Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Training Set Reduction Methods for Protein Secondary Structure Prediction in Single-Sequence Condition

Zafer Aydin, Yucel Altunbasak
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250, USA
aydinz@ece.gatech.edu,
yucel@ece.gatech.edu

Isa Kemal Pakatci, Hakan Erdogan
Faculty of Natural Sciences and Engineering
Sabanci University
Tuzla Istanbul, 34956 Turkey
isakemal@su.sabanciuniv.edu,
haerdogan@sabanciuniv.edu

Abstract— **Orphan proteins are characterized by the lack of significant sequence similarity to database proteins. To infer the functional properties of the orphans, more elaborate techniques that utilize structural information are required. In this regard, the protein structure prediction gains considerable importance. Secondary structure prediction algorithms designed for orphan proteins (also known as single-sequence algorithms) cannot utilize multiple alignments or alignment profiles, which are derived from similar proteins. This is a limiting factor for the prediction accuracy. One way to improve the performance of a single-sequence algorithm is to perform re-training. In this approach, first, the models used by the algorithm are trained by a representative set of proteins and a secondary structure prediction is computed. Then, using a distance measure, the original training set is refined by removing proteins that are dissimilar to the given protein. This step is followed by the re-estimation of the model parameters and the prediction of the secondary structure. In this paper, we compare training set reduction methods that are used to re-train the hidden semi-Markov models employed by the IPSSP algorithm [1]. We found that the composition based reduction method has the highest performance compared to the alignment based and the Chou-Fasman based reduction methods. In addition, threshold-based reduction performed better than the reduction technique that selects the first 80% of the dataset proteins.**

I. INTRODUCTION

Prediction of protein function using amino acid sequences greatly accelerates experimental elucidation of protein function. The typical approach to predict the function of a protein is to compare its amino acid sequence to sequences of the database proteins. This is performed using such pairwise alignment algorithms as Smith-Waterman [2] and their efficient approximations (*e.g.,* BLAST [3] and FASTA [4]). If there is significant similarity to a known protein, the function can be estimated with a high degree of confidence. However, there are thousands of new proteins that are discovered at an unprecedented rate, in which the sequence based methods fail to predict their function because there are no database proteins even with weak sequence similarity. In addition, for most of these proteins, neither profile-based methods [5], [6] nor iterative search methods (PSIBLAST [7], SAM [8]) are applicable. For such orphan proteins, the function prediction cannot depend solely on sequence-level comparisons. An alternative procedure to extract functional information is to compare protein structures, for the function of a protein is mainly determined by its structural conformation. It is known that the protein structure is better conserved under mutation than the amino acid sequence. Even when the amino acid sequences diverge significantly, their structures are often conserved during evolution. For such scenarios, structure prediction algorithms can help to estimate the function.

Accurate prediction of the regular elements of protein 3D structure is important for precise prediction of the whole 3D structure. In secondary structure prediction, one is mainly concerned with the assignment of secondary structure elements (the α-helix {H}, the β-strand {E} and the loop {L}) to each amino acid residue as shown in Fig. 1.

G	K	C	.....	N	T	F	V	←	Amino Acid Sequence
|	|	|		|	|	|	|		
L	L	E	.....	H	H	H	H	←	Secondary Structure Labels

Fig. 1. Secondary Structure Prediction

There are two types of protein secondary structure prediction algorithms. A single-sequence algorithm does not use information about other (homologous) proteins. The algorithm should be suitable for a sequence with no similarity to any other protein sequence. Algorithms of another type are explicitly using sequences of homologous proteins, which often have similar structures. The prediction accuracy of such an algorithm should be higher than one of a single-sequence algorithm due to incorporation of additional evolutionary information from multiple alignments or alignment profiles [9]. Single-sequence methods are important because for orphan proteins, any method of secondary structure prediction performs as a single-sequence method. Hence, developing better prediction methods for single-sequence condition has a definite merit as it helps to improve the functional annotation of orphan proteins.

Secondary structure prediction methods often employ such machine learning tools as neural networks (NN), support vector machines (SVM), and hidden Markov models (HMM). There are essential steps in the development of a machine learning based predictor. The first step is called feature set selection, where the most informative correlations and patterns are identified. This allows us to choose a model that reflects the dependencies within structural elements. Feature set selection is followed by the training phase where the model parameters are estimated using proteins with

1-4244-0788-5/07/$20.00 ©2007 IEEE

known secondary structures. Finally, in the testing phase, the performance is evaluated by making predictions for new test samples. The content of the training set is important. It has been shown that the reduction methods, which refine the training set by excluding structurally dissimilar proteins improve the prediction accuracy. In this paper, we compare training set reduction methods that are used to re-train the hidden semi-Markov models employed by the state-of-the-art IPSSP algorithm [1].

II. Iterative Protein Secondary Structure Parse (IPSSP) Algorithm

Amino acid and DNA sequences have been successfully analyzed using hidden Markov models (HMM). For a comprehensive introduction to HMMs, see [10]. In a hidden semi-Markov model (HSMM), a transition from a hidden state into itself cannot occur, and a hidden state can emit a whole string of symbols rather than a single symbol. The hidden states of the model used in protein secondary structure prediction are the structural states {H, E, L} designating α-helix, β-strand and loop segments, respectively. Transitions between the states are characterized by a probability distribution. At each hidden state, an amino acid segment with uniform structure is generated according to a given length distribution, and the segment likelihood distribution.

The IPSSP algorithm utilizes three HSMMs and an iterative training procedure to refine the model parameters. The steps of the algorithm can be summarized as follows:

IPSSP Algorithm

1. For each HSMM, compute the posterior probability distribution that defines the probability of an amino acid to be in a particular secondary structure state. This is achieved by using the posterior decoding algorithm (also known as the forward-backward algorithm).
2. For each HSMM, compute a secondary structure prediction by selecting the secondary structure states that maximize the posterior probability distribution.
3. For each HSMM, reduce the original training set using a distance measure that compares the training set proteins to the predictions computed in step 2. Then, train each HSMM using the reduced dataset and compute secondary structure predictions as described in steps 1 and 2.
4. Repeat step 3 until convergence. At each iteration, start from the original dataset and perform reduction.
5. Take the average of the three posterior probability distributions and compute the final prediction as in step 2.

It has been observed that performing the dataset reduction step only once (*i.e.*, one iteration) generated satisfactory results [1].

III. Training Set Reduction Methods

In this section, we describe three dataset reduction methods that are used to refine the parameters of an HSMM: composition based reduction, alignment based reduction and reduction using Chou-Fasman parameters. In each method, the dataset reduction is based on a similarity (or a distance) measure. We considered two types of decision boundaries to classify proteins as similar or dissimilar. The first approach selects the first 80% of the proteins in the original dataset that are similar to the input protein. The second approach applies a threshold and selects proteins accordingly.

A. Composition Based Reduction

In this method, the distance between the predicted secondary structure and the secondary structure segmentation of a training set protein is computed as follows:

$$D = \max(|H_p - H_t|, |E_p - E_t|, |L_p - L_t|), \quad (1)$$

where H_p, E_p, and L_p denote the composition of α-helices, β-strands and loops in the predicted secondary structure, respectively. Similarly H_t, E_t, and L_t represent the composition of α-helices, β-strands and loops in the training set protein. Here, the composition is defined as the ratio of the number of secondary structure symbols in a given category to the length of the protein. For instance, H_p is equal to the number of α-helix predictions divided by the total number of amino acids in the input protein. This measure allows us to reduce the training set to proteins that belong to the same SCOP class [11]. Thus, for example, a prediction with high α-helix content is expected to generate a training set that contains proteins in all-α class. After, sorting the proteins in the training set, we considered two possible approaches to construct the reduced set: (1) selection of the first 80% of the proteins with the lowest D values; (2) selection of the proteins that satisfy $D < 0.35$ [1].

B. Alignment Based Reduction

In this method, first, pairwise alignments of the given protein to training set proteins are computed. Then proteins with low alignment scores are excluded from the training set. As in the composition based method, two approaches are considered to obtain the reduced dataset: (1) selection of the first 80% of the proteins with the highest alignment scores; (2) selection of the proteins with alignment scores above a threshold. Here, the threshold is computed by finding the alignment score that corresponds to the threshold used in the composition based reduction method. In the following sections, we will give more details on pairwise alignment settings.

1) Alignment Scenarios: We considered the following cases:

- Alignment of secondary structures (SS)
- Alignment of amino acid sequences (AA)
- Joint alignment of amino acid sequences and secondary structures (AA+SS)

In the first case, the aligned symbols are the secondary structure states, which take one of the three values: H, E, or L. In the second case, the symbols are the amino acids and finally, in the third case, the aligned symbols are the pairs of amino acid and secondary structure type.

[1]The threshold is found experimentally [1].

2) Score Function: The score of an alignment is computed by summing the scores of the aligned symbols (matches and mismatches) as well as the gapped regions. This is formulated as follows:

$$S = \sum_{k=1}^{r} (\alpha M_{aa}(a_k, b_k) + \beta M_{ss}(c_k, d_k)) + G \qquad (2)$$

where S is the alignment score, r is the total number of match/mismatch pairs, G is the total score of the gapped regions, a_k, b_k represent the k^{th} amino acid pair of the aligned proteins (the input and the training set protein, respectively), c_k, d_k denote the k^{th} secondary structure pair of the aligned proteins, $M_{aa}(.)$ is the amino acid similarity matrix, $M_{ss}(.)$ is the secondary structure similarity matrix, and finally, the parameters α, and β determine the weighted importance of the amino acid and secondary structure similarity scores, respectively. To compute possible alignment variations described in the previous section, α and β take the following values: (1) $\alpha = 0, \beta = 1$ to align secondary structures; (2) $\alpha = 1, \beta = 0$ to align amino acid sequences; (3) $\alpha = 1, \beta = 1$ to align amino acid and secondary structures in a joint manner.

3) Similarity Matrices: We used the BLOSUM30 table [12] as the amino acid similarity matrix and the Secondary Structure Similarity Matrix (SSSM) [13] shown in Table I.

TABLE I
SECONDARY STRUCTURE SIMILARITY MATRIX, WHICH IS USED TO SCORE THE SIMILARITY OF TWO SECONDARY STRUCTURE SYMBOLS.

M_{ss}	H	E	L
H	2	-15	-4
E	-15	4	-4
L	-4	-4	2

4) Gap Scoring: When a symbol in one sequence does not have any counterpart (or match) in the other sequence, then that symbol is aligned to a gap symbol '-'. Allowing gap regions in an alignment enables us to better represent the similarity between the aligned sequences in a biologically meaningful manner. In the state-of-the-art gap scoring, opening a gap is penalized more than extending it. For example, in the "affine gap scoring", which is one of the most widely used gap scoring techniques, starting a gap is scored by the parameter g_o, and extending a gap region is scored by g_e. In that case, the total gap score in (2) is computed as:

$$G = N_o g_o + N_e g_e, \qquad (3)$$

where N_o is the total number of gap openings, and N_e is the total number of gap extensions. In this work, we set the parameters g_o, and g_e to -12, and -2, respectively.

5) Optimum Alignment: Given a scoring function, the computation of the optimum (best scoring) alignment can be found using a dynamic programming approach. In this paper, we used the Smith-Waterman algorithm to compute the local alignment between a pair of proteins. Further details on the alignment algorithms and dynamic programming can be found in Durbin *et al.* [14].

6) Score Normalization: After computing the raw score of an alignment, it is useful to normalize it to a statistically meaningful range. In this paper, we normalized the alignment score by the average length of the aligned proteins. In that case, the normalized score is computed as $2\frac{rawscore}{l_1+l_2}$, where l_1, and l_2 are the lengths of the aligned proteins.

C. Reduction using Chou-Fasman parameters

In this method, the training set reduction is based on the Chou-Fasman distance measure, which is defined as:

$$D_{cf} = \sum_{k \in H,E,L} \left\{ \frac{1}{l_p} \sum_{j=1}^{l_p} f_k(q(j)) - \frac{1}{l_t} \sum_{j=1}^{l_t} f_k(h(j)) \right\}. \qquad (4)$$

Here, l_p is the length of the input protein, l_t is the length of the training set protein, $q(j)$ is the j^{th} amino acid of the input protein, $h(j)$ is the j^{th} amino acid of the training set protein, and $f_k(z)$ is the Chou-Fasman coefficient that reflects the propensity of the amino acid of type z to be in the secondary structure state k. These coefficients can be computed as described in [15]. In this formulation, the secondary structure information of the proteins is not used and each amino acid is allowed to take three possible secondary structure states. In a slightly modified version of this method, we defined the Chou-Fasman distance using the secondary structure information as follows:

$$D_{cf,2} = \left\{ \frac{1}{l_p} \sum_{j=1}^{l_p} f_{k(q(j))}(q(j)) - \frac{1}{l_t} \sum_{j=1}^{l_t} f_{k(h(j))}(h(j)) \right\}, \qquad (5)$$

where $k(q(j))$ is the predicted secondary structure state for the j^{th} amino acid of the input protein, and $k(h(j))$ is the secondary structure state for the j^{th} amino acid of the training set protein. In Chou-Fasman based reduction, we computed the reduced dataset by selecting the first 80% of the proteins with the lowest Chou-Fasman distances.

IV. RESULTS AND DISCUSSION

In our simulations, we used the EVA set of "sequence-unique" proteins [16] derived from the PDB database [17]. We removed sequences shorter than 30 amino acids and arrived to a set of 2720 proteins. To reduce eight secondary structure states used in the DSSP notation to three, we used the following conversion rule: H, G to H; E, B to E; I, S, T, ' ' to L. We used the PDB_SELECT dataset to compute the Chou-Fasman coefficients (*i.e.*, the function $f(.)$ in (4) and (5)) as in [15]. Here, the coefficients reflect the propensity of an amino acid to be either in H, E, or L state, which are defined using the above conversion rule.

We evaluated the performances of the methods by a leave-one-out cross validation experiment (jacknife procedure). At each step, a protein is chosen as the test example and is taken out from the dataset. The remaining proteins form the training set and are used to estimate the parameters of the hidden semi-Markov model (*i.e.*, transition, length and emission distributions). Since the true secondary structures were available, we used the maximum-likelihood estimation procedure, in which the observed frequencies for the desired

quantities are divided by a proper normalization factor to compute the probability values. After estimating the model parameters, we predicted the secondary structure sequence of the test protein and repeated the leave-one-out procedure until all the proteins in the test set are evaluated. To save computation time, we restricted our test data to the first 600 proteins in the dataset, which gave a good approximation to the true result. Then, we computed the performance measures by taking the true secondary structures of the proteins as reference. To evaluate the performance, we chose the three-state-per-residue accuracy (Q_3) as the overall sensitivity measure, which is computed as the total number of correctly predicted amino acids in all dataset proteins divided by the total number of amino acids in the dataset.

From the results shown in Tables II and III, the composition based reduction method performs better than the other reduction methods. This is mainly because of the fact that composition based reduction does not impose strong constraints, which serves to compensate for the errors made in the initial secondary structure prediction. In addition, threshold based reduction is slightly better than the reduction that selects the first 80% of the most similar proteins. Among the methods being compared, the composition based reduction method with thresholding gave the most accurate result, where the secondary structure prediction accuracy is improved by 0.6% compared to the condition with no re-training. Another advantage of the composition based method is its low computational complexity.

Comparing the alignment based reduction methods, the best result is obtained by the method that aligns secondary structures. Joint alignments of amino acid sequences and secondary structures did not perform better than secondary structure alignments. This is not surprising because in single-sequence condition the input protein is not statistically similar to dataset proteins at the amino acid level. Therefore, the discriminative power of the amino acid similarity matrix is weaker than the secondary structure similarity matrix.

TABLE II

SENSITIVITY MEASURES OF THE TRAINING SET REDUCTION METHODS. THE TOP 80% OF THE PROTEINS ARE CLASSIFIED AS SIMILAR TO THE INPUT PROTEIN.

Method	$Q_3(\%)$
Composition Based	67.01
Alignment Based (SS)	67.00
Alignment Based (AA+SS)	66.92
Alignment Based (AA)	66.69
Chou-Fasman Based (D_{cf})	66.65
No Re-training	66.59
Chou-Fasman Based ($D_{cf,2}$)	66.50

V. CONCLUSIONS

We showed that the training set reduction followed by the re-estimation of the model parameters improves the secondary structure prediction accuracy in single-sequence condition. Among the methods being compared, the composition based reduction technique with thresholding generated the most accurate results. This is mainly because of the fact that composition based reduction does not impose strong constraints, which serves to compensate for the errors made in the initial secondary structure prediction. As a future work, we are planning to optimize the threshold parameter used to construct the reduced dataset. In addition, the methods analyzed can be applied to the second class of prediction algorithms, which utilize evolutionary information in the form of alignment profiles or multiple alignments.

TABLE III

SENSITIVITY MEASURES OF THE TRAINING SET REDUCTION METHODS. THE DATASET PROTEINS ARE CLASSIFIED AS SIMILAR TO THE INPUT PROTEIN BY APPLYING A THRESHOLD.

Method	$Q_3(\%)$
Composition Based	67.17
Alignment Based (SS)	67.12
Alignment Based (AA+SS)	67.06

REFERENCES

[1] Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction for a single sequence using hidden semi-Markov models," *BMC Bioinformatics*, vol. 7, no. 178, 2006.

[2] T. Smith and M. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195–197.

[3] S. F. Altschul, W. Gish, W. Miller, E. Y. Myers, and D. J. Lipman, "A basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410.

[4] W. R. Pearson, "Rapid and sensitive sequence conparisons with FASTP and FASTA," *Methods in Enzymology*, vol. 183, pp. 63–98.

[5] M. Gribskov, A. McLachlan, and D. Eisenberg, "Profile analysis: Detection of distantly related proteins," *P.N.A.S., USA*, vol. 84, pp. 4355–4358.

[6] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Mol. Biol.*, vol. 235, pp. 1501–1531.

[7] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402.

[8] K. Karplus, C. Barrett, and R. Hugley, "Hidden Markov models for detecting remote homologies," *Bioinformatics*, vol. 14, pp. 846–856.

[9] D Frishman and P Argos, "Seventy-five percent accuracy in protein secondary structure prediction," *Proteins*, vol. 27, pp. 329–335, 1997.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," vol. 77, no. 2, pp. 257–286, 1989.

[11] A G Murzin, S E Brenner, T Hubbard, and C Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, pp. 536–540, 1995.

[12] S. Henikoff and J.G. Henikoff, "Amino acid substitution matrices from protein blocks," *P.N.A.S. USA*, vol. 89, pp. 10915–10919, 1992.

[13] A. Wallqvist, Y. Fukunushi, L. R. Murphy, A. Fadel, and R. M. Levy, "Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases," *Bioinformatics*, vol. 16, no. 11, pp. 988–1002, 2000.

[14] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1981.

[15] P. Chou and G. Fasman, "Empirical predictions of protein conformation," *Annu. Rev. Biochem.*, vol. 47, pp. 251–276, 1978.

[16] "Eva: secondary structure (intro):," http://cubic.bioc.columbia.edu/eva/doc/intro_sec.html.

[17] "The protein data bank," http://www.rcsb.org/pdb.