Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

Prediction of protein solvent profile using SVR

Zheng Yuan[1,2,†], Timothy L. Bailey[1,2,‡]
[1]Institute for Molecular Bioscience University of Queensland, Brisbane, Australia
[2]ARC Centre in Bioinformatics, University of Queensland, Brisbane, Australia
[†]z.yuan@imb.uq.edu.au, [‡]t.bailey@imb.uq.edu.au

***Abstract*—We describe a support vector regression (SVR) approach to predict the accessible surface area (ASA) of a protein from its sequence. Our approach encodes each protein residue as a vector of amino acid propensities derived from a multiple alignment of the subject protein with homologous proteins. The vector consists of the log-likelihood ratios of each of the twenty amino acids in the residue's multiple alignment column. Using a reference set of proteins of known structure and, hence, known ASA, we trained an SVR model. Each training sample consists of the fifteen log-likelihood vectors in a window of width fifteen surrounding a residue, along with the "true" ASA value, computed from the known structure. To apply the model to proteins of unknown structure, only the subject protein sequence is required. Our method uses PSI-BLAST to simultaneously determine a set of (putative) homologs and compute the log-likelihood vectors needed to encode the subject protein. We show that this method provides substantially improved accuracy in predicting ASA when compared with an earlier method.**

***Keywords*—protein; solvent profile; accessible surface area; support vector regression**

I. Introduction

Determining which residues are on the surface of a protein is extremely important to understanding its function. To a large extent, only residues that are solvent accessible can interact with the protein's environment. For example, active sites are always located on the protein's surface.

The absolute solvent accessibility (ASA) of a protein residue is the surface area of the residue available for interaction with the solvent. Determining the ASA of a residue is relatively easy when the protein's structure is known. However, since most proteins do not have solved structure, we are interested in the problem of predicting solvent accessibility directly from the sequence of the protein.

A great deal of research has looked at the related problem of classifying residues into two or more solvent accessibility classes [1], [2], [3], [4]. Two-class (buried/accessible) prediction methods have achieved accuracies as high as 75% when multiple sequence alignment data is used. In this paper, we address the more difficult problem of predicting the ASA value of each residue, rather than simply predicting its class. Of course, any ASA prediction method can also be used to solve the class prediction task.

Our goal is to predict the ASA of each residue in a protein of unknown structure. In previous work [5], we used support vector regression (SVR) [6], [7] to predict ASA values. We based the prediction on the linear context of the residue–the identities of the residues on either side of it in the linear protein chain. In this work, we seek to improve the accuracy of the method by incorporating information implicit in a multiple alignment of the subject protein with (putatively) homologous proteins.

In our previous approach we encoded the context of a residue as a sequence of fifteen context vectors, $X = (x_{-7}, x_{-6}, \ldots, x_{-1}, x_0, x_1, \ldots, x_7)$. These vectors represented the amino acids in a window of width fifteen around the subject residue (with x_0 representing the subject residue itself.) Each context vector, x_i, represented a single amino acid, encoded as a length-21 unary vector. Positions one through twenty in the vector correspond to the twenty common amino acids listed in their order in the IUPAC single-letter code [8]. The twenty-first position in the vector corresponds to the other letters in the IUPAC code (B, U, X, Z). (E.g., "A"$= (1, 0, \ldots, 0)$, "C"$= (0, 1, 0, \ldots, 0)$, ..., "other" $= (0, \ldots, 0, 1)$.)

In the current work, we encode each context residue as a length-21 vector of real numbers representing the information contained in that residue's column in a multiple alignment. To derive this encoding, we first perform a multiple alignment of the subject protein with a set of homologous proteins. Then, for each column of the multiple alignment, we calculate the observed frequency of each of the twenty amino acids. We convert these frequencies to log-likelihood ratios by dividing them by the relative overall frequencies of the amino acid and taking logarithms. This results in a length-20 vector of log-likelihood ratios encoding each context residue. That is, the residue at position i in the subject protein is encoded as $x_i = (ll_A, ll_C, \ldots, ll_Y)$, where ll_a is the log-likelihood of residue a appearing aligned with residue i in the subject protein. A twenty-first component is set to zero for most residues. However, we mask low complexity and coiled-coil regions in the input sequences. Masked residues are encoded using the unary-encoding described above.

II. Methods

We perform SVR on a set of training samples, $T = \{(X_i, y_i)\}$, derived from a large reference set of proteins of known structure. Each amino acid in each protein

0-7803-8439-3/04/$20.00©2004 IEEE

in the reference set contributes one sample to T. Each sample consists of the linear context sequence, X_i, and a calculated ASA value, y_i. We describe the formulation of the problem and the encoding method for deriving X_i below. Using the same encoding method, the ASA of each residue of a target protein of unknown structure can be estimated from the X_i for the target protein.

A. Support Vector Regression

We formulate the ASA prediction problem as a support vector regression problem following [7]. Each residue in the proteins in the training set is encoded, as described in the next section, into a vector we will call X. Then, X is mapped (non-linearly) onto an m-dimensional feature space. A linear model is constructed in this feature space. The predicted ASA value will be given by

$$f(X) \ = \ \sum_{j=1}^{m} w_j \Phi(X) + b.$$

Here, $\Phi(X)$ is the non-linear mapping and b is the "bias". The regression parameters w_j and b are estimated by minimizing the sum of the norm of the "weights", $\|w\|^2$, and the empirical risk on the training samples. In particular, Vapnik's ϵ-insensitive loss function is used here to quarantine the errors. It is defined by

$$L_\epsilon(y - f(X)) \ = \ \begin{cases} 0, & \text{if } |y - f(X)| \leq \epsilon \\ |y - f(X)| - \epsilon, & \text{otherwise.} \end{cases}$$

The empirical risk function is

$$\frac{C}{n} \sum_{i=1}^{n} L_\epsilon(y_i - f(X_i)),$$

where C is a (user-setable) regularization constant and n is the number of training samples. Only those errors that are larger than ϵ contribute to the risk function. The size of C determines the relative contributions of the model complexity and error terms to the risk function.

To reduce the complexity of the optimization problem, slack variables ζ and ζ^* are introduced to measure the deviation of samples outside the ϵ-insensitive zone. Thus, SVR is formulated as:

$$\text{minimize:} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} (\zeta_i + \zeta_i^*)$$

$$\text{subject to:} \quad \begin{cases} f(X_i) - y_i \leq \epsilon + \zeta_i \\ y_i - f(X_i) \leq \epsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0, \text{ for i = 1,...,m} \end{cases}$$

This can be transformed to the dual problem. Its solution is given by

$$f(X) \ = \ \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(X_i, X) + b,$$

where the dual variables are constrained to $[0, C]$ and the kernel function,

$$K(X, X') = \Phi(X) \bullet \Phi(X').$$

In this study we use the radial basis function

$$K(X, X') = e^{-\gamma \|X - X'\|^2}.$$

The user-settable parameter gamma determines the radius of kernel.

In our implementation, we use the SVM_Light [9] package to perform support vector regression.

B. Residue encoding

To encode a subject protein of known or unknown structure, we begin by creating a multiple alignment. We use PSI-BLAST [10] run for three iterations against the NCBI non-redundant protein database to create the multiple alignment. For each residue i in the subject protein, PSI-BLAST directly outputs the length-twenty vector, x_i, of log-likelihood ratios described in the introduction. The values of x_i in a window of length fifteen are used to compose the context sequence $X_i = (x_{i-7}, x_{i-6}, \ldots, x_i, x_{i+1}, \ldots, x_{i+7})$. Each residue is thus mapped onto an $15 \cdot 21 = 315$ dimensional feature space.

C. Reference protein set

We prepared a high-quality dataset of 945 non-redundant protein chains from the Protein Data Bank (PDB) [11] extracted using PDB-REPRDB [12]. The pairwise identity of all proteins in the dataset is not more than 25%. All proteins are at least 60 amino acids long. For structures solved using X-ray crystallography, only those with resolution ≤ 2.0Å^2 and R-factor ≤ 0.2 are included in our dataset. For the NMR structures, we use only the first model in its PDB entry. (Datasets are available at URL http://www.uq.edu.au/ uqgyuan/embs2004.)

For proteins in the reference set, we compute the ASA of each residue using the SURFace algorithm [13]. (We refer to this value as the "true" ASA.) Our previous research [5] has shown that accuracy is improved by predicting normalized ASA rather than raw ASA. Hence, we normalize the calculated ASA by dividing it by the ASA value of the subject amino acid, R, in the tripeptide Ala-R-Ala, as given in [14]. This gives the value of y_i for the subject residue. Absolute ASA values are easily recovered by multiplying by the same normalizatIon constants.

TABLE I
COMPARISON OF ACCURACY OF ASA PREDICTIONS FOR THE TWO SVR METHODS.

method	*C*	*mean absolute error* ($\AA^2$)	*correlation coefficient*
S1	5.0	31.28 ± 0.04	0.597 ± 0.002
S2	2.0	30.26 ± 0.05	0.621 ± 0.002
M1	2.0	27.52 ± 0.05	0.679 ± 0.002
M2	5.0	$\mathbf{26.82 \pm 0.05}$	$\mathbf{0.693 \pm 0.002}$

D. Measurement of prediction performance

We use a testing methodology related to cross-validation to measure the accuracy of our prediction method. We divide the reference protein set into three groups each containing 315 protein chains. Three-fold cross-validation would use two groups for training and one group for testing. Because of the large amount of computer time required for training, we instead trained on one group and tested on two, as follows. For each group, we create a set of samples by encoding the residues and computing the ASA as described above. We then learn the SVR parameters using one group of samples, and test using the other two groups. This procedure is repeated three times, using each group of samples as the training set once. The final results are the average of the three rounds of tests.

To measure the prediction performance of SVR, the ASA absolute error is calculated for each residue, defined as the absolute value of the difference in "true" and predicted ASA. We report the mean absolute error for each round of testing. For each protein, we also compute the Pearson's correlation coefficient between "true" and predicted ASA values.

We also consider the accuracy of our regression models when used as a classifier. We choose various threshold values on "true" ASA for dividing residues into two solvent accessibility classes (buried/exposed). We plot the ROC curve [15] for the predicted ASA values of all test residues.

E. Comparison with other methods

Our previous research [5] shows that our single-sequence SVR method is superior to other known methods for predicting ASA. We compare our new method, therefore, against our earlier single-sequence SVR method run on the same dataset.

Both the single-sequence and our current method have three user-settable parameters. For the single-sequence SVR method, we used the best parameters as determined in our earlier research. For our current, multiple-sequence SVR method, we did not optimize over the possible user-setable parameters. We used

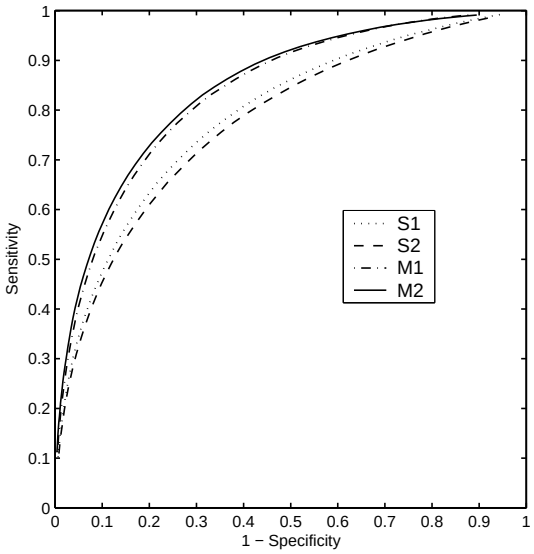

Fig. 1. **ROC analysis of ASA classification accuracy.** The figure shows the ROC plots for the same data as in Table I. Residues with "true" ASA values greater than 45$\AA^2$ were classified as exposed; otherwise, they were classified as buried. The ROC curves plot the sensitivity (true positives/positives) versus 1 - specificity (false positives/negatives) for each possible predicted ASA threshold.

$\epsilon = 0.01$, $\gamma = 0.01$, and tried two values of C, $C = 2$ and $C = 5$. This puts our new method at a relative disadvantage, making any inference of superiority more conservative.

III. RESULTS

ASA prediction accuracy using the multiple alignment and single sequence SVR is summarized in Table I. In the table, methods S1 and S2 are the single-sequence SVR method with two different values of C; similarly, methods M1 and M2 are the multiple-sequence SVR method. The average absolute errors in predicted ASA for all sequences is shown for the two methods is shown. The mean correlation coefficient of predicted ASA and "true" ASA is also shown, averaged over the sequences in the test groups. Results are expressed as mean $\pm$ the standard error. The best results for each metric are highlighted in bold type. Under both accuracy metrics–mean absolute error and average correlation coefficient–the multiple-sequence method, regardless of the choice of C, is substantially better. Overall, the multiple-sequence method using the larger value of C ($C = 5$), which increases the importance reducing training errors as compared with model complexity, is most accurate according to all two metrics. Clearly the use of the encoding incorporating multiple sequence alignment information improves prediction accuracy.

We also studied the accuracy that can be obtained when the predicted ASA values are used in a classifier. Using the same results as in Table I, we applied six different thresholds for classifying residues as buried versus exposed: 5, 25, 35, 45, 75, 100 $Å^2$. In each case, the ROC curves for the current method (M1 and M2) "dominated" the curves for our previous, single sequence method (S1 and S2). Figure 1 shows a typical example. All the points on the curves for methods M1 and M2 lie above the corresponding points on the curves for S1 and S2. This means that, for any desired level of sensitivity, the specificity of the new method is always higher on this data.

IV. Discussion

We have described a new method for predicting the ASA of protein residues from primary sequence information. We have shown that it is more accurate under various measures than a previous method we described and showed to be more accurate than the state-of-the-art. Therefore, we believe our current method to be better than any existing method. We will make our ASA predictor available as a web service via URL `http://bioinformatics.org.au`.

The improved accuracy of our method comes from incorporating the extra information present in multiple alignments. The function of proteins constrains the evolution of their residues, and this signal is present when we align the homologs present in a non-redundant database with a subject protein. One can speculate that this reduces the noise present when we look at the sequence of a single protein and try to predict ASA based solely on local sequence context.

The improvement in prediction accuracy comes at a cost in computation time relative to our previous method. The earlier method used a unary encoding for each residue. This resulted in a very sparse encoding. Our current method encodes each residue as a real vector. Training the SVR takes much longer as a result. One line of future research is to explore ways to speed up the SVR training. Computing the ASA of a novel protein takes only a minute or so plus the time for a PSI-BLAST search of the non-redundant database. This time is more important than the training time, which only need be done once, or at most periodically when the PDB has grown substantially since the last training.

References

[1] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins*, vol. 20, pp. 216–226, 1994.

[2] J.A. Cuff and G.J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins*, vol. 40, pp. 502–511, 2000.

[3] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, "Prediction of coordination number and relative solvent accessibility in proteins," *Proteins*, vol. 47, pp. 142–153, 2002.

[4] Z. Yuan, K. Burrage, and John S. Mattick, "Prediction of protein solvent accessibility using support vector machines," *Proteins*, vol. 48, pp. 566–570, 2002.

[5] Z. Yuan and B. Huang, "Prediction of protein accessible surface areas by support vector regression," *Proteins*, in press.

[6] A. Smola and B. Scholkopf, *A tutorial on support vector regression*, Available: `http://www.neurocolt.com`, 1998.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[8] JCBN, "Nomenclature and symbolism for amino acids and peptides," *European Journal of Biochemistry*, vol. 138, pp. 9–37, 1984.

[9] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Sch olkopf, C. Burges, and A. Smola, Eds., pp. 42–56. MIT, 1999.

[10] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinhui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman, "Gapped-BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 1997.

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[12] T. Noguchi and Y. Akiyama, "Pdb-reprdb: a database of representative protein chains from the protein data bank (pdb) in 2003," *Nucleic Acids Research*, vol. 31, pp. 492–493, 2003.

[13] A. Nicholls, K. Sharp, and B. Honig, "Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons.," *Proteins*, vol. 11, pp. 281–296, 1991.

[14] S. Ahmad, M.M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," *Proteins*, vol. 50, pp. 629–635, 2003.

[15] R.M. Centor, "Signal detectability: The use of roc curves and their analyses," *Medical Decision Making*, vol. 11, pp. 102–106, 1991.