Protein Science (1999), *8*:1733–1742. Cambridge University Press. Printed in the USA.

Stable proline box motif at the N-terminal end of α-helices

A.R. VIGUERA[1] AND L. SERRANO
EMBL, Meyerhofstrasse 1, Heidelberg 69117, Germany

(RECEIVED October 26, 1998; ACCEPTED May 13, 1999)

Abstract

We describe a novel N-terminal α-helix local motif that involves three hydrophobic residues and a Pro residue (*Pro-box motif*). Database analysis shows that when Pro is the N-cap of an α-helix the distribution of amino acids in adjacent positions changes dramatically with respect to the average distribution in an α-helix, but not when Pro is at position N1. N-cap Pro residues are usually associated to Ile and Leu, at position N′, Val at position N3 and a hydrophobic residue (h) at position N4. The side chain of the N-cap Pro packs against Val, while the hydrophobic residues at positions N′ and N4 make favorable interactions. To analyze the role of this putative motif (sequence fingerprint hPXXhh), we have synthesized a series of peptides and analyzed them by circular dichroism (CD) and NMR. We find that this motif is formed in peptides, and that the accompanying hydrophobic interactions contribute up to 1.2 kcal/mol to helix stability. The fact that some of the residues in this fingerprint are not good N-cap and helix formers results in a small overall stabilization of the α-helix with respect to other peptides having Gly as the N-cap and Ala at N3 and N4. This suggests that the *Pro-box* motif will not specially contribute to protein stability but to the specificity of its fold. In fact, 80% of the sequences that contain the fingerprint sequence in the protein database are adopting the described structural motif, and in none of them is the helix extended to place Pro at the more favorable N1 position.

Keywords: α-helix capping; proline; proline–valine interaction

Protein design from scratch has often found as a mayor difficulty the simultaneous stabilization of several conformations, giving rise to disperse and inhomogeneous set of species. These can have a similar secondary structure content but different packing and tertiary contacts (*molten globules*). It is, therefore, crucial in any design exercise to consider alternative folded conformations and to stabilize specifically the desired one (Shakhnovich, 1998). Productive strategies would imply increasing not only the stability of the target structure, but also increasing the energy gap between this and other possible conformations. The same reasoning is valid for structure prediction. Finding sequence motifs that determine the local structure of the polypeptide chain could be helpful in this respect.

Much is already known about the rules that govern helix formation, and it is possible to predict with reasonable accuracy the helical content of monomeric sequences in the absence of tertiary interactions (Muñoz & Serrano, 1995c). However, not all interactions between amino acids have been already determined, and more interesting, there is a large number of local motifs at the ends of α-helices that have been described and need to be checked experimentally (Aurora & Rose, 1998). A strategy to find new local motifs is to identify combinations of amino acids in the protein structure database that are found with higher frequency than expected from a random distribution (Harper & Rose, 1993; Muñoz et al., 1995; Viguera & Serrano, 1995a, 1995b; Prieto & Serrano, 1997). To eliminate context effects and to precisely calculate the energy associated to a given interaction, the sequence fingerprint obtained from the statistical analysis of the protein database must be studied isolated from the rest of the protein. One possibility is to introduce the designed sequence in a host polyalanine-based peptide. Changes in the helical content of the peptides analyzed by a helix/coil transition algorithm allow to assign energies to given interactions.

There are several local motifs at α-helix ends involving two or more residues that have been identified and experimentally characterized. At the N-terminus the *Capping box* (Harper & Rose, 1993; Lyu et al., 1993), the *Hydrophobic staple* (Seale et al., 1994; Muñoz et al., 1995; Muñoz & Serrano, 1995a), and vari-

Reprint requests to: Luis Serrano, EMBL Structural Biology Program, Meyerhofstrasse 1, Heidelberg 69117, Germany.
[1]Present address: Dpto. Bioquímica y Biología Molecular, Unidad de Biofísica, U.P.V., Bilbao, Spain.

Abbreviations: 2D, two-dimensional; CD, circular dichroism; COSY, 2D scalar correlated spectroscopy; DQF-COSY, double-quantum filter-COSY; HPLC, high-performance liquid chromatography; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser enhancement spectroscopy; ppm, parts per million; ROESY, rotating frame Overhauser effect spectroscopy; RMSD, root-mean-square deviation; TFE, trifluoroethanol; TOCSY, total correlation spectroscopy; TSP, sodium 3-trimethylsylyl (2,2,3,3-2H4)propionate.

ants of those (Lacroix et al., 1998). At the C-terminus the *Schellman* (Schellman, 1980; Milner-White, 1988; Aurora et al., 1994; Viguera & Serrano, 1995a) and the *Pro-capping* (Prieto & Serrano, 1997) motifs. Other local motifs have been described recently (Aurora & Rose, 1998; Penel et al., 1999), but so far, not analyzed experimentally.

Throughout the following text we will define the α-helix positions as previously indicated (Aurora et al., 1994).

```
 N′ Ncap-N1 N2 N3 N4-...Nc C4 C3 C2 C1 Ccap C′
STC  STC H  H  H  H     H  H  H  H  H  STC STC
```

where STC indicates nonhelical angles, and H reflects helical angles.

In this work we describe and analyze by CD and NMR, a new N-terminal local *motif* (*N-Proline box* motif). This motif is based in the presence of Pro at position N-cap that makes unnecessary the presence of a hydrogen bond partner and in the stabilizing interaction of three accompanying hydrophobic side chains at positions N′, N3, and N4.

Materials and methods

Helix description

In proteins, helical segments have average ϕ,ψ backbone dihedral angles of $-64 \pm 7°$ and $-41 \pm 7°$, respectively (Presta & Rose, 1988). Residues at the N- and C-termini, termed N-cap, and C-cap by Richardson and Richardson (1988), define helix boundaries. Each N-cap and C-cap residue could make one additional main-chain intrahelical hydrogen bond, but departs from the standard ϕ,ψ helical values (Presta & Rose, 1988).

Protein database analysis

The protein structures database used is based on the one described by Hobohm and Sander (1994), including proteins with less than 25% homology and is implemented in the program WHATIF (Vriend, 1990). The conformational searches were done with the SCAN3D option of the same program using Kabsch and Sander's (1983) definition of a secondary structure.

The selected motif contains seven consecutive residues. The two first amino acids have nonhelical angles (*STC*), and the last five are helical (*H*). The probability of finding a certain fingerprint sequence associated with this structural motif ($P_{\text{fingerprint}}$) is calculated by multiplying the individual probabilities of each residue type of the fingerprint (P_{ind}). Depending on if the residues being analyzed are located in the first two positions of the fingerprint or in the remaining five, P_{ind} will be different [$(STC)_{\text{ind}}$, $P(H)_{\text{ind}}$, respectively]. To calculate $P(STC)_{\text{ind}}$, we have divided the protein database into three-residue segments and counted the number of those in which the central position is nonhelical ($N_{\text{conf}} = 39{,}233$). The individual probability of a specific residue type ($P(STC)_{\text{ind}}$) is the number of the segments containing this residue type in the central position (N_{res}), divided by the total number of three-residue segments (N_{conf}) (Equation 1).

$$P_{\text{ind}} = N_{\text{res}}/N_{\text{conf}}. \quad (1)$$

In the same way, $P(H)_{\text{ind}}$ is calculated for different residues at the five last positions of our fingerprint. In this case $N_{\text{conf}} = 17{,}753$.

The number of cases expected (N_{expected}) for each sequence fingerprint is calculated multiplying $P_{\text{fingerprint}}$ times N_{total} (the total number of seven-residue segments with the last five residues in helical conformation, 1,339).

$$N_{\text{expected}} = P_{\text{fingerprint}} * N_{\text{total}}. \quad (2)$$

Peptide synthesis

The peptides were synthesized at the EMBL peptide service by solid-phase synthesis methods. Peptide homogeneity, compositions, and molecular weight were checked by analytical HPLC, amino acid analysis, and matrix assisted laser desorption time-of-flight mass spectrometry.

CD and aggregation analysis

Peptide concentration was determined by ultraviolet absorbance (Gill & Hippel, 1989). The CD spectra were acquired in a JASCO-710 dichrograph calibrated with (1S)-(+)-10-camphorsulfonic acid, using the continuous mode with a 1 m bandwidth, 1 s response, and a scan speed of 50 nm/min. Thirty scans were averaged to give the final spectrum. The temperature of the cuvette was kept constant at 278 K. The peptides were analyzed at pH 3.5 in the presence, or absence, of 1 M NaCl. To check the concentration dependence of the ellipticity at 222 nm, two spectra at 10 μM (5 mm pathlength cuvette) and 0.5 mM (0.1 mm pathlength cuvette) peptide concentrations were done. No concentration dependence was observed in the molar ellipticity, in any of the cases. The helical content of the peptides was estimated from the ellipticity at 222 nm (Chen et al., 1974).

$$\%\ \text{helical content} = 100 * \left[\frac{\theta_{222} - \theta_{222\text{RC}}}{39{,}500(1 - 2.57/n)} - \theta_{222\text{RC}} \right] \quad (3)$$

where θ_{222} is the observed ellipticity at 222 nm, $\theta_{222\text{RC}}$ is the ellipticity at 222 nm of the random coil state (ellipticity found for peptide AVV in the absence of salt), and n is the number of peptide bonds ($n = 17$ in our case).

Calculation of the interaction free energies

The change in free energy for α-helix formation upon mutation in peptides cannot be precisely calculated using a standard two-state model because of the existence of different conformations in equilibrium in solution. A more precise estimation can be obtained by fitting a helix/coil transition algorithm to the changes in helical content detected by far-ultraviolet (UV) CD. The one-sequence version of the helix/coil transition algorithm AGADIR1s (Muñoz & Serrano, 1997), recently modified to include local motifs, ionic strength effects, and long-range electrostatics (AGADIR1s-2; Lacroix et al., 1998), was used to fit the far-UV CD data of the peptides to the experimental value (within a $\pm 2\%$ difference). AGADIR1s-2 correctly predicts the helical content of the control peptides (GAA, AAA, GAAA, and GVA) and, therefore, it is possible to determine the contribution of the different interactions being analyzed, just by modifying their energy contribution in the algorithm. Essentially, the free-energy contribution of the target interaction is increased, or decreased, until the experimental helical content of the peptide is predicted within a certain margin.

The rest of the parameters: hydrogen bond, intrinsic propensities, capping effects, side-chain–side-chain interactions within the

helix, and electrostatic interactions with the helix dipole as well as between charged residues, were the same as previously described (Lacroix et al., 1998).

NMR analysis

NMR samples were prepared in a $H_2O/^2H_2O$ 9:1 (by vol.) mixture at pH 3.5. DQFCOSY, TOCSY, ROESY, and NOESY spectra were performed in a Bruker AMX 500 MHz spectrometer at 278 K using standard procedures (Wüthrich, 1986). Sodium 3-trimethylsylyl (2,2,3,3-2H4) propionate was used as an internal reference. The proton resonances were assigned by the sequential assignment procedure (Wüthrich, 1986). The mixing time in the NOESY spectra was 150 ms for aqueous solution and 100 ms for the 30% TFE samples. The CαH proton conformational shifts were obtained by subtracting the random-coil chemical shift values (Merutka et al., 1995) from those measured in the peptide. In the case of Gly, we show the difference with the average values of its two protons when they are separated.

Structure calculation by distance geometry

NOEs identified in the NOESY spectrum (100 ms mixing time) of a 1.2 mM preparation of IVV peptide in 30% TFE solution at pH 3.5 were classified by visual inspection into three intensity categories: strong, medium, and weak. Distances of 4, 5, and 6 Å were assigned for the corresponding pairs of protons, respectively. The distance geometry program DIANA (Guntert et al., 1991) was used to find a set of structures compatible with the observed NOEs. Protons for which stereochemical assignment was not possible were submitted to pseudo-atom conversion within the same program. Fifteen structures were selected with no distance violation and a value for the target function between 4×10^{-6} and 6×10^{-2}. Local RMSD for backbone atoms of residues 4 to 7 was lower than 0.15 Å.

Results

Database analysis

Table 1 shows the results of our statistical analysis of the protein structures database. In the protein database, there are 90 helices that contain Pro at position N-cap, and 79 cases are expected, suggesting that this location is slightly preferred with respect to other nonhelical conformations. However, as previously described (Richardson & Richardson, 1988; Dasgupta & Bell, 1993), Pro at position N1 is clearly more favored, being found four times more frequently than expected from a random distribution. Interestingly, the probability of finding Pro as the N-cap residue increases when there are apolar residues (Ala, Val, Ile, Leu, Met, and Phe; h residues) at positions N′, N3, and N4 (sequence fingerprint h-P-X-X-h-h). In this sequence fingerprint, there is an overwhelming presence for Ile and Leu at N′, and for Val at position N3 (the sequence I/LPXXVh, where X is any residue, is found 30 times above the random expected number). This preference is not due to sequence homology between the proteins containing this motif (Table 1: data not shown). On the other hand, when Pro is at position N1, we do not observe the same residue preferences (Table 1).

Table 1. *Statistical analysis of the protein database*[a]

	STC/STC/HHHHH			
Fingerprint	Cases	Observed	Expected	*f*
hXXXXXX	15,362	464	365	1.3
(I/L)PXXXX	410	28	9.6	2.9
XXXXhXX	15,383	330	424	0.8
XXXXXhX	15,632	620	424	1.5
hXXXhXX	3,951	100	107	0.9
hXXXXhX	3,925	224	107	2.1
XPXXXXX	2,637	90	79	1.1
XPXXhXX	723	35	23	1.5
XPXXXhX	726	40	23	1.7
XPXXhhX	211	17	7	2.4
hPXXXXX	756	47	20	2.3
hPXXhXX	213	22	6	3.7
hPXXXhX	182	20	6	3.3
hPXXhhX	51	10	2	5.6
(I/L)PXXVX	38	7	0.6	12.0
(I/L)PXXVh	11	6	0.2	30.0
XXPXXXX	2,637	112	32	3.5
XhPXXhh	57	1	1	1.0

[a]Note: h is hydrophobic, p is polar, X is any residue. STC is any conformation except helix and H is helix conformation. Probabilities: h (Leu, Ile, Val, Met, Phe) in STC = 0.2531; at position N-cap is 0.08364; in H = 0.2942. Prob. (I/L) Leu, Ile in STC = 0.1208. Prob. Val in H = 0.0656. Pro in STC = 0.0549; Pro in H = 0.024. The proteins having the (I/L)PXXVh motif are: 2er7 (226–231), 1csh (118–123), 2abh (296–301), 1byb (103–108, 464–469), and 1cmb (28–33).

We have analyzed the dihedral angle distribution for all the cases containing this motif in the protein structures database (Table 2). This analysis shows that Pro at the N-cap position is conformationally restricted with dihedral angles corresponding to a PPII conformation and its side chain pointing up toward the C-terminus. The ψ angle for Ile/Leu residue at position N′ is restricted in an extended conformation, while the Val side chain at position N3 has only one rotamer that results in its side chain pointing toward the N-terminus of the helix. As a result of these restrictions, the side chain of Val packs against the side chain of Pro, and the side chain of Ile/Leu at N′ interacts with the side chain of the hydrophobic residue at position N4. The N′–N4 interaction is reminiscent of a *hydrophobic staple* motif (Muñoz et al., 1995). This is illustrated in Figure 1. These side-chain–side-chain interactions could stabilize the α-helix as they do in the *hydrophobic staple* motif, explaining the higher than expected statistical preference in the protein database. We define this putative local motif (h-P-X-X-h-h) as the "*N-Pro-box*" motif.

Peptide design

The Pro-box motif has been sequentially introduced on a model polyalanine-based peptide. Different combinations of the sequence fingerprint residues have been used as controls to evaluate the local effects due to the intrinsic preferences of the amino acids to be in certain conformations.

Table 2. *Analysis of the dihedral angles and side-chain rotamers in protein Pro-boxes*

		N′		N-cap		*N* + 1		*N* + 2		*N* + 3		*N* + 4	
Protein	Sequence	ψ	χ^1	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	χ^1	ϕ
2er7 (226–231)	LPLKVA	163	−99	−54	146	−56	−47	−59	−38	−78	−49	−175	−59.7
1csh (118–123)	LPATVL	140	−66	−75	149	−50	−48	−70	−25	−70	−42	−179	−62.9
2abh (296–301)	LPSHVV	150	−98	−64	153	−62	−32	−66	−30	−74	−43	173	−68.3
1byb (103–108)	IPDSVV	172	63	−55	134	−56	−34	−63	−38	−73	−40	−167	−58.5
1byb (464–469)	IPQWVV	124	56	−66	144	−56	−35	−68	−27	−68	−44	−172	−64.9
1pbe (247–2520)	LPIEVL	161	−67	−55	136	−45	38	−59	−41	−76	−38	177	−63.5
1cmb (28–33)	IPAEVL	123	−177	−56	149	−51	−42	−70	−28	−71	−37	176	−62.5
Average		147 ± 19		−61 ± 8	146 ± 7	−56 ± 6	−38 ± 7	−65 ± 5	−32 ± 6	−73 ± 4	−42 ± 4	−178 ± 7	−63 ± 3

```
     1         10        17
GAA  GGPKAAAAKARAAKAGY- NH2
AAA  GAPKAAAAKARAAKAGY- NH2
IAA  GIPKAAAAKARAAKAGY- NH2
GAV  GGPKAAVAKARAAKAGY- NH2
AAV  GAPKAAVAKARAAKAGY- NH2
IAV  GIPKAAVAKARAAKAGY- NH2
GVA  GGPKAVAAKARAAKAGY- NH2
IVA  GIPKAVAAKARAAKAGY- NH2
GVV  GGPKAVVAKARAAKAGY- NH2
AVV  GAPKAVVAKARAAKAGY- NH2
IVV  GIPKAVVAKARAAKAGY- NH2
GAL  GGPKAALAKARAAKAGY- NH2
AAL  GAPKAALAKARAAKAGY- NH2
IAL  GIPKAALAKARAAKAGY- NH2
GVL  GGPKAVLAKARAAKAGY- NH2
IVL  GIPKAVLAKARAAKAGY- NH2
GAAA GGAKAAAAKARAAKAGY- NH2
GAVA GGAKAVAAKARAAKAGY- NH2
```

Tyr at the end of the sequence, separated from the putative helical region by a Gly residue, was introduced to allow an accurate determination of peptide concentration, without the interference of the aromatic residue on the far-UV CD spectra (Chakrabartty et al., 1993).

Positions 2, 6, and 7 of the peptides have been mutated independently. Gly, Ala, and Ile occupy position 2. Ile at position 2 is part of the motif we are analyzing, while Ala and Gly are controls. Pro cannot occupy position N2 of an α-helix without disrupting it. The reason behind this is that there are steric constraints that force the residue preceding a Pro to adopt extended ϕ dihedral angles (Flory, 1988). As a result, the residues at position 2 can only be found at positions N′ and N-cap of a putative regular α-helix. Because Gly is a good N-cap residue, while Ala and Ile are not (Doig et al., 1994; Muñoz & Serrano, 1995b), we expect to find Gly at the N-cap position, and Ala and Ile outside of the helix at position N′. At position 6, Ala is a control of the putative Pro–Val interaction. Position 7 bears Ala, Val, or Leu, to check for the specificity of the interaction with residue N′ in our local motif,

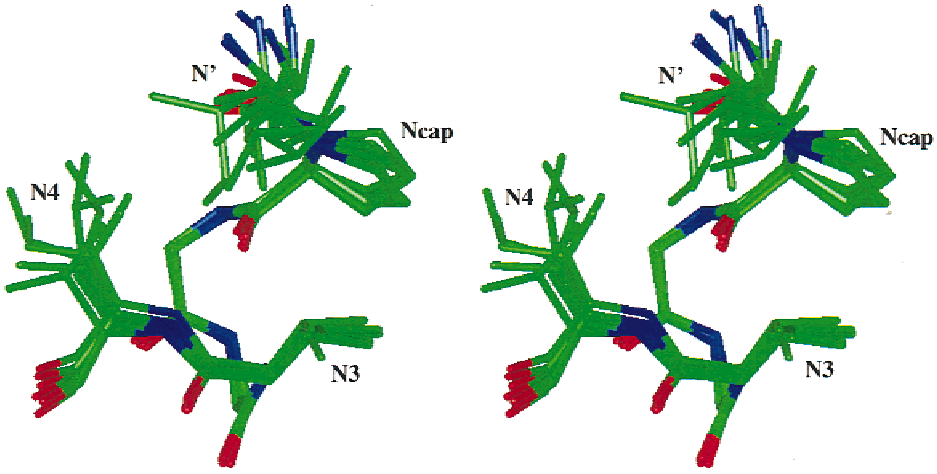

Fig. 1. Superimposition of the region corresponding to the fingerprint sequence lPXXVX (l = leu + ile, P = pro, V = val, X = any amino acid) adopting the described motif in known protein structures. Corresponding entries in the database are: 2er7, 1csh, 2nad, 1byb, 1pbe, 1cmb, 1gox. Identification and representation have been performed with the program WHATIF.

because any of the three can make a *Hydrophobic staple* motif with Ile at position 2 of the peptide (Muñoz & Serrano, 1995a; Muñoz et al., 1995). The two peptides with Ala at position 3 instead of Pro, and Ala, or Val, at position 6 (GAAA, GAVA), are controls to check for the reliability of AGADIR in predicting helix destabilization upon mutating a Val into an Ala in an α-helix.

CD analysis

The CD measurements were performed at low pH, 3.5. The helix content of the peptides was low in water solution in the absence of salt, due to the repulsion of the positively charged amino acids and the destabilizing effect of the free N-terminus (Lacroix et al., 1998). At high ionic strength (1 M NaCl), the helix content was reasonable, and further calculations were done with these data. No aggregation of the peptides was observed under any of the tested conditions. Peptide AVV in the absence of salt has been used to obtain the random coil ellipticity of the peptide series (ellipticty at 222 nm of ~1,378).

Table 3. *CD data, helix percentage and helix content predicted by AGADIR*[a]

	e222	%hel	Agadir[b]	Agadir[c]	Agadir[d]	Agadir[e]
GAA	−7,211	27	26	26	26	26
AAA	−4,096	17	18	18	18	18
IAA	−4,933	20	19	22	19	22
GAV	−3,171	14	14	14	14	14
AAV	−2,336	12	10	11	10	11
IAV	−3,435	15	11	14	11	14
GVA	−3,501	15	16	16	17	17
AVA	−2,078	11	12	12	13	13
IVA	−4,302	18	12	14	15	17
GVV	−1,575	9	9	9	10	10
AVV	−1,281	8	7	7	8	8
IVV	−2,702	13	7	9	9	12
GAL	−6,837	26	24	24	24	24
AAL	−4,710	19	16	17	16	17
IAL	−6,296	24	18	22	18	23
GVL	−3,761	16	14	14	16	16
IVL	−5,260	21	11	14	14	19
GAAA	−9,514	34	33	33	33	33
GAVA	−4,105	17	20	20	20	20

[a]The ellipticity values at 222 nm for 100 and 0% helix are −33,528 and 1,378, respectively.

[b]AGADIR1s-2 prediction (Lacroix & Serrano, 1998).

[c]AGADIR1s-2 prediction after consideration of a full hydrophobic staple contribution when Pro is the N-cap residue (Ile-Ala −0.55 kcal/mol; Ile-Leu −0.8 kcal/mol; Ile-Val −0.75 kcal/mol). In the previous version of AGADIR the hydrophobic staple contribution was multiplied by 1 when the N-cap residue is Ser, Thr, Asn, or Asp, and by 0.5 with any other N-cap residue (see supplementary material in Lacroix et al., 1998). This is based on the analysis of several different peptides containing this motif with, or without a good N-capping residue.

[d]AGADIR1s-2 prediction after consideration of an $i,i+3$ interaction (−0.4 kcal/mol) between Pro and Val, when Pro is the N-cap residue.

[e]AGADIR1s-2 prediction after considering both contributions simultaneously (1.15 kcal/mol). The magnitude of the error, in the determination of the free energy contribution, is large for those peptides having a small helical content. However, the Pro-Val $i,+3$ interaction is found in several of the peptides, and we can reproduce their helical content with the same interaction energy without any fitting. We estimate that the average error is ±0.25 kcal/mol.

Table 3 summarizes the CD results, and Figure 2 shows the CD spectra for some of the more important peptides in this study (GAL, IAL, GVL, and IVL, see below).

All peptides with a Gly at position 2 are more helical than those with Ala and Ile at the same position, as expected from the better N-capping properties of Gly (Doig et al., 1994). However, there are two important exceptions that are peptides IVV and IVL, in which the full sequence fingerprint (I/LPXXVh) has been introduced. Mutating Ala into Ile at position 2 results in a higher helical content, independently of which residues are found at positions 6 and 7. This could be due to better N-capping properties of Ile compared to Ala (Doig et al., 1994), if Pro is occupying position N1 in the α-helix conformation. Alternatively, the formation of a *Hydrophobic Staple* motif between Ile2 side chain at position N′ and the hydrophobic residue at position N4 (Ala, Val, or Leu) could explain this result (Muñoz et al., 1995; Muñoz & Serrano, 1995a). Formation of a *Hydrophobic Staple* motif between residues at positions 2 and 7 (N′ and N4) is supported by the fact that Leu at position 7 does not reduce the helical content of the corresponding peptides when compared to Ala, or even increases it (IAL and IVL). A similar result was found in a series of peptides used to analyze the *Hydrophobic staple* motif and was explained by the existence of a favorable hydrophobic interaction of Leu at position N4 with the residue at position N′ (Gly, Ala, and Ile) (Muñoz et al., 1995; Muñoz & Serrano, 1995a). Val, at position 6 or 7, reduces the helical content with respect to Ala, as expected from its poor helical propensity (Padmanabhan et al., 1990; Muñoz & Serrano, 1995b; Petukhov et al., 1998), not compensated in the case of position 7 by the formation of a *Hydrophobic staple*.

NMR analysis in 30% (v/v) trifluoroethanol (TFE)

To check for the formation of the "*N-Pro box*" motif described at the beginning of the Results section we have analyzed by NMR the

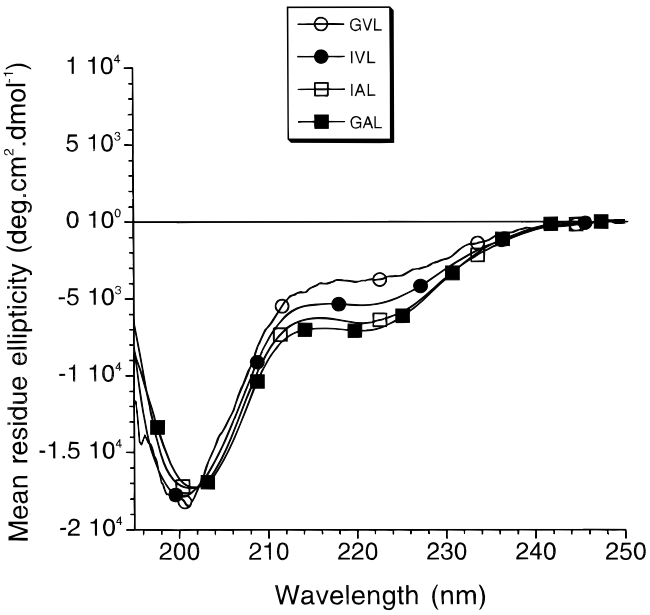

Fig. 2. Far-UV CD spectra of some of the studied peptides in aqueous solution (1 M NaCl pH 3.5).

two peptides containing the full motif (IVV and IVL). This analysis has been done in 30% TFE because the spectra in 1 M NaCl are of very poor quality, and in aqueous solution the peptides are not structured.

Figure 3 shows the difference in the conformational chemical shifts of the Cα protons with respect to random coil values (Merutka et al., 1995), for peptides IVV and IVL. The conformational shifts are the same within the error for all amino acids except for position 7 (Val in IVV and Leu in IVL). This reflects a difference in the random coil conformational preferences of Leu and Val. Although Val in the random coil is populating mainly extended regions in the Ramachandran plot, Leu populates the helical region equally well (Serrano, 1995; Swindells et al., 1995; Smith et al., 1996). The final result is that the change in the CαH chemical shift when folding into an α-helix is larger for Val than for Leu (Serrano, 1995).

Table 4 shows the long-range NOEs that we found in 30% TFE for peptide IVV, and Figure 4 shows a region from the NMR spectrum showing some of the critical long-range NOEs described in Table 4. NOEs are observed between Pro3 and Val6 side chains, as well as between Ile2 and Pro3. Regretfully, no NOEs could be observed between the Ile side chain and any of the Val side chains, because of signal overlapping. A set of structures, compatible with the observed NOEs, has been calculated by distance geometry with the program DIANA (Guntert et al., 1991). The α-helix at the N-terminus is very well defined, while the C-terminus of the helix is not due to severe overlapping of the NOEs (Fig. 5). Pro is the N-cap residue and the Val side chains adopt the expected *trans* rotamer (see Table 2), with Val6 packing against the Pro side chain. This rotamer is, in fact, the preferred one in α-helices for β-branched residues (Dunbrack & Karplus, 1994), while for linear, or aromatic side chains two or more rotamers are allowed. This reduces the entropic cost of the Pro–Val interaction and could be one of the clues for the special preference for Val at N3, among other hydrophobic residues.

Free energies of interaction

We have estimated the free energies of interaction by using a new version of the helix/coil transition theory algorithm AGADIR (Muñoz & Serrano, 1997). This new version, AGADIR1s-2 (La-

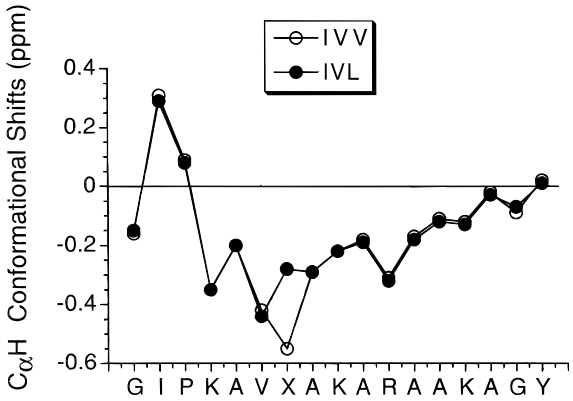

Fig. 3. Plot of the conformational shifts of the Cα protons defined as $\delta_{measured} - \delta_{coil}$ in 30% TFE solution for peptides IVL and IVV.

Table 4. *Observed NOEs in a 100 ms mixing time NOESY experiment of peptide IVV in 30% TFE solution*[a]

2 Ile			
Hγ	3 Pro	Hδ	4.0
3 Pro			
Hα	4 Lys	HN	3.0
Hβ1	5 Ala	HN	4.0
Hβ1	5 Ala	Hβ	5.0
Hβ2	6 Val	HN	4.5
Hβ2	6 Val	Hγ2	4.5
Hγ	6 Val	Hγ2	5.0
Hδ1	6 Val	Hβ	5.0
Hδ1	6 Val	Hγ2	5.0
4 Lys			
HN	5 Ala	HN	5.0
HN	7 Val	Hγ2	5.0
Hα	7 Val	HN	5.0
Hα	7 Val	Hβ	4.0
Hα	7 Val	Hγ2	4.0
Hβ	7 Val	HN	5.0
Hβ	8 Ala	HN	5.0
Hγ	7 Val	HN	5.0
5 Ala			
HN	6 Val	HN	5.0
Hα	8 Ala	HN	5.0
Hβ	6 Val	Hα	5.0
6 Val			
HN	7 Val	HN	4.0
Hα	9 Lys	HN	4.0
Hα	9 Lys	Hβ	5.0
Hα	9 Lys	Hγ	5.0
Hα	10 Ala	HN	5.0
7 Val			
HN	8 Ala	HN	4.0
Hα	10 Ala	HN	4.0
Hα	10 Ala	Hβ	4.0
Hα	11 Arg	HN	5.0
7 Val			
Hγ1	10 Ala	HN	5.0
Hγ1	11 Arg	HN	5.0
Hγ1	11 Arg	Hδ	5.0
Hγ1	11 Arg	Hϵ	5.0
8 Ala			
HN	9 Lys	HN	4.0
9 Lys			
HN	10 Ala	HN	4.0
HN	13 Ala	HN	5.0
Hα	12 Ala	Hβ	5.0
10 Ala			
Hγ1	11 Arg	HN	4.0
Hγ1	14 Lys	Hβ	5.0
11 Arg			
HN	12 Ala	HN	4.0
Hα	14 Lys	HN	5.0
12 Ala			
HN	13 Ala	HN	4.0
13 Ala			
HN	14 Lys	HN	4.0
Hα	18 Gly	HN	5.0
Hβ	18 Gly	HN	5.0
14 Lys			
HN	15 Ala	HN	4.0
Hα	17 Tyr	Hβ	5.0
Hα	17 Tyr	Hγ	5.0
Hα	17 Tyr	Hζ	6.0
15 Ala			
HN	16 Gly	HN	4.0
Hβ	17 Tyr	Hζ	6.0
16 Gly			
HN	17 Tyr	HN	4.0
Hβ	17 Tyr	Hγ	6.0

[a]Distances were assigned by visual inspection of the intensities.

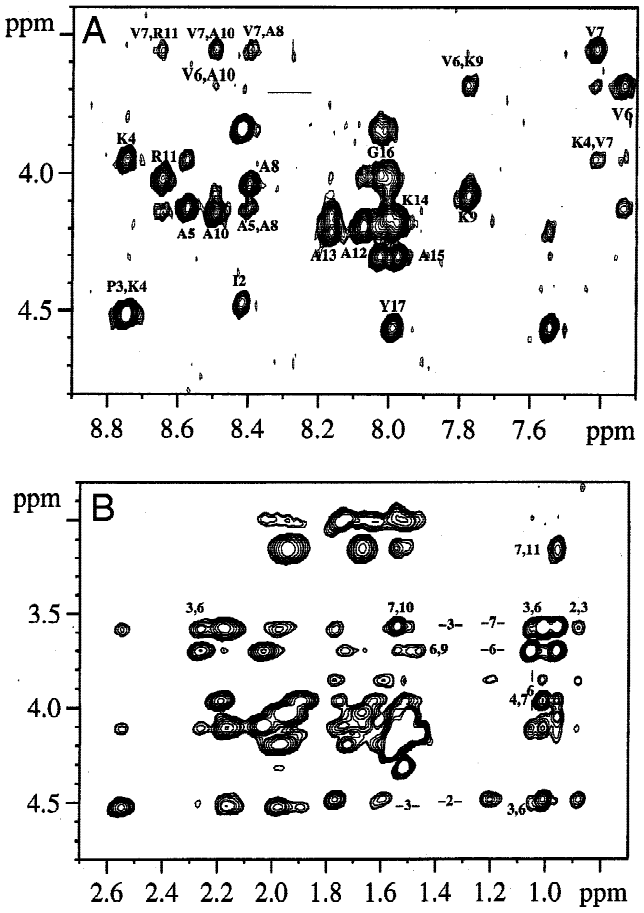

Fig. 4. Selected regions of the 2D ^{1}H-NMR spectra showing some of the NOEs described in the manuscript. **A:** Amide-CαH. **B:** Cα-methyl side-chain region. The position of the intraresidue NOEs is shown as single numbers. The nonsequential NOEs are shown as two numbers separated by a comma.

croix et al., 1998), includes previously described local motifs: *Hydrophobic staple*, *Schellman* motif, and *Pro-capping* motif, new variants of those and newly described side-chain–side-chain interactions. It takes into consideration a position dependence of the helical propensities for some of the 20 amino acids (Petukhov et al., 1998). A new electrostatic model with all electrostatic interactions up to 12 residues in distance in the helix and random coil conformations, as well as the effect of ionic strength has been implemented.

In Table 3 we show AGADIR1s-2 prediction of the helical content for all the peptides. The algorithm correctly predicts (±2% of the experimental value) the helical content of peptides: GAA, AAA, IAA, GAV, AAV, GVA, AVA, GVV, AVV, GAL, GVL, and GAAA. For peptides AAL and GAVA, the predicted and experimental values are quite close (less than ±3%, when taking into account the decimals). However, peptides IAV, IVA, IVV, IAL, and IVL are all predicted to have less helical content (4 to 10%) when compared with the experimental data. One possibility to explain these results is that the N-cap contribution of Ile or the position dependence of the helical propensity of Val are not estimated correctly. The N-cap contribution of Ile in helices has been calibrated in AGADIR using the peptide series analyzed by Baldwin and coworkers (Doig et al., 1994). Making Ile more favorable as N-cap residue to correctly predict peptides IAV, IVA, IVV, IAL, and IVL results in an overestimation of the helical content of peptides IAA and IAL (data not shown). Therefore, the N-cap contribution of Ile cannot explain the low helical content predicted for these peptides. Moreover, NMR analysis of peptide IVV indicates that Pro occupies the N-cap position, while Ile is at position N′. In the case of Val, we have previously determined the position dependence of its intrinsic helical propensity (Petukhov et al., 1998). In any case, we could not find any combination of values for Val intrinsic propensity at different helical positions that will explain the behavior of all the peptides analyzed here (data not shown). Therefore, there must be other factors that can explain these differences:

1. There are favorable $i,i+3$ Pro–Val and/or $i,i+4$ Pro–Val or Pro–Leu, interactions when both residues are in the helical conformation.

2. Pro, when being the N-cap residue, can interact with residues at position $i+3$ and/or $i+4$, as we have seen in the NMR analysis.

3. Pro, when being the N-cap residue, can promote the formation of a *hydrophobic staple* motif as efficiently as good N-cap residues: Asn, Asp, Ser, and Thr. In AGADIR1s-2 the *Hydrophobic staple* contribution, when the N-cap residue is not Asn, Asp, Ser, and Thr, is considered to be half of its normal value (Lacroix et al., 1998).

To determine which of these possibilities is correct, we have introduced, or modified, the corresponding interactions in AGADIR1s-2. We can reproduce the helical content of some of the peptides mentioned above by introduction of favorable $i,i+3$ Pro–Val and/or $i,i+4$ Pro–Val or Pro–Leu interactions when both residues are in the helical conformation. However, this result is an overestimation of some of the peptides initially predicted correctly (data not shown). Moreover, in the helical conformation the Pro side chain is pointing toward the N-terminus far away from the N3 Val side chain. Consideration of cases (2) and (3), separately, correctly predicts some of the above-mentioned peptides, but not all of them (Table 2). It is only the combination of these two possibilities that allows a correct prediction of all the peptides analyzed (Table 3). Comparison of the predicted helical content at a residue level for peptides GVL and IVL illustrates that when having the *Pro-box* motif sequence, Pro tends to occupy the N-cap position. On the other hand, when there is a good N-cap residue before Pro in the sequence (Gly in this case) it results in Pro occupying the $N+1$ position (Fig. 6).

Discussion

There are several ways in which a Pro residue can be found at the beginning of an α-helix. In one of these patterns, Pro is occupying the N-cap position. An analysis of the protein database shows that the presence of a Pro at the N-cap position of an α-helix is usually associated to hydrophobic amino acids at positions N′, N3, and N4. This could suggest the presence of a new N-terminal local motif (sequence fingerprint: hPXXhh), which could stabilize the helical conformation and/or define the helix N-terminus (*Pro-box* motif). However, the appearance in the protein structure data bank of

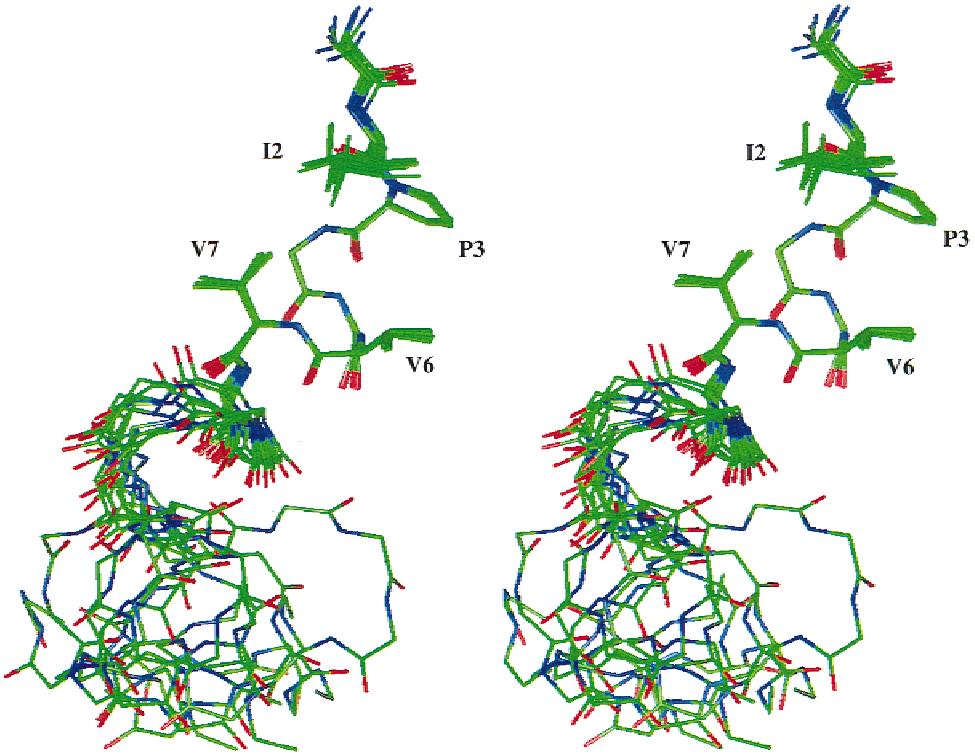

Fig. 5. Superimposition of the 15 best structures obtained by distance geometry analysis using the program DIANA (Guntert et al., 1991) of the NOEs obtained in a 100-ms mixing time NOESY experiment of the IVV peptide. Representation was done using the program InsightII (Biosym Technologies, San Diego, California).

certain amino acid combinations more frequently than expected should be regarded with caution. For example, it is normal to find in helices a nice hydrophilic/hydrophobic pattern that is more related to tertiary considerations than to local effects. The distinction between tertiary and local reasons for the frequent association of a structural and a sequence motif can only be done by isolating the given sequence fingerprint in the absence of the protein context.

The CD and NMR analysis we have performed on several polyalanine-based peptides containing the *Pro-box* motif fingerprint sequence (hPXXhh) and the corresponding controls indicates that this sequence corresponds to a local helical motif. This motif is similar in its organization to the *Hydrophobic staple* combined with a *Capping box* motif. In both cases, there is a favorable interaction between residues at positions N-cap and N3 (a side-chain–side-chain hydrophobic interaction in the *N-terminal Pro-box* motif and two side-chain–main-chain hydrogen bonds in the *Capping box* motif), as well as between N′ and N4. The *Hydrophobic staple* contribution for the three pairs analyzed here: Ile-Ala, Ile-Val, and Ile-Leu seems to be the same than that in other peptides containing a *Capping box* motif (Muñoz & Serrano, 1995a).

In globular proteins, Pro occurs rarely inside α-helices, and in those rare cases the helices are kinked (Barlow & Thornton, 1988). The reason for it is that the Pro main chain cannot make an $i, i-4$ hydrogen bond. However, Pro ϕ angle is constrained to approximately $-65°$, which is ideal for a helical conformation. This should make Pro an excellent amino acid in the first turn of the α-helix where the amide groups are not involved in main-chain–main-chain hydrogen bonding. However, Pro is statistically preferred at position N1 but not at other helical positions. This is explained because the pyrrolidine ring conformationally restricts the ψ angle of the preceding residue in an extended conformation incompatible with helical angles (Schimmel & Flory, 1968). This results in an entropic gain but is incompatible with the preceding residue adopting a helical conformation. For the same reason, Pro should be in principle also compatible with the N-cap position, although statistically this is not the case. The reason for it is that Pro side chain will not provide a hydrogen bond to any of the unsatisfied main-chain amide groups in the first helical turn as Asn, Asp, Ser, or Thr do, nor it will facilitate their solvation as Gly does (Serrano & Fersht, 1989). However, by restricting the ψ angle of the preceding residue it poses its side chain to point toward the helix and, therefore, favors the formation of a side-chain–side-chain interaction with residue N4 (*Hydrophobic staple* motif). Therefore, when Pro occupies the N-cap position there must be a selection toward residues at positions N′ and N4 that make favorable interactions. Another important point is that Pro at the N-cap has its side chain oriented toward the side chain of residue N3 (Fig. 1), while at position N1 it points toward the N-terminus of the helix. As a result, when Pro is the N-cap residue it can establish side-chain–side-chain interactions with the N3 residue, but not when it is at position N1. The overwhelming preference for Val at N3 can be explained by the fact that the rotamer normally adopted by Val in helices (*trans* conformer; Dunbrack & Karplus, 1994) is the one found in this interaction. In addition, due to the nature of Pro side chain, its packing with the Val side chain buries a large surface

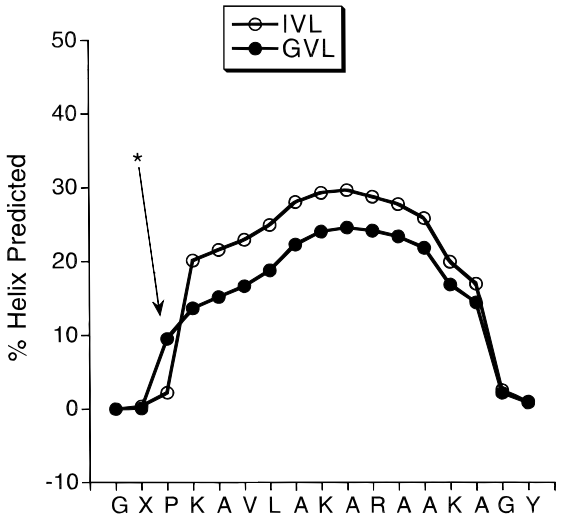

Fig. 6. AGADIR1s-2 prediction of the helical content at a residue level for peptides GVL and IVL. The asterisk illustrates the difference in the helical behavior of Pro in the two peptides. While in peptide GVL Pro adopts the helical conformation, in peptide IVL Pro tends to occupy the N-cap position. The prediction has been done after incorporating the interactions described in this paper.

area, while this is not the case for other longer hydrophobic residues (data not shown).

Despite of the above favorable interactions (~−1.2 kcal/mol), Pro as N-cap together with hydrophobic amino acids at positions N′, N3, and N4 is not very common in proteins. A simple explanation for this is that compared with other capping motifs, the *N-terminal Pro-box* is not very stabilizing due to the low N-cap propensity of Pro and the poor helix propensity of Val, even considering the favorable interaction with Val at N3. This can be exemplified by the fact that peptide GAA, which has Gly as the N-cap and Pro at position N1, is more helical than peptide IVL having the full *N-terminal Pro-box*. However, the presence of this motif will clearly prevent helix elongation, because a Pro in helical angles is very unfavorable and will also eliminate competing conformations. This is shown by the fact that out of 11 cases in the protein structure database containing the *N-terminal Pro-box* fingerprint, 8 are located at the N-terminal of α-helices.

Acknowledgments

A.R.V. was supported by a Fellowship of the Spanish Ministry of Education and Science. This work was partly supported by an EU Grant (BIO4-CT97-2086).

References

Aurora R, Rose GD. 1998. Helix capping. *Protein Sci 7*:21–38.

Aurora R, Srinivasan R, Rose GD. 1994. Rules for alpha-helix termination by glycine. *Science 264*:1126–1130. [Published Erratum appears in *Science.* 1994. *264*:1831; see Comments.]

Barlow DJ, Thornton JM. 1988. Helix geometry in proteins. *J Mol Biol 201*:601–619.

Chakrabartty A, Kortemme T, Padmanabhan S, Baldwin RL. 1993. Aromatic side-chain contribution to far-ultraviolet circular dichroism of helical peptides and its effect on measurement of helix propensities. *Biochemistry 32*:5560–5565.

Chen YH, Yang JT, Chau KH. 1974. Determination of the helix and beta form of proteins in aqueous solution by circular dichroism. *Biochemistry 13*:3350–3359.

Dasgupta S, Bell JA. 1993. Design of helix ends. *Int J Pept Protein Res 41*:499–511.

Doig AJ, Chakrabartty A, Klingler TM, Baldwin RL. 1994. Determination of free energies of N-capping in alpha-helices by modification of the Lifson-Roig helix-coil therapy to include N- and C-capping. *Biochemistry 33*:3396–3403.

Dunbrack RL Jr, Karplus M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol 1*:334–340.

Flory PJ. 1988. *Statistical mechanics of chain molecules*. Oxford: Oxford University Press.

Gill SC, Hippel PH. 1989. Calculation of protein extinction coefficients from aminoacid sequences data. *Anal Biochem 182*:319–326.

Guntert P, Braun W, Wuthrich K. 1991. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J Mol Biol 217*:517–530.

Harper ET, Rose GD. 1993. Helix stop signals in proteins and peptides: The capping box. *Biochemistry 32*:7605–7609.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

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

Lacroix E, Viguera AR, Serrano L. 1998. Elucidating the folding problem of α-helices: Local motifs, long-range electrostatics, ionic strength dependence and prediction of NMR parameters. *J Mol Biol 284*:173–191.

Lyu PC, Wemmer DE, Zhou HX, Pinker RJ, Kallenbach NR. 1993. Capping interactions in isolated alpha helices: Position-dependent substitution effects and structure of a serine-capped peptide helix. *Biochemistry 32*:421–425.

Merutka G, Dyson HJ, Wright PE. 1995. "Random coil" 1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J Biomol NMR 5*:14–24.

Milner-White EJ. 1988. Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with alpha-helices and beta-bulge loops. *J Mol Biol 199*:503–511.

Muñoz V, Blanco FJ, Serrano L. 1995. The hydrophobic-staple motif and a role for loop-residues in alpha-helix stability and protein folding. *Nat Struct Biol 2*:380–385.

Muñoz V, Serrano L. 1995a. Analysis of $i,i+5$ and $i,i+8$ hydrophobic interactions in a helical model peptide bearing the hydrophobic staple motif. *Biochemistry 34*:15301–15306.

Muñoz V, Serrano L. 1995b. Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J Mol Biol 245*:297–308.

Muñoz V, Serrano L. 1995c. Helix design, prediction and stability. *Curr Opin Biotechnol 6*:382–386.

Muñoz V, Serrano L. 1997. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers 41*:495–509.

Padmanabhan S, Marqusee S, Ridgeway T, Laue TM, Baldwin RL. 1990. Relative helix-forming tendencies of nonpolar amino acids. *Nature 344*:268–270.

Penel S, Hughes E, Dolg AJ. 1999. Side-chain structures in the first turn of the α-helix. *J Mol Biol 287*:127–143.

Petukhov M, Muñoz V, Yumoto N, Yoshikawa S, Serrano L. 1998. Position dependence of non-polar amino acid intrinsic helical propensities. *J Mol Biol 278*:279–289.

Presta LG, Rose GD. 1988. Helix signals in proteins. *Science 240*:1632–1641.

Prieto J, Serrano L. 1997. C-capping and helix stability: The Pro C-capping motif. *J Mol Biol 274*:276–288.

Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of alpha helices. *Science 240*:1648–1652. [Published Erratum appears in *Science.* 1988. *242*:1624.]

Schellman C. 1980. In: Jaenicke R, ed. *Protein folding.* New York: Elsevier/North Holland. pp 53–61.

Schimmel PR, Flory PJ. 1968. Conformational energies and configurational statistics of copolypeptides containing L-proline. *J Mol Biol 34*:105–120.

Seale JW, Srinivasan R, Rose GD. 1994. Sequence determinants of the capping box, a stabilizing motif at the N-termini of alpha-helices. *Protein Sci 3*:1741–1745.

Serrano L. 1995. Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. *J Mol Biol 254*:322–333.

Serrano L, Fersht AR. 1989. Capping and alpha-helix stability. *Nature 342*:296–299.

Shakhnovich EI. 1998. Protein design: A perspective from simple tractable models [In Process Citation]. *Folding Design 3*:R45–R58.

Smith LJ, Bolin KA, Schwalbe H, MacArthur MW, Thornton JM, Dobson CM. 1996. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J Mol Biol 255*:494–506.

Swindells MB, MacArthur MW, Thornton JM. 1995. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat Struct Biol 2*:596–603.

Viguera AR, Serrano L. 1995a. Experimental analysis of the Schellman motif. *J Mol Biol 251*:150–160.

Viguera AR, Serrano L. 1995b. Side-chain interactions between sulfur-containing amino acids and phenylalanine in alpha-helices. *Biochemistry 34*:8771–8779.

Vriend G. 1990. WHATIF: A molecular modeling and drug design program. *J Mol Graph 8*:52–56.

Wüthrich K. 1986. *NMR of proteins and nucleic acids.* New York: John Wiley & Sons.