PepExplorer: a similarity-driven tool for analyzing *de novo* sequencing results

Felipe V Leprevost[1], Richard H Valente[2,3], Diogo L Borges[1], Jonas Perales[2,3], Rafael Melani[4], John R Yates III[5], Valmir C Barbosa[6], Magno Junqueira[4], Paulo C Carvalho[1*]

1. Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil;
2. Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil;
3. Instituto Nacional de Ciência e Tecnologia em Toxinas (INCTTox / CNPq), Brazil
4. Proteomics Unit, Rio de Janeiro Proteomics Network, Department of Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil;
5. Department of Chemical Physiology, The Scripps Research Institute, La Jolla, CA, United States;
6. Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil;

*Corresponding author

Running Title
PepExplorer: *de novo* sequencing post-processing tool.

KEYWORDS
PepExplorer; similarity-driven tool; *de novo* sequencing; protein inference; bioinfomatics; *Bothrops jararaca*; plasma.

ABSTRACT

Peptide Spectrum Matching (PSM) is the current gold standard for protein identification by mass spectrometry-based proteomics. PSM compares experimental mass spectra against theoretical spectra generated from a protein sequence database to perform identification, but protein sequences not present in a database can not be identified unless their sequences are in part conserved. The alternative approach, *de novo* sequencing, can infer a peptide sequence directly from a mass spectrum, but interpreting long lists of peptide sequences resulting from large-scale experiments is not trivial. With this as motivation, PepExplorer was developed to use rigorous pattern recognition to assemble a list of homologue proteins using *de novo* sequencing data coupled to sequence alignment to allow biological interpretation of the data. PepExplorer can read the output of various widely adopted *de novo* sequencing tools and converge to a list of proteins with a global false-discovery rate (FDR). To this end, it employs a radial basis function neural network that considers precursor charge states, *de novo* sequencing scores, peptide lengths, and alignment scores to select similar protein candidates, from a target-decoy database, usually obtained from phylogenetically related species. Alignments are performed using a modified Smith-Waterman algorithm tailored for the task at hand. We have verified the effectiveness of our approach on a reference set of identifications generated by ProLuCID when searching for *Pyrococcus furiosus* mass spectra on the corresponding NCBI RefSeq database. We then modified the sequence database by swapping amino acids until ProLuCID was no longer capable of identifying any proteins. By searching the mass spectra using PepExplorer on the modified database, we have been able to recover most of the identifications at a 1% FDR. Finally, we have employed PepExplorer to disclose a comprehensive proteomic assessment of the *Bothrops jararaca* plasma, a known biological source of natural inhibitors of snake toxins. PepExplorer is integrated into the *PatternLab for Proteomics* environment, which makes available various tools for downstream data analysis, including resources for quantitative and differential proteomics.

INTRODUCTION

Very often, groundbreaking discoveries with a significant impact on the biotechnological and biomedical fields have emerged from studying "non-canonical" organisms. For example, the study of *Thermus aquaticus* allowed us to ultimately pave the way to modern molecular biology with the characterization of that organism's thermostable DNA polymerase (1). The characterization of the Green Fluorescent Protein (GFP) in *Aequoria victoria* led to a revolution in cellular biology and to a Nobel Prize being awarded to Osamu Shimomura, Martin Chalfie, and Roger Tsien. In Brazil, Sergio Ferreira's work on the venom of the Brazilian poisonous snake *Bothrops jararaca* enabled the development of the first angiotensin-converting enzyme inhibitor drug (Captopril) for the treatment of hypertension (2).

In scenarios such as these, proteomics has the prospective to allow a better understanding of the complexity of biological systems and the process of evolution than the study of the genetic code alone. It enables the characterization of molecular processes according to their protein content, facilitating new discoveries. In proteomics, the most frequently used strategy for protein identification is the so-called Peptide Spectrum Matching (PSM), i.e., comparing experimental mass spectra obtained by fragmenting peptides in a mass spectrometer to theoretical spectra generated from a sequence database. In general, the identification process follows from the sequence whose theoretical spectrum yields the highest matching score according to some empirical or probabilistic function. Examples of search engines adopting this strategy are SEQUEST (3), X!Tandem (4), and Mascot (5).

Back in the 1990's, establishing a cutoff score for confident identification relied mostly on user experience; for example, given a specific charge state, Washburn et al established cross-correlation and deltaCn cutoff values for SEQUEST in order to allow the selection of a subset of confident identifications from LCQ data; this has since been termed "the Washburn criterion". In what followed, target-decoy databases were implemented to allow for more sophisticated refinements in filtering the data (6). In 2007, Elias and Gygi published a seminal paper on the target-decoy approach to shotgun proteomics (7) which

Page: 3

ultimately firmed this approach as a standard and motivated the development of several statistical filters capable of converging to a list of confident identifications satisfying a user-specified False-Discovery Rate (FDR) with significantly more sensitivity than the conservative Washburn criterion. Such statistical filters include mixtures of probabilities (8), quadratic discriminant analysis (9), semi-supervised learning with support vector machines (10), and Bayesian logic (11) using a semi-labeled decoy analysis to account for overfitting (12). With so many advances, the PSM work-flow has become the gold standard as, it is a very sensitive and least error-prone method when a database is available with the corresponding proteins. The latter limits the application of PSM to those organisms for which accurate sequence databases were previously established. If a peptide's sequence is not contained within the sequence database, it cannot be identified by the PSM method. However, efforts in developing error-tolerant PSM approaches such as implemented in Mascot have made it possible to handle minor sequence modifications constrained by a simle set of rules. Nevertheless, increasing the search space in the PSM approach leads to loss in sensitivity (13).

Even though the concept of computer-aided *de novo* sequencing predates that of PSM (14), advances in the quality of mass spectrometry data and the power of computer hardware have allowed it to re-emerge at the heart of a highly active field. *De novo* sequencing is unbiased insofar as it is not constrained by a sequence database, being therefore complementary to PSM. On the other hand, it has remained the most error-prone of the two methods (15). Notwithstanding the challenges of *de novo* sequencing, a few recent and notable improvements in computer-aided *de novo* analysis are PepNovo (16), which combines graph theory with machine learning, pNovo+ (17), which is optimized for high-resolution HCD data, NovoHMM (18), relying on Hidden Markov Models for increased sensitivity, and PEAKS (19), which uses a spectrum graph model by performing dynamic programming on the mass values regardless of the presence of an observed fragment ion. By considering the complementarities of different fragmentation strategies [e.g., Collision Induced Dissociation (CID), Electron Transfer Dissociation (ETD) (20), and

Page: 4

Electron Capture Dissociation (ECD) (21)], computational proteomic scientists have also demonstrated significant advances in *de novo* accuracy (22). In particular, the Bandeira group has continually pushed the limits and redefined the notion of what *de novo* sequencing can do, by introducing the spectral networks paradigm (23–25). Briefly, this strategy can assemble mass spectra into spectral pairs by joining overlapping spectra obtained from sample aliquots digested by different enzymes. As a consequence, it reduces noise and significantly improves protein coverage. Its latest version also combines data from different fragmentation techniques.

These algorithm developments have improved *de novo* sequencing shifting the bottleneck to post-sequence processing of data. This is because the output of a *de novo* software is a long list of highly similar full and partial peptide sequence and scores. An initial attempt to overcome these limitations consisted of a tag approach that was a hybrid between *de novo* sequencing and database searching: short sequence tags were derived from tandem mass spectra and used to search a sequence database (26). In what followed, a modified version based on the FASTA homology search tool was proposed for homology-driven proteomics (27). This strategy was implemented as part of the CIDentify tool, whose novelty was to account, in the alignment score, for limitations of mass spectrometry sequencing like switching between leucine and isoleucine or other combinations of amino acids having the same mass. The next steps were taken mainly by the Shevchenko group through the introduction of the MS-Blast algorithm, which relies on a different set of scores and uses the PAM30MS substitution matrix, itself tailored for mass-spectrometry-based proteomics (28, 29). For a complete review on *de novo* sequencing and homology searching we suggest reference (30).

The current *de novo* post-processing paradigm presents several limitations that are similar to those of the early PSM work-flow. Output files generally consist of a peptide list with corresponding scores,

demanding an experienced user to assess trustworthy identifications. If the same peptide is analyzed by different mass spectrometers it can generate different scores, which makes data comparison between different groups a challenging task. In a sense, problems are similar to when adopting the early Washburn criterion. Assembling protein information from a list of peptides is not a simple task, and usually not performed by state-of-the-art *de novo* tools. While there are great tools for doing this at the PSM level, there is still a lack of similar tools for *de novo* sequencing.

To tackle the shortcomings aforementioned, and in line with our strong interest in diversity-driven proteomics (29), we present a methodology for post processing *de novo* sequencing data that infers protein identification by statistically mapping *de novo* sequencing results to a protein sequence database. Our approach begins with the use of Gotoh's version of the Smith-Waterman algorithm, based on affine gap scoring (31) for increased scalability, to align *de novo* sequences against those in a target-decoy database. Then a radial basis function neural network (RBF-NN) is used to rank results according to alignment score, *de novo* score, precursor charge state, and peptide length. Finally, a heuristic method is used to present protein identification results in a user-friendly, interactive report. The resulting algorithm was implemented as the software termed PepExplorer. In essence, its goal is somewhat similar to that of post-processing tools such as DTASelect (9), Percolator (10), and SEPro (11), but with an extra layer of complexity inherent from *de novo* sequencing. PepExplorer can handle the output of several widely adopted *de novo* tools, such as PepNovo, pNovo+, and PEAKS, and accepts a generic format to enable result analysis from a broader range of tools once they are run through simple parsers. Likewise, the software accepts a series of database formats for input analysis. Taken together, these features are not found in other tools. PepExplorer is freely available to the scientific community and is provided with the necessary documentation.

The effectiveness of our methodology has been verified in two distinct scenarios, the first in a real but controlled experiment and the other pertaining to a comprehensive profiling of the plasma components of *Bothrops jararaca*, a venomous viper endemic to Brazil, southern Paraguay, and northern Argentina. The first scenario's purpose was to validate the effectiveness of the tool on a published *Pyrococcus furiosus* dataset (11). We note that this organism is recognized by the proteomics community as being well suited for benchmarking, since it allows for the rigorous testing of identification algorithms at the peptide and protein levels (32, 33). We modified the *P. furiosus* sequence database in such a way that no more peptides are identified by the PSM approach and by another widely adopted error-tolerant search tool Mod-A (34), we then show that we could recover protein identifications using our tool. The *B. jararaca* scenario has allowed us to explore uncharted territory, as this organism has an incomplete sequence database and therefore required us to rely on those of orthologous organisms. In particular, *B. jararaca* plasma was chosen as it is a main research model studied at the Laboratory of Toxinology (FIOCRUZ, Brazil) and several natural inhibitors of snake toxins have already been identified/characterized from this biological matrix (35–37).

METHODS

***Bothrops jararaca* plasma sample preparation**

B. jararaca plasma was supplied to the Laboratory of Toxinology (FIOCRUZ, Brazil) during the experimental procedures described in the research project, approved by the Ethics Committee of the Butantan Institute (555/2008), of the Biomedical Science Institute of the University of São Paulo (138/2009). This project was also approved by the Brazilian Institute for Environment and Renewable Natural Resources, a Brazilian Ministry of the Environment's enforcement agency (IBAMA, License 01/2009). Protein concentration was determined by the bicinchoninic acid assay (BCA) (38) and 40 μg were submitted to complete dryness by lyophilization. Next, 20 μL of a 0.25% (m/v) *Rapi*Gest SF (Waters, USA) in 50 mM ammonium bicarbonate solution were added to solubilize the proteins, followed by heating for 5 min at 100 °C. Disulfide bridges were reduced with 20 mM dithiothreitol for 30 min at 60 °C followed by cysteine alkylation with 68 mM iodoacetamide for 15 min at room temperature, in the dark. Four microliters of a 0.2 μg/μL, in 50 mM acetic acid, porcine trypsin solution (cat. number V511 - Promega, USA) were added and incubation proceeded for 22.5 h at 37 °C followed by 45 min at 56 °C. The reaction was stopped by the addition of 2.4 μL 5% (v/v) trifluoroacetic acid in water. For *Rapi*Gest removal, samples were incubated for 45 min at 37 °C, centrifuged at 16,000x*g* for 10 min at room temperature. The supernatant was collected and desalted/concentrated with *in-house* made columns packed with Poros R2 resin (Life Technologies, USA), eluted with 60% acetonitrile in 0.1% (v/v) trifluoroacetic acid and completely dried using a SpeedVac (Thermo Scientific, USA) vacuum centrifuge concentrator. Samples were resuspended in 30 μL of 1% (v/v) formic acid and submitted to a 10 min ultrasonic bath cycle before analysis by nanoLC-MS/MS.

Mass-spectrometry analysis

The sample was analyzed in technical triplicate by LC-MS/MS. Tryptic digests were separated by reversed-phase capillary liquid chromatography coupled to nano-electrospray high-resolution mass

spectrometry for identification. For each sample, 2 μL of desalted tryptic peptide digest were initially applied to a 2-cm long (100 μm internal diameter) trap column packed with 5 μm, 200 A Magic C18 AQ matrix (Michrom Bioresources, USA) followed by separation on a 30-cm long (75 μm internal diameter) column that was packed with the same matrix, directly on a self-pack 15 μm PicoFrit empty column (New Objective, USA). Chromatography was carried out on an EASY-nLC II instrument (Thermo Scientific, USA). Samples were loaded onto the trap column at 2,000 nL/min while chromatographic separation occurred at 200 nL/min. Mobile phase A consisted of 0.1 % (v/v) formic acid in water while mobile phase B consisted of 0.1% (v/v) formic acid in acetonitrile and gradient conditions were as follows: 2 to 40% B during 162 min; up to 80% B in 4 min, maintaining at this concentration for 2 min. more. Eluted peptides were directly introduced to the LTQ XL/Orbitrap mass spectrometer (Thermo, San Jose) for analysis. Source voltage was set to 1.9 kV, capillary temperature to 200 C and tube lens voltage to 100 V. FTMS full and MSn AGC target values were set to 500,000 and 200,000, respectively. MS1 spectra were acquired on the Orbitrap analyzer (300 to 1,700 *m/z*) at a 60,000 resolution (for *m/z* 445.1200). Tandem mass spectra were acquired from the 10 most intense ions by HCD fragmentation (minimum signal required of 10,000; isolation width of 2.0; normalized collision energy of 45.0 and activation time of 30 s) followed by MS2 acquisition on the Orbitrap analyzer at a 15,000 resolution. Dynamic exclusion option was enabled and set with the following values for each parameter: repeat count = 1; repeat duration = 30 s; exclusion list size = 500; exclusion duration = 45 s and exclusion mass width = 10 ppm. Charge state rejection was enabled for unassigned charges and for those equal to one.

Preparation of sequence databases used for similarity-driven identification and PSM

Reference sequences for *P. furiosus* were obtained from UniProt and those for *Reptilia* together with *Amphibia* are from the NCBI RefSeq, all downloaded in June 2013. The sequences obtained from the *Reptilia* and *Amphibia* databases were merged into a single structure and then joined by 127 sequences of common mass spectrometry contaminants as well as, for each database entry, a reversed version of the

corresponding sequence (a decoy sequence). The final *P. furiosus* and *Reptilia* plus *Amphibia* databases had 4,347 and 302,287 sequences, respectively.

Three *P. furiosus* proof-of-concept databases were generated by repeatedly adding "mutations" and insertions to the sequence database. These databases are referenced as PFU_Gap25_Substitution15, PFU_Gap20_Substitution10, and PFU_Gap15_Substitution8. In the PFU_Gap25_Substitution15 database, for example, an amino acid is inserted at every 25th position; likewise, every 15th amino acid is replaced by some other, randomly chosen amino acid. These databases provide increasing distances from the original database and thereby present the algorithms with different levels of difficulty. As no new proteins were added to obtain any of them, each of these databases has the same number of sequences as the initial *P. furiosus* database. Our goal has been to modify the native sequences from the organism's database to simulate the appearance of different, but phylogenetically close, organisms that would render PSM useless.

Peptide spectrum matches and quality assessment

The mass spectra were exported to the MS2 format (39) from the RAW files using PatternLab's RawReader module (available at: http://proteomics.fiocruz.br/Softwares.aspx). The ProLuCID (40) search engine was used to compare experimentally obtained spectra against theoretical spectra generated from a sequence database and select the most similar. Briefly, the search was limited to fully tryptic peptide candidates as we imposed only carbamidomethylation as a fixed modification. The search engine accepted peptide candidates within a 50 ppm tolerance from the measured precursor *m/z, 550* ppm for the MS2, and used XCorr and ZScore as the primary and secondary search engine scores, respectively.

The validity of the peptide spectrum matches was assessed using the Search Engine Processor (SEPro) (11). Identifications were grouped by charge state ($\leq$ +2 and > +2), resulting in two distinct subgroups. For each result, the ProLuCID XCorr, DeltaCN, and ZScore values were used to generate a Bayesian discriminator. The identifications were sorted in nondecreasing order according to the discriminator score.

A cutoff score was established to accept an FDR of 1% based on the number of labeled decoys. This procedure was independently performed on each data subset, resulting in a false-positive rate that was independent of charge state. Additionally, a minimum six-amino acid residues-long sequence was required. Results were post-processed to only accept matches with less than 5 ppm and proteins supported by at least two spectral counts. This last filter led to a 0% FDR in the search results.

De novo sequencing

De novo sequencing was performed using PEAKS Studio 6.0 (Bioinformatics Solutions Inc., ON, Canada). The parent mass error tolerance was 7 ppm, the fragment mass error tolerance was 0.05 Da. Carbamidomethylation of cysteine was considered as a fixed modification. Acceptable results required an ALC score of at least 70% and a TLC of at least 5, and were exported to a CSV file using the export option built into the software.

Blind PTM search with Mod-A

Mod-A was used to search the original and modified versions of the PFU dataset using its automatic precursor mass detection mode and allowing for arbitrary modifications in the peptides. The parameter files used by Mod-A are included in the supplementary files.

PepExplorer algorithm

The PepExplorer algorithm was coded in C# 4.5. It has a graphical user interface but can also be executed from the command prompt, which enables it to work in cluster environments. The algorithm's work-flow can be summarized in four steps: *de novo* result parsing, sequence alignment, result filtering, and result presentation (Figure 1). Below we detail each of these steps. All parameters can be adjusted using the graphical user interface (Figure 2).

[Suggested place for Figure 1]

[Suggested place for Figure 2]

De novo result parsing

PepExplorer currently contains parsers for three widely adopted *de novo* sequencing algorithms: PepNovo, pNovo+, and Peaks. PepExplorer treats the *de novo* algorithm with an abstraction layer that allows for the inclusion of new parsers upon request. The software also allows the analysis of a list of peptides by copying and pasting them in the corresponding text box found in the *de novo* Output box (Figure 2). However, in this scenario its neural network runs in a simplified mode by not considering precursor charge states, scan numbers, and *de novo* score quality.

Sequence alignment

PepExplorer relies on Gotoh's version of the Smith-Waterman algorithm (31), built into its core for aligning peptide sequences against a target-decoy sequence database specified by the user. The user can specify several alignment parameters, such as the open gap and extend gap penalties, the number of *de novo* sequence results to be considered per spectrum, and a substitution matrix of choice. For this study the values were 13, 5, 1, and the PAM30MS matrix, respectively.

These default open gap and extend gap parameters resulted from a grid search also made available in PepExplorer through the *Advanced Analysis* menu. This enables the algorithm to explore the landscape of combinations of these two parameters within user-predefined bounds and report the combination yielding the highest number of alignments under a user-defined FDR. For this work, we performed the grid search in the PFU dataset allowing both the open gap and the extend gap penalties to vary from 2 to 30. The grid search results are available as part of the online supplementary files in the software's website (see availability section).

Result filtering with the RBF-NN

Each sequence alignment obtained in the previous step is internally treated by PepExplorer as an alignment object containing the following properties: peptide length, *de novo* score, precursor charge, number of gaps, identifier, similarity and alignment scores. All these parameters are used for result quality

assessment, together with complementary information relevant to report assembly, such as scan number, raw file name, and details on the sequence alignment.

As a first step, these alignment objects are separated into two lists: those originating from peptide ions with charge state less than or equal to two, and those from peptide ions with charge state greater than two. Each of these lists is handled by a different RBF-NN. This enables convergence to a list of alignment objects satisfying an FDR that is independent of precursor charge state.

Given a list of alignment objects, the RBF-NN is based defining six clusters; to this end, PepExplorer relies on the k-means++ algorithm (41) applied to the normalized values (i.e., between 0 and 1) of the alignment score, the *de novo* score, and the peptide length of each alignment object. We note that k-means++ employs a "careful seeding" to address the NP-hard problem of minimizing the average squared distance between points in the same cluster. The "careful seeding" is performed by choosing the first cluster center randomly among the data points to be clustered. Subsequent cluster centers are chosen from locations coinciding with the remaining data points with probability proportional to each point's squared distance to the closest existing center. The algorithm then continues with the established k-means optimization procedure. The initial "careful seeding" is justified by a faster convergence and finding better solutions. PepExplorer runs k-means++ 50 times in search of the best solution.

The RBF-NN is then trained to capture the nonlinear relationship between target and decoy alignment objects. The network we used is a single-hidden layer feed-forward neural network whose three input nodes forward the input signal to the hidden nodes directly, with no weights. The kernel transfer function used in the jth hidden node is

$$\varphi_j(x) = \exp\left(\frac{-\| x - \mu_j \|^2}{2\sigma_j^2}\right),$$

where μ_j is the jth cluster center determined by k-means++ and σ_j is a width parameter given by the smallest Euclidean distance between any two cluster centers. The latter is to better capture the local-

ness and thus the smoothness and continuity of the fitted function. The connections from the six hidden nodes to the single output node are weighted and the value of the output node is given by

$$y(x) = \sum_{j=1}^{6} w_j \varphi_j(x),$$

where w_j is the connection weight between the jth hidden node and the output node. During training, $y(x)$ is either $+1$ or -1, depending respectively on whether the alignment object in question corresponds to a target sequence or a decoy sequence (alignment objects mapping to sections of sequences found in both target and decoy sequences are not considered). The weights of the RBF-NN equations are determined by linear regression using a least-squares objective function. All identifications are sorted in a nondecreasing order according to the classification function. Finally, a cutoff score can be established to achieve an FDR based on the decoy identifications.

Result presentation

Results are presented in the form of a dynamic, interactive report that allows the user to sort them according to a criterion of choice and interact with it by setting parameters of interest. The report can quickly adjust to a user-specified FDR or provide a list of maximum-parsimony alignments, since all alignments are stored to enable the algorithm to quickly converge to various settings. Among the threshold parameters we highlight the global FDR, the minimum alignment count (the closest to spectral count), the maximum alignment parsimony, using distinct RBF-NN for precursors of different charge states, and the minimum identifier. The report is provided as two interactive panels, the upper one being related to protein information, the lower to identification data. The upper panel provides information such as protein identifier, protein length, coverage percentage, sequence count, alignment count, and description. By selecting a protein, detailed information is made available in the lower panel of all alignments that mapped to it such as the scan number, file name, *de novo* score, precursor charge state, identifier, similarity, number of gaps, alignment score, sequence found in the database, and sequence provided by the *de novo* sequencing tool (Figure 3). By selecting a row of interest in this lower panel, a

new window displaying the sequence alignment is made available. In this window, selecting a row in the upper panel with the protein information, a graphical coverage report is displayed (Figure 4). This report is integrated with the cloud service of PatternLab for Proteomics (42), enabling the use of the Infer Domains function to instantly access predicted on-demand protein domains inferred with HMMER3 over Pfam-A (43).

[Suggested place for Figure 3]

[Suggested place for Figure 4]

RESULTS

PFU proof of concept

A Venn Diagram showing the overlap of the protein identifications from ProLuCID / SEPro, Mod-A, and PepExplorer on the unmodified PFU database is found in Figure 5. We recall that only proteins having two or more spectral counts were considered.

[Suggested place for Figure 5]

We further manually examined the non-decoy proteins uniquely identified by PepExplorer; by individually analyzing each case, based on spectral quality, alignment scores, and coverage, we feel comfortable in considering them as correctly identified, even though it was not found among our gold-standard, ProLuCID.

The results of these tools applied to the modified versions of the PFU dataset is discriminated in Table I,

Table 1. Performance of ProLuCID, Mod-A, and PepExplorer on the four PFU datasets. The first number stands for how many proteins having at least two spectral counts were identified under a 1% FDR. The number in parentheses stands for the average sequence coverage.

	# Proteins (Average Coverage)			
	PFU	PFU_Gap_25_Substitution_15	PFU_Gap_20_Substitution_10	PFU_Gap_15_Substitution_8
ProLuCID	585 (0.16)	45 (0.04)	7 (0.04)	0
Mod-A	499 (0.16)	63 (0.06)	45 (0.05)	0
PepExplorer	311 (0.17)	190 (0.17)	143 (0.17)	102 (0.17)

Page: 15

All results are made available as part of the supplementary files or in the PepExplorer website.

***Bothrops jararaca* plasma proteomic assessment**

PepExplorer generated 3,862 alignments (1% FDR), corresponding to 1,333 peptides mapping to 199 proteins arranged into 86 protein groups. The ProLuCID / SEPro pipeline identified 349 spectra corresponding to 83 peptides mapping to 17 proteins arranged into 12 protein groups (0% FDR). All protein groups identified by ProLuCID were present in the PepExplorer results. Moreover, all but five proteins identified by ProLuCID had their identifiers contained in the PepExplorer results. These five remaining identifications share peptides or have at least 80% identity with one protein provided by PepExplorer. The detailed lists of identifications, SEPro files, and PepExplorer files are provided as supplementary files.

A 100% overlap between our similarity-driven approach and a PSM approach might not occur because of the convergence strategy adopted by PepExplorer, as it will opt for proteins having higher number of alignment mappings to converge to a maximum-parsimony list. On the hand, when we compare the average sequence coverage obtained for the same proteins identified by the PepExplorer and PSM approaches, there is an approximate 64% increase using the former approach (Suppl. Table 1).

Recently, Morais-Zani and co-workers (44) have analyzed the plasma composition of juvenile and adult *B. jararaca* snakes seeking for ontogenetic variability. They have used an experimental strategy consisting of two-dimensional electrophoresis (2DE) separation followed by mass spectrometry analysis and protein identification by PSM, using MASCOT as search engine. The authors were able to report 8 plasma protein groups, where one of them might be due to sample contamination during collection (beta-actin). With the exception of transferrin, all plasma proteins reported by them were also detected by our PSM approach (ProLuCID/SEPro); furthermore, we were also able to identify other proteins such as fibronectin 1, alpha-2-macroglobulin, apolipoprotein B100, fibrinogen beta chain, and small serum

protein (Suppl. Table 1). One possible explanation for our extended list of identifications by PSM might reside in our experimental approach (*shotgun* proteomics) as opposed to theirs (2DE).

Finally, when we compare the PepExplorer results (for proteins displaying a sequence count greater than two) we were able to identify all plasma protein families above and additional ones, namely: gamma phospholipase inhibitor type IV, plasminogen, ceruplasmin, IgG Fc-binding protein-like, complement C4-B-like, inter-alpha-trypsin inhibitor heavy chains H4- and H3-like, Ig lambda light chain variable region, calnexin, multiple EGF-like domains protein 6, collagen alpha-1(XXIV) chain, kininogen-1, anionic trypsin-like, fibrinogen alpha chain-like, Ig gamma-1 chain C region and heparin cofactor 2 (Suppl. Table 1). Supplementary Figure 1 exemplifies peptide identifications provided by PepExplorer that were missed by the PSM approach.

DISCUSSION

Error-tolerant, similarity-driven tools have as an ultimate goal to list "true" identifications. However, defining what it means for an identification to be true is far from trivial, so in general one seeks to define trueness based on how the sequences in question differ according to some measure relating to the evolutionary distance between them. Among the first successful attempts at quantifying an evolutionary distance we highlight the Point Accepted Mutation (PAM) divergence, defined for two given sequences as the average number of accepted point mutations per one hundred amino acids required to convert one sequence into the other without any insertions or deletions (45). The PAM matrices are substitution matrices that summarize an expected evolutionary change at the amino-acid level through log-odds substitution ratios. Theoretically, this approach is designed to compare sequences that are within a known evolutionary PAM divergence in evolution. Conversely, it is common experience that PAM matrices are, in general, very effective in finding "true alignments" that reflect biological phenomena even though PAM divergences do not always correspond to true evolutionary distances. In the experiment at hand, we chose one of the so-called low-order PAM matrices (e.g., PAM30MS), which theoretically

should favor "closer" sequences and therefore such true alignments. Future versions and tools should incorporate strategies for automatically selecting substitution matrices tailored for the problem at hand. This could ultimately help in determining a subset of sequences for maximizing the sensitivity of the algorithm. We argue that the current version of PepExplorer helps by showing which peptides (and ultimately proteins) can be taken into consideration confidently enough. However, selecting an adequate substitution matrix remains an issue for the user's careful consideration.

The results provided herein can be used to compare three paradigms for performing spectral identification: PSM, an error tolerant / blind-PTM approach, and a similarity-driven approach. The strategies are shown to be complementary, each having advantages and disadvantages. For example, the PSM approach was found to be the most sensitive one on the PFU dataset. This happened because we were dealing with a model organism, so by fully (and tightly) relying on the restrictions provided in the sequence database the best sensitivity could be achieved. However, its performance rapidly degraded as more distractions and modifications were inserted into the database. Although Mod-A has not outperformed PSM on the original PFU database, it has been able to retain significantly more identifications as more distractions were inserted in the database. Mod-A most likely did not outperform the PSM approach on the original PFU dataset because the latter takes into account many more possibilities, resulting in a larger search space and sacrificing sensitivity (13). On the other hand, on higher organisms it would not be surprising if Mod-A outperformed PSM, as it will tolerate amino-acid substitutions and unanticipated post-translational modifications. Indeed, taking into account multiple PTMs can also quickly degrade the performance of *de novo* tools, and for this reason Mod-A will always provide results that are complementary to those of PepExplorer. Finally, PepExplorer presented the least sensitive results on the original PFU dataset, as *de novo* approaches are know to be error-prone. On the other hand, the alignment paradigm is able to effectively retain the results as distractions are included in the database.

Finally, we would like to point out some potential applications of PepExplorer. Our algorithm is used to pinpoint a subset of *de novo* results that are similar to the database at hand. Yet, there can be several *de novo* results, having a very high *de novo* sequencing score, that are not included in the PepExplorer output. These results should be given special attention: what PepExplorer discards could actually turn out to be truly novel molecules, given the high confidence of the *de novo* results.

CONCLUDNG REMARKS

PepExplorer is recommended for large-scale shotgun proteomic experiments, that is, those in which a considerable number of spectra are generated, as in the datasets presented. Its use is not recommended for analyzing small collections of spectra such as those obtained when analyzing a 2D gel spots. In such cases MS-BLAST (28) should be used instead.

A key realization brought about by modern biotechnology has been that underneath the myriad unknown organisms lies great potential (46). Current strategies inspired by this realization include exploring extreme biomes for the so-called extremophiles, a peculiar class of organisms that are generally responsible for the biosynthesis of molecular components useful for pharmaceutical or industrial applications. Perhaps one of the best examples has been the discovery of *Thermus aquaticus* and its heat-resistant polymerases, elected by *Science* in 1989 as the "Molecule of the year" (47) and which have since aided in the development of biotechnology tools and ultimately facilitated the engineering of more effective drugs. The molecular characterization of venoms has also resulted in the engineering of new drugs (48). In conclusion, the literature is full of examples demonstrating the vast richness of biomolecular components and drug candidates that are naturally produced by different organisms already existing in our fauna and flora.

Recent advances in proteomic technologies are significantly impacting similarity-driven proteomics and consequently the exploration of novel organisms. Recently, Coon and coworkers benchmarked a new

hybrid mass spectrometer, the Orbitrap Fusion (Thermo, San Jose, California). The authors mention events in which they identified up to 19 sequences within less than a second, enabling them to achieve a 90% coverage of the yeast proteome in one hour (49). Through this, the authors have raised the bar, in terms of number of proteins identified per minute, to 70. High scanning rates coupled to ever-increasing resolving power are ingredients to boost the performance of *de novo* sequencing algorithms. As the general quality of predicted peptides is increasing, we foresee *de novo* sequencing playing a key role in efficiently handling data from organisms with no genomic information available.

The field of genomics is also constantly going through significant advances. For example, next-generation sequencers are enabling the single cell transcriptome (50) and personal genomics (51). Indeed, the coupling of *omics* sciences such as proteomics and metabolomics with next-generation sequencers will pave the way to true systems biology approaches as these strategies are complementary to each other. The ever-growing amount of data on sequenced organisms, powered by next-generation sequencers, adds to similarity-driven approaches as even more organisms will have their genomic information available. However, instrument time, expertise in data analysis, and financial resources are current bottlenecks for many groups.

Here we described a new methodology for dealing with *de novo* sequencing approaches, taking into account rigorous statistical criteria. We clearly demonstrated its efficiency in a controlled but real experiment with the PFU modified database, and then presented the most comprehensive proteomic profile of the *B. jararaca* plasma. Efforts such as the present work are necessary as they expand the possibilities of what can be achieved in proteomics and in the study of organism biology. In the near future we plan to automate the integration of data between different strategies like PSM and *de novo*, aiming at a wider perspective for mass-spectral analyses.

Page: 20

Availability of PepExplorer, the raw data, and results

PepExplorer and supplementary files, including the *B. jararaca* raw data and all the results described in this work, are made freely available for academic purposes at our website, http://proteomics.fiocruz.br/supplementaryfiles/pepexplorer/. To view the full PSM results, installation of SEPro is required (http://pcarvalho.com/patternlab/downloads/windows/seprocessor/). PepExplorer is required for viewing it results (http://proteomics.fiocruz.br/software/pepexplorer/).

Acknowledgments

References

1. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239, 487–491

2. Opie, L. H., and Kowolik, H. (1995) The discovery of captopril: from large animals to small molecules. *Cardiovasc. Res.* 30, 18–25

3. 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

4. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinforma. Oxf. Engl.* 20, 1466–1467

5. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567

6. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* 13, 378–386

7. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207–214

8. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74, 5383–5392

9. Cociorva, D., L Tabb, D., and Yates, J. R. (2007) Validation of tandem mass spectrometry database search results using DTASelect. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* Chapter 13, Unit 13.4

10. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* 4, 923–925

11. Carvalho, P. C., Fischer, J. S. G., Xu, T., Cociorva, D., Balbuena, T. S., Valente, R. H., Perales, J., Yates, J. R., 3rd, and Barbosa, V. C. (2012) Search engine processor: Filtering and organizing peptide spectrum matches. *Proteomics* 12, 944–949

12. Barboza, R., Cociorva, D., Xu, T., Barbosa, V. C., Perales, J., Valente, R. H., França, F. M. G., Yates, J. R., 3rd, and Carvalho, P. C. (2011) Can the false-discovery rate be misleading? *Proteomics* 11, 4105–4108

13. 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. *Bioinforma. Oxf. Engl.* 29, 1343–1344

14. Biemann, K., Cone, C., Webster, B. R., and Arsenault, G. P. (1966) Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.* 88, 5598–5606

15. Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., and Zhang, X. (2006) Performance evaluation of existing de novo sequencing algorithms. *J. Proteome Res.* 5, 3018–3028

16. Frank, A., and Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 77, 964–973

17. 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

18. Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., and Buhmann, J. M. (2005) NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.* 77, 7265–7273

19. 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. RCM* 17, 2337–2342

20. Coon, J. J., Ueberheide, B., Syka, J. E. P., Dryhurst, D. D., Ausio, J., Shabanowitz, J., and Hunt, D. F. (2005) Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 102, 9463–9468

21. Zubarev, R. A., Horn, D. M., Fridriksson, E. K., Kelleher, N. L., Kruger, N. A., Lewis, M. A., Carpenter, B. K., and McLafferty, F. W. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* 72, 563–573

22. Zubarev, R. A., Zubarev, A. R., and Savitski, M. M. (2008) Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? *J. Am. Soc. Mass Spectrom.* 19, 753–761

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

24. Guthals, A., and Bandeira, N. (2012) Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Mol. Cell. Proteomics MCP* 11, 550–557

25. Guthals, A., Clauser, K. R., and Bandeira, N. (2012) Shotgun protein sequencing with meta-contig assembly. *Mol. Cell. Proteomics MCP* 11, 1084–1096

26. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399

27. Taylor, J. A., and Johnson, R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom. RCM* 11, 1067–1075

28. 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

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

30. Ma, B., and Johnson, R. (2012) De novo sequencing and homology searching. *Mol. Cell. Proteomics MCP* 11, O111.014902

31. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162, 705–708

32. Vaudel, M., Burkhart, J. M., Breiter, D., Zahedi, R. P., Sickmann, A., and Martens, L. (2012) A complex standard for protein identification, designed by evolution. *J. Proteome Res.* 11, 5065–5071

33. Yates, J. R., 3rd, Park, S. K. R., Delahunty, C. M., Xu, T., Savas, J. N., Cociorva, D., and Carvalho, P. C. (2012) Toward objective evaluation of proteomic algorithms. *Nat. Methods* 9, 455–456

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

35. Estevão-Costa, M. I., Rocha, B. C., de Alvarenga Mudado, M., Redondo, R., Franco, G. R., and Fortes-Dias, C. L. (2008) Prospection, structural analysis and phylogenetic relationships of endogenous gamma-phospholipase A(2) inhibitors in Brazilian Bothrops snakes (Viperidae, Crotalinae). *Toxicon Off. J. Int. Soc. Toxinology* 52, 122–129

36. Tanaka-Azevedo, A. M., Tanaka, A. S., and Sano-Martins, I. S. (2003) A new blood coagulation inhibitor from the snake Bothrops jararaca plasma: isolation and characterization. *Biochem. Biophys. Res. Commun.* 308, 706–712

37. Valente, R. H., Dragulev, B., Perales, J., Fox, J. W., and Domont, G. B. (2001) BJ46a, a snake venom metalloproteinase inhibitor. Isolation, characterization, cloning and insights into its mechanism of action. *Eur. J. Biochem. FEBS* 268, 3042–3052

38. Smith, P. K., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M. D., Fujimoto, E. K., Goeke, N. M., Olson, B. J., and Klenk, D. C. (1985) Measurement of protein using bicinchoninic acid. *Anal. Biochem.* 150, 76–85

39. McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J., Venable, J., Graumann, J., Johnson, J. R., Cociorva, D., and Yates, J. R., 3rd (2004) MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom. RCM* 18, 2162–2168

40. Xu, T., Venable, J. D., Park, S., Cociorva, D., Lu, B., Liao, L., Wohlschlegel, J., Hewel, J., and Yates, J. R. (2006) ProLuCID, a Fast and Sensitive Tandem Mass Spectra-based Protein Identification Program. *Mol Cell Proteomics* 5, S174

41. Arthur, D., and Vassilvitskii, S. (2007) in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07. (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA), pp 1027–1035.

42. 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

43. Eddy, S. R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform. Int. Conf. Genome Inform.* 23, 205–211

44. De Morais-Zani, K., Grego, K. F., Tanaka, A. S., and Tanaka-Azevedo, A. M. (2013) Proteomic Analysis of the Ontogenetic Variability in Plasma Composition of Juvenile and Adult Bothrops jararaca Snakes. *Int. J. Proteomics* 2013, 135709

45. Dayhoff, M. O. (1979) *Atlas of protein sequence and structure, 5, 5,* (National Biomedical Research Foundation, Washington, D.C.)

46. Brock, T. D. (1997) The value of basic research: discovery of Thermus aquaticus and other extreme thermophiles. *Genetics* 146, 1207–1210

47. Guyer, R. L., and Koshland, D. E., Jr (1989) The Molecule of the Year. *Science* 246, 1543–1546

48. Fox, J. W., and Serrano, S. M. T. (2007) Approaching the golden age of natural product pharmaceuticals from venom libraries: an overview of toxins and toxin-derivatives currently involved in therapeutic or diagnostic applications. *Curr. Pharm. Des.* 13, 2927–2934

49. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., and Coon, J. J. (2014) The one hour yeast proteome. *Mol. Cell. Proteomics MCP* 13, 339–347

50. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382

51. Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., Stamatoyannopoulos, J. A., and Akey, J. M. (2012) Personal and population genomics of human regulatory variation. *Genome Res.* 22, 1689–1697

Page: 25
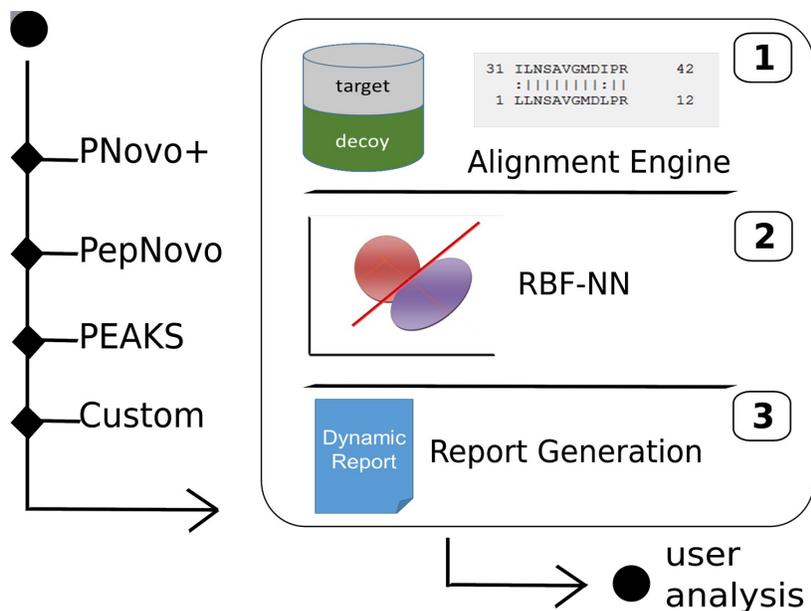
Figure 1. A *de novo* tool is used to generate candidate sequences from mass spectra.

The output from the *de novo* tool, together with a target-decoy database, serves as input to PepExplorer. PepExplorer uses the Smith-Waterman algorithm to align the *de novo* sequences against the target-decoy database (1). An RBF neural network is employed to rank the *de novo* alignments according to a confidence score that takes into account the *de novo* sequencing score, the alignment score, and the number of amino acids contained in the peptide (2). Finally, a dynamic report is generated (3).
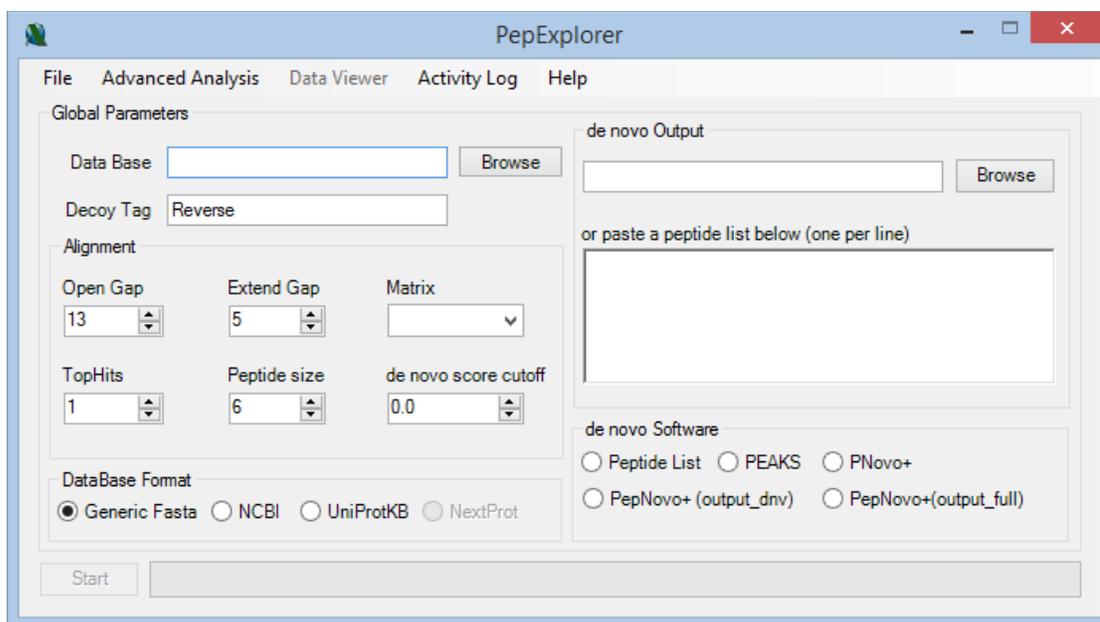
Figure 2- The PepExplorer graphical user interface.

Figure 3- Graphical user interface of the results browser. The results browser is composed of two

panels. The upper panel displays information related to protein identification. By clicking on a protein of

interest, further details on the peptides and their corresponding alignments are displayed in the lower panel.
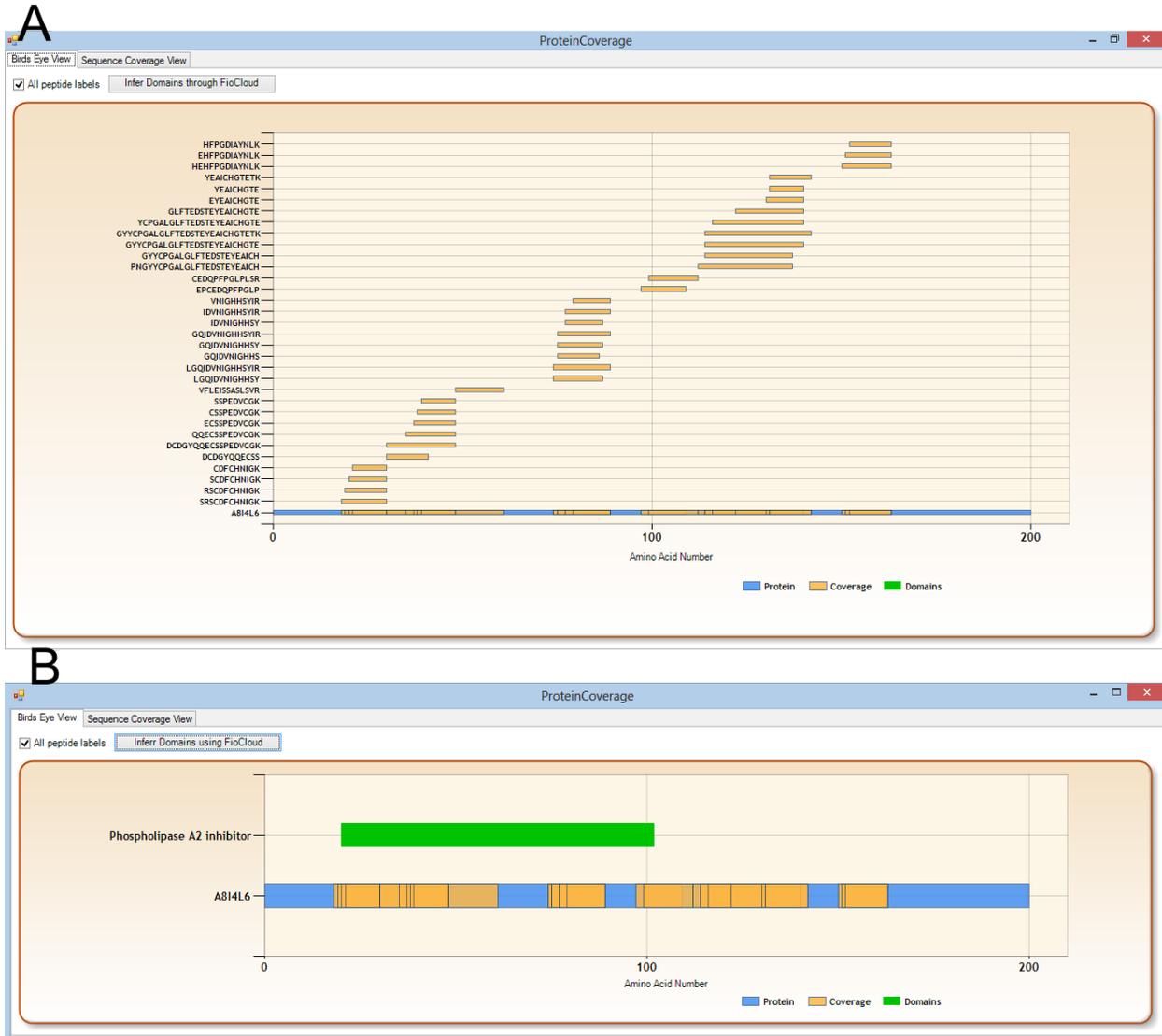

Figure 4- Example of report provided by PepExplorer for each identified protein. In (A) the graphical report of the protein sequence coverage shows the extension of the area covered by predicted peptides. In (B) the result from the domain inferred by the cloud service running HMMER3 over Pfam-A on the fly is shown.
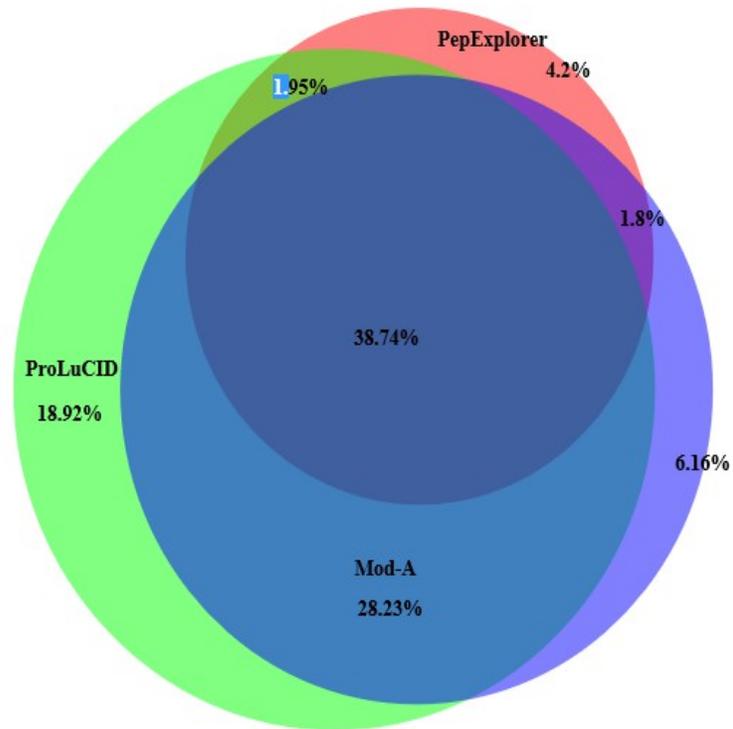
Figure 5. A Venn diagram comparing the protein identification overlap of ProLuCID, Mod-A, and PepExplorer in the PFU dataset using the unmodified database.