Data in brief 28 (2020) 104834

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A non-redundant dataset of inter-residue lysine-lysine solvent accessible surface distances in homo-oligomeric protein complexes

Aljaž Gaber

Department of Chemistry and Biochemistry, Faculty of Chemistry and Chemical Technology, University of Ljubljana, Večna pot 113, Ljubljana SI 1000, Slovenia

ARTICLE INFO

Article history:
Received 9 October 2019
Received in revised form 7 November 2019
Accepted 11 November 2019
Available online 18 November 2019

Keywords:
Chemical cross-linking
Homo-oligomers
Solvent accessible surface distances

ABSTRACT

This article contains inter-residue solvent accessible surface distances between lysines in a comprehensive dataset homo-oligomeric protein complex structures, downloaded from 3D Complex database. Solvent Accessible Surface Distances were calculated with Jwalk algorithm. To avoid unnecessary redundancy due to symmetry, we calculated only distances originating from the first subunit of each protein complex. Redundancy was further reduced by including only the shortest of the two possible inter-subunit alternatives in the final non-redundant dataset. For each protein complexes we also calculated weight ob subunits, number of lysines radius of gyration and average distances.

This data can be used for structural analyses of homo-oligomeric protein complexes and for further optimization of distance-based restraints, such as those based on data obtained from chemical cross-linking coupled with mass spectrometry.

E-mail address: aljaz.gaber@fkkt.uni-j.si.

https://doi.org/10.1016/j.dib.2019.104834

Specification Table

Subject	Structural Biology
Specific subject area	Protein structure analysis
Type of data	A collection of files containing inter-residue lysine-lysine solvent accessible surface distances and a table containing average distances and structural parameters of corresponding protein complexes.
How data were acquired	Distances were calculated with software Jwalk v 1.3 [1].
Data format	Raw and Analyzed
Parameters for data collection	We calculated inter-residue distances between lysines of a non-redundant set of homo-oligomeric protein complexes with high resolution. Only residue pairs that had at least one endpoint residue in the first subunit were considered to avoid unnecessary redundancy due to symmetry.
Description of data collection	Distances were calculated with software Jwalk v 1.3 [1]. Jwalk calculated Solvent accessible surface distances (SASDs) in four steps: Placing protein on a grid, calculating the solvent accessible surface, calculating solvent accessible paths between C_α-atoms of lysine-lysine residue pairs of interest, using Bread-First Search to find the shortest distance for each residue pair. Euclidean distances were calculated as the distance of the straight line between C_α-atoms.
Data source location	University of Ljubljana, Ljubljana, Slovenia
Data accessibility	With the article.
Related research article	Aljaž Gaber, Gregor Gunčar, Miha Pavšič Proper evaluation of chemical cross-linking-based spatial restraints improves the precision of modeling homo-oligomeric protein complexes BMC Bioinformatics 10.1186/s12859-019-3032-x

Value of the Data

- This data is useful for structural analyses of homo-oligomeric protein complexes.
- Calculation of solvent accessible surface distances presented in this data set is often computationally too expensive and thus avoided. This dataset enables investigators to use a large pre-calculated set of solvent accessible surface distances.
- This data set can be used for investigation of symmetry and for further optimizations of distance-based restraints in computational modeling of homo-oligomeric protein complexes.

1. Data

Most proteins form homo-oligomeric protein complexes. One of the most commonly employed methods for studying their structure is chemical cross-linking coupled with mass spectrometry (XL-MS). We have recently investigated different scoring approaches for proper evaluation of distance-based restraints based on XL-MS data in modelling [2]. The data set of distances used in this analysis is presented in this report.

The raw_distances.tar.gz file contains 13,110 text files, one for each PDB in the analyzed data set. Each text file is named XXXX_X_crosslink_list.txt, where XXXX_X is the unique PDB biounit identifier. Files contain Solvent Accessible Surface Distances (SASDs) and Euclidean distances (EUCs) for each lysine-lysine residue pair in each PDB. Both intra-subunit and inter-subunit alternative are included. To reduce redundancy due to symmetry, only pairs that have at least one endpoint in the first subunit (*i.e.* the first chain) are included. Each line in these document presents a standard Jwalk [1] output for a residue pair, containing Index, Model, Atom1, Atom2, SASD and EUC plus the difference between the distances and the intra/inter-subunit assignation. To ensure our list is comprehensive and that no pairs are filtered out, SASDs of non-accessible residue pairs are assigned a distance of 9999 Å.

The non-redundant_distances.tar.gz file also contains 13,110 text files. Files were generated from corresponding XXXX_X_crosslink_list.txt and named XXXX_X_non_redundant.txt. Raw distances were grouped by residue numbers. For example: distances 1A-2A, 1A-2B, 2A-1B, which were initially written as three separate outputs were now concatenated to a single line containing residue numbers (Res1 and Res2), intra-subunit distances (SASD_intra and Euc_intra) and inter-subunit distances (SASD_inter and EUC_inter).

Table 1 contains a list of all 13,110 complexes along with some basic characteristics and average distances per complex. Each row contains the following information: unique PDB biounit identifier (PDB_code), total number of amino-acid residues in the complex (AA_total), the number of amino-acid residues per subunit (AA_per_sub), total weight of the complex (weight), weight per subunit (weight_per_sub), Radius of gyration (Rg), symmetry (corrected_sym) and number of subunits (corrected_nsub), number of lysine residues per subunit (num_lys), average intra- and inter-subunit SASD and EUC (Avg_intra_SASD, Avg_intra_EUC, Avg_inter_SASD, Avg_inter_EUC, respectively). The number of amino-acid residues per subunit, symmetry and number of subunits was taken from 3DComplex database [3], other properties were calculated as described below.

2. Experimental design, materials, and methods

2.1. Extraction of representative data set of homo-oligomeric protein complexes

First, we downloaded all protein complexes from 3D Complex database (version 6, based on PDB database on March 1st, 2015) that matched the following criteria: homo-oligomers (homomers only) with at most 90% sequence similarity (QS90) and resolution of 2.5 Å or lower. Second, we excluded structures that were found to have incorrect stoichiometry during a community based manual inspection (annotations “YES” and “PROBYES” in PiQSi database [4]). Third, we removed proteins that had more than one structure (with different stoichiometries) in the data set (1 521). During our calculations we also excluded structures with split polypeptide chains (593) and structures for which the distances could not have been calculated in three days (13). The final data set has 13 110 structures.

2.2. Calculation of raw distances

Distances were calculated with Jwalk v 1.3 [1]. Jwalk calculates SASDs between C_α-atoms and uses a Breath-First Search to find the shortest distance. During the process of calculating SASD we also calcualted Euclidean distance as the distance of the straight line between C_α-atoms. We calculated inter-residue distances between all lysine residue pairs that had at least one residue in the first subunit. This was done to avoid redundancy due to symmetry, which is inherent to most homo-oligomeric protein complexes. All residues pairs were initially assigned a distance of 9999.0 Å to prevent Jwalk algorithm from filtering out the results if SASD could not be calculated because either of the residues in the pair was buried or otherwise non-accessible.

2.3. Filtering out non-redundant distances

Inter-subunit distances in raw distance files are redundant, because there are at least two alternatives for each residue pair. In a dimer, for example, both distances 1A-2B and 1B-2A are included, although one is a symmetric equivalent of the other. In non-redundant distance files, we only kept the shortest of the distances.

Furthermore, whenever SASD was not calculated, the number was replaced with slash (/). This happened for three reasons: either of residues was non-accessible to the solvent, either of residues was missing 3D coordinates in the PDB because it was part of a flexible loop or terminal regions or because both residues had the same residue number (intra-subunit connection does not exist). If SASD could not be calculated, EUC was also removed.

The code used for filtering is available in get_non-redundant_distances.py.

2.4. Calculation of additional protein complex properties and average distances

Weight per subunit (weight_per_sub) was calculated with Biopython [5]. This weight and the number of amino-acid residues (AA_per_sub) were multiplied with the number of subunits to obtain the total weight of protein complex and the total number of amino-acid residues (weight and AA_total, respectively). Radius of gyration calculation was adapted from PyMOL (https://pymolwiki.org/index.php/Radius_of_gyration). The number of lysines was extracted from the number of unique residues in non-redundant distance files. Average intra- and inter-subunit SASD and EUC (Avg_intra_SASD, Avg_intra_EUC, Avg_inter_SASD, Avg_inter_EUC, respectively) were calculated from non-redundant distance file.

The code used in analysis is available in data_analysis.py.

Acknowledgments

This work was supported by the Slovenian Research Agency (grant J1-7119).

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104834.

References

[1] J.M.A. Bullock, J. Schwab, K. Thalassinos, M. Topf, The importance of non-accessible crosslinks and solvent accessible surface distance in modelling proteins with restraints from crosslinking mass spectrometry, Mol. Cell. Proteom. (2016) 12–15, https://doi.org/10.1074/mcp.M116.058560.

[2] A. Gaber, G. Gunčar, M. Pavšič, Proper evaluation of chemical cross-linking-based spatial restraints improves the precision of modeling homo-oligomeric protein complexes, BMC Bioinf. 20 (2019) 464, https://doi.org/10.1186/s12859-019-3032-x.

[3] E.D. Levy, J.B. Pereira-Leal, C. Chothia, S.A. Teichmann, 3D complex: a structural classification of protein complexes, PLoS Comput. Biol. 2 (2006) e155, https://doi.org/10.1371/journal.pcbi.0020155.

[4] E.D. Levy, PiQSi: protein quaternary structure investigation, Structure 15 (2007) 1364–1367, https://doi.org/10.1016/j.str.2007.09.019.

[5] P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J.L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (2009) 1422–1423, https://doi.org/10.1093/bioinformatics/btp163.