D454–D458 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 9 November 2017*
doi: 10.1093/nar/gkx1084

KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling

Guillaume Postic[1,2,3,4], Jérôme Gracy[5], Charlotte Périn[1,2,3,4], Laurent Chiche[5] and Jean-Christophe Gelly[1,2,3,4,*]

[1]INSERM, U 1134, DSIMB, 6, rue Alexandre Cabanel, 75739, Paris Cedex 15, France, [2]Université Paris Diderot, Sorbonne Paris Cité, UMR_S 1134, 75739 Paris, France, [3]Institut National de la Transfusion Sanguine, 75739 Paris, France, [4]Laboratory of Excellence GR-Ex, 75739 Paris, France and [5]CNRS UMR 5048, INSERM U1054, Centre de Biochimie Structurale, Université Montpellier, 34090 Montpellier, France

Received September 15, 2017; Revised October 14, 2017; Editorial Decision October 16, 2017; Accepted October 20, 2017

ABSTRACT

Knottins, or inhibitor cystine knots (ICKs), are ultra-stable miniproteins with multiple applications in drug design and medical imaging. These widespread and functionally diverse proteins are characterized by the presence of three interwoven disulfide bridges in their structure, which form a unique pseudoknot. Since 2004, the KNOTTIN database (www.dsimb.inserm.fr/KNOTTIN/) has been gathering standardized information about knottin sequences, structures, functions and evolution. The website also provides access to bibliographic data and to computational tools that have been specifically developed for ICKs. Here, we present a major upgrade of our database, both in terms of data content and user interface. In addition to the new features, this article describes how KNOTTIN has seen its size multiplied over the past ten years (since its last publication), notably with the recent inclusion of predicted ICKs structures. Finally, we report how our web resource has proved usefulness for the researchers working on ICKs, and how the new version of the KNOTTIN website will continue to serve this active community.

INTRODUCTION

Inhibitor cystine-knots (ICKs) form a family of ultra-stable miniproteins, found in a wide variety of organisms, with confirmed and potential medical applications. They are characterized by the presence of at least three interwoven disulfide bridges, which form an intramolecular knot and confer them structural and functional resistance to high temperature, enzymatic degradation, extreme pH and mechanical stress. ICKs are ∼30–50 residue long, which make them easily accessible to chemical synthesis. The loops connecting the disulfide bridges show a high variability of sequence, which results in a broad range of functions covered by ICKs, from channel blockage to inhibition of enzymes. All these properties have lead to use ICKs as scaffolds for the engineering of various pharmaceutical and imaging agents (1,2). Examples are the US FDA-approved Linzess® (linaclotide, Allergan/Ironwood Pharmaceuticals; www.linzess.com) (3), which is used to treat irritable bowel disease with constipation, and Prialt® (ziconotide, Azur Pharma; www.prialt.com), which is used to treat chronic pain.

The ICK family is made of three groups: (i) knottins, which represents the majority of ICKs—so that the two names are often used interchangeably—and are characterized by having disulfide bonds between the knot-forming cysteines III and VI going through cystines I-IV and II-V; (ii) cyclotides, which have the same disulfide connectivity as knottins, but have their backbone cyclized via an N-terminal to C-terminal peptide bond; and (iii) growth factor cystine-knots, which is the smallest group of the ICK family and includes proteins with a different connectivity than knottins and cyclotides. Since its launch in 2004, our KNOTTIN database (4) concentrates sequences, structures and bibliographical data about ICKs, except the few proteins belonging to the growth factor cystine-knots group. Data about cyclotides can also be found in the more specialized database CyBase (5,6).

Our knottin-dedicated database is valuable given the specific properties of these miniproteins in terms of sequence, structure and function. Indeed, the knottins very low sequence identity between families and high sequence plasticity (except for the cysteines) require specific procedures to correctly identify and classify these proteins. In the same way, their structural pseudoknot, very low content of regular secondary structures and particular protein hy-

[*]To whom correspondence should be addressed. Email: jean-christophe.gelly@univ-paris-diderot.fr

drophobic core formed by disulfide bonds, require adapted 2D and 3D representation. The whole diversity of knottins functions (such as neurotransmitters, analgesics, anthelmintics, anti-erectile dysfunction, antimalarials, antimicrobials, antitumor agents, protease inhibitors, toxins and insecticides) also has to be properly represented and documented, by gathering relevant functional annotations and bibliographic data. Finally, the active community of researchers working on the applied and theoretical aspects of knottins needs easy access to softwares dedicated to the analysis of knottins, which is provided by the computational tools available on our KNOTTIN website. Here, we present an upgraded version of KNOTTIN (www.dsimb.inserm.fr/KNOTTIN/), 10 years after its last publication (7). This report describes the new features of the database, the way it has evolved and shown usefulness over the past decade, and future developments.

DESCRIPTION

Website navigation

The content of the KNOTTIN database can be browsed with the horizontal navigation bar of the web interface. The 'Experimental 3D structures' and 'Sequences & 3D models' menus give access to experimental and theoretical models, respectively. In these pages, users can select one or several proteins, to visualize either their aligned sequences or 3D structures. The latter can also be done under the 'Sequence alignments' menu, which displays pre-compiled multi-sequence alignment files. In each of these three menus, the knottins are grouped by families, which have been determined based on sequence similarity and biological activity. Each knottin of the database is also categorized based on the length of its loops between the knot-forming cysteines (i.e. five loops for the knottins, and six for the cyclotides). Thus, a sequence (a)b.c(d)[e] is attributed to each knottin, the letters 'a' to 'e' being the loops lengths. This nomenclature of ICKs has been introduced with the initial release of the database in 2004.

Querying the system

The database can also be searched. Under the 'Sequence search' menu, amino acids sequences can be searched with the BLAST algorithm. The theoretical and experimental models of KNOTTIN can also be searched, under the 'Conformation search' menu, based on different structural features, such as torsion angles, secondary structure content and solvent accessibility. Sequences and structures can also be accessed via the 'Information search' menu, which offers users the possibility of searching the database by using different criteria, such as family, keywords, crystallography techniques or the aforementioned knottins nomenclature. The database also gathers the literature about the knottin proteins, which can be accessed under the 'Article search' menu based on criteria, such as articles authors, publication date and keywords. Bibliographic data about knottins functions, folding, synthesis, modeling, and biotechnological applications can also be browsed under the different sections of the 'Bibliography' menu.

Specific tools

Besides the database, the KNOTTIN website is also a platform that regroups under the 'Tools' menu softwares dedicated to the analysis of knottin proteins. These computational methods are Knoter1D and Knoter3D (7), which are aimed at identifying knottins based on their 1D sequences and 3D structures, respectively. The third section of the 'Tools' menu is a portal to our Knoter1D3D web server (8) for the prediction of knottin 3D structures based on their sequence. The KNOTTIN web site also provides access to statistics about the database content, citations and web traffic. General information about knottins and the database usage are also available for users. Finally, the 'Links' menu contains hyperlinks to knottin-related web resources—such as the Cyclotide Webpage (www.cyclotide.com), CyBase (www.cybase.org.au), Tox-Prot (www.uniprot.org/program/Toxins), MvirDB (9) and the ConoServer (10,11)—and other protein databases and web servers.

PRACTICAL USE

Over the past decade, KNOTTIN has been cited by multiple articles and reviews. In most of the cases, our database is cited either as a source of information on knottins, or when introducing knottin proteins and their structural characteristics, various functions, or their presence in a wide range of species. Numerous citations of KNOTTIN are also related the families of knottins defined in the database, to which authors refer when identifying or quantifying knottins of interest. This shows that our database, throughout the years, has successfully served as a useful overview on the field of inhibitor cystine knots. In addition to the data stored in KNOTTIN, the computational tools available on the website have also been cited, in particular Knoter1D and Knoter3D (for example in (12–14)). Our database has also been cited by three patents, including one describing cystine knot peptides engineered for anti-thrombotic therapies (15) and which suggests using the KNOTTIN's conformation search to determine folding patterns.

NEW FEATURES

Since its launch, the KNOTTIN database has seen its size grow (from 385 sequences in 2004, to 3320 in 2017, and from 85 to 214 natives structures), as the number of available sequences in UniProt and native structures in the PDB increased. However, the amount of experimentally determined structures of knottins remains relatively limited compared to the number of sequences, due to the difficulties related to the purification and crystallography of these proteins. This lack of available structures is a critical concern regarding the knottin-based drug design, which mainly lies on the study of 3D structures. To overcome this issue, one of the main features of this upgraded version of KNOTTIN is the systematic and automatic modeling and inclusion in the database of theoretical models produced with our aforementioned Knoter1D3D tool. This addition of predicted structures for every knottin sequence has increased by one order of magnitude the size of our database in terms of

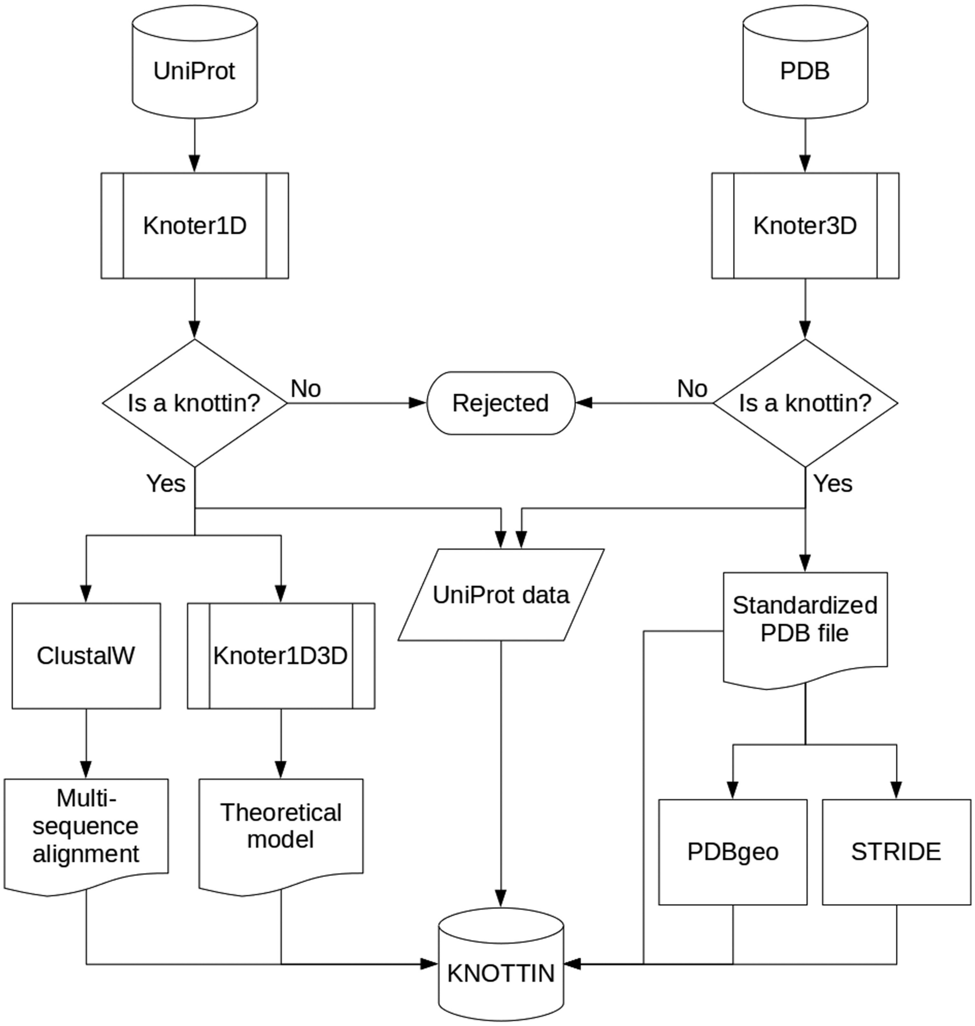
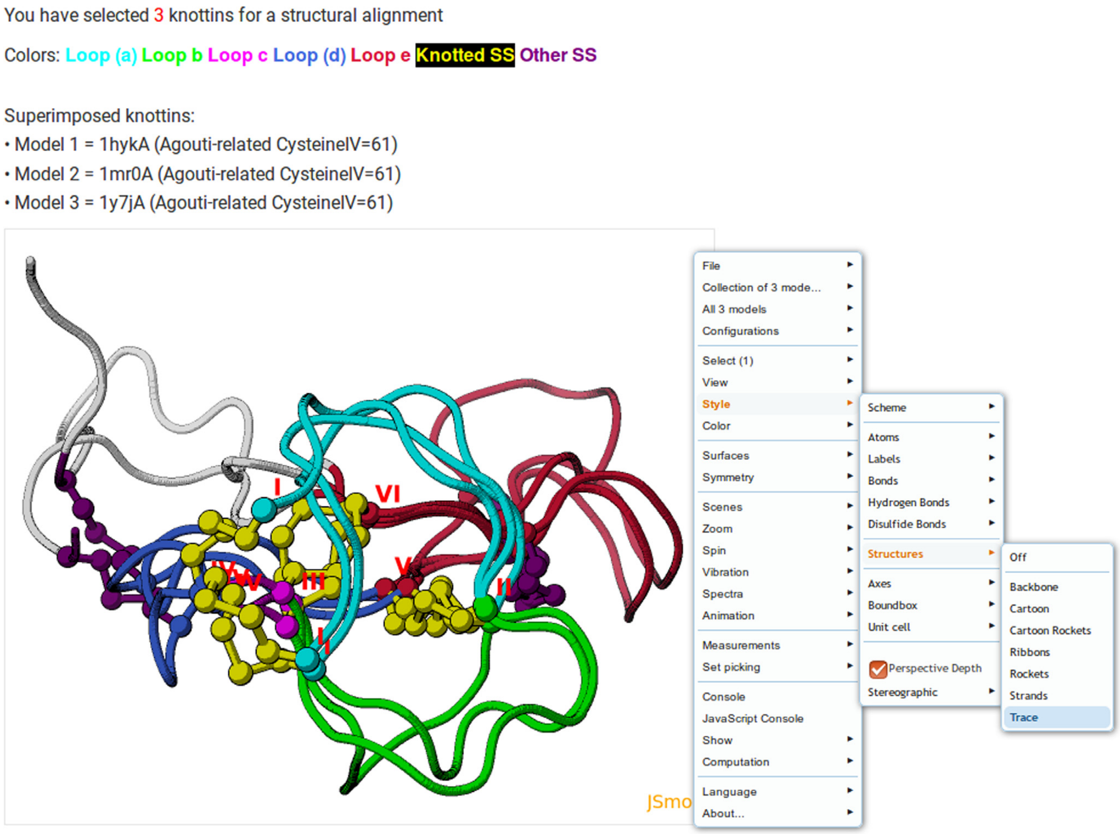


Figure 1. Flowchart describing how the KNOTTIN database is generated. The Knoter1D and Knoter3D processes have been defined in the previous release of KNOTTIN (7); the Knoter1D3D process is also described in our previous work (8). The UniProt data are automatically extracted, by using a Perl script, from the corresponding UniProt web pages.

structures, with currently contains >3000 theoretical models. In details, these theoretical models constitute a valuable source of structural data, especially for knottin families for which there is no experimental structure available, such as the bacterial knottins.

The prediction of knottin structures has been integrated as a step in the pipeline for generating the data of KNOTTIN (Figure 1). When a protein sequence from UniProt is identified as a knottin by Knoter1D, it is used as an input for Knoter1D3D to produce theoretical models, which will be added to the database—along with related UniProt data (such as, sequence, descriptor, species, tissue, authors, PMID, keywords) and a multi-sequence alignment with other knottins of the same family computed with ClustalW (16). The rest of the KNOTTIN pipeline concerns the detection and the subsequent addition of native knottin structures to the database. Thus, when a knottin structure is identified in the PDB by Knoter3D, the coordinate file is 'standardized': each residue is renumbered so that the knotted cysteines (except Cys IV) correspond to the positions 20, 40, 60, 80 and 100; the coordinate file is also reoriented by superimposition of the knotted cysteines onto those of the structure of the squash seed trypsin inhibitor (PDB code: 2btcI). These coordinate files are added to the database, along with data regarding structural properties (such as torsion angles, secondary structures, solvent accessibility) computed with STRIDE (17) and PDBgeo (described in (18)).

It should be noted that predicted structures of knottins can also be found in other databases of protein theoretical models (such as SWISS-MODEL (19) and ModBase (20)), but these data are do not match the models from KNOTTIN, neither quantitatively nor qualitatively. Indeed, according to the Protein Model Portal (www.proteinmodelportal.org), which contains models from SWISS-MODEL and ModBase, there are only 468 predicted structures of knottins in these two databases. The much greater number of models in KNOTTIN is explained by the fact that our Knoter1D3D comparative modeling procedure can accurately predict knottin structures when the template-to-sequence identity is as low as 10% (8), which is rather common among knottins. Moreover, our use of



You have selected 3 knottins for a structural alignment

Colors: Loop (a) Loop b Loop c Loop (d) Loop e Knotted SS Other SS

Superimposed knottins:

- Model 1 = 1hykA (Agouti-related CysteineIV=61)
- Model 2 = 1mr0A (Agouti-related CysteineIV=61)
- Model 3 = 1y7jA (Agouti-related CysteineIV=61)

Figure 2. Visualization of the structural superimposition of three native structures of knottins belonging to the 'Agouti-related' family (PDB codes: 1hykA, 1mr0A and 1y7jA). A right click on the JSmol viewer allows users to modify the representations of structures, or to perform other actions.

a modeling method optimized for knottins necessarily improves the quality of the theoretical models, compared to those from other databases generated with regular comparative modeling procedures, which are not adapted to the particular structural features of knottins.

This new version of the KNOTTIN database also comes with several technical improvements. The whole web interface has been entirely redesigned with the aim of being more user-friendly and compatible with modern devices. Notably, natives and predicted structures can now be visualized on the website thanks to the JavaScript-based molecular viewer JSmol (21), which therefore does not require users to have Java installed. By default, structures are displayed as a 'cartoon' representations, with each knottins loops colored differently. The numbering of the cysteines I to VI is displayed, and the disulfide bonds ('SS') are represented as 'balls and sticks' and colored differently, depending on whether they are knotted or not. The regular secondary structure elements are hidden by default, but users have the possibility to change our graphical presets by right-clicking the viewer (Figure 2)—as they would do with a local molecular viewer. This interactivity of JSmol also allows saving the session with the current parameters, as well as exporting the molecule as an image.

Finally, some other improvements have been brought to the KNOTTIN website, such as a page in the 'Statistics' menu dedicated to the articles that cite KNOTTIN, and an updated version of the multi-sequence alignments viewer Mview (22). The possibility of downloading the whole database content (sequences, native and predicted structures, and multi-sequence alignments) as a compressed archive is also a new feature. It has been implemented with the aim of being useful for researchers wanting to carry out statistical studies about knottins. These data can also be used as training or benchmark datasets in the development of new computational methods dedicated to knottin proteins. Under the same 'Data' menu, users can now contribute to the maintenance and update of the database, by proposing either a new protein sequence (optionally with additional information or coordinate file) or a published article about knottins. This new functionality is achieved through web forms that users can fill out; the input is then manually verified, before being integrated to the database.

CONCLUSIONS AND PERSPECTIVES

This new version of the KNOTTIN website is distinct from the former by the updated content of its database, as well as by its new interface and the inclusion of new data types. The KNOTTIN database now contains more than 3000 sequences of knottins, and has greatly extended its reach with the addition of predicted structures for all of these sequences. To cope with the daily increase of the number of sequence in the UniProt database, future efforts will be put in the full automation of the update pipeline. Regarding the latest data, it is interesting to observed that, while sequences have been found in animals, plants, fungi, bacteria and viruses, knottins are still absent from archaea, which converges with previous findings (13). Therefore, particular attention will be paid to new data about these organisms, which represent one of the three domains of life. Finally, KNOTTIN is also a web server providing a user-friendly access to our knottin-specific tools. Following this direction, the platform will integrate the future computational methods we will develop for the analysis of knottin proteins.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their helpful comments. We thank Marie-Paule Lefranc for authorizing the use of the expression 'Collier de Perles' which originally refers to standardized 2D representations in IMGT, the international ImMunoGenetics information system® (http://www.imgt.org).

FUNDING

Funding for open access charge: INSERM.
Conflict of interest statement. None declared.

REFERENCES

1. Ackerman,S.E., Currier,N.V., Bergen,J.M. and Cochran,J.R. (2014) Cystine-knot peptides: emerging tools for cancer imaging and therapy. *Expert Rev. Proteomics*, **11**, 561–572.
2. Kintzing,J.R. and Cochran,J.R. (2016) Engineered knottin peptides as diagnostics, therapeutics, and drug delivery vehicles. *Curr. Opin. Chem. Biol.*, **34**, 143–150.
3. Lembo,A.J., Schneier,H.A., Shiff,S.J., Kurtz,C.B., MacDougall,J.E., Jia,X.D., Shao,J.Z., Lavins,B.J., Currie,M.G., Fitch,D.A. *et al.* (2011) Two randomized trials of linaclotide for chronic constipation. *N. Engl. J. Med.*, **365**, 527–536.
4. Gelly,J.-C., Gracy,J., Kaas,Q., Le-Nguyen,D., Heitz,A. and Chiche,L. (2004) The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucleic Acids Res.*, **32**, D156–D159.
5. Mulvenna,J.P., Wang,C. and Craik,D.J. (2006) CyBase: a database of cyclic protein sequence and structure. *Nucleic Acids Res.*, **34**, D192–D194.
6. Wang,C.K.L., Kaas,Q., Chiche,L. and Craik,D.J. (2008) CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res.*, **36**, D206–D210.
7. Gracy,J., Le-Nguyen,D., Gelly,J.-C., Kaas,Q., Heitz,A. and Chiche,L. (2007) KNOTTIN: the knottin or inhibitor cystine knot scaffold in 2007. *Nucleic Acids Res.*, **36**, D314–D319.
8. Gracy,J. and Chiche,L. (2010) Optimizing structural modeling for a specific protein scaffold: knottins or inhibitor cystine knots. *BMC Bioinformatics*, **11**, 535.
9. Zhou,C.E., Smith,J., Lam,M., Zemla,A., Dyer,M.D. and Slezak,T. (2007) MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.*, **35**, D391–D394.
10. Kaas,Q., Westermann,J.-C., Halai,R., Wang,C.K.L. and Craik,D.J. (2008) ConoServer, a database for conopeptide sequences and structures. *Bioinformatics*, **24**, 445–446.
11. Kaas,Q., Yu,R., Jin,A.-H., Dutertre,S. and Craik,D.J. (2012) ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.*, **40**, D325–D330.
12. Haney,R.A., Ayoub,N.A., Clarke,T.H., Hayashi,C.Y. and Garb,J.E. (2014) Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. *BMC Genomics*, **15**, 366.
13. Islam,S.A., Sajed,T., Kearney,C.M. and Baker,E.J. (2015) PredSTP: a highly accurate SVM based model to predict sequential cystine stabilized peptides. *BMC Bioinformatics*, **16**, 210.
14. Rong,M., Liu,J., Zhang,M., Wang,G., Zhao,G., Wang,G., Zhang,Y., Hu,K. and Lai,R. (2016) A sodium channel inhibitor ISTX-I with a novel structure provides a new hint at the evolutionary link between two toxin folds. *Sci. Rep.*, **6**, 29691.
15. Cochran,J.R., Silverman,A.P. and Kariolis,M.S. (2014) Cystine knot peptides binding to alpha IIb beta 3 integrins and methods of use. U.S. Patent No. 8, 778, 888. Washington, DC: U.S. Patent and Trademark Office.
16. Thompson,J.D., Gibson,T.J. and Higgins,D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics*. doi:10.1002/0471250953.bi0203s00.
17. Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.
18. Gracy,J. and Chiche,L. (2005) PAT: a protein analysis toolkit for integrated biocomputing on the web. *Nucleic Acids Res.*, **33**, W65–W71.
19. Biasini,M., Bienert,S., Waterhouse,A., Arnold,K., Studer,G., Schmidt,T., Kiefer,F., Cassarino,T.G., Bertoni,M., Bordoli,L. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–W258.
20. Pieper,U., Webb,B.M., Dong,G.Q., Schneidman-Duhovny,D., Fan,H., Kim,S.J., Khuri,N., Spill,Y.G., Weinkam,P., Hammel,M. *et al.* (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **42**, D336–D346.
21. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Isr. J. Chem.*, **53**, 207–216.
22. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.