

*Structural bioinformatics***GOR V server for protein secondary structure prediction**Taner Z. Sen,<sup>1,2</sup> Robert L. Jernigan,<sup>1,2</sup> Jean Garnier<sup>3</sup> and Andrzej Kloczkowski<sup>1,\*</sup><sup>1</sup>L.H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University Ames, IA 50011, USA,<sup>2</sup>Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USAand <sup>3</sup>INRA—Unite MIG, Bat. 233, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France

Received on February 23, 2005; revised on March 21, 2005; accepted on March 22, 2005

Advance Access publication March 29, 2005

**ABSTRACT**

**Summary:** We have created the GOR V web server for protein secondary structure prediction. The GOR V algorithm combines information theory, Bayesian statistics and evolutionary information. In its fifth version, the GOR method reached (with the full jack-knife procedure) an accuracy of prediction  $Q_3$  of 73.5%. Although GOR V has been among the most successful methods, its online unavailability has been a deterrent to its popularity. Here, we remedy this situation by creating the GOR V server.

**Availability:** The GOR V server is freely accessible to public users and private institutions at <http://gor.bb.iastate.edu/>

**Contact:** [kloczkow@iastate.edu](mailto:kloczkow@iastate.edu)

**INTRODUCTION**

Structural information can provide insight into protein function, and therefore, high-accuracy prediction of protein structure from its sequence is highly desirable. The availability of structural information may expedite drug design efforts and provide a more detailed understanding of protein–protein interaction networks. Secondary structure prediction methods are also useful for motif detection in globular (Rost, 2001) and membrane proteins (Chen and Rost, 2002; Chen *et al.*, 2002), or for enhancing homology modeling (Schwede *et al.*, 2003).

The protein secondary structure prediction problem has been intensively studied by many research groups for over three decades. The first prediction methods developed by Chou and Fasman (1974), Lim (1974a,b) and Garnier *et al.* (1978) reached an accuracy of ~60%. Some of the most successful recent methods based on neural networks such as PhD (Rost, 2003) and PSIPRED (Jones, 1999) reported an accuracy of above 76%. Frishman and Argos (1997) reached an accuracy of 74.8% using PREDATOR. Some secondary structure predictions even reached an accuracy of 77% (Levin and Garnier, 1988; Petersen *et al.*, 2000). Support vector machines (Hua and Sun, 2001) and recently fragment databases (Cheng *et al.*, 2005) were also successfully used for the secondary structure prediction, among others. Although the GOR V method is around 5% less accurate (when  $Q_3$  values are compared) than the widely used neural-network based methods of PhD and PSIPRED, it may provide complementary information because it is based on different approaches, such as information theory and Bayesian statistics.

The secondary structure predictions are usually compared with DSSP (Kabsch and Sander, 1983) assignments of secondary structure

from crystallographically determined coordinates. Although DSSP defines eight different structural elements, these eight states are commonly translated into three secondary structure states:  $\alpha$ -helix,  $\beta$ -sheet and coil. This translation is usually performed in the following manner: (1)  $\alpha$ -helix in the three letter code corresponds to H ( $\alpha$ -helix), G ( $3_{10}$  helix) and I ( $\pi$ -helix) from the DSSP 8-letter code, (2) sheet corresponds to B (bridge—single residue sheet), and E (extended  $\beta$ -strand) in DSSP nomenclature and finally, (3) coil in 3-letter code corresponds to the remaining three DSSP states: T ( $\beta$ -turn), S (bend) and C (coil). In the GOR V output,  $\alpha$ -helix is represented by a letter H,  $\beta$ -sheet by E and coil by C.

**IMPLEMENTATION**

The GOR (Garnier–Osguthorpe–Robson) method uses both information theory and Bayesian statistics for predicting the secondary structure of proteins (Garnier *et al.*, 1978). Over the years, the method has been improved by including larger databases and more detailed statistics, which account not only for amino acid composition but also for amino acid pairs and triplets (Garnier and Robson, 1989; Garnier *et al.*, 1996; Gibrat *et al.*, 1987). These changes were gradually integrated into the first four versions of GOR. The fourth version of GOR algorithm, GOR IV, has been available for many years online at: <http://abs.cit.nih.gov/gor/>. However since GOR IV does not utilize evolutionary information its accuracy measured by  $Q_3$  is (similar to other single sequence-based prediction methods) ~65%.

In the most recent GOR version, GOR V (Kloczkowski *et al.*, 2002), several additional improvements were incorporated into the prediction methodology. The most crucial change in the algorithm was the inclusion of evolutionary information using PSI-BLAST (Altschul *et al.*, 1997). Multiple alignments are generated using PSI-BLAST after five iterations based on the non-redundant database (Benson *et al.*, 1999). The idea behind incorporating multiple sequence alignments into GOR is to increase the information content for improved discrimination among secondary structures. Note that only the sequence information of these multiple alignments is being used in GOR V, and not the secondary structure information of these aligned sequences. In GOR V, the prediction accuracy  $Q_3$  using full jackknifing reached 73.5%. The segment overlap (Zemla *et al.*, 1999), which is a measure of normalized secondary structure segments, was 70.8%.

Although GOR V is one of the better secondary structure methods, which provide high prediction accuracy, the public was not able to use GOR V for secondary structure predictions for their own

\*To whom correspondence should be addressed.

sequences. Now, we have taken the initiative to set up the GOR V server, available to everyone.

Since GOR V is based on completely different principles (such as information theory) than most of the other secondary structure prediction methods, we believe that its inclusion on metasevers for secondary structure prediction would be beneficial, and could improve the overall accuracy of the prediction of metasevers. In our recent work on protein binding site prediction (Sen *et al.*, 2004), we have combined several orthogonal methods, such as support vector machines, threading, conservatism of conservatism and phylogenetic trees, and developed a consensus method that has an accuracy of prediction better than each of the individual methods. This shows that consensus predictions benefit from the inclusion of predictions that are not perfect but based on fundamentally different principles.

The GOR V server is based on the database of Cuff and Barton (1999, 2000) of 513 sequentially non-redundant domains, which contains 84 107 residues. To ensure that such a set was representative of available proteins, non-redundancy was defined with stringent tests. Instead of employing a simple percentage of sequence identity between pairs of proteins, a range of sequence alignments and subsequent clusterings were performed. After randomization, only the aligned sequences with a Z-score <5 were considered dissimilar to hinder homology among sequences. Details of this data set can be found in Cuff and Barton (1999, 2000).

The GORV server works in the following manner. When the input sequence is provided by the user, the GORV server that was trained on 513 proteins calculates the helix, sheet and coil probabilities at each residue position and makes an initial prediction based on the structural states having highest probabilities. After this initial prediction, heuristic rules are applied. These rules include converting helices shorter than five residues and sheets shorter than two residues to coil. For a more detailed discussion of these heuristic rules, please refer to the original GOR V paper (Kloczkowski *et al.*, 2002). As output, the user receives the secondary structure prediction for the input sequence and the probabilities for each secondary state element at each position. The prediction results are shown in the web browser, which should stay open during the run, and are also sent to the e-mail address previously provided by the user. Any run-time error message will appear in the web browser, and if any problem arises, the user can contact the system administrator via the e-mail provided on the web page.

For a sequence of 100 amino acids, the secondary structure prediction takes ~1 min. However, the most time consuming steps are PSI-BLAST alignments, that in some cases—for many hits and slowly converging iterations may take considerable time. We have also tested the GOR V server for sequences up to 300 amino acids successfully. Currently, the server is a Linux box with RedHat Enterprise 3.0 system installed with 4.5GB RAM and 140GB memory. The program code is compiled using the Intel Fortran Compiler 8.0.034, and the web interface is established with a CGI script written using HTML and PERL. In the future, we will enhance the GOR V server both in hardware or software for improved performance, especially if user demand necessitates it.

## ACKNOWLEDGEMENTS

T.Z.S., R.L.J. and A.K. were supported by NIH grants R01GM072014 and R21GM066387.

## REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.
- Benson,D.A. *et al.* (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.
- Chen,C.P. and Rost,B. (2002) State-of-the-art in membrane protein prediction. *Appl. Bioinformatics*, **1**, 21–35.
- Chen,C.P. *et al.* (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.
- Cheng,H. *et al.* (2005) Prediction of protein secondary structure by mining fragments database. *Polymer* (in press).
- Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry (Mosc.)*, **13**, 222–245.
- Cuff,J.A. and Barton,G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508–519.
- Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
- Frishman,D. and Argos,P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329–335.
- Garnier,J. and Robson,B. (1989) Prediction of protein structure and the principles of protein conformation. In: Fasman,G.D. (ed.), Plenum Press, New York, pp. 417–465.
- Garnier,J. *et al.* (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.
- Garnier,J. *et al.* (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.*, **266**, 540–553.
- Gibrat,J.F. *et al.* (1987) Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. *J. Mol. Biol.*, **198**, 425–443.
- Hua,S. and Sun,Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.
- Jones,T.D. (1999) Protein secondary structure prediction based on position specific matrices. *J. Mol. Biol.*, **292**, 195–202.
- Kabsch,W. and Sander,C. (1983) A dictionary of secondary structure. *Biopolymers*, **22**, 2577–2637.
- Kloczkowski,A. *et al.* (2002) Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins*, **49**, 154–166.
- Levin,J.M. and Garnier,J. (1988) Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Acta*, **955**, 283–295.
- Lim,V. (1974a) Structural principles of the globular organization of protein chains: a stereochemical theory of globular protein secondary structure. *J. Mol. Biol.*, **88**, 857–872.
- Lim,V. (1974b) Algorithm for prediction of  $\alpha$ -helical and  $\beta$ -structural regions in globular proteins. *J. Mol. Biol.*, **88**, 873–894.
- Petersen,T.N. *et al.* (2000) Prediction of protein secondary structure at 80% accuracy. *Proteins*, **41**, 17–20.
- Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
- Rost,B. (2003) Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem. Anal.*, **44**, 559–587.
- Schwede,T. *et al.* (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
- Sen,T.Z. *et al.* (2004) Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. *BMC Bioinformatics*, **5**, 205.
- Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins (Suppl 3)*, 22–29.