Bioinformatics, 33(13), 2017, 2042–2044
doi: 10.1093/bioinformatics/btx122
Advance Access Publication Date: 2 March 2017
Applications Note

Sequence analysis

PeptideMapper: efficient and versatile amino acid sequence and tag mapping

Dominik Kopczynski[1], Harald Barsnes[2,3], Pål R. Njølstad[4,5], Albert Sickmann[1,6,7], Marc Vaudel[4,8,*] and Robert Ahrends[1,*]

[1]Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V, Dortmund 44227, Germany, [2]Proteomics Unit, Department of Biomedicine, [3]Computational Biology Unit, Department of Informatics, [4]KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway, [5]Department of Pediatrics, Haukeland University Hospital, Bergen, Norway, [6]College of Physical Sciences, University of Aberdeen, Old Aberdeen AB24 3UE, UK, [7]Medizinische Fakultät, Ruhr-Universität Bochum, Bochum 44801, Germany and [8]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 5, 2016; revised on February 1, 2017; editorial decision on February 26, 2017; accepted on March 1, 2017

Abstract

The mapping of amino acid sequences is an essential task in bioinformatics. Notably, the mapping of peptide sequences on a proteome is required for the post-processing of proteomics results. However, this step can quickly become a bottleneck when working with extensive numbers of peptides or large protein sequence databases. Here, we present PeptideMapper, a novel amino acid sequence mapper for both peptide sequences and *de novo* sequencing identification results. By taking advantage of the latest advances in pattern matching, PeptideMapper achieves unprecedented performance (i.e. up to 1000× faster than state-of-the-art software) in terms of memory footprint and execution speed, with regards to both the indexing and the querying of protein sequence databases.

Availability and Implementation: PeptideMapper is implemented in the open source Java CompOmics framework under the permissive Apache 2.0 license https://github.com/compomics/compomics-utilities.

Contact: robert.ahrends@isas.de or marc.vaudel@uib.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High throughput proteomics sample characterization techniques have provided researchers with vast amounts of identification results. These often need to be remapped to a reference proteome, e.g. after an update of the protein sequences, enrichment with sequence variants, or simply to avoid biases in protein inference algorithms. However, the mapping of peptides to proteins can often become a bottleneck in pipelines dealing with large datasets and/or databases, e.g. in fields such as metagenomics and proteogenomics. Several indexing approaches have been created to map peptides against a reference proteome in sublinear time, e.g. using the q-gram index (Chen *et al.*, 2013), but since many indexes require a contiguous sequence of at least six amino acids (i.e. 6-gram), the mapping of shorter amino acid sequences is not possible. Here, we propose a new approach for indexing and querying proteomes, called PeptideMapper. PeptideMapper is based on the FM-Index (Ferragina and Manzini, 2000), a modern, flexible and compact index already employed in genetic sequencing mapping DNA reads against the genome (Langmead *et al.*, 2009; Li and Durbin, 2009). While it allows the mapping of amino acid sequences of any length to protein sequences at an unprecedented speed, we also implemented the possibility to map sequence tags. Sequence tags are amino acid sequences presenting mass gaps, unresolved series of amino acids (Mann and Wilm, 1994). Notably, our implementation accounts for (i) indistinguishable

amino acids such as leucine/isoleucine, (ii) amino acid combinations such as X representing any amino acid, (iii) post-translational modifications (PTMs) and (iv) sequence variants.

2 Materials and methods

PeptideMapper is implemented in Java as part of the open source CompOmics framework (Barsnes *et al.*, 2011). The details of the implementation and the benchmarking can be found in Supplementary Information. Briefly, an index of all provided protein sequences (i.e. provided as FASTA files) is created, and this index can then be quickly queried for amino acid sequences or tags, retrieving a list of the mapped peptides together with their location in the corresponding protein sequences. The index is created as follows: (i) compute the suffix array (SA) (Manber and Myers, 1990) of the complete proteome, (ii) compute its Burrows and Wheeler (1994) transform (BWT) and (iii) store the BWT in a wavelet tree (WT) (Grossi *et al.*, 2003). Note that we sample the SA and derive the missing entries by using the last-to-front (LF)-mapping function on the BWT (Ferragina and Manzini, 2000). To query the index, our approach utilizes the backward search, finding all occurrences of a pattern in a text in $\mathcal{O}(p)$, where p is the size of the pattern. The query function was further extended to manage amino acids of indistinguishable mass, amino acid combinations, as well as variants. Finally, the mapping of sequence tags was implemented, including PTMs. The performance of PeptideMapper was evaluated on a standard desktop computer using the databases listed in Table 1, including a proteogenomics (Crappé *et al.*, 2014; Menschaert *et al.*, 2013) and metaproteomics (Tanca *et al.*, 2013) database.

Table 1. For every of the six databases used for benchmarking, this table provides the date of release, the number of protein sequences, the number of residues and the number of 'X' residues found in the protein sequences

Database	*Date*	*# Sequences*	*# Amino acids*	*# X*
UniProt yeast	May 2016	6 721	3 025 143	0
UniProt mouse	May 2016	16 813	9 474 758	79
UniProt human	July 2015	20 207	11 326 153	670
Proteogenomics	2015	83 721	13 851 427	0
Metaproteomics	Jan. 2013	55 152	100 955 085	2 561 698
All UniProt prot.	July 2016	551 705	197 114 987	8 027

3 Results

PeptideMapper is easily implementable in both server and desktop applications, either via command line or as external library. As a proof-of-principle, PeptideMapper has been integrated in the user friendly PeptideShaker (Vaudel *et al.*, 2015) and DeNovoGUI (Muth *et al.*, 2014) tools, for processing proteomics search engine and *de novo* identification results, respectively. As shown in Figure 1A, the memory footprint of PeptideMapper is around 32 MB for the canonical yeast, mouse and human proteomic databases. It remains below 250 MB for the large proteogenomics and metaproteomics databases, and below 521 MB for the concatenation of all UniProt proteomes, enabling in memory storage even in such extreme cases. Allied to very efficient querying methods, this makes short query times affordable on most setups, see Figure 1B. Both the indexing speed and the index size evolve linearly with the database size, however, as demonstrated in Figure 1C, query times also increase linearly depending on dataset size and the prevalence of ambiguous amino acids such as 'X' that can represent any amino acid. Sequence tags can be inferred by *de novo* sequencing tools and can be mapped to proteomes using sequence alignment tools (Vaudel *et al.*, 2012). However, this approach does not take the precursor mass and possible PTMs into account. An alternative was proposed with TagRecon (Dasari *et al.*, 2010), but TagRecon only supports a single tag structure (Mass-Sequence-Mass), a limited set of PTMs, and its performance is heavily dependent on cleavage restrictions. For details on the benchmark comparison, please consider the Supplementary Material and the PeptideMapper web site. In contrast, PeptideMapper supports virtually any sequence tag, any set of PTMs, and, as demonstrated in Figure 1C, it can quickly map large datasets onto a proteome without any cleavage restriction. Yet, it can be anticipated that the query time will increase depending on the complexity of the query, e.g. due to the inclusion of numerous variable modifications and variants.

4 Conclusion

PeptideMapper is a very efficient and versatile approach for indexing entire proteomes, and is particularly well-suited for challenging

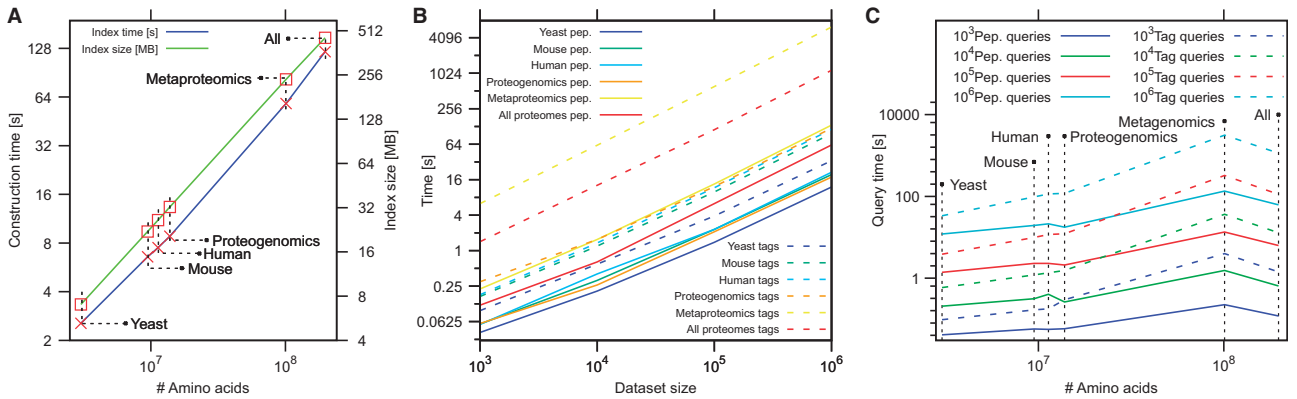

Fig. 1. Performance benchmark of PeptideMapper using various datasets and databases: the UniProt yeast, mouse and human proteomes, plus metagenomics and metaproteomics databases, and finally, the concatenation of all UniProt proteomes. (**A**) Indexing time (blue) and size (green), in seconds (s) and megabytes (MB) respectively, plotted against the database size in terms of the total number of amino acids. (**B**) Query time for the different peptides and tags datasets in solid and dashed lines, respectively, plotted against the dataset size in terms of the number of peptides or tags. (**C**) Query time for the different peptides and tags datasets in solid and dashed lines, respectively, plotted against the database size in terms of the number of amino acids. See main text and Supplementary Information for details

experimental setups involving large databases. It has the potential to overcome several bottlenecks in proteomic data analysis workflows and is especially equipped for the handling of *de novo* sequencing results. PeptideMapper is however not limited to proteomics, and we invite the community to broaden its application.

Funding

The supports by the Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen, the Senatsverwaltung für Wirtschaft, Technologie und Forschung des Landes Berlin and the Bundesministerium für Bildung und Forschung and BMBF (Code 031A534B) to DK and RA, the Norwegian Research Council to HB and the Bergen Research Foundation to HB are gratefully acknowledged. We thank the ISAS lipidomics team together with members of the protein dynamics lab and our colleges from MPC (Bochum, Germany) for thoughtful comments and discussions.

Conflict of Interest: none declared.

References

Barsnes,H. *et al*. (2011) compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics*, **12**, 1.

Burrows,M. and Wheeler,D.J. (1994) A block-sorting lossless data compression algorithm. Technical report, Digital Equipment Corporation.

Chen,C. *et al*. (2013) A fast peptide match service for UniProt knowledgebase. *Bioinformatics*, **29**, 2808–2809.

Crappé,J. *et al*. (2014) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res*., **43**, e29.

Dasari,S. *et al*. (2010) TagRecon: high-throughput mutation identification through sequence tagging. *J. Proteome Res*., **9**, 1716–1726.

Ferragina,P. and Manzini,G. (2000). Opportunistic data structures with applications. In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pp. 390–398.

Grossi,R. *et al*. (2003). High-order entropy-compressed text indexes. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pp. 841–850.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*., **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Manber,U. and Myers,G. (1990). Suffix Arrays: A New Method for On-Line String Searches. In: *1st Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '90, pp. 319–327.

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

Menschaert,G. *et al*. (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics*, **12**, 1780–1790.

Muth,T. *et al*. (2014) DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J. Proteome Res*., **13**, 1143–1146.

Tanca,A. *et al*. (2013) Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS ONE*, **8**, 1–14.

Vaudel,M. *et al*. (2012) Current methods for global proteome identification. *Exp. Rev. Proteomics*, **9**, 519–532.

Vaudel,M. *et al*. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol*., **33**, 22–24.