Proceedings of the 28th Annual Hawaii International Conference on System Sciences — 1995

Use of Multiple Alignments in Protein Secondary Structure Prediction

V. Di Francesco, P. J. Munson and J. Garnier*

Analytical Biostatistics Section, Laboratory of Structural Biology, Division of Computer Research Technology and *Fogarty International Center, National Institutes of Health, Bethesda, MD 20892-5626
valedf@helix.nih.gov, munson@helix.nih.gov, garnier@darwin.bu.edu

Abstract

Using a new database of 20 proteins not included in any of the previously used training datasets, we have incorporated multiple alignment information from homologous proteins into two well-characterized prediction methods: COMBINE (a jury method) and the Q-L (or quadratic-logistic) method. It is found that the increase in accuracy from the use of related proteins is similar for both methods (5.8% and 6.3%, respectively) yielding a per residue prediction accuracy (Q_3) of 68.7% and 69.0%, respectively, for a three state prediction. Most of the improvement came from consideration of averaging, profiling or consensus predictions. Of this improvement, a small amount (0.5%) came from recognition that "gap-permissive" positions in the alignment are most frequently in the coil state. Our finding is consistent with the hypothesis of a common secondary structure for the aligned family, and that improved accuracy is due to reduced noise in the prediction.

Introduction

The present databases of amino acid sequences contain a relatively large number of homologous sequences of proteins that are evolutionarily related and usually perform the same function. This means also that these homologous proteins share the same overall three dimensional structure. When one of the members of this protein family has a known structure, we may apply the principle of the protein modeling by homology [1]. If no structure is known for that family, some structural data can nevertheless be extracted from the amino acid sequences, concerning the secondary structure as made of α-helix, (H), β-strand or extended structure (E) and non periodic structure, coil (C). When a single sequence is used, the accuracy of the best existing automated methods is 63-65% of correctly predicted residues in the three states, H, E and C [see reviews in 2-3]. However if the sequences of several members of the same family are known, it has been shown that secondary structure predictions can be significantly improved [4-6]. These enhancements in accuracy rely on multiple alignments of the homologous sequences which currently must have at least 25% of identical residues [5], a value corresponding to the threshold used for sequences longer than 80 residues exhibiting similar secondary structures [7]. We do not imply that lower percentages of identity between sequences might not also correspond to similar secondary and tertiary structures, but rather that the present alignment methods are not able to distinguish such proteins from others having a very different fold. Furthermore the quality of alignments between such distant sequences degrades with evolutionary distance. The treatment of the multiple alignment of sequences differs with the authors [4-6], but generally, these methods account for the individual prediction of secondary structure for each aligned sequence and at each position in the alignment. This is done through a profile [6] or through a consensus prediction [4, 5]. As the improvement brought to the prediction by the consensus prediction is very significant for the GOR or SIMPA algorithms [4, 5], we wanted to address the question whether more advanced and accurate prediction algorithms, such as COMBINE [8] or the quadratic-logistic algorithm [9], can also benefit from the multiple alignment and how this compares with the profile algorithm used by others and data already obtained. For this we used a data base of 20 proteins of high X-ray resolution, not bearing significant homology to any of the proteins used in developing the parameters for these different algorithms.

We also wished to investigate the role of gaps in homologous sequence alignment in possibly augmenting prediction accuracy. Others have shown that presence of gaps in the alignment may be associated with coil propensity, and it is reasonable to assume that the exposed, loop regions of proteins are more permissive of mutational insertions and deletions, and these loop regions commonly assume random coil conformation [10].

Table I. Database of 20 proteins having less than 25% pairwise sequence identity, taken from the Brookhaven Data Bank, April 1993 release.

PDB Name	Protein Name	Sequence Length	Resolution (Å)	Number of Homologous Sequences
1ace	acetylcholinesterase	535	2.8	9
1col	colicin A	204	2.4	2
1cox	cholesterol oxydase	502	1.8	3
1dfn	defensinHNP-3	30	1.9	5
1f3g	phosphocarrier	150	2.1	7
1gly	glucoamylase	470	2.2	6
1gmf	colony-stimulating factor	119	2.4	3
1hdd	engrailed homeodomain	60	2.8	14
1hrh	h domain of HIV-1 reverse transcriptase	136	2.4	11
1msb	mannose binding protein, lectin domain	115	2.3	10
1nsb	neuraminidase	390	2.2	13
1pi2	serine proteinase inhibitor	63	2.5	16
2cpk	cAMP-dependent protein kinase	350	2.7	16
2hip	high potential iron sulfur protein	71	2.5	3
2ifb	intestinal fatty acid binding protein	131	1.96	11
2pk4	human plasminogen kringle 4	80	2.25	23
2sar	endoribonuclease sa	96	1.8	2
2scp	sarcoplasmic calcium binding protein	174	2.0	2
3fgf	fibroblast growth factor	146	1.77	13
5p21	ras P21 protein	166	1.35	15
Total		3988		

Methods

Database

We used a dataset which consists of 20 proteins (Table I) from the Brookhaven Protein Data Bank [11], release April 1993, whose structures have recently been determined with high resolution (2.8 Å or less, 0.148< R factor<0.230) and bear less than 25% identity with any of the proteins used in calibrating the prediction methods. The total number of residues is 3988. We will refer to these 20 sequences as the 'test' sequences.

The observed secondary structure was taken from the *hssp* database [7], which uses the *dssp* algorithm [12]. These sequences were cleaned by removing several spurious crystallographic homodimers, and one fusion protein. The *dssp* secondary structure assignment was kept for those residues with helix conformation (H) and extended conformation (E). If 3 residues in a row having G conformation were adjacent to the N or C terminus of a helix segment, they were converted to H; any 4 residues in a row having G conformation were converted to H. The remaining residues were assigned coil conformation (C). Residues not seen in the electron density map were assigned a coil structure. The secondary structure composition of the test sequences is the following: 29.4% residues in helical conformation, 20.8% in extended conformation and 49.7% in coil conformation. The homologous sequences were selected from the homologues in the *hssp* database to obtain a nearly uniform spread of the percent of identity from the test sequences thereby avoiding including a large number of nearly identical sequences. Homologous sequences were taken from SWISS-PROT or GenBank. When only a small region of the whole sequence was homologous to the test sequence, we kept this region and disregarded the rest of the sequence. The multiple alignments were done with CLUSTAL [13] using default settings, running on a DEC Alpha Model 500 workstation. We used a homology threshold that depends on the length of the test sequence, as suggested by Schneider and Sander [7].

Table II. Prediction accuracy, Q3, for individual proteins

Protein	Sequence Length	COMBINE	Consensus	Consensus + gaps *	Quadratic Logistic	Quadratic Logistic+ profiles	Quadratic Logistic+ profiles+ gaps*
1ace	535	63.36	72.34	72.90	62.62	71.40	71.96
1col	204	65.69	66.18	66.18	71.57	67.65	69.61
1cox	502	58.17	62.35	62.55	57.57	62.75	62.75
1dfn	30	66.67	66.67	66.67	46.67	63.33	63.33
1f3g	150	50.67	56.00	56.00	58.67	66.67	66.00
1gly	470	64.89	71.28	73.40	64.68	71.06	71.70
1gmf	119	72.27	78.99	78.99	74.79	74.79	74.79
1hdd	60	65.00	83.33	83.33	60.00	70.00	70.00
1hrh	136	62.50	63.24	62.50	64.71	66.18	65.44
1msb	115	65.22	73.04	73.04	72.17	78.26	79.13
1nsb	390	53.08	64.62	64.87	55.13	65.90	66.67
1pi2	63	74.60	77.78	77.78	80.95	80.95	80.95
2cpk	350	65.14	72.57	72.86	63.71	78.00	76.57
2hip	71	63.38	60.56	63.38	50.70	59.15	59.15
2ifb	131	41.22	45.80	46.56	46.56	51.15	53.44
2pk4	80	86.25	93.75	93.75	85.00	92.50	92.50
2sar	96	75.00	77.08	77.08	70.83	75.00	76.04
2scp	174	64.94	66.09	65.52	52.30	57.47	58.62
3fgf	146	62.33	66.44	67.12	58.90	60.27	63.01
5p21	166	68.67	69.88	69.88	69.28	68.07	66.87
Average per chain		64.45	69.40	69.72	63.34	69.03	69.43
Standard Deviation		9.42	10.28	10.13	10.63	9.36	9.00
Average per residue		62.46	68.28	68.73	62.34	68.61	68.98

* Uses coil probability of 0.5 when gaps are present, see *Methods.*

Predictive methods

1. COMBINE is an expert system [8], which associates the predictions from three methods: the pattern of hydrophobic residues [8], GOR III using directional residue pair informations [14] and SIMPA a prediction algorithm based on peptide similarity [15]. The consensus prediction [5] uses the predictions derived from homologous sequences at each aligned position. A voting (majority) rule is applied to obtain the final predicted state: the one with the highest probability index from the COMBINE confidence scales. A probability index P_{xi} for conformation x at position i is obtained by normalizing the probability values of the confidence scale to sum to 1.
2. The quadratic-logistic method, Q-L, [9] uses a maximum-likelihood logistic regression which incorporates both the conventional "linear" effects of residues contained in a 17 residue window with the "quadratic" effect of all residue pairs within the same window. The inherent overparameterization of quadratic models is effectively dealt with using penalized likelihood techniques, parameter selection and imposition of reasonable constraints on the quadratic parameters. In particular, the periodic information inherent in the alpha-helix (period 3.6) and beta sheet (period 2.0) is encoded in the quadratic model. The quadratic parameters have been shown to significantly augment the predictive power of the linear, "directional" coefficients. The advantages of the Q-L method are that it uses statistically optimal simultaneous calibration of all parameters through the maximum likelihood step, it is extremely versatile compared with linear models, and there is a well-developed theory for the characteristics of logistic models in the context of categorical variables. The quadratic logistic method produces true probability estimates for each of three structural states. Extending the Q-L model to

Table III. Number of predicted residues by various methods

Observed State	Predicted State H	E	C	Predicted State H	E	C	Total
	COMBINE			Quadratic-Logistic			
H	803	82	289	678	144	352	1174
E	234	246	351	175	306	350	831
C	395	146	1442	295	186	1502	1983
Total	1432	474	2082	1148	636	2204	
	COMBINE+consensus			Quadratic-Logistic+profiles			
H	840	70	264	731	113	330	1174
E	177	255	399	133	332	366	831
C	267	88	1628	194	116	1673	1983
Total	1284	413	2291	1058	561	2369	
	COMBINE+consensus+gaps*			Quadratic-Logistic+profiles+gaps*			
H	827	66	281	716	102	356	1174
E	174	252	405	130	329	372	831
C	245	76	1662	170	107	1706	1983
Total	1246	394	2348	1016	538	2434	

* Uses coil probability of 0.5 when gaps are present, see *Methods*.

incorporate aligned sequences can be done in one of two ways. One can use the aligned sequences by combining the predictions for each homologue as an average or consensus. Alternatively, the aligned sequences may first be combined into a sequence of profiles describing the proportion of each residue observed at each position. Since the original Q-L model treats each position of the original sequence as a vector of 20 zero-one or dummy variables (one for each amino acid), one can easily replace this with the vector of proportions **p** ($0<p_i<1$, i=1,...,20) of the residues in the observed profile. Thus, observation of alanine in the original sequence would be represented as the index vector (1,0,0,...,0) whereas if one alanine and one cysteine are observed at the same position in the alignment, the profile is represented as (0.5,0.5,0,...,0), assuming an alphabetical ordering of the one letter amino acid codes. Details of the profile Q-L method are described elsewhere (manuscript in preparation). Although both consensus and profile approaches are possible with the Q-L method, we cite only the results obtained with the profile approach, since early investigations (not shown) showed virtually no difference between the two approaches.

Use of gaps in homologous sequences

Based on the observation that helical or extended residues (mainly protein cores) are more conserved than coil regions [10, 16] and assuming that such conservation is reflected in the multiple alignments, we wanted to incorporate this idea into our prediction schemes with homologous sequences. In the case of the consensus prediction, any gap position in the aligned sequences was predicted as coil, with a confidence scale probability equal to 0.5. In the case of the Q-L algorithm with profiles, the estimated probability for the coil state was set to 0.5. We determined empirically the value 0.5 using a database of proteins described previously [6].

Results and Discussion

Results for individual proteins are presented in Table II and global results for six different predictive schemes are presented in Tables III-V in the form of matrices for observed and predicted percentage and number of residues for the three conformations, H, E and C. In Table II, one

Table IV. Predicted structural states as percentages of observed states by various methods*

Observed State	Predicted State H	E	C	Predicted State H	E	C	Total
	COMBINE			Quadratic-Logistic			
H	68	7	25	58	12	30	100
E	28	30	42	21	37	42	100
C	20	7	73	15	9	76	100
	COMBINE+consensus			Quadratic-Logistic+profiles			
H	72	6	22	62	10	28	100
E	21	31	48	16	40	44	100
C	13	4	82	10	6	84	100
	COMBINE+consensus+gaps^			Quadratic-Logistic+profiles+gaps^			
H	70	6	24	61	9	30	100
E	21	30	49	16	40	45	100
C	12	4	84	9	5	86	100

*Table III divided by the total number of residues in the three observed conformations. The numbers on the diagonal of each matrix correspond to the sensitivity index (correct/observed).

^ Uses coil probability of 0.5 when gaps are present, see *Methods.*

can observe that both COMBINE and Q-L methods give almost identical Q_3 accuracy (percent of correctly predicted residues in H, E and C per total residues). More interesting is that they both benefit by about the same amount, 5.8% and 6.3% respectively, from the consensus or profile prediction. As expected Combine and Q-L performed better than GOR III and SIMPA, (Q3: 62.5% and 62.4% respectively). For GOR III and SIMPA the total average improvement with consensus with a different data set was 7.6% [5]. This improvement is slightly higher than that found here for COMBINE and Q-L.

Various attempts to improve the consensus algorithm were not very successful, however a 0.5% (COMBINE) or 0.4% (Q-L) improvement was obtained (Table II) by assigning the coil probability to 0.5, wherever a deletion in the multiple alignment was seen. For the same proteins our predictive accuracies are somewhat lower than those obtained by Rost *et al.* published in two different papers [6, 17] . However there are several significant differences between prediction accuracies in those two reports. For instance 2scp is predicted with 52% accuracy in [6] and 65% in [17]. These differences make the comparison with our results difficult. Also, we do not use the same set of homologous sequences as Rost *et al.* [6, 17]. One can see that in Tables I-II, the 8 proteins having 2 to 6 homologous proteins showed an average improvement of 4.4% for COMBINE (4.9% for Q-L) whereas those having 7 to 23 homologous sequences (12 proteins) showed an average improvement of 7.6% for COMBINE and consensus (7.9% for Q-L) on a residue basis; per chain the difference of improvements is similar. The quality of the alignment has also been shown to be important [5] and possibly alignments given by *hssp* [7] could be better than the ones we obtained from CLUSTAL for a given set of proteins. In our case, results of prediction accuracy with *hssp* alignments were practically identical to those obtained with CLUSTAL (data not shown). A point of interest in this study was the comparison of two different methods of prediction, the quadratic-logistic approach [9] and a jury method COMBINE [8]. In earlier studies, these methods were shown to have similar crossvalidated Q_3 values; this fact is confirmed in the present study (Table II).

The Q-L method tends to underpredict helices but its helix predictions have more specificity than COMBINE (Tables III and V). Both methods overpredict the coil conformation and underpredict the extended conformation, but COMBINE has a slightly higher specificity or probability index (correct/predicted) for the extended and coil conformation (Table V). Interestingly these trends

Table V. Observed structural states as percentages of predicted states by various methods*

Observed State	Predicted State H	E	C	Predicted State H	E	C
	COMBINE			Quadratic-Logistic		
H	56	17	14	59	23	16
E	16	52	17	15	48	16
C	28	31	69	26	29	68
total	100	100	100	100	100	100
	COMBINE+consensus			Quadratic-Logistic+profiles		
H	65	17	12	69	20	14
E	14	62	17	13	59	15
C	21	21	71	18	21	71
total	100	100	100	100	100	100
	COMBINE+consensus+ gaps^			Quadratic-Logistic+profiles+ gaps^		
H	66	17	12	70	19	15
E	14	64	17	13	61	15
C	20	19	71	17	20	70
total	100	100	100	100	100	100

*Table III divided by the total number of residues in each predicted conformation. The numbers on the diagonal of the matrices correspond to the probability index (correct/predicted) or specificity index.
^ Uses coil probability of 0.5 when gaps are present, see *Methods*.

persist in consensus and profile predictions with or without gap probability adjustment. Both methods benefit equally from the consensus or profile algorithms. The improvements come mainly in the prediction of coil (Table IV) which becomes more sensitive, whereas the specificity improves for helix and extended conformations (Table V). The difficulty of using the sensitivity index (correct/observed) is seen for coil improvement: the increase in sensitivity does not come from a more efficient prediction (the probability index does not vary much) but because more residues are predicted in coil, 2082 to 2348 for COMBINE and 2204 to 2434 for the Q-L method. The imbalance and underprediction of the extended conformation is one of the principal source of inaccuracy which is not solved here.

One remarkable aspect of these studies is the quality of improvement obtained by averaging the predicted conformations at each position of the alignment. A rationale for this behavior has already been proposed (18). The multiple alignments should position residues as they are in the three dimensional structure, consequently having the same secondary structure and largely the same contact residues or force field effect due to the tertiary structure. By averaging the predicted secondary structures at each position in the multiple alignment on all the homologous sequences, one should improve the prediction just as when one measures a physical property from several experiments and takes the average of individual measures. The mean of the predictions will have a value with less noise than each single prediction and should be more accurate. If this is a correct interpretation, then one can predict that the more sequences are available (and to some extent, the more distantly related they are), the more accurate the average prediction will be.

Acknowledgment

We want to thank and acknowledge the contribution of Dr. Ljubomir Buturovic for writing the consensus program for COMBINE. We also acknowledge the contribution of Dr. Raul Porrelli for his work on the original Q-L program.

References

1. Browne N.J., North A.C.T., Philips D.C., Brew K., Vanaman T.C. and Hill R.L. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol., 42, 65-86. 1969.

2. Garnier J. and Levin J.M. The protein structure code: what is its present status? CABIOS, 7, 133-142, 1991.

3. Rost B., Sander C. and Schneider R. Progress in protein structure prediction? TIBS, 18, 120-123, 1993.

4. Zvelebil M.J., Barton G.J., Taylor W.R. and Stenberg M.J.E. Prediction of protein secondary structure and active sites using alignment of homologous sequences. J.Mol. Biol., 195, 957-961. 1987.

5. Levin J.M., Pascarella S., Argos P. and Garnier J. Quantification of secondary structure prediction improvement using multiple alignments. Prot. Eng. 6, 849-854, 1993.

6. Rost B. and Sander C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232, 584-599, 1993.

7. Schneider R. and Sander C. Database of homology-derived structures and the structurally meaning of sequence alignment. Proteins: Struct. Funct. Genet., 9, 56-68, 1991.

8. Biou V., Gibrat J. F., Levin J. M., Robson B. and Garnier J. Secondary structure prediction: combination of three different methods. Prot. Eng., 2, 185-191. 1988.

9. Munson P.J., Di Francesco V. and Porrelli R. Protein secondary structure prediction using periodic-quadratic-logistic models: statistical and theoretical issues. Proc. 27th Annual Hawaii Int. Conf. on System Sciences, vol V, 375-384, 1994.

10. Chothia C. and Lesk A.M. The relation between the divergence of sequences and structure in proteins, EMBO J., 5, 823-826, 1986.

11. Bernstein F.C., Koetzle T.F., Williams G.J., Meyer E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T. and Tasumi M. The protein data bank: a computer based archival file for macromolecular structures. J. Mol. Biol., 112, 535-542, 1977.

12. Kabsch W. and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers, 22, 2577-2637, 1983.

13. Higgins D.G. and Sharp P.M. Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS, 5, 151-153, 1989.

14. Gibrat J.F., Garnier J. and Robson B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol., 198, 425-443, 1987.

15. Levin J.M. and Garnier J. Improvements in a secondary structure prediction method based on search for local sequence homologies and its use as a model building tool. Biochim. Biophys. Acta, 955, 283-295, 1988.

16. Greer J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins: Struct. Funct. Genet., 7, 317-34, 1990.

17. Rost B., Sander C. and Schneider R. Evolution and neural networks- Protein secondary structure prediction above 71% accuracy. Proc. 27th Annual Hawaii Int. Conf. on System Sciences, vol V, 385-394, 1994.

18. Levin J.M., Pascarella S., Argos P. and Garnier J. Quantification of secondary structure prediction improvement using distantly related proteins. Protein Structure by distance analysis. Bohr H. and Brunak S. ed., IOS Press, Amsterdam, 302-314. 1994.