Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Secondary structures at polypeptide-chain termini and their features

Rajasri Bhattacharyya,† Debnath Pal†‡ and Pinak Chakrabarti*

Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Kolkata 700 054, India

† RB and DP contributed equally to this work.
‡ Present address: Department of Structural Biology and Crystallography, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 Jena, Germany.

Correspondence e-mail: pinak@boseinst.ernet.in

Received 10 May 2002
Accepted 25 July 2002

An analysis of secondary structures (α-helices and β-strands) in the two terminal regions of polypeptide chains reveals features different from those observed over the whole protein structure. Compared with the overall distribution, the helices in the N-terminal region tend to be smaller and have higher propensities to contain Gln and Leu, while the C-terminal helices are longer and have a greater proportion of Lys and Glu. As a strand, the C-terminal region is never found in the interior of parallel β-sheets and has a higher propensity to be at the edge of antiparallel β-sheets. In contrast, compared with the whole structure the N-terminal region has a higher propensity to be in the interior of parallel β-sheets. Compared with the overall distributions, terminal helices and strands show distinct periodicities in length. The Schellman motif, which is a prevalent C-capping motif in helices, is not common in C-terminal helices. There are other observations that can be used in the design of helical peptides: more residues beyond the C-terminus of helices are used for capping interactions than residues before the N-terminus. Consideration of the distribution of terminal strands in the interior and at the edge of β-sheets suggests a sequential folding mechanism beginning at the N-terminus of the polypeptide chain.

1. Introduction

The non-random occurrence of certain amino-acid residues at the first and last positions of polypeptide chains is well established (Berezovsky *et al.*, 1999; Pal & Chakrabarti, 2000). Met is over-represented in the first position, which is the consequence of the initiation of translation from Met and *N*-formyl Met in eukaryotic and prokaryotic cells, respectively. The retention or excision of this residue by aminopeptidases depends primarily on the second amino-acid residue: small residues favour the removal of Met, while large, hydrophobic and charged residues seem to prevent removal (Hirel *et al.*, 1989; Tsunasawa *et al.*, 1985). The chemical structure of the N-terminal amino acid of the mature protein considerably influences its half-life (Bachmair *et al.*, 1986; Varshavsky, 1996). Though less prominent, at the other end of the polypeptide chain there is some preference for the last position to be occupied by residues such as Lys, Arg and Gln (Berezovsky *et al.*, 1999; Brown *et al.*, 1990; Pal & Chakrabarti, 2000). The efficiency of translation termination is known to depend on the last two amino acids of the nascent peptide (Björnsson *et al.*, 1996) and there are strong biases in the upstream and

Table 1
Numbers of terminal secondary structures and their flanking positions for which information on sequence (and coordinates, given in parentheses, if different from the number based on sequence) is available.

Although the same positional labels are used for strands, the concept of capping, as applicable to helix, is not applicable to strands.

Structure	N′′′′	N′′′	N′′	N′	Ncap	Nterm	Cterm	Ccap	C′	C′′	C′′′	C′′′′
Helix	13 (6)	27 (21)	45 (37)	66 (57)	74	74	134	134	108 (97)	80 (65)	41 (29)	19 (14)
Strand	18 (12)	38 (26)	65 (49)	96 (82)	152	152	114	114	69 (61)	43 (33)	23 (17)	14 (9)

downstream nucleotides surrounding stop codons (Brown *et al.*, 1993; Tate & Mannering, 1996), possibly signifying a recognition site (which is larger than the triplet codon) for the polypeptide-release factors (Nakamura *et al.*, 1996).

Like the sequence, the folding pattern of the chain termini is also not random. With one end free, the two terminal segments of a chain have lesser structural constraints than the interior during folding; also, because of the charge, the termini are located on the protein surface (Thornton & Sibanda, 1983). It is a moot point whether the termini have any role in protein folding. Considering the first secondary structure that is encountered from the terminus within the terminal ten residues of the chain, it was found that the amino-terminal (N-terminal) region adopts an extended β-strand, while the carboxy-terminus (C-teminus) is usually helical (Thornton & Chakauya, 1982). Pal & Chakrabarti (2000) observed that none of the terminal residues is a part of any regular secondary structure, but there is a greater proclivity towards assuming a β conformation for the next six residues in the N-terminal region, possibly owing to electrostatic and hydrogen-bond interaction involving the free amino group, which stabilizes an extended structure near the terminus that can propagate for a few residues, forming a β-strand (Chakrabarti & Pal, 2001). The residue preceding the C-terminus also has a high propensity to be in the β conformation, but prior to this α-helix is preferred to β-strand in a ratio which is not much different from the average α:β ratio of residues in all the structures.

The terminal secondary structures (α-helix and β-strand) in polypeptide chains with a free end are under lesser constraint from tertiary interactions and may have different length distributions, residue compositions and a different frequency of occurrence in parallel and antiparallel β-sheets, as well as interior and edge strands. 'Capping' interactions, both hydrogen bonding involving the >N—H groups in the first turn of the helix and >C═O groups in the last turn (which are not engaged in typical helix hydrogen bonding) as well as hydrophobic interactions (between two residues: one within the helix and another beyond) contribute to helix stability (Aurora & Rose, 1998) and it is worth knowing whether terminal helices have features different from those of interior helices. These and other related issues are addressed in this paper by considering helices and strands which have one end lying within the terminal six residues of polypeptide chains in known structures. The design of peptides and small proteins with a small assortment of secondary-structural elements (Dahiyat & Mayo, 1997; DeGrado, 1997; Regan & Wells, 1998) may benefit from this analysis.

2. Methods

Atomic coordinates were obtained from the Protein Data Bank (PDB) now operated by the Research Collaboratory for Structural Bioinformatics (Berman *et al.*, 2000). 432 chains (in 418 files) were selected using PDB_SELECT (Hobohm & Sander, 1994) from PDB files (as of March 2000) with an R factor $\leq 20\%$, a resolution ≤ 2.0 Å and sequence identity less than 25%. The names of the PDB files used can be found in Bhattacharyya *et al.* (2002).

The secondary-structural elements in a chain were determined using the *DSSP* algorithm (Kabsch & Sander, 1983). As one or more of the terminal residues may be disordered and not seen in the electron-density map and consequently would have no records corresponding to its coordinates, there was a need to define terminal secondary structures not on the basis of what is given under the ATOM coordinates list, but from the sequence information of the 'whole' chain provided under the SEQRES records of the PDB files. When the first (for the N-terminal end) or the last (C-terminal) secondary structure that is encountered in a chain is an α-helix (*DSSP* notation 'H') or β-strand ('E') and it occurs at a position within the first (or last) six residues as given in the sequence under the SEQRES record, it is identified. Thus, if six (or more) terminal residues were missing from the ATOM coordinates list, the chain was excluded from our analysis even if the observed model starts (or ends) with 'H' or 'E' as the secondary structure. The numbers of secondary structures and preceding/following positions for which sequence and structural information could be retrieved are given in Table 1. All the terminal secondary structures were pooled and analyzed for their length, residue composition, hydrogen bonding *etc.*, the latter being determined using the program *HBPLUS* (McDonald & Thornton, 1994).

Helices and their flanking residues are labelled as follows (Presta & Rose, 1988; Richardson & Richardson, 1988):

N′′—N′—Ncap—N1—N2—N3— $\cdots$ —C3—C2—C1—Ccap—C′—C′′.

The positions N1 through C1 were identified by noting the helical residues which are denoted by 'H' in the *DSSP* output; the immediately flanking positions were labelled Ncap and Ccap. The first three residues, N1, N2 and N3, were grouped together as the N-end region of the helix, while C1, C2 and C3 constitute the C-end region. The α-helix (and similarly β-sheet) propensity was calculated as the proportion of a particular amino acid in α-helix divided by the proportion of all amino acids in α-helix (Chou & Fasman, 1974). The local propensity of a residue to be in the N-end or C-end region was

likewise calculated as the proportion of the residue in a particular region of helix divided by the proportion of all residues in helix in that particular region (Chakrabarti & Pal, 2001). The propensity of a residue to occur at a position before or after the helix was calculated as the ratio of the fraction (based on the total number of chains having sequence information, not necessarily coordinates, for that position) of the residue to occur at the position to the fraction of the residue in the data set. Different sets of propensities were examined for statistical significance by calculating the z values as given in Karpen *et al.* (1992). If $|z| \geq 1.96$ (5% significance level), the observed number of occurrences was considered to deviate significantly from its expected value. Negative values of z indicate under-representation and positive values indicate over-representation.

β-Strands were identified as a stretch of at least two residues with tag 'E' in the *DSSP* output. It is not always straightforward to identify a strand with parallel, antiparallel or mixed β-sheet, or whether it is an edge or an interior strand. For example, a strand that is edge at one end can be interior at the other, being part of two different β-sheets. Consequently, we assigned the type of β-sheet and the location (edge/interior) based on the characteristics of the maximum number of residues in the strand.

Two hydrophobic residues around the helix terminus were considered to be in contact if the distance between any two side-chain C atoms was within 4.5 Å. The two terminal secondary structures were considered to be in contact if there were at least two residues (from either structures) which had atoms within 4.5 Å of each other. The protein structures were classified according to the convention of *CATH* (Orengo *et al.*, 1997). Cartoon representations of molecules were generated using *MOLSCRIPT* (Kraulis, 1991).

3. Results

α-Helices and β-strands which start or end within the first or last six residues of the polypeptide chains, with no regular secondary structure other than the turn conformation preceding or following along the chain, have been identified. The numbers of α-helices found in the N- and C-terminal regions in 432 chains are 74 and 134, respectively; the corresponding numbers for β-strands are 152 and 114.

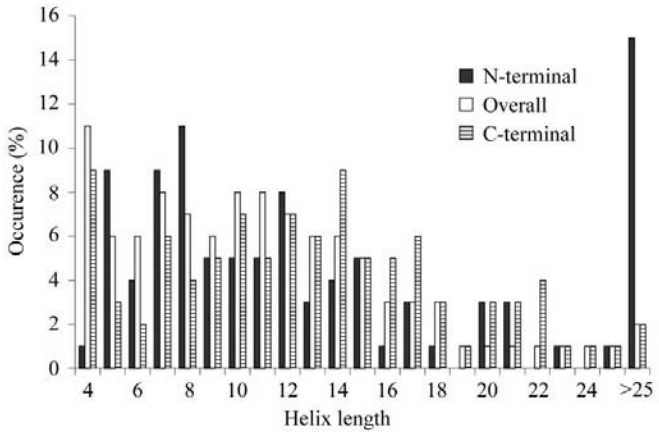

Figure 1
Frequency of occurrence of helices of different lengths in the whole structure (overall) and in the N-terminal and C-terminal regions.

3.1. Helix length

The distribution of length for all the helices (2275 in total) in the structures (Fig. 1) essentially follows the features noted earlier (Barlow & Thornton, 1988; Zhu & Blundell, 1996), with most of the helices being 12 residues or less in length. There seems to be a shift towards a smaller number of residues in the helices in the N-terminal region (20% of helices being of length 7 or 8), whereas at the other end the helices are longer, about a third of them are of length 13–17. Penel *et al.* (1999) observed that helices show a preference for an integral number of turns, with 'favoured lengths' of 6, 7, 10, 11, 13, 14, 17, 18, 21, 22, 24, 25, 28, 29 or 31 residues (the omitted numbers are of 'disfavoured length'). Although there are peaks at 4, 7, 10 and 11, the overall frequency distribution in our case is not as clear. [It needs to be pointed out that Penel *et al.* (1999) consider 3_{10}-helices, which are often found at the termini of α-helix, as part of the α-helix, while we exclude these when calculating the helix length.] However, the peaks of the distributions for the terminal regions are much more distinct. The peaks for the C-terminal region occur at lengths 4, 7, 10, 14 and 17, while those for the N-terminal region are shifted, occurring at 5, 8, 12, 15 and 20.

The N-terminal region is conspicuous in having a large percentage of long helices of length greater than 25 residues. In fact, out of 41 such helices in the whole database, 11 are in the N-terminal region and three are in the C-terminal region. The N-terminal long helices are from structures containing an up–down bundle of helices or with few secondary structures. Interestingly, in the latter category, when the structure has two helices, the C-terminal helix is much shorter than the N-terminal helix (PDB codes 1lts, 1mof, 1nkd, 1svf and 1wdc).

3.2. α-Helix propensity

Only residues with significant z values ($\geq$1.96) were selected to find residue-propensity differences in terminal secondary structures relative to those located in other regions. Compared with the overall values, the propensity for N-terminal helix is significantly higher for Gln (Table 2*b*); Leu also has a larger value. (If we consider the N-terminal helices longer than 25 residues separately from the shorter helices, the above observation holds true for both groups.) For the C-terminal helices there is a slightly higher preference for Lys and Glu.

The beginning of the helix in the N-terminal region and the end of the helix in the C-terminal region are likely to be different from those in interior helices. As such, we grouped the first three (N1–N3) and last three (C1–C3) residues to form the N-end and C-end regions of helix and compared the distribution of residues in these two ends in terminal helices with those found in these ends in all helices taken together. For this, we computed the local helix propensities for these

Table 2
Propensities and z values of residues in N- and C-terminal β-strands, α-helices (and some regions and positions relative to the helix), compared with the respective secondary structures in general.

P stands for propensity and *Z* for the z values of the amino-acid residues. Only the residues which are over-represented ($Z \geq 1.96$) in the terminal secondary structures are tabulated.

(*a*) β-Strand.

	Overall		N-terminal		C-terminal	
Residue	*P*	*Z*	*P*	*Z*	*P*	*Z*
Phe	1.5	13.0	1.6	3.2	1.6	2.8
Tyr	1.5	13.1	1.7	4.0	1.4	1.9
Ile	1.8	25.0	2.0	6.7	1.6	3.3
Val	1.9	33.9	2.2	9.2	2.1	7.4
Thr	1.2	7.4	1.4	3.0	1.3	2.1

(*b*) α-Helix.

	Overall		N-terminal		C-terminal	
Residue	*P*	*Z*	*P*	*Z*	*P*	*Z*
Ala	1.5	22.6	1.4	3.6	1.6	7.4
Glu	1.4	14.0	1.4	2.8	1.5	5.9
Gln	1.3	10.8	2.2	7.6	1.4	3.0
Lys	1.2	7.1	1.3	2.2	1.4	4.5
Leu	1.4	16.8	1.6	5.5	1.2	3.0

(*c*) N-end.

	Overall		N-terminal	
Residue	*P*	*Z*	*P*	*Z*
Ala	0.9	−3.8	1.4	2.2
Pro	7.5	50.0	3.4	3.7
Glu	1.4	10.5	1.7	2.9
Asp	1.5	10.4	1.7	2.1

(*d*) C-end.

	Overall		C-terminal	
Residue	*P*	*Z*	*P*	*Z*
Ser	1.1	1.0	1.7	2.9
Lys	1.3	5.8	1.6	3.3
Leu	1.2	4.4	1.3	2.1

Table 2 (continued)
(*e*) Ncap.

	Overall		N-terminal	
Residue	*P*	*Z*	*P*	*Z*
Ser	2.4	17.7	2.4	3.1
Asp	2.4	16.3	2.7	3.7
Asn	2.3	14.8	4.0	5.8
Thr	2.0	11.5	3.0	4.2

(*f*) N′.

	Overall		N-terminal	
Residue	*P*	*Z*	*P*	*Z*
Met	1.5	4.0	5.3	4.8
Leu	1.2	3.5	2.8	3.9

(*g*) Ccap.

	Overall		C-terminal	
Residue	*P*	*Z*	*P*	*Z*
Ala	1.2	3.1	2.2	4.9

(*h*) C′.

	Overall		C-terminal	
Residue	*P*	*Z*	*P*	*Z*
Gln	1.0	−0.5	2.1	2.5
Lys	1.8	4.4	2.9	3.2

ends (Chakrabarti & Pal, 2001). Given that a residue is in a helix, the value would indicate how likely it is to be located in a particular end compared with the rest of the helix; a value of one indicates no preference for an amino acid (the values will be less than one if an amino acid is disfavoured and greater than one if it is favoured) in a particular end of helix. The overall distribution for the two ends (Table 2*c* and 2*d*) is essentially the same as has been reported earlier (Chakrabarti & Pal, 2001). However, the N-end of N-terminal helices has a lower preference for Pro and a higher preference for Ala: in fact, based on z values, from being under-represented in general helices Ala becomes over-represented in terminal helices. Similarly, the C-end of C-terminal helices shows a higher local propensity for Ser, Lys and Leu. If the local propensities for the three individual positions at the two ends are considered, Pro is observed overwhelmingly at N1 of N-terminal helices and Ala at N1 and N2, while values for Glu and Asp increase from N1 through N3. No such position-specific preferences are observed at the C-end of C-terminal helices.

3.3. Residues flanking terminal helices and their conformation

The conventional nomenclature for helices and their flanking residues is as follows (Aurora & Rose, 1998):

$$-\mathrm{N''}-\mathrm{N'}-\mathrm{Ncap}-\mathrm{N1}-\mathrm{N2}-\mathrm{N3}-\cdots-\mathrm{C3}-\mathrm{C2}-\mathrm{C1}-\mathrm{Ccap}-\mathrm{C'}-\mathrm{C''}-,$$

where N1 through C1 define the helix, flanked by the primed residues. Ncap and Ccap are bridge residues, with non-helical φ, ψ angles, but nevertheless make one additional intrahelical hydrogen bond. Residues with side chains capable of providing a hydrogen bond to the unsatisfied main-chain groups at the beginning (Asn, Asp, Ser and Thr) or end (His and Asn) of α-helices are frequently found at these positions (Argos & Palau, 1982; Kumar & Bansal, 1998; Richardson & Richardson, 1988) and numerous experiments demonstrate that capping stabilizes helices in proteins (Bell *et al.*, 1992; Serrano & Fersht, 1989; Thapar *et al.*, 1996).

The general trend in propensities of residues to occur at N′, Ncap, Ccap and C′ positions in all helices taken together is similar to that seen in the values calculated by Kumar & Bansal (1998) for positions within and around helices and for the Ncap position given by Doig *et al.* (1997) (data not shown).

However, for the N-terminal helices the propensities of good Ncap residues (notably, Asn, Asp and Thr) increase, while that for Ser remains the same (Table 2*e*). At the adjacent N′ position, Met and Leu exhibit a large propensity value. Even at the other end of the helix, there are some differences in the propensities observed for C-terminal helices compared with the overall values (Table 2*g* and 2*h*): there is an increase for Ala at the Ccap position and for Lys and Gln at C′.

Gly, which is clearly over-represented at the Ccap and C′ positions of general helices (*z* values of 16.9 and 17.4, respectively), as has also been noticed by Kumar & Bansal (1998), does not occur as abundantly in these positions of C-terminal helices (*z* values of 0.6 and 1.7). About 40% of all helices terminate with a residue (overwhelmingly Gly) with conformation in the left-handed helical region (Gunasekaran *et al.*, 1998; Karpen *et al.*, 1992); in fact, Gly has a high occurrence at both Ccap and C′ positions (Kumar & Bansal, 1998). There are two recurrent capping motifs, the Schellman and the α_L motif (Schellman, 1980; Aurora *et al.*, 1994), with their typical patterns of hydrogen bonding and hydrophobic interactions but a common feature, which is a positive value of the backbone dihedral angle, φ, of the residue at C′, normally occupied by Gly. In contrast, in the C-terminal helices, where the φ, ψ angles at C′ can be calculated based on the available coordinates, ∼25% of residues have a positive φ (data not shown) and only 19% residues are Gly. This, along with the absence of many downstream residues, does not allow the occurrence of typical C-terminal capping motifs in C-terminal helices.

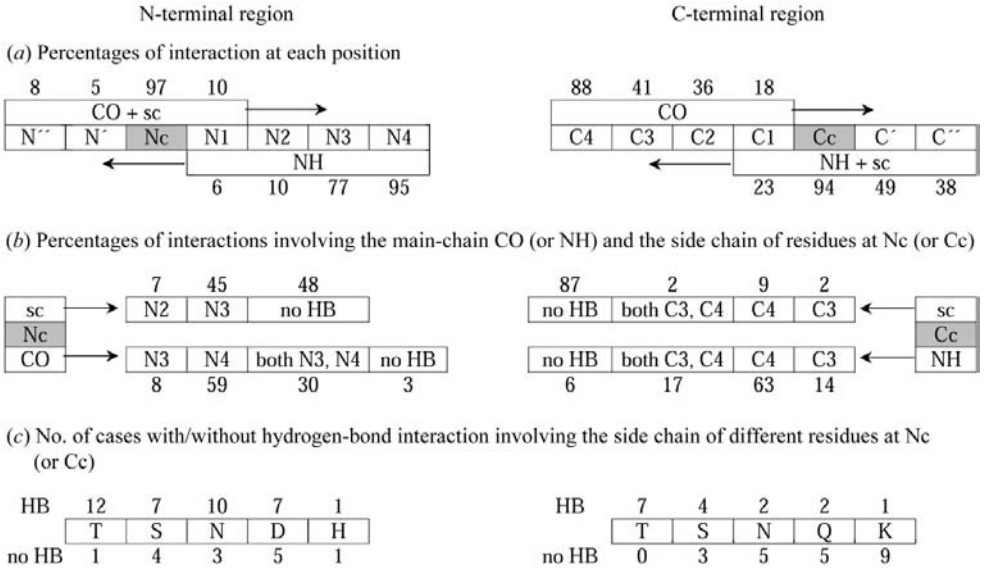

Figure 2
(*a*) Hydrogen-bond interactions around the N-terminal region of N-terminal helices and the C-terminal region of C-terminal helices. CO and NH stand for main-chain groups (arrows point to the other residues they are interacting with); sc means that the side chain acts as a hydrogen-bond donor (interacting with CO) or acceptor (with NH). The coordinates for all the positions, N′′, N′, C′ and C′′, are not available for all helices and the total number found at these positions are used to calculate the percentages. Ncap and Ccap positions (shown here as Nc and Cc, respectively) are shaded and the details of their interactions are given in (*b*); the different residues at these positions having at least one case of side-chain interaction are enumerated in (*c*). Other than the Nc position (and to some extent, Cc), other locations do not generally use sc for intrahelical hydrogen bonding.

3.4. Helix-capping interactions in the terminal helices

The fulfilment of the hydrogen-bonding potential of the NH groups of positions N1–N4 of N-terminal helices and the CO groups of positions C1–C4 of C-terminal helices by three flanking residues is shown in Fig. 2. It can be seen that the majority of groups at positions N/C1 and N/C2 are not hydrogen bonded, the situation being more pronounced at the N-terminal end. In addition to the 18% of the CO groups at C1 being engaged by flanking residues, a further 10% are hydrogen bonded to a remote group beyond the helix; such an interaction with a remote residue is less common at other positions. Approximately 95% of the residues at the Ncap and Ccap positions are involved in interaction with one or more main-chain atoms of residues at the two helical ends; the main chain is normally involved, but the side chain of about 50% Ncap residues is also utilized (Fig. 2*b*). The most common pattern associated with Ncap is the interaction of its main chain with N4 and the side chain with N3. Not all the side chains with an acceptor group at Ncap are used in hydrogen bonding (Fig. 2*c*). Almost all Thr side chains participate in hydrogen bonding and this is true even for the Ccap position.

Beyond the Ncap and Ccap positions of the helix, very few residues at the N′ and N′′ positions are involved in capping through hydrogen bonds, whereas at the other end about half of the C′ residues and 38% of the C′′ residues take part in hydrogen bonds (Fig. 2*a*). Thus, compared with residues upstream of the N-terminal helices, more downstream residues are involved in capping interactions for the C-terminal helices.

Helix-capping motifs have been identified both on the basis of patterns in hydrogen bonding and hydrophobic interactions (Aurora & Rose, 1998). If *h*, *p* and *x* represent a hydrophobic, polar and any residue, respectively, then two common N-terminal motifs can be represented as *hx-pxhx* and *hx-pxph* (where positions before the hyphen are N′ and Ncap, and N1–N4 follow the hyphen). There is a hydrophobic interaction between N′ and N3 in the former and N′ and N4 in the latter. As the propensity of two hydrophobic residues Met and Leu to occur at the N′ position is very high (Table 2*f*), we checked whether these residues contribute to the hydrophobic interaction present in the two motifs. Indeed, in 18 of the 20 cases involving Met and Leu at N′, the residues have hydrophobic contacts with the N4 residue (N3 being polar) or with both N3

and N4. The hydrophobic residues in the most common Schellman motif (*hxp-x*G*h*, G being Gly) observed in C-terminal capping occur at positions C3 and C″ (Schellman, 1980; Aurora & Rose, 1998). There are only four cases of Gly (at C′) followed by a hydrophobic residue in C-terminal helices. Thus, hydrophobic interaction is not likely to contribute to C-terminal capping.

3.5. β-Sheet type

The involvement of the terminal regions in different types of β-sheets is given in Table 3, along with the overall pattern. The changes that take place at the two terminals relative to the overall values are clearly reflected in the propensity values. The C-terminal region is less likely to be found in parallel β-sheets and is not found at all as the interior strand of such a sheet. It has a higher propensity to be the edge strand of antiparallel β-sheets. It may also be mentioned here that Sternberg & Thornton (1977) have also observed that the C-terminal strands preferentially occupy an 'external' position; however, they did not explicitly state how the terminal strands were defined and the type of β-sheet they were located in was not considered. The N-terminal region, on the other hand, shows a higher propensity to occur in the interior of parallel β-sheets. The occurrence of both the terminal regions on strands in mixed β-sheets (a parallel strand on one side and an antiparallel on the other) is higher compared with that overall, although the absolute percentage of occurrence is still quite low (7–8%).

3.6. β-Strand length and β-sheet propensity

The overall frequencies of occurrence of β-strands decrease as their lengths become longer (Fig. 3) (Penel *et al.*, 1999; Zhu & Blundell, 1996). The frequencies of strand lengths with even and uneven residue numbers do not differ significantly, contrary to the conclusions of Sternberg & Thornton (1977). However, marked periodicity is observed when the terminal strands are considered and there are distinctions even between the two termini. The distribution for the N-terminal strands has peaks at odd numbers, while that for the opposite end show maxima at even numbers. This preference may be a reflection of the type of β-sheet in which the two terminal regions are involved. 85% of C-terminal strands are found in antiparallel β-sheets, while about a third of N-terminal strands

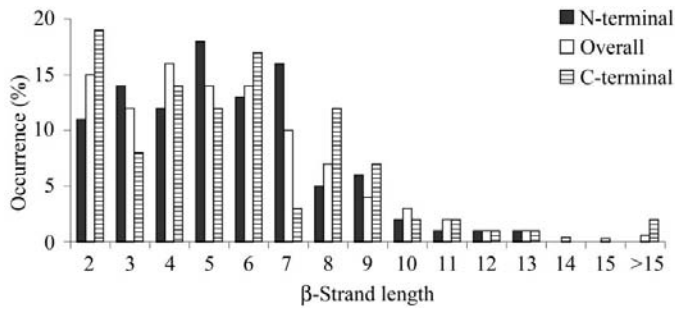

Figure 3
Frequency of occurrence of β-strands of different lengths in the whole structure (overall) and in the N- and C-terminal regions.

Table 3
Percentage composition of different types of β-strands.

In each row the total number of occurrence is given, followed by the percentage distribution in different types of β-strands. The propensity value is obtained as the ratio of the composition of the terminal to the overall value (given in the same column).

		Antiparallel		Parallel		
	Total No.	Interior	Edge	Interior	Edge	Mixed
Overall	3477	19.8	49.6	11.3	13.8	5.6
N-terminal	152	11.8	49.3	18.4	12.5	7.9
C-terminal	114	17.5	67.5	0	7.9	7.0
Propensity						
N-terminal		0.60	0.99	1.63	0.91	1.41
C-terminal		0.88	1.36	0.0	0.57	1.25

are in parallel sheets (Table 3). In an antiparallel sheet, if one residue provides a pair of hydrogen bonds to one side, the adjacent residue provides another pair to the other side (Richardson, 1981). Consequently, in an antiparallel sheet with three or more strands, two residues from each strand are needed to be in register so that the first pair of hydrogen bonds can link the edge strands to the one inside; with the addition of two more residues the next pair of hydrogen bonds can be formed and so on. This may be the reason why the C-terminal strands, mostly found in antiparallel sheets, prefer even numbers of residues. In a parallel β-sheet, the NH and CO groups of one residue in a strand form hydrogen bonds with the CO and NH groups of two different residues with a residue in between; *i.e.* odd numbers (three against one) of residues need to be in register. Consequently, the N-terminal strands, with a higher percentage of occurrence in parallel β-sheets, have preferences for lengths with an odd number of residues.

Compared with the N-terminal strands, a higher percentage of strands in the C-terminal region are found to have lengths of $\geq$8 residues. This may also be a consequence of the preference of such strands to occur in antiparallel β-sheets, which are likely to contain longer strands than those found in parallel β-sheets (Richardson, 1981). Changes in the β-sheet propensity of residues on going from the overall structure to the two terminal regions occurring in β-sheet (Table 2*a*) generally mirror the trend that one can associate with the type of β-sheets preferred at the two termini. Thus, Val, Ile and Tyr, which have high propensity values for parallel β-sheets (Chakrabarti & Pal, 2001) also have high values in the N-terminal region.

3.7. Contact between terminal secondary structures

Table 4 provides the number of cases with four distinct combinations of the two elements of terminal secondary structure. A β-strand occurring at both termini is found to be the most common and a chain starting with an α-helix and ending with a β-strand is least often observed. 38 cases of the αα structure are found in the ratio 3:1 in all-α and α/β classes of proteins, whereas 58 cases of the ββ pattern are observed in all-β and α/β classes in the ratio 1.5:1.

The proximity between the terminal regions in the three-dimensional structures has been considered previously (Christopher & Baldwin, 1996; Ptitsyn, 1981; Thornton & Sibanda, 1983) without any particular reference to their secondary structures. The percentages of different terminal secondary structures which are close to each other are given in Table 4. It can be seen that the contact between the structural elements is larger (60.5%) with a helix occurring at both ends, while with different types of secondary structures occupying the two ends the contact can take place in only about 20% of cases. Furthermore, consideration of domains present in different molecules suggests that only for single-domain proteins can the terminal secondary structures be in contact. Two examples of $\alpha\alpha$ and $\beta\beta$ structures in contact are shown in Fig. 4.

Table 4
Secondary structures (at two termini) in contact.

	No. of cases (No. in contact)					
		No. of domains†			% in contact, relative to	
Secondary structures‡	Total	1	2	3	Total No.	No. of single-domain proteins
$\alpha\alpha$	38 (23)	37 (23)	1		60.5	62.2
$\beta\beta$	58 (25)	53 (25)	5		43.1	47.2
$\alpha\beta$	11 (2)	9 (2)	2		18.1	22.2
$\beta\alpha$	34 (7)	26 (7)	7	1	20.6	26.9

† The total number of cases is broken up according to the number of domains [as given in *CATH* (Orengo *et al.*, 1997)] each protein contains. In six out of the 16 cases of multi-domain proteins, the two terminal secondary structures are present in separate domains. ‡ The two terminal secondary structures are indicated.

4. Discussion

The length and stability of secondary structures are considerably controlled by tertiary interactions. Consequently, a great deal of structural information needed for the design of small proteins can be gleaned from an analysis of secondary structures at two ends of polypeptide chains, which are likely to be influenced to a lesser degree by the tertiary context than the protein interior. Moreover, the disposition of the terminal secondary-structural elements relative to the overall fold and the proximity of the two terminal structures in a chain provide valuable information on the folding process.

4.1. α-Helix

Compared with the overall distribution, the N-terminal helices in proteins tend to be shorter and the C-terminal helices longer (Fig. 1). However, in short polypeptide chains made up of very few secondary structures (two helices), the N-terminal helix is found to be longer than the C-terminal helix. Some examples can be seen in the structures of enterotoxin (PDB code 1lts; Sixma *et al.*, 1993), viral coat protein (1mof; Fass *et al.*, 1996), paramyxovirus SV5 fusion protein (1svf; Baker *et al.*, 1999) *etc*. It is likely that the asymmetry in the length of the two helices in the polypeptide chain may be a consequence of the biological role of these segments.

As first noticed by Srinivasan (1976), helices (over the whole structure, as well as the terminal helices) are found to occur with a near-integral number of turns (the distribution for the N-terminal helices differs to some extent from the other two categories in Fig. 1). Penel *et al.* (1999) suggested that with a favoured length, hydrophobic residues occupy optimum positions in the C-terminal end of the helix, leading to its termination with a Schellman or α_L C-capping motif. By this logic, the periodicities in length should not be present in isolated helices. As we observe the periodicity in C-terminal helices in general lacking typical C-capping motifs (discussed later), the existence of the latter cannot be the correct explanation. Rather, an integral number of helix turns places the two terminal residues in a helix on the same side, enabling the residues beyond to make efficient contact with the rest of the molecule. A surface helix is less likely to have a disfavoured length, which will make the chain enter and exit the helix on two opposite sides.

The propensities of residues to occur in and around α-helices can be different in helices at polypeptide-chain termini. The propensity of Gln and Leu to be in the N-terminal helix is considerably higher, while the values for Lys and Glu increase in the C-terminal helices (Table 2). Grouping the three residues at the N1–N3 positions together to define

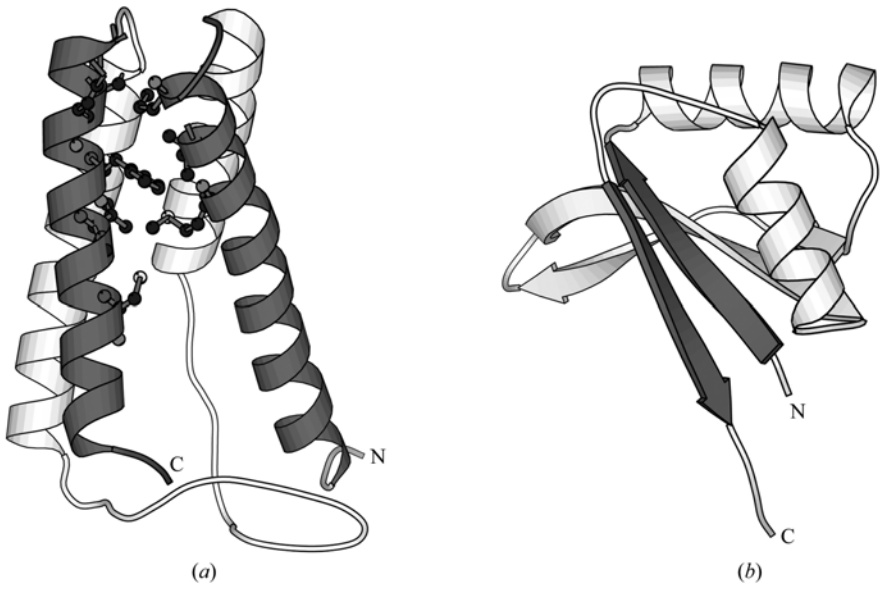

Figure 4
Two examples where the secondary structures from the two ends of the chain are in contact: (*a*) α-helices in PDB file 1bbh (subunit *A*) (side chains in contact are shown), (*b*) β-strands from 2bop (subunit *A*).

the N-end and grouping C1–C3 to define the C-end, residues within helices are distributed in such a way that compared with the general distribution the N-end of N-terminal helices has a lesser preference for Pro and a higher preference for Ala, while the C-end of C-terminal helices has a higher preference for Ser, Lys and Leu. Of the good Ncap residues, Asp, Asn, Thr and Ser, the first three show an increase in propensity in N-terminal helices, indicating that this helix-start signal is present in a greater number of N-terminal helices. These residues can use their side chain to form a hydrogen bond with the main-chain NH group at N3 (Fig. 2). At the adjacent N′ position, Met and Leu show a remarkable increase in propensity (Table 2). These residues are involved in hydrophobic interaction with the residue at N3 and/or N4, which is a prevalent N-terminal capping motif (Aurora & Rose, 1998). As mentioned in §1, Met is very common as the first residue in a polypeptide chain and for N-terminal helices the residue can also be used to provide stability to the structure through capping interactions.

Relative to other residues, Lys has a higher than expected number of occurrences at the last position of the polypeptide chain and also at nearby positions (Berezovsky *et al.*, 1999). This may be a consequence of the structural role the residue plays near the helical C-terminus. For C-terminal helices, Lys is found more near the carboxy end and at C′ (Table 2*d* and 2*h*). Compared with the overall value, the propensity of Gly to occur at the Ccap and C′ positions of C-terminal helices is lesser. As a result, the Schellman motif of C-terminal capping, which needs a Gly residue (in left-handed α-helical conformation) at C′, is not common for such helices. In connection with this, it can be mentioned that experimental analysis of peptides containing the Schellman motif has not revealed any strong contribution to helix stability or its ability to be formed without the presence of the helix-promoting solvent trifluoroethanol (Viguera & Serrano, 1995). It appears that the Schellman motif is important in the tertiary context of the structure in leading the chain out of the helix in the right direction and is not required for capping an isolated helix. The capping at the C-terminal end is mostly satisfied by hydrogen bonding involving the main-chain atoms at C′ and C″ after the helix. At the N-terminal end, on the other hand, N′ and N″ positions before the helix are hardly used (Fig. 2). This suggests a lesson for the design of isolated helical peptides: beyond the helix there is a need to have a greater number of residues at the C-terminal end than at the N-terminal end. The need to have about two residues after the end of the helix may be the reason why a larger proportion of residues in the last two positions of the polypeptide chain are found in β-sheet than in α-helix (Pal & Chakrabarti, 2000).

4.2. β-Sheet

Like the C-terminal helices, the β-strands in the C-terminus tend to be longer than the average length of strands (Fig. 3). Compared with the general distribution, the C-terminal strand has a higher propensity to occur at the edge of antiparallel β-sheet, whereas the N-terminal strand has a preference to occur as the interior strand of parallel β-sheet, a location totally avoided by a strand from the other end of the chain (Table 3). Strands at both the ends have a slightly higher propensity to be located in mixed β-sheets. The relatively higher preference of N-terminal strands for parallel sheets and C-terminal strands for antiparallel sheets may result in differences in periodicities observed in the distributions of strand lengths at the two termini. The N-terminal strands have an inclination to contain odd numbers of residues, while the C-terminal strands are inclined to contain even numbers (Fig. 3); this may be a reflection of the minimum numbers of residues that are needed to generate the hydrogen-bonding pattern observed in parallel (three residues) and antiparallel (two residues) β-sheets. (This has been elaborated on in §3.) Owing to the higher occurrence of N-terminal strands in parallel β-sheets, residues (Val, Ile and Tyr) which have a higher propensity to occur in such sheets are also found to occur more in the N-terminal strands (Table 2*a*).

4.3. Protein folding

Although the proposition of Anfinsen (1973) that the conformation of a protein is determined solely from its amino-acid sequence and that the folding of the peptide chain to the native three-dimensional structure is spontaneous are generally accepted, a detailed picture of the folding process is still missing. In this context, there are two possibilities: either the nascent peptide chain of a protein folds while it is being synthesized, or the folding occurs only after the synthesis of the chain has been completed. In other words, folding could be cotranslational or post-translational (Tsou, 1988). It has been shown with the β-subunit of bacterial luciferase that protein folding does occur cotranslationally (contributing to the rapid formation of the native structure in the cell), suggesting that the emerging polypeptide may adopt conformations that predispose the folding of the complete polypeptide (Fedorov & Baldwin, 1995). In other words, a sequential folding mechanism may be operative, at least for the multidomain complex proteins, and this significantly reduces the possibility of misfolding (Netzer & Hartl, 1997) and is kinetically more favourable than a non-sequential folding process (Tsai *et al.*, 1999; Rumbley *et al.*, 2001). From a statistical analysis of known protein structures, Alexandrov (1993) has shown that the N-terminal region is more compact than the C-terminal; this was taken to be an indication that proteins fold during their synthesis beginning from the N-terminus, so that the N-terminal residues interact more among themselves. Results presented in Table 3 provide another structural argument for the N-terminal initiation of protein folding (especially for proteins rich in β-sheets). A higher percentage of C-terminal strands are found at the edge of β-sheets than in the interior, indicating that in a large number of structures the C-terminal end of the polypeptide chain aligns itself to the edge of an already-formed β-sheet.

Data provided in Table 4 offer insight into another issue related to the folding process, *viz*. whether the interactions between the secondary structures at the two ends of the chain

can guide the process. Ptitsyn (1981) suggested that interactions between terminal regions may be important in the folding pathway. At least in some systems the two termini may be brought into close proximity as a result of the steps involved in biosynthesis. Thus, the heat-shock protein SSB (Hsp70) has been shown to associate with translating ribosomes and to dissociate upon the addition of an inhibitor (puromycin) of protein synthesis (Nelson *et al.*, 1992), which may indicate that the termini are essentially tethered to the ribosome and are released from the ribosome–SSB complex at the same time. Table 4 shows that when the polypeptide chain contains two helices at the termini, in 60.5% of cases the two helices are in contact (Fig. 4). Only in single-domain proteins can the terminal secondary structures interact. Chances of interactions diminish (43.1%) with β-strands at two termini and are rather low when one secondary structure is α and the other is β. This suggests that there may be cooperativity in the formation of the terminal helices in the all-α class of single-domain proteins. In agreement with this hypothesis, the N- and C-terminal helices in cytochrome *c* are formed in the early phase of folding (Roder *et al.*, 1988) and are also present in the compact cooperatively folded intermediate characterized during the refolding of apomyoglobin (Jennings & Wright, 1993).

5. Summary

The effect of tertiary interactions on secondary structures can be clearly visualized in a situation where there are very few such interactions. The terminal regions, with one free end, provide an ideal context to study the features that will be useful for the design of polypeptides with few secondary structures. This analysis has shown that compared with the overall values there can be differences in the propensities of residues to occur in and around terminal helices. C-terminal helices do not exhibit the typical capping interactions. There are periodicities in the distribution of lengths for the terminal secondary structures. Terminal strands have different propensities to be in the interior and at the edge of β-sheets. This difference in the relative location of terminal strands and the observation that the two terminal secondary structures can be in contact with each other have implications for the folding mechanism.

We thank the Department of Biotechnology for computational facilities and the Council of Scientific and Industrial Research for a grant.

References

Alexandrov, N. (1993). *Protein Sci.* **2**, 1989–1991.
Anfinsen, C. B. (1973). *Science*, **181**, 223–230.
Argos, P. & Palau, J. (1982). *Int. J. Pept. Protein Res.* **19**, 380–393.
Aurora, R. & Rose, G. D. (1998). *Protein Sci.* **7**, 21–38.
Aurora, R., Srinivasan, R. & Rose, G. D. (1994). *Science*, **264**, 1126–1130.
Bachmair, A., Finley, D. & Varshavsky, A. (1986). *Science*, **234**, 179–186.
Baker, K. A., Dutch, R. E., Lamb, R. A. & Jardetzky, T. S. (1999). *Mol. Cell*, **3**, 309–319.
Barlow, D. J. & Thornton, J. M. (1988). *J. Mol. Biol.* **201**, 601–619.
Bell, J. A., Becktel, W. J., Sauer, U., Baase, W. A. & Matthews, B. W. (1992). *Biochemistry*, **31**, 3590–3596.
Berezovsky, I. N., Kilosanidze, G. T., Tumanyan, V. G. & Kisselev, L. L. (1999). *Protein Eng.* **12**, 23–30.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235- 242.
Bhattacharyya, R., Samanta, U. & Chakrabarti, P. (2002). *Protein Eng.* **15**, 91–100.
Björnsson, A., Mottagui-Tabar, S. & Isaksson, L. A. (1996). *EMBO J.* **15**, 1696–1704.
Brown, C. M., Dalphin, M. E., Stockwell, P. A. & Tate, W.P. (1993). *Nucleic Acids Res.* **21**, 3119–3123.
Brown, C. M., Stockwell, P. A., Trotman, C. N. A. & Tate, W. P. (1990). *Nucleic Acids Res.* **18**, 2079–2086.
Chakrabarti, P. & Pal, D. (2001). *Prog. Biophys. Mol. Biol.* **76**, 1–102.
Chou, P. Y. & Fasman, G. D. (1974). *Biochemistry*, **13**, 211–222.
Christopher, J. A. & Baldwin, T. O. (1996). *J. Mol. Biol.* **257**, 175–187.
Dahiyat, B. I. & Mayo, S. L. (1997). *Science*, **278**, 82–87.
DeGrado, W. F. (1997). *Science*, **278**, 80–81.
Doig, A. J., MacArthur, M. W., Stapley, B. J. & Thornton, J. M. (1997). *Protein Sci.* **6**, 147–155.
Fass, D., Harrison, S. C. & Kim, P. S. (1996). *Nature Struct. Biol.* **3**, 465–469.
Fedorov, A. N. & Baldwin, T. O. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 1227–1231.
Gunasekaran, K., Nagarajaram, H. A., Ramakrishnan, C. & Balaram, P. (1998). *J. Mol. Biol.* **275**, 917–932.
Hirel, P. H., Schmitter, J., Dessen, P., Fayat, G. & Blanquet, S. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 8247–8251.
Hobohm, U. & Sander, C. (1994). *Protein Sci.* **3**, 522–524.
Jennings, P. A. & Wright, P. E. (1993). *Science*, **262**, 892–896.
Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
Karpen, M. E., De Haseth, P. L. & Neet, K. E. (1992). *Protein Sci.* **1**, 1333–1342.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Kumar, S. & Bansal, M. (1998). *Proteins*, **31**, 460–476.
McDonald, I. K. & Thornton, J. M. (1994). *J. Mol. Biol.* **238**, 777–793.
Nakamura, Y., Ito, K. & Isaksson, L. A. (1996). *Cell*, **87**, 147–150.
Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M. & Craig, E. A. (1992). *Cell*, **71**, 97–105.
Netzer, W. J. & Hartl, F. U. (1997). *Nature (London)*, **388**, 343–349.
Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.
Pal, D. & Chakrabarti, P. (2000). *Biopolymers*, **53**, 467–475.
Penel, S., Morrison, R. G., Mortishire-Smith, R. J. & Doig, A. J. (1999). *J. Mol. Biol.* **293**, 1211–1219.
Presta, L. G. & Rose, G. D. (1988). *Science*, **240**, 1632–1641.
Ptitsyn, O. B. (1981). *FEBS Lett.* **131**, 197–202.
Regan, L. & Wells, J. (1998). *Curr. Opin. Struct. Biol.* **8**, 441–442.
Richardson, J. S. (1981). *Adv. Protein Chem.* **34**, 167–339.
Richardson, J. S. & Richardson, D. C. (1988). *Science*, **240**, 1648–1652.
Roder, H., Elöve, G. A. & Englander, S. W. (1988). *Nature (London)*, **335**, 700–704.
Rumbley, J., Hoang, L., Mayne, L. & Englander, S. W. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 105–112.
Schellman, C. (1980). *Protein Folding*, edited by R. Jaenicke, pp. 53–61. New York: Elsevier.
Serrano, L. & Fersht, A. R. (1989). *Nature (London)*, **342**, 296–299.
Sixma, T. K., Kalk, K. H., van Zanten, B. A. M., Dauter, Z., Kingma, J., Witholt, B. & Hol, W. G. J. (1993). *J. Mol. Biol.* **230**, 890–918.
Srinivasan, R. (1976). *Ind. J. Biochem. Biophys.* **13**, 192–193.
Sternberg, M. J. E. & Thornton, J. M. (1977). *J. Mol. Biol.* **110**, 285–296.
Tate, W. P. & Mannering, S. A. (1996). *Mol. Microbiol.* **21**, 213–219.

Thapar, R., Nicholson, E. M., Rajagopal, P., Waygood, E. B., Scholtz, J. M. & Klevit, R. E. (1996). *Biochemistry*, **35**, 11268–11277.
Thornton, J. M. & Chakauya, B. L. (1982). *Nature (London)*, **298**, 296–297.
Thornton, J. M. & Sibanda, B. L. (1983). *J. Mol. Biol.* **167**, 443–460.
Tsai, C.-J., Maizel, J. V. Jr & Nussinov, R. (1999). *Protein Sci.* **8**, 1591–1604.
Tsou, C.-L. (1988). *Biochemistry*, **27**, 1809–1812.
Tsunasawa, S., Stewart, J. W. & Sherman, F. (1985). *J. Biol. Chem.* **260**, 5382–5391.
Varshavsky, A. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 12142–12149.
Viguera, A. R. & Serrano, L. (1995). *J. Mol. Biol.* **251**, 150–160.
Zhu, Z.-Y. & Blundell, T. L. (1996). *J. Mol. Biol.* **260**, 261–276.