Published online 8 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W571–W574***
doi:10.1093/nar/gkp338

SuperLooper—a prediction server for the modeling of loops in globular and membrane proteins

Peter W. Hildebrand[1,*], Andrean Goede[2], Raphael A. Bauer[3], Bjoern Gruening[3], Jochen Ismer[1], Elke Michalsky[3] and Robert Preissner[3]

[1]Institute of Medical Physics and Biophysics, [2]Institute of Biochemistry and [3]Institute of Physiology, Charité, University of Medicine, Berlin, Germany

Received February 20, 2009; Revised April 16, 2009; Accepted April 21, 2009

ABSTRACT

SuperLooper provides the first online interface for the automatic, quick and interactive search and placement of loops in proteins (LIP). A database containing half a billion segments of water-soluble proteins with lengths up to 35 residues can be screened for candidate loops. A specified database containing 180 000 membrane loops in proteins (LIMP) can be searched, alternatively. Loop candidates are scored based on sequence criteria and the root mean square deviation (RMSD) of the stem atoms. Searching LIP, the average global RMSD of the respective top-ranked loops to the original loops is benchmarked to be <2 Å, for loops up to six residues or <3 Å for loops shorter than 10 residues. Other suitable conformations may be selected and directly visualized on the web server from a top-50 list. For user guidance, the sequence homology between the template and the original sequence, proline or glycine exchanges or close contacts between a loop candidate and the remainder of the protein are denoted. For membrane proteins, the expansions of the lipid bilayer are automatically modeled using the TMDET algorithm. This allows the user to select the optimal membrane protein loop concerning its relative orientation to the lipid bilayer. The server is online since October 2007 and can be freely accessed at URL: http://bioinformatics.charite.de/superlooper/

INTRODUCTION

Loop prediction is generally one of the most challenging tasks in protein structure determination and modeling (1–17). The preferred conformation of loops often remains unclear even when the rest of the protein is resolved at high resolution. This is due to the high flexibility of loops that is often related to their function (18). Loops are regularly involved in the recognition and binding of modulators or associated proteins. Medically highly relevant interactions, such as the coupling of receptors to G proteins are mediated by membrane protein loops (19). Therefore, the knowledge of the conformation or the conformational space of a loop is essentially important to understand the mechanisms to activate or deactivate membrane receptors and transporters, or more broadly to model protein–protein or protein–ligand interactions.

For loop modeling, two different methods, *ab initio* (1,3,5,8,15–17) and comparative modeling (6,9,14) are applied. *Ab initio* methods calculate possible loop conformations with the help of various energy functions and minimizations. These methods do not depend on large template libraries, but are generally time consuming, and are therefore less appropriate for interactive searches. Comparative modeling approaches allow quick searches, but the quality of prediction largely depends on the availability of a suitable template loop structure. Thus, the potential of comparative modeling methods grows, as the diversity of available templates enlarges (14). It is estimated that, at the moment, the conformation of any loop up to the length of 14 residues is already represented very well by protein fragments in the RCSB Protein Data Bank (PDB) (12,20). Therefore, the performance of knowledge-based methods to find the native loop conformation particularly depends on the size of the loop databank and on the scoring function.

We have developed a scoring function for knowledge-based loop predictions that performs very well compared with other methods (14). Based on this scoring function, we now setup SuperLooper, a web application that provides a very simple, quick, user-friendly and reliable way to fill in a missing loop. No extra software has to be installed and no databank has to be downloaded to get the program started. For user guidance, the candidate loops can be visualized by a JMol (http://www.jmol.org/)

*To whom correspondence should be addressed. Tel: +49 304 5025 8155; Email: peter.hildebrand@charite.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

plug-in. Moreover, the web server provides information on sequence identities or proline and glycine exchanges between the template and the target, as well as close distances between a selected loop and the remainder of the protein. Finally, the membrane planes are automatically detected and visualized using the TMDET algorithm (21). Thus, the specificities of membrane protein loops arising from the positioning at the membrane–water interface can be respected, too (22).

METHODS

To allow the searches to be performed in real time, we have improved the scoring procedure that is the most time consuming process of our method (14). The search for the appropriate loop is now performed in a three-step process, described below. This hierarchical principle causes that the most CPU intensive calculations are performed on relatively small datasets.

(1) Up to 100 000 candidates with the required loop length are preselected from the two databases LIP (loops in proteins, ~500 000 000 protein segments) and LIMP (loops in membrane proteins, ~180 000 loops). The stem atoms (two main chain atoms preceding and following the loop, respectively) of candidate loops must fit the stem atoms of the target structure with a maximum deviation of 0.75 Å for each atom pair.
(2) The best 500 candidates are chosen by a specific 'goodness value' that allows a quick estimation of the steric fit of loop candidates to a target protein, described in detail in our previous analysis (14).
(3) Finally, the loop candidates are ranked by a score that includes the sequence similarity between loop candidate and target sequence, as well as the root mean square deviation (RMSD) of the stem atoms. To assure that the 50 top listed loops cover a maximum of the plausible conformational space, candidates with identical sequences and similar backbone conformations (RMSD < 1.0 Å) are further excluded from the list. For the benchmarks described in the following, only the top-ranked loop was considered in each case.

RESULTS

Performance

Using the test dataset of the Sali lab (15), we have shown previously that the accuracy of the method underlying SuperLooper performs better than other methods in particular for longer loops (14). The performance of SuperLooper was now benchmarked applying a new test dataset that was recently published to benchmark four commercially available programs for loop sampling Prime (Schrödinger, LLC), Modeler (Accelrys Software, Inc.), ICM (Molsoft, LLC) and Sybyl (Tripos, Inc.) (7). The outcome of that study is that Prime, an *ab initio* method performs best especially with increasing loop lengths. To compare our results with this study, protein structures with the same PDB entry as in the test datasets were first of all excluded from LIP. In the next step, loop candidates coming from proteins with very similar sequences were also excluded from LIP. Similarity here means 'different versions of the same protein or slightly mutated variants'. This criterion is assessed by a sliding window technique as described previously (14). As a result, top-ranked loops show a global RMSD (main chain atoms) to the original loops of <1.3 Å for loops up to six residues or <3.0 Å for loops shorter than 10 residues.

Best results are obtained, when loops with nearly identical sequences or close homologs are available. This, however, is presently not always the case for longer loops. To compare the performance of SuperLooper with that of the above mentioned tools, the analysis was repeated for loops with 11- and 12-residues length using a sequence identity limit of 90%. As a result, the average performance of SuperLooper at loop lengths 11 and 12 (RMSD = 2.6 and 4.0, respectively) is comparable with that of Prime (RMSD = 3.7 and 3.5, respectively). At loop length 11 homologous templates with sequence identities ranging from 32% to 82% are detected by SuperLooper for 9 of 14 tested loops. The average global RMSD of the modeled to the native loops is 0.7. For the remaining five template loops (with no homologous template available) the RMSD is 5.9. At loop length 12 homologous templates with sequence identities ranging from 58% to 95% are found for 4 of 10 tested loops. The average global RMSD of the modeled to the native loops is 0.6. For the remaining six template loops, the RMSD = 6.3. Thus, SuperLooper clearly outperforms Prime at these critical loop lengths if a homologous template is available. If no homologue is found, the *ab initio* method Prime performs usually better.

In conclusion, the performance of knowledge based methods such as SuperLooper clearly depends on the size and actuality of the data base in use. SuperLooper is thus regularly updated. More detailed data on actual benchmarks of SuperLooper are available from http://bioinformatics.charite.de/superlooper/. Better results can always be obtained when not only the top ranked loop is considered. Thus, the user is encouraged to visually inspect the loops to determine, which is most reasonable. SuperLooper was, therefore, implemented with a user-friendly interface to visualize and select the proper loop structure from a list of proposed conformations.

Server implementation

SuperLooper is implemented as an easy to use web application combining an interactive query of the loop database with a 3D visualization of the results. At the query site, the stem amino acids of the uploaded PDB file have to be provided together with the destined amino acid sequence. The result site provides all information necessary for the user to select the appropriate loop from a list of candidates ranked from the LIMP and LIP data bases (Figure 1). Loop candidates can be selected from both data bases provided. Due to the extensive size, the quality of loop predictions taken from the LIP data base generally

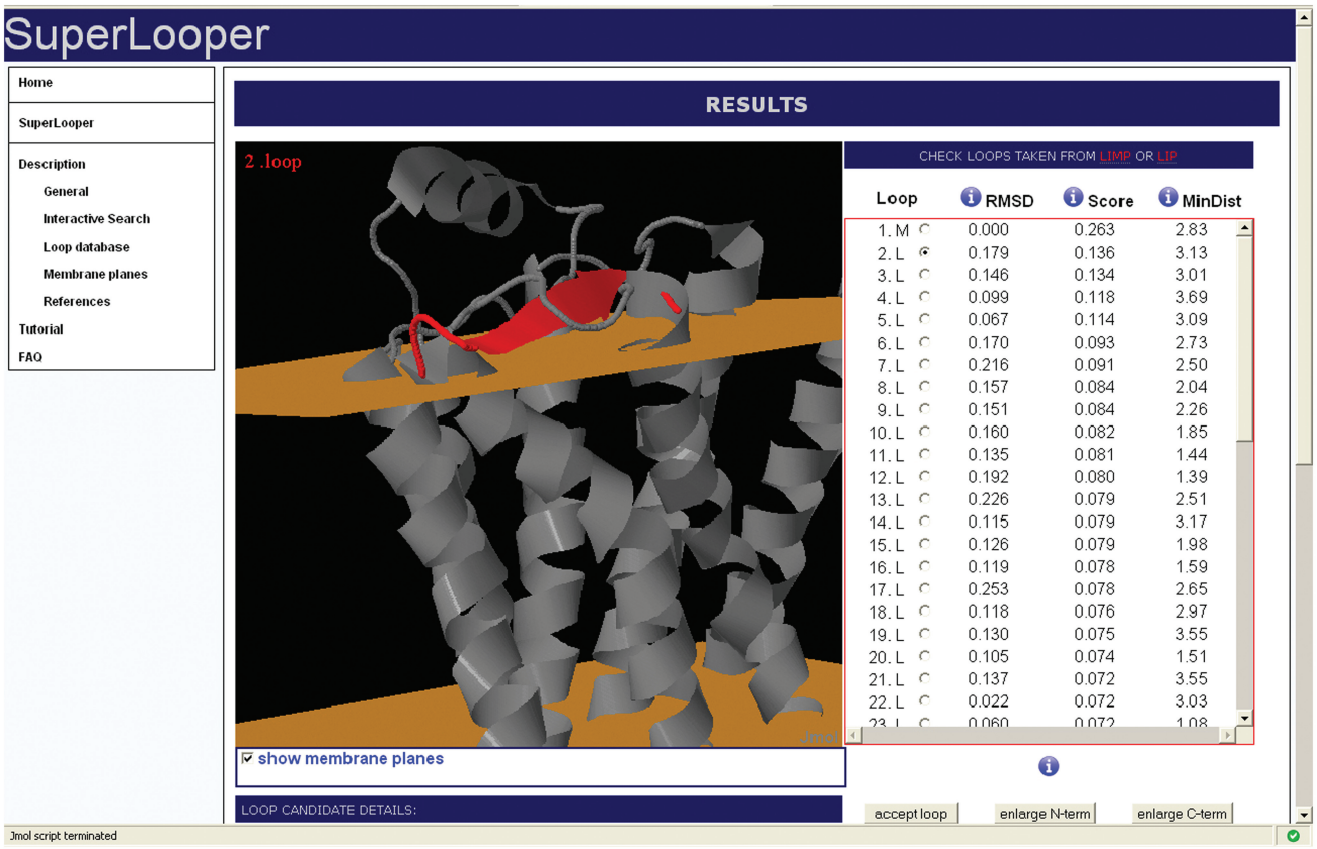

Figure 1. Alternative conformations (red) for loop 2 of the human β_2-adrenergic receptor (2rh1.pdb) can be selected from the list calculated by SuperLooper considering the predicted membrane planes (yellow).

ranges above that of predictions with the LIMP data base. Nevertheless, considering the specific amino acid composition of transmembrane helix caps and loops (22) candidates taken from the LIMP data base should always be checked first, when a membrane loop is to be modeled.

If no appropriate loop is found, the search may be expanded easily in N- or C-terminal direction up to a final loop length of 35 amino acids. To generally avoid unfavorable loop conformations and steric hindrance, the positions of proline and glycine exchanges in the selected loop are highlighted as well as distances <2.4 Å to the rest of the protein. The percentage sequence identity of a template loop is always noted to inform the user about the probability that the native loop conformation is actually matched. A membrane protein loop should be selected with respect to its relative orientation to the lipid bilayer indicated by the protein viewer. The expansions of the lipid bilayer are predicted applying the TMDET algorithm (21,23).

Technical notes

The web application uses PHP and AJAX. Membrane planes are calculated on a remote server (TMDET) connected via web service (21). The web site uses Jmol (http://jmol.sf.net) for visualization, and therefore needs a Java JRE, freely available from http://java.net. The web application uses the PDB-file format as the default input and output format, and is designed to be used with Internet Explorer 7 and Firefox 2.0–3.0. The web application is also compatible with IE 6, but tends to be unstable on some computers regarding some combinations of JRE and IE 6.

ACKNOWLEDGEMENTS

We would like to thank Dr Tusnady for kindly providing the TMDET algorithm. We thank Stefanie Neumann for helpful discussions.

FUNDING

European Union (ProFIT); Deutsche Forschungsgemeinschaft (SFB449, SFB740, DFG GRK1360). Funding for open access charge: SFB449.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Sellers,B.D., Zhu,K., Zhao,S., Friesner,R.A. and Jacobson,M.P. (2008) Toward better refinement of comparative models: predicting loops in inexact environments. *Proteins*, **72**, 959–971.

3. Soto,C.S., Fasnacht,M., Zhu,J., Forrest,L. and Honig,B. (2008) Loop modeling: sampling, filtering, and scoring. *Proteins*, **70**, 834–843.
4. Olson,M.A., Feig,M. and Brooks,C.L. 3rd. (2008) Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions. *J. Comput. Chem.*, **29**, 820–831.
5. Rapp,C.S., Strauss,T., Nederveen,A. and Fuentes,G. (2007) Prediction of protein loop geometries in solution. *Proteins*, **69**, 69–74.
6. Peng,H.P. and Yang,A.S. (2007) Modeling protein loops with knowledge-based prediction of sequence-structure alignment. *Bioinformatics*, **23**, 2836–2842.
7. Rossi,K.A., Weigelt,C.A., Nayeem,A. and Krystek,S.R. Jr. (2007) Loopholes and missing links in protein modeling. *Protein Sci.*, **16**, 1999–2012.
8. Zhu,K., Pincus,D.L., Zhao,S. and Friesner,R.A. (2006) Long loop prediction using the protein local optimization program. *Proteins*, **65**, 438–452.
9. Fernandez-Fuentes,N., Zhai,J. and Fiser,A. (2006) ArchPRED: a template based loop structure prediction server. *Nucleic Acids Res.*, **34**, W173–W176.
10. Lasso,G., Antoniw,J.F. and Mullins,J.G. (2006) A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics*, **22**, e290–e297.
11. Monnigmann,M. and Floudas,C.A. (2005) Protein loop structure prediction with flexible stem geometries. *Proteins*, **61**, 748–762.
12. Fernandez-Fuentes,N., Querol,E., Aviles,F.X., Sternberg,M.J. and Oliva,B. (2005) Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. *Proteins*, **60**, 746–757.
13. Jacobson,M.P., Pincus,D.L., Rapp,C.S., Day,T.J., Honig,B., Shaw,D.E. and Friesner,R.A. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins*, **55**, 351–367.
14. Michalsky,E., Goede,A. and Preissner,R. (2003) Loops In Proteins (LIP)—a comprehensive loop database for homology modelling. *Protein Eng.*, **16**, 979–985.
15. Fiser,A. and Sali,A. (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, **19**, 2500–2501.
16. Forrest,L.R. and Woolf,T.B. (2003) Discrimination of native loop conformations in membrane proteins: decoy library design and evaluation of effective energy scoring functions. *Proteins*, **52**, 492–509.
17. Barth,P., Schonbrun,J. and Baker,D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl Acad. Sci. USA*, **104**, 15682–15687.
18. Lawson,Z. and Wheatley,M. (2004) The third extracellular loop of G-protein-coupled receptors: more than just a linker between two important transmembrane helices. *Biochem. Soc. Trans.*, **32**, 1048–1050.
19. Scheerer,P., Park,J.H., Hildebrand,P.W., Kim,Y.J., Krauss,N., Choe,H.W., Hofmann,K.P. and Ernst,O.P. (2008) Crystal structure of opsin in its G-protein-interacting conformation. *Nature*, **455**, 497–502.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics.*, **21**, 1276–1277.
22. Hildebrand,P.W., Preissner,R. and Frömmel,C. (2005) Structural features of transmembrane helices. *FEBS Lett.*, **559**, 145–151.
23. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.