BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2163–2164
doi:10.1093/bioinformatics/btr348

Structural bioinformatics

Advance Access publication June 11, 2011

Xwalk: computing and visualizing distances in cross-linking experiments

Abdullah Kahraman*, Lars Malmström and Ruedi Aebersold*

Department of Biology, Institute of Molecular Systems Biology, Swiss Federal Institute of Technology (ETH Zurich), CH-8093 Zurich, Switzerland

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Chemical cross-linking of proteins or protein complexes and the mass spectrometry-based localization of the cross-linked amino acids in peptide sequences is a powerful method for generating distance restraints on the substrate's topology.

Results: Here, we introduce the algorithm Xwalk for predicting and validating these cross-links on existing protein structures. Xwalk calculates and displays non-linear distances between chemically cross-linked amino acids on protein surfaces, while mimicking the flexibility and non-linearity of cross-linker molecules. It returns a 'solvent accessible surface distance', which corresponds to the length of the shortest path between two amino acids, where the path leads through solvent occupied space without penetrating the protein surface.

Availability: Xwalk is freely available as a web server or stand-alone JAVA application at http://www.xwalk.org.

Contact: abdullah@imsb.biol.ethz.ch; aebersold@imsb.biol.ethz.ch

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

Received on March 18, 2011; revised on May 18, 2011; accepted on June 6, 2011

1 INTRODUCTION

In computational structural biology, distance restraints from chemical cross-linking experiments have so far been employed as an upper limit on the Euclidean distance between a pair of cross-linked amino acids (Kaimann *et al.*, 2008; Shandiz *et al.*, 2007). However, deducing the 'cross-linkability' of an amino acid pair by measuring the length of a Euclidean distance vector disregards the fact that the vector often penetrates segments of the protein. Potluri *et al.* (2004) have recognized this problem and implemented a short-cut algorithm that computes the shortest path between two cross-linked amino acids by using vertices from a protein surface triangulation and convex hull, while Zelter *et al.* (2010) have explicitly modeled the cross-linker molecule onto existing protein structures. We have implemented Xwalk, which resembles the approach taken by Potluri *et al.,* but instead uses grids and a search algorithm to compute the length of the shortest path (Fig. 1), which shall be referred to as solvent accessible surface distance (SASD). Our code is the only of its kind being open source and available in form of a web server.

*To whom correspondence should be addressed.

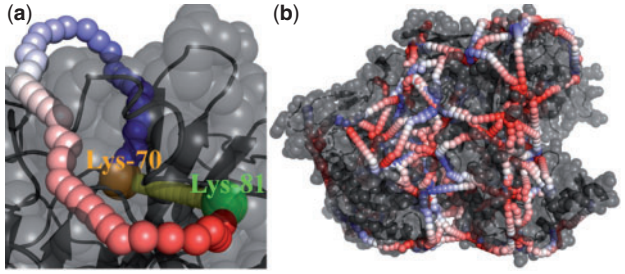

Fig. 1. (**a**) Shortest SASD path illustrated on the example of human prothrombin (PISA-Id: 1dx5, chain E). The C_β atoms of Lys-70 (orange sphere) and Lys-81 (green sphere) have a Euclidean distance of 9.1 Å (yellow vector), which by value would have been in the cross-link range for DSS or BS^3. However, the shortest path with an SASD of 59.2 Å reveals that the Euclidean distance vector actually penetrates the protein, leaving the only option to connect both amino acids via a long detour over the protein surface (chain of spheres colored blue to red for distances of 0–59 Å., respectively). (**b**) Argonaut protein from the RNA-induced silencing complex (RISC) with 271 virtual intra-protein cross-links. Both figures were rendered with PyMOL (http://www.pymol.org).

2 IMPLEMENTATION

Xwalk was written in the JAVA programming language. It is based on the CleftXplorer modelling package (Kahraman *et al.*, 2010) and uses the breath-first search algorithm on local grid representations of the protein and its surrounding solvent to calculate the shortest SASD path between two atoms of two amino acids on the protein surface (Fig. 1). Xwalk can run in two modes (Supplementary Material), namely in validation mode in which Xwalk verifies experimentally measured cross-links on an existing protein structure, or in production mode, in which Xwalk reports a list of *in silico* predicted theoretically possible virtual cross-links (vXL) that might be observed in a cross-linking experiment. Both modes are identical except for step 1.c in the Supplementary Material, which in production mode is replaced by the specification of generic identifiers of amino acids to be cross-linked *in silico*.

Xwalk checks that the cross-linked amino acids and the entire SASD path is solvent accessible. Furthermore, Xwalk takes the dynamic disorder of protein segments within X-ray structures into account. Therefore, it increases the maximum distance range of a cross-linker spacer arm by the sum of the mean atomic displacement of the cross-linked amino acids. The mean atomic displacement $\langle x \rangle$ of a single amino acid is inferred from the Debye–Waller formula $B = 8\pi^2 \left\langle x^2 \right\rangle$, where B is the atomic B factor of the cross-linked

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

amino acid as given in a PDB file. Moreover, Xwalk holds the option to discard all side chains from the distance calculation to account for their conformational change when reacting with the cross-linker molecule. At the same time, the solvent accessible surface area is expanded by increasing the solvent radius to 2.0 Å to avoid path calculations through molecular 'tunnels' that arise due to the side chain depletion.

The output of Xwalk is either a list of vXL or a PyMOL script (http://www.pymol.org) displaying the shortest SASD path as a list of dummy atom entries in a PDB file (Fig. 1). The list of vXL is a list of atom pairs sorted by SASD with information on their amino acid number and name, chain identifier and atom name, along with their distances in the PDB sequence, their Euclidean distance and SASD. Furthermore, an *in silico* trypsin digestion can be requested, in which case the associated shortest tryptic peptide sequences are reported. The source code of Xwalk is available under a Creative Commons license together with the executable at http://www.xwalk.org. The same site provides also an easy to use web interface to the basic functionalities of Xwalk with a Jmol viewer applet (http://www.jmol.org) as a visualization tool for the shortest paths.

3 CROSS-LINKING THE PDB

A cross-linking experiment can only yield cross-links if the proteins under study have particular amino acids that are within a certain distance from each other and solvent accessible. However, the number of cellular proteins that have such characteristics is not known neither is the number of cross-links one can expect per protein or protein complex.

To estimate these numbers, we have run Xwalk in production mode on a non-homologous protein dataset and simulated the most common cross-linking reagents DSS and BS3 with both having a maximum distance cut-off of 22.4 Å (11.42 Å N–N distance in DSS $+2\times5$ Å CB-NZ distance in lysine), discarding side chains and measuring distances between Cβ atoms. The protein dataset consisted of 1621 X-ray protein structures from the PISA server (Krissinel and Henrick, 2007), where protein homology was defined by the H-level or the superfamily-level in the CATH (Orengo *et al.*, 1997) or SCOP data base (Murzin *et al.*, 1995), respectively. Each protein in the dataset was selected to have the highest annotated domain coverage and the highest number of protein chains within its homology class, while setting an upper bound of 20 protein chains for oligomeric protein complexes.

In the entire dataset, we calculated 30 266 unique vXL (excluding vXL that are found between equivalent amino acids in homomers, as these cannot be distinguished in real cross-linking experiments). Of these, 25 751 were intra-protein and 4515 were inter-protein vXL. The number of the unique intra- and inter-protein vXL increases for one to five unique protein chains from 15 to 45 and 2 to 24, respectively. In all, 18% of proteins had no vXL at all, while 40 protein structures had more than 100 vXL. The highest number of unique vXL in the dataset, namely 271 vXL, was found in the monomeric structure of the RNA-induced silencing complex (RISC) associated argonaut protein (PDB-Id: 1u04, see Fig. 1b) and in the bacteriophage DNA polymerase–DNA terminal protein complex (PDB-Id: 2ex3).

The benefit of Xwalk and SASD becomes apparent when the above analysis is repeated with the conventional Euclidean distance. The repetition with a 22.4 Å Euclidean distance cutoff resulted in 65 447 vXL, i.e. more than twice as many as with SASD. Of these, 35 181 vXL had a SASD larger than 22.4 Å that differed on average by >8 Å. Of these, >100 vXL's had a distance difference of >50 Å (see exemplary Fig. 1a). These numbers suggest that Xwalk is able to reduce the false positive prediction of cross-links by $>50\%$. The large discrepancy emphasizes the importance of an adequate model for a cross-linker molecule in cross-linking experiments.

Despite the smaller number of false positives with SASD, we have observed that the number of vXL usually exceeds the number of experimental cross-links by at least one order of magnitude (Leitner *et al.*, 2010). Most of the theoretically predicted but experimentally unobserved cross-links may be missed because of their low abundance, unfavorable chromatographic, ionization and fragmentation properties or due to their unsuitable peptide length. Another issue arises in cases in which segments of the protein structure are missing, such as in intrinsically disordered proteins or proteins with flexible loops. These regions will have missing atom coordinates that are currently ignored by Xwalk and may lead to lower SASD than expected.

ACKNOWLEDGEMENTS

We thank Manfred Claassen, Alexander Leitner, Franz Herzog, Thomas Walzthöni for their valuable input into details of the Xwalk algorithm.

Funding: ETH Zurich; the Commission of the European Communities through the PROSPECTS consortium (EU FP7 projects 201648, 233226); SystemsX.ch – The Swiss Initiative for Systems Biology in part.

Conflict of Interest: none declared.

REFERENCES

Kahraman,A. *et al.* (2010) On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins* **78**, 1120–1136.

Kaimann,T. *et al.* (2008) Molecular model of an alpha-helical prion protein dimer and its monomeric subunits as derived from chemical cross-linking and molecular modeling calculations. *J. Mol. Biol.*, **376**, 582–596.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.

Leitner,A. *et al.* (2010) Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell Proteomics*, **9**, 1634–1649.

Murzin,A.G. *et al.* (1995) Scop - a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. *et al.* (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Potluri,S. *et al.* (2004) Geometric analysis of cross-linkability for protein fold discrimination. *Pac. Symp. Biocomput.*, **9**, 447–458.

Shandiz,A.T. *et al.* (2007) Intramolecular cross-linking evaluated as a structural probe of the protein folding transition state. *Biochemistry*, **46**, 13711–13719.

Zelter,A. *et al.* (2010) Isotope signatures allow identification of chemically cross-linked peptides by mass spectrometry: a novel method to determine interresidue distances in protein structures through cross-linking. *J. Proteome Res.*, **9**, 3583–3589.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016