β-Sheet propensity and its correlation with parameters based on conformation

Debnath Pal and Pinak Chakrabarti*

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

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

The dispersion of the main-chain and side-chain conformations in the φ, ψ, χ_1 space for all residues have been estimated in terms of three parameters corresponding to the entropy (S) of the distribution, the volume (D_V) and the area (D_A) the points are enclosed in. These parameters are inversely correlated with Chou and Fasman β-sheet propensities, P_β (Gly and Pro excluded), suggesting that residues with greater dispersion in the conformational space are weak β-sheet formers. It was also found that different residues have different relative populations in the bridging region (intervening between the helical and β-sheet regions) which may lie on the pathway for interconversion between α and β conformations. The energy barrier for this transformation, as obtained from the population of residues in the bridging region relative to the β region, is directly correlated to P_β. Residues with high P_β have branched side chains, which have greater steric interactions with the main-chain atoms resulting in a shrinking of the available conformational space (first correlation) and a steeper energy gradient beyond the allowed space (second correlation) compared with linear residues. It is proposed that if residues exist in an extended conformation when the polypeptide chain is synthesized, a stretch of residues with high P_β, because of the high energy barrier for their conversion into the α conformation, will continue to remain in the extended conformation and will ultimately constitute a β-strand in the folded structure.

1. Introduction

Deciphering the rules by which amino-acid sequence determines protein folding is a major goal of biochemical research. Towards this goal, it is necessary to have a code based on physicochemical properties of individual residues that explains their secondary-structural propensities. A quantitative energetic description of many factors that determine the stability of α-helical secondary structure is available (Chakrabartty & Baldwin, 1995; Parthasarathy *et al.*, 1995). Our understanding of the elements that determine β-sheet stability is much less advanced. This also shows up in the lower success rate that is generally achieved in predicting β-strands than α-helices by the current algorithms for secondary-structure prediction. A likely explanation for this is that helices form locally, whereas β-sheets have elements of both secondary (a stretch of residues form a β-strand) and tertiary (two or more strands have to come together to constitute a β-sheet) structures. As a result, although different amino-acid residues have been found to have measurably different propensities for forming β-sheets (Kim & Berg, 1993; Minor & Kim, 1994*a*; Smith *et al.*, 1994), the value is also found to depend on

research papers

whether the residue is located in the central or the edge strand of the β-sheet (Garratt *et al.*, 1991; Minor & Kim, 1994*b*) and also on the interaction energy between cross-strand pairs of side chains (Smith & Regan, 1995; Wouters & Curmi, 1995). The effect of the context is, however, averaged out while determining the statistically derived Chou & Fasman (1978) type propensity values, which may thus be considered to be intrinsic to a residue independent of the surroundings. In this paper, we offer an explanation for the relative β-sheet propensities in terms of some conformational features of different residues.

Although the secondary structure is defined using only the main-chain torsion angles (φ and ψ; Ramachandran & Sasisekharan, 1968), because of the interdependence of these angles and the side-chain torsion angle χ_1 (Chakrabarti & Pal, 1998), any conformation-based parameter that tries to explain the propensity values must be derived using all three torsion angles. Indeed, we have recently shown how the distribution of points in φ, ψ, χ_1 conformational space can be used to estimate the loss of main-chain conformational entropy of different residues on protein folding (Pal & Chakrabarti, 1999). Some of the parameters derived, such as the absolute entropy (S) of the distribution of points for a residue and the volume occupied by the points (which we call the volume of dispersion, D_V), along with an equivalent area term (D_A), are shown here to be correlated with the β-sheet propensity. Information on energetics can also be obtained from these distributions, throwing some light on the mechanism for the formation of β-sheets. The factors introduced here may facilitate the development of better algorithms for the prediction of β-sheets and the manipulation of such structures.

2. Methods

Main-chain and side-chain torsion angles (φ, ψ, χ_1) were calculated (as previously reported; Pal & Chakrabarti, 1999) for a set of refined non-redundant protein structures from the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The same data set was used to identify, using *DSSP* (Kabsch & Sander, 1983), the residues located in β-sheets and calculate the Chou & Fasman (1978) type β-sheet propensities.

2.1. Entropy (*S*), volume (*D*_V) and area (*D*_A) of dispersion

The absolute entropy (S) and the volume of dispersion (D_V) for the distribution of φ, ψ, χ_1 points for individual residues (except Gly, Ala and Pro) were calculated as before (Pal & Chakrabarti, 1999), but with one minor modification. For reasons outlined previously (Pal & Chakrabarti, 1999), the g^- conformation of Leu was excluded in the computation of D_V; here, however, the whole χ_1 space was considered and this provided a higher D_V value for Leu.

D_V was calculated (Pal & Chakrabarti, 1999) by determining the means and the standard deviations of φ, ψ and χ_1 angles in individual clusters in the three-dimensional distribution of points; the product of the standard deviations in a

cluster gave its volume dispersion and the summation over all the clusters gave the total volume, which on normalization (dividing by the value of Ile, which has the smallest volume) provided D_V for a residue. To include residues (Gly, Ala and Pro) with no χ_1 torsion angle of the side chain or a restricted one, a similar calculation was performed in two-dimensional φ, ψ space to obtain D_A, the area of dispersion. Although χ_1 was dispensed with, its influence on φ, ψ was taken into account in an indirect manner by considering the φ, ψ distribution of non-Gly/Ala/Pro residues in the three discreet χ_1 conformational states (Chakrabarti & Pal, 1998): g^+ (combining all residues with χ_1 in the range −120 to 0°), g^- (0 to 120°) and t (−240 to −120°). The area enclosing the points in individual clusters was estimated as before and was summed over the whole φ, ψ space at a given χ_1 state. The weighted average (based on the population of each conformational states) of the three values provided the parameter D_A (normalized), which can be taken as the closest two-dimensional representation of the φ, ψ, χ_1 information. For Gly and Ala, D_A is just the area of the φ, ψ distribution; for Pro, two conformational states of the side chain were used (χ_1 value positive or negative).

2.2. Bridging region

The two prominent clusters of points in the φ, ψ map (Ramachandran & Sasisekharan, 1968) can be demarcated using the following ranges (Chakrabarti & Pal, 1998) in ψ: −120 to 60° and 60 to 240° (φ is in the range −180 to 0° for both). Residues with the α-helical conformation form a cluster in the former region and those with the β-sheet conformation cluster in the latter. The bridging region links the two clusters and was identified for each residue as follows. For non-Gly/Ala/Pro residues, the φ, ψ, χ_1 space (φ in the range −180 to 0° only) was divided into 10 × 10 × 10° grids. Grids containing one or more helical residues (any helical conformation, as identified by the program *DSSP*; Kabsch & Sander, 1983) were identified as α and those filled with β-sheet residues (with *DSSP* notation E) were tagged β. Any isolated grid (with no occupied adjacent grids) was neglected. The α and β clusters thus delineated had rough boundaries and small cavities in some instances. Consequently, smoothing was performed by filling in the grids which had at least two occupied neighbours. The total numbers of points with helical (α) and sheet (β) conformations were counted; the means of the ψ values and their standard deviations in the α and β clusters corresponding to the three conformational states of the side chain were computed. The number of points occupying the bridging region was determined by scanning the grids along the ψ direction (Fig. 1), starting at mean(α) + 1σ(α) and going up to mean(β) − 1σ(β). The result does not change in any significant way if 0.5σ or 1.5σ is used instead of 1σ to define the scan range. For Gly, Ala and Pro, as the distribution is considered in two dimensions, the grids tend to have a larger number points than in three dimensions and the α and β clusters become more diffused. As a result, a cut-off value [the

percentage of residues (having a negative value of φ) that must occupy a grid for it to be included; 0.05% for Gly and Ala and 0.01% for Pro] was used to define the core α and β regions.
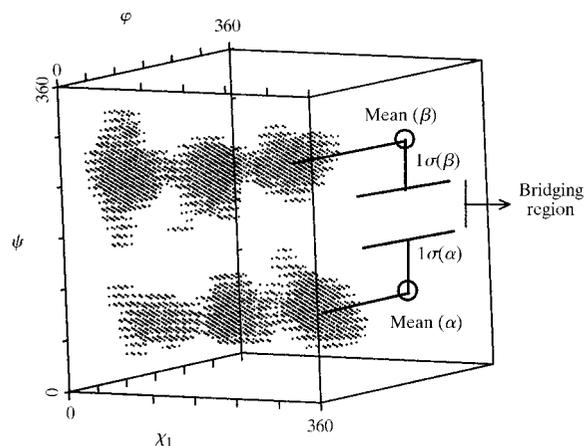
Assuming that the Boltzmann distribution governs the number (N) of points in different regions of the conformational space, the energy ($\Delta E = E_{bridging} - E_\beta$) of the bridging region relative to the β region is given by

$$N_{bridging}/N_\beta = \exp(-\Delta E/RT).$$

3. Results and discussion

3.1. Entropy, volume and area as measures of dispersion of points and correlation with β-sheet propensity

In an attempt to estimate the loss in conformational entropy on folding, we calculated the absolute entropy S of the distribution of points in the three-dimensional φ, ψ, χ_1 space for each residue (Pal & Chakrabarti, 1999). As another measure of this distribution we also calculated the volume, D_V, which encompasses these points. Interestingly enough, both these parameters show strong inverse correlation with P_β, the β-sheet propensity (Table 1 and Fig. 2), suggesting that the residues with a compact distribution have a high propensity value. As our method considered the combined φ, ψ, χ_1 distribution, it excluded residues with no or restricted χ_1 (Gly, Ala and Pro). To include these residues, a parameter D_A (involving area, instead of volume) was calculated. For the three special residues, D_A gives an indication of the spread of points in two dimensions; for the remainder, D_A also retains in an indirect way the influence of the side chain, as three φ, ψ distributions corresponding to the three χ_1 conformational states were used in its calculation. This factor D_A also has a high inverse correlation with P_β (Table 1). Compared with S or D_V, the use of D_A was only of minor advantage, as only Ala

Table 1
Parameters representing conformational dispersion and their correlation coefficients with β-sheet propensity values P_β.

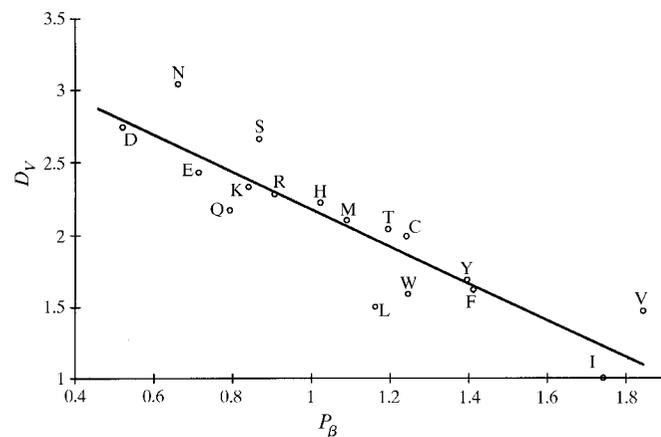Values of D_V and D_A are normalized relative to the smallest value in the set.

Parameters	P_β	D_V	Entropy S	D_A
Ser	0.87	2.66	6.84	2.89
Cys	1.24	1.99	6.23	2.79
Met	1.09	2.10	5.98	2.48
Glu	0.71	2.43	6.41	2.51
Gln	0.79	2.17	6.26	2.37
Lys	0.84	2.33	6.62	2.36
Arg	0.91	2.28	6.49	2.46
Leu	1.16	1.5	6.24	1.65
Asp	0.52	2.74	6.62	3.40
Asn	0.66	3.04	6.80	3.73
His	1.02	2.22	6.47	3.14
Phe	1.41	1.62	6.36	2.37
Tyr	1.40	1.69	6.45	2.54
Trp	1.25	1.59	6.09	2.13
Val	1.84	1.47	5.81	1.21
Ile	1.74	1.00	5.60	1.00
Thr	1.20	2.04	6.32	1.69
Pro	0.46	—	—	0.85
Ala	0.75	—	—	3.94
Gly	0.67	—	—	7.67
Correlation	—	−0.89	−0.77	−0.77†

† Excluding Gly and Pro.

could be added to the linear relationship with P_β; the exclusion of Gly and Pro is not unexpected, as the former, with points distributed in all the four quadrants of the φ, ψ map (Ramachandran & Sasisekharan, 1968), has a D_A much larger than the rest, whereas the latter, with a restricted φ, has a smaller D_A.

3.2. Bridging region

Residues which have a more diffused distribution of points (larger S, D_V or D_A) are poor β-sheet makers. One may realise the physical significance of this observation by looking at the χ_1-dependent φ, ψ plots (Chakrabarti & Pal, 1998) of these
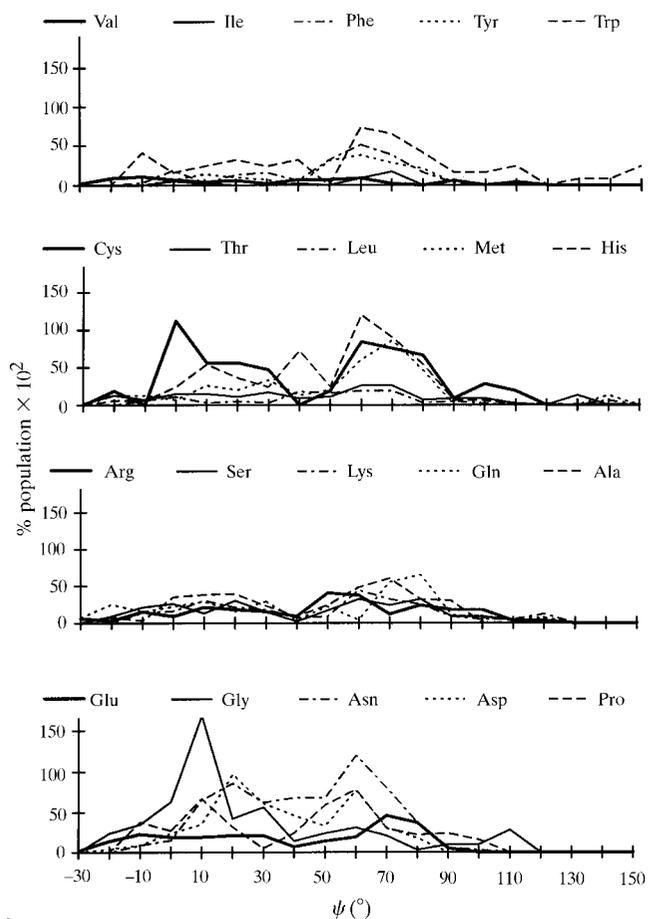

Figure 1
Definition of the bridging region. Shown are the grids corresponding to the α and β regions of Ser. The means of the distribution in the α and β clusters for the three conformational states of the side chain (χ_1) and the associated standard deviations (σ) are used to define the space along the ψ direction that has to be scanned for determining the residues in the bridging region. (0–360° ranges of angles shown in the figure actually correspond to φ, −180 to 180°; ψ, −120 to 240°; χ_1, −240 to 120°).

Figure 2
Plot of D_V against P_β (details given in Table 1). The fitted line has the equation $D_V = -1.28P_\beta + 3.46$. (If S and D_A are used instead of D_V, the equations are $S = -0.69P_\beta + 7.09$ and $D_A = 1.65P_\beta + 4.26$.)

residues (for example, Asp and Asn); the greater spread of points for these residues may, to a large extent, be a consequence of their occupying to a greater degree the bridging region connecting the α-helix and β-sheet regions. A strong β-sheet former like Ile or Val lacks points in this region. Consequently, we delineated the α and β regions and the intervening region in the φ, ψ, χ_1 space and counted the number of points each enclosed (Table 2). Fig. 3, where the spread of points along the ψ direction spanning the bridging region is depicted, shows that in general the residues with higher P_β have shallower distribution.

3.3. Energetics of interconversion between α and β conformations

What are the implications for the difference in the distributions of points in the bridging region for different residues? It is known that the low-energy regions of molecular potential-energy surfaces can be recognized and mapped from the observed distributions (Bürgi & Dunitz, 1983, 1988). Since the bridging region straddles the α and β regions, it can be expected to lie on the reaction path for the transformation of the α conformation to the β conformation and *vice versa*.

Table 2
Distribution of residues in the helical (α), β-sheet (β) and the bridging regions and the energy (ΔE) of the bridging region relative to the β region (at 300 K).

Residue	Number			ΔE (kcal mol^{-1})
	α	β	Bridging	
Ser	2178	2180	115	1.77
Cys	391	512	63	1.26
Met	801	576	53	1.43
Glu	2742	1279	106	1.49
Gln	1621	901	88	1.40
Lys	2418	1624	107	1.63
Arg	1825	1269	83	1.64
Leu	3399	2703	76	2.14
Asp	2400	1747	198	1.31
Asn	1529	1278	201	1.11
His	761	709	87	1.26
Phe	1400	1596	64	1.93
Tyr	1310	1460	46	2.08
Trp	569	529	55	1.36
Val	1942	3359	41	2.64
Ile	1760	2334	15	3.03
Thr	1934	2440	88	1.99
Pro	1432	1460	157	1.34
Ala	4071	1936	231	1.28
Gly	1213	1028	152	1.15

Because of the large number of data points used in this analysis, one may expect them to be distributed between the β and bridging or between the α and bridging regions following the Boltzmann distribution. Accordingly, the number of points in the bridging region relative to that in the β region provides an estimate of the energy barrier for the interconversion between the β and α conformations. Data provided in Table 2 show that this energy barrier (ΔE) for the conversion of the β conformation to the α conformation could be as high as 3 kcal mol^{-1}[1] (a large enough value to preclude a thermal equilibrium between the α and β conformations) for Ile and Val, the residues with the highest β-sheet propensity, whereas the residues with low propensity have a low barrier (Fig. 4). Furthermore, ΔE and P_β values are strongly correlated (Fig. 5), with the three residues Gly, Ala and Pro also falling into the general pattern. This suggests that the propensity of a residue to be in the β structure is directly related to the energy barrier that separates its α conformation from the β conformation.

3.4. Formation of β-sheets

From the above, it can be inferred that kinetic control (energy barrier) is operative during the formation of β-sheets. If it is assumed that in the growing polypeptide chain in the ribosome the residues are added with the backbone in the extended conformation [which has been shown to be the dominant conformation for a single residue peptide, Ac-Ala-NHMe (Ac, acetyl; Me, methyl) using molecular-dynamics simulation; Hermans, 1993], the residues with high barriers will continue to remain in the extended conformation unless they are adjacent to residues which have a higher propensity
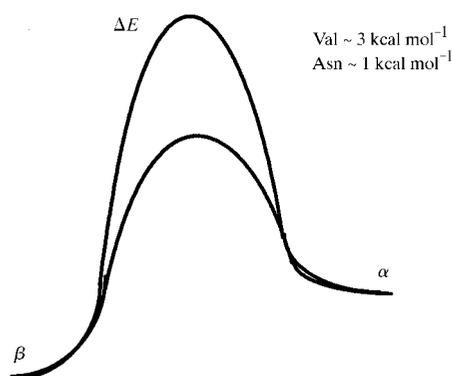

Figure 3
Variation of the population density [given by the fraction of points in each 10° step to the total number (with a negative value of φ), multiplied by 10⁴] in the bridging region, along the ψ axis. The scan range is explained in Fig. 1. Residues are grouped in the descending order of P_β values.

---

[1] 1 kcal mol^{-1} = 4.184 kJ mol^{-1}.

to take up a folded conformation (helix or turn), so that the latter group of residues, in a co-operative way, can make the former cross the energy barrier and occupy the α-conformation. It is possible that a stretch of residues with high β-sheet propensity will exist in the extended conformation and because of hydrophobic or electrostatic compatibility and other specific interactions (Garratt *et al.*, 1991; Minor & Kim, 1994*b*; Smith & Regan, 1995; Wouters & Curmi, 1995) such stretches will come together forming a β-sheet.

3.5. Implications for protein folding

There are examples to show the existence of free-energy barriers in polypeptide-chain conformational space, so that a protein might fold into a structure that is not the most stable thermodynamically (Baker & Agard, 1994). From our analysis, it can be suggested that residues like Val or Ile which have a high energy barrier for conformational transitions can maintain the conformation they are in and, by being located at

some crucial regions in the chain, can guide the folding process along the kinetic pathway. Moreover, the kinetically driven formation of the β-sheet may not necessarily yield the thermodynamically most stable structure. The amyloid fibril formation from soluble proteins that underlies a range of fatal diseases may provide an example. The core structure of all amyloid fibrils consists of β-sheets with the strands perpendicular to the long axis of the fibre (Blake & Serpell, 1996). It has been proposed (Booth *et al.*, 1997) that a partly folded molten globule-like form of the protein that retains some of the β-sheet elements of the native structure self-associates through the β-domain to initiate fibril formation. It is noteworthy that in this mechanism the molten globule (which is thermodynamically less stable than the native state) contains some β-strands that provide the template for the development of the stable intermolecular β-sheet leading to amyloidosis. Because of the 'trapping' of the β-strands giving rise to cross β-sheet, the individual protein molecules cannot proceed to the native state.

Figure 4
A schematic representation of the potential-energy diagram, showing the energy barriers between the α and β structures for two residues with distinct β-sheet propensities.
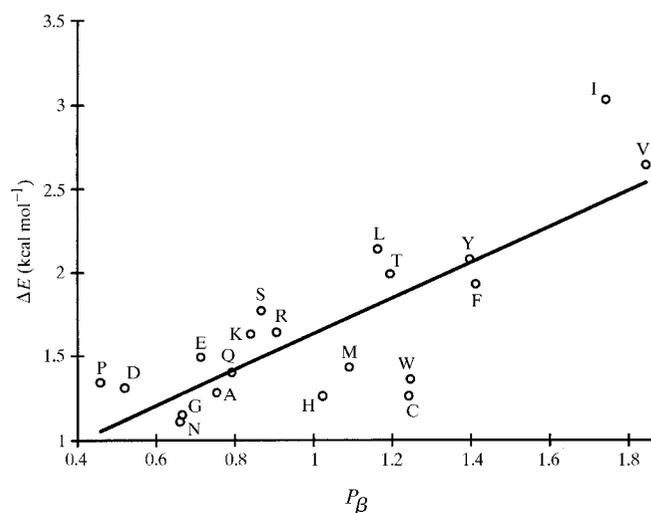
4. Summary

We have estimated the volume/area and entropy of the distribution of φ, ψ, χ_1 torsion angles of different residues and found them to correlate in an inverse fashion with the β-sheet propensities (Gly and Pro are exceptions). The larger values of the former set of parameters (indicating a greater dispersion in the conformational space) for residues with low propensities are also reflected in those residues having a greater relative population in the bridging region (linking the α and β regions) and *vice versa*. Assuming the bridging region to be on the reaction pathway for the interconversion between the α and β conformations, the relative proportion of the occurrence of a residue in this region compared with the α or β region provides an estimate of the energy barrier for the interconversion. By converting the relative population of residues in the bridging region relative to the β region into the equivalent energy (ΔE) term, it is found that these energy barriers bear direct correlation with β-sheet propensities. The two correlations mentioned above can be reconciled on the basis of the chemical structures of the side chains. The groups of residues with high P_β are branched (at C_β) aliphatic or aromatic (branched at C_γ) residues. Branching in the side chain close to the main chain means that there will be greater steric clash in these cases (compared with linear residues), resulting in a reduction in the available conformational space and a concomitant lowering of entropy (first correlation). On the same steric grounds, there will be a higher energy barrier linking one backbone conformation (corresponding to a secondary structure) to another (second correlation). If the polypeptide chain is synthesized with residues in the extended conformation, then a portion of the chain containing residues having high barriers is likely to have an extended conformation and will ultimately constitute a strand of the β-sheet in the folded structure. This also implies that when polypeptide structures are being modelled, residues with high β-sheet propensities can be put in the extended conformation before

Figure 5
ΔE (kcal mol^{-1}) plotted against P_β (the correlation coefficient between the two variables is 0.80). The equation from the regression analysis is $\Delta E = 1.07 P_\beta + 0.56$.

subjecting them to molecular-dynamics simulations. Finally, we have derived a new set of parameters, which in spite of having high correlation with β-sheet propensity values, are conceptually distinct and may improve our understanding of protein structure.

References

Baker, D. & Agard, D. A. (1994). *Biochemistry*, **33**, 7505–7509.

Bernstein, F., Koetzle, T. K., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blake, C. C. F. & Serpell, L. C. (1996). *Structure*, **4**, 989–998.

Booth, D. R., Sunde, M., Bellotti, V., Robinson, C. V., Hutchinson, W. L., Fraser, P. E., Hawkins, P. N., Dobson, C. M., Radford, S. E., Blake, C. C. F. & Pepys, M. B. (1997). *Nature (London)*, **385**, 787–793.

Bürgi, H.-B. & Dunitz, J. D. (1983). *Acc. Chem. Res.* **16**, 153–161.

Bürgi, H.-B. & Dunitz, J. D. (1988). *Acta Cryst.* B**44**, 445–448.

Chakrabarti, P. & Pal, D. (1998). *Protein Eng.* **11**, 631–647.

Chakrabartty, A. & Baldwin, R. L. (1995). *Adv. Protein Chem.* **46**, 141–176.

Chou, P. Y. & Fasman, G. D. (1978). *Adv. Enzymol.* **47**, 45–148.

Garratt, R. C., Thornton, J. M. & Taylor, W. R. (1991). *FEBS Lett.* **280**, 141–146.

Hermans, J. (1993). *Curr. Opin. Struct. Biol.* **3**, 270–276.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kim, C. A. & Berg, J. M. (1993). *Nature (London)*, **362**, 267–270.

Minor, D. L. Jr & Kim, P. S. (1994*a*). *Nature (London)*, **367**, 660–663.

Minor, D. L. Jr & Kim, P. S. (1994*b*). *Nature (London)*, **371**, 264–267.

Smith, C. K. & Regan, L. (1995). *Science*, **270**, 980–982.

Smith, C. K., Withka, J. M. & Regan, L. (1994). *Biochemistry*, **33**, 5510–5517.

Pal, D. & Chakrabarti, P. (1999). *Proteins*, **36**, 332–339.

Parthasarathy, R., Chaturvedi, S. & Go, K. (1995). *Prog. Biophys. Mol. Biol.* **64**, 1–54.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Adv. Protein Chem.* **23**, 283–437.

Wouters, M. A. & Curmi, P. M. G. (1995). *Proteins*, **22**, 119–131.