Supporting information to "2018 YPIC challenge: A case study in characterizing an unknown protein sample"

Lindsay Pino[1], **Andy Lin**[1], **Wout Bittremieux**[*,1,2,3]

[1]Department of Genome Sciences, University of Washington, Seattle WA 98195, USA; [2]Department of Mathematics and Computer Science, University of Antwerp, 2020 Antwerp, Belgium; [3]Biomedical Informatics Network Antwerpen (biomina), 2020 Antwerp, Belgium

*Corresponding author: wout.bittremieux@uantwerpen.be, +32 3 265 34 07.

Contents

1 Simulated digestion of English dictionary 2

List of Figures

1 Simulated peptide length 3

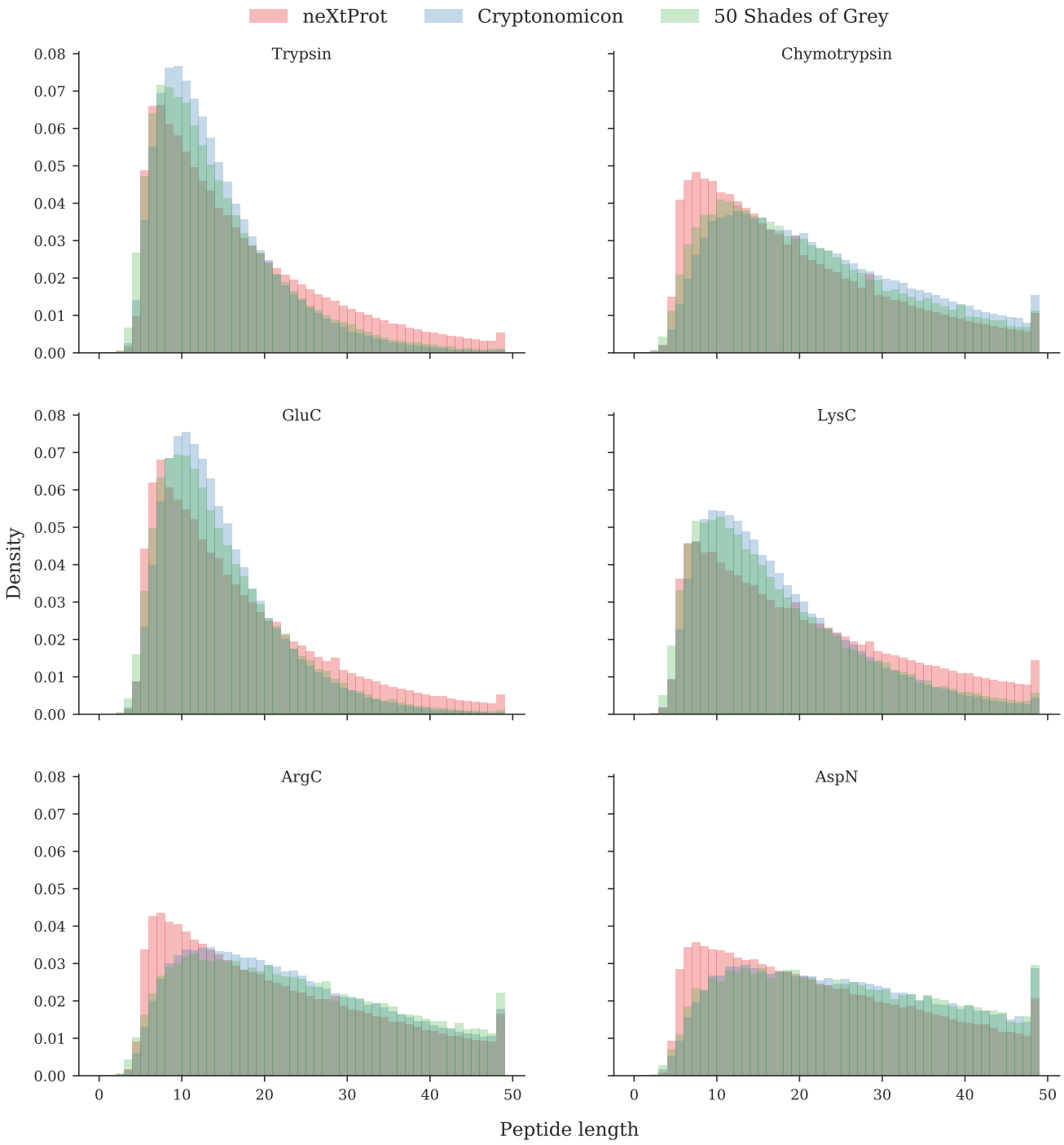
1 Simulated digestion of English dictionary

When analyzing a protein of unknown sequence, one key decision is to determine which digestion enzyme to use. To help inform our decision we simulated the digestion of various corpuses using multiple proteases to determine whether they would generally yield peptides whose lengths are amenable to detection by mass spectrometry (supplementary figure 1).

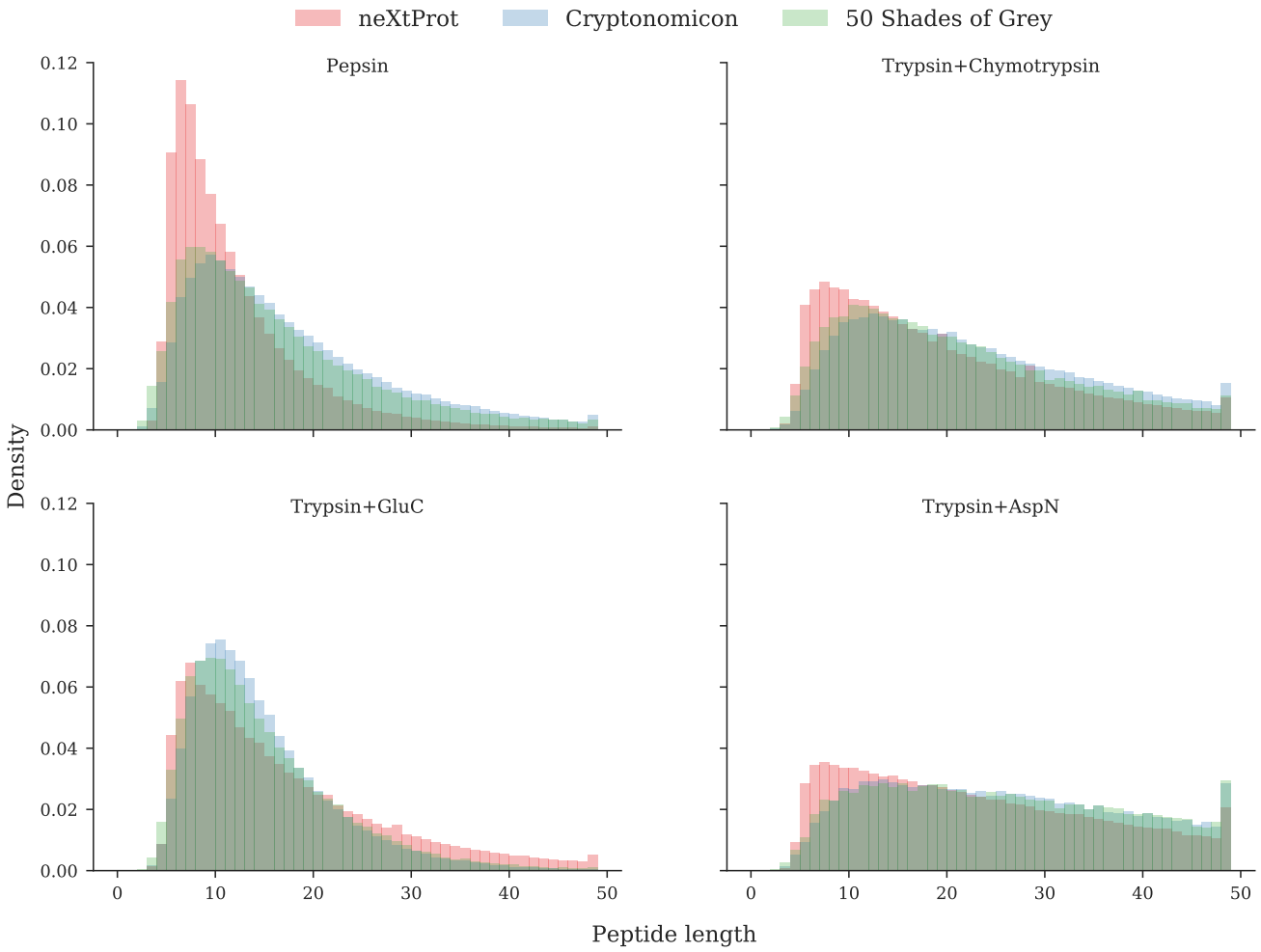
A simulated digestion of neXtProt [1], a database of human proteins, with trypsin while allowing for a single missed cleavage showed that a large portion of the resulting peptides will have a length between 6 and 10 amino acids, with the mode of the peptide length distribution at 7. While peptides of length 7 to 10 are perfectly amenable to detection by mass spectrometry, there is a significant tail in the distribution of peptide lengths. For example, approximately 15 % of the peptides consist of 30 or more amino acids, which is not ideal for detection by mass spectrometry.

An important issue with using neXtProt is that peptides in the human proteome are not expected to be a good proxy for 'peptides' found in human language. One possible reason is that the frequency of amino acids found in the human proteome is unlikely to be the same as the frequency of letters found in the English language. Therefore, we also simulated digestions of the text within two different English fiction novels with a varying vocabulary complexity: Cryptonomicon by Neal Stephenson [3] and 50 Shades of Grey by EL James [2]. Text files of the novels were found online and words containing the letters 'B' or 'K' were removed while the letters 'O' and 'U' were replaced by the letter 'K', in accordance with the challenge's instructions. Additionally, the letters 'J', 'X', and 'Z' were removed as these characters do not represent valid amino acids.

We found that a simulated tryptic digestion of these two novels yielded peptides whose lengths are slightly different than the length of peptides generated from neXtProt. The majority of English peptides is slightly longer than those generated from neXtProt (the mode of the peptide length distribution is 10 for Cryptonomicon and 7 for 50 Shades of Grey), while the English peptides include less very long peptides than neXtProt, as indicated by the right tail of the peptide length distributions. As a result, we found that trypsin is a suitable enzyme to digest the synthetic protein consisting of two English sentences. Additionally, we explored the digestion of neXtProt, Cryptonomicon, and 50 Shades of Grey with alternative proteases, including chymotrypsin, Glu-C, Lys-C, Arg-C, Asp-N, and pepsin, as well as combined digestions using two different proteases. These simulations again indicate that peptides generated from English are typically slightly longer than those generated from human proteins.



Supplementary Figure 1: Length of simulated peptides for various corpuses using various proteases including trypsin, chymotrypsin, Glu-C, Lys-C, Arg-C, Asp-N, pepsin, and a combined digestion with trypsin. NeXtProt is a database of human proteins, whereas Cryptonomicon and 50 Shades of Grey are two English fiction novels.



Supplementary Figure 1: Length of simulated peptides for various corpuses using various proteases including trypsin, chymotrypsin, Glu-C, Lys-C, Arg-C, Asp-N, pepsin, and a combined digestion with trypsin. NeXtProt is a database of human proteins, whereas Cryptonomicon and 50 Shades of Grey are two English fiction novels.

References

[1] Gaudet, P., Michel, P.-A., Zahn-Zabal, M., Cusin, I., et al. "The neXtProt Knowledgebase on Human Proteins: Current Status." In: *Nucleic Acids Research* 43.D1 (Jan. 28, 2015), pp. D764–D770. DOI: 10.1093/nar/gku1178.

[2] James, E. L. *Fifty Shades of Grey*. Fifty shades trilogy 1. New York: Vintage Books, 2015. 514 pp.

[3] Stephenson, N. *Cryptonomicon*. 1st ed. New York: Avon Press, 1999. 918 pp.