

Classification Comparison of Prediction of Solvent Accessibility From Protein Sequences

Huiling Chen¹, Huan-Xiang Zhou², Xiaohua Hu³, Ilhoi Yoo³

¹Department of Physics, Drexel University, Philadelphia, Pennsylvania 19104.

²Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32312.

³College of Information Science and Technology, Drexel University, Philadelphia, Pennsylvania 19104
thu@cis.drexel.edu

Abstract

The prediction of residue solvent accessibility from protein sequences has been studied by various methods. The direct comparison of these methods is impossible due to the variety of datasets used and the difference in structure definition. In this paper we choose 5 classification approaches (decision tree (DT), Support Vector Machine (SVM), Bayesian Statistics (BS), Neural Network (NN) and Multiple Linear Regression (MLR)) for predicting solvent accessibility based on the same dataset and using the same structure definition so that we can directly compare different methods. We evaluate these methods in a cross-validation test on 2148 unique proteins using single sequences and multiple sequences approaches with a cutoff of 20% for two-state definition of solvent accessibility. According to the experiment results, SVM and NN are both the best predictors with accuracy 79%, correlation coefficient 0.59, 2~4% superior to other three methods on multiple sequences prediction. A further test result on a blind test set from Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP5) is consistent with this result. On single sequence prediction, DT, BS and MLR perform about the same at 71~72% with correlation coefficient 0.43. The improvement over the baseline model that use only the identity of target residue is small. Local sequence seems embed very little information on accessibility. Separate training according to protein size improves the prediction when there are sufficiently large dataset available. The consensus prediction combining the 5 approaches is not significantly better than the best single method.

Keywords: protein structure prediction, classification comparison, ensemble prediction, solvent accessibility.

1 Introduction

Knowledge of a protein's 3-dimensional structure is essential for a full understanding of its functionality. However, only a tiny fraction of the enormous number of proteins that have been sequenced have their structures determined. As large-scale gene sequencing projects continue to widen the sequence-structure gap, developing reliable and generally applicable structure prediction

methods has become an urgent problem and one of the most important tasks of theoretical biologist. Prediction of the 3D structure from protein sequence should be feasible based on the well-established credo that protein sequence uniquely determines protein structure (Anfinsen 1973). However, an accurate structure prediction is currently only possible for proteins with significant sequence similarity to proteins of known structure by homology modelling. For some of the remaining sequences, the structure prediction may be feasible by threading (sequence-structure alignment) in the absence of significant sequence similarity. For most of the sequences without related proteins of known structure, there is not yet a reliable method available for structure prediction. Thus, the simplification of the problem, reducing 3D structure to 1D features, may be useful and regarded as the first-step in understanding the protein-folding problem. The prediction of secondary structure is the most familiar and well-defined aspect of the problem. The prediction of solvent accessibility, i.e. amino acid residue on the surface or in the interior of protein molecule, is another aspect of the problem.

The studies of solvent accessibility in proteins have led to numerous insights into protein structures. It has already been shown that in proteins, the hydrophobic free energies are directly related to the accessible surface area of both polar and non-polar groups (Ooi *et al.*, 1987). In the final folded structure of a protein, the hydrophilic side-chains have access to the aqueous solvent, but the contact between the hydrophobic side-chains and solvent is minimized. The discovery by Chan and Dill (1990) shows that burial of core residues may be a strong driving force in protein folding. The prediction of residue solvent accessibility can aid in elucidating the relationship between protein sequence and structure.

In addition to providing insight into the organization of 3D structure, prediction of residue solvent accessibility has many other applications. First, it has been observed that the distribution of surface residues of a protein is correlated with its subcellular environments and, consequently, using the information of surface residues has made an improvement in the prediction of protein subcellular location (Andrade *et al.*, 1998). Second, solvent accessibility has also been used to predict the position of protein hydration sites, which may play an important part in a protein's function (Ehrlich *et al.*, 1998). Third, accurately predicting the buried residues is useful for identifying those residues, when appropriately mutated, are most likely to give rise to a temperature-sensitive phenotype (Varadarajan *et al.*, 1996).

Prediction of solvent accessibility has been implemented by various methods by examining a window of residues centered at the test residue and using amino acid identity (single sequence prediction) or the sequence profile (multiple sequence prediction) of these residues as input attributes. Rost and Sander(1994) used a Neural Network method, Tompson and Goldstein (1996) applied Bayesian statistics, Li and Pan (2001) developed a multiple linear regression method, Naderi-Manesh *et al.*(2001) introduced a method based on information theory, and most recently Yuan *et al.* (2002) implemented support vector machine. Due to the variety of datasets used and the difference in structure definition, the direct comparison of these methods is impossible. The prediction accuracies reported are approximately 70-72% for single sequence prediction and 73-76% for multiple sequence prediction for two-state (buried/exposed) prediction. It is very useful to compare the different approaches to find out which methods have better performance. In this study, we made a direct comparison of 5 classification methods, decision tree (DT), support vector machine (SVM), neural network (NN), multiple linear regression (MLR), and Bayesian statistics (BS), on the same database and using the same structure definition and combined them for consensus prediction.

2 Data set

The 2148 unique proteins (sequence identity <25%) was selected from the FSSP database (Holm and Sander, 1996) (last update Oct 30, 2001) with sequence length ≥ 90 amino acid residues. We excluded the short chain effect based on the fact that most of the residues in a small protein are exposed to solvent, thus hydrophobic residues are less likely to be sheltered in a hydrophobic interior. The residue surface area was extracted from the DSSP files of Kabsch and Sander(1983). Percent solvent accessibility was computed by normalizing the accessible surface area over the maximum values of amino acids obtained by Shrake and Rupley(1973). A cutoff 20% was used to define the two states(buried/exposed), which evenly splits the dataset. Sequence profiles used in multiple sequence prediction were extracted from the position-specific scoring matrices produced by PSI-BLAST (Altschul *et al.*,1997). We performed a five-fold cross-validation test on this dataset and a blind test on 21 proteins from CASP5 without significant sequence similarity to any protein of known structure.

3 Classification Methods

In our study, 5 classification methods : decision tree (DT), support vector machine (SVM), Neural network (NN), multiple linear regression (MLR) and Bayesian statistics (BS) were compared on the same dataset and using the same structure definition.

Decision tree (DT): Decision tree has been used for a wide range of applications including bioinformatics such as gene finding(Salzberg *et al.*,1998), pattern recognition in genome(Delamarche *et al.*,1999), prediction of protein cellular localization sites(Horton *et al.*,1997), prediction of secondary structure(Selbig *et al.*,1999). A Decision tree approach recursively split the data sets into different

subtrees based on the values of the one attribute (or a few attributes) until all (or almost) the data points in the nodes are in the same category. Its main advantage is that the rules derived from the decision trees are easier to understand how the decision tree has arrived at its decision. In our comparison study, we used C5.0 by Quinlan (<http://www.rulequest.com>).

Support Vector Machine (SVM): SVM proposed by Vladimir Vapnik based on statistical learning theory has quickly become one of the most popular classification and regression methods. It has been used extensively in a lot of application domains such as microarray data analysis (Furey *et al.*,2000), protein structure prediction(Ding and Dubchak,2001; Hua and Sun, 2001) because of its flexibility in choosing a similarity function, the ability to handle large feature spaces and very good accuracy. Support Vector Machines perform binary classification and regression estimation tasks based on the structural risk minimization principle. SV machines create a classifier with minimized VC dimension. If the VC dimension is low, the expected probability of error is low as well, which means good generalization. Yuan *et al.*(2002) implemented SVM^{light} (Joachims 1999) in predicting solvent accessibility and reported better result than the neural network(Rost and Sander,1994) and Bayesian statistics (Tompson and Goldstein,1996) based on a set of 126 proteins. In our study, we compared these methods on a more extensive dataset of 2184 proteins.

Neural Network (NN): NN is a computer software to simulate the human neuron connectivity. Each neuron has a certain number of input nodes with a associated *weight* value for each of them. The weights are an indication of the importance of the incoming signal. The *net value* of the neuron is then calculated. Each neuron has its own threshold value, and if the *net* is greater than the threshold, the neuron fires (or outputs a 1), otherwise it stays quiet (outputs a 0). The output is then fed into all the neurons in the next level. Many types of neural network have been developed over the years including back-propagation, the delta rule and Kohonen learning. In our comparison study, a two-stage feed-forward, back-propagation neural network proposed by Jones(1999) was used. The results of this approach were found to be better than the popular PHD (NN by Rost and Sander) in protein secondary structure prediction.

Bayesian Statistics (BS): Bayesian statistics uses probability theory to manage uncertainty by explicitly representing the conditional dependencies between the different knowledge components. In protein structure prediction, Bayes' theorem allows us to express the conditional probability of a particular structure given knowledge of the corresponding sequence in terms of the conditional probability of the sequence given the particular structure. Because of the low occurrence probability for any stretch of residues in protein sequences, statistically significant results for the burial probability of a residue inside a particular stretch of residues cannot be obtained from any training set. Therefore, assumptions must be made. The simplest assumption is that the probability for a type of residue to appear in a site within a segment of accessibility states is independent of

neighboring positions. In our study, we used Thompson and Goldstein's method(1996).

Multiple Linear Regression (MLR): Multiple linear regression is a very powerful statistical tool that is relatively complex but of immense use. Multiple linear regression is to obtain the least squares equation to predict some response. With multiple linear regression, there are multiple predictors. This procedure performs linear regression on the selected dataset. This fits a linear model of the form: $Y = c_0 + c_1X_1 + c_2X_2 + \dots + c_pX_k + e$ where Y is the dependent variable (response), p is the number of predictors and X_1, X_2, \dots, X_k are the independent variables and e is random error. $c_0, c_1, c_2, \dots, c_k$ are known as the regression coefficients, which have to be estimated from the data. In our study, we implemented Li and Pan's method(2001) in predicting residue solvent accessibility.

Accuracy measurement: Prediction accuracy was measured by percentage of correctly predicted residues and the correlation coefficient between the observed and the predicted states, as given by

$$\frac{N \sum o_i p_i - \sum o_i \sum p_i}{\sqrt{N \sum o_i^2 - (\sum o_i)^2} \sqrt{N \sum p_i^2 - (\sum p_i)^2}}$$

where $o = \{0,1\}$ is the observed state, $p = \{0,1\}$ is the predicted state, N is the number of residues predicted.

4 Result

The baseline for prediction of solvent accessibility Appropriate evaluation of the performance of any prediction requires an analysis of the worst prediction (random) as a baseline. On residue solvent accessibility prediction, we used a more challenging baseline by Richardson & Barlow(1999). The method makes predictions solely on the basis of the exposure category in which an amino acid is most often found. In particular, those exposed >50% of the time in the training set is predicted to be exposed; the rest is predicted to be buried.

Table I lists the properties of the 20 standard amino acids and their average occurrence and probability for exposure based on our dataset. The 20 amino acids can be grouped into five main classes based on the polarity or tendency to interact with water of their R group at biological pH(near pH 7.0)(Nelson&Cox). The statistical data confirms that the polar (hydrophilic) side chains tend to be on the surface of a protein and the non-polar (hydrophobic) side chains tend to be in the interior, except for Cys, Pro and Gly. The reasons for non-polar R groups of Pro and Gly tend to be exposed are easily explained from their structures (Nelson&Cox). And two Cys are readily oxidized to form a disulfide bond. The disulfide-linked residues are strongly hydrophobic.

According to the statistical data, the baseline method predicts A, V, L, I, M, F, W, Y, C to be buried, and G, P, S, T, N, Q, K, R, H, D, E to be exposed. The accuracy is 69.6% compared to 52% by random prediction. The method is simple and takes no account of the local sequence surrounding a residue. It should be used as a

baseline by which more sophisticated approaches can be judged.

Amino acid	Hydropathy index	exposure (%)	Occurrence (%)
Nonpolar R group (hydrophobic)			
Gly G	-0.4	53.3	7.3
Ala A	1.8	40.2	8.1
Val V	4.2	27.5	7.0
Leu L	3.8	26.7	9.1
Ile I	4.5	23.9	5.7
Met M	1.9	33.4	2.1
Pro P	1.6	64.4	4.7
Aromatic R group (hydrophobic)			
Phe F	2.8	26.8	4.0
Trp W	-0.9	31.2	1.4
Tyr Y	-1.3	39.3	3.6
Polar, uncharged R group (hydrophilic)			
Ser S	-0.8	59.2	6.0
Thr T	-0.7	57.0	5.7
Cys C	2.5	31.5	1.4
Asn N	-3.5	69.7	4.4
Gln Q	-3.5	74.6	3.8
Positively R charged (hydrophilic)			
Lys K	-3.9	88.0	5.9
Arg R	-4.5	75.8	5.0
His H	-3.2	57.7	2.3
Negatively R charged (hydrophilic)			
Asp D	-3.5	76.6	5.9
Glu E	-3.5	82.5	6.6

Table I: Properties and statistics of the standard amino acids. Hydropathy index is a scale combining hydrophobicity and hydrophilicity of R groups; it can be used to measure the tendency of an amino acid to seek an aqueous environment(- value) or a hydrophobic environment(+ value). (Kyte & Doolittle,1982)

Window size	Accuracy (%)	Window size	Accuracy (%)
1	69.6	11	71.0
3	70.3	13	71.0
5	70.6	15	71.1
7	70.9	17	71.1
9	71.0	19	71.1

Table II: Prediction accuracy using different window size

Effect of different window sizes The approaches used to predict solvent accessibility are similar to those used to predict secondary structure. The rationale is clear: in both cases the investigator hopes that the local amino acid sequence has a strong influence on the one-dimensional property under consideration. Increasing the window size can provide more local information. We chose different window size using DT method to investigate the results. From Table II, we found that window size has a very small influence on prediction accuracy. Using window size=1

Data set	No seq.	% Exp.	Joint Training					Separate Training						
			BL	BS	MLR	DT	NN	BS	MLR	DT	NN	SVM	WE	MW
Set 1	886	63	74.3	74.8	76.7	76.8	79.5	76.0	75.5	77.5	79.6	79.5	80.1	79.0
			0.46	0.48	0.51	0.52	0.56	0.48	0.46	0.52	0.56	0.56	0.57	0.55
Set 2	883	51	73.7	75.6	76.3	77.4	78.9	75.6	76.4	77.1	79.2	79.8	80.2	79.0
			0.48	0.51	0.53	0.55	0.58	0.51	0.53	0.54	0.58	0.60	0.60	0.58
Set 3	379	46	71.9	74.1	75.4	75.7	78.1	74.7	75.5	76.3	78.4	78.9	79.4	78.3
			0.45	0.49	0.51	0.52	0.56	0.49	0.51	0.52	0.57	0.58	0.59	0.56
All	2148	52	73.4	75.0	76.2	76.8	78.9	75.5	76.0	77.0	79.1	79.5	80.0	78.8
			0.46	0.50	0.52	0.54	0.58	0.51	0.52	0.54	0.58	0.59	0.60	0.58

Table III. Results of multiple sequence prediction. The results are given as percentage of correctly predicted residues and correlation coefficient between the observed and predicted states. The whole dataset were divided into three subsets according to protein sequence length: set1(90-199aa), set2(200-439aa), set3(≥ 440 aa). Separate training: to train and test each subsets separately and the results were summed up. Joint training: to train and test the whole dataset and the results were divided into 3 subsets. BL: baseline method, BS: Bayesian statistics, MLR: multiple linear regression, NN: neural network, SVM: support vector machine, WE: weighted ensemble, MW: “majority win” consensus.

DT predicts at the same way as baseline method does and gives the same result of 69.6%. Half of the improvement from $w=1$ to $w=15$ was gained using $w=3$. There is no improvement using window size >15 . Similar results were gained by other methods. The results suggest that local sequence doesn’t have as strong influence on solvent accessibility as on secondary structure. The preference of a residue for solvent exposure or burial mainly depends on the property of its side chain. For maximum performance, we selected 15 as the window size for DT in the following computing process. The window size used for BS and MLR methods was 13, for NN and SVM was 15, according to their authors.

Single sequence prediction We compare the performance of three methods for predicting solvent accessibility using single sequence prediction. The performance of BS(71.2%/0.42), MLR(71.6%/0.43) and DT(71.5%/0.43) are very close. The methods, which take into account of the influence of neighboring residues, didn’t do much better than the baseline(BL) (69.6% and 0.39). The result suggests that the influence of neighboring residues on its solvent accessibility is weak. Local sequence seems embed very little information on accessibility. The problem is not with methods.

Multiple sequence prediction One key to more accurate predictions of 1D structure, such as secondary structure or solvent accessibility is the use of evolutionary information because profiles of residue substitutions in naturally evolved protein families are highly specific for details of a particular protein structure (Rost and Sander, 1995). It has been shown that PSI-BLAST profiles produced more accurate secondary structure predictions than the sequence profiles generated by multiple sequence alignment (Jones 1999). In addition, PSI-BLAST profiles are very convenient to use because they are generated as part of the search process. Using the intermediate PSI-BLAST profiles as a direct input eliminates the lengthy process of extracting the sequences and producing an explicit multiple sequence alignment as a separate step. We thus chose sequence profiles produced by PSI-BLAST in multiple sequence prediction.

To evaluate these methods on multiple sequence prediction, we developed a new baseline method taking

into account of sequence profiles. At each position i , the probability of exposure P_i^E is weighted by the frequencies of residue substitutions f_{ij} from the sequence profiles.

P_j^E is the average probability of exposure for the j^{th} amino acid type. f_{ij} is a 20-D vector with the component for residue in position i as 1 and the others as 0 for using single sequence prediction and the compositions of 20 amino acids in position i for multiple sequence prediction.

$$P_i^E = \sum_{j=1}^{20} P_j^E \cdot f_{ij} \quad P_i^B = \sum_{j=1}^{20} P_j^B \cdot f_{ij}$$

An exposed state is predicted if the weighted probability of exposure is greater than that of burial. The performance of the baseline(73.4%) is about 4% improvement over the baseline for single sequence prediction. The result directly shows that sequence profiles improve the prediction markedly.

The results of multiple sequence prediction are listed in Table III. For BS, MLR and DT, the improvement over single sequence prediction is 4~5% but the improvement over baseline is only 2~3%. The results suggest that evolutionary information incorporated in the sequence profiles improve the prediction significantly, but the local sequence embeds very little information on solvent accessibility. NN and SVM achieve 79% accuracy, apparently outperform the other three methods.

Effect of protein size on prediction Simple geometry dictates that the smaller the protein, the lower the ratio of volume to surface area, it is tempting to speculate that the prediction may be protein size dependent. The conjecture was verified by two evidences. First, the chain length and prediction accuracy for NN has a negative correlation coefficient -0.093, which shows higher accuracies are associated with shorter chains although the relationship is relatively weak. Second, adding the chain length as an attribute improved the prediction of NN by 1.1%. We then divided the dataset into three subsets according to the chain length. Each subset was trained and tested separately and the results were compared to that of joint training using the whole dataset (Table III). The results show separating training did improve the performance of

different methods although the improvement is not significant. The only exception is MLR, which performs worse on set1 in separate training. The reason is probably that the set1 is not sufficiently large for MLR (the number of residues is only about half of the other two sets). This was further verified by testing MLR on a smaller dataset of 883 proteins (Shan *et al.*,2001). The performance of MLR on the three subsets all decreased while NN all improved in separate training. This should attribute to their different mechanisms of training. MLR method needs larger dataset to calculate the coefficients. In conclusion, when there is a sufficiently large dataset, grouping the data with similar properties (here the protein size) for training can improve the prediction. In addition, since the training data are reduced, the training process is also speeded.

Ensemble classification combining different methods
 Since improvement by any single method is difficult, an easier way to improve the prediction is to form consensus prediction from different methods. Here we used the weighted ensemble as following based on the accuracy of each classification method i , and its prediction confidence level C_i . For example, SVM performed the best, NN the 2nd, so we gave SVM the highest weight W_i , and NN the second, etc.

$$\sum_{i=1}^5 \pm W_i C_i \begin{cases} >0 & +1 \text{ (exposed)} \\ <0 & -1 \text{ (buried)} \end{cases}$$

We assigned exposed state as positive, and burial state as negative, so the final predicted state depends on the sign of weighted ensemble. As shown in Table III, the ensemble classification achieved higher accuracy than the best single method, but the improvement was not significant. However, the weighted ensemble classification is much better than the simple ‘majority win’ ensemble.

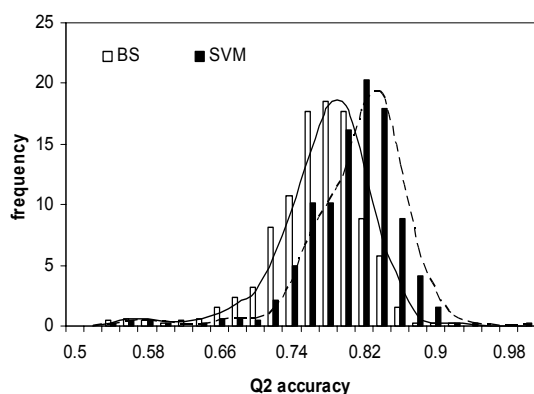


Figure I Distribution of accuracy for BS and SVM

Comparison of methods with statistical tests The level of success achieved by different methods varies very little. Statistical tests must be done to check whether one result is significantly better than another result. FigI showed the distribution of accuracies for SVM and BS of the 2148 protein chains. The distributions are approximately Gaussian with a mean of 0.796 and a standard deviation 0.052 for SVM and a mean of 0.757 and a standard deviation 0.055 for BS. A standard t-test assuming equal

variances was performed for any pair of methods. Using a threshold of P values $\alpha=0.01$, the tests accept the null hypothesis that there is no difference in the means between NN and SVM (P=0.15), between BS and MLR(P=0.02), between SVM and WE(P=0.10), and reject all the others. We then conclude that NN and SVM are both the best predictors. Their performances are significantly better than BS and MLR in predicting residue solvent accessibility. The performance of DT is better than BS and MLR but worse than NN and SVM. The weighted ensemble predictor didn't do much better than NN and SVM.

Blind test on targets from CASP5 The community wide experiment on the critical assessment of techniques for protein structure prediction (CASP) which has been run every two years since 1994 offers a means to evaluate prediction methods entirely blindly (<http://predictioncenter.llnl.gov>). Despite the number of targets is limited, the CASP experiment does offer an opportunity for methods to be fairly compared against each other. We applied the 5 predictors on solvent accessibility prediction on the targets from CASP5 to further test and compare these methods. Of a total of 53 targets whose structures have been solved so far, 21 targets had no significant sequence similarity to proteins of known structures. The test results (Table IV) on the 21 targets confirmed that NN and SVM achieved the best performance at 78~79%, DT achieved about 77%, and BS and MLR at about 76%.

Data set	No. seq	BS	MLR	DT	NN	SVM	WE
uniq	21	76.5	76.4	77.2	78.9	78.3	78.9
hom	31	78.1	78.5	79.5	81.2	81.6	82.1

Table IV. Results of blind test set from CASP5. uniq set: targets without significant sequence similarity to proteins of known structure. Homo set: targets homologous to proteins of known structure.

Currently homology modelling is still the most accurate method for predicting many aspects of protein structure. Thus, an analysis of the conservation of solvent accessibility within a family of homologous 3D structures gives us an upper limit for the accuracy of predicting solvent accessibility. Rost and Sander(1994) have shown that solvent accessibility is less conserved in 3D homologues than is secondary structure and hence is predicted less accurately from homology modelling. The accuracy for two-state solvent accessibility prediction by sequence alignment was 83.8%, by structural alignment was 84.8%. For comparison, we tested the methods on the rest of 32 targets which are homologous to proteins of known structure and the best predictor achieved 82% accuracy, which is very close to that of homology modelling.

5 Conclusion

We applied and compare different classification approaches to predict protein solvent accessibility. The methods use local sequence centered at the target residue. On single sequence prediction, DT, BS and MLR perform about the same. The improvement over the baseline model

that use only the identity of target residue is small. Local sequence seems embed very little information on accessibility. The problem is not with methods. On multiple sequence prediction, sequence profiles improve the prediction markedly. The best methods, NN and SVM, achieved the level of success of 79% with correlation coefficient of 0.59. Separate training according to protein size improves the prediction when there are sufficiently large dataset available. Ensemble classification based on the five methods didn't significantly improve the prediction over the best single method.

6 Acknowledgement

This work was supported in part by grant GM58187 from the National Institutes of Health (to H.-X. Z.).

7 References

- Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), *Nucleic Acids Res.* 25:3389-3402.
- Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* 1998; 276:517-525.
- Anfinsen C.B. Principles that govern the folding of protein chains. *Science* 181:223-230, 1973.
- Chan, H.S. and Dill, K.A.(1990) *Proc. Natl. Acad. Sci. USA*, 87, 6388-6392.
- Delamarche C, Guerdoux-Jamet P, Gras R, Nicolas J. *Biochimie* 81: 1065-1072 (1999).
- Ding CHQ, Dubchak I. *Bioinformatics* 2001; 17:349-358
- Ehrlich,L., Reczko, M., Bohr,H. and Wade,R.C. (1998) *Protein Engng*, 11, 11-19.
- Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. *Bioinformatics* 2000; 16:906-914.
- Holm L. and Sander C. (1996) Mapping the protein universe. *Science* 273:595-602
- Horton P, Nakai K. *Intelligent Systems in Molecular Biology* 5:147-152(1997).
- Hua S, Sun Z. *J Mol Biol* 2001;308:397-407
- Joachims T., *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- Jones D. Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol* (1999); 292:195-202.
- Kabsch W, Sander C. Dictionary of protein secondary structures: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* 22: 2577-2637 (1983).
- Kyte J.& Doolittle, R. F.(1982)*J. Mol. Biol.* 157, 105-132.
- Li X., Pan X-M. New methods for accurate prediction of solvent accessibility from protein sequence. *Proteins* 42:1-5 (2001)
- Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibility with information theory. *Proteins* 42: 452-459 (2001).
- Nelson D., Cox M. *Lehninger Principles of Biochemistry* (3rd ed.) Page 118
- Ooi T, Oobatake M, Nemethy G, Scheraga HA,(1987) Accessible surface areas as a measure of the thermodynamics parameters of hydration of peptides. *Proc. Natl. Acad. Sci USA* 84:3086-3090.
- Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. *Protein Eng* 12: 1051-1054 (1999).
- Rost B, and Sander C: Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20, 216-226, 1994
- Rost B., Sander C. Progress of 1D protein structure prediction at last. *Proteins* 23:295-300 (1995).
- Salzberg SL, Delcher AL, Fasman KH, Henderson J. A decision tree system for finding genes in DNA. *J. Comput. Biol.* 5: 667-680 (1998).
- Selbig J, Mevissen T, Lengauer T. Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics* 15: 1039-1046 (1999)
- Shan Y., Wang G, Zhou H. Fold recognition and accurate query-template alignment by a combination of Psi-Blast and threading. *Proteins* 42: 23-37 (2001).
- Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms: lysozyme and insulin. *J Mol Biol* 79: 351-371 (1973).
- Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins* 25: 38-47 (1996).
- Varadarajan R, Nagarajaram HA and Ramakrishnan C(1996) *Proc. Natl. Acad. Sci. USA* 93, 13908-13913.
- Yuan Z., Burrage K., Mattick J. Prediction of protein solvent accessibility using support vector machines. *Proteins* 48:566-570 (2002).