Volume 277, number 1,2, 159-163 FEBS 09280 December 1990

Prediction of prolyl residues in *cis*-conformation in protein structures on the basis of the amino acid sequence

Cornelius Frömmel[1] and Robert Preissner[2]

[1]*Free University Berlin, Inst. f. Kristallographie, Takustr. 6, W-1000 Berlin 33, Germany* and [2]*Charité, Humboldt University, Inst. Biochem., Hessesche Str. 4–6, Berlin, O-1040, Germany*

Received 8 October 1990

In proteins most peptide bonds are in *trans*-conformation: the torsion angle $\omega = 180°$. Only few show *cis*-conformation in known protein structures ($\omega = 0°$). Most of them are prolyl residues. About 6% of about 4000 prolyl residues are in *cis*-conformation. Between *trans*- and *cis*-prolyl residues significant differences are observed in the surrounding sequences. E.g. there are large amounts of aromatic residues N-terminally in case of *cis*-prolyl residues, but in the case of *trans*-prolyl residues more aromatic amino acids occur C-terminally. But in all cases there are only complex patterns which are indicative of *cis*- and *trans*-conformation, respectively. Considering the neighbours (± 6 residues) of prolyl residues and their physicochemical properties we find 6 different patterns which allow one to assign correctly about 75% of known *cis*-structured prolyl residues, whereby no false positive one is predicted.

Pattern research; Protein structure; Prediction; *Cis*-peptide bond; Proline

1. INTRODUCTION

The importance of sequence templates for improvement of three-dimensional structure prediction of proteins is reflected by a growing number of publications in this field [1]. Sequence patterns were used to find nucleotide binding sites, zinc fingers, leucine zippers, calcium binding sites and other structural motifs, for overview see [2]. The question arises whether it is possible to predict other spatial motifs on the basis of sequence patterns. One interesting local structure is the *cis*-conformation of the peptide bond. In initial steps to model protein structures in X-ray crystal structure analysis and in protein design, often only two torsional angles - ϕ (rotation about N-Cα) and ψ (rotation about Cα-C(O)) - are considered as degrees of freedom in the main chain conformation. The third torsional angle - ω (rotation about N-C(O) of the peptide bond) - of the backbone in the peptide chain is not treated explicitly as variable. In Fig. 1 the difference in the trace of the peptide chain resulting from *cis*- or *trans*-conformation of one prolyl-peptide unit is shown. I should be remembered, that the Cα-Cα distance in *cis*-conformations is nearly 1 Å shorter than *trans*, 2.9 Å versus 3.8 Å. In prolyl units in the peptide chain, the thermodynamic equilibrium between *cis*- and *trans*-conformation is about 1:20 to 1:5 [3], [4]. This corresponds to an energy difference of about 2–5 kcal/mol in favor of the *trans*-conformation [5]. The *cis–trans* isomerisation of prolyl residues is quite slow. The energy barrier for the transition between *cis*- and *trans*-conformation for prolyl residues is about 13 kcal/mol [5]. This slow reaction is catalyzed by a class of enzymes, the *cis–trans*-prolyl isomerases [6,7,8]. The *cis–trans* isomerization of prolyl residues determines the rate-limiting steps of protein folding as shown by design of mutants without prolyl residues and refolding experiments [9]. Relations may exist between sequence patterns surrounding prolyl residues and the recognition mechanism of the *cis–trans*-isomerase, which might be identical or similar to cyclophyllin, linked to the immunosuppressive action of cyclosporine A (a *cis*-peptide ring) [7,8]. Also the influence of local sequence patterns on the equilibrium between *cis*- and *trans*-form was demonstrated by NMR-experiments [4,9]. In consequence there are several good reasons to search for sequence patterns to predict *cis*-prolyl residues merely on the basis of the primary structure of proteins. Examining the frequencies of occurrence of *cis*-prolyl residues [3] reveals no clear relationship between amino acid type imide bonded to proline. So we extended the analysis on 6 amino acids flanking the prolyl residue.

2. MATERIALS AND METHODS

The protein structures were taken from the Protein Data Bank (Brookhaven, NY, USA) [10]. The definition of the torsional angles in the peptide chain is given in Fig. 2. The ω-angles of the protein structures were calculated using the following atoms: Cα(i), C′(i), N(i), Cα(i + 1). O(i) was not used because of its weak electron density in X-ray structure analysis. For a prolyl residue *i* the frequency of flanking amino acids was recorded in positions

Correspondence address: R. Preissner, Charité, Humboldt University, Dept. of Research, Schumannstr. 20–21, Berlin, O-1040, Germany

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

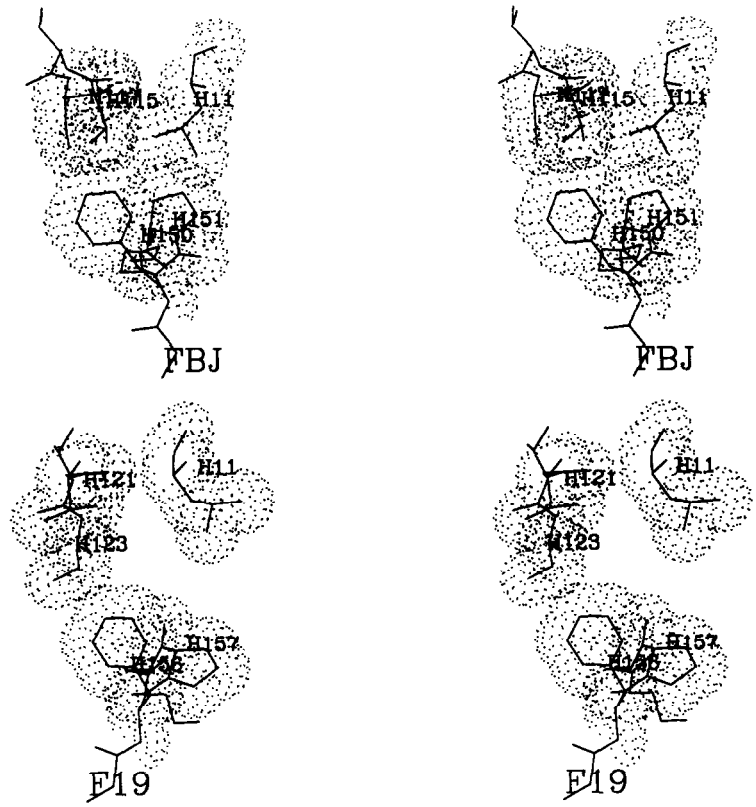

Fig. 1. Comparison of the structure in case of *cis*-prolyl residue (Pro H 151, Brookhaven, 1FBJ, *R*-value = 0.19, top) and *trans*-conformation of prolyl residue (Pro H 157, Brookhaven 1F19, *R*-value = 0.30, bottom) in the same position (only the numbering is different) in the heavy chain of antibodies. It is the switch region in immunoglobulins responsible for the ball-and-socket joint of the V_H chain to the C_{H1} chain, where prolyl residues should always be in the *cis*-conformation. All known immunoglobulin sequences and all known T-cell-receptor sequences are highly conserved in this region.

$i-6,\ldots,i-1,i+1,\ldots,i+6$. Prolyl residues which do not have six neighbours in both directions are not included in this analysis. Because no additional information could be detected in the remote surrounding the length of the patterns was fixed to ±6 amino acids. The construction of the property matrix and the search within the database was performed using the program PAT [11]. Each amino acid in a protein sequence is described by a vector of physico-chemical properties like hydrophobicity; positively, negatively, generally charged; polar; small, large; aliphatic, aromatic; presence and absence of certain amino acids. A list of the occurring amino acids for each position of the pattern is translated into a property matrix. This property matrix is used when searching in sequence databases, for instance those extracted from the Protein Data Bank [10]. In case of match of one property in several sequences the example is recorded. To avoid unbiased weights we take only one example in the case of identical sequences around prolyl residues. Using this method, consensus patterns are more easily detected in sequences with low similarity. For further details see [12].

3. RESULTS AND DISCUSSION

The scatter of all ω-angles depends on the crystallographic refinement method. Procedures using Diamond's real space refinements often show very sharp distributions (±0.1°). The lack of atomic resolution, which makes independent refinement of atomic coordinates impossible, was combatted by reducing the number of positional parameters from three to less than one per atom using torsion angles as independent variables [13]. Examples for such Protein Data Bank entries are: 1ACX, 1CAC, 1FDH, 1GPD, 1HCO, 1PYP, 1LDH. More advanced techniques, such as that of Hendrickson and Konnert [14], aim to solve the problem by effectively augmenting the number of observations by introducing known stereochemical parameters rather than by reducing the number of variables. These restrained refinement procedures result in broader distribution of ω (±45°), e.g. 1HDS, 1NXB. But in both cases the classification to *trans*- or *cis*-conformation is unambiguous. We also see the clear correlation between the resolution of the protein structures and their *cis*-prolyl content: the lower the resolution the lower the *cis*-prolyl content [3] (resolution below 1.5 Å: 8.1%; between 1.5 Å and 2.2 Å: 7.9%; above 2.2 Å: 5.0%). As a consequence, X-ray structures at lower resolution should be interpreted cautiously: 1F19 shows *trans* PRO L 136, H 151 (resolution 2.8 Å)! We predict *cis*-conformation in these positions as observed in better resolved X-ray structures.

In Table I the neighbours of *cis*-prolyl residues are given. Redundancies are eliminated. It is obvious that practically all 20 residue types are allowed at each position.

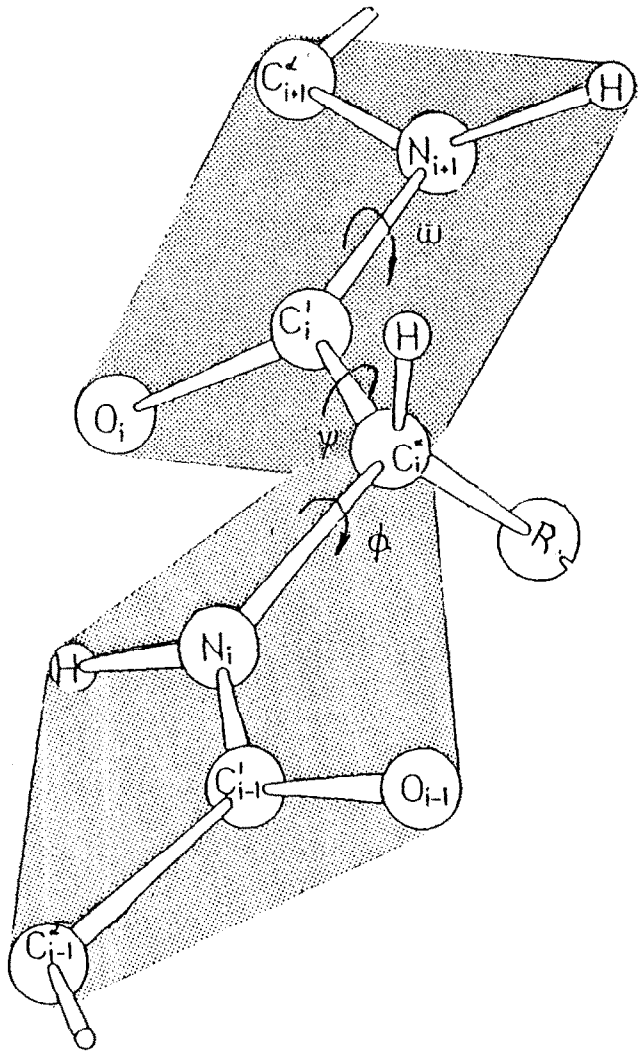

Fig. 2. Definition of the torsional angles of residue *i* at the protein backbone.

In Table II the significant preference values of amino acids in the neighbourhood of *cis*-prolyl residues are given. In the very complex pattern there are some significant features: The N-terminal region near *cis*-prolyl residues differs stronger in amino acid frequencies from N-terminal region near *trans*-prolyl residues than the C-terminal region does. Close neighbours of prolyl residues are more indicative of *cis*-prolyl structure than residues situated more distantly. An accumulation of aromatic residues at the N-terminal region (Tyr, Phe) is evident. But the occurrence of these residues at this position is not conclusive for the existence of *cis*-prolyl residues in the three-dimensional structure. In consequence only a procedure to detect patterns based on properties of amino acid residues can be applied [11].

Using PAT [12], six complex property patterns were found which indicate the location of *cis*-prolyl residues (Table III). Using these 6 patterns we can predict correctly most (176, ≈75%) of known *cis*-prolyl residues by the program PAT. Excluding obviously wrong protein structure (see Fig. 1), we do not predict any proline in *trans*-conformation as *cis*. Further improvement of the method can be achieved first by the analysis of possible patterns around *trans*-prolyl residues and secondly by an increased number of examples in the data base. This number will be increased by both new protein structure determinations and by re-estimation of known structures because the procedures used to develop structures from X-ray data have biased the *cis*-prolyl residue content [3].

The complex pattern around *cis*-prolines probably reflects different reasons. Some special pattern is related to the stability of the local (e.g. phenylalanyl-

Table I

Occurrence of amino acids in the neighbourhood of *cis*-prolyl residues

	Position relative to the *cis*-prolyl residue											
AA	−6	−5	−4	−3	−2	−1	+1	+2	+3	+4	+5	+6
Ala	3	12	2	7	2	5	16	9	9	6	14	5
Arg	3	2	4	2	0	3	5	2	4	3	3	1
Asn	2	6	8	9	2	8	11	11	5	7	8	5
Asp	6	5	4	7	2	3	5	11	7	6	8	10
Cys	4	0	10	1	2	0	6	5	2	5	2	2
Gln	6	8	3	3	8	4	3	5	3	2	2	2
Glu	5	2	3	7	5	4	6	4	7	5	8	6
Gly	16	12	10	10	14	9	11	9	4	14	8	12
His	1	0	2	3	4	3	4	1	1	1	3	4
Ile	8	15	10	8	3	3	2	10	14	1	6	4
Leu	13	9	11	1	2	11	12	4	12	3	2	8
Lys	10	1	7	1	8	3	5	6	3	16	7	6
Met	1	0	3	2	0	2	1	0	2	2	0	2
Phe	5	5	2	8	12	12	1	0	11	6	5	3
Pro	4	1	3	5	10	4	2	10	2	7	2	7
Ser	6	12	12	13	16	12	8	10	7	13	5	11
Thr	11	9	7	18	9	8	5	11	11	12	12	15
Trp	0	3	3	3	1	2	0	0	2	1	2	2
Tyr	3	2	6	3	10	21	8	4	4	6	3	3
Val	13	16	10	9	10	3	9	8	10	4	20	12

AA = Amino acid type (three-letter code). Subunits, different X-ray data for one structure and identical sequences are counted only once.

Table II

Preference of amino acid residues in the neighbourhood of prolyl residues in *cis*-conformation compared with their abundance in the neighbourhood of prolyl residues in total

		Position relative to the *cis*-prolyl residue											
AA	%	−6	−5	−4	−3	−2	−1	+1	+2	+3	+4	+5	+6
Ala	6.2	0.28		0.18		0.24	0.45	2.11					0.56
Arg	2.2		0.52		0.46	0.00							0.18
Asn	5.7	0.42				0.34			1.96				
Asp	5.0			0.54		0.32	0.52						
Cys	2.7		0.00	4.03	0.52		0.00	3.74	3.46		2.79		
Gln	3.4					1.98					0.38		0.52
Glu	4.3		0.46	0.46									
Gly	8.9									0.56	1.84		
Ile	5.8		2.30				0.43	0.38		1.97	0.13		
Leu	6.1				0.12	0.28			0.40		0.34	0.19	
Lys	5.0		0.13		0.14		0.46			0.37	2.15		
Phe	4.9			0.44	1.96	2.62	2.65	0.27	0.00	2.25			
Pro	4.0		0.17	0.42		1.99		0.47	1.98	0.33		0.33	
Ser	8.7						1.96						
Thr	8.8				2.19								2.17
Tyr	5.0		0.52			1.99	2.93				2.06	0.48	0.50
Val	8.6						0.30				0.50	2.54	

Due to the low number of tryptophan, methionine and histidine, respectively, there are no significant data for these amino acids at all. In the table only significant values different from 1.0 are given. AA = amino acid type (three-letter code). The second column gives the total content (%) of this amino acid type in the neighbourhood of *cis*-prolyl residues.

The preference value is estimated as the quotient of number of the district amino acid residue at the given position flanking *cis*-prolyl-residue divided by the number of the same amino acid residues at the given position flanking any prolyl residue.

Table III

The properties of the amino acid sequences around *cis*-prolyl residues *i*

Group	Properties of the neighbour		Properties of the neighbourhood $i \pm 6$ residues	Number of observations (total/indep.)
	$i-1$	$i+1$		
1	hydrophobic aromatic	no prolyl residue	only few glycine residues, polarity changing	60/26
2	only leucine		prolyl residues forbidden, quite polar, ionic charges forbidden	34/7
3	only threonine		prolyl residues forbidden, polarity changing	15/8
4	only serine		aromatic glycine and prolyl residues very seldom (*i*-2 polar)	18/10
5	polar	hydrophobic	glycine residues very seldom, ionic charges very seldom	33/9
6	only glycine		no proline residue, ionic charges very seldom, small and hydrophobic residues preferred	16/6

prolyl peptides prefer more the *cis*-conformation than other dipeptides of proline [15]) and global conformation of the protein, respectively. Furthermore, we guess that the specificity of peptidyl prolyl isomerase is also of some concern. These enzymes are important in the protein folding to overcome the often rate limiting step of *trans-cis*-isomerisation.

In summary we describe here a prediction method for *cis*-conformation of prolyl residues. This method will be valuable for protein structure prediction as well as for experimental structure determination by X-ray analysis and nuclear magnetic resonance.

Acknowledgements: The authors would like to thank Dr P. Bork and Dr C. Grunwald for guidance at the work with their program package PAT. The program package can be purchased with a set of property patterns on request from Dr P. Bork (*Address:* Zentralinst. f. Molekularbiologie, Dept. of Biomathematic, R.-Rössle-Str. 10, Berlin, O-1115, Germany). This work was supported by Berliner Senat.

REFERENCES

[1] Levin, J.M. and Garnier, J. (1988) Biochim. Biophys. Acta 955, 283–295.
[2] Hodgman, T.C. (1989) CABIOS 5/1, 1–13.
[3] Stewart, D.E., Sarka, A. and Wampler, J.E. (1990) J. Mol. Biol. 214, 253–260.
[4] Schmid, F.X., Grafl, R., Wrba, A. and Beintema, J.J. (1986) Proc. Natl. Acad. Sci. USA 83, 872–876.
[5] Richardson, J.S. (1981) Adv. Prot. Chem. 34, 167–339.
[6] Fischer, G., Bang, H. and Mech, C. (1984) Biomed. Biochim Acta 43, 1101–1111.
[7] Fischer, G., Wittman-Liebold, B., Lang, K., Kiefhaber, T. and Schmid, F.X. (1989) Nature 337, 476–478.
[8] Takahashi, N., Hayano, T. and Suzuki, M. (1989) Nature 337, 437–475.
[9] Kiefhaber, T., Grunert, H.-P., Hahn, U. and Schmid, F.X. (1990) Biochemistry 29, 6475–6480.
[10] 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.
[11] Bork, P. and Grunwald, C. (1989) Studia Biophys. 129, 231–240.
[12] Bork, P. (1989) FEBS Lett. 257, 191–195.
[13] Diamond, R. (1971) Acta Crystallogr., Sect. A A27, 436–452.
[14] Konnert, J.H. (1976) Acta Crystallogr., Sect. A A32, 614–617.
[15] Harrison, R.K. and Stein, R.L. (1990) Biochemistry 29, 3813–3816.