost, F.V., Lima, D.B., Crestani, J., Perez-Riverol, Y., Zanchin, N., Barbosa, V.C., and Carvalho, P.C. 2013. Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics. *J. Proteomics* 89:179-182.

Leprevost, F.V., Valente, R.H., Lima, D.B., Perales, J., Melani, R., Yates, J.R., Barbosa, V.C., Junqueira, M., and Carvalho, P.C. 2014. PepExplorer: A similarity-driven tool for analyzing *de novo* sequencing results. *Mol. Cell Proteomics* 13:2480-2489.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. 2003. PEAKS: Powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17:2337-2342.

Muth, T., Kolmeder, C.A., Salojärvi, J., Keskitalo, S., Varjosalo, M., Verdam, F.J., Rensen, S.S., Reichl, U., de Vos, W.M., Rapp, E., and Martens, L. 2015. Navigating through metaproteomics data: A logbook of database searching. *Proteomics* [Epub ahead of print].

Na, S., Bandeira, N., and Paek, E. 2012. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell Proteomics* 11:M111.010199.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., and Finn, R.D. 2012. The Pfam protein families database. *Nucleic Acids Res.* 40:D290-D301.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)* 230:1350-1354.

Seidler, J., Zinn, N., Boehm, M.E., and Lehmann, W.D. 2010. *De novo* sequencing of peptides by MS/MS. *Proteomics* 10:634-649.

Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K.G. 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 73:1917-1926.