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

Micha Johannes Birklbauer[1], Fränze Müller[2], Manuel Matzinger[2], Karl Mechtler[2,3,4], and Viktoria Dorfer[1]

1) Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria
2) Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna, Austria
3) Institute of Molecular Biotechnology (IMBA), Vienna BioCenter (VBC), Vienna, Austria
4) Gregor Mendel Institute (GMI), Vienna BioCenter (VBC), Vienna, Austria

Introduction

Cross-linking mass spectrometry has emerged as a prominent tool for the identification of protein-protein interactions and for gaining insights into the native structures of proteins. Over the last decades the field of cross-linking has seen continuous growth and the development of cleavable cross-linking reagents allowed studying systems up to human proteome-wide scale. However, while non-cleavable crosslinkers exert properties attractive for biological applications, their use always has been limited by computational data analysis tools not being able to handle the extremely large search spaces of non-cleavable cross-linking experiments. We here present MS Annika 3.0, an updated and improved version of our cross-linking search engine that efficiently tackles this so-called *n-squared* search space problem and allows identification of non-cleavable crosslinks beyond human proteome-wide scale.

Methods

Identifying crosslinks from non-cleavable reagents requires smart handling of the given protein database as the search space grows with its square, potentially yielding trillions of peptide candidate pairs to consider. In MS Annika every peptide and mass spectrum is encoded as a high-dimensional sparse vector and the whole protein database can therefore be represented as a large sparse matrix which allows efficient scoring of millions of candidates within a fraction of a second by multiplying this matrix with a spectrum vector. This super-fast algorithm is the core of the MS Annika non-cleavable search, identifying likely peptide candidates and significantly decreasing the search space. The top candidates are re-scored with our in-house developed peptide search engine MS Amanda and possible peptide pairs are combined to crosslink-spectrum-matches. Results are validated using a transparent target-decoy approach and can additionally be exported for more sophisticated validation with tools like xiFDR.

Results and Discussion

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. 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.

Innovative aspects

- Enabling non-cleavable cross-linking experiments up to proteome-wide scale
- Representation of peptides and mass spectra as sparse vectors and fast search based on these representations allowing scoring of millions of peptide candidates
- Extremely efficient search in both speed and memory, enabling searches being performed on normal office laptops