Published in final edited form as:

Bioinformatics. 2007 October 1; 23(19): 2628–2630. doi:10.1093/bioinformatics/btm379.

Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: Combining GOR V and Fragment Database Mining (FDM)

Haitao Cheng[1,2], **Taner Z. Sen**[1,3], **Robert L. Jernigan**[1,3], and **Andrzej Kloczkowski**[1,3]

[1] Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA

[2] Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA

[3] L.H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA

Summary

One of the challenges in protein secondary structure prediction is to overcome the cross-validated 80% prediction accuracy barrier. Here, we propose a novel approach to surpass this barrier. Instead of using a single algorithm that relies on a limited data set for training, we combine two complementary methods having different strengths: Fragment Database Mining (FDM) and GOR V. FDM harnesses the availability of the known protein structures in the Protein Data Bank and provides highly accurate secondary structure predictions when sequentially similar structural fragments are identified. In contrast, the GOR V algorithm is based on information theory, Bayesian statistics, and PSI-BLAST multiple sequence alignments to predict the secondary structure of residues inside a sliding window along a protein chain. A combination of these two different methods benefits from the large number of structures in the PDB and significantly improves the secondary structure prediction accuracy, resulting in Q3 ranging from 67.5 to 93.2%, depending on the availability of highly similar fragments in the Protein Data Bank.

1 INTRODUCTION

Accurate prediction of protein secondary structure is essential for many bioinformatics applications. It allows structural alignments based on secondary structure topology (Krissinel and Henrick, 2004), provides certain structural understanding of proteins when homologous tertiary structures are not available in the PDB (Wray and Fisher, 2007) [especially for membrane proteins (Kashlan et al., 2006)], and leads to more accurate tertiary structure predictions (Jayaram et al., 2006; Meiler and Baker, 2003). For tertiary structure predictions, we encounter two main limitations: (1) for certain proteins, tertiary structure prediction methods cannot provide reliable 3D models, (2) when a model can be built, the model resolution can vary widely from 2–3 Å to tens of Angstroms (Moult, 2006). In contrast, secondary structure prediction methods always provide a secondary structure model, though with a varying accuracy. Improved secondary structure prediction can also lead to enhanced structural

Correspondence to: Taner Z. Sen.

Associate Editor: Burkhard Rost

Present address: USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA 50011-3260 and Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011-3260, USA.

Conflict of Interest: none declared.

searches and comparisons, as well as to the identification of distant homologies. Many tertiary structure prediction methods, such as fold recognition or de novo (ab initio) modeling require the derivation of sufficient structural constraints that include as precise as possible a description of secondary structure. Of course information about the secondary structure is most useful for template-free de novo modeling since it leads to significant reduction of the conformational space in Monte Carlo simulations. The least important is for comparative modeling, especially if the Protein Data Bank contains homologous proteins with sufficiently high sequence identity.

The most popular parameter measuring the accuracy of prediction is Q3, which counts the percentage of residues correctly assigned to three secondary structure categories: alpha-helices, beta-strands, and coil. However, secondary structure prediction methods based on a single method cannot surpass a virtual barrier of around 80% Q3 accuracy (Kihara, 2005). In an effort to overcome this barrier, we choose here to combine two complementary approaches developed recently by us (Sen et al., 2006). One approach, Fragment Database Mining (FDM) (Cheng et al., 2005) mines PDB structures and searches for sequence-based similarities to collect structural fragments for a prediction. Another method, GOR V (Kloczkowski et al., 2002; Sen et al., 2005), is based on statistical preferences for a given residue in the center of the sliding window along a protein chain to assume a specific secondary structure.

The FDM method (Cheng et al., 2005) is our recently developed secondary structure prediction method inspired by the success of Rosetta software (Simons et al., 1997), which uses structural fragments to build 3D models. For a given query sequence, FDM BLASTs (Altschul et al., 1990) the sequence against the PDB (Berman et al., 2000) and obtains a set of structural fragments that are sequentially similar to the query. Then, the FDM program assigns weights to each fragment based on the identity score of the alignment and calculates which structural assignment is most probable for a given site. The performance of FDM is excellent; however, only when highly similar fragments are available.

The GOR V method (Kloczkowski et al., 2002; Sen et al., 2005), on the other hand, is the latest version of a successful and pioneering secondary structure prediction method based on information theory and Bayesian statistics. Since the introduction of GOR in 1978 (Garnier et al., 1978), the training database set has been significantly enlarged, the statistics of pairs of residues have been added and many other improvements have been proposed (Garnier and Robson, 1989; Garnier et al., 1996; Gibrat et al., 1987). The performance of GOR V, which additionally uses evolutionary information contained in multiple sequence alignments, is presently comparable to the best cross-validated secondary structure prediction methods such as PHD (Rost, 1996, 2001) and PSIPRED (Jones, 1999). For example, the prediction accuracy measured by Q3 is 73.5% for GOR V, 71.9% for PHD and 76.6% for PSIPRED.

We have combined FDM and GOR V in the following manner: we define a sequence identity threshold to distinguish highly similar fragments obtained with the BLAST search from those less similar. In our previous work, we found the optimum value for the sequence identity score to be 55% (Sen et al., 2006). We have chosen the fragments (up to 5000 for the CDM server) with sequence identity above this threshold as an input for FDM. Because of this initial fragment sifting, FDM predicts the secondary structure only for sites having highly similar fragments. The secondary structure for the remainder of the sites is then predicted by GOR V, since GOR V relies on statistical averages and not on the sequence similarity of available fragments. The details of the CDM method and its performance are discussed in our recent paper (Sen et al., 2006). The Q3 prediction accuracy of CDM ranges from 67.5 to 93.2% depending on the availability of both sequentially similar structural fragments and multiple sequence alignments. These results demonstrate that CDM is one of the best secondary structure methods currently available, and we expect its accuracy to improve as the Protein Data Bank includes more structures. Here the lower end of the Q3 range, 67.5%, refers to the cross-validated GOR V Q3 predictions based on single sequences without evolutionary information from multiple sequence alignments. The higher end of the range, 93.2%, refers to the case when all the sequences are predicted by FDM. Note that the prediction by FDM requires the availability of highly similar fragments in the PDB, and therefore such Q3 cannot be cross-validated, but is strongly biased by the type and distribution of the sequences present in the PDB. However, it is important to note that the CDM secondary structure predictions will approach this upper limit of Q3, as more and more proteins become available in the PDB, significantly enhancing performance of FDM in the future.

In order to further validate this point, we culled 76 proteins recently deposited in the PDB, with <30% sequence similarity between each other, and an average length of 235 amino acids. We used the BLAST nr database generated before the deposition of these 76 proteins to the PDB. We calculated and compared Q3 values from GOR V, FDM, and CDM against PSIPRED predictions. Note that none of the Q3 values is cross-validated here, and they may contain biases since we did not check for similarity of these 76 sequences with the training data sets for all servers. All servers were treated equally because none of them contained information on these newly deposited PDB sequences. We found that for these sequences, the PSIPRED predicted Q3 =0.796 and CDM Q3 =0.755, at the sequence identity cutoff of 55%. However, the most striking feature in these calculations was the FDM value of Q3 computed as a function of sequence identity cutoff shown in Table 1. The increase of the sequence identity cutoff allows FDM to use fragments with higher similarities, and in turn, leads to higher values of Q3 for the predicted regions. These results substantiate our view that with the availability of more proteins in the PDB, the Q3 values of the secondary structure predictions by FDM will be constantly increasing and will exceed performance of other methods that do not rely on the structural templates.

2 IMPLEMENTATION

A user can obtain the secondary structure prediction of a sequence using our new CDM server. On the homepage of the CDM server (http://gor.bb.iastate.edu/cdm), the user is asked to enter his/her e-mail address and sequence information as a series of one-letter amino acid codes, up to 5000 residues in length. Once the information is submitted, the server checks the reliability of the e-mail address and the sequence information, and then sends a confirmation page to the browser (or an error notice if there is a problem). At this point, the server accepts the job and the user can close the web browser anytime without disturbing the job run. Another perl script then runs BLAST against pdb, and PSI-BLAST against the nr (non-redundant) database. The results of these searches are then used as inputs to FDM and GOR V, respectively. When the FDM, GOR V, and CDM runs are completed, the following information is sent to the user's e-mail address (as html links to the output files on the server): the secondary structure predictions of FDM, GOR IV, GOR V, and CDM; the secondary structure prediction weights for each site for GOR V; the fragment alignments and their identity scores used by FDM. The predictions for FDM, GOR V, and CDM are provided in two formats: either as a single line (for each method of prediction), or formatted so that each line contains up to 80 residues. These two formats should be sufficient for most users to facilitate their visualization of the prediction results.

The CDM server uses the RedHat Enterprise 3.0 system, built on a Dell Xeon with 4.6 GB memory. The server side CGI script is a combination of html and perl, and the program code is written in C++ (FDM and CDM) and Fortran (GOR V). The server is housed at the LH Baker Center for Bioinformatics and Biological Statistics, Iowa State University.

3 CONCLUSION

We have combined two complementary algorithms having different strengths, FDM and GOR V, to improve the performance of the secondary structure prediction. We have developed the CDM web server available for public and private use. As we showed in our previous work (Sen et al., 2006), combining FDM and GOR V benefits from the availability of experimentally determined structures and considerably enhances the secondary structure prediction. We are also planning to register our CDM server with EVA (EValuation of Automatic protein structure prediction) initiative (Eyrich et al., 2001) to benchmark the CDM performance in predicting protein secondary structure.

Acknowledgments

The authors acknowledge the financial support provided by the NIH grants 1R01GM072014 and R33GM066387.

References

Altschul SF, et al. Basic local alignment search tool. J Mol Biol 1990;215:403–410. [PubMed: 2231712]

Berman HM, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. [PubMed: 10592235]

Cheng H, et al. Prediction of protein secondary structure by mining structural fragment database. Polymer 2005;46:4314–4321. [PubMed: 19081746]

Eyrich VA, et al. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243. [PubMed: 11751240]

Garnier, J.; Robson, B. The GOR method for predicting secondary structures in proteins. In: Fasman, GD., editor. Prediction of Protein Structure and the Principles of Protein Conformation. Plenum Press; New York: 1989. p. 417-465.

Garnier J, et al. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120. [PubMed: 642007]

Garnier J, et al. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540–553. [PubMed: 8743705]

Gibrat JF, et al. Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. J Mol Biol 1987;198:425–443. [PubMed: 3430614]

Jayaram B, et al. Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucleic Acids Res 2006;34:6195–6204. [PubMed: 17090600]

Jones TD. Protein secondary structure prediction based on position specific matrices. J Mol Biol 1999;292:195–202. [PubMed: 10493868]

Kashlan OB, et al. Distinct structural elements in the first membrane-spanning segment of the epithelial sodium channel. J Biol Chem 2006;281:30455–30462. [PubMed: 16912051]

Kihara D. The effect of long–range interactions on the secondary structure formation of proteins. Protein Sci 2005;14:1955–1963. [PubMed: 15987894]

Kloczkowski A, et al. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 2002;49:154–166. [PubMed: 12210997]

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 2004;60:2256–2268. [PubMed: 15572779]

Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci USA 2003;100:12105–12110. [PubMed: 14528006]

Moult J. Rigorous performance evaluation in protein structure modelling and implications for computational biology. Philos Trans R Soc Lond, B, Biol Sci 2006;361:453–458. [PubMed: 16524833]

Rost B. PHD: Predicting one-dimensional protein structure by profile-based neural networks. Comput Methods Macromol Sequence Anal 1996;266:525–539.

Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218. [PubMed: 11551180]

Sen TZ, et al. GOR V server for protein secondary structure prediction. Bioinformatics 2005;21:2787–2788. [PubMed: 15797907]

Sen TZ, et al. A Consensus Data Mining secondary structure prediction by combining GOR V and Fragment Database Mining. Protein Sci 2006;15:2499–2506. [PubMed: 17001039]

Simons KT, et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225. [PubMed: 9149153]

Wray LV, Fisher SH. Functional analysis of the carboxy-terminal region of Bacillus subtilis TnrA, a MerR family protein. J Bacteriol 2007;189:20–27. [PubMed: 17085574]

Table 1

The Q3 values for FDM predictions as a function of the sequence identity cutoff

Sequence identity cutoff	Percentage of residues predicted by FDM	Q3 for the regions predicted by FDM
0.4	76.4	0.794
0.5	57.2	0.836
0.6	46.0	0.875
0.7	41.1	0.891
0.8	40.0	0.898
0.9	32.9	0.898