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

BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
Pages 660-665

ON α-HELICES TERMINATED BY GLYCINE
1. IDENTIFICATION OF COMMON STRUCTURAL FEATURES

Robert Preißner[a,b] and Peer Bork[c]

[a] Charité, Humboldt Universität, Institut für Biochemie, Hessische Str. 4-6, Berlin, O-1040, Germany

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

[c] Zentralinstitut für Molekularbiologie, Biomathematik, Robert-Roessle-Str. 10, Berlin, O-1115, Germany

Received August 26, 1991

Summary: About one third of all helices is terminated by residues with a positive torsion angle ϕ. 74% of them are glycines. This strong propensity can be explained by typical bifurcated three-center hydrogen bonds which are only compatible with a positive torsion angle ϕ, causing helix termination. An algorithm was developed to identify these structural features in α-helices. 158 out of 456 helices in 79 different well-refined protein structures examined in our analysis were found to have a glycine with this special conformation which have been conserved remarkably during evolution.

Hydrogen bonds (H-bonds) are essential for the stabilization of proteins (1) and form characteristic patterns in secondary structures (2). About 90% of all H-bonds in α-helices are bifurcated (3). The resultant typical three-center H-bond patterns (fig.1) allow to study helix ends from a new aspect. Usually, helices and their ends can be defined by torsion angles ϕ, ψ (4) or by specific main chain H-bonds (5) or by Cα-carbon positions (6). Preferences for helices (e.g. (7)) were tabulated and glycine turned out to occur prevalently at the C-terminal ends (C-caps) (8). A comparative analysis of local conformations revealed that these glycines are often in the rare αL-conformation with a positive torsion angle ϕ (positive ϕ) (9-11). The enlarged database of highly resolved tertiary structures allows a detailed analysis of the underlying structural features. Here we explain the high content of glycines in C-caps by a unique three-center H-bond pattern and show its evolutionary conservation in structurally related protein sequences.

Abbreviations: H-bonds, hydrogen bonds; C-cap, C-terminal residue of the helix; positive ϕ, positive torsion angle between backbone atoms N and Cα; PDB, Brookhaven Protein Data Bank (Release 52).

0006-291X/91 $1.50

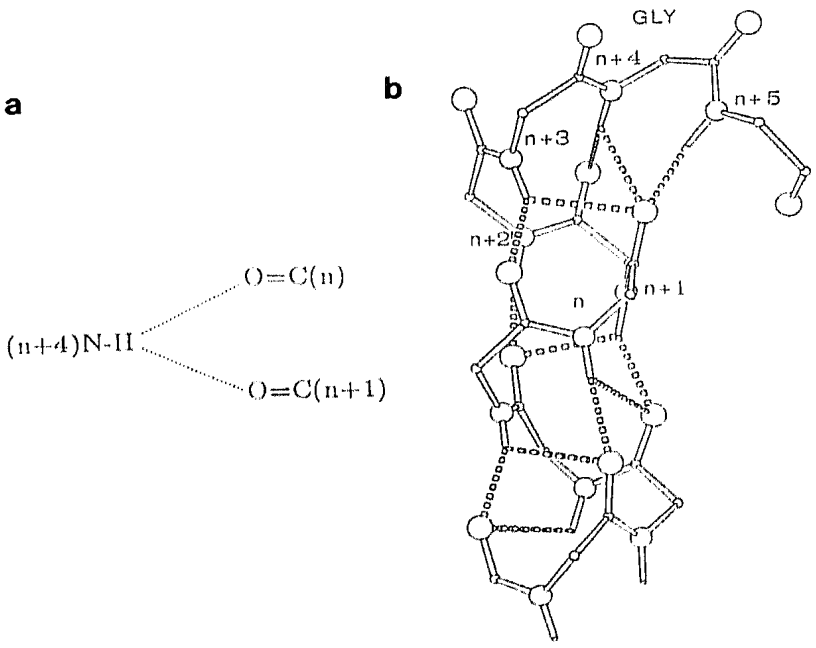

Fig.1. Representation of three-center hydrogen bonds.
a) schematic representation.
b) helix nomenclature with H-bond pattern. The C-terminal N-H atoms are numbered.

Materials and Methods

Since the period in α-helices is 3.6 and not a whole number, most of the H-bonds are bifurcated (3). This means that the N-H group points into the direction between two carbonyl oxygen atoms. In these bifurcated three-center H-bonds two acceptors (n)-C=O, (n+1)-C=O share one donor (n+4)-N-H (fig. 1b). The hydrogen atom lies in the plane defined by the nitrogen and the two carbonyl oxygen atoms. Along the helix they form a typical zigzag pattern.

The geometrical criteria for three-center H-bonds were developed by examining a representative small set of protein structures refined to a resolution of 1.6 Å or better (3). In the Brookhaven Protein Data Bank (PDB) (12) many protein structures determined at lower resolution than 1.6 Å exhibit poor H-bond geometries (and corresponding energies) and allow only approximate assignment of secondary structure. Nearly all of the parameters of H-bond patterns such as their energy contributions, ϕ-, ψ-values can be generated by DSSP (5).

To generalize the data found in the small dataset of highly resolved structures (3) simple rules were derived which allow the use of DSSP (5). To prevent the problem of redundancy we generated a reduced database containing 79 well-refined protein structures, excluding closely related ones (13) from the whole PDB. The following criteria identified exactly the same C-caps as found by the rule of specific H-bond patterns:
i) Only helices with more than five residues are considered (definition by DSSP).
ii) A glycine with a positive ϕ is obligatory at the C-cap; (positions (n+1) and (n+2) from the helix ends are taken into account because the definition of precise helix termination is sometimes difficult).
iii) Solvent accessibility from (n-8) to (n+2) (counted from the C-caps) alternates with a period of three or four residues.

These features were used to check all DSSP files for C-caps having the putative H-bond patterns. To make sure that the frequent occurrence of glycines at C-caps is statistically significant, the database HSSP (14) was utilized. It contains files

corresponding to PDB entries. All sequences from the protein sequence database SWISSPROT that exhibit significant homology to the considered protein are given in a multiple sequence alignment. The sequence variability is computed in each position of all 79 proteins. From these a relative variability weighted against the average variability was calculated. The HSSP database improves the judgement of the degree of evolutionary conservation, because its base of alignend master sequences is by two orders of magnitude larger than that of PDB (14).

Results and Discussion

The H-bonds at the C-caps were studied and glycine (n+4) was found to be prevalent with the following arrangement (fig.1b):

1. (n)-C=O $\cdots$ H-N-(n+3),
2. (n)-C=O $\cdots$ H-N-(n+4),
3. (n)-C=O $\cdots$ H-N-(n+5) and
4. (n+1)-C=O $\cdots$ H-N-(n+4).

The first H-bond usually occurs in 3_{10} helices and the second is the major component in 3.6_{13} α helices. Both are known to be present at helix ends (5). The third H-bond was found to be associated with the occurence of glycines (9), whereas the fourth was so far not described to be typical for glycine (n+4) in C-cap position. The occurrence of these H-bonds in all examples detected implies a common structural signal. All these C-caps as extracted from the most different proteins superimpose very well (fig.2). This is mirrored in the values of torsion angles ϕ and ψ. Up to position (n+2) (or position -2 counted from the glycine) typical α-helical torsion angles were observed. Uniform slight distortions occur at position

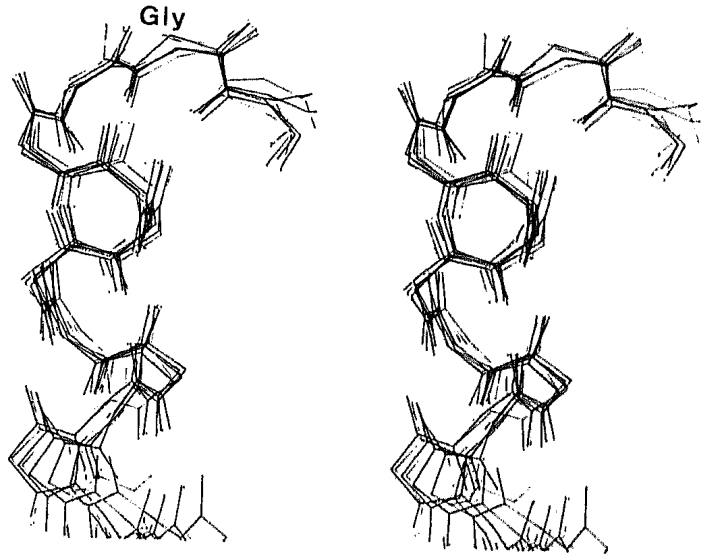

Fig.2. Stereo view on a multiple structural alignment of helix C-termini detected by H-bond patterns in highly resolved protein structures (for sequence alignment see fig.3). Glycine is marked to indicate the C-terminus.

Prot[1]	n[2]	AA[3]	ϕ Acc[4]	ψ Var[5]	AA	ϕ Acc	ψ Var	AA	ϕ Acc	ψ Var	AA	ϕ Acc	ψ Var	AA	ϕ Acc	ψ Var	AA	ϕ Acc	ψ Var
CSE	15	K	-54	-47	V	-72	-32	Q	-65	-38	A	-64	-33	Q	-66	-20	G	118	11	F	-129	111
			94	58		0	34		17	47		90	26		77	27		52	0		61	19
CSE	113	W	-54	-39	A	-63	-40	T	-66	-35	T	-86	-20	N	-109	20	G	86	22	M	-68	143
			55	22		0	12		25	38		118	46		76	27		31	47		1	41
CSE	141	N	-68	-46	A	-58	-46	Y	-65	-42	A	-60	-32	R	-64	20	G	99	29	V	-98	131
			84	49		0	23		52	52		92	41		103	53		32	23		0	41
GRS	38	R	-71	-4	A	-58	-47	A	-61	-39	E	-59	-29	L	-87	2	G	102	-3	A	-67	142
			58	36		0	21		20	28		129	42		54	41		63	17		10	51
GRS	205	I	-71	-49	L	-60	-47	S	-61	-48	A	-62	-33	L	-83	-4	G	110	3	S	-76	145
			0	30		0	33		22	44		21	46		9	6		56	0		5	38
MBD	145	K	-70	-39	Y	-55	-50	K	-57	-49	E	-51	-52	L	-67	-14	G	78	33	Y	-108	140
			56	19		2	0		160	12		160	13		81	21		70	11		37	0
PRK	236	Y	-61	-46	L	-70	-37	M	-64	-42	T	-69	-25	L	-76	-12	G	78	17	K	-78	-29
			5	29		6	24		10	24		15	31		60	47		62	48		127	40
RNT	25	K	-61	-44	L	-63	-37	H	-63	-46	E	-64	-37	D	-80	-4	G	64	34	E	-117	157
			93	37		22	40		41	43		117	37		93	46		69	30		114	25
TLN	309	A	-57	-51	F	-65	-38	D	-65	-39	A	-63	-31	V	-97	4	G	72	24	V	-98	100
			0	14		0	6		75	29		1	0		0	0		22	0		3	20
Con[6]		X	-63	-41	L,A	-63	-41	X	-64	-42	X	-63	-34	X	-81	-2	G	89	21	X	-94	111
			44	29		3	19		45	32		84	28		60	30		59	18		30	32

Fig.3. Multiple sequence alignment of the helix C-termini corresponding to the superposition displayed in fig.2. Some descriptors such as torsion angles, accessibility and variability are introduced for each residue.
[1] The proteins given by their Brookhaven Protein Data Bank code (12).
[2] Position of the terminal glycine.
[3] Amino acids shown in one letter code. X denotes an arbitrary residue.
[4] Solvent accessibility calculated by DSSP (5).
[5] Position dependent variability derived from an aligned set of homologous sequences in HSSP (14). The lower the value the better is the conservation of the respective residue in related sequences.
[6] A consensus line is introduced to represent dominating residues and patterns for the different descriptors (mean values are given).

(n+3) (-1 from glycine) with average (ϕ,ψ) pairs of (-81°,-2°) (see consensus in fig.3). In addition a striking accumulation of hydrophobic amino acids at position (n) (position -4 from glycine) is obvious. Especially leucine and alanine occur very frequently. This high hydrophobicity leads to a decreased dielectric constant which, in turn, increases the energy contribution of the H-bonds. Fig.3 gives a sequence alignment of a few C-caps including the above-mentioned structural features like torsion angles or solvent accessibility.

Sometimes glycine is shifted by one position against the hydrophobic (inaccessible) face of the helix. In this manner the accessibility pattern appears blurred in the consensus average, which can be overcome by a detailed analysis at the sequence level (15).

A total of 456 helices with a length of at least 6 residues were collected in 79 proteins. In fig. 4 the propensity of all amino acids in the database is displayed: 1. to occur at C-caps, 2. to adopt positive ϕ, 3. to occur at C-caps with positive

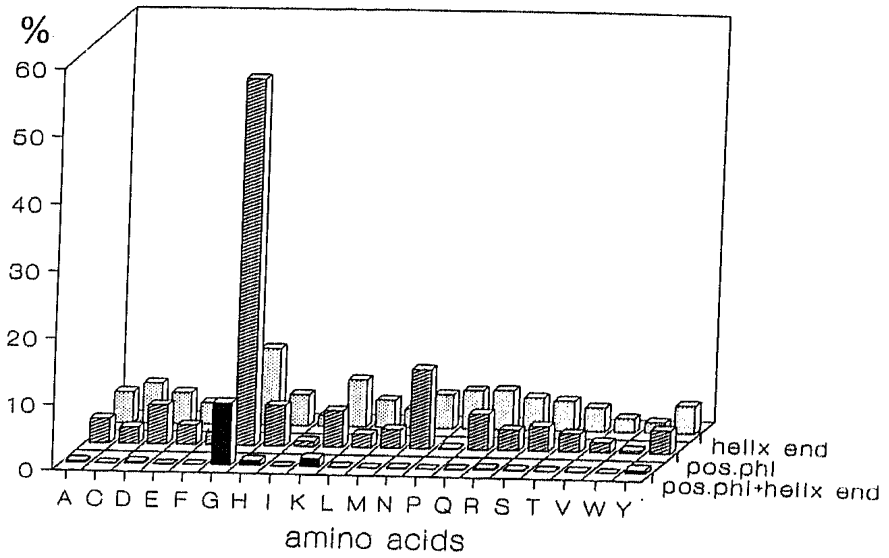

Fig.4. Preferences of amino acids (one letter code) in the studied 79 different proteins
(i) to occur at helix C-termini (helix end),
(ii) to adopt a positive torsion angle ϕ (pos. phi) and
(iii) to occur at helix C-termini with positive torsion angle ϕ.

ϕ. There is a strong coupling between the first two features in the case of glycine: The rate of glycines with a positive ϕ compared to all glycines in the PDB increases from 54% to 91% at helix ends. Positive ϕ-angles at C-caps go along with glycines (158), lysines (17), asparagines (8), glutamines (7), from which glycines make up 74%. This surprising preference of glycine with a positive ϕ for helix ends indicates a functional meaning. For example, this structural feature occurs in all DNA-binding helix-turn-helix motifs of distantly related procaryotic repressors of known structure (16) in which glycine is one of the most conserved residues. This type of helix termination may contribute to the flexibility of the motif as seen in Mellitin (17). Mellitin, included in the venom of the honey bee consists of one long helix which is kinked at a glycine with a positive ϕ (18). Different structures of Mellitin (17-19) confirm the observations of a variable bent angle of the helix. This movement may be correlated with membrane binding as predicted for four helix bundle apolipoproteins (20). Such a functional meaning should find its expression in a high evolutionary conservation at the sequence level. Indeed, inspecting HSSP (14) glycine at C-caps was found to be 20% more conserved than in the average. The described structural feature is widespread among topologically different proteins. About 30% of all C-caps contain a glycine with a positive ϕ (158 of 456 helices in the nonredundant database, 659 of 2314 helices in the whole PDB).

Summing up, this signal of helix disruption is so strong that a recognition at the sequence level should be feasible (15).

Acknowledgments: The authors are indebted to W. Saenger and J. Reich for helpful comments and critical reading of the manuscript.

References

1. Baker, E.N. and Hubbard, R.E. (1984) Prog. Biophys. molec. Biol. 44, 97-179.
2. Richardson, J.S. and Richardson, D.C. (1989) in: Prediction of Protein Structure and Principles of Protein Conformation, (G. Fasman, Ed.) Plenum, New York, 1-98.
3. Preißner, R., Egner, U. and Saenger, W. (1991) FEBS submitted.
4. Ramachandran, G.N. and Sasisekharan, V. (1968) Adv. Protein Chem. 23, 283-437.
5. Kabsch, W. and Sander C. (1983) Biopolymers 22, 2577-2637.
6. Levitt, M. and Greer, J. (1977) J. Mol. Biol. 114, 181-239.
7. Levin, J.M. and Garnier J. (1988) Biochim. Biophys. Acta 955, 283-295.
8. Richardson, J.S. and Richardson, D.C. (1988) Science 240 1648-1652.
9. Schellman, C. (1980) in: Protein Folding, R. Jaenicke, Ed. Elsevier/North-Holland Biomedical Press, 53-61.
10. Presta, L.G. and Rose, G.D. (1988) Science 240, 1632-1641.
11. Efimov, A.V. (1991) Protein Engineering 4, 245-250.
12. 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.
13. Rooman, M.J. and Wodak, S.J. (1991) Proteins 9, 69-78.
14. Sander, C. and Schneider, R. (1991) Proteins 9, 56-68.
15. Bork, P. and Preißner, R. (1991) BBRC 180, 666-672.
16. Steitz, T.A. (1990) Quart.Rev.Biophys. 23, 205-280.
17. Teikichi, I., Nobuhiro, G. and Inagaki, F. (1991) Proteins 9, 81-89.
18. Terwillinger, T.C. and Eisenberg, D. (1982) J. Biol. Chem. 257, 6016-6022.
19. Gribskov, M., Wesson, L. and Eisenberg, D. (1990) in (12).
20. Breiter, D.R., Kanost, M.R., Benning, M.M., Wesenberg, G., Law, J.H., Wells, M.A., Rayment, I. and Holden, H.M. (1991) Biochemistry 30, 603-608.