BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 18 2003, pages 2500–2501
DOI: 10.1093/bioinformatics/btg362

ModLoop: automated modeling of loops in protein structures

András Fiser[1,*] and Andrej Sali[2,*]

[1]Department of Biochemistry and Seaver Foundation Center for Bioinformatics, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461, USA and [2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, 600 16th Street, Suite N472D, University of California, San Francisco, CA 94143-2240, USA

Received on May 1, 2003; revised on June 23, 2003; accepted on July 8, 2003

ABSTRACT

Summary: ModLoop is a web server for automated modeling of loops in protein structures. The input is the atomic coordinates of the protein structure in the Protein Data Bank format, and the specification of the starting and ending residues of one or more segments to be modeled, containing no more than 20 residues in total. The output is the coordinates of the non-hydrogen atoms in the modeled segments. A user provides the input to the server via a simple web interface, and receives the output by e-mail. The server relies on the loop modeling routine in MODELLER that predicts the loop conformations by satisfaction of spatial restraints, without relying on a database of known protein structures. For a rapid response, ModLoop runs on a cluster of Linux PC computers.

Availability: The server is freely accessible to academic users at http://salilab.org/modloop

Contact: andras@fiserlab.org

The function of a protein is generally determined by shape, dynamics and physiochemical properties of its solvent exposed molecular surface. Likewise, functional differences among the members of the same protein family are usually a consequence of the structural differences on the protein surface. In a given fold family, structural variability is a result of substitutions, insertions and deletions of residues among members of the family. Such changes frequently correspond to exposed loop regions that connect elements of secondary structure in the protein fold. Thus, loops often contribute to binding sites and determine the functional specificity of a given protein framework. Consequently, the accuracy of loop modeling (Oliva *et al*., 1997; Fine *et al*., 1986; Xiang *et al*., 2002; van Vlijmen and Karplus, 1997; Rapp and Friesner, 1999; Martin and Thornton, 1996) is a major factor determining the usefulness of comparative protein structure models (Fiser *et al*., 2002; Al Lazikani *et al*., 2001) in studying interactions between the protein and its ligands. Loop modeling can also be useful in refining low- and medium-resolution structures determined by X-ray crystallography and NMR spectroscopy.

Prediction of loop conformations by optimization of an objective function was implemented in MODELLER (Fiser *et al*., 2000). The method optimizes the positions of all non-hydrogen atoms of a loop in a fixed environment. The optimization relies on a protocol consisting of the conjugate gradient minimization and molecular dynamics simulation with simulated annealing. The pseudo-energy function contains terms from a molecular mechanics force field as well as restraints based on statistical distributions derived from known protein structures. Bonds, angles, some dihedral angles and improper dihedral angles are restrained by the corresponding terms in the CHARMM-22 potential function (MacKerell *et al*., 1998). The mainchain and sidechain dihedral angles as well as non-bonded atom pairs are restrained by statistical potentials extracted from many known protein structures (Sali and Blundell, 1993; Sali and Overington, 1994; Fiser *et al*., 2000; Melo and Feytmans, 1997).

Evaluation of the method relied on 40 randomly selected loops of known structure at each length from 1 to 14 residues. The accuracy was determined by building loops both in the native and distorted loop environments because only approximate loop environments are available in real comparative modeling applications. The errors in loop predictions increase with loop length and environment distortion (Fig. 1).

The modeling protocol generates a number of independently optimized conformations, starting with random initial conformations. The final loop prediction is the optimized conformation that has the lowest pseudo-energy score. Currently, the number of independent optimizations is limited to 300. Because each individual optimization typically takes a few minutes of CPU time on a Pentium processor, the ModLoop web server runs on a cluster of Pentium nodes.

*To whom correspondence should be addressed.

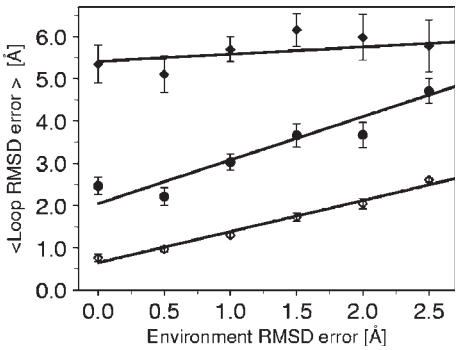

Fig. 1. The accuracy of loop modeling as a function of environment distortion (Fiser *et al*., 2000). Average prediction accuracy is shown for 40 loops of 4 (open circles), 8 (filled circles) and 12 residues (filled squares). The root-mean-square (RMS) error of the loops is calculated for the four mainchain atom types upon the optimal superposition of the whole model on the native structure. The RMS error upon superposition of only the loop atoms tends to be ∼1.5 times smaller for 8-residue loops than the errors shown (Fiser *et al*., 2000). The distortion of the environment is measured by the RMS deviation for the three residues flanking both sides of the loop. Error bars indicate the standard error of the mean.

The ModLoop server requires as input a coordinate file in the Protein Data Bank (PDB) format, as well as the starting and ending positions of the loops. The user can specify several loops as input. These loops will be optimized simultaneously, which is particularly useful if conformations of multiple interacting loops need to be predicted. Given the rapid decrease in the prediction accuracy as the loop length increases, the total number of residues in all selected loops is currently limited to 20. The resulting coordinate file in the PDB format is sent back to the user by e-mail.

The ModLoop server is a useful addition to the already available protein structure modeling servers (http://salilab.org/bioinformatics_resources.shtml). It has been utilized in several applications, including modeling of the active site loop conformations in dehydrogenases with various substrate specificities (Wu *et al*., 1999; Vernal *et al*., 2002), predicting the conformation of the linker loop in an artificial construct of a circularly permuted cyanovirin protein (Barrientos *et al*., 2001), and rationalizing the observed functional impact of various mutants of the zebrafish winged helix protein No Soul/Foxi1 (Lee *et al*., 2003).

REFERENCES

Al Lazikani,B., Jung,J., Xiang,Z. and Honig,B. (2001) Protein structure prediction. *Curr. Opin. Chem. Biol.*, **5**, 51–56.

Barrientos,L.G., Campos-Olivas,R., Louis,J.M., Fiser,A., Sali,A. and Gronenborn,A.M. (2001) 1H, 13C, 15N resonance assignments and fold verification of a circular permuted variant of the potent HIV-inactivating protein cyanovirin-N. *J. Biomol. NMR*, **19**, 289–290.

Fine,R.M., Wang,H., Shenkin,P.S., Yarmush,D.L. and Levinthal,C. (1986) Predicting antibody hypervariable loop conformations. II: minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins*, **1**, 342–362.

Fiser,A., Do,R.K. and Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.

Fiser,A., Feig,M., Brooks,C.L.,III. and Sali,A. (2002) Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.*, **35**, 413–421.

Lee,S.A., Shen,E.L., Fiser,A., Sali,A. and Guo,S. (2003) The zebrafish forkhead transcription factor Foxi1 specifies epibranchial placode-derived sensory neurons. *Development*, **130**, 2669–2679.

MacKerell,A.D.,Jr, Bashford,D., Bellott,M., Dunbrack,R.L.,Jr, Evanseck,J.D., Field,M.J., Fischer,S., Gao,J., Guo,H., Ha,S. *et al*. (1998) All-atom empirical potential for molecular modleing and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.

Martin,A.C. and Thornton,J.M. (1996) Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.*, **263**, 800–815.

Melo,F. and Feytmans,E. (1997) Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, **267**, 207–222.

Oliva,B., Bates,P.A., Querol,E., Aviles,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814–830.

Rapp,C.S. and Friesner,R.A. (1999) Prediction of loop geometries using a generalized born model of solvation effects. *Proteins*, **35**, 173–183.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sali,A. and Overington,J.P. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.*, **3**, 1582–1596.

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.

Vernal,J., Fiser,A., Sali,A., Muller,M., Jose,C.J. and Nowicki,C. (2002) Probing the specificity of a trypanosomal aromatic alpha-hydroxy acid dehydrogenase by site-directed mutagenesis. *Biochem. Biophys. Res. Commun.*, **293**, 633–639.

Wu,G., Fiser,A., ter Kuile,B., Sali,A. and Muller,M. (1999) Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl Acad. Sci. USA*, **96**, 6285–6290.

Xiang,Z., Soto,C.S. and Honig,B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl Acad. Sci. USA*, **99**, 7432–7437.