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

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.

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 upper limit of the Euclidean distance vector, and permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

Supplementary information: Supplementary data are available at Bioinformatics online.

© 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.
A cross-linking experiment can only yield cross-links if the proteins (PDB-Id: 2ex3). The bacteriophage DNA polymerase–DNA terminal protein complex (PDB-Id: 1u04, see Fig. 1b) and in monomeric structure of the RNA-induced silencing complex (RISC) of unique vXL in the dataset, namely 271 vXL, was found in the protein structures had more than 100 vXL. The highest number respectively. In all, 18% of proteins had no vXL at all, while 40 for one to five unique protein chains from 15 to 45 and 2 to 24, Of these, 25 751 were intra-protein and 4515 were inter-protein vXL. As these cannot be distinguished in real cross-linking experiments). vXL that are found between equivalent amino acids in homomers, for oligomeric protein complexes. Its homology class, while setting an upper bound of 20 protein chains 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 Walzthomi 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 lysyls in binding pockets of unrelated proteins. Proteins 78, 1120–1136.

Kaiman,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., 38, 582–596.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. J. Mol. Biol., 372, 714–737.

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, 586–596.

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

Pollard,S. et al. (2004) Geometric analysis of cross-linkability for protein fold discrimination. Pac. Symp. Biocomput., 9, 474–485.

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

Zehner,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.