Bioinformatics, 33(21), 2017, 3471–3472
doi: 10.1093/bioinformatics/btx318
Advance Access Publication Date: 17 May 2017
Applications Note

OXFORD

Sequence analysis

The neXtProt peptide uniqueness checker: a tool for the proteomics community

Mathieu Schaeffer[1,†], Alain Gateau[2,†], Daniel Teixeira[2], Pierre-André Michel[2], Monique Zahn-Zabal[2] and Lydie Lane[1,2,*]

[1]Department of Human Protein Science, Faculty of Medicine, University of Geneva, Geneva, Switzerland and [2]CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, 1211 Geneva 4, Switzerland

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on February 22, 2017; revised on May 5, 2017; editorial decision on May 8, 2017; accepted on May 10, 2017

Abstract

Summary: The neXtProt peptide uniqueness checker allows scientists to define which peptides can be used to validate the existence of human proteins, i.e. map uniquely versus multiply to human protein sequences taking into account isobaric substitutions, alternative splicing and single amino acid variants.
Availability and implementation: The pepx program is available at https://github.com/calipho-sib/pepx and can be launched from the command line or through a cgi web interface. Indexing requires a sequence file in FASTA format. The peptide uniqueness checker tool is freely available on the web at https://www.nextprot.org/tools/peptide-uniqueness-checker and from the neXtProt API at https://api.nextprot.org/.
Contact: lydie.lane@sib.swiss

1 Introduction

Most proteomics experiments aiming to identify proteins in complex samples rely on proteolytic digestion followed by separation of the resulting peptides by liquid chromatography and their identification by tandem mass spectrometry (MS). Since the link between peptides and their protein precursors is lost, peptide-to-protein mappings must be obtained and critically evaluated, even when there is strong evidence that peptide identifications are correct. To ensure that a peptide unambiguously maps to a protein, isobaric substitutions of isoleucine to leucine that cannot be distinguished by most current MS techniques must be taken into account, as well as possible sequence variations arising from single amino acid variants (SAAVs) and alternative splicing. This is especially important for projects such as the HUPO Human Proteome Project (HPP), whose aim is to experimentally validate the existence of *in silico*-predicted human proteins (Omenn *et al.*, 2016).

The neXtProt knowledgebase on human proteins currently contains >42 000 isoforms produced by alternative splicing and >5 million variants (Gaudet *et al.*, 2017), which generates an incommensurate number of proteoforms. There is currently no tool taking this diversity into account when mapping peptides to proteins. The peptide uniqueness checker tool on the neXtProt platform was developed to meet this need.

2 Materials and methods

The peptide indexer (pepx) maps proteomics peptides to a given set of protein sequences using an *n*-mer-based index: all protein sequences are scanned with a sliding window of *n* amino acids and each *n*-mer found is written in an index pointing to the list of identifiers of sequences containing it. Peptides to be mapped are scanned with the same *n*-mer sliding window. For each peptide all *n*-mers must be found in the same protein isoform sequence to return a match. This method allows protein variations to be taken into account comprehensively, while minimizing combinatorial explosion.

For proteomic peptides from shotgun MS experiments which are typically 7–30 aa in length, the index for 6-mers offers the best trade-off between speed, memory and performance. The pepx program also builds indexes for 3-, 4- and 5-mers which can for instance be used to

search for short linear motifs. To further guarantee the confidence in uniqueness pepx can build special indexes where isobaric amino acids I (Ile) and L (Leu) are merged and replaced with the ambiguity code J. Thus no false positive will occur when two peptides differing only by I/L substitutions exist in different proteins.

3 Usage

As the reference knowledgebase for HPP, neXtProt validates the existence of human proteins based on several criteria, including peptide identification data from mass spectrometry-based proteomics experiments. According to the latest HPP guidelines, a protein is validated if two unique peptides of at least 9 aa in length are reported (Deutsch *et al.*, 2016). At each neXtProt release, all splice isoforms are indexed, and pepx is used to assess peptide uniqueness: a peptide is considered to be unique if all the matching isoform sequences derive from a single neXtProt entry.

For the so-called 'missing' proteins for which no experimental evidence has previously been reported, HPP requests to further check peptide uniqueness by taking all possible SAAVs into account (Deutsch *et al.*, 2016). Pepx can be used for this check by using an index containing the >5 million SAAVs from neXtProt; the index building step takes a few hours and index size is about 4 Gb. A single substitution per 6-mer is accepted in order to limit the size of the index to a manageable size. For example, PNVLLA with known variants P->L, P->S and V->A will generate entries for sequences PNVLLA, LNVLLA, SNVLLA and PNALLA, but not SNALLA. The position of the SAAV within the 6-mer is recorded in the indexes, allowing the variant path to be rapidly reconstructed when displaying matches. Since pepx has successfully been used to validate the identification of missing proteins in human sperm (Vandenbrouck *et al.*, 2016), the HPP board requested that this tool be accessible to the whole human proteomics community.

A dedicated web interface, the neXtProt peptide uniqueness checker, and a corresponding API (application programming interface) service have been developed. The list of peptides, which can be typed in a text area or imported from a text file, is sent via an http request to the API (Fig. 1(1)) which then queries the 6-mer index created by pepx (Fig. 1(2)) and returns the identifiers of the isoform matched (Fig. 1(3)). Because pepx retrieves isoform sequences that map to the 6-mers included in a given peptide and not to the full extent of the peptide, it occasionally returns false positives. An exact string search using the neXtProt API is performed as a validation step to ensure that the entire sequence of the peptide is present in the retrieved sequences. Another validation step is performed to check that the variant at the indicated position in the matching entry exists and justifies the match (Fig. 1(4)). The validated matches (Fig. 1(5)) are sent from the API server to the client in JSON (JavaScript Object Notation) format (Fig. 1(6)). The results are displayed in boxes, each box containing the resulting matches for one peptide, taking into account SAAVs (lower panel) or not (upper panel). Usually, in MS data analysis, peptide uniqueness is not evaluated at the level of protein isoforms, but at the level of genes or protein entries. Therefore, by default, all the matching isoforms that belong to a same entry are merged, and matches are displayed as neXtProt entry accession numbers followed by the corresponding gene names. A button in each box allows the user to toggle between this default view and the isoform view, which displays matches at the level of isoform sequences, and, in case of additional mappings due to variants, the variant involved. A color code allows entry-specific peptides (in green) to be quickly distinguished from peptides matching

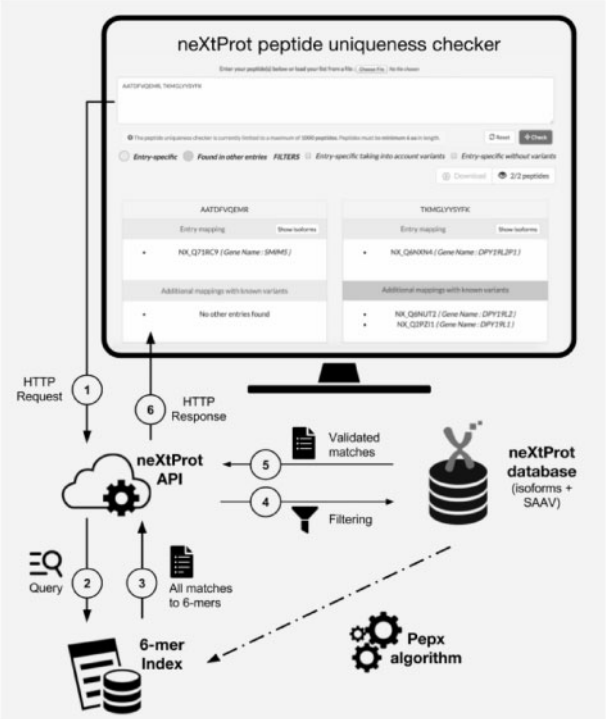

Fig. 1. neXtProt peptide uniqueness checker workflow. Of the two peptides submitted in this example, one is entry-specific (left) while the other loses its specificity if variants are taken into account (right)

several entries (in blue). Appropriate filters allow only entry-specific peptides to be displayed, either taking variants into account or not. The results displayed can be downloaded in CSV (Comma-separated values) format.

The use of this tool is recommended in the latest HPP guidelines (Deutsch *et al.*, 2016).

Acknowledgements

The neXtProt server is hosted by Vital-IT, the SIB Swiss Institute of Bioinformatics' Competence Centre in Bioinformatics and Computational Biology. The authors thank Amos Bairoch for his insight on the article; Frédéric Nikitin, Anne Gleizes and Valentine Rech de Laval for their help in integrating the uniqueness checker in the neXtProt interface. They also thank the HUPO HPP collaborators for their valuable feedback.

Funding

This work was supported by SIB Swiss Institute of Bioinformatics and University of Geneva.

Conflict of Interest: none declared.

References

Deutsch,E.W. *et al.* (2016) Human proteome project mass spectrometry data interpretation guidelines 2.1. *J. Proteome Res.*, **15**, 3961–3970.

Gaudet,P. *et al.* (2017) The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.*, **45**, D177–D182.

Omenn,G.S. *et al.* (2016) Metrics for the human proteome project 2016: progress on identifying and characterizing the human proteome, including post-translational modifications. *J. Proteome Res.*, **15**, 3951–3960.

Vandenbrouck,Y. *et al.* (2016) Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *J. Proteome Res.*, **15**, 3998–4019.