BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 1 2014, pages 125–126
doi:10.1093/bioinformatics/btt617

Structural bioinformatics

Advance Access publication October 30, 2013

AnchorMS: a bioinformatics tool to derive structural information from the mass spectra of cross-linked protein complexes

Shannon L.N. Mayne and Hugh-G. Patterton*

Advanced Biomolecular Research Cluster, University of the Free State, P.O. Box 339, Bloemfontein 9300, South Africa

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Mass spectrometry is being increasingly used in the structural elucidation of mega-Dalton protein complexes in an approach termed MS3D, referring to the application of MS to the study of macromolecular structures. This involves the identification of cross-linked residues in the constituent proteins of chemically cross-linked multi-subunit complexes. AnchorMS was developed to simplify MS3D studies by identifying cross-linked peptides in complex peptide mixtures, and to determine the specific residues involved in each cross-link. When identifying cross-linked peptide pairs (CLPP), AnchorMS implements a mathematical model to exclude false positives by using a dynamic score threshold to estimate the number of false-positive peak matches expected in an MS/MS spectrum. This model was derived from CLPPs with randomly generated sequences. AnchorMS does not require specific sample labeling or pre-treatment, and AnchorMS is especially suited for discriminating between CLPPs that differ only in the cross-linked residue pairs.

Availability: AnchorMS was coded in Python, and is available as a free web service at cbio.ufs.ac.za/AnchorMS.

Contact: patterh@ufs.ac.za

Received on April 15, 2013; revised on October 20, 2013; accepted on October 22, 2013

1 INTRODUCTION

Many essential protein complexes are composed of multiple subunits, where an understanding of the structural arrangement and interactions between the constituent subunit proteins is needed for a mechanistic insight into the function of the complex. However, large multi-subunit protein complexes are often not amenable to structural elucidation by traditional methods such as X-ray crystallography or nuclear magnetic resonance. Mass spectrometry (MS) has been applied to enzymatic digests of chemically cross-linked protein complexes to identify cross-linked peptides. The identity of cross-linked peptides and the cross-linked residues indicates possible interaction surfaces. This allowed the determination of the orientation and contact points of subunits within large protein complexes, such as the 15-subunit 670-kDa complex of Pol II with TFIIF (Chen *et al*., 2010).

This approach is termed 'MS3D' and requires specialized software for the analysis of the MS and the MS/MS spectra (Rappsilber, 2011; Stengel *et al*., 2012). Perhaps the most challenging aspect of this type of analysis is the correct identification of low-abundance cross-linked peptide pairs (CLPPs) in complex sample mixtures. A number of software tools have been released to address this difficulty, but many are orientated toward niche experimental designs and sample treatments (Mayne and Patterton, 2011).

Furthermore, some implement sophisticated scoring schemes and probability models to estimate the rate of false-positive identification. Generally, these are either trained on specific empirical datasets (Li *et al*., 2012; McIlwain *et al*., 2010; Walzthoeni *et al*., 2012), or mathematically derived from theoretical principles of gas-phase peptide chemistry (Xu and Freitas, 2007; Zhang *et al*., 2002). However, if the experimental training dataset is insufficiently diverse, bias may be introduced into the resulting scoring scheme or probability model.

AnchorMS was developed as an alternative general tool for identifying CLPPs in MS and MS/MS spectra.

2 MATHEMATICAL MODEL FOR FALSE-POSITIVE MATCHES

The identification of CLPPs involves the comparison of observed spectra (*b*- and *y*-series) with spectra predicted for all possible CLPPs in the sample, which is based on the known sequences of the proteins in the sample and the cross-linking reagent specificity. False-positive peak matches occur when the peak value of non-identical peptide ions is sufficiently similar within a given tolerance. In complex sample mixtures, a fraction of the peptides and CLPPs, as well as their fragmentation ions, is likely to have similar mass to charge ratios (m/z values). We considered how often such similar m/z values occurred in a typical MS3D experiment, and simulated this *in silico*.

To avoid the introduction of bias, we used randomly generated CLPP sequences of varying lengths, assuming a uniform distribution of residues. A range of precursor charges (1+ to 5+) and tolerance values (in ppm and in Da) was also selected. Fragment charge states, up to and including the precursor charge, were considered. A decoy CLPP was generated for each randomly generated CLPP by shuffling the original sequence. This allowed the comparison of two CLPPs with identical amino acid compositions, a condition most likely to produce identical fragment masses, and thus similar MS/MS spectra. The predicted MS/MS spectrum for each CLPP and its decoy were compared, and the relationship between the number of false-positive peak matches was investigated as a function of precursor size, precursor charge and the matching tolerance.

Figure 1 shows a mathematical model for the number of false positives derived from the aggregated simulation data. A clear dependence was observed between the number of false peak

*To whom correspondence should be addressed.

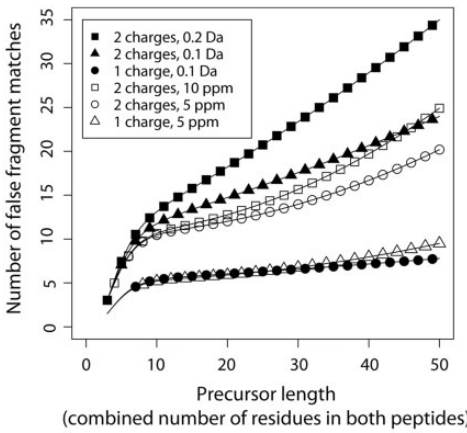

Fig. 1. The number of false-positive peak matches expected in the MS/MS spectrum as a function of the CLPP length. This relationship is shown for different precursor ion charges and matching tolerances. Each datum point is the average of 1000 random sequence/sequence shuffled pairs

matches and the CLPP length, ion charge and tolerance. The difference in the number of false-positive matches observed between absolute and relative tolerance modes was particularly noticeable. The model is not dependent on residue composition, and considers the aggregate of the residue compositions that were sampled.

AnchorMS implemented this calibrated mathematical model as a dynamic false-positive threshold, which is calculated for each putative CLPP spectrum assignment. Those that score below the threshold value are excluded by AnchorMS.

3 IDENTIFYING THE CROSS-LINKED RESIDUES

The identity and position of the cross-linked residues in a CLPP is clear-cut when there is only one residue that is reactive to the cross-linking reagent in each of the peptides. Here, the number of direct MS/MS peak-to-peak matches indicates the best CLPP identification. A similar method is used in the identification of post-translational modifications (Beausoleil *et al*., 2006). However, in the case of multiple reactive residues on each peptide, several cross-linking combinations are possible. AnchorMS approaches this ambiguity by also considering the number of unique MS/MS peaks associated with each possible combination. The cross-linking combination with the largest number of unique peaks that match the experimental MS/MS spectrum is selected as the correct match.

4 USING AnchorMS

AnchorMS is implemented as a set of Python scripts with a PHP: Hypertext Preprocessor (PHP) front-end. The source code is available under a GNU General Public License (GPL-3.0). The AnchorMS web service is freely accessible at cbio.ufs.ac.za/AnchorMS.

The AnchorMS web page displays a series of numbered activities to systematically guide the user through all steps required for an analysis. The residue sequences of the proteins involved in the cross-linking experiment must be uploaded as FASTA format, and the cross-linking reagent and protease used for digestion must be selected. Custom reagents and protease can also be defined. The MS and MS/MS spectra data files must be uploaded in text formats widely used in proteomics (ms2, mgf, mzXML, mzData or mzML) (Mayne and Patterton, 2011). The allowed post-translational modifications and the matching tolerance values must also be selected. The instrument mass accuracy must be supplied, which is applied to both precursor and fragment ions, and significantly influences the discriminating power of AnchorMS. AnchorMS will identify cross-linked peptides based on the MS peak list, and confirm the assignment using the MS/MS peak lists. The positions of cross-linked residues are then identified using the MS/MS peak lists. AnchorMS reports the identified CLPPs, as well as the residues that were cross-linked. The number of matching peaks in the MS/MS spectrum, as well as the number of unique matching peaks expected for a specific cross-linking configuration, where several are possible, is also reported. The false-positive match threshold value is also displayed for each identified CLPP. Additionally, the maximum distance between cross-linked residues is inferred and displayed. This value may be used for further structure analysis. When the analysis has been completed, a link to the web page listing the results is e-mailed to the user. This results page remains available for a period of time.

ACKNOWLEDGEMENT

The authors thank L. du Preez for his helpful comments on the manuscript.

Funding: University of the Free State Strategic Research Funds.

Conflict of Interest: none declared.

REFERENCES

Beausoleil,S.A. *et al.* (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.*, **24**, 1285–1292.

Chen,Z.A. *et al.* (2010) Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.*, **29**, 717–726.

Li,W. *et al.* (2012) SQID-XLink: implementation of an intensity-incorporated algorithm for cross-linked peptide identification. *Bioinformatics*, **28**, 2548–2550.

Mayne,S.L. and Patterton,H.G. (2011) Bioinformatics tools for the structural elucidation of multi-subunit protein complexes by mass spectrometric analysis of protein-protein cross-links. *Brief Bioinform.*, **12**, 660–671.

McIlwain,S. *et al.* (2010) Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. *J. Proteome Res.*, **9**, 2488–2495.

Rappsilber,J. (2011) The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.*, **173**, 530.

Stengel,F. *et al.* (2012) Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. *Mol. Cell. Proteomics*, **11**. R111.014027.

Walzthoeni,T. *et al.* (2012) False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods*, **9**, 901–903.

Xu,H. and Freitas,M.A. (2007) A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data. *BMC Bioinformatics*, **8**, 133–142.

Zhang,N. *et al.* (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**, 1406–1412.