Review Letter

How good are predictions of protein secondary structure?

Wolfgang Kabsch and Christian Sander

Department of Biophysics, Max Planck Institute of Medical Research, Jahnstr. 29, 6900 Heidelberg, FRG

Received 17 February 1983

The three most widely used methods for the prediction of protein secondary structure from the amino acid sequence are tested on 62 proteins of known structure using a program package and data collection not previously available. None of these methods predicts better than 56% of the residues correctly, for a three state model (helix, sheet and loop). The algorithms of Robson et al. [J. Mol. Biol. (1978) 120, 97–120] and Lim [J. Mol. Biol. (1974) 88, 873–894] are the best of those tested. New methods, now under development, can be tested against this benchmark.

Protein structure *Secondary structure prediction* *Amino acid sequence*

1. INTRODUCTION

The explosive increase in our knowledge of DNA sequences (currently at about 1 Megabase) has led to increased use of protein secondary structure predictions from the amino acid sequence. Typically, one wants to know what structural type of protein the DNA codes for and whether the protein is related to one of known function or structure. The most interesting practical use has been the prediction of antigenic oligopeptides as potential vaccines [3,4]. For any of these uses it is important to know how well secondary structure prediction methods work.

Assessment of available prediction methods is best made by comparing predictions with the crystallographically determined structure. Such comparisons have been made [5], but have been hindered by two facts:

(i) There are ambiguities in two of the best known methods, those of Chou [6] and Lim [2], in that they often give different results in the hands of different people and are therefore not programmable without extension or modification;

(ii) There is considerable variation in the definitions of secondary structure given by crystallographers.

We have now solved both of these difficulties and report the results of a completely objective, up-to-date assessment of the most widely used prediction methods on 62 proteins with more than 10000 residues. For a three-state definition of secondary structure (helix, sheet, loop/turn) the overall prediction accuracy for new protein structures does not exceed 56% for the best of these methods and is only 50% for the most widely used (Chou) method. We caution against the over-interpretation of predictions made by presently available methods and provide a benchmark against which new methods, now under development, can be tested.

2. METHODS

Ambiguities in the method of Chou [6] were overcome by selecting possible secondary structure segments such that the sum of preference parameters over *all* chosen segments is maximal; technically, this is a difficult optimization problem

Published by Elsevier Science Publishers B.V.
00145793/83/$3.00

but was achieved by a recursive algorithm which was added to a program written by C. Oefner [7]. Turn prediction, done separately by Chou [8], was not included. Conceivably the overall success of Chou's method can be improved by rules for eliminating overlaps of predicted turns with predicted helix/sheet residues. Ambiguities in the method of Lim [2] were overcome by a simplified iterative procedure for segment selection which was added to a program written by J.A. Lenstra [9]. The (unambiguous) method of Robson was used as programmed by the authors [1].

Known methods not compared here include: Nagano [10] (bad beta prediction in our hands); Maxfield and Scheraga [11] (similar to Robson's, reportedly 57% accurate for five states); Ptitsyn and Finkelstein [12] (new version just published [13]); Palau and Argos [14] (reportedly 56% accurate for four states).

Objective and accurate assignment of secondary structure was achieved by a pattern recognition algorithm [15] which extracts hydrogen-bonded features from the full atomic coordinates as deposited with the Protein Data Bank [16].

3. RESULTS AND DISCUSSION

Predictive success is given in table 1 for each protein and each method as the percentage of residues predicted correctly in a three state description of secondary structure. The result of the comparison is similar to that of Busetta and Hospital [17] who have 47% success for Chou and 57% success for Robson on 34 proteins. The method of Lim has a surprising 65% success rate for protein structures known in 1974 when his method was published, but this drops to 56% for proteins elucidated after 1974. The difference can be understood to be due to special rules tailored to particular proteins in Lim's method.

Structure predictions can be evaluated in more

Table 1

Predictive success of the three most widely used secondary structure prediction methods: details for each protein, averaged over the three structure states

(a) 24 protein structures pre-1974					
fraction correct (%) Chou [6]	Robson [1]	Lim [2]	IDEN	RES	protein structure NAME
54	59	72	1CPV	108	CALCIUM-BINDING PARVALBUMIN B
60	55	51	2B5C	85	CYTOCHROME B5 (OXIDIZED)
43	54	56	1CYT	103	CYTOCHROME C (OXIDIZED).
43	43	70	1C2C	112	CYTOCHROME C2 (FERRI)
35	54	50	1FDX	54	FERREDOXIN (PEPTOCOCCUS AEROGENES)
65	57	69	2RXN	54	RUBREDOXIN (OXIDIZED, FE(III))
40	44	71	11NS	51	INSULIN (A AND B CHAIN)
58	63	64	7LYZ	129	LYSOZYME (HEN EGG WHITE, TRICLINIC)
58	56	68	1SNS	142	STAPHYLOCOCCAL NUCLEASE (COMPLEX)
57	66	66	1RNS	124	RIBONUCLEASE-S
47	61	69	1CPA	308	CARBOXYPEPTIDASE A
50	52	59	2TLN	316	THERMOLYSIN
53	54	74	2GCH	236	GAMMA CHYMOTRYPSIN A
57	53	66	1PTN	223	BETA-TRYPSIN (NATIVE AT PH 3)
54	57	67	1SBT	275	SUBTILISIN BPN'
55	59	58	1EST	240	TOSYL-ELASTASE
53	55	70	8PAP	212	PAPAIN
59	62	71	3CNA	237	CONCANAVALIN A
46	56	51	4LDH	329	LACTATE DEHYDROGENASE, APO ENZYME M4
51	72	71	1MBN	153	MYOGLOBIN (FERRIC IRON - METMYOGLOBIN)
30	54	63	1ECD	136	HEMOGLOBIN (ERYTHROCRUORIN DEOXY)
47	57	62	2MHB	287	HEMOGLOBIN (HORSE, AQUO MET)
37	55	72	1LHB	148	HEMOGLOBIN(MET) (SEA LAMPREY)
71	74	72	3PTI	58	TRYPSIN INHIBITOR
51	57	65		4120	SUBTOTAL PRE-1974
(b) 33 protein structures post-1974					
fraction correct (%) Chou [6]	Robson [1]	Lim [2]	IDEN	RES	protein structure NAME
51	53	49	1ABP	306	L-ARABINOSE-BINDING PROTEIN
53	46	64	1HIP	85	OXIDIZED HIGH POTENTIAL IRON PROTEIN
46	69	51	1bbb	103	CYTOCHROME b562 (E. COLI, OXIDIZED)
55	44	66	1bbc	104	CYTOCHROME c550
49	66	61	2SIC	82	CYTOCHROME c551 (OXIDIZED)
36	49	72	1FXC	96	FERREDOXIN (SPIROLINA PLAT. 59°C)
60	59	61	3FXN	133	FLAVODOXIN (DESULF.)
59	47	60	1AZU	125	AZURIN
56	52	57	1PLC	99	PLASTOCYANIN
56	44	59	1FPT	36	AVIAN PANCREATIC POLYPEPTIDE
45	41	59	1GCN	29	GLUCAGON (PH 6-7)
50	47	50	1BP2	123	PHOSPHOLIPASE A2
32	56	52	1LZM	164	LYSOZYME (BACTERIOPHAGE T4)
54	59	62	1APR	324	ACID PROTEASE (RHIZOPUS NIVEUS)
46	53	60	1APP	323	ACID PROTEINASE (PENICILLOPEPSIN)
48	52	52	1ALP	198	ALPHA LYTIC PROTEASE
56	60	52	1CGA	181	PROTEINASE A FROM STREPTOMYCES GRISEUS
51	56	51	2ACT	218	ACTINIDIN
54	60	60	1FAB	428	LAMBDA IMMUNOGLOBULIN FAB
57	54	60	1REI	107	BENCE-JONES IMMUNOGLOBULIN
44	55	61	3PGM	230	PHOSPHOGLYCERATE MUTASE (DE-FREE ENZ)
50	62	59	1TIM	246	TRIOSE PHOSPHATE ISOMERASE
50	53	61	1CAC	256	CARBONIC ANHYDRASE FORM C
46	51	50	1DFR	162	DIHYDROFOLATE REDUCTASE (COMPLEX)
47	55	58	1GPD	333	D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
39	44	52	4ADH	374	APO-LIVER ALCOHOL DEHYDROGENASE
45	49	48	2GRS	451	GLUTATHIONE REDUCTASE
56	72	64	2SOD	151	CU,ZN SUPEROXIDE DISMUTASE
52	69	50	1HBL	152	LEGHEMOGLOBIN (YELLOW LUPIN)
37	33	44	1CRN	46	CRAMBIN
48	54	77	1OVO	56	OVOMUCOID THIRD DOMAIN
51	63	54	2SSI	107	STREPTOMYCES SUBTILISIN INHIBITOR
68	65	69	1CTX	71	ALPHA CONOTOXIN
42	42	46	1MLT	26	MELITTIN
50	61	60	1NXB	62	NEUROTOXIN B
52	73	65	2ADK	194	ADENYLATE KINASE
55	54	54	1PHD	293	RHODANESE
46	42	40	2PAB	114	PREALBUMIN (HUMAN PLASMA)
49	55*	56*		6636	SUBTOTAL POST-1974
50	56	59		10756	TOTAL FOR 57 PROTEINS

All methods are compared according to how well they predict the three states α-helix, β-sheet and loop (everything else) or older (a) and newer (b) protein structures. Fraction correct is the number of residues predicted correctly in any state divided by the total number of residues. The protein name is preceded by the protein data bank [16] identifier IDEN and the number of residues RES. The * indicates the percentage of correctly predicted residues one can expect in applying the methods of Robson and Lim to newly determined sequences

detail by calculation of assorted quality indices [5] which indicate how well a *particular* state is predicted, whether there is over- or underprediction etc. All of these indices can be calculated from the predicted/observed matrix in table 2 which indicates, say, how many of the 2295 observed helical residues are correctly predicted as helical (H) and how many are wrongly predicted as loop/turn (L) or sheet (E, for extended); or, how many of the 2684 residues predicted as helical by Lim are sheet, loop or helical in the crystallographic structure. One such quality index for each state is the 'fraction correct of observed' in table 2. For example, we see that 74% of the observed loop residues are correctly predicted by Lim, while only 36% of the observed sheet residues are correct; this imbalance is related to an overall underprediction (1690/2295) of sheet and an overprediction (6389/5421) of loop residues in Lim's method.

Table 2

Predictive success of the three most widely used secondary structure prediction methods: details of sheet (E), loop/turn (L) and helix (H) prediction averaged over all proteins

	Chou [6]			Robson [1]			Lim [2]		
observed/predicted matrix (number of residues)									
	predicted			predicted			predicted		
	E	L	H	E	L	H	E	L	H
observed E	1195	883	417	1244	689	362	820	1151	324
observed L	1361	2900	1160	1161	3057	1203	576	4027	818
observed H	894	878	1275	570	761	1715	294	1211	1542
observed/predicted total (number of residues)									
	E	L	H	E	L	H	E	L	H
observed	2495	5421	3047	2295	5421	3047	2295	5421	3047
predicted	3450	4861	2654	2975	4507	3281	1690	6389	2684
fraction correct (%)									
	E	L	H	E	L	H	E	L	H	
of observed	52	53	42	54	56	56	36	74	51	(a)
of predicted	35	65	45	42	68	52	49	63	57	(b)
rms error in predicting secondary structure content (residues per 100 residues)									
	E	L	H	E	L	H	E	L	H	
	17	14	16	12	16	16	13	14	13	

(a) Number of residues correctly predicted in state S divided by number of residues observed in state S = percentage of correct predictions when state S is observed

(b) Number of residues correctly predicted in state S divided by number of residues predicted in state S = percent probability of correct prediction when state S is predicted. The latter 'probability of correct prediction', *PC*(S), is practically useful in predicting unknown secondary structure

Suppose you have predicted a residue as helical and want to know the chances of being right. For a particular method, the average 'fraction correct of predicted' (table 2) defined as:

$$PC(\mathrm{S}) = \frac{N\ (\text{correctly predicted in state S})}{N\ (\text{predicted in state S})} \times 100$$

is a direct measure of the *probability of correct prediction* having predicted a residue to be in state S. Curiously, *PC*(S) does not appear among the quality indices commonly used [5], but is perhaps the most useful in prediction practice (after all, in a truly unknown protein structure no reference can be made to observed states). For example, when Lim's method predicts a sheet strand, we can estimate from table 2 that there is a 49% chance of correct prediction. Note the high probability of correct loop prediction of 63–68% which is related to the high fraction (50%) of observed loop residues.

Suppose you do not care about the details of secondary structure assignments but merely want to use a secondary structure prediction method to predict the helix/sheet *content* of a protein; for example, for comparison with spectroscopic determinations (such as circular dichroism). The root-mean-square average difference between predicted and observed secondary structure content for the 62 proteins is 12–17 residues/100 residues (table 2). For example, a prediction by Robson of sheet content has a typical uncertainty of ± 12%. An uncertainty of this size renders present comparisons of predicted secondary structure content with circular dichroism experiments useless in all but extreme cases.

We conclude that one may expect a success rate, for three states, of about 50% with Chou's method and of 55–56% with either Robson's or Lim's method. In any event, an error rate of 44% is unacceptable for many purposes and newly developing methods must do better. We estimate that empirical–statistical prediction of secondary structure alone may eventually reach 70% accuracy for three states; higher accuracy will, in our opinion, only come with a protein-folding theory aiming at prediction of the complete three-dimensional structure.

ACKNOWLEDGEMENTS

A detailed comparison for 9 proteins is reported in C. Oefner's thesis [7]. We thank G.E. Schulz for his active role in organizing the implementation of published prediction methods in our laboratory and the Deutsche Forschungsgemeinschaft for financial support to the project 'Protein Structure Theory'. The automated program package or predictions based on it are available on a collaborative basis.

REFERENCES

[1] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[2] Lim, V.I. (1974) J. Mol. Biol. 88, 873–894.
[3] Mueller, G.M., Shapira, M. and Arnon, J. (1981) Proc. Natl. Acad. Sci. USA 79, 569–573.
[4] Pfaff, E., Mussgay, M., Boehm, H.O., Schulz, G.E. and Schaller, H. (1982) EMBO J. 1, 869–874.
[5] Schulz, G.E. and Schirmer, R.H. (1979) Principles of Protein Structure, Ch.6, Springer-Verlag, New York.
[6] Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45–148.
[7] Oefner, C. (1982) Thesis (Diplomarbeit), University of Heidelberg.
[8] Chou, P.Y. and Fasman, G.D. (1979) Biophys. J. 25, 367–383.
[9] Lenstra, J.A. (1977) Biochim. Biophys. Acta 491, 333–338.
[10] Nagano, K. (1977) J. Mol. Biol. 109, 251–274.
[11] Maxfield, F.R. and Scheraga, H.A. (1979) Biochemistry 18, 697–704.
[12] Finkelstein, A.V. and Ptitsyn, O.B. (1971) J. Mol. Biol. 62, 613–624.
[13] Ptitsyn, O.B. and Finkelstein, A.V. (1983) Biopolymers 22, 15–25.
[14] Palau, J., Argos, P. and Puigdomenech, P. (1982) Int. J. Pept. Prot. Res. 19, 394–401.
[15] Kabsch, W. and Sander, C. (1983) Biopolymers, in press.
[16] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[17] Busetta, B. and Hospital, M. (1982) Biochim. Biophys. Acta 701, 111–118.