doi: 10.1093/nar/gkw297

SL2: an interactive webtool for modeling of missing segments in proteins

Jochen Ismer[1], Alexander S. Rose[1], Johanna K. S. Tiemann[1], Andrean Goede[2], Robert Preissner[2] and Peter W. Hildebrand[1,*]

[1]Institute of Medical Physics and Biophysics, University Medicine, Berlin, 10117 Berlin, Germany and [2]Institute of Physiology & Experimental Clinical Research Center, University Medicine, Berlin, 13125, Germany

Received January 29, 2016; Revised March 30, 2016; Accepted April 11, 2016

ABSTRACT

SuperLooper2 (SL2) (http://proteinformatics.charite.de/sl2) is the updated version of our previous webserver SuperLooper, a fragment based tool for the prediction and interactive placement of loop structures into globular and helical membrane proteins. In comparison to our previous version, SL2 benefits from both a considerably enlarged database of fragments derived from high-resolution 3D protein structures of globular and helical membrane proteins, and the integration of a new protein viewer. The database, now with double the content, significantly improved the coverage of fragment conformations and prediction quality. The employment of the NGL viewer for visualization of the protein under investigation and interactive selection of appropriate loops makes SL2 independent of third-party plug-ins and additional installations.

INTRODUCTION

Structural biology is an established but still emerging research field of life sciences, as reflected by the exponential rise of atomic models deposited in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (1). However, in more than one half of all entries deposited in the RSCB PDB segments are missing (2). These missing segments are often located in flexible and functionally important regions of proteins such as loops or turns, not resolved by X-ray crystallography or single particle cryo-electron microscopy. These regions have to be modeled to obtain a more complete structural model for further analysis of the structure, e.g. for molecular dynamics simulations (3).

Loop regions are one of the most demanding regions in homology modeling workflows. A prominent example are G protein coupled receptors (GPCRs), which constitute the largest protein family in the human genome. The number of available templates for modeling of GPCRs has increased dramatically in the last decade facilitating the generation of homology models for structure-based drug design. The common topology of the transmembrane-spanning regions, even of distantly related GPCRs, allows homology modeling of these regions and docking of small rigid orthosteric ligands with close to experimental accuracy. However, predictions of long or flexible loops remain unsolved problems, as evaluated recently by the community-wide GPCR Dock assessment (4). As the sequence similarity within loop regions is generally much lower than within other parts of proteins, specialized methods are required for modeling.

Loop modeling approaches can be divided into *ab initio* (5–8), fragment-based, (9–12) or a mixture of both methods (13,14). *Ab initio* based methods utilize molecular mechanics force fields to determine possible loop conformations. These methods are generally CPU-intensive but capable of predicting currently unknown loop conformations. Fragment-based methods on the other hand are less CPU-intensive and thus faster, but depend on known structures and precalculated fragment databases to find loop conformations. It remains unclear which method provides the better predictions. Some studies find that both methods perform on a similar level (9,12), while others describe advantages to either *ab initio* (15) or fragment-based (16) methods. As fragment-based methods generally provide results much faster, they are well suited for web-based tools such as SuperLooper (17), allowing instant visualization and control of the results.

The quality of fragment-based loop predictions using depends on the completeness of the fragment database. Independent studies have shown that the conformational space for short loops up to 12–14 residues is covered by structural fragments derived from the RCSB PDB (18,19). Enlargement of fragment databases may thus particularly enhance prediction of longer loops. Depending on the method used, also the prediction of shorter loops might benefit from a larger pool of available templates, e.g. when the exact fit of the stem atoms of the template loop to the gap is an evalua-

[*]To whom correspondence should be addressed. Tel: +49 30450524190; Fax: +49 30450524952; Email: peter.hildebrand@charite.de
Present address: Alexander S. Rose, San Diego Supercomputer Center, University of California, San Diego, CA 92093-0743, USA.

tion criterion. The database of globular and membrane proteins has more than doubled since our previous publication (17). In order to benefit from this enlargement of available structures we updated our fragment database.

Fragment-based tools such as SuperLooper depend on databases too large to distribute as stand-alone programs (∼80 GB in the case of SL2). The rapid delivery of a large number of possible loop conformation makes web-based tools a perfect candidate. The database remains on a server and the user is able to choose a suitable loop from listed results using a web-based molecule viewer. Here, we use NGL (20) for protein and fragment visualization, which adopts capabilities of modern web browsers, such as WebGL for molecular graphics. NGL allows interactive display of even large molecular complexes and is unaffected by the retirement of third-party plug-ins such as Flash or Java-Applets. This viewer offers comprehensive molecular visualization through a graphical user interface so that life scientists can easily access and utilize available structural data without any further installations (20).

Thus, SL2 benefits from the significantly enlarged database of fragments and new fast molecule viewer. Due to the improved coverage of the conformational loop space, the quality of prediction, measured by the backbone root mean square deviation (RMSD), has improved by 20% on average compared to our previous version (17). The new version of our fragment-based web-application for loop modeling SL2 thus has an improved performance in loop prediction as well as an up-to-date visualization.

UPDATE OF THE LIP AND LIMP DATABASE

The loop database (LIP) is composed of all possible fragments of 3–35 amino acids length extracted from the RSCB PDB entries in December 2015. Here, not only loops are considered but also fragments derived from secondary structure elements like helices and β-sheets. For each fragment, the amino acid sequence, PDB identifier, chain identifier, residue number of stem atoms and a geometrical fingerprint is stored.

Geometrical fingerprint matching is used as a criterion to estimate the sterical fit of stem atoms of N- and C-termini of each database fragment to the C- and N-terminal stem atoms of a gap in a protein structure. The geometrical fingerprints of both the stem atoms of each database fragment and the stem atoms of the gap are composed of the distance between the N- and C-terminal stem atoms and three angles defining their relative orientation (Figure 1). Compared to our previous version, we slightly altered the geometric fingerprint. Previously, we used a combination of two distances and two angles for scoring, resulting in a higher weighting of the fit of the residue where the angle was measured. In SL2, we solved this problem employing distance and three angles.

Since the first release of SuperLooper in 2008, the number of entries deposited in the RSCB PDB has more than doubled from 54 543 structures to 114 693 in 2015. A total of 901 609 231 fragments with a length of 3 to 35 residues was extracted from this enlarged pool of template structures (Figure 2A). Because more short than long overlapping fragments are extracted from a given template structure, the number of fragments decreases linearly with length. For loops with three amino acids, more than 30 million fragments are stored in the database, for 35 amino acids 24 million fragments are available. To benefit from the continuous growth of the RCSB PDB an update protocol was imple-

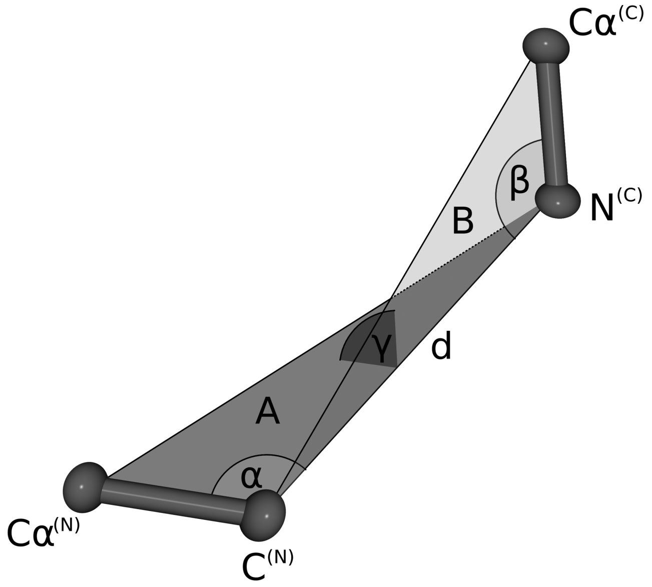


Figure 1. Schematic representation of the geometrical fingerprint: The geometrical fingerprint is characterized by the distance d between the N-terminal C- and the C-terminal N atom and the following three angles: α defined by the line between C$\alpha^{(N)}$, $C^{(N)}$ and d, β is spanned by the line between $N^{(C)}$, C$\alpha^{(C)}$ and d, γ is the angle between the two planes A (defined by C$\alpha^{(N)}$, $C^{(N)}$ and $N^{(C)}$) and B (C$\alpha^{(C)}$, $C^{(N)}$ and $N^{(C)}$).

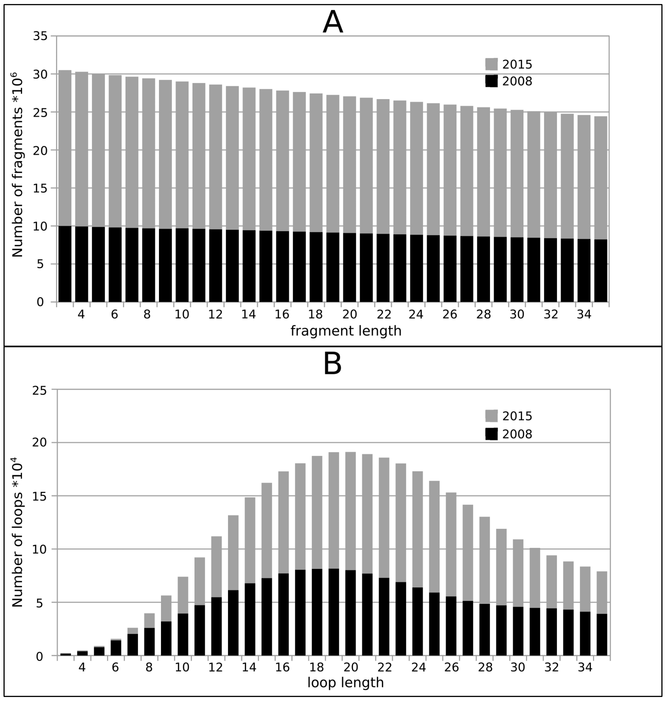


Figure 2. Length dependency of the number of fragments stored in our previous (black) and present fragment (gray) database; (**A**) loops in proteins (LIP), and (**B**) loops in membrane proteins (LIMP).

mented that automatically adds novel fragments to the LIP or LIMP database every three months.

Due to (partial) embedding into the lipid bilayer, loops of membrane proteins have a more hydrophobic amino acid composition compared to loops of globular proteins (21). Tools developed for the prediction of loops connecting transmembrane helices were indeed found to enhance prediction of GPCR loops (22). In SL2, such loops can be selected from LIMP, which is a collection of fragments extracted from loops of all helical transmembrane proteins. Loops were defined as parts without regular fold, thus also containing kinks, bulges or re-entrant loops (23). To allow selection of membrane protein loops taking the lipid bilayer into account, the extension of the lipid bilayer is indicated by two parallel planes (as described below).

The number of membrane protein structures deposited in the RCSB PDB rose from 805 (in 2008) to 2298 (in 2015) according to the Protein Data Bank of Transmembrane Proteins (24). As a result, the loops stored in LIMP doubled from 179 580 to 378 839. For LIMP is composed mainly of loop structures, the length distribution differs from LIP where the fragments also include helical fragments and fragments derived from β-sheets. In LIMP (Figure 2B), few loop templates are available for short loops of 3–5 amino acids in length. The number of loops stored in LIMP increases markedly to a maximum of 20 000 up to a length of 20 residues before it decreases again.

SEARCH PROCEDURE

To start the search the stem residues flanking the N- and C-terminus of a missing (or existing) loop in a protein model and the amino acid sequence have to be provided. As in our previous version, the search procedure is based on a stepwise approach which minimizes the calculation time. Fragments with appropriate sequence length, and with geometrical fingerprints of the fragment and the gap matching with an accuracy of at least 0.75 Å RMSD distance are selected. This RMSD value is subsequently used to determine the top 1000 loop candidates. These loop candidates are then rescored by the parameters 'sequence similarity between missing segment and template loop' and 'fingerprint matching of the template loop to the gap in the model.' Only one representative of fragments with identical primary structure and high tertiary structure similarity (with backbone RMSD < 0.5 Å) is kept in the results list to maximize the conformational space of fragments used for further calculations. The top 100 loop candidates are finally displayed in the results list. Suitable candidates can be selected from that list by visual inspection.

VISUALIZATION AND USER INTERFACE

For visual inspection of results, we employed the NGL viewer which works without installation of additional plug-ins (20). As a common graphical user interface for the NGL viewer (Figure 3) the search mask and the results list were implemented within JavaScript. A protein structure uploaded via the file selection dialog is instantly loaded to the NGL viewer. The stem residues of the gap in the protein model must either be typed into the according search field or can be selected by clicking them in the NGL viewer. The sequence of the missing segment must be typed or copied into the search mask. If the membrane protein-specific LIMP data base (Membrane DB) is not checked, the LIP data base will be searched. After the submission button is pressed, the search is started. Depending on the loop length, results are expected to appear after few seconds or up to half a minute in the results list.

The top hit will automatically be loaded into the gap of the protein model depicted in the viewer window. Alternative loop conformations can be selected from the results table containing the 100 best loop candidates. For each candidate, the score ranging from 0 to 0.455, the RCSB PDB entry-code and sequence of the template protein, the number of clashes, and the sequence identity between target and template are listed. If no appropriate loop is found, the user can select 'Decrease N-terminal stem' or 'Increase C-terminal stem' to add a residue to the loop and shift the stem atoms of the gap, accordingly. As an additional visual control, for helical membrane proteins, the position of the lipid bilayer can be calculated ('Calculate membrane planes'), employing the web-service TMDET (25).

There is an option to display the complete list of loop candidates at the same time as visualizing the conformation space of the loop. Loop candidates can be colored according to score, sequence identity or clashes by selecting the corresponding color scheme from the dropdown menu. The completed structure (initial model plus selected loop) can be downloaded by clicking the download button. Alternatively, the complete list of loops can be downloaded for further analysis.

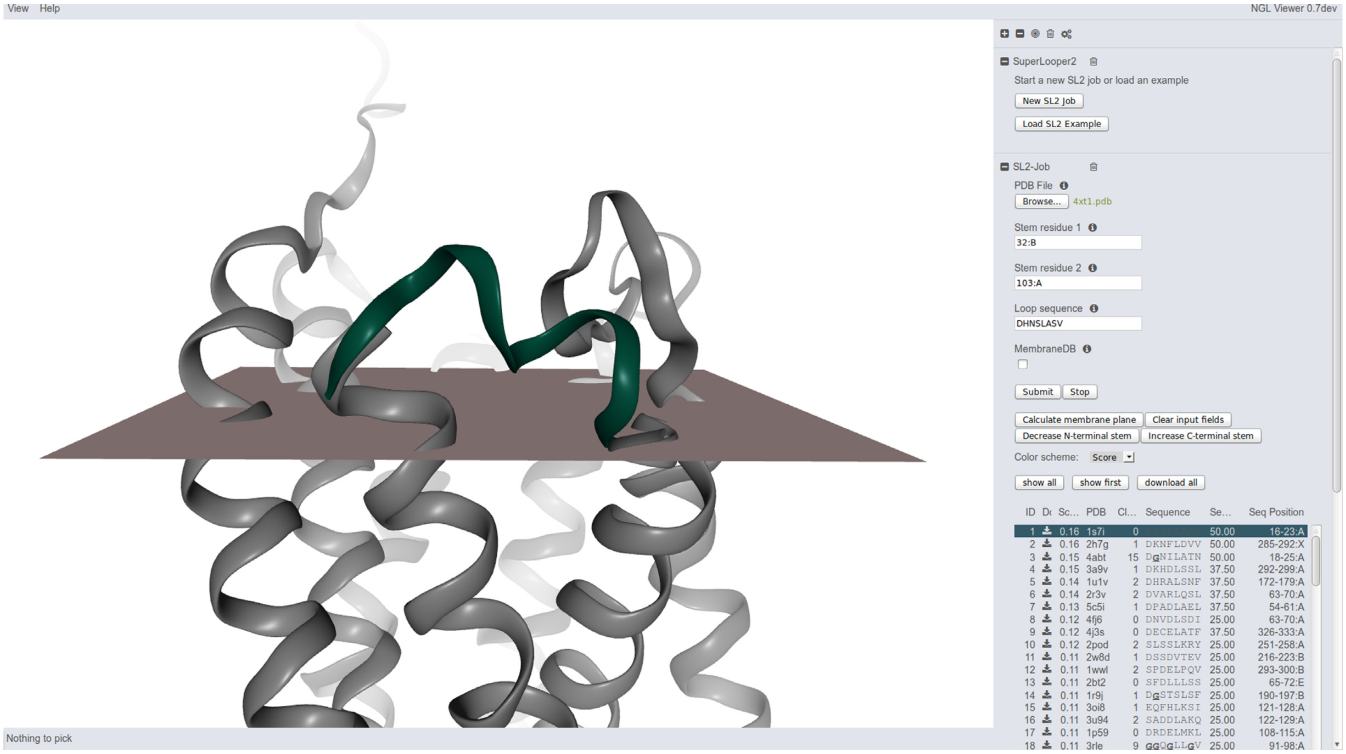
TECHNICAL ASPECTS

Visualization is carried out by the NGL viewer (20). To use the full feature set of the NGL viewer an up-to-date web browser (tested on the recent versions of Firefox, Google Chrome, Safari, IE and Edge) is recommended. The specialized graphical user interface is written in JavaScript. For job handling a simple python job server based on the Flask framework (http://flask.pocoo.org/) is used.

PERFORMANCE, LIMITATIONS AND OUTLOOK

The updated version of our fragment based web-application tool for loop modeling, SL2, benefits from an enlarged fragment database and a new user interface including an updated protein viewer. As a result of the enlarged fragment database the prediction quality has been further improved. Using the same dataset (15) and validation procedure as in our previous publication (17), an average gain in prediction quality by 20% is observed for loops of 3–16 residues length (Figure 4). A drop of the backbone RMSD between experimentally determined and modeled loops (only the top hit was considered) starts to become evident for loops with eight residues length. This implies that the coverage of possible loop conformations has been further optimized starting with this length.

Despite the gain of prediction quality, the top hit results obtained by SL2 sometimes deviate from the experimentally determined structure even for short loops. There are several



SuperLooper2

Start a new SL2 job or load an example

New SL2 Job

Load SL2 Example

SL2-Job

PDB File: Browse... 4xt1.pdb

Stem residue 1: 32:B

Stem residue 2: 103:A

Loop sequence: DHNSLASV

MembraneDB: □

Submit | Stop

Calculate membrane plane | Clear input fields

Decrease N-terminal stem | Increase C-terminal stem

Color scheme: Score

show all | show first | download all

ID	Dc	Sc...	PDB	Cl...	Sequence	Se...	Seq Position
1	⬇	0.16	1s7i	0	DEAKLAAV	50.00	16-23:A
2	⬇	0.16	2h7g	1	DKNFLDVV	50.00	285-292:X
3	⬇	0.15	4abt	15	DGNILATN	50.00	18-25:A
4	⬇	0.15	3a9v	1	DKHDLSSL	37.50	292-299:A
5	⬇	0.14	1u1v	2	DHRALSNF	37.50	172-179:A
6	⬇	0.14	2r3v	2	DVARLQSL	37.50	63-70:A
7	⬇	0.13	5c5i	1	DPADLAEL	37.50	54-61:A
8	⬇	0.12	4fj6	0	DNVDLSDI	25.00	63-70:A
9	⬇	0.12	4j3s	0	DECELATF	37.50	326-333:A
10	⬇	0.12	2pod	2	SLSSLKRY	25.00	251-258:A
11	⬇	0.11	2w8d	1	DSSDVTEV	25.00	216-223:B
12	⬇	0.11	1wwl	2	SPDELPQV	25.00	293-300:B
13	⬇	0.11	2bt2	0	SFDLLLSS	25.00	65-72:E
14	⬇	0.11	1r9j	1	DGSTSLSF	25.00	190-197:B
15	⬇	0.11	3oi8	1	EQFHLKSI	25.00	121-128:A
16	⬇	0.11	3u94	2	SADDLAKQ	25.00	122-129:A
17	⬇	0.11	1p59	0	DRDELMKL	25.00	108-115:A
18	⬇	0.11	3rle	9	GGQGLLGV	25.00	91-98:A

Figure 3. Screenshot of the SL2 results page (NGL viewer). Structure of the human cytomegalovirus GPCR US28 (PDB-ID: 4xt1) in a gray cartoon representation with top ranked loop (green) and calculated membrane planes. The list of loop candidates filling the gap 94 to 103 in the GPCR structure is displayed as table on the right hand just below the search mask.

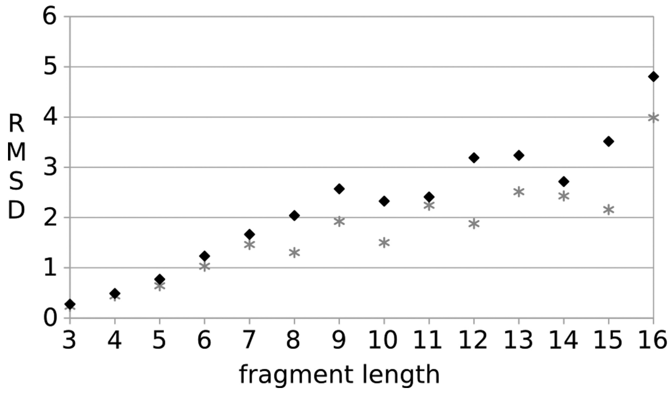


Figure 4. Comparison of benchmarks of our previous (17) (black rhombus) and updated version SL2 (gray star) using a standard loop dataset (15).

possible reasons for this. First, many loops are highly flexible or are even located in structurally disordered regions of proteins (26,27). The conformations suggested by SL2 may thus indicate alternative loop conformations not observed by protein X-ray structure crystallography (e.g. Figure S6 in (28)). Second, as scoring of the loops mainly depends on the stem residues, experimentally caused distortions of these stem atoms may prevent selection of a specific conformation (29). Prediction quality drops with loop length, mainly due to the increased conformational space. A promising strategy to enhance prediction quality of longer loops would be inclusion of additional experimental constraints such as mass spectrometry (30,31) or electron density maps from single particle cryo-electron microscopy (32).

FUNDING

Deutsche Forschungsgemeinschaft [Sfb740/B6, DFG HI 1502/1-2, BI 893/8]; 'Norddeutscher Verbund für Hoch- und Höchstleistungsrechner' (HLRN) (in part). Funding for open access charge: Sfb740 / B6.
Conflict of interest statement. None declared.

REFERENCES

1. Rose,P.W., Bi,C., Bluhm,W.F., Christie,C.H., Dimitropoulos,D., Dutta,S., Green,R.K., Goodsell,D.S., Prlic,A., Quesada,M. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
2. Brandt,B.W., Heringa,J. and Leunissen,J.A.M. (2008) SEQATOMS: a web tool for identifying missing regions in PDB in sequence context. *Nucleic Acids Res.*, **36**, 255–259.
3. Rose,A., Theune,D., Goede,A. and Hildebrand,P.W. (2014) MP:PD - A data base of internal packing densities, internal packing defects and internal waters of helical membrane proteins. *Nucleic Acids Res.*, **42**, 347–351.
4. Kufareva,I., Katritch,V., Stevens,R.C. and Abagyan,R. (2014) Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure*, **22**, 1120–1139.
5. Fiser,A., Do,R.K. and Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.
6. Spassov,V.Z., Flook,P.K. and Yan,L. (2008) LOOPER: a molecular mechanics-based algorithm for protein loop prediction. *Protein Eng. Des. Sel.*, **21**, 91–100.

7. Liang,S., Zhang,C. and Zhou,Y. (2014) LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J. Comput. Chem.*, **35**, 335–341.
8. Tang,K., Zhang,J. and Liang,J. (2014) Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLoS Comput. Biol.*, **10**, e1003539.
9. Michalsky,E., Goede,A. and Preissner,R. (2003) Loops In Proteins (LIP)–a comprehensive loop database for homology modelling. *Protein Eng.*, **16**, 979–985.
10. Messih,M.A., Lepore,R. and Tramontano,A. (2015) LoopIng: a template-based tool for predicting the structure of protein loops. *Bioinformatics*, **31**, btv438.
11. Fernandez-Fuentes,N., Zhai,J. and Fiser,A. (2006) ArchPRED: A template based loop structure prediction server. *Nucleic Acids Res.*, **34**, 173–176.
12. Peng,H.-P. and Yang,A.-S. (2007) Modeling protein loops with knowledge-based prediction of sequence-structure alignment. *Bioinformatics*, **23**, 2836–2842.
13. Deane,C.M. and Blundell,T.L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.*, **10**, 599–612.
14. van Vlijmen,H.W. and Karplus,M. (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.*, **267**, 975–1001.
15. Rossi,K.A., Weigelt,C.A., Nayeem,A. and Krystek,S.R. (2007) Loopholes and missing links in protein modeling. *Protein Sci.*, **16**, 1999–2012.
16. Choi,Y. and Deane,C.M. (2010) FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, **78**, 1431–1440.
17. Hildebrand,P.W., Goede,A., Bauer,R.a, Gruening,B., Ismer,J., Michalsky,E. and Preissner,R. (2009) SuperLooper–a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Res.*, **37**, W571–W574.
18. Du,P., Andrec,M. and Levy,R.M. (2003) Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng. Des. Sel.*, **16**, 407–414.
19. Fernandez-Fuentes,N. and Fiser,A. (2006) Saturating representation of loop conformational fragments in structure databanks. *BMC Struct. Biol.*, **6**, 15.
20. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
21. Hildebrand,P.W., Preissner,R. and Frömmel,C. (2004) Structural features of transmembrane helices. *FEBS Lett.*, **559**, 145–151.
22. Goldfeld,D.A., Zhu,K., Beuming,T. and Friesner,R.A. (2011) Successful prediction of the intra- and extracellular loops of four G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 8275–8280.
23. Kauko,A., Illergård,K. and Elofsson,A. (2008) Coils in the membrane core are conserved and functionally important. *J. Mol. Biol.*, **380**, 170–180.
24. Kozma,D., Simon,I. and Tusnady,G.E. (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**, D524–D529.
25. Tusnády,G.E., Dosztányi,Z. and Simon,I. (2005) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.
26. Elgeti,M., Rose,A.S., Bartl,F.J., Hildebrand,P.W., Hofmann,K.-P. and Heck,M. (2013) Precision vs flexibility in GPCR signaling. *J. Am. Chem. Soc.*, **135**, 12305–12312.
27. Rose,A.S., Elgeti,M., Zachariae,U., Grubmüller,H., Hofmann,K.P., Scheerer,P. and Hildebrand,P.W. (2014) Position of transmembrane helix 6 determines receptor G protein coupling specificity. *J. Am. Chem. Soc.*, **136**, 11244–11247.
28. Scheerer,P., Heck,M., Goede,A., Park,J.H., Choe,H.-W., Ernst,O.P., Hofmann,K.P. and Hildebrand,P.W. (2009) Structural and kinetic modeling of an activating helix switch in the rhodopsin-transducin interface. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 10660–10665.
29. Lessel,U. and Schomburg,D. (1999) Importance of anchor group positioning in protein loop prediction. *Proteins Struct. Funct. Genet.*, **37**, 56–64.
30. Rappsilber,J. (2011) The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.*, **173**, 530–540.
31. Schneidman-Duhovny,D., Kim,S. and Sali,A. (2012) Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct. Biol.*, **12**, 17.
32. Wang,R.Y.-R., Kudryashev,M., Li,X., Egelman,E.H., Basler,M., Cheng,Y., Baker,D. and DiMaio,F. (2015) *De novo* protein structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods*, **12**, 335–338.