Proteome-wide Non-Cleavable Crosslink Identification Using Sparse Matrix Multiplication with MS Annika 3.0

Micha J. Birklbauer[1], Fränze Müller[2], Sowmya S. Geetha[3,4,5], Manuel Matzinger[2], Karl Mechtler[2,6,7], Viktoria Dorfer[1]

1 Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg
2 Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna
3 Max Perutz Labs (MPL), Vienna BioCenter (VBC), Vienna
4 Max Perutz Labs (MPL), Department of Chromosome Biology, University of Vienna, Vienna
5 Vienna BioCenter PhD Program, a Doctoral School of the University of Vienna and the Medical University of Vienna, Vienna BioCenter (VBC), Vienna
6 Institute of Molecular Biotechnology (IMBA), Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna
7 Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna

Der Wissenschaftsfonds.

Cross-linking mass spectrometry

has emerged as a prominent tool for identification of protein-protein interactions and for gaining insight into the structures of proteins. [1] We previously published MS Annika, a cross-linking search engine which can accurately identify cross-linked peptides in MS2 and MS3 spectra from a variety of different MS-cleavable crosslinkers. [2][3]

Non-cleavable crosslink identification

MS Annika 3.0 is an updated and improved version - that additionally to cleavable crosslinkers - supports identification of cross-linked peptides from non-cleavable reagents using a sparse matrix multiplication-based search algorithm as depicted in **Figure 1** and **2**. This algorithm can efficiently handle beyond human proteome-wide studies on commodity hardware. MS Annika 3.0 is available free of charge for Proteome Discoverer 3.1 at:

https://ms.imp.ac.at/?action=ms-annika

Results

We compared MS Annika 3.0 to other commonly used cross-linking search engines and show that MS Annika is on par or better in terms of crosslink identifications while providing a more robust false discovery rate (FDR) estimation, reporting 75% less false positives than competing tools on average (**Figure 3**). Most importantly we could show that MS Annika is able to accurately identify more than 430 unique crosslinks at 1% estimated FDR from an experiment with *C. elegans* nuclei, using the full C. elegans proteome of over 26 000 proteins for search (**Figure 4**), which allowed us to conclude a comprehensive structural analysis of the Box C/D complex, enhancing our understanding of its assembly and functional dynamics (**Figure 5**).

MS Annika 3.0

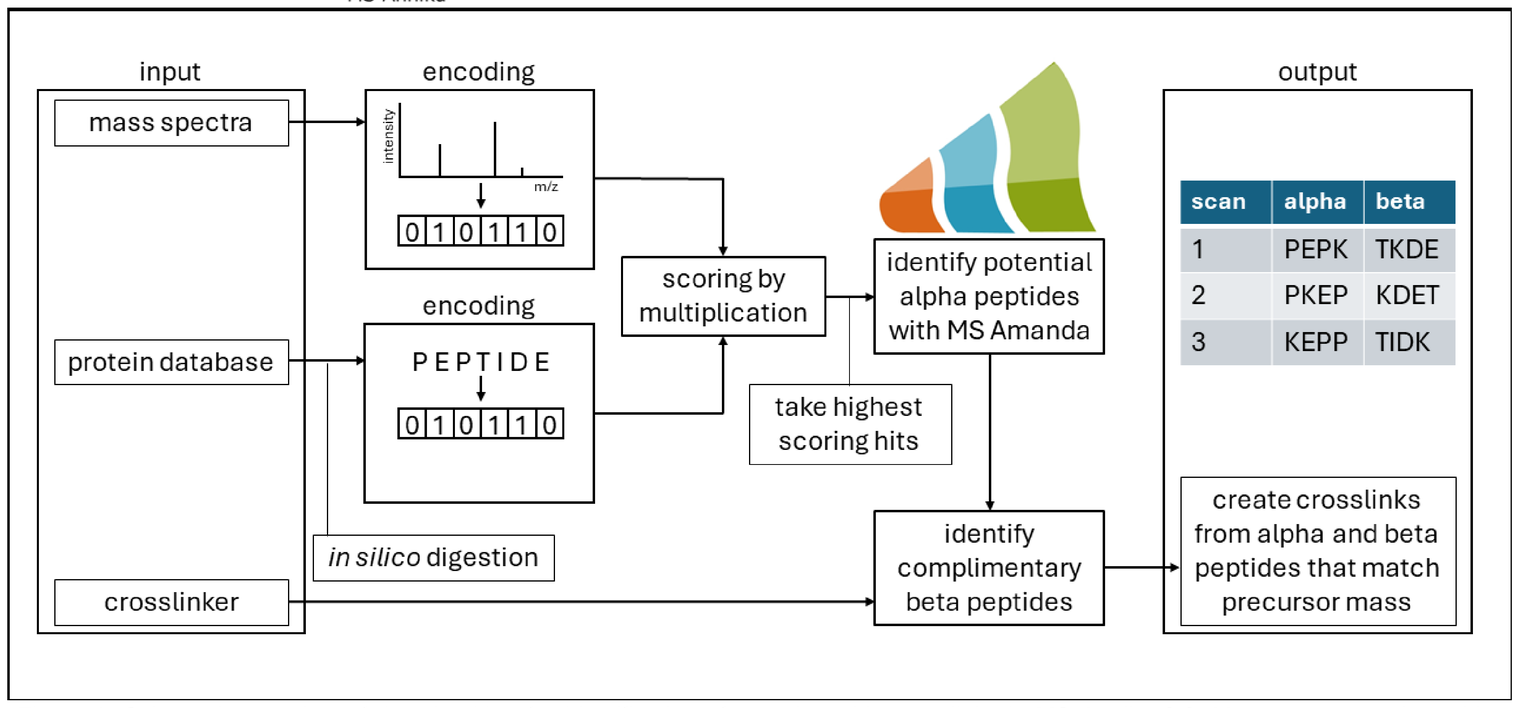

Figure 1: Schematic overview of the algorithm for identification of non-cleavable crosslinks in MS Annika 3.0. Mass spectra and peptides arising from the in-silico digestion of the protein database are encoded as sparse vectors (as shown in Figure 2) and subsequently scored by matrix multiplication. The highest scoring hits are considered for the identification of potential alpha peptides with our in-house developed peptide search engine MS Amanda [4][5]. Identified alpha peptides are used to find complimentary beta peptides and ultimately alpha and beta peptides matching the mass spectrum's precursor mass are combined to crosslinks.

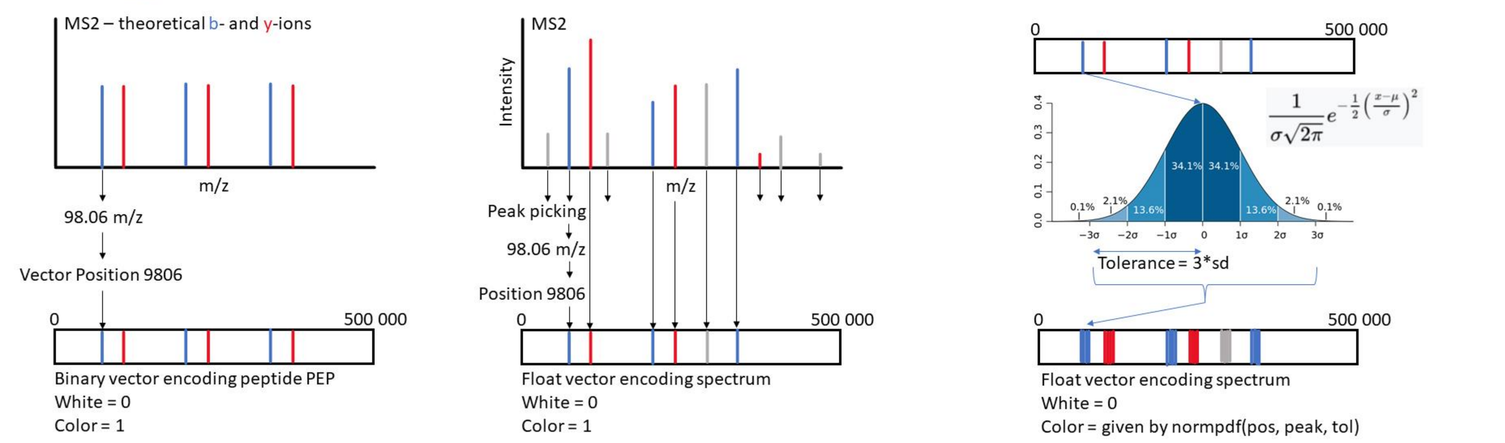

Figure 2: Encoding of peptides and mass spectra as sparse vectors. Peptide encoding: theoretical ions of the peptide are calculated and the resulting m/z values are binned in 0.01 Da windows to vector indices representing each ion. Mass spectrum encoding: peaks are binned using 0.01 Da windows, but every peak is modelled as gaussian distribution that incorporates instrument tolerance.

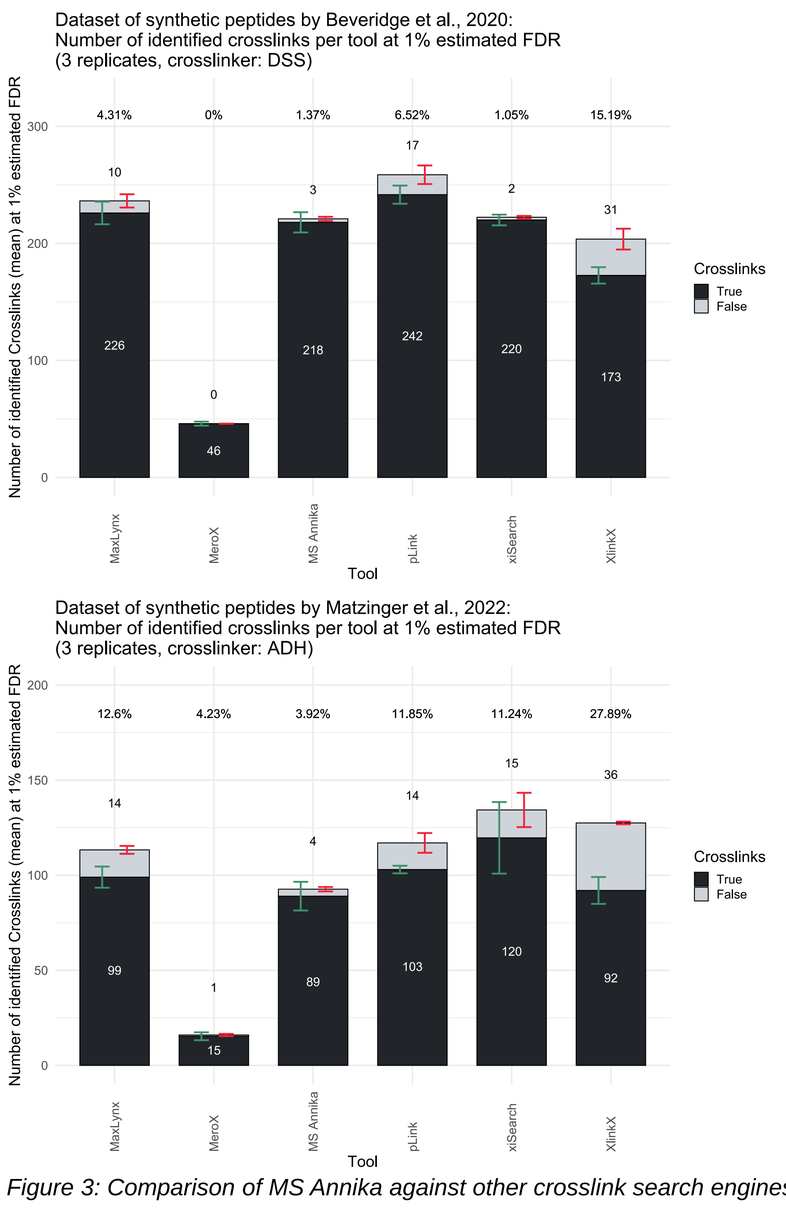

Figure 3: Comparison of MS Annika against other crosslink search engines.

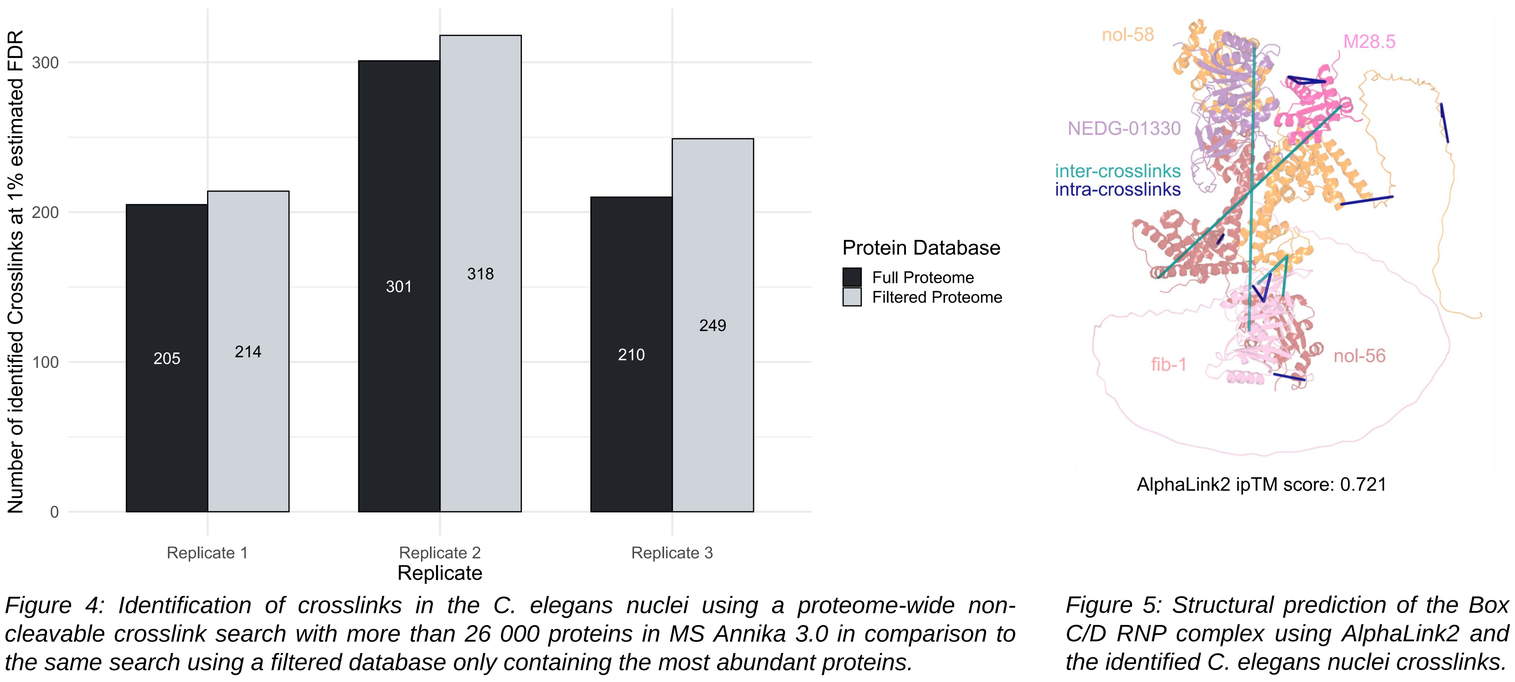

Figure 4: Identification of crosslinks in the C. elegans nuclei using a proteome-wide non-cleavable crosslink search with more than 26 000 proteins in MS Annika 3.0 in comparison to the same search using a filtered database only containing the most abundant proteins.

Figure 5: Structural prediction of the Box C/D RNP complex using AlphaLink2 and the identified C. elegans nuclei crosslinks.

Acknowledgements

This research project has received funding from the Austrian Science Fund (FWF), project number 35045.

References

[1] Matzinger, M. & Mechtler, K. (2021) Cleavable Cross-Linkers and Mass Spectrometry for the Ultimate Task of Profiling Protein-Protein Interaction Networks in Vivo. J. Proteome Res., 20, 78 – 93
[2] Pirklbauer, G. J., Stieger, C. E., Matzinger, M., Winkler, S., Mechtler, K. & Dorfer, V. (2021) MS Annika: A New Cross-Linking Search Engine. J. Proteome Res., 20, 2560 – 2569
[3] Birklbauer, M. J., Matzinger, M., Müller, F., Mechtler, K. & Dorfer, V. (2023) MS Annika 2.0 Identifies Cross-linked Peptides in MS2-MS3-Based Workflows at High Sensitivity and Specificity, J. Proteome Res., 22, 3009 – 3021
[4] Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S. & Mechtler, K. (2014) MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra, J. Proteome Res., 13, 3679 – 3684
[5] Dorfer, V., Strobl, M., Winkler, S. & Mechtler, K. (2021) MS Amanda 2.0: Advancements in the standalone implementation. Rapid Communications in Mass Spectrometry, 35, e9088

http://bioinformatics.fh-hagenberg.at
Micha J. Birklbauer: micha.birklbauer@fh-hagenberg.at