Protein Science (1994), *3*:2395–2410. Cambridge University Press. Printed in the USA.

Conservation of polyproline II helices in homologous proteins: Implications for structure prediction by model building

ALEXEI A. ADZHUBEI[1,2] AND MICHAEL J.E. STERNBERG[1]

[1] Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, P.O. Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, United Kingdom
[2] CRC Biomolecular Structure Unit, The Institute of Cancer Research, Cotswold Road, Sutton, Surrey SM2 5NG, United Kingdom

(RECEIVED July 20, 1994; ACCEPTED October 7, 1994)

Abstract

Left-handed polyproline II (PPII) helices commonly occur in globular proteins in segments of 4–8 residues. This paper analyzes the structural conservation of PPII-helices in 3 protein families: serine proteinases, aspartic proteinases, and immunoglobulin constant domains. Calculations of the number of conserved segments based on structural alignment of homologous molecules yielded similar results for the PPII-helices, the α-helices, and the β-strands. The PPII-helices are consistently conserved at the level of 100–80% in the proteins with sequence identity above 20% and RMS deviation of structure alignments below 3.0 Å. The most structurally important PPII segments are conserved below this level of sequence identity. These results suggest that the PPII-helices, in addition to the other 2 secondary structure classes, should be identified as part of structurally conserved regions in proteins. This is supported by similar values for the local RMS deviations of the aligned segments for the structural classes of PPII-helices, α-helices, and β-strands. The PPII-helices are shown to participate in supersecondary elements such as PPII-helix/α-helix. The conservation of PPII-helices depends on the conservation of a supersecondary element as a whole. PPII-helices also form links, possibly flexible, in the interdomain regions. The role of the PPII-helices in model building by homology is 2-fold: they serve as additional conserved elements in the structure allowing improvement of the accuracy of a model and provide correct chain geometry for modeling of the segments equivalenced to them in a target sequence. The improvement in model building is demonstrated in 2 test studies.

Keywords: conserved regions; homology modeling; mobile conformation; protein structure; regular secondary structure

A major cluster (Adzhubei et al., 1987a; Richardson & Richardson, 1989) in the conformational distribution in ϕ,ψ angles space is termed polyproline II (PPII) because of its similarity with the left-handed helical conformation of the homopolymer of *trans*-proline (Cowan & McGavin, 1955; Arnott & Dover, 1968). This cluster, however, is populated with all types of residues including, but not restricted to, proline. Taken together, the major clusters in the distribution (αR, β, PPII, αL, β-αR *trans*) combine up to 90% of the residue composition of globular proteins with the β and the PPII accounting for approximately the same proportion of 20% (Adzhubei et al., 1987b). These results though represent *single* residue conformations, i.e., residues unrelated to their neighbors. Analysis of the possible regular (periodic) structures represented in globular proteins showed, apart from the α-helices, the β-strands, and the 3_{10}-helices, high occurrence of the left-handed helices of 4 or more residues in length (Adzhubei & Sternberg, 1993). The ϕ,ψ angles specifying these left-handed helices were in the PPII cluster observed in the distribution of individual residues. This structure, termed the polyproline II (PPII) helix, appeared to be the only regular structure class significantly populated in globular proteins, which was not included in the currently used secondary structure classification schemes (Kabsch & Sander, 1983; Richards & Kundrot, 1988; Sklenar et al., 1989). The PPII-helices can be identified with the examples of the collagen-like helix found in globular proteins (Ananthanarayanan et al., 1987). The experimental data obtained by other research groups also showed that the PPII conformation can be structurally important for polypeptides (Siligardi et al., 1991; Makarov et al., 1992; Woody, 1992) and proteins (Lim & Richards, 1994; Sreerama & Woody, 1994; Yu et al., 1994). It has been suggested that the PPII-helices should

Reprint requests to: Michael J.E. Sternberg, Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, P.O. Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, UK; e-mail: m_sternberg@icrf.icnet.uk.

be classified as a regular secondary structure and, as such structure, used in homology model building.

An important question, however, remained unresolved. From the initial series of structures of homologous proteins, e.g., globins (Perutz et al., 1968; Fermi et al., 1984) and serine proteinases (Mauguen et al., 1982), it was observed that blocks of regular secondary structure in proteins tend to form conformationally conserved regions (Subramanian et al., 1977; Lesk & Chothia, 1980; Chothia & Lesk, 1986). The extent of structure similarity is related to the level of sequence identity (Holm et al., 1992; Flores et al., 1993; Hilbert et al., 1993). The variation in Cartesian and dihedral geometry observed for the PPII-helices is similar to the spread for the α-helices, the β-strands, and the 3_{10}-helices (Adzhubei & Sternberg, 1993). Thus, to establish the PPII-helices as a structural class closely corresponding to the other 2 periodic secondary structures, it is necessary to carry out the analysis of their conservation in homologous proteins.

Any additional information on the evolutionarily conserved regions is of importance for the protein structure prediction and analysis. In structure prediction by homology, the identification of structurally conserved regions (SCRs) in proteins and assigning their backbone conformation to the respective parts of a modeled structure is one of the starting conditions (Greer, 1981, 1990; Blundell et al., 1987; Sutcliffe et al., 1987a, 1987b). In addition, the methods developed to identify folds and structural motifs in proteins, as well as to assign proteins to different families rely strongly on conservation of secondary structure (Blundell & Johnson, 1993; Orengo et al., 1993; Yee & Dill, 1993). The absence of insertions and deletions within secondary structures is used also for the refinement of sequence alignment techniques (Barton & Sternberg, 1987).

In this work we will show that the PPII-helices follow, in homologous protein structures, the same pattern of behavior as the other regular structures. The analysis will mainly concentrate on homologous structures with sequence identity of 50–20%, representing the most important part of the similarity range. At these levels of sequence identity, SCRs can be clearly identified. There is also the possibility of major discrepancies in the conformations of variable regions (VRs), with insertions/deletions most likely to be observed in VRs. Structures with sequence identity above 50% have highly similar backbone conformations and cannot be used as a source of relevant data. In the "twilight zone" of sequence identity, below 20%, sequence alignments are unreliable (Sander & Schneider, 1991) and such structures are normally not included in our analysis.

Implications for modeling of the extension of SCRs to accommodate the PPII-helices could include better approximation for the conformation of VRs. VRs in protein structure often have substantial length, and if structurally conserved elements could be identified in VRs, this will provide useful breaking points. Modeling then could be done for shorter segments. The modeling examples included in this work aim to show how the PPII-helices can be utilized to simplify modeling and increase its accuracy.

Methods

The data set

Three protein families were analyzed for the level of conservation of the polyproline II helices: serine proteinases, aspartic proteinases, and immunoglobulin (Ig) constant domains. The initial selection criteria for the structures included in the dataset were the sequence identity for every pairwise alignment less than 55% and RMS deviation for the pairwise structural alignments of the selected structures lower than 4.5 Å. Three groups representing 3 protein families were established. Multiple structure alignments were performed and a reference structure was identified for each family on the basis of the lowest mean RMS with the other molecules. At the final selection stage, the sequence identities were recalculated according to structural alignments of the reference structure with every other structure in a family group. For the structures retained in the family groups, the sequence identities calculated in this manner fell within the 55–20% interval (see Table 1). Thus, structural alignments were used to collect data on the conservation of the PPII-helices covering a wide range of sequence identities. For the reference structures, RMS deviations with the other member structures of a family stayed $\leq$3.5 Å, with the only exception of 2sga displaying an RMS of 4.1 Å from the reference structure 1tld of serine proteinases. 2sga was also among the 4 molecules in the dataset whose sequence identity with related reference structures was below 20% (Table 1). These molecules displayed high structural dissimilarity with the rest of the structures in the families. The separate pairwise structural alignments with the reference structures were therefore composed and used for further analysis for 2sga and 2alp of the serine proteinases and for 3hvp of the aspartic proteinases. These structures were included in the dataset to provide information on the level of conservation of the PPII-helices at the margin of the twilight zone of sequence identity.

Initial sequence alignments were composed using the programs GAP (GCG) and ALIGN (PIR). Structural alignments were calculated by the program MULSTR (Pickett et al., 1992), implementing the modified algorithm of Taylor and Orengo (1989a, 1989b). The program produced reliable results for all levels of sequence identity, exemplified by the alignment of residues forming the catalytic triad, the substrate specificity pocket, and other functionally important residues between the structures with low homology in the serine proteinases (Fig. 3B). The RMS deviation matrix, constructed for each family in order to identify a reference structure, was computed according to McLachlan (1972). The sequence identities for aligned structures were calculated as part of the general analysis of the alignment data with the package written in FORTRAN and C, running under UNIX.

Secondary structure identification

Secondary structure definitions of the β-sheets, the α-helices, and the 3_{10}-helices were assigned according to DSSP (Kabsch & Sander, 1983). For further analysis, the 3_{10}-helices were included in the class α-helices. The PPII helices were defined using the regular segment search (RSS) algorithm with the peptide group (C^{α}-C^{α}) structural unit geometry and the 2-step classification, both introduced in Adzhubei and Sternberg (1993). The technique involved monitoring the deviation of torsion angles ϕ, ψ and α from their mean values for the initial assignment of conformational types. The assessment of the hydrogen bonding patterns was used as the final criterion, with no periodic main-chain to main-chain hydrogen bonds allowed in PPII segments. The PPII-helices comprising 4 or more C^{α} positions

Table 1. *The data set: representative protein structures for the 3 families*[a]

Protein family	PDB code	Resolution Å	Sequence identity %	Reference
Serine proteinases				
β-Trypsin, bovine[b]	1tld	1.50		Bartunik et al., 1989
α-Chymotrypsin, bovine	4cha	1.68	46.08	Tsukada and Blow, 1985
Tonin, rat	1ton	1.80	42.86	Fujinaga and James, 1987
Kallikrein A, porcine	2pka	2.05	40.54	Bode et al., 1983
Elastase, porcine	3est	1.65	38.46	Meyer et al., 1988
Trypsin, *Streptocmyces griseus*	1sgt	1.70	35.55	Read and James, 1988
Elastase, human neutrophil	1hne	1.84	34.13	Navia et al., 1989
Protease II, rat mast cell	3rp2	1.90	33.95	Remington et al., 1988
Proteinase A, *S. griseus*	2sga	1.50	19.25	Moult et al., 1985
α-Lytic protease, *Lysobacter enzymogenes*	2alp	1.70	17.24	Fujinaga et al., 1985
Aspartic proteinases				
Penicillopepsin, *Penicillium janthinellum*[b]	3app	1.80		James and Sielecki, 1983
Endothiapepsin, *Endothia prasitica*	4ape	2.10	54.69	Pearl and Blundell, 1984
Rhizopuspepsin, *Rhizopus chinensis*	2apr	1.80	40.89	Suguna et al., 1987
Pepsin, porcine	5pep	2.34	32.24	Cooper et al., 1990
Chymosin B (renin), bovine	1cms	2.30	28.80	Gilliland et al., 1990
HIV protease[c] (synthetic)	3hvp	2.80	16.33	Miller et al., 1989
Immunoglobulin constant domains				
IgG1 FC fragment, CH3 domain, human[b]	1fc1	2.90		Deisenhofer, 1981
IgG FAB fragment, CL domain, human	2fb4	1.90	31.31	Marquart et al., 1980
IgG FAB fragment, CH1 domain, human	2fb4	1.90	29.17	Marquart et al., 1980
IgA FAB fragment, CL domain, mouse	2fbj	1.95	28.71	Suh et al., 1986
IgA FAB fragment, CH1 domain, mouse	2fbj	1.95	24.21	Suh et al., 1986
IgG1 FC fragment, CH2 domain, human	1fc1	2.90	23.16	Deisenhofer, 1981
Class I MHC[d], β2 domain, human	1hsa	2.10	22.68	Madden et al., 1992
Class I MHC[d], α3 domain, human	1hsa	2.10	18.95	Madden et al., 1992

[a] Here and in other tables and figures: PDB code, protein code in the Brookhaven Protein Data Bank (Bernstein et al., 1977); sequence identity, as calculated from structural alignments for pairs with the reference structure.
[b] Reference structures for family subsets.
[c] Synthetic enzyme corresponding to the HIV-protease type 1, isolate SF2.
[d] Histocompatibility antigen.

were considered. Shorter 3-residue PPII-helices were only included in the definition if they were equivalenced to longer helices in homologous structures.

Analysis of structural alignments

Pairwise structural alignments of the reference structure with members of the family were inspected. Comparisons were made starting from the first aligned residue at the N-terminus of a shorter sequence because some of the N-terminal regions included fragments absent in the other sequences. A PPII-helix was considered to be conserved in one chain if the equivalenced segment in the other chain was aligned with at least a 50% overlap of segment lengths (see Fig. 1). The proportion of conserved segments of structural class k for a pair of aligned structures A and B was calculated according to the equation:

$$P_{AB}^{consk} = (N_A^{consk} + N_B^{consk})/(N_A^k + N_B^k), \quad (1)$$

where N_A^{consk} and N_B^{consk} are the numbers of segments k in A and B, and N_A^k and N_B^k refer to the total number of segments of class k. The proportion of conserved secondary structure segments in multiple alignments was calculated in a different manner:

$$PM^{consk} = N^{consk}/N^{possiblek}, \quad (2)$$

where N_{SSk}^{cons} is the number of conserved positions of the secondary structure class k, and $N_{SSk}^{possible}$ are all possible positions of class k segments according to the multiple alignment.

To calculate local RMS deviation of the regular secondary structure blocks in the aligned proteins, a window of 4 residues sliding by 1 residue at a time was used. The RMS deviations were computed for each window position corresponding to continuous segments of an aligned, identical secondary structure in both chains. For equivalenced residues, if full window lengths in both molecules were found to be continuous segments of an identical secondary structure type, segments were superimposed and the local RMS deviations in Cartesian space (RMSC) calculated. The local RMS deviations in the torsion angles ϕ,ψ space (RMST) were also calculated for such windows,

$$\mathrm{RMST}_{win} = \left(\sum_{i=1}^{N} \sum_{j=\phi\psi} (t_{ij}^a - t_{ij}^b)^2 / N_{win} \right)^{1/2}, \quad (3)$$

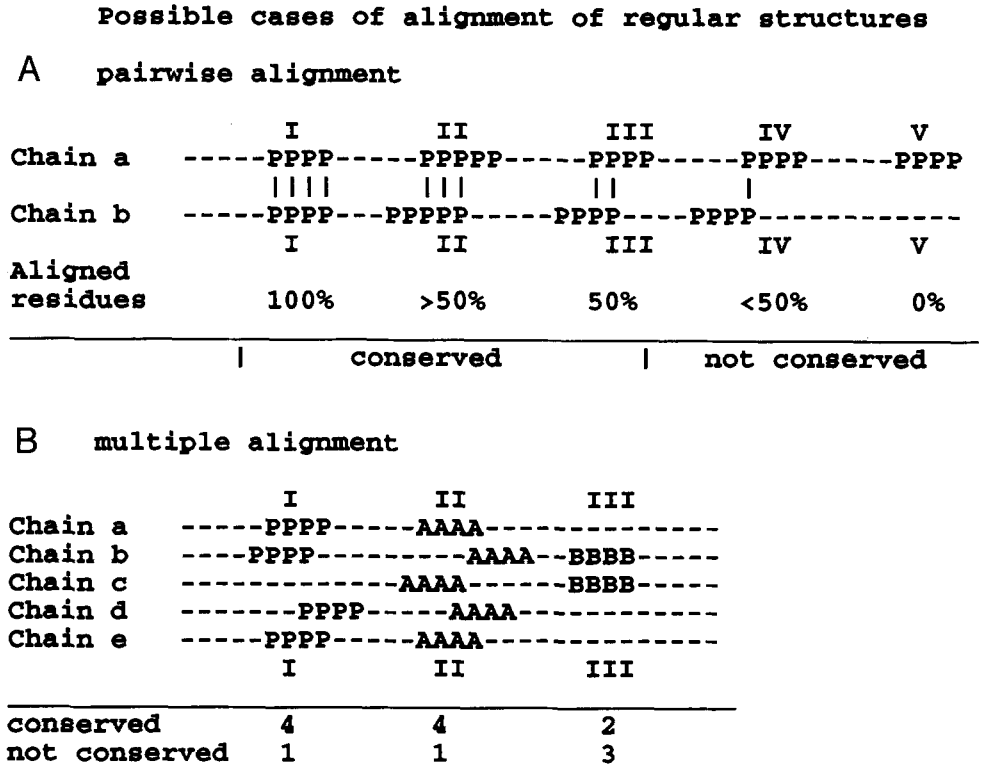


Fig. 1. Illustration of the method to quantify structural conservation. The proportion of conserved residues is calculated differently for pairwise (**A**) and multiple (**B**) alignments. For a pairwise alignment, a conserved structural block has no less than 50% overlap of its segments. In multiple alignments, the 50% rule still applies and all possible positions of the segments in a block are taken into account. **A:** According to Equation 1 in Methods, $P_{ab}^{consP} = (N_a^{consP} + N_b^{consP})/(N_a^P + N_b^P) = (3 + 3)/(5 + 4) = 0.67$. **B:** According to Equation 2 in Methods, $PM^{consk} = N^{consk}/N^{possiblek}$, $PM^{consP} = 4/5 = 0.80$; $PM^{consA} = 4/5 = 0.80$; $PM^{consB} = 2/5 = 0.40$.

where t^a and t^b are the coordinates of aligned structures A and B, respectively, and N_{win} is the number of structural units in a window. The mean local RMSC and RMST were calculated for secondary structure segments as

$$\langle \mathrm{RMSC}_{seg}^{loc} \rangle = \sum_{i=1}^{nw} \mathrm{RMSC}_i^w / nw;$$

$$\langle \mathrm{RMST}_{seg}^{loc} \rangle = \sum_{i=1}^{nw} \mathrm{RMST}_i^w / nw, \quad (4)$$

where nw is the number of windows of a full length in a secondary structure segment. The mean local RMSC and RMST were calculated for each aligned pair of proteins and for the entire family. They were used as a measure of the local structure deviation of aligned polypeptide chains.

To normalize the values of RMSC, it was necessary to determine the RMSC distributions for nonhomologous secondary structures. Accordingly, the local RMSC deviation matrices were constructed for all segments of the α-helices, the β-strands, and the PPII-helices. The length of segments was set equal to a window size of 4 residues. A subset of the database of nonhomologous structures (Adzhubei & Sternberg, 1993), including molecules with all 3 types of regular structure, the α, the β, and the PPII, was used. The RMSC group frequencies were calculated and histograms of the resulting distributions are shown in Figure 2. The distributions have different standard deviations from the mean for different secondary structure classes. The value ($\langle$RMSC$\rangle$ + std) was therefore taken to estimate the upper level of RMSC typical for a structure class. Thus the relative RMSC of the aligned blocks of regular structures were calculated as:

$$\langle \mathrm{RMSC}_{seg}^{rel} \rangle = \langle \mathrm{RMSC}_{seg}^{loc} \rangle / \mathrm{RMSC}_{stn}^{SSk} \quad (5)$$

$$\mathrm{RMSC}_{stn}^{SSk} = \langle \mathrm{RMSC}_{distr}^{SSk} \rangle + \mathrm{STD}_{distr}^{SSk}, \quad (6)$$

where $\langle \mathrm{RMS}_{distr}^{SSk} \rangle$ is the mean RMS for a secondary structure class k calculated from its RMS distribution and $\mathrm{STD}_{distr}^{SSk}$ is its standard deviation.

Modeling

The modeling package written by Paul Bates (Bates & Sternberg, 1992), based on the approach of Jones and Thirup (1986) was used to carry out the homology modeling. In particular, the program 3D-JIGSAW was used to scan the PDB structures database and generate models of variable regions from sequence alignments. Because the aim of modeling is to build a segment of the target structure using the information from the parent structure, the search in the PDB database for segments with low

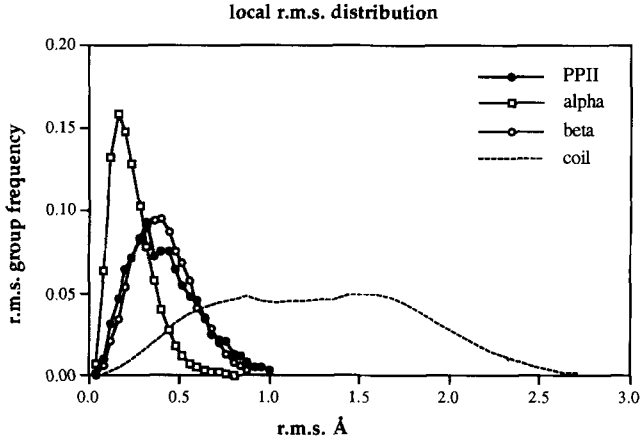


Fig. 2. The distributions of local RMSC calculated from the data set of nonhomologous structures for the structural classes polyproline II helices (PPII), α-helices (alpha), β-strands (beta), and for the segments not included in any secondary structure class (coil). The distribution for the PPII-helices follows closely that for the β-strands, but the standard deviations point to a higher conformational mobility of the PPII-helices: 0.197 for the PPII and 0.166 for the β. The α-helices are clearly the most conformationally rigid structures, with standard deviation of 0.120. The RMS distributions for the 3 regular structural classes are markedly different from the distribution for nonregular coil.

RMS deviation from the fixed ends of a parent loop was performed. The search was based on the RMS of 4 C^{α} positions in the parent structure, 2 positions on both the N- and the C-termini, checked against the cutoff of 2.0 Å. Final selection of a model segment from the produced list was based on the RMS deviation of C^{α} positions, the RMS between C=O vectors of the equivalenced peptide groups, and the dihedral angles between least-square planes of the listed segments and the parent loop. Dayhoff scores of sequence similarity with the target structure were also checked.

Results

Structural conservation

The proteins representing the families, chosen on the basis of sequence and structure similarity (Table 1), display the wide range of sequence similarities with the reference structures within the medium identity interval of 50–20%. This range enables us to analyze the degree of conservation of the PPII-helices for different levels of sequence identity. It was assumed that molecules with sequence similarity higher than 55% had practically identical backbone conformations in the regions of periodic structures. They were therefore not relevant to this analysis and were not included in the representative members of the families.

Conservation of the PPII-helices in aligned structures

The multiple alignments, showing positions of the PPII-helices and the other secondary structures, are presented in Figure 3. There is a strong tendency for the PPII-helices to be conserved in homologous structures. Qualitatively, the level of their conservation is comparable to that of the other structures. To quantify this we have calculated the proportion of conserved PPII segments for 3 families from the pairwise alignments data, see Table 2 (P^{cons}, Equation 1 in Methods) and from the multiple alignments (Fig. 4) (PM^{cons}, Equation 2 in Methods). For all 3 families, the results indicate that the level of conservation of the PPII-helices is comparable to that of the α-helices and the β-strands.

A higher sequence identity does not necessarily lead to higher conservation levels, as demonstrated by the data of the PPII overall conservation for 3 families plotted in Figure 5. Even for the sequence identity of about 30%, the PPII-helices are conserved at the level of 80–100%. A decrease of sequence identity below 25% is normally followed by declining conservation of the PPII segments. However conservation at the level of 50–70% is not uncommon in such cases but drops sharply if the RMS deviation of aligned structures is higher than 3.0 Å. For immunoglobulins, where high conservation levels correspond to lower sequence identity compared to the other 2 subsets, an 80–100% conservation is observed for as low sequence identities as 23–24%.

β-Strands display a generally higher degree of conservation compared to α-/3_{10}-helices and PPII-helices. This is probably an effect of the regular interchain hydrogen bonds, restricting possible absence of single β-strands, which could result in destabilization of the fold. The β-type hydrogen bonds are also more likely to fix rigidly positions of β-strands in the protein structure. Our dataset includes predominantly β-structure proteins, where α-helices do not play a major structural role. This could lead to a lower level of conservation of the α-helices, with only their functionally important segments retaining conservation comparable to that of the β-strands. As a result, the overall conservation of the α-helices could be reduced.

Table 2 also shows that the conservation of the PPII segments correlates with the proportion of residues forming structurally aligned pairs, providing an indirect measure of the number of insertions/deletions in alignments. The PPII conservation tends to stay at a steadily high level when the proportion of the aligned residues is above 94%. A sharp decrease of the conservation level is observed for values lower than 90%.

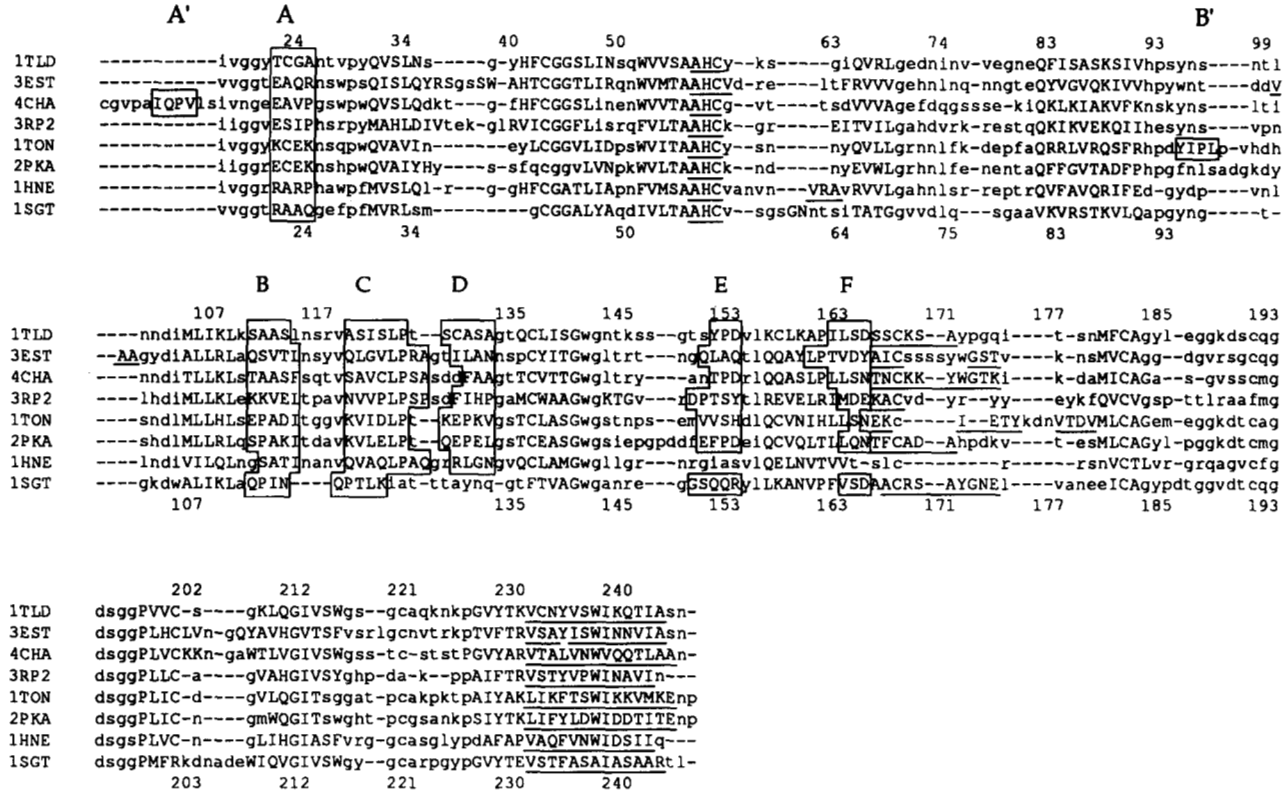
Local superposition of secondary structures

Local superpositions of secondary structures were used to (1) enable comparison for segments of both identical and different length, and (2) assess local deviations in backbone conformation (C^{α}s) of the aligned regular parts of molecules. In this way, equivalenced segments corresponding to the sliding window of size 4 residues were treated as independent segments for which the best possible alignment was found. Superpositions were performed and local RMS deviations calculated for all aligned blocks of the identical regular structure types. The local RMS deviations in Cartesian space (RMSC) and in the torsion angles ϕ, ψ space (RMST) for 3 protein family subsets are listed in Table 2.

To achieve accurate comparison of the structural deviations associated with conserved blocks of different secondary structure classes, it was necessary to take into account the differences in conformational diversity observed for regular secondary structures. These differences can be expressed in terms of RMS deviations and a correction factor accounting for the disparity in the observed levels of RMS deviations in structural classes should be used. A correction factor was introduced in the form of standard RMSC for each secondary structure class (see Methods) and relative RMSC calculated (Figs. 6, 7).

A

Serine proteinases. Sequence identity 46-20 %.



Aspartic proteinases. Sequence identity 55-29 %.

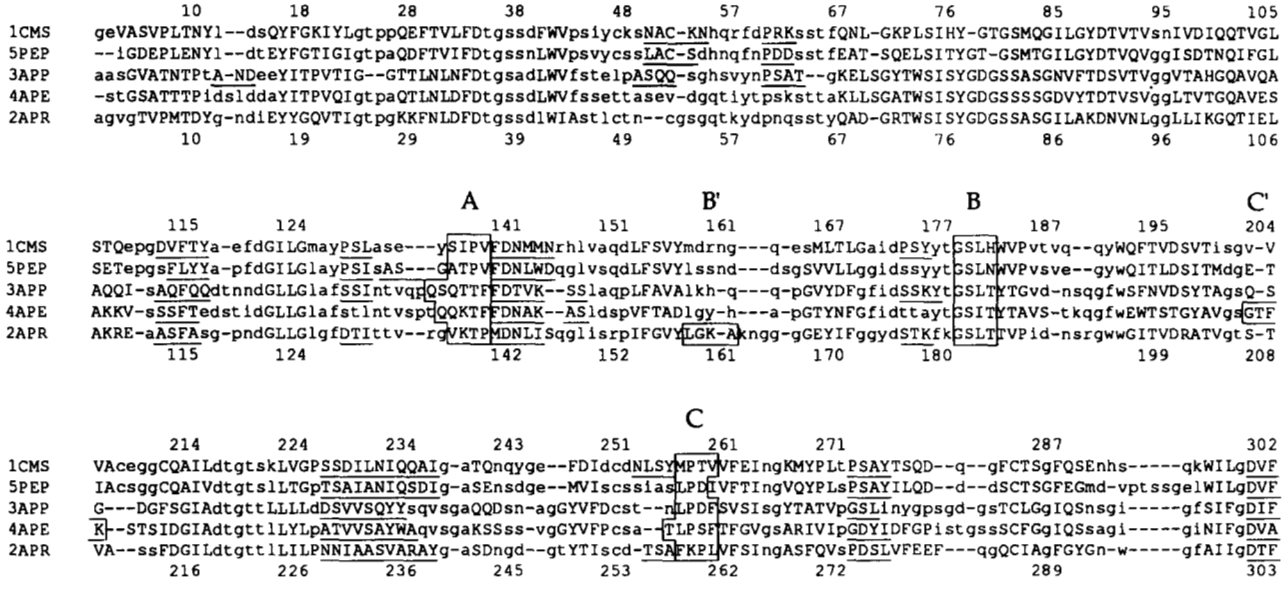
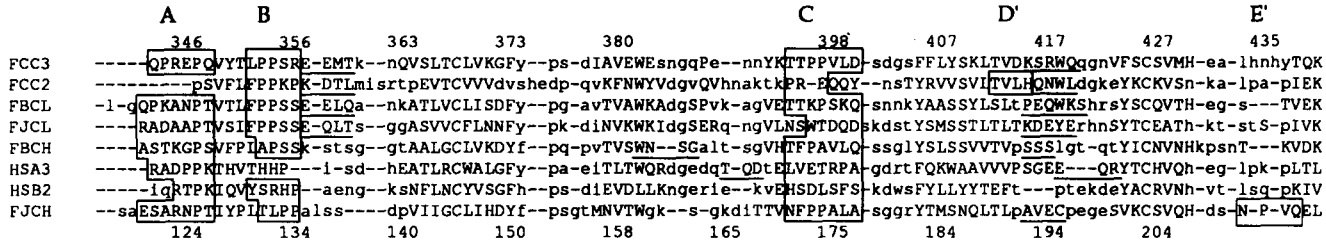


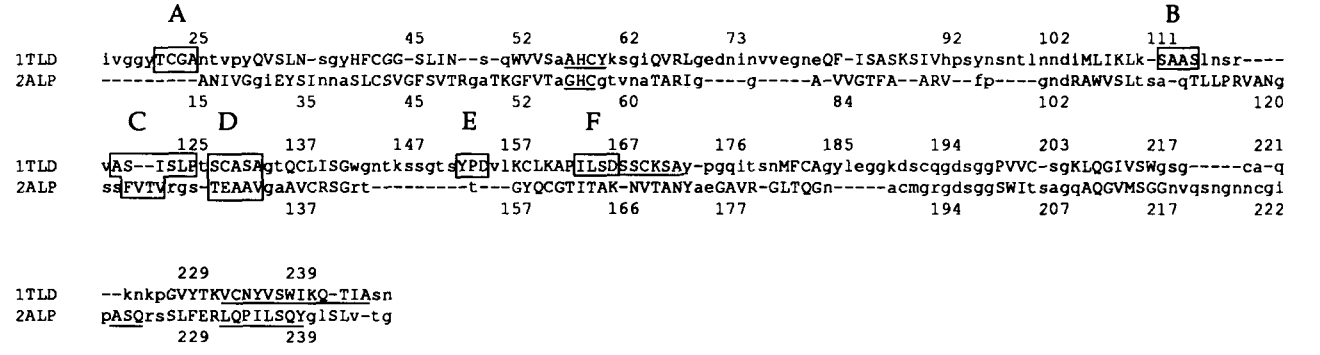
Fig. 3. A: Multiple structure alignments for the 3 families. **B:** Examples of pairwise structure alignments with low sequence identity to reference structures. Regions of regular secondary structure (α, β, and PPII) are shown in uppercase, polyproline II helices are boxed, and α-helices are underscored. In (A), PPII-helices form blocks of conserved structures. In (B), the PPII-helices responsible for important function, i.e., interdomain links, are conserved. (*Figure continues on facing page.*)

Immunoglobulin constant domains. Sequence identity 31-19 %.



B

1TLD - 2ALP. Sequence identity 17%.



3APP - 3HVP. Sequence identity 16%.

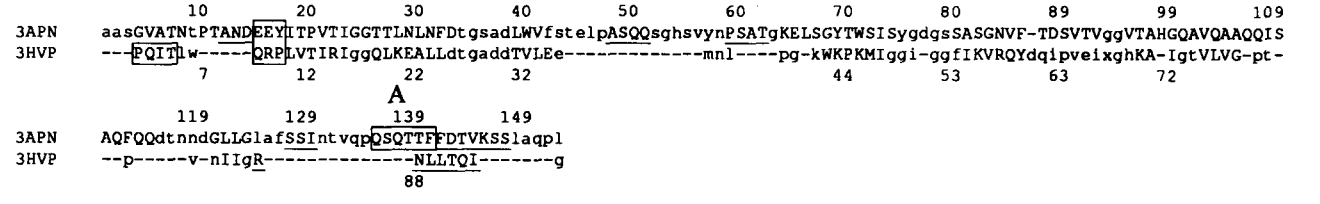


Fig. 3. *Continued.*

The results of calculations of both direct (Table 2) and relative RMSC (Fig. 6), as well as the mean RMSC and RMST shown in Figure 7, suggest that local conformational deviations in the conserved PPII-helices fall within the range observed for other secondary structures. This similar range of conformational distortions is readily identifiable for the PPII, α and β in Figure 6, where the local RMSC data is plotted for the 3 analyzed families. In the subset of immunoglobulin constant domains the PPII-helices have lower local structure deviations compared to the β-sheets (see Fig. 7).

Overall, these results confirm our previous conclusions of the similar degree of conformational stability in the PPII, α, and β. The relative structure deviations stayed at close levels for all 3 structure classes, with no significant difference for the α-helices. However, the low level of the PPII structure deviations in Ig constant domains was present also in the relative RMSC data, which points at its highly conserved character in this family.

It should be noted that although local RMS deviations in Table 2 and Figure 6 represent comparisons of the conserved secondary structure elements, their values display a wide spread for the molecules with sequence identities below 40%. A considerably smaller degree of local distortions is observed in the conserved secondary structures for sequence identity levels above 40% (see Fig. 6). Therefore, it could be suggested that relative local conformational stabilization of conserved secondary structures is only reached at the level of sequence conservation of 40% and higher.

Thus, an extensive comparison of the conservation levels and the RMS deviation patterns yielded similar results for the PPII-helices and other regular secondary structures in 3 families. Generally, the pattern of their conservation in homologous

Table 2. *Conservation of the PPII-helices calculated from the pairwise alignments, and local RMS deviations for the regular secondary structures*[a]

Protein family PDB code	Sequence identity (%)	RMS (Å)	Aligned residues[b] (%)	PPII conserved (P^{cons} %)	PPII segments conservation[c]	PPII RMSC (Å)	PPII RSMT (deg.)	α RSMC (Å)	α RSMT (deg.)	β RSMC (Å)	β RSMT (deg.)
Serine proteinases											
1tld-4cha	46.08	1.170	97.0	100.0	A B C D E F	0.119	22.08	0.131	11.58	0.165	14.69
1tld-1ton	42.86	1.394	97.0	83.0	A b′ B C D E f	0.155	18.54	0.106	13.85	0.191	19.97
1tld-2pka	40.54	1.260	99.0	100.0	A B C D E F	0.172	18.85	0.195	13.70	0.211	20.63
1tld-3est	38.46	1.140	99.0	100.0	A B C D E F	0.311	22.09	0.063	08.27	0.181	16.92
1tld-1sgt	35.55	1.553	95.0	90.0	A B C d E F	0.308	38.25	0.127	09.35	0.204	19.47
1tld-1hne	34.13	1.271	93.0	80.0	A B C D e f	0.276	25.59	0.176	17.07	0.153	15.98
1tld-3rp2	33.95	1.215	96.0	100.0	A B C D E F	0.148	15.01	0.086	10.27	0.232	20.46
1tld-2sga	19.25	4.098	72.0	50.0	[e]	[f]	[f]	0.216	37.40	0.431	40.88
1tld-2alp	17.24	5.319	78.0	50.0	a b C D e f	[f]	[f]	0.175	33.47	0.499	43.66
Aspartic proteinases											
3app-4ape	54.69	1.535	99.0	86.0	A B c′ C	0.227	15.12	0.123	17.28	0.196	22.45
3app-2apr	40.89	1.906	97.0	86.0	A b′ B C	0.236	27.68	0.093	14.83	0.228	21.63
3app-5pep	32.24	1.979	94.0	100.0	A B C	0.271	36.48	0.160	19.04	0.266	28.70
3app-1cms	28.80	1.887	96.0	100.0	A B C	0.264	26.89	0.137	13.24	0.257	38.79
3app-3hvp	16.33	3.151	56.0	50.0[d]	[e]	[f]	[f]	0.200	32.06	0.472	56.64
IG constant domains											
1fc1(CH3)-2fb4(CL)	31.31	1.671	97.0	100.0	A B C	0.220	17.32	0.202	16.94	0.174	18.81
1fc1(CH3)-2fb4(CH1)	29.17	1.751	94.0	100.0	A B C	0.264	25.17	[f]	[f]	0.301	28.90
1fc1(CH3)-2fbj(CL)	28.71	1.713	99.0	100.0	A B C	0.106	14.56	0.208	13.06	0.230	22.00
1fc1(CH3)-2fbj(CH1)	24.21	2.202	93.0	86.0	A B C e′	0.206	33.87	[f]	[f]	0.306	31.56
1fc1(CH3)-1fc1(CH2)	23.16	1.510	93.0	80.0	B C d′	0.157	16.29	0.163	17.17	0.323	29.39
1fc1(CH3)-1hsa(beta2)	22.68	2.302	95.9	100.0	A B C	0.132	28.37	[f]	[f]	0.280	29.42
1fc1(CH3)-1hsa(alpha3)	18.95	1.897	93.0	80.0	A b C	0.195	22.75	[f]	[f]	0.210	21.25

[a] Here and in figures: RMS, RMS deviation for structurally aligned molecules; aligned residues, number of structurally equivalenced residues; PPII, left-handed polyproline II helices; α, α-helices; β, β-strands; RMSC, local RMS deviation in Cartesian space; RMST, local RMS deviation in ϕ,ψ torsional angle space.

[b] Proportion of the structurally aligned residues in the reference structure.

[c] As identified in Figure 3, conserved segments are shown in uppercase, nonconserved segments are indicated with lowercase.

[d] Including 3-residue PPII-helices.

[e] Not shown in Figure 3.

[f] No alignment of secondary structure type possible.

structures follows the same rules as do the α-helices and the β-strands.

A detailed description of the PPII-helices occupying structurally similar positions in the analyzed families is given in the following sections.

Specific protein families

Serine proteinases

The PPII-helices in bovine trypsin, used as the reference structure, are mainly positioned at the molecule surface and form exposed structural elements (Fig. 8A; Kinemage 1). This seems to be the main characteristic feature of the PPII-helix A (Thr 21–Ala 24, Fig. 3), lying close to the N-terminus and separated from other regular structure segments. This is also true for the short PPII-helix E (Tyr 151–Asp 153) that forms a part of the exposed region between β-strands 1 and 2 in the second domain of the molecule. In addition to this general exposed location, certain PPII-helices can be associated with a specific structural role. PPII-helix F (Ile 162–Asp 165), which is also exposed, serves as a connecting segment between the β-strand 1 and the α-helix in the second domain. It thus participates in a supersecondary structural element β-PPII-α, the unusual aspect of which is the transition of a left-handed PPII-helix to a right-handed α-helix at the point of an overlapping residue. This type of supersecondary structure was identified as commonly found for the PPII-helices (Adzhubei & Sternberg, 1993). A different role of the PPII-helices in the structure of trypsin is the formation, with some interruptions, of the whole block connecting 2 domains in the molecule. An inspection of the relative orientation of the PPII-helices B (Ser 110–Ser 113) and following in the sequence C (Ala 119–Pro 124) shows that they lie in the same plane at the approximately 90° angle to each other provided by the right-hand turn at the bending point. The PPII-helix D Ser 127–Ala 132, which follows immediately, continues this supersecondary motif (Fig. 8A; Kinemage 1). Consequently, the interdomain block formed mainly with the PPII-helices and connecting β-strand 6 of the first domain and β-strand 1 of the second domain can be identified. The whole block lies closely to the molecule surface and has high degree of exposure. Looking at the positions of the PPII-helices relative to the active site one can

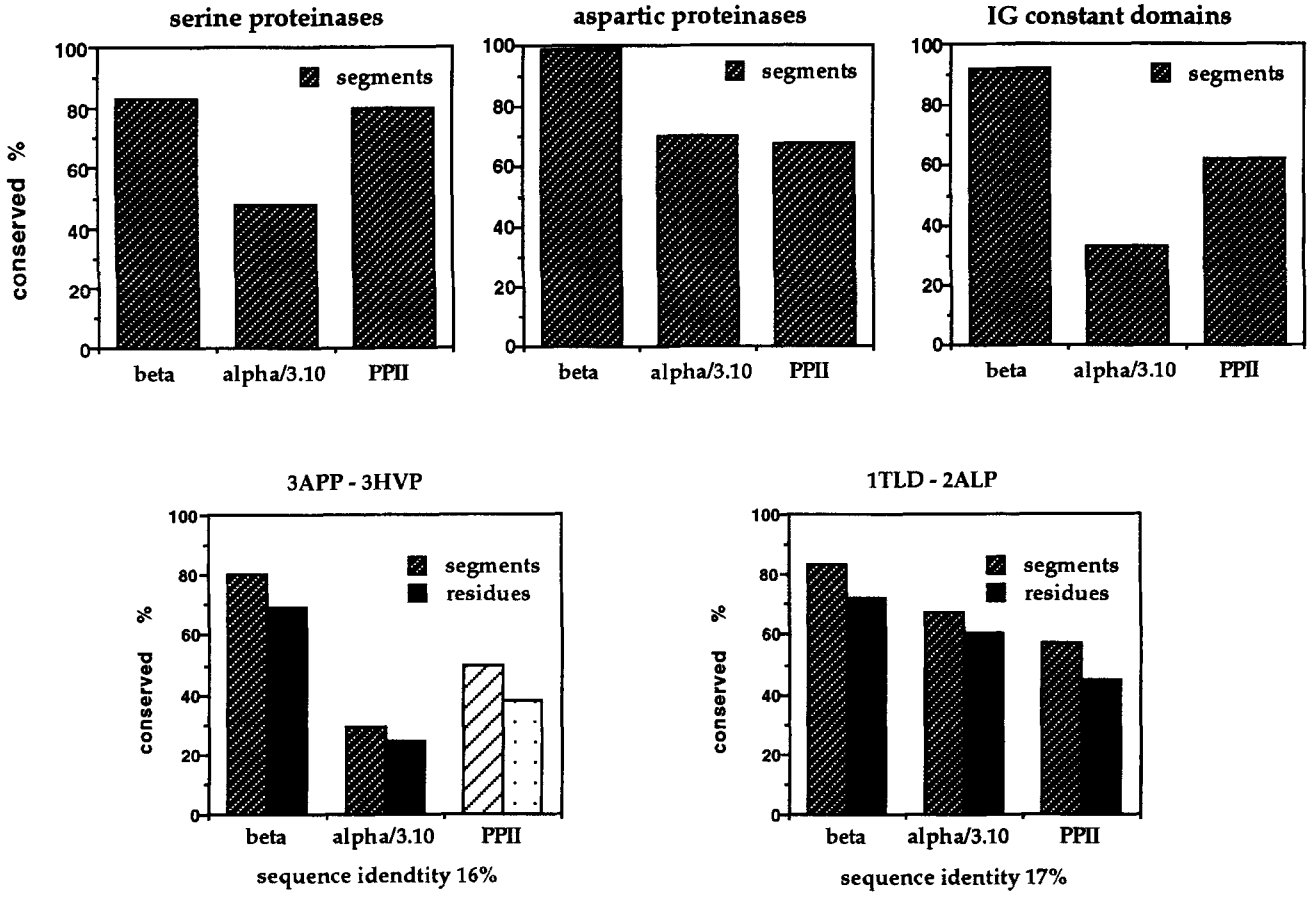


Fig. 4. Levels of conservation for secondary structure classes of the α-helices, the β-strands, and the PPII-helices calculated from the multiple alignments using Equation 2 in the Methods. The degree of conservation can vary for different secondary structure classes and protein families but does not normally drop below 50%. The number of conserved PPII-helices in 3hvp includes 3-residue segments, no PPII-helices of length 4, and more residues are conserved there (see text: Specific protein families).

notice that they represent structural elements most remote to the residues of the catalytic triad. Indeed, being comparatively evenly placed on the molecule surface, PPII-helices form the first, external layer of regular structure.

From the rest of serine proteinases in the family, the 100% level of conservation of the structural motifs involving PPII-helices is found for α-chymotrypsin (4cha), tonin (1ton), kallikrein A (2pka), porcine elastase (3est), and proteinase 2 (3rp2). This does not include the PPII-helix Ile 6–Val 9, located in chain A of α-chymotrypsin absent in other molecules (see Fig. 3). There is less, but still substantial conservation for human neutrophil elastase (1hne) and *Streptomyces griseus* trypsin (1sgt) (Fig. 3). One particular difference of 4cha, 3est, 1hne, and 3rp2 from the structural features described for 1tld is a longer PPII-helix C in the interdomain motif and the presence of a loop connecting it with PPII-helix D of the interdomain block. In 1ton and 2pka however the length and orientation of PPII-helices is identical to 1tld. Length of the PPII-helix in β-PPII-α motif can vary for different molecules, depending on the length of the β-strand in a particular structure. In 1ton, the β-PPII-α motif is formed with a distorted 3_{10}-helix rather than an α-helix and an additional PPII-helix appears between residues Tyr 94–Leu 95B, which is not found in other molecules. This is most probably due to differences in structure caused by a chain break between residues Leu 95B and Pro 95K in 1ton. An interesting feature of the PPII-helices A', A, and E in α-chymotrypsin is that they occur at the start of each of the 3 chains. This could be compared with PPII-helix A in other structures that also lies close to the N-terminus.

In trypsin 1sgt, with its 35% sequence identity with 1tld, PPII-helix D in the interdomain block is not conserved. The chain, though, forms 1 turn of a distorted left-handed helix, which, for hypothetical modeling purposes, could be approximated by a PPII-helix. Two proteins that have shorter polypeptide chains, with structure, as well as sequence being distinctly different from 1tld, are proteinase A (2sga) and α-litic proteinase (2alp). The majority of PPII-helices, like many other structural features of 1tld, are not conserved in 2alp. The domain-linking PPII segments are however conserved (Fig. 3), although they are distorted. The situation is similar for 2sga. This probably points at the structurally most important location of the PPII helical segments, performing a common function of linking structural domains, retained across the family.

Thus, the interdomain structural block in serine proteinases, formed by PPII-helices, displays a high level of conservation for molecules with sequence identities ranging from 46 to 35%. Al-

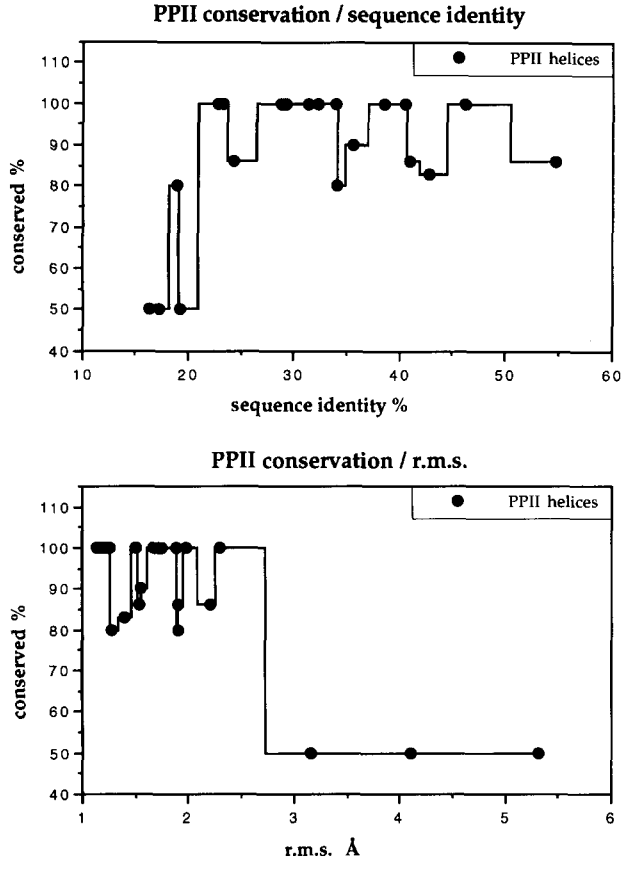


Fig. 5. Overall conservation of the PPII-helices in the dataset incorporating 3 families calculated from the pairwise alignments. The average proportion of conserved blocks reaches its plateau, with fluctuations from 80 to 100%, for the levels of sequence identity above 25% and RMS deviations below 3.0 Å.

though the rest of the PPII-helices are also conserved for high sequence identity levels, the situation becomes less predictable when it drops to 35–30%. This is associated with high divergence of the local RMS for this sequence identity level, as shown in Figure 6. Proteinase 2 at 34% identity with trypsin displays a 100% conservation of all PPII-helices. The conformation of the chain of human neutrophil elastase, at the same level of sequence identity, deviates from trypsin especially in the exposed regions of structure. Unlike in 1tld, there is no α-helix corresponding to the β-PPII-α motif and the relative PPII-helix is also not conserved. The PPII-helix E is distorted and is not included in the set of identified PPII segments (Fig. 3). However, if less rigorous criteria are applied the left-handed helical structure of this segment should be considered as conserved.

In summary, the PPII-helices are conserved in serine proteinases for the range of sequence identity 46–34%, although some of them may be distorted at its lower levels. PPII-helices that are not conserved in this identity interval mainly participate in supersecondary structure formations and are absent from the structure as a part of the nonconserved supersecondary block. At low levels of sequence identity, less than 30%, and the dissimilarity of structures associated with it, PPII-helices performing vital functions, mainly domain linkage, are conserved.

Aspartic proteinases

An important structural feature of the PPII-helices in aspartic proteinases is that nearly all of them are incorporated in the standard highly conserved supersecondary motif formed by a short 3_{10}-helix, an intermediate segment of 1–4 residues, and a PPII-helix. The proper hydrogen bonding in 3_{10}-helices might not be formed in particular structures, but their geometry, if only with minor distortions, is always retained. The result of occurrence of such a motif in a structure is chain reversal. This pattern of recurring motifs is clearly manifested in penicillopepsin (3app), which was chosen as a reference structure for the family (Fig. 3). The first 3_{10}-PPII motif, formed with PPII-helix A (Gln 135–Phe 140), connects a β-strand and an α-helix in the first domain of penicillopepsin. The most important is probably the second 3_{10}-PPII motif connecting 2 β-strands and formed by PPII-helix B (Gly 177–Thr 180). It serves as a link between the 2 domains in the molecule, the function similar to that of the interdomain motif in serine proteinases, also formed by PPII-helices. The motif linking 2 domains is conserved across the whole subset of aspartic proteinases. The third standard motif,

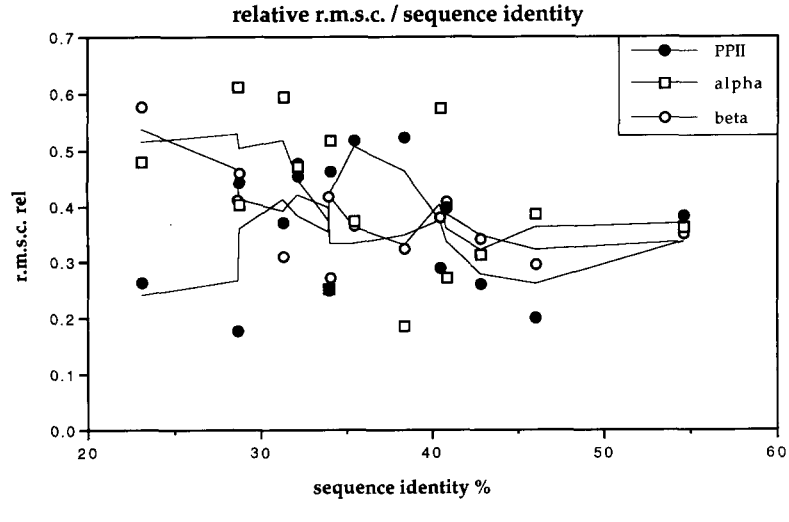


Fig. 6. Range of relative local RMS deviations in the conserved α-helices, β-strands, and PPII-helices calculated for all analyzed structures. RMSC rel, relative local RMS deviations in Cartesian space. The distribution for PPII is similar to α and β. Relative stabilization of local conformations of the conserved blocks of secondary structures, with a much lower range of distortions, is observed for sequence identities above 40%.

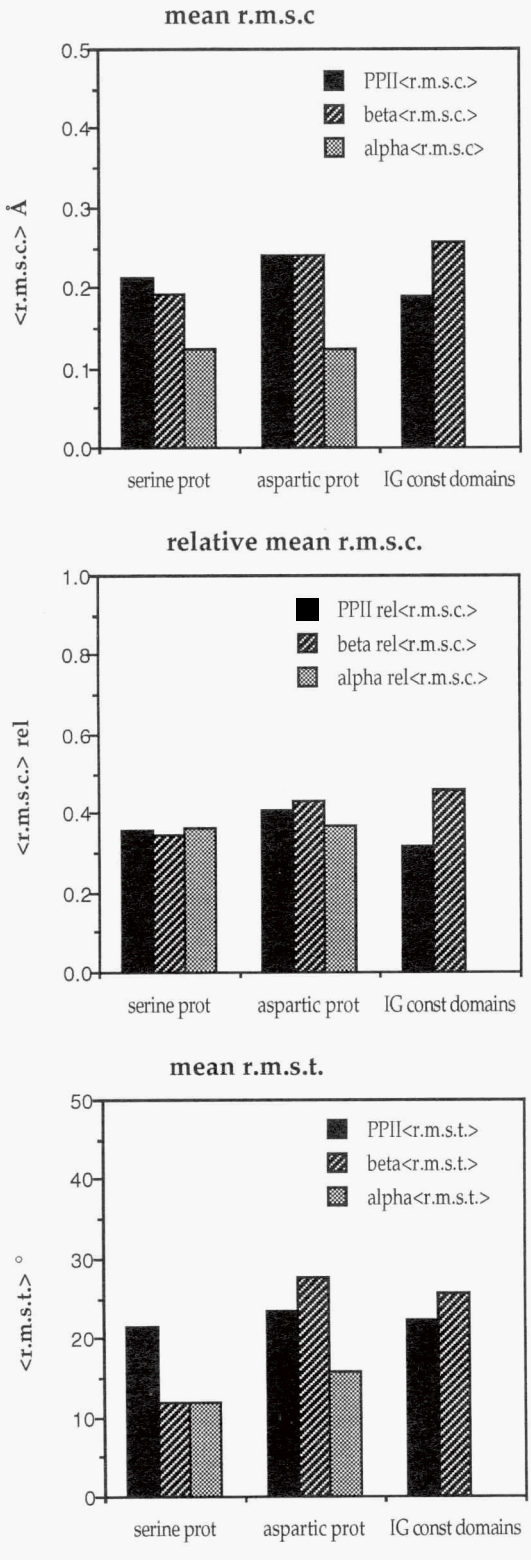


Fig. 7. Mean local RMS in the secondary structure classes α, β, and PPII, for the 3 families. The comparison of the ⟨RMSC⟩ and the relative ⟨RMSC⟩ values shows similar degree of structural divergence in the conserved blocks for all 3 secondary structure classes. A lower level of relative ⟨RMSC⟩ for PPII in Ig constant domains suggests fewer structural deviations in this class compared to β-strands. In serine proteinases, the ⟨RMST⟩ for PPII is higher compared to α and β, pointing at a higher conformational dissimilarity. Even so, it still stays at the level occupied by both β and PPII in the other 2 families.

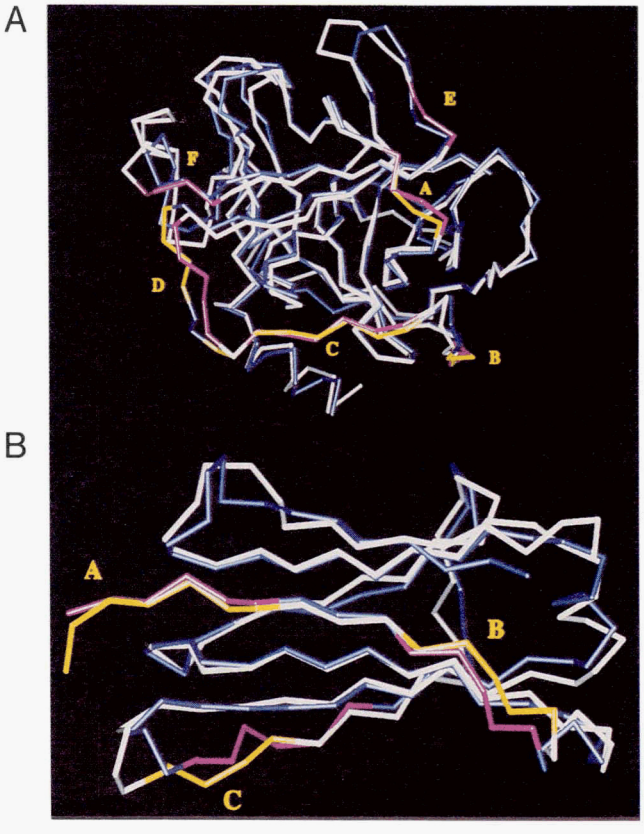


Fig. 8. Structural alignments. The aligned molecules shown represent low levels of sequence identity, with substantial structural deviations. The reference structures are shown in white, with the PPII-helices in magenta. The aligned structures are in blue with the PPII-helices shown in yellow. PPII-helices are labeled according to the notation in Figure 3. **A:** 1tld (white)–1hne (blue), 34% sequence identity. The interdomain motif formed with the PPII-helices B, C, and D is shown; 80% of the PPII-helices are conserved. The structural deviation in this pair is higher than for other members of the family of serine proteinases. The loop flanked by the PPII-helices C and D was modeled in 1hne from the parent segment in 1tld. **B:** 1fc1/CH3 (white)–2fbj/CL (blue), sequence identity 29%. The PPII-helix A in 1fc1/CH3 forms an interdomain link corresponding to the PPII-helical switch peptide 2fbj/CL. The PPII-helices are highly conserved in immunoglobulins and here the conservation is 100%. Note the PPII-helix C in 2fbj/CL, which is 2 residues shorter than in 1fc1/CH3. Color images produced using program PREPI by Dr. S. Islam, ICRF.

which includes PPII-helix D (Leu 253–Phe 256), forms a connection between the 2 β-strands in the second domain.

The 3_{10}-PPII motifs are conserved for all structures included in the family (see Fig. 3), with some differences that do not affect the overall shape of the motifs. The length of intermediate segments as well as the PPII-helices can vary for the first and the third occurrence of 3_{10}-PPII motif. The motif linking 2 domains is most conserved, with the length of the PPII-helix and the intermediate segment identical for all structures. The relative location of PPII-helices is on the surface of domains, in symmetric positions respective to the active site. It is possible that an additional degree of flexibility, apart from that provided by the PPII-helix in the interdomain link, is imparted by the PPII-helices in each domain.

There are, however, 2 PPII-helices, the Leu 158–Ala 161 in rhizopuspepsin (2apr) and the Gly 202–Lys 204 in endothiapep-

sin (4ape), that are not retained in other structures. The PPII-helix in 2apr participates in a β-PPII-α motif that is not formed in other structures. In 4ape, with its longer chain, the additional PPII-helix also does not have an equivalenced segment in other structures.

Because the active enzyme of HIV protease (3hvp) is formed by 2 molecules, its structure was aligned with the N-domain of penicillopepsin (3app). The RMS deviation of the 2 aligned structures, with sequence identity at 16%, is high and reaches 3.15 Å. Although main structural features similar to that of 3app can be traced in 3hvp, some of the structural elements are missing in its shorter chain (Fig. 3). Several α-helices and β-strands are not conserved, and the details of relative orientation of other β-strands are different. No structural region in HIV protease could be aligned to the part of the structure of penicillopepsin, which includes the PPII-helix of the first domain. It should be noted however that the 2 PPII-helices positioned immediately at the N-terminus of 3hvp will serve as the interface between 2 molecules of the active enzyme. They thus mimic the PPII helical interdomain link in the rest of aspartic proteinases. Hence even though no direct structural similarity could be found between the N-terminal PPII-helices in HIV protease and the interdomain PPII segment in penicillopepsin, the topological and functional similarity is clear. This fact could prove valuable for homology modeling.

Thus, the PPII-helices in aspartic proteinases participate in 3_{10}-PPII supersecondary motifs, which are conserved across the family. The 2 PPII-helices that are not conserved are located in structural blocks dissimilar with the corresponding parts of homologous structures. One of these PPII-helices is located in a nonconserved supersecondary motif. Although at the sequence identity level below 20%, in 3hvp, only 1 short PPII-helix at the N-terminus is directly conserved, the PPII conformation of the linking region is retained.

Immunoglobulin constant domains

The relative location of PPII-helices in Ig constant domains follows the pattern from previous results. Firstly, the PPII-helices form interdomain links. The structure of a curved PPII-helix is adopted by switch peptides connecting variable and constant domains in FAB fragments (Fig. 9A). The interdomain link CH2-CH3 in the FC fragment is also formed by 2 PPII-helices (Fig. 9C). The second common feature is the location of PPII-helices on the domain surface, where they participate in the first layer of regular structures.

The CH3 domain of the FC fragment of human immunoglobulin IgG1 (1fc1) serves as a reference structure for the family. Because some residues are missing at its C-terminus, no structure identification was performed for this part of the domain. Fifteen residues are also missing at the N-terminus of the CH2 domain of 1fc1, and this region of CH2 was excluded from structure comparisons.

The curved PPII-helix A (Gln 342–Gln 347, Fig. 3), located at the N-terminus, corresponds to the PPII-helical switch peptides in FABs. PPII-helix B (Leu 351–Arg 355), following closely to the first one, connects a short β-strand and a 3_{10}-helix, thus forming a β-PPII-3_{10}-supersecondary element. Together these 2 PPII-helices span along the domain surface forming a flexible interdomain link (see Fig. 8B and Kinemage 2). The long PPII-helix C (Thr 393–Asp 399) also lies on the surface at the same side of the domain as the first 2 helices, at the approximately 30° angle to them. PPII-helix C is immediately followed by a reverse turn.

PPII-helices A, B, and C are conserved practically in all structures, even in those with low levels of sequence identity close to 20% (see Fig. 3). The exception is the α3 domain of 1hsa, at 19% sequence identity with reference structure, where PPII-helix B is not conserved. In the α3 and the β2 domains of 1hsa, the

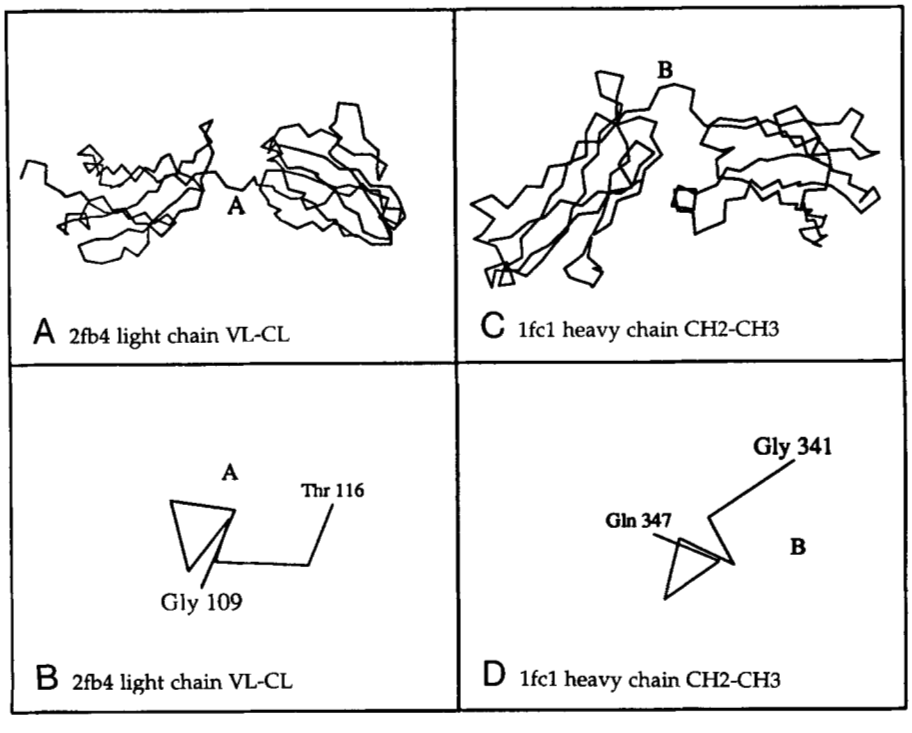


Fig. 9. PPII-helices as interdomain structure in immunoglobulins. **A:** Switch peptide formed by the PPII-helix (A) in the structure of IgG FAB fragment 2fb4. **B:** The curved PPII-helix A in the switch peptide that serves as a domain-domain link in 2fb4 (VL-CL). **C:** The PPII-helix (B) forming an interdomain link in the IgG1 FC fragment 1fc1. **D:** The 2 adjacent PPII-helices B of the interdomain link in 1fc1 (CH2-CH3). Diagrams were prepared by MOLSCRIPT (Kraulis, 1991).

right-handed 3_{10}- or α-helix normally following the PPII-helix B in Ig constant domains is absent, and the topology of supersecondary structure of this part of the domains is different. With the decrease of sequence identity for the $\alpha 3$ domain, this results in the absence of the corresponding PPII-helix.

Compared to the CH3 domain of 1fc1, the interdomain PPII-helix A forms a longer structure of switch peptide in CL domains of FAB fragments. In the CH1 domains of human Ig FAB (2fb4) and mouse IgA FAB (2fbj), the PPII structure of switch peptides is longer and more curved. PPII-helix C is represented by a short distorted structure in the CH2 domain of 1fc1, and in the CL domain of 2fbj (see Fig. 8B and Kinemage 2) it is shorter compared to other domains.

An additional PPII-helix, Asn 209–Gln 212, not formed in other structures, was found in the CH1 domain of 2fbj. The topology of the chain however is conserved here, with the 2 subsequent left-handed turns in the reference structure, the CH3 domain of 1fc1, mimicking the PPII-helix in the CH1 domain of 2fbj. Thus, for modeling purposes the chain conformation could be approximated with a PPII-helix.

To summarize, PPII-helices are highly conserved in Ig constant domains, even at the lower levels of sequence identity. They are mostly found at the N- and C-termini of domains, serving as linking structures in switch peptides and in the similar peptides connecting CH2 and CH3 domains.

PPII-helices in modeling

The benefits of introducing the new class of elements in regular SCRs in proteins lie mainly in the reduction of size and number of VRs, which can therefore increase the accuracy of modeling. In practice, when a PPII-helix occurs in a region of protein structure that previously was treated as a nonconserved loop, only shorter parts of this loop will now be scanned against the database in order to find suitable candidate fragments for modeling.

The other source of improvements in modeling quality are PPII-helices themselves. When PPII-helices are treated as part of loops, the geometry of a modeled chain is likely to be misrepresented. It happens because chirality of the chain is not taken into account and the left-handed PPII-helix could be easily modeled with a right-handed α- or 3_{10}-helix.

Two examples are shown here. A loop in elastase, flanked by 2 PPII-helices, was modeled from the shorter loop in trypsin, and a loop in pepsin between a 3_{10}-helix and a PPII-helix was modeled starting from a longer loop in penicillopepsin.

In the sequence of trypsin 1tld, the conserved PPII-helices C and D are separated by 1 residue (see Fig. 3). The corresponding 7-residue segment LPAQGRR in elastase 1hne was modeled (see Fig. 8A and Kinemage 1). For a database scan, the ends of the parent loop in 1tld were fixed at residue pairs 123, 124 in PPII-helix C and 127, 128 in PPII-helix D. From the set of suitable fragments found in the database, a fragment in the immunoglobulin 2fbj H-chain that satisfied the RMS criteria and had high sequence similarity with the target sequence was fitted to the target structure and its sequence mutated (see Methods). The superposition of the modeled structure and the native loop in 1hne showed their high similarity (Fig. 10A). The RMS deviation of the C^{α} atoms of 2 segments is 0.9 Å.

As shown in Figure 3, any attempt to model the same loop in 1hne when PPII-helices are not used as conserved elements would imply the first conserved elements to be the β-strand 6 of the first domain, and the β-strand 1 of the second domain. The length of VR would be equal to 25 residues, thus ruling out the possibility of a direct search of the database for suitable model segment. We tried to model different combinations of sections of the loop but failed to produce model structures with RMS and sequence similarity comparable to the results of modeling with PPII-helices as SCRs. The introduction of PPII-helices yielded a markedly higher accuracy in modeling, enabling sharp reduction of the length of a VR.

The loop lying between 2 conserved helices in penicillopepsin 3app, the short 3_{10}-helix, Ser 127–Ile 129, and PPII-helix A (Gln 135–Phe 140), was used to model a corresponding shorter segment in pepsin 5pep (Fig. 3). For a PDB database scan, a fragment in the parent structure 3app that started from residues 127, 128 of the 3_{10}-helix and included the loop and residues 137, 138 of the PPII-helix, was taken. The fixed pair of residues at the PPII-helical end of the parent structure was shifted 1 residue along the sequence to avoid assigning PPII structural class to the glycine in the target sequence. Gly residues are shown to be highly unfavorable for the PPII-helices (Adzhubei & Sternberg, 1993). In fact, the PPII-helices are the most unfavorable secondary structure for Gly. It is thus recommended that, when assigning PPII-helices to target sequence, Gly residues should be left outside the boundaries of a PPII segment. The target sequence in 5pep, PSISASGAT (see Fig. 3), included 9 residues. The database search yielded a fragment in cytochrome *c* (1ccr) that satisfied both criteria of the RMS and sequence similarity. A part of this segment connecting the conserved 3_{10}- and PPII-helices was used to model the loop. The RMS deviation of the C^{α} atoms of superimposed native and model structures is 0.4 Å (Fig. 10B).

The other aspect of the importance of PPII structure for modeling can be seen after a subsequent attempt had been made to model the same segment in 5pep without accounting for the PPII-helix. Starting from the same 3_{10}-helix end of the parent segment, its other end was assigned to the first residue of an α-helix in 3app directly following PPII-helix A used as an SCR in the previous modeling run (see Fig. 3). The PPII-helix was treated as a variable region. The best of the resulting fragments found in the structure of influenza virus hemagglutinin 2hmg is shown in Figure 10C, superimposed with the native structure. The RMS deviation is 1.77 Å. Here, high RMS is combined with the questionable resemblance to the target structure. The most incorrectly modeled though is the PPII-helix itself: the corresponding modeled structure is a right-handed α-helix.

Discussion

An important feature of the α-helices and the β-strands in proteins is their tendency to occupy the same relative positions and retain similar length in homologous structures. This pattern, being trivial for the levels of sequence identity of 50% and higher, is formulated as the principle of conservation of main secondary structure blocks for the lower levels of sequence identity and is essential for modeling. It is apparent that an expansion of structurally conserved core, with new elements added to it, is of primary importance in the situation when the geometry of whole structural blocks can deviate sharply in the molecules under comparison. A conserved character of such secondary structures as the α-helices, the β-sheets, and the 3_{10}-helices can be

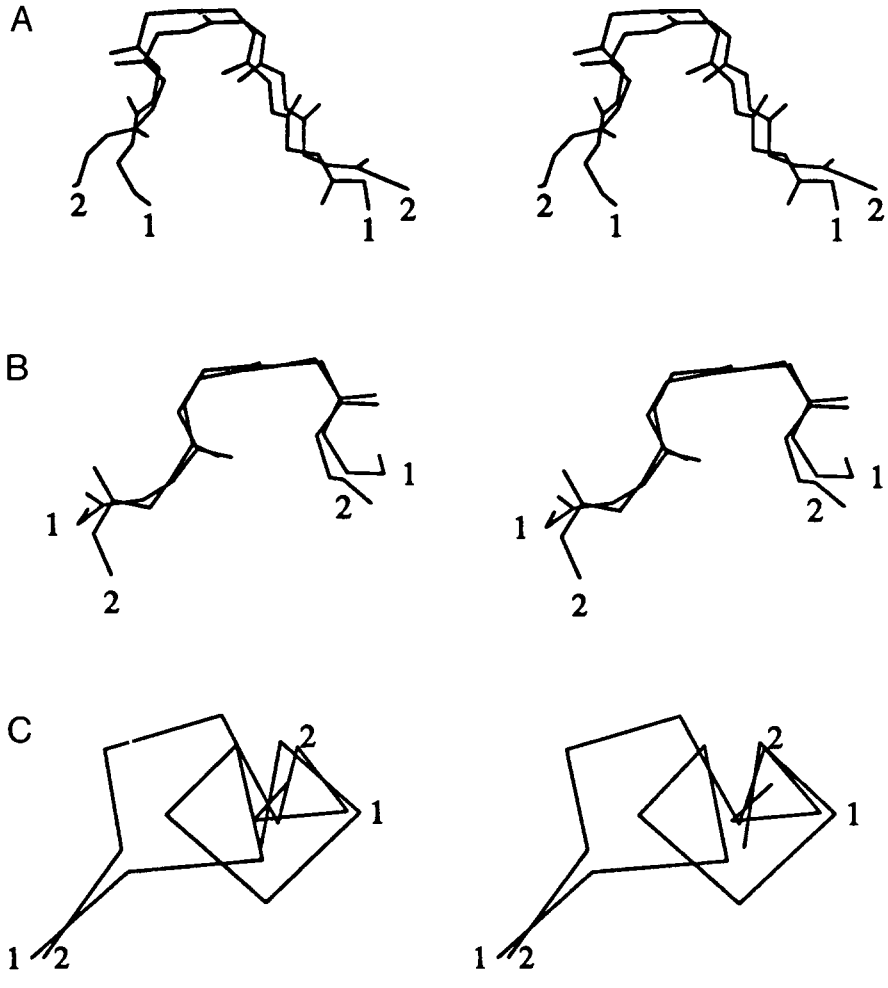


Fig. 10. PPII-helices in modeling. **A:** Superimposed structures of the modeled (1) and the native (2) loop in 1hne, RMS 0.9 Å. The model is based on the assumption that the PPII-helices flanking the loops are conserved and form SCRs. **B:** Superimposed structures of the modeled (1) and the native (2) loop in 5pep, RMS 0.4 Å. The conserved 3_{10}-helix is located at the N-terminus and the conserved PPII-helix at the C-terminus of the loop. **C:** C^{α}-tracing of the superimposed modeled (1) and native (2) segments in 5pep. Here, modeling was performed for the same loop as in (B), but the PPII-helix was not considered as a conserved structure and its conformation was not assigned to the corresponding sequence in the target segment. The RMS of the model with the native segment is high at 1.77 Å. The part of the modeled segment equivalenced with the left-handed PPII-segment in the native structure is formed by a right-handed α-helix, which makes the model inadequate. Stereo diagrams were prepared by MOLSCRIPT (Kraulis, 1991).

easily predicted: they form the spatial backbone of a molecule, thus determining the structure–function relationship. If both the function and sequence are similar, the building blocks are also likely to be similar. The PPII-helices however cannot be equalled with other regular structures in their characteristic features. More flexible, found mainly on the molecule surface, the PPII-helices probably perform quite a different role in protein structure compared to relatively rigid blocks of the α-helices and the β-sheets. PPII-helices form structural elements that can be termed *flexible blocks*, serving as connections between *building blocks* and may be capable of performing minor structural adjustments important for function.

The presence of PPII-helices as flexible structural elements is shown in this work, most importantly as the predominant structure of interdomain links in all 3 protein families analyzed here. However, it is exactly these properties of high conformational mobility that make any a priori conclusions about the conservation of the PPII-helices unreliable. Perhaps the PPII conformation of a mobile element is flexible to the extent where it would be not conserved in a homologous molecule.

The results of this work however show that in terms of conservation in evolution the PPII-helices in protein structure behave similarly to the α-helices and the β-strands. The PPII-helices are normally conserved down to the low levels of sequence identity of 30–20%. Even if at the lower end of the sequence identity range the structure is distorted, the left-handed conformation of the chain and its characteristic geometry are retained. This allows assignment of the PPII structure to segments for modeling purposes with high degree of confidence. It is noteworthy that the other tendency in conservation of the PPII-helices is associated with their role as part of supersecondary structure elements. These supersecondary elements, where a PPII-helix normally serves as a flexible link with other secondary structures, i.e., the α-helices and the β-sheets, were first identified by Adzhubei and Sternberg (1993). Their presence in protein structure is confirmed by the results of this work. When participating in a supersecondary element, a PPII-helix is conserved so long as the element as a whole is conserved. The absence of an α-helix from such supersecondary element in a homologous structure will lead to the associated PPII-helix also being absent in this structure. The explanation of this behavior probably lies in the extremely close connection formed by the 2 structures, with a 1-residue overlap of the left-handed and the right-handed helical conformations (Adzhubei & Sternberg, 1993). As the next step in the analysis of the PPII-helices, we plan to identify and classify supersecondary motifs incorporating this structure.

The conservation of PPII-helices seems also to be related to their role in the structure of a specific molecule. The PPII-helices forming key structural elements, e.g., interdomain links, are

conserved even at low levels of sequence identity, as demonstrated for all 3 protein families analyzed here.

Because the PPII-helices are mostly located on the molecule surface and do not participate in intramolecular hydrogen bonding networks, regular hydrogen bonds with water are important (see Adzhubei & Sternberg, 1993). The central role of water for maintaining the PPII-conformation was confirmed by Monte Carlo (Eisenhaber et al., 1992) and molecular dynamics (Sreerama & Woody, 1992) calculations. Because of their strong interactions with water the PPII-helices can be seen as key points for the structure of a hydrating layer of water molecules. A detailed study of the PPII–water interactions in crystal structures would without doubt provide deeper insight into their role of linking flexible blocks.

PPII-helices located on the protein surface can also serve as sites of intermolecular interactions. Their ability to form hydrogen bonds directed to the outside of the molecule can provide a flexible link between 2 structures. In addition to their role of interdomain links, the PPII-helices tend to participate in the regions connecting major structural parts of a molecule. A good illustration of this is the immunoglobulin hinge region where the PPII conformation was confirmed by X-ray (Marquart et al., 1980) and NMR (Kessler et al., 1991). Thus, in immunoglobulins the PPII-helices form connecting blocks for virtually all structural domains.

Although the 23 structures from 3 protein families we have examined represent a relatively small dataset compared to the number of available protein structures, its size reflects the analytical rather than statistical direction of this work. The abundance of PPII segments in the selected protein structures allows one to trace their conservation in different structural environments, i.e., as the part of a supersecondary element, as a conserved region in loops, etc. Because the conservation of PPII-helices remained stable, we conclude that their conservation pattern does not depend on the immediate structural environment and the structural class of a molecule. Further support for these conclusions is provided in subsequent work (Adzhubei et al., in prep.), where a conserved PPII-helix was identified in the structure of DNA-binding α-helical proteins from the family of HMG-box domains.

The analysis of conservation in evolution and the modeling results therefore suggest that the PPII-helices belong to structurally conserved regions in proteins and should be regarded as such for purposes of modeling by homology and other structural studies. The results presented here also support the importance of the PPII-helices as a secondary structure class that should be accounted for in any comprehensive secondary structure classification scheme.

Acknowledgments

We thank Professor V.G. Tumanyan and Dr. N.G. Esipova (The Engelhardt Institute of Molecular Biology, Moscow) for useful discussions. We also thank Dr. P. Bates (ICRF) for helpful discussions and the homology modeling computer package, and Dr. S. Islam (ICRF) for the graphics program PREPI.

References

Adzhubei AA, Eisenmenger F, Tumanyan VG, Zinke M, Brodzinski S, Esipova NG. 1987a. Third type of secondary structure: Noncooperative mobile conformation. Protein Data Bank analysis. *Biochem Biophys Res Commun 146*:934–938.

Adzhubei AA, Eisenmenger F, Tumanyan VG, Zinke M, Brodzinski S, Esipova NG. 1987b. Approaching a complete classification of protein secondary structure. *J Biomol Struct Dynam 5*:689–704.

Adzhubei AA, Lauton CA, Neidle S. 1994. An approach to protein modelling based on an ensemble of structures solved by NMR: A structural model for the Sox-5 HMG-box protein.

Adzhubei AA, Sternberg MJE. 1993. Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol 229*:472–493.

Ananthanarayanan VS, Soman KV, Ramakrishnan C. 1987. A novel secondary structure in globular proteins comprising the collagen-like helix and β-turn. *J Mol Biol 198*:705–709.

Arnott S, Dover SD. 1968. The structure of poly-L-proline II. *Acta Crystallogr B 24*:599–601.

Barton GJ, Sternberg MJE. 1987. Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng 1*:89–94.

Bartunik HD, Summers LJ, Bartsch HH. 1989. Crystal structure of bovine beta-trypsin at 1.5 Å resolution in a crystal form with low molecular packing density. Active site geometry, ion pairs and solvent structure. *J Mol Biol 210*:813–828.

Bates PA, Sternberg MJE. 1992. From protein sequence to structure. In: Rees AR, Sternberg MJE, Wetzel R, eds. *Protein engineering—A practical approach*. Oxford, UK: Oxford University Press. pp 117–141.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Blundell TL, Johnson MS. 1993. Catching a common fold. *Protein Sci 2*:877–883.

Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature 326*:347–352.

Bode W, Chen Z, Bartels K, Kutzbach C, Schmidt-Kastner G, Bartunik H. 1983. Refined 2 Å X-ray crystal structure of porcine pancreatic kallikrein A, a specific trypsin-like serine proteinase. Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. *J Mol Biol 164*:237–282.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J 5*:823–826.

Cooper JB, Khan G, Taylor G, Tickle IE, Blundell TL. 1990. X-ray analyses of aspartic proteases. II. Three-dimensional structure of the hexagonal crystal form of porcine pepsin at 2.3 Å resolution. *J Mol Biol 214*:199–222.

Cowan PM, McGavin S. 1955. Structure of poly-L-proline. *Nature 176*: 501–503.

Deisenhofer J. 1981. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9- and 2.8-Å resolution. *Biochemistry 20*:2361–2370.

Eisenhaber F, Adzhubei AA, Eisenmenger F, Esipova NG. 1992. Hydration of polyproline II type lefthelical conformation. Monte Carlo study. *Biophysica (Moscow) 37*:62–67.

Fermi G, Perutz MF, Shaanan B, Fourme R. 1984. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol 175*:159–174.

Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci 2*:1811–1826.

Fujinaga M, Delbaere LTJ, Brayer GD, James MNG. 1985. Refined structure of alpha-lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. *J Mol Biol 184*:479–502.

Fujinaga M, James MN. 1987. Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1.8 Å resolution. *J Mol Biol 195*:373–396.

Gilliland GL, Winbourne EL, Nachman J, Wlodaver A. 1990. The three-dimensional structure of recombinant bovine chymosin at 2.3 Å resolution. *Proteins Struct Funct Genet 8*:82–101.

Greer J. 1981. Comparative model-building of the mammalian serine proteases. *J Mol Biol 153*:1027–1042.

Greer J. 1990. Comparative modelling methods: Application to the family of the mammalian serine proteases. *Proteins Struct Funct Genet 7*: 317–334.

Hilbert M, Bohm GRJ. 1993. Structural relationships of homologous proteins as a fundamental principle in homology modelling. *Proteins Struct Funct Genet 17*:138–151.

Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. 1992. A database of protein structure families with common folding motifs. *Protein Sci 1*:1691–1698.

James MNG, Sielecki AR. 1983. Structure and refinement of penicillopepsin at 1.8 Å resolution. *J Mol Biol 163*:299–361.

Jones TA, Thirup S. 1986. Using known substructures in protein model building and crystallography. *EMBO J 5*:819–822.

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

Kessler H, Mronga S, Muller G, Moroder L, Huber R. 1991. Conformational analysis of a IgG1 hinge peptide derivative in solution determined by NMR spectroscopy and refined by restrained molecular dynamics simulations. *Biopolymers 31*:1189–1204.

Kraulis J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol 136*:225–270.

Lim WA, Richards FM. 1994. Critical residues in an SH3 domain from Sem-5 suggest a mechanism for proline-rich peptide recognition. *Nature Struct Biol 1*:221–225.

Madden DR, Gorga JC, Strominger JL, Wiley DC. 1992. The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. *Cell 70*:1035–1048.

Makarov AA, Lobachov VM, Adzhubei IA, Esipova NG. 1992. Natural polypeptides in left-handed helical conformation. *FEBS Lett 306*:63–65.

Marquart M, Deisenhofer J, Huber R, Palm W. 1980. Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 Å and 1.0 Å resolution. *J Mol Biol 141*:369–391.

Mauguen Y, Hartley RW, Dodson EJ, Dodson GG, Bricogne G, Jack A. 1982. Molecular structure of a new family of ribonucleases. *Nature 297*:162–164.

McLachlan AD. 1972. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr A 28*:656.

Meyer E, Cole G, Radhakrishnan R. 1988. Structure of native porcine pancreatic elastase at 1.65 Å resolution. *Acta Crystallogr B 44*:26–38.

Miller M, Schneider J, Sathyanarayana BK, Toth MV, Marshall GR, Clawson L, Selk L, Kent SB, Wlodawer A. 1989. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. *Science 246*:1149–1152.

Moult J, Sussman F, James MNG. 1985. Electron density calculations as an extension of protein structure refinement. *Streptomyces griseus* protease at 1.5 Å resolution. *J Mol Biol 182*:555–566.

Navia MA, McKeever BM, Springer JP, Lin TY, Williams HR, Fluder EM, Dorn CP, Hoogsteen K. 1989. Structure of human neutrophil elastase in complex with a peptide chloromethyl ketone inhibitor at 1.84 Å resolution. *Proc Natl Acad Sci USA 86*:7–11.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.

Pearl L, Blundell TL. 1984. The active site of aspartic proteinases. *FEBS Lett 174*:96–101.

Perutz MF, Miurhead H, Cox JM, Goaman LC, Mathews FS, McGandy EL, Webb LE. 1968. Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 Å resolution: (1) X-ray analysis. *Nature 219*:29–32.

Pickett SD, Saqi MAS, Sternberg MJE. 1992. Evaluation of the sequence template method for protein structure prediction. *J Mol Biol 228*:170–187.

Read RJ, James MN. 1988. Refined crystal structure of *Streptomyces griseus* trypsin at 1.7 Å resolution. *J Mol Biol 200*:523–551.

Remington SJ, Woodbury RG, Reynolds RA, Matthews B, Neurath H. 1988. The structure of rat mast cell protease II at 1.9 Å resolution. *Biochemistry 27*:8097–8105.

Richards FM, Kundrot CE. 1988. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins Struct Funct Genet 3*:71–84.

Richardson JS, Richardson DC. 1989. Principles and patterns of protein conformation. In: Fasman GD, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 1–98.

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

Siligardi G, Drake AF, Maskagni P, Rowlands D, Brown F, Gibbons WA. 1991. Correlations between the conformations elucidated by CD spectroscopy and the antigenic properties of four peptides of the foot-and-mouth disease virus. *Eur J Biochem 199*:545–551.

Sklenar H, Etchebest C, Lavery R. 1989. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins Struct Funct Genet 6*:46–60.

Sreerama N, Woody RW. 1992. Molecular dynamics simulations of polypeptide conformations in water: A comparison of α, β and PII conformations. *Biophys J 61*:A462.

Sreerama N, Woody RW. 1994. Poly(Pro)II helices in globular proteins: Identification and circular dichroic analysis. *Biochemistry 33*:10022–10025.

Subramanian E, Swan I, Liu M, Davies DR, Jenkins JA, Tickle IJ, Blundell TL. 1977. Homology among acid proteases: Comparison of crystal structures at 3 Å resolution of acid proteases from *Rhizopus chinensis* and *Endothia parasitica*. *Proc Natl Acad Sci USA 74*:556–559.

Suguna K, Bott RR, Padlan EA, Subramanian E, Sheriff S, Cohen GD, Davies DR. 1987. Structure and refinement at 1.8 Å resolution of the aspartic proteinase from *Rhizopus chinensis*. *J Mol Biol 196*:877–900.

Suh SW, Bhat TN, Navia MA, Cohen GH, Rao DN, Rudikoff S, Davies DR. 1986. The galactan-binding immunoglobulin FAB J539: An X-ray diffraction study at 2.6-Å resolution. *Proteins Struct Funct Genet 1*:74–80.

Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987a. Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng 1*:377–384.

Sutcliffe MJ, Hayes F, Blundell TL. 1987b. Knowledge based modelling of homologous proteins, part II: Rules for the conformation of substituted sidechains. *Protein Eng 1*:385–392.

Taylor WR, Orengo CA. 1989a. A holistic approach to protein structure alignment. *Protein Eng 2*:505–519.

Taylor WR, Orengo CA. 1989b. Protein structure alignment. *J Mol Biol 208*:1–22.

Tsukada H, Blow DM. 1985. Structure of α-chymotrypsin refined at 1.68 Å resolution. *J Mol Biol 184*:703–711.

Woody RW. 1992. Circular dichroism and conformation of unordered polypeptides. *Adv Biophys Chem 2*:37–79.

Yee DP, Dill KA. 1993. Families and the structural relatedness among globular proteins. *Protein Sci 2*:884–899.

Yu H, Chen JK, Feng S, Dalgarno DC, Brauer AW, Schreiber SL. 1994. Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell 76*:933–945.