*Published online 28 October 2016* *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D243–D249***
*doi: 10.1093/nar/gkw976*

# LinkProt: a database collecting information about biological links

**Pawel Dabrowski-Tumanski[1,2,†], Aleksandra I. Jarmolinska[2,3,†], Wanda Niemyska[2,4,†], Eric J. Rawdon[5], Kenneth C. Millett[6] and Joanna I. Sulkowska[1,2,*]**

[1]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093, Warsaw, Poland, [2]Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097, Warsaw, Poland, [3]College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Banacha 2c, 02-097, Warsaw, Poland, [4]Institute of Mathematics, University of Silesia, Bankowa 14, 40-007, Katowice, Poland, [5]Department of Mathematics, University of St. Thomas, Saint Paul, MN 55105, USA and [6]Department of Mathematics, University of California, Santa Barbara, CA 93106, USA

Received September 08, 2016; Revised October 06, 2016; Editorial Decision October 07, 2016; Accepted October 13, 2016

## ABSTRACT

**Protein chains are known to fold into topologically complex shapes, such as knots, slipknots or complex lassos. This complex topology of the chain can be considered as an additional feature of a protein, separate from secondary and tertiary structures. Moreover, the complex topology can be defined also as one additional structural level. The LinkProt database (http://linkprot.cent.uw.edu.pl) collects and displays information about protein links — topologically non-trivial structures made by up to four chains and complexes of chains (e.g. in capsids). The database presents deterministic links (with loops closed, e.g. by two disulfide bonds), links formed probabilistically and macromolecular links. The structures are classified according to their topology and presented using the minimal surface area method. The database is also equipped with basic tools which allow users to analyze the topology of arbitrary (bio)polymers.**

## INTRODUCTION

Nowadays the existence of protein chains with complex topologies is firmly confirmed. Non-trivial structures identified in proteins with complex topologies include (open) knots and slipknots (1), complex lassos (2,3), cysteine knots (4) and various other structures defined by taking into account protein-metal bonds (5,6). Their function still puzzles researchers, but their statistical analysis is facilitated by various databases (7–10).

In most cases, topological complexity enters in between the secondary and tertiary structural level, requiring one to define which bonds are 'topologically significant' (6). However, the complexity can also be defined on the quaternary structural level, as the topologically non-trivial arrangement of whole chains. One example of this phenomenon is in various domain-swapped proteins (11,12). On the other hand, the observation of an appropriate 'spatial proximity' of protein chains in an arc repressor has led to the design of a knotted protein (13). These examples are crucially different from other proteins with defined quaternary structure e.g. haemoglobin. One therefore needs a precise measure of such a difference. The natural way to define the topological complexity of a set of linear chains (e.g. proteins) is the notion of a link, i.e. the generalization of the notion of a knot into structures with many components. In fact, one can generalize the standard chain closure techniques known for knotted proteins (1,14,15) to define links in proteins. Such links are probabilistic, meaning that there is a certain probability of identification of a definite link type, depending on the number of different chain closures realizing this particular topology.

Although this is an entirely new characterization of proteins, similar studies were performed independently during the writing of this work (16). LinkProt also includes those results. Moreover, there exist at least two other ways of defining links in proteins. One can also define 'macromolecular links' in which each component is built out of many separate chains, e.g. in the chainmail structure of bacteriophages and virus capsids (17–19), or in protein catenanes (20–23). The existence of links in such systems is known to introduce extra stability to the capsids of bacteriophages.

The last way to define links in proteins is to consider covalent loops formed by the main protein chain and disulfide bonds as individual chains. Since components of such links are well-defined, there is no ambiguity in defining their topology and therefore we call these structures 'determin-

*To whom correspondence should be addressed. Tel: +48 22 55 43 675; Fax: +48 22 822 02 11 (Ext 320); Email: jsulkowska@chem.uw.edu.pl
†These authors contributed equally to the paper as first authors.

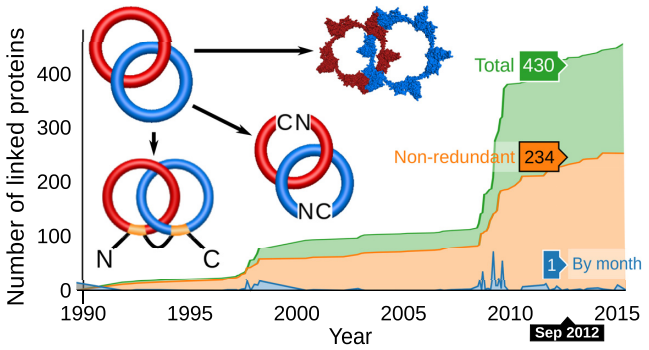


**Figure 1.** The timeline of the number of new LinkProt entries. The total number (green curve), total number of sequentially non-redundant structures (orange curve) and non-redundant entries in each month (blue curve) are presented. In the top-left corner various realizations of the Hopf link topology are presented — the deterministic link (bottom-left), probabilistic (center) and macromolecular (top-right).

istic links'. The first approach for finding such links was due to Mislow (5,6). However, in his papers, he investigated components containing at least two disulfide bonds each, considering the existence of linked covalent loops closed by only one disulfide bridge as improbable. The first example of a two-chain protein with linked covalent loops (formed by the main chain closed by only one disulfide bridge) was given in 2007 (24) and many examples of linked covalent loops within a single protein chain have been discovered only recently (as shown by Dabrowski-Tumanski et al) as a result of the analysis of proteins with lassos (2,3). Similarly to the macromolecular links, deterministic links were shown to introduce additional stability to the structure (24). The three different realizations (deterministic, probabilistic and macromolecular) of the simplest link, i.e. the Hopf link, are presented in the top-left corner of Figure 1.

The number of known proteins with links is growing, and almost each month entirely new proteins (i.e. sequentially non-homologous to any known structure) are added to the RCSB PDB database (Figure 1). However, up to now, such structures have never been assembled, analyzed and classified in a way that would enable their systematic analysis. The LinkProt database (http://linkprot.cent.uw.edu.pl) aims to fill this gap. The goal of the database is to store and analyze the link topologies (either probabilistic or deterministic) of one-, two-, three- and four-chain complexes and the macromolecular link topologies for appropriate complexes. To analyze each system, we combine the strength of the *minimal surface analysis* introduced in the study of complex lasso proteins (2,3) with a open-chain knot determination approach, which has been used, for example, in defining knots in proteins (1,7). As a result, the LinkProt database is a powerful tool for the statistical analysis of protein topology, useful for biologists, biotechnologists, biophysicists and mathematicians. Moreover, the database is equipped with the unique server submission function which allows users to analyze the link topology of any two-chain protein structure or arbitrary polymer. To ensure the robustness of link classification, HOMFLY-PT polynomials for almost 24 000 topologically different links (including chirality and orientation) were calculated. The database is compatible with other databases and servers concerning non-trivial topologies in proteins, such as KnotProt (7) (http://knotprot.cent.uw.edu.pl) and LassoProt (3) (http://lassoprot.cent.uw.edu.pl), facilitating its easy use in conjunction with them.

## MATERIALS AND METHODS

### Link detection

To identify the link type, the HOMFLY-PT polynomial is calculated (25,26) using its implementation by Ewing and Millett (27). Moreover, a complete *minimal surface analysis* (2,3), revealing the piercings through the surfaces spanning the closed loops, is performed. The HOMFLY-PT polynomial distinguishes between different chiralities and orientations of the links, allowing one to split the major topological classes into smaller subclasses. The polynomial can be calculated only for closed loops. In the case of deterministic or macromolecular links, the closed loops are defined in a straightforward fashion based on the position of the cysteine, amide, ester or thioester bridge. In the case of probabilistic links, the chains have to be closed first. This is done by connecting the termini of each chain on a huge sphere surrounding the protein structure. To avoid crossings of added intervals on the sphere, conversely to standard procedure used in protein knot determination, the termini of each chain are connected in one point (separate for each chain) (14,15). The links are detected based on the native positions of CA atoms.

### Link classification and notation

Rolfsen notation (28) is used for all prime links with fewer than 10 crossings in their minimal crossing presentation, unless a classical name exists (e.g. Hopf link). Hoste-Thistlethwaite notation (http://katlas.org, http://indiana.edu/~linkinfo/) is used for prime links with 10 or 11 crossings. For non-alternating links with higher numbers of crossings (which have not yet been fully catalogued), arbitrary names were given (like L14n1, flower link, etc). Depending on the chirality and orientations (which are determined based on the protein chain directions), this major topological type can be split into smaller subtypes, denoted by arbitrarily assigned numbers (e.g. $6_1^2.1$). The composition of links and unions of links are denoted by # and $\bigcup$, respectively. In total, polynomials for almost 24 000 topologically different links were calculated. A list of assigned names and a corresponding schematic image of each link is available in the 'Classification of links' section in the online manual.

### Proteins in database

The database contains the topologically non-trivial structures found in the entire RCSB PDB database, including crystal, Nuclear Magnetic Resonance and Electron Microscopy structures, also including structures with missing atoms. In such cases, the gap in the main chain is filled with a straight line segment. The database self-updates every Wednesday. For each structure, its substructures consisting of one-, two-, three- and four chains are analyzed. Moreover, all of the analyzed (trivial and non-trivial) structures are available as a list of entries.

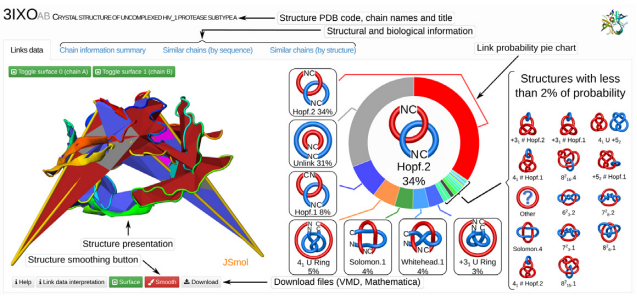


**Figure 2.** Example page of a link data interpretation. Left, structure of proteins with minimal surfaces spanning the covalent loops. Right, pie chart presenting possible link types for the HIV-1 protease with PDB code 3IXO. All detected links whose probability is shown on the pie chart are displayed as icons.

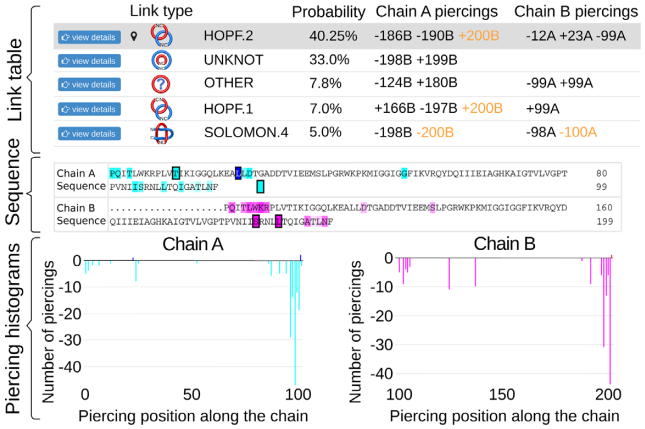


**Figure 3.** Top, the table of possible link types with the link probability and piercing residues. Middle, the colored sequence presenting the piercing residues. Bottom, the piercing histograms.

### Graphical presentation

The pie charts and histograms are generated using Plotly. Structures and surfaces are visualized using JSmol (HTML5/JavaScript version). The input files for Mathematica and VMD (29) software are calculated by appropriate transformation of JSmol code by our software. The smoothed structure (with the same link type) is obtained by computing the running average of the positions of three consecutive atoms.

### Database technicalities

The database is written in Python with a Flask framework dynamically generating HTML pages using Apache2 with WSGI. The data are stored using a SQLite3 database. Information about the proteins are downloaded from the PDB using RESTful services, and the PFAM and EC data using the SIFTS service. The whole service is installed on multicore Linux nodes.

## DATABASE INTERFACE AND DATA PRESENTATION

### Single link data presentation

Each structure is presented in the same manner, consisting of (i) a protein structure presentation and link probability pie chart (Figure 2), and (ii) a table of the possible link types, a list of the sequences and images of the piercing histograms (Figure 3). Moreover, the user can change tabs and see the basic biological and structural information for a given protein. All these are described in the following sections.

*Structure presentation and link probability pie chart.* The structure of the protein is shown in the left panel of Figure 2. The user can use all of the standard JSmol capabilities, including rotation or zoom. Since the structures are analyzed using the minimal surface analysis, the triangulated minimal surface is also shown for each link component. The surfaces show explicitly where the link-forming piercings are located. The surfaces can be turned on and off by clicking on the appropriate buttons above the structure. To facilitate easier viewing, the user can smooth the structure by clicking the corresponding button below the presentation.

The pie chart in the right panel of Figure 2 shows the various link types that have been detected. Upon pointing the cursor at any part of the pie chart, the corresponding link type is presented in the center of the chart. Clicking on that part of the chart fixes the link type and changes the JSmol presentation to show the surfaces corresponding to the chosen link type. For each link type, the structure with a given surface can be downloaded for further analysis as a VMD or Mathematica file by clicking the Download button. In some cases, one of the parts of the pie-chart contains an 'Other' class. Currently, the LinkProt database distinguishes all prime links up to 11 crossings (in minimal crossing presentation) for links with four or fewer components, most composite links up to 11 crossings and some links with more than 11 crossings. Nevertheless, some chain closures create complicated link types whose HOMFLY-PT polynomials are not included in the library. In such cases, this type is classified as 'Other'.

*Link table, sequence and histograms.* The table of possible link types is presented below the graphical presentation (Figure 3). The table contains the miniature of the link picture, the names of the chains forming the link and the indices of residues piercing the surface for each chain (the most probable ones). In some cases, one of the piercings is created by the segment extending from one of the termini. To distinguish this type of piercing, the corresponding data is shown in orange.

Below the table, the sequence of each chain is presented. In the sequence, the piercing residues are highlighted, where the intensity of the color denotes the probability that this particular residue is the piercing one (depending on the termini extension, the piercing residue can vary). Hovering the mouse pointer over a letter in the sequence results in the index in the sequence appearing over that letter. Moreover, clicking on the letter results in a violet bead being displayed in the appropriate location in the structure above. In this way, the user can localize in the structure as many residues as desired. Clicking one more time on the letter deactivates appropriate bead. The same information is also presented below the sequence as histograms for each of the piercing residues. In contrast to the sequence presentation, the his-

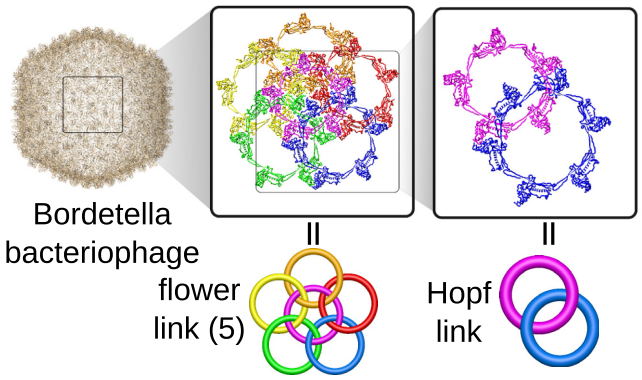


**Figure 4.** Example of macromolecular links with their schematic representations: left panel flower link (5), right panel Hopf link.

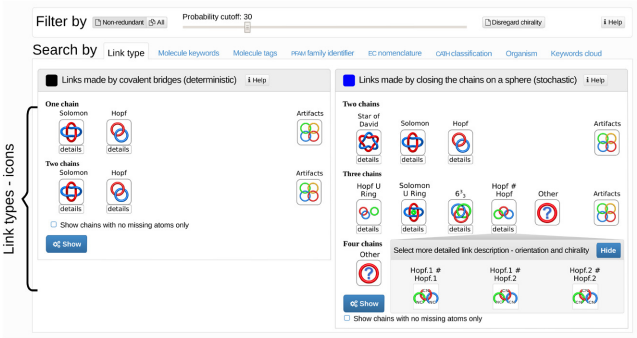


**Figure 5.** Presentation of the Search tab. The top of the page shows filters: sequence similarity, the probability cut-off (for probabilistic links) and disregard chirality (to distinguish subtypes in a given link type, depending on its chirality and chain orientation). Middle, left and right panels present icons of deterministic and probabilistic links. The search panel also contains macromolecular links and link types which are not shown here.

togram shows also the sign of the piercing (as defined in (3)).

*Structural and biological information.* Apart from the strictly topological data, LinkProt also shows basic biological and structural information, such as the EC number (for enzymes), the PFAM and CATH family of the protein or chains similar by sequence or by structure. This information can be obtained by choosing the appropriate tab at the top of the page.

*Macromolecular links.* The macromolecular links form a separate category (Figure 4). These structures are displayed similarly. However, for ease of viewing, no minimal surface is shown and only the simplifications of the chains are displayed. In such structures, it is very common for some chains to be homologous, in which case only the sequences of essentially different chains are presented. Since the link type is determined unequivocally for macromolecular links, instead of the piercing histogram (which would not add any new information), the animated structure is shown.

## Searching structures and browsing the database

The database has two major tabs: Browse and Search. In the Browse mode the user can display all of the entries currently deposited in the database. In the default view, the PDB code, chain names and PDB title are displayed along with the prescribed topological type. The structures deposited in the LinkProt database can also be listed (PDB code and chain names) with their topological type. This can be useful for automated, computer-aided analysis. To facilitate the understanding of the symbols representing link types, the user can also use the graphical view, in which the PDB code and the chain names are displayed along with images of the links. The different techniques for link detection are denoted using colored squares (black—deterministic, blue—probabilistic).

To find a protein with a desired topological type, the user can use the Search tab, where different filters are designed to display selected types of structures. In the Search tab, three panels are displayed—containing deterministic, probabilistic and macromolecular links (Figure 5). In each panel, the possible link topological types are presented. By default, the non-redundant set of proteins (measured by sequence similarity) are displayed. This is convenient to various statistical analyses. The user can, however, display all proteins contained in the database by clicking the All button in the top-left corner of the page.

In the case of probabilistic links, the exact number of link types depends on the probability cut-off set displayed on the slider. By default, the links with at least 30% probability are displayed. The default cut-off of 30% is used to ensure the robustness of the link types — links with at least 30% probability are conserved in the homology cluster, at least for two-chain links. The probability cut-off can be set individually by using the slider, which also changes the content of all other data. This gives users the capability to adjust the list of the proteins to their needs. Note that in special cases, the same structure can be classified into multiple sets (e.g. Hopf link and Solomon link). Note also that the number of possible link types increases with the number of components. Therefore, the link types for structures with four chains have, in general, lower probability than the link types for two-component links.

The use of the HOMFLY-PT polynomial in link identification allows us to distinguish subtypes within a given link type, depending on the link's chirality and chain orientation. Since these characteristics can be biologically relevant (as the biology imposes an orientation on the chains), the user can select only the structures with a desired topological subtype by clicking on the Details button below the general link type picture. Note that the probability cut-off adjusted at the top part of the page applies to the subtype only. To display the Hopf link structures with 50% probability while disregarding the possible orientations (subtypes), for example, one must click on the Disregard chirality button at the top of the page. After selecting the desired types/subtypes and clicking the Show button, the list of structures fulfilling the given conditions will be displayed.

Besides the geometrical and topological filters described previously, the user can also search structures according to their biological features, such as EC number, PFAM access code, molecular keyword, etc. These can be chosen in the

separate tab in the Search part of the database. Note that the probability cut-off set in the main tab applies also for the displayed set, so the user can easily perform searches to understand the biological meaning and uniqueness of a given link type.

Finally, the user can check the topology of a desired PDB structure by typing its PDB code in the field on the top right corner of any subpage. After doing so, the list of non-trivial structures stored in the LinkProt database will be displayed.

*Artifact structures.* The LinkProt database analyzes all possible structures from the RCSB PDB database, including those with gaps. The gaps are filled with straight line segments, which, in the case of large gaps, can change the topological type. Therefore, the possibly topologically non-trivial structures with large gaps (larger than six residues) are classified separately as artifacts.

## RESULTS

As of the fall of 2016, the LinkProt database contains 456 structures, 101 representing deterministic and 357 probabilistic links. However, the number of such structures constantly rises (Figure 1), showing the increasing need of analyzing such structures. Moreover, the trend for non-redundant structures has mostly the same slope as the total number of linked structures deposited in the RCSB PDB database. Thus, such structures are getting more attention and even more complex structures could exist and will be crystallized in future. A comparison of their topological and biological or biophysical properties may reveal previously unnoticed influences of topology on function and evolution of proteins. Some preliminary observed correlations are also shown below.

### Classification and new results

The LinkProt database contains link classifications for up to four chains and categorizes separately probabilistic and deterministic links. Those data are further classified based on sequence similarity. In the set of non-redundant proteins, the one-chain deterministic links are the Hopf or Solomon links and there are two deterministic two-chain links, only one of which was reported previously (24).

Analysis of probabilistic links indicates that above the probability cut-off (30%) no new type of links are observed. For this value, which is used as the default, there are three major two-chain types (Hopf, Solomon and Star of David links) and taking into account their chirality and orientation this number rises to five classes. In the case of three-chain links, four major types are represented: two of them consisting of three chains ($6_3^3$ and connected sum of two Hopf links—Hopf # Hopf), while the two others are split sums of the Hopf or Solomon links with a ring. When considering the chirality and orientation, the number of classes rises to six. There are no four-chain structures with the probability of a single link type above the 30% cut-off level. However, the human apolipoprotein a-i (PDB code 1AV1) has 20% probability of the link type L12n1.

Enzymes constitute roughly half of the analyzed structures, as can be seen from statistics shown on the EC nomenclature tab of the Search page. This percentage is far less than in the case of protein knots. Most of the link-enzymes are involved in DNA, RNA or nucleotide handling.

On the other hand, the great majority of probabilistic link proteins come from the human immunodeficiency virus, making such structures especially interesting. Structurally, such proteins are classified as Mainly Beta (roughly 2/3 of proteins classified by CATH database), most of them being Beta Barrel proteins. Both of these statistics can be accessed through the appropriate tabs of the Search page.

## SERVER — ANALYZE OWN (BIO)POLYMERS

Apart from the database, LinkProt gives users the unique opportunity to check the topology of their own protein or any other (bio)polymer consisting of one or two chains closed deterministically. The user can upload the file in two formats: PDB or an ASCII file consisting of atom index and its X, Y and Z coordinates. The server can automatically detect the location of SS bridges or they can be chosen manually. The output is presented in the same manner as in the database, including the pie-chart and downloadable files. The results are stored for 2 weeks.

### Server advanced options

By default, the server validates the structure by checking the CA-CA distance and the bridge length. This feature can be turned off in the advanced server options, which can be especially useful when analyzing non-protein structures.

## APPLICATIONS

The ability to identify links in proteins opens an entirely new way of describing proteins. For biologists and biotechnologists, the link topology was shown to introduce additional stabilization (Dabrowski-Tumanski *et al*). Therefore, the existence of the protein in the LinkProt database can explain its chemo-physical properties, such as elevated stability or give a clue to how such stability can be introduced artificially. This applies both to deterministic links and to macromolecular links, especially in virus capsids. On the other hand, the probability of link formation can be yet another feature of proteins which distinguishes between similar families. Moreover, the topology can influence other biological properties and such influences can be explored using the tools available in the LinkProt database. The probability of link formation can also be a powerful parameter (reaction coordinate) for biophysicists simulating many chain complexes.

The server can be used by biotechnologists to design structures with a desired topology, which can be further used, e.g. in drug design or to build superstable molecular complexes as in the case of macromolecular links or RNA origami. Moreover, it can be used to uncover and study links in proteins made via small particles or ions. Since our technique is fast, it can be used by CAPRI or CASP participants to verify predicted structures of a molecular complex or even to choose a complex with a new fold for following competitions.

Finally, our huge links library can be useful for mathematicians and physicists wishing to quickly identify the

topology of a system (set of curves in space), just by submitting to the LinkProt server. Also, the link table given in the online manual can be treated as yet another table of links. Mathematical link databases, such as the Knot Atlas (http://katlas.org) do not show information for different chiralities or orientations of the components within a given link type. In physical settings, such information may be critical and LinkProt's classification is sensitive to chirality and component orientations.

## ONLINE DOCUMENTATION

The database is supported by extensive, well-organized online documentation which includes: detailed descriptions of the link detection methods, a list of different topological types and subtypes identified by LinkProt with examples, instructions for how to search, browse and interpret the results from the database, as well as containing a section on applications. Database statistics, such as lists of non-trivial structures, the newest complex linked protein deposited in the PDB and an art gallery are also included.

### Comparison with other databases

To our knowledge, the LinkProt is the first database about links formed by one, two, three or four protein chains or by networks of protein chains. In comparison to other databases of knotted structures, such as Protein Knots (http://knots.mit.edu/), pKNOT (http://pknot.life.nctu.edu.tw) and KnotProt (http://linkprot.cent.uw.edu.pl) this database: (i) presents all detected linking, deterministic and probabilistic, for all molecular complexes in the PDB, (ii) offers search options using molecular keywords, molecular tags, PFAM access code and CATH topology, (iii) detects broken protein chains and uses this information in the analysis, and (iv) allows for user submission of structures for linking analysis.

## SUMMARY

The LinkProt database collects information about links—topologically non-trivial structures made by up to four chains, as well as complexes of chains (e.g. capsids) and represents their complexity using linking and the minimal surface area method. The database presents deterministic links (with loops closed, e.g. by two disulfide bonds), probabilistic links (formed by the whole chain) and macromolecular links. The presented link classification is extensive, taking into account all available biological information and thus goes beyond the standard link classifications used in the mathematics. Moreover, LinkProt is equipped with a submission server which enables users to analyze their own (bio)polymers. To our knowledge, this is the first database about biological links in proteins. Furthermore, it can be used to analyze links formed in small particles or ions. We hope that the versatility of the database will stimulate new discoveries and methods in various areas of research.

## ACKNOWLEDGEMENTS

We would like to thank Pawel Pasznik for his help with constructing the server and for helpful discussions about database construction. Also, we want to thank Joanna Macnar and Grzegorz Rajchel for help in preparing figures and GR for evaluating the macromolecular links data.

## FUNDING

National Science Centre [#2012/07/E/NZ1/01900 to J.I.S.]; European Molecular Biology Organization Installation Grant [#2057 to J.I.S.]; Foundation for Polish Science [130/UD/SKILLS/2015 to W.N.]; University of Warsaw [#120000-501/86-DSM-112 700 to P.D.-T.]; National Science Foundation [#1418869 to E.J.R.]. Funding for open access charge: Foundation for Polish Science (Fundacja na rzecz Nauki Polskiej) [130/UD/SKILLS/2015], Scientific work financed from the budget for science in the years 2016-2019, 0003/ID3/2016/64 Ideas Plus.
*Conflict of interest statement*. None declared.

## REFERENCES

1. Sułkowska,J.I., Rawdon,E.J., Millett,K.C., Onuchic,J.N. and Stasiak,A. (2012) Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Nat. Acad. Sci. U.S.A.*, **109**, E1715–E1723.
2. Niemyska,W., Dabrowski-Tumanski,P., Kadlof,M., Haglund,E., Sułkowski,P. and Sulkowska,J.I. (2016) Complex lasso: new entangled motifs in proteins. *Scient. Rep.*, in press.
3. Dabrowski-Tumanski,P., Niemyska,W., Pasznik,P. and Sulkowska,J.I. (2016) LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res.*, **44**, W383–W389.
4. Craik,D.J., Daly,N.L. and Waine,C. (2001) The cystine knot motif in toxins and implications for drug design. *Toxicon*, **39**, 43–60.
5. Liang,C. and Mislow,K. (1994) Knots in proteins. *J. Am. Chem. Soc.*, **116**, 11189–11190.
6. Liang,C. and Mislow,K. (1995) Topological features of protein structures: knots and links. *J. Am. Chem. Soc.*, **117**, 4201–4213.
7. Jamroz,M., Niemyska,W., Rawdon,E.J., Stasiak,A., Millett,K.C., Sułkowski,P. and Sulkowska,J.I. (2015) KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.*, **43**, D306–D314.
8. Kolesov,G., Virnau,P., Kardar,M. and Mirny,L.A. (2007) Protein knot server: detection of knots in protein structures. *Nucleic Acids Res.*, **35**(Suppl. 2), W425–W428.
9. Lai,Y.-L., Yen,S.-C., Yu,S.-H. and Hwang,J.-K. (2007) pKNOT: the protein KNOT web server. *Nucleic Acids Res.*, **35**(Suppl. 2), W420–W424.
10. Lai,Y.-L., Chen,C.-C. and Hwang,J.-K. (2012) pKNOT v. 2: the protein KNOT web server. *Nucleic Acids Res.*, **40**, W228–W231.
11. Liu,Y. and Eisenberg,D. (2002) 3D domain swapping: as domains continue to swap. *Protein Sci.*, **11**, 1285–1299.
12. Shameer,K., Shingate,P.N., Manjunath,S., Karthika,M., Pugalenthi,G. and Sowdhamini,R. (2011) 3DSwap: curated knowledgebase of proteins involved in 3D domain swapping. *Database*, **2011**, bar042.
13. King,N.P., Jacobitz,A.W., Sawaya,M.R., Goldschmidt,L. and Yeates,T.O. (2010) Structure and folding of a designed knotted protein. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 20732–20737.
14. Millett,K.C., Rawdon,E.J., Stasiak,A. and Sułkowska,J.I. (2013) Identifying knots in proteins. *Biochem. Soc. Trans.*, **41**, 533–537.
15. Sułkowska,J.I., Noel,J.K., Ramírez-Sarmiento,C.A., Rawdon,E.J., Millett,K.C. and Onuchic,J.N. (2013) Knotting pathways in proteins. *Biochem. Soc. Trans.*, **41**, 523–527.
16. Baiesi,M., Orlandini,E., Trovato,A. and Seno,F. (2016) Linking in domain-swapped protein dimers. *Sci. Rep.*, **6**, 33872.
17. Duda,R.L. (1998) Protein chainmail: catenated protein in viral capsids. *Cell*, **94**, 55–60.
18. Gan,L., Speir,J.A., Conway,J.F., Lander,G., Cheng,N., Firek,B.A., Hendrix,R.W., Duda,R.L., Liljas,L. and Johnson,J.E. (2006) Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure*, **14**, 1655–1665.

19. Helgstrand,C., Wikoff,W.R., Duda,R.L., Hendrix,R.W., Johnson,J.E. and Liljas,L. (2003) The refined structure of a protein catenane: the HK97 bacteriophage capsid at 3.44 Å resolution. *J. Mol. Biol.*, **334**, 885–899.
20. Cao,Z., Roszak,A.W., Gourlay,L.J., Lindsay,J.G. and Isaacs,N.W. (2005) Bovine mitochondrial peroxiredoxin III forms a two-ring catenane. *Structure*, **13**, 1661–1664.
21. Lee,B.I., Kim,K.H., Park,S.J., Eom,S.H., Song,H.K. and Suh,S.W. (2004) Ring-shaped architecture of RecR: implications for its role in homologous recombinational DNA repair. *EMBO J.*, **23**, 2029–2038.
22. Zimanyi,C.M., Ando,N., Brignole,E.J., Asturias,F.J., Stubbe,J. and Drennan,C.L. (2012) Tangled up in knots: structures of inactivated forms of E. coli class Ia ribonucleotide reductase. *Structure*, **20**, 1374–1383.
23. Smeulders,M.J., Barends,T.R., Pol,A., Scherer,A., Zandvoort,M.H., Udvarhelyi,A., Khadem,A.F., Menzel,A., Hermans,J., Shoeman,R.L. *et al.* (2011) Evolution of a new enzyme for carbon disulphide conversion by an acidothermophilic archaeon. *Nature*, **478**, 412–416.
24. Boutz,D.R., Cascio,D., Whitelegge,J., Perry,L.J. and Yeates,T.O. (2007) Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. *J. Mol. Biol.*, **368**, 1332–1344.
25. Freyd,P., Yetter,D., Hoste,J., Lickorish,W.R., Millett,K. and Ocneanu,A. (1985) A new polynomial invariant of knots and links. *Bull. Am. Math. Soc.*, **12**, 239–246.
26. Przytycki,J.H. and Traczyk,P. (1988) Invariants of links of Conway type. *Kobe J. Math.*, **4**, 115–139.
27. Ewing,B. and Millett,K.C. (1991) A load balanced algorithm for the calculation of the polynomial knot and link invariants. In: Rassias,GM (ed). *The Mathematical Heritage of CF Gauss*. World Scientific Publishing, Singapore, pp. 225–266.
28. Rolfsen,D. (1976) *Knots and Links, Mathematics Lecture Series*. AMS Chelsea Publishing, American Mathematical Society, Providence, Rhode Island.
29. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.