BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 14 2003, pages 1849–1851
DOI: 10.1093/bioinformatics/btg249

RVP-net: online prediction of real valued accessible surface area of proteins from single sequences

Shandar Ahmad[1,*], M. Michael Gromiha[2] and Akinori Sarai[1]

[1]Department of Biochemical Engineering and Science, Kyushu Institute of Technology, IZUKA, 820 8502, Fukuoka-ken, Japan and [2]Computational Biology Research Center (CBRC), AIST, 2-41-6, Koto-ku, Tokyo 135 0064, Japan

Received on December 19, 2002; revised on March 11, 2003; accepted on April 7, 2003

ABSTRACT

Summary: RVP-net is an online program for the prediction of real valued solvent accessibility. All previous methods of accessible surface area (ASA) predictions classify amino acid residues into exposure states and named them buried or exposed based on different thresholds. Real values in some cases were generated by taking the mid points of these state thresholds. This is the first method, which provides a direct prediction of ASA without making exposure categories and achieves results better than 19% mean absolute error. To facilitate batch processing of several sequences, a standalone version of this tool is also provided.

Availability: Online predictions are available at http://www.netasa.org/rvp-net/. Standalone version of the program can be obtained from the corresponding author by E-mail request.

Contact: shandar@bse.kyutech.ac.jp

INTRODUCTION

Solvent accessibility is one of the key properties of amino acids residues in proteins, which can be predicted, with a reasonable degree of accuracy (e.g., Rost and Sander, 1994; Pollastri *et al*., 2002). However all methods of accessible surface area (ASA) prediction have hitherto been based on classifying residues into exposure states and then making the category predictions. In the past, we provided an online ASA category prediction using a linear neural network and showed that relatively simpler networks with small data sets could be comparable to rigorous networks in terms of their prediction accuracy (Ahmad and Gromiha, 2002). We have also shown that ASA plays a major role in determining the probability of an amino acid residue to bind to DNA (Ahmad *et al*., 2002). We recently reported the first method for predicting real valued solvent accessibility without a priori classification into exposure states (Ahmad *et al*., 2003). In that work, we showed that our method of ASA prediction carries more information than even a 100% correct classification into two state categories (Rost and Sander, 1994). Although some of these methods provide a real value by averaging several category predictions, this was the first algorithm, which provided real valued ASAs without making classifications. Quality of predictions was measured by mean absolute error (MAE), which is defined as absolute per residue difference between the predicted and experimental values of relative ASA. We tested the validity of our new method, using different data sets of varying sizes. Non-redundant lists of proteins have been provided by Rost and Sander (1993), Carugo (2000), Cuff and Barton (2000), and Manesh *et al*. (2001). We also used two methods of ASA calculation from three-dimensional structures of proteins viz. DSSP (Kabsch and Sander, 1983) and ASC (Eisenhaber and Argos, 1993). We found that the change in the method of ASA calculation and variation of data sizes within the attempted range affected the prediction quality by less than 2% MAE. For the online predictions, we have used the largest of these data sets (Cuff and Barton, 2000), which could enhance its predictability and reliability. Neural network was trained for more than 80 000 residues provided in this data set and resultant MAE of prediction was found to be nearly 19%. Correlation between the predicted and desired values of ASA has been found to be 0.480 for this data set. Correlation between the desired and predicted solvent accessibility in binary states was previously reported as 0.432 based on single sequence information by Rost and Sander (1994) compared to a value of 0.472, obtained by us between the desired and predicted real values of solvent accessibility for the same data set. Main advantages of real valued predictions as against the state predictions are as follows:

1. Real valued predictions provide more information. Even if there were 100% correct prediction in two-state classification, the mean doubt in ASA prediction is 25%. Since our MAE of predictions is just 19%, real valued predictions carry at least 6% more information than a 100% correct two-state prediction.

*To whom correspondence should be addressed.

Table 1. A comparison between available ASA prediction servers and RVP-net

Property	NETASA	PHD,[1] JPRED[2] and other ASA prediction servers	RVP-net ASA prediction server
Algorithm			
Neural network type	Feed forward multi-layer, classification neural network	Similar to NETASA	Feed forward multi-layer function mapping neural network
Neural network output	Binary (0 or 1)	Binary (one, two or more units)	Real Value (between 0 and 1)
Data used	Training: 30 proteins, Test: 185 proteins	Upto 512 proteins, larger training data sets	Four independent data sets of 126, 215, 338 and 512 proteins,
ASA calculation from structure	ASC	DSSP and others	ASC and DSSP
ASA processing for training	Two or three states of solvent accessibility	Two, Three and 10 States	Absolute ASA scaled to relative values by a factor of extended state ASA of the residue
Neural network error function	Based on relative number of correct predictions in either state	Similar to NETASA	Per residue absolute difference between predicted and experimental value of ASA
Prediction server			
Sequence inputs accepted	One letter code unformatted sequence	FASTA and other standard formats	FASTA, SwissProt, PIR format or unformatted single letter code
Prediction type	Binary or ternary state at selected threshold	Binary prediction to 10 state prediction. 10 state thresholds separated unevenly by n^2 law	Realative ASA, absolute ASA in Å^2 and binary prediction at fixed threshold
Results provided by	Immediately online	Mainly by E-mail	Immediately online
Speed	Fast	Slower due to alignment calculations	Fast
Standalone program availability	Yes	Mostly no	Yes

[1]Rost and Sander (1994).
[2]Cuff and Barton (2000).

2. Same value of ASA threshold may not be significant for all residues. Some residues have significantly lower mean ASA and hence a state classification at the same threshold for all residues may not be justified. By providing a real value prediction, our method enables us to classify residues into categories, if needed, at different thresholds for different residues within the same prediction.

In view of this, RVP-net provides a useful tool for the prediction of solvent accessibility and hence will be helpful in estimating structure and function of proteins with unknown three-dimensional structures. A detailed comparison between our previously published server Netasa (www.netasa.org/netasa/) and other available category prediction servers on the one hand and the present RVP-net on the other is provided in Table 1. RVP-net server is flexible to the input formats as it can take the sequence input in a two-column format (residue number) or a single running sequence. Numeral characters, spaces and next line characters are ignored by the sequence-parsing program. Single letter code is expected for residues and all residues must be entered in upper case. Lower case and non-residue characters are treated as unknown residues. Largest size of sequence accepted is 700. The RVP-net web server allows one query at a time and in some cases, users may be interested in making large-scale predictions. Therefore, we also distribute a standalone version of this program to enable users to make batch predictions of solvent accessibility without needing to access the internet. Online prediction and the standalone programs are provided free for academic use.

NOTE ADDED IN THE PROOF

After the acceptance of this paper, RVP-net predictions are also available in graphical formats.

REFERENCES

Ahmad,S. and Gromiha,M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, **18**, 819–824.

Ahmad,S., Gromiha,M.M. and Sarai,A. (2002) Specificity and predictability of DNA binding in terms of one-dimensional properties of proteins. *Protein Sci.* (suppl.), **11**, 53.

Ahmad,S., Gromiha,M.M. and Sarai,A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.

Carugo,O. (2000) Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng.*, **13**, 607–609.

Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.

Eisenhaber,F. and Argos,P. (1993) Improved strategy in analytical surface calculation for molecular system—handling of singularities and computational efficiency. *J. Comp. Chem.*, **14**, 1272–1280.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bond and geometrical features. *Biopolymers*, **22**, 2577–2637.

Manesh,H.N., Sadeghi,M., Arab,S. and Movahedi,A.M. (2001) Prediction of protein surface accessibility with information theory. *Proteins*, **42**, 452–459.

Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2002) Prediction of coordination number and relative solvent accessibility. *Proteins*, **47**, 142–153.

Rost,B. and Sander,C. (1993) Improved prediction of protein secondary structure by using sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.