Vol. 180, No. 2, 1991
October 31, 1991

BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
Pages 666-672

ON α-HELICES TERMINATED BY GLYCINE
2. RECOGNITION BY SEQUENCE PATTERNS

Peer Bork[a] and Robert Preißner[b]

[a] Zentralinstitut für Molekularbiologie, Biomathematik, Robert-Rössle-Str.10, O-1115 Berlin-Buch, Germany

[b] Institut für Kristallographie, Freie Universität Berlin, Takustr. 6, D-1000 Berlin 33, Germany

Received August 26, 1991

Summary: Consensus sequence patterns were constructed to describe helix ends with a characteristic conformation caused by specific three-center hydrogen bonds. This special type of hydrogen bond pattern comprises about one third of all helices and mostly contains glycine with a positive torsion angle ϕ at the helix ends . After a simple clustering procedure 6 resulting consensus sequence patterns were able to identify 501 out of 575 helix ends in the Brookhaven Protein Data Bank, showing the above-mentioned features. The patterns did not detect any false segment, but numerous sequence segments not identified by structural criteria were recognized. It is likely that they are indeed helices terminated by glycine with a positive torsion angle ϕ.

The aim of *ab initio* prediction of protein tertiary structure from the sequence arose many years ago when the first X-ray structure became available (1). At present thousands of sequences and about a hundred different, relatively highly resolved tertiary structures of proteins are known. In spite of ingenious new approaches the prediction of secondary structure remains a difficult task (2-4). Reasons are, for example, the difficulties in including long range interactions, the insufficient number of known tertiary structures of high resolution to obtain sharper rules for predictions or, more generally, the still fragmentary knowledge of energetical, evolutionary as well as folding constraints. Therefore approaches for autonomous structural motifs seem to have more success than general prediction methods (5-7). In all these cases pattern recognition methods were used (for reviews see (8, 9)) which weight important positions and which are able to describe correlated features. Since specific sequence signals also seem to exist for local spatial motifs such as glycosylation and phosphorylation sites (10) or *cis*-prolyl

Abbreviations:
positive ϕ, positive torsion angle between backbone atoms N and Cα;
PDB, Brookhaven Protein Data Bank (Release 52).

0006-291X/91 $1.50

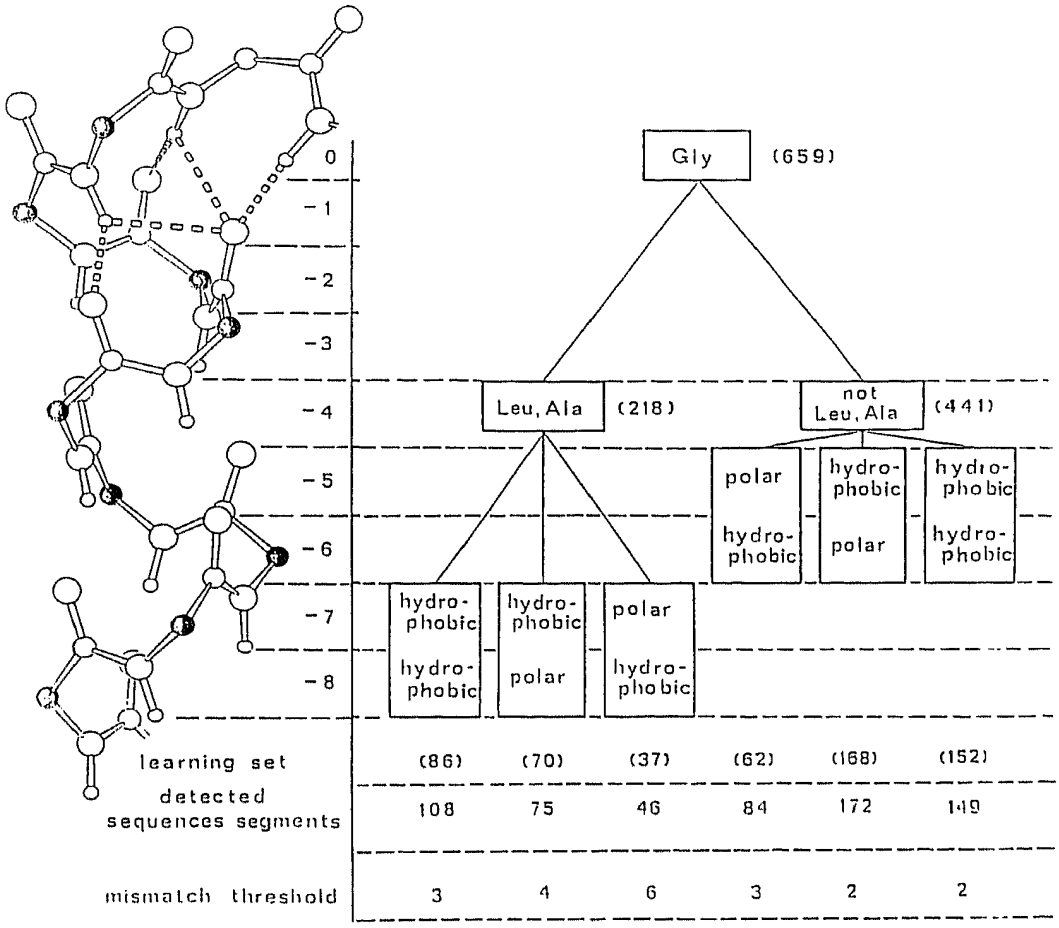


Fig.1. Cluster of derived consensus patterns.
A schematic representation of an α-helix backbone with C-terminal H-bond patterns is given to indicate the position dependent sequence signals used for the clustering procedure. Cα-atoms are highlighted. The number of sequence segments which were found by structural criteria (learning set) is shown for each subtree. The combination of two polar amino acids in position -7 and -8 or -5 and -6 respectively is neglected which leads to a small difference between the number of all considered helix ends and those of the resultant patterns. 87% of the sequence segments included in the learning set could be redetected by the patterns. Therefore, the row of detected sequence segments mirrors a remarkable number of additional findings even if the abolute number is similar to those of the learning set. For each pattern a mismatch theshold is given. Up to this theshold all findings are correct and no 'false positive' (i.e. sequence segments that are not helix ends with a positive torsion angle ϕ) were recognized.

residues (11) there should be a good chance of detailed description of helix ends by sequence patterns. About one third of all α-helices are terminated by glycine with a positive torsion angle ϕ (positive ϕ) forming specific three-center hydrogen bonds (12-16). Here we present our studies on sequence consensus patterns for this common helix end conformation (fig.1).

Materials and Methods

The successive steps of pattern construction are outlined in a flow chart (see fig.2a). A comprehensive learning set for our sequence pattern description was derived from the whole Brookhaven Protein Data Bank (PDB) (17). The resulting

redundancy does not affect the construction of the sequence patterns. In a first step we extracted a library of sequence segments from the PDB corresponding to helix ends stabilized by putative three-center hydrogen bonds (18). Using DSSP (19) we screened for glycine with a positive ϕ at helix ends (for details see (16)).

Since a single glycine is not sufficient for establishing helix ends we started a simple clustering procedure by successive recording of correlated sequence features. The accumulation of leucine and alanine in the position -4 (glycine=0; fig.1) was found to be the best correlated sequence feature. Therefore, this position was selected for definition of subpatterns. For the sequence segments of the branch containing leucine or alanine in position -4 (fig.1) a peak of hydrophobic amino acids in positions -7 and -8 was recorded, which is in agreement with the hydrophobic moment of α-helices (20) resulting from characteristic polarity patterns (21). Thus, the positions -7, -8 were included in the clustering procedure (fig.1).

In the sequence segments belonging to the other subtree (without leucine, alanine in position -4) the best peaks were obtained at position -5 and -6, where in addition to the cumulation of certain hydrophobic amino acids a clear correlation with accessibility could be found. This points to a shift in helix orientation relative to the solvent. Therefore these positions were used to derive subpatterns (fig.1).

Although there exist several complex pattern recognition methods as well as different amino acid similarity matrices (22) such as that of Dayhoff (23), we chose a very simple description of the resulting 6 patterns based on the observed frequencies of amino acids in each position (fig.2b). In each of the 13 positions included (9 N-terminal and 3 C-terminal counted from the glycine at the helix end) all amino acids occurring more than once were initially not penalized (position 0 in fig.1). When an amino acid was observed only once the penalty 1 was introduced to weight possible 'outliers' at each position. To amino acids, which do not appear in a position the penalty of 2 was assigned. The importance of positions essential for the clustering procedure was emphazised by penalties of 4 (see fig.2c). This weighting procedure is empirically based on experience with property patterns (24). To use our pattern search programs (e.g. (25)) and to consider the differences between the sequence parts of the PDB and usual sequence databases (only 31% of the corresponding sequences are identical (26)) we converted the sequences given in the PDB in a similar way as described in (27).

Results and Discussion

The sequences of all tertiary structures deposited in the PDB were screened for matching against the 6 constructed sequence patterns (fig.1c). For each pattern

---

Fig.2. One examplary pattern.

a) The flow chart outlines the successive steps of pattern construction.

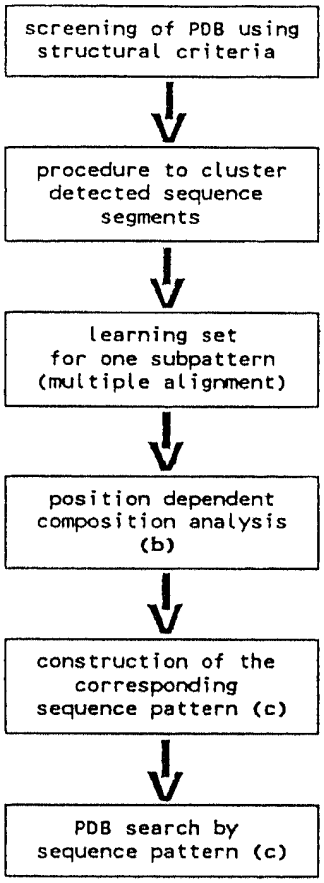
b) Position dependent numbers of occurrence of amino acids (one letter code in the left column) at position -9 to +3 counted from glycine. These numbers are derived from a learning set of 29 nonidentical helices terminated by a glycine with positive torsion angle ϕ.

c)Corresponding sequence pattern according to the stated rules (see Materials and Methods).

d) Alignment of nonhomologous sequence segments recognized by the pattern in the PDB. Protein names, PDB codes (incl. chains) and the sequence positions are given. All segments of the learning set could be redetected by the pattern with the exception of a region in crambin (1CRN). In addition a segment of xylose isomerase (3XIA) was indicated. This structure has a resolution of 2.5 Å and the predicted glycine is located five residues after a short helix adopting a positive torsion angle ϕ. Interestingly, a distant related structure (4XIA) has in this region a longer helix terminated by the corresponding glycine with a positive ϕ. Because of amino acid substitutions the corresponding segment of 4XIA is recognized by another pattern.

a maximum mismatch number is shown. Below these thresholds no 'false positive' was detected, i.e. all of the recognized sequence segments are indeed helix ends with the demanded features. In a reduced database of 79 well-refined different protein structures (16,28) about 83% of the helix ends as characterized by

a



b

A	3	0	13	7	4	19	4	6	3	0	8	3	2
C	0	0	0	0	0	0	0	0	0	0	1	0	2
D	3	3	0	1	0	0	2	1	0	0	1	2	1
E	3	1	0	4	0	0	1	5	3	0	1	3	1
F	0	0	3	4	0	0	0	0	1	0	2	0	1
G	1	8	0	0	0	0	0	0	1	28	1	1	4
H	1	0	0	0	1	0	0	0	3	0	0	1	1
I	0	0	4	1	2	0	3	0	0	0	2	3	1
K	1	8	0	1	3	0	4	9	1	0	0	1	3
L	1	0	4	0	3	9	4	0	3	0	1	1	0
M	1	0	0	1	0	0	0	1	3	0	1	0	1
N	3	2	0	0	1	0	0	0	1	0	1	0	0
P	2	0	0	0	0	0	0	0	0	0	1	3	3
Q	3	2	0	3	1	0	1	1	2	0	1	1	1
R	1	0	0	3	3	0	0	1	0	0	1	1	1
S	2	1	0	1	0	0	1	1	1	0	0	0	2
T	3	3	0	2	1	0	4	1	3	0	2	7	1
V	0	0	3	0	4	0	3	1	2	0	2	0	3
W	0	0	0	0	2	0	0	0	0	0	0	0	0
Y	0	0	1	0	2	0	1	1	1	0	1	1	0

c

A	0	4	0	0	0	0	0	0	0	9	0	0	0
C	2	4	3	2	1	4	2	2	2	9	0	2	0
D	0	0	4	1	2	4	0	1	2	9	1	0	1
E	1	1	4	0	2	4	1	0	0	9	1	0	0
F	2	4	0	0	2	4	2	2	1	9	0	2	1
G	1	0	4	2	2	4	2	2	1	0	0	0	0
H	1	3	4	2	1	4	2	2	0	9	2	1	1
I	2	4	0	1	0	4	0	2	2	9	0	0	1
K	1	0	4	1	0	4	0	0	1	9	2	1	0
L	1	4	0	2	0	0	0	2	0	9	1	1	2
M	1	4	3	1	2	4	2	1	0	9	1	2	1
N	0	0	4	2	1	4	2	2	1	9	1	2	2
P	0	4	4	2	2	4	2	2	2	9	1	0	0
Q	0	0	4	0	1	4	1	1	0	9	1	1	1
R	1	1	4	0	0	4	2	1	2	9	1	1	1
S	0	1	4	1	2	4	1	0	1	9	2	2	0
T	0	0	4	2	1	4	0	1	0	9	0	0	1
V	2	4	0	2	0	4	0	1	0	9	0	2	0
W	2	4	3	2	0	4	2	2	2	9	2	2	2
Y	2	4	1	2	0	4	1	1	1	9	1	1	2

d

subtilisin	1cse-1	107	S	G	I	E	W	A	T	T	N	G	M	D	V
glycolate oxidase	1gox-1	90	A	T	A	R	A	A	S	A	A	G	T	I	M
glycolate oxidase	1gox-2	38	E	D	A	R	L	A	V	Q	H	G	A	A	G
glycolate oxidase	1gox-2	94	T	D	V	F	K	A	L	A	L	G	A	A	G
phosphofructokinase	1pfk-1	107	M	G	A	M	R	L	T	E	M	G	F	P	C
phosphofructokinase	1pfk-1	176	L	T	L	A	A	A	I	A	G	G	C	E	F
434 repressor	1r69	15	N	Q	A	E	L	A	Q	K	V	G	T	T	Q
thermitase	1tec-1	116	N	G	I	T	Y	A	A	D	Q	G	A	K	V
triose-P isomerase	1tim-1	109	Q	K	V	A	H	A	L	A	E	G	L	G	V
tryptophan synthase	1wsy-3	26	E	Q	V	S	A	A	V	R	A	G	A	A	G
tryptophan synthase	1wsy-4	203	T	K	A	Q	I	L	D	K	E	G	R	L	P
cytochrom c	2ccy-1	92	T	K	L	A	A	A	A	K	A	G	P	D	A
citrate synthase	2cts	89	E	G	L	F	W	L	L	V	T	G	Q	I	P
citrate synthase	2cts	151	S	N	F	A	R	A	Y	A	E	G	I	H	R
lactate dehydrogenase	2ldx	235	G	G	Y	E	V	L	D	M	K	G	Y	T	S
cytochrom c2	3c2c	73	P	K	A	F	V	L	E	K	S	G	D	P	K
glutathion reductase	3grs	16	A	S	A	R	R	A	A	E	L	G	A	R	A
glutathion reductase	3grs	427	Q	G	F	A	V	A	V	K	M	G	A	T	K
phosphofructokinase	3pfk	83	K	G	I	E	Q	L	K	K	H	G	I	Q	G
phosphofructokinase	3pfk	106	Q	G	A	K	K	L	T	E	H	G	F	P	C
phosphoglycerate kinase	3pgk	371	D	T	A	T	V	A	K	K	Y	G	V	T	D
phosphoglycerate mutase	3pgm	95	D	K	A	Q	T	L	K	K	F	G	E	E	K
thermolysin	3tln	237	N	K	A	A	Y	L	I	S	Q	G	G	T	H
D-xylose isomerase	3xia	36	D	T	V	Q	R	L	A	G	L	G	A	H	G
malate dehydrogenase	4mdh-1	161	A	K	A	Q	I	A	L	K	L	G	V	T	S
D-xylose isomerase	4xia-1	118	H	N	I	D	L	A	A	E	M	G	A	E	T
alcohol dehydrogenase	8adh	226	D	K	F	A	K	A	K	E	V	G	A	T	E
catalase	8cat-1	259	R	D	L	F	N	A	I	A	T	G	N	Y	P

structural criteria (16) could be recognized by comparison with the 6 consensus sequence patterns. The rate increased to 87% for the whole PDB (501 out of 575 helix ends). The main body of the 74 helix ends which were not redetected contain proline within the helix. The penalties for proline had to be higher than average because of its helix breaking features. Many of the few prolines which occur in helices (15, 29) were neglected in this way, but the patterns could be sharpened. A second group of helix ends not redetected are those which contain 'outliers' in different positions.

In addition to the correctly identified helix ends we have recognized 131 segments not included in the learning set, but found to be helix ends with the demanded features as suggested by inspection of the respective structures. Often, the required glycine with a positive ϕ was located three residues after helix ends (mainly in structures with resolution lower than 2.0 Å). In other cases the number of helical residues was below 6, so that these segments did not comply with the criteria for inclusion in the learning set (see fig.3). Sometimes segments as detected were not confirmed by analysis in DSSP (19), but in better resolved structures of homologous proteins the respective regions were found to be indeed helix ends with a positive ϕ. A few selected examples are aligned in fig.3. The corresponding superposition (fig.4) is not as good as for helices of this type in highly resolved protein structures (see accompanied paper (16)). Nevertheless, the similarities in helix termination are evident (see also ϕ, ψ in fig.3). They become closer for helix ends detected by one subpattern (fig.5). The r.m.s. deviations between the super-

Prot[1]	n[2]	AA[3]	ϕ	ψ	AA	ϕ	ψ	AA	ϕ	ψ	AA	ϕ	ψ	AA	ϕ	ψ	AA	ϕ	ψ
			Acc[4]	Var[5]		Acc	Var		Acc	Var		Acc	Var		Acc	Var		Acc	Var
1fx1	30	Q	-77	-28	L	-74	-57	A	-54	-41	N	-64	-18	A	-91	3	G	100	15	Y	-72	-176
			52	35		2	29		44	48		120	40		47	43		71	23		31	56
5adh	311	M	-66	-16	L	-65	-33	L	-90	-28	L	-63	-33	S	-60	-42	G	151	4	R	-67	153
			127	15		34	54		70	8		149	8		63	46		36	27		34	5
6adh	236	K	-58	-12	A	-91	-68	K	-50	-34	E	-68	-24	V	-91	-13	G	126	26	A	-100	101
			44	39		0	17		117	26		131	32		20	13		11	0		4	30
2cts	44	D	-64	-42	M	-54	-26	M	-92	-17	Y	-113	2	G	-136	33	G	78	27	M	62	44
			127	17		39	50		75	45		123	0		29	0		53	0		80	0

Fig.3. Multiple alignment of detected sequence segments predicted to form specific three-center hydrogen bonds at the C-terminus. These regions were not included in the learning set. Flavodoxin (1FX1) has a resolution of 2.0 Å, alcohol dehydrogenases (5ADH), (6ADH) 2.9 Å and citrate synthase (1CTS) 2.7 Å.

[1] The proteins are given with their Brookhaven Protein Data Bank code (17).

[2] Position of the terminal glycine.

[3] Amino acids are shown in one letter code. X denotes an arbitrary residue.

[4] Solvent accessibility was calculated by DSSP (19).

[5] Position dependent variability derived from an aligned set of homologous sequences in HSSP (30). The lower the value the better is the conservation of the respective residue in related sequences.

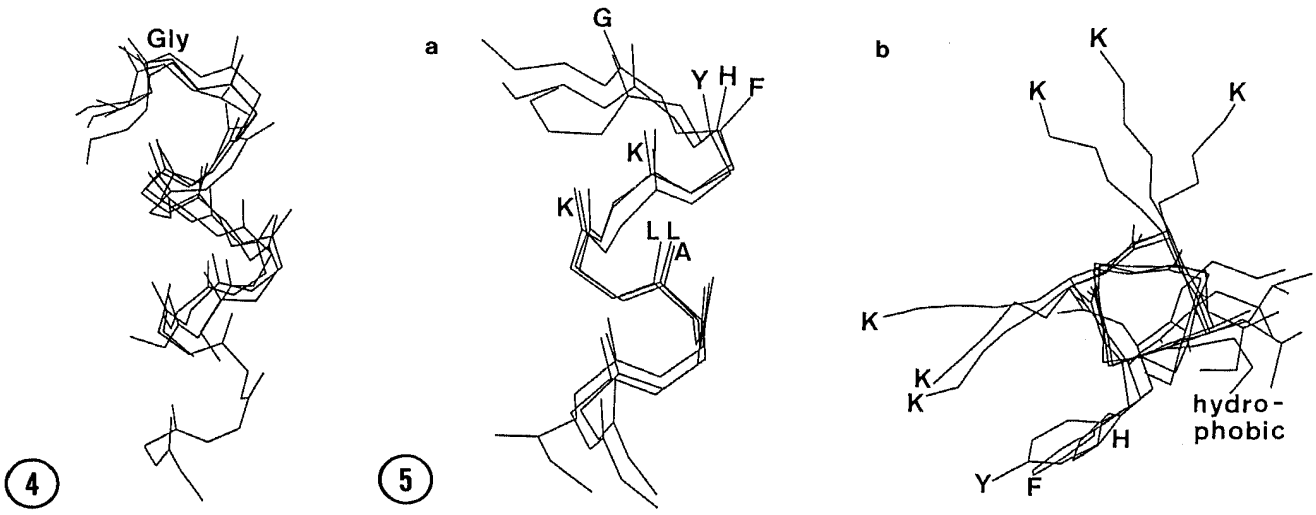


Fig.4. Structural alignment of the four segments displayed in fig.3. Despite the low resolutions the similar geometry in helix termination is obvious.

Fig.5. Superposition of three selected helices detected by the pattern shown in fig.2.
a) Side view of the following helices (PDB code and number of the glycine): 3PFK 93, 3PGK 381, 3PGM 105. Only backbone atoms are drawn. At the carbonyl-groups the last amino acids of the helices are labeled with one letter code.
b) Top view along the helix-axis on the three superimposed helices. The side chains of the above-mentioned residues are drawn and marked with one letter code. It is shown that glycine is located at the hydrophobic face of the helix.

imposed backbone atoms (fig.5a) of otherwise topologically unequivalent proteins are small. The sequence similarity is expressed in structural coincidence of the corresponding side chains (fig.5b).

Common signals of formation or disruption of structural elements such as helices seem to exist in a broad variety of topologically and functionally different proteins. Resulting sequence patterns allow tertiary structure prediction of local motifs. The examination of further structural elements from this point of view is in progress.

The set of patterns and the surrounding software running on VMS-compatible systems are available from the authors on request.

Acknowledgments

The authors are grateful to J. Reich, W. Saenger and C. Sander for helpful suggestions and critical reading of the manuscript. We thank E. Wolf for technical assistance. The work was partly supported by BMFT.

References

1. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Philips, D.C. (1958) Nature 181, 662-666.
2. Kabsch, W. and Sander, C. (1983) FEBS Lett. 155, 179-182.
3. Nishikawa, K. (1983) Biochim. Biophys. Acta 748, 285-299.

4. Garnier, J. (1990) Biochimie 72, 513-524.

5. Taylor, W.R. and Thornton, J.M. (1984) J. Mol. Biol. 173, 487-514.

6. Cohen, F.F., Arbanel, R.M., Kuntz, I.D. and Fletterick, R.J. (1986) Biochemistry 25, 266-275.

7. Rooman, M. and Wodack, S. (1988) Nature 335, 45-49.

8. Taylor, W.R. (1988) Prot. Eng. 2, 77-86.

9. Barton, G.J. and Sternberg, M.J.E. (1990) J. Mol. Biol. 212, 389-402.

10. Nakai, K. and Kanehisa, M. (1988) J. Biochem. 104, 693-699.

11. Frömmel, C. and Preißner, R. (1990) FEBS Lett. 277, 159-163.

12. Schellman, C. (1980) in: Protein Folding (Jaenicke, R. ed.) Elsevier/North-Holland Biomedical Press, 53-61.

13. Presta, L.G. and Rose, G.D. (1988) Science 240, 1632-1641.

14. Richardson and Richardson (1989) in: Prediction of Protein Structure and Principles of Protein Conformation, (Fasman, G. ed.) Plenum, New York, 1-98.

15. Richardson and Richardson (1988) Science 240, 1648-1652.

16. Preißner, R. and Bork, P. (1991) BBRC 180, 660–665.

17. 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.

18. Preißner, R., Egner, U. and Saenger, W. (1991) FEBS Lett. submitted.

19. Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577-2637.

20. Eisenberg D., Weiss, R.M. and Terwillinger, T.C. (1982) Nature 299, 371-374.

21. Schiffer, M. and Edmundson, A.B. (1967) Biophys. J. 7, 121-135.

22. Risler, J.L., Delorme, M.O., Delacroix, H. and Henaut, A. (1988) J. Mol. Biol. 204, 1019-1029.

23. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) in: Atlas of protein sequence and structure Vol.5, Natl. Biomed. Res. Found., Washington, D.C., 345-352.

24. Bork, P. and Grunwald, C. (1990) Eur. J. Biochem. 191, 347-358.

25. Bork, P. and Rohde, K. (1990) Biochem. Biophys. Res. Comm. 171, 1319-1325.

26. Lesk, A.M., Boswell, D.R., Lesk, V.I., Lesk, V.E. and Bairoch, A. (1989) Prot. Seq. Data Anal. 2, 295-308.

27. Pattabiraman, N., Namoodiri, K., Lowrey, A. and Gaber, B.P. (1990) Prot. Seq. Data Anal. 3, 387-405.

28. Rooman, M.J. and Wodak, S.J. (1991) Proteins 9, 69-78.

29. Woolfson, D.N. and Wiliams, D.H. (1990) FEBS Letters 277, 185-188.

30. Sander, C. and Schneider, R. (1991) Proteins 9, 56-68.