Protein Engineering vol. 16 no. 5 pp. 331–339, 2003
DOI: 10.1093/protein/gzg046

Stranded in isolation: structural role of isolated extended strands in proteins

Narayanan Eswar[1], C.Ramakrishnan and N.Srinivasan[2]

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

[1]Present address: Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94143-2240, USA

[2]To whom correspondence should be addressed.
E-mail: ns@mbu.iisc.ernet.in

Reasons for the formation of extended-strands (E-strands) in proteins are often associated with the formation of β-sheets. However E-strands, not part of β-sheets, commonly occur in proteins. This raises questions about the structural role and stability of such isolated E-strands. Using a dataset of 250 largely non-homologous and high-resolution (<2 Å) crystal structures of proteins, we have identified 518 isolated E-strands from 187 proteins. The two most distinguishing features of isolated E-strands from β-strands in β-sheets are the high preponderance of prolyl residues occuring in isolated E-strands and their high exposure to the surroundings. Removal of regions with polyproline conformation from the dataset did not significantly reduce the propensity of prolyl residues to occur in isolated E-strands. Isolated E-strands are often characterized by their main-chain amide and carbonyl groups involved in hydrogen bonding with polar side chains or water. They are often flanked by irregular loop structures and are less well conserved, than β-sheet forming β-strands, among homologous protein structures. It is suggested that isolated β-strands have many characteristics of loop segments but with repetitive (ϕ,ψ) values falling within the β-region of the Ramachandran map.

Keywords: β-sheet/β-strand/extended strand/hydrogen bonding/protein structures

Introduction

The requirement of polar groups in proteins to be satisfied by hydrogen bonding can be viewed as a director of protein folding (Rose and Wolfenden, 1993). As most of the amino acid residues in the interior of protein structures are known to lack polar side chains (Chothia, 1976; Miller *et al*., 1987), it is conceivable that most of the polar groups at the interior are situated at the backbone of the polypeptide chain. These polar groups of the polypeptide backbone (NH and C=O groups) are known often to be satisfied by virtue of the formation of helical and β-sheet structures in proteins (Baker and Hubbard, 1984; Stickle *et al*., 1992). Formation of characteristic hydrogen bonding patterns involving the amide and carbonyl groups of the polypeptide main chain is an essential feature of the formation of α-helices, β-sheets and β-turns in proteins (Pauling and Corey, 1951; Pauling *et al*., 1951; Venkatachalam, 1968). Indeed, an important driving factor for the formation of α-helix in proteins is suggested to be the formation of intra-segment hydrogen bonding (Presta and Rose, 1988). Deviation from the characteristic hydrogen bonding patterns in α-helices and β-sheets is known to result in distortions in these structures (Richardson *et al*., 1978; Barlow and Thornton, 1988). These regions of distortion are often found to be solvated. For example, the kink produced by a proline residue in the middle of an α-helix and the existence of a β-bulge in β-sheets are well known.

The amino acid residue preferences and van der Waals stabilizing interactions are also characteristics of α-helices and β-strands in proteins (Street and Mayo, 1999). The conformational entropy for the rotation of side chains is suggested to be a key feature in the preference or otherwise of an amino acid type to occur in α-helix or β-sheet form (Presta and Rose, 1988; Creamer and Rose, 1992; 1994; Stapley and Doig, 1997). For example, interactions between the side chains in positions i and $i + 3$ (and $i + 4$) in α-helices (Creamer and Rose, 1995) and interactions between side chains across β-strands involved in the formation of a β-sheet are known to contribute to the stabilization of these structures (Lifson and Sander, 1980; Otzen and Fersht, 1995; Smith and Regan, 1995; Wouters and Curmi, 1995).

The β-sheet is generally considered as a 'secondary structure' although it is known to be distinct from the other kinds of regular secondary structures. The distinction stems from the fact that it requires spatially neighbouring regions of the protein, in extended conformation, to become aligned to form the characteristic inter-strand hydrogen bonds. However, it may be inappropriate to refer to the β-strand as a secondary structure as, unlike other kinds of secondary structure, there are no intra-segment hydrogen bonds. Often, it is tempting to associate the role of formation of a main-chain region in the extended conformation (extended strands or E-strands) with that of β-sheets.

In this paper, we draw attention to the regions of proteins in extended conformation that are not involved in the formation of a β-sheet. As the description of an extended strand does not involve the hydrogen bonding of amide and carbonyl groups of the backbone, unless involved in the formation of a β-sheet, the role of such extended structures in proteins is puzzling. Also, as these E-strands are not participating in the formation of β-sheet there is no possibility of inter-strand interaction between non-polar residues like the one first observed by Lifson and Sander (1980). We have surveyed a large number of known protein structures and found that such isolated extended strands commonly occur in proteins and share characteristics of loops and β-sheets in proteins. These E-strands are distinct from the polyproline type II extended conformation whose occurrence in globular protein structures has been extensively studied (Soman and Ramakrishnan, 1983; Adzhubei *et al*., 1987a–c; Ananthanarayanan *et al*., 1987; Adzhubei and Sternberg, 1993). The polyproline type II conformation is

somewhat similar to that of a single strand of collagen with characteristic (ϕ,ψ) values of around (–65°,140°) and is distinct from that of a β-strand which has approximate (ϕ,ψ) values of (–115°,130°). Various features of polyproline type II-related structures (also referred to as 'mobile' or M conformations by Esipova and co-workers) as seen in the known crystal structures of proteins have been analysed extensively by Esipova and co-workers (Adzhubei *et al*., 1987a–c; Vlasov *et al*., 2001). In particular, they have made several detailed analyses of length, residue and tetrapeptide sequence distributions and have made comparisons of the extents of occurrence of this structure with that of α-helix and β-sheet (Adzhubei *et al*., 1987a–c; Vlasov *et al*., 2001). As can be seen during the course of the present analysis, the isolated E-strands described here are distinguished from the polyproline type II-related structures as the (ϕ,ψ) values of isolated E-strands are closer to those of β-sheets than polyproline type II structures.

Materials and Methods

Dataset used

A dataset of 250 highly resolved (resolution <2.0 Å) and non-homologous protein structures derived from the Protein Data Bank (PDB) (Bernstein *et al*., 1977; Berman *et al*., 2000) was used for the analysis. In the case of proteins with identical or very similar polypeptide chains, only one of them was considered. The chain used in such cases is shown as a fifth character in the complete list of PDB codes of the proteins used as follows: 1aan, 1aazA, 1abe, 1abk, 1acf, 1acx, 1afgA, 1ahc, 1ak3A, 1alc, 1ald, 1alkA, 1amp, 1ankA, 1aozA, 1apmE, 1arb, 1arp, 1ars, 1ast, 1bbhA, 1bbpA, 1bgc, 1bgh, 1bmdA,1brsD, 1bsaA, 1byb, 1cbn, 1ccr, 1cewI, 1cgt, 1chmA, 1cmbA, 1cot, 1cpcA, 1cpcB, 1cpn, 1cseE, 1cse I, 1csh, 1ctf, 1cus, 1ddt, 1dfnA, 1dmb, 1dri, 1dsbA, 1eca, 1esl, 1ezm, 1fas, 1fdn, 1fgvH, 1fiaA, 1fkf, 1flp, 1flv, 1fna, 1frrA, 1fus, 1fxl, 1fxd, 1gd1O, 1gia, 1gky, 1glqA, 1glt, 1gog, 1gox, 1gp1A, 1gpr, 1hel, 1hip, 1hleA, 1hleB, 1hoe, 1hpi, 1hsbA, 1hsbB, 1hslA, 1huw, 1hvkA, 1hyp, 1iag, 1ifb, 1isaA, 1isuA, 1lcf, 1lec, 1lib, 1lis, 1lldA, 1ltsA, 1ltsC, 1ltsD, 1mba, 1mbd, 1mdc, 1mjc, 1molA, 1mpp, 1nar, 1nbaA, 1nlkR, 1npc, 1nscA, 1olbA, 1onc, 1opaA, 1ovaA, 1pda, 1pgb, 1phc, 1php, 1pii, 1pk4, 1pmy, 1poc, 1poh, 1ppa, 1ppbH, 1ppbL, 1ppfE, 1ppt, 1prn, 1ptf, 1ptsA, 1r69, 1rbp, 1rdg, 1rec, 1ris, 1rnh, 1ropA, 1sacA, 1sbp, 1sgt, 1shaA, 1shfA, 1shg, 1sim, 1sltA, 1smrA, 1srdA, 1stn, 1tca, 1ten, 1tfg, 1tgn, 1tgsI, 1tgxA, 1thbA, 1tml, 1ton, 1trb, 1trkA, 1ubq, 1utg, 1whtA, 1whtB, 1×ib, 1ypiA, 256bA, 2acq, 2act, 2alp, 2apr, 2bbkH, 2bbkL, 2bmhA, 2cab, 2ccyA, 2cdv, 2chsA, 2ci2I, 2cmd, 2cpl, 2ctvA, 2cy3, 2cyp, 2end, 2fcr, 2gbp, 2gstA, 2had, 2hbg, 2hmqA, 2lh7, 2lhb, 2ltnA, 2ltnB, 2lzm, 2mcm, 2mltA, 2mnr, 2msbA, 2ohxA, 2ovo, 2pabA, 2pia, 2plt, 2por, 2prk, 2rhe, 2rspA, 2sarA, 2scpA, 2sga, 2sn3, 2spcA, 2trxA, 2tscA, 2wrpR, 2ztaA, 351c, 3app, 3b5c, 3bcl, 3blm, 3c2c, 3chy, 3cla, 3cox, 3dfr, 3dni, 3drcA, 3ebx, 3est, 3grs, 3il8, 3mdsA, 3psg, 3rp2A, 3rubL, 3rubS, 3sdhA, 3tgl, 4azuA, 4bp2, 4cpv, 4enl, 4fxn, 4gcr, 4i1b, 4icb, 4insC, 4insD, 4mt2, 4tnc, 5chaA, 5cpa, 5fd1, 5p21, 5pti, 5rubA, 6ldh, 7acn, 7rsa, 8dfr, 8fabA, 8fabB, 9wgaA.

Identification of secondary structural elements

A stretch of at least four consecutive residues was identified as an E-strand if all the (ϕ,ψ) values in this region lie within the region defined by: $-180° < \phi < -30°$, $60° < \psi < 180°$ or $-180° < \psi < -150°$ (Gunasekaran *et al*., 1998). A strand in the extended conformation is qualified to be a polyproline II type of structure if the ϕ-values at each of the residues of the segment are greater than –90°. The polyproline II type conformation has a close resemblance to that of a single strand of collagen and is known to occur in globular protein structures (Soman and Ramakrishnan, 1983; Adzhubei *et al*., 1987a–c; Ananthanarayanan *et al*., 1987; Adzhubei and Sternberg, 1993; Vlasov *et al*., 2001). The E-strands thus picked up were further separated into two classes, namely, *isolated* (those not in register with another E-strand by means of hydrogen bonding characteristic of β-sheets) and *aligned* E-strands (those in register with another E-strand forming a β-sheet), using an algorithm of secondary structure assignment based on the relative positions of the Cα atoms (Ramakrishnan and Soman, 1982; Soman and Ramakrishnan, 1986). The E-strands not part of the β-sheet are referred as 'isolated' solely to reflect the fact that there is no hydrogen bonding interaction between the main-chain polar atoms of the strand with another strand of extended conformation. The aligned E-strands are also referred as β-strands as they participate in the formation of the β-sheet. From the β-strands *edge* β-*strands* were then defined as those segments of extended conformation which are in register with only one other β-strand, as opposed to *inner* β-*strands* which possess segments in register on either side. Identification of hydrogen bonding is based on the method used by Overington *et al*. (Overington *et al*., 1990) involving distances between putative donors and acceptors and hydrogen bonding interaction energy.

Helices were identified in a manner similar to the E-strands with a criterion that at least four contiguous residues were in the α_R region (defined by $-140° < \phi < -30°$, $-90° < \psi < 45°$) (Gunasekaran *et al*., 1998). 3_{10} helices were differentiated from α-helices by using the procedure of Ramakrishnan and Soman (Ramakrishnan and Soman, 1982). Further, a stretch of at least four consecutive residues which does not fall into any of the categories described above was classified as a *loop* and the remaining non-secondary structural non-loop residues were termed *random coil residues*. The results of identification of secondary structures using the Cα position-based and (ϕ,ψ)-based methods were very similar to those obtained using other methods such as DSSP (Kabsch and Sander, 1983).

In the discussions, the symbols β_E, β_B, E_I and PPII refer to the edge β-strand, inner β-strand, isolated E-strand and polyproline II regions, respectively.

Generation of all the neighbouring molecules in the crystal lattice

We also investigated the interactions, if any, between the isolated E-strands and the neighbouring molecules in the crystal lattice (our dataset contains no NMR structures). For every protein structure with at least one isolated E-strand we generated the fractional coordinates using the cell dimensions given in the coordinate file. Using the space group information, the equivalent points are automatically recognized from the library of equivalent points stored against every space group. The fractional coordinates of all the atoms corresponding to every equivalent point are generated. Further, translations by –1, 0 and +1 are made along each of the fractional *x*-, *y*- and *z*-axes to generate the entire system of neighbouring molecules (including those in the adjacent unit cells) around a given molecule. Finally, all the generated coordinate sets are converted to the original orthogonal ångstroms coordinate system using the cell dimensions. For example, if the space

Table I. Extent of occurrence of the various secondary structural segments in the dataset used

	No. of segments	No. of residues involved	Average length of segments	Peak of the length distribution	% of segments at the peak
α-Helix	1483	17277	11.7	10	8.4
3_{10}-Helix	119	630	5.3	4	41.2
Isolated E-strand	518	2564	5.0	4	50.8
Polyproline helix	56	241	4.3	4	76.8
Edge β-strand	1103	6892	6.3	4	23.9
Inner β-strand	791	5822	7.4	6	19.0
Loop segment	1960	15422	7.9	4	24.0

group of a given entry is such that it has four equivalent points [including the original (x, y, z)] and each of the equivalent points can result in a set of 3 × 3 × 3 (= 27) neighbouring molecules to result in the generation of 4 × 27 (= 108) coordinate sets. We cross-referenced our results with those given in PQS server (Henrick and Thornton, 1998) and the results were found to be absolutely consistent. Interaction between the main-chain polar atoms of putative isolated E-strands in the original coordinate set and the neighbouring copies in the crystal lattice was analysed. Further, if a crystal structure has more than one molecule in the asymmetric unit, interaction between the putative isolated E-strand and the other molecule(s) present in the asymmetric unit was also analysed.

Results and discussion

The dataset of 250 proteins was probed to identify the various structural elements, namely α- and 3_{10}-helices, isolated E-strands, edge and inner β-strands and loops, which resulted in a total of 6030 segments consisting 48 848 amino acid residues. The results of the search are summarized in Table I. Over half the segments (56%) identified and 61% (N = 29 991) of the residues fall under the well recognized secondary structural elements, α-helices (N = 1483) and β-sheets [edge (β_E) + inner β-strands (β_B), N = 1894] and close to 33% (N = 1960) of the segments consisting of 15 422 residues classify as loops. A major proportion of the rest is comprised of the 518 segments of isolated E-strands (E_I), which is the subject of this paper. Fifty-six segments were identified as similar to polyproline type II helices (PPII) studied by Esipova and co-workers (Adzhubei *et al*., 1987a–c; Vlasov *et al*., 2001). Results obtained from analysing the distribution of lengths of these segments indicate that α-helical regions and β-sheet forming β-strands (β_E + β_B strands) tend to form longer segments than either the E_I or PPII. The α-helices have an average length of 11.7 residues (per segment) (Barlow and Thornton, 1988; Kumar and Bansal, 1998) while the β_E and β_B strands have average lengths of 6.3 and 7.4, respectively (Sternberg and Thornton, 1977). In contrast, other regular structures such as the 3_{10}-helices (Ramakrishnan and Soman, 1982), E_I and PPII strands (Soman and Ramakrishnan, 1983, 1986; Adzhubei *et al*., 1987a–c; Adzhubei and Sternberg, 1993; Vlasov *et al*., 2001) are observed to be shorter with average lengths of 4–5 residues per segment. It is also observed that segments of irregular regions in proteins, termed loops, tend to be long with an average length of close to eight residues per segment (Martin *et al*., 1995). Table I also gives the peak of the length distribution for each type of structure and the percentage of examples represented by the peak. It can be seen that the peak of the length distribution is at four residues per segment for the majority of the structures, with the exception of only α-helices and inner β-strands. In the case of α-helices, although the peak occurs at 10 residues per segment, the percentage of examples represented by the peak is very small (~8%). These facts indicate that short segments of regular structures are ubiquitously found in proteins.

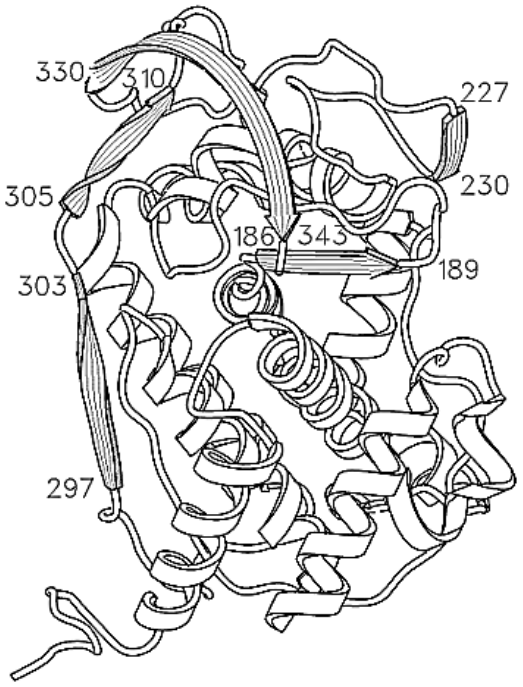

Fig. 1. A representative example from the dataset for one of the longest isolated E-strand segments of length 14 residues from the fungal peroxidase (pdb = 1arp, 330–343). All the isolated E-strands in the protein are shown as striped arrows and start and end of these strands are marked by their residue number. This figure was prepared using SETOR (Evans, 1993).

The 518 segments of isolated E-strands identified from the dataset contain a total of 2564 amino acid residues. The length of these segments varies from four to 14 residues per segment. It is found that close to 51% of these segments are just four residues long, supporting the earlier observation of Soman and Ramakrishnan (1986) that the E_I segments in protein structures are often short. One of the longest examples of E_I strands exists in the structure of the fungal peroxidase (PDB code = 1arp, 330–343) (Kunishima *et al*., 1994) shown in Figure 1, which runs to a length of 14 residues.

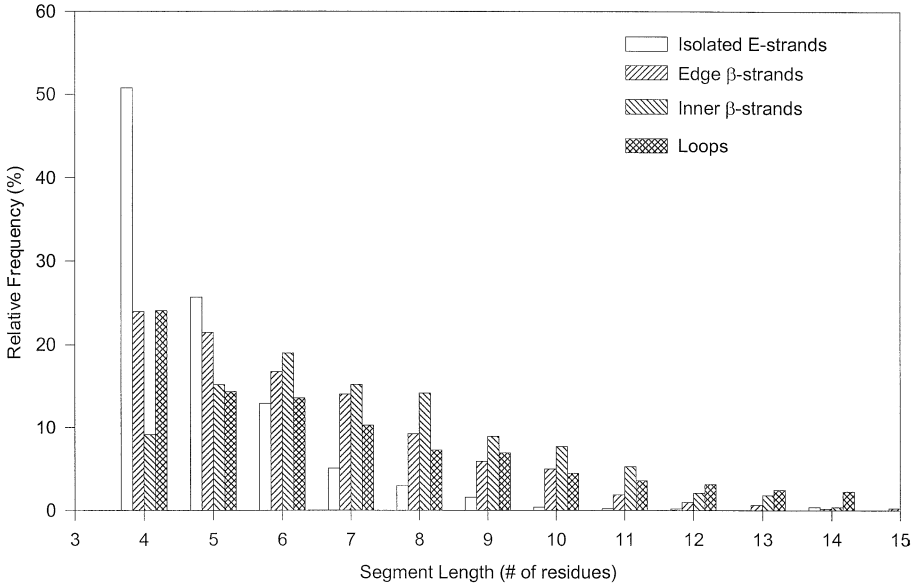

Fig. 2. Distribution of lengths of segments representing E_I strands, β_E and β_B strands and loops. The progressive decrease in the frequencies of occurrence of longer E_I, β_E and loop segments is contrasted by the peak of β_B segments at the bin representing six residues.

Table II. Propensity of amino acid residues to occur in various extended segments and loops

Residue	Isolated E-strand	Polyproline type helix	Edge β-strand	Inner β-strand	Loop segment
Ala	0.74	1.77	0.73	0.84	0.74
Arg	1.02	0.41	1.00	0.81	0.94
Asn	0.76	0.52	0.70	0.57	1.36
Asp	0.78	1.09	0.61	0.66	1.32
Cys	1.27	1.63	1.12	1.17	1.13
Gln	0.94	0.58	0.81	0.84	0.85
Glu	0.92	0.67	0.81	0.63	0.86
Gly	0.37	0.20	0.45	0.54	1.73
His	0.92	0.40	1.02	0.96	1.04
Ile	1.18	0.39	1.42	1.67	0.67
Leu	1.01	0.93	1.12	1.20	0.63
Lys	0.98	0.62	0.85	0.83	0.93
Met	1.01	0.89	1.07	1.27	0.57
Phe	1.18	1.04	1.29	1.29	0.81
Pro	2.27	6.24	1.00	0.76	1.36
Ser	0.96	0.68	0.94	0.93	1.24
Thr	1.23	0.60	1.36	1.19	1.07
Trp	0.90	0.56	1.47	1.22	0.80
Tyr	1.01	0.67	1.41	1.46	0.88
Val	1.25	0.47	1.68	1.72	0.66

A comparison of the lengths of E_I strands with the other extended segments, the β_E and β_B strands and the loops, is shown in Figure 2. It can be seen that the trend is towards shorter segments for the isolated E-strands and edge β-strands and also the loops, shown by a gradual decline in the proportion of segments populating bins corresponding to longer segments. On the other hand, the peak for the inner β-strands lies at six residues, which agrees with the results of Sternberg and Thornton (Sternberg and Thornton, 1977).

Propensities of amino acid residues to occur in various extended segments and loops

The propensities of the 20 amino acid residues to occur in the various kinds of extended segments and loops were calculated in order to assess the preferences exhibited by individual residues for specific types of structures. The propensities were calculated using the standard Chou–Fasman approach (Chou and Fasman, 1974). The results are shown in Table II. It can be seen that, in general, the hydrophobic residues are preferred over the polar residues in all the three extended segments, E_I, β_E or β_B strands. It is widely known that β-branched residues such as Val, Ile and Thr show a high propensity to occur in β-sheets (Chou and Fasman, 1974; Lifson and Sander, 1979; Munoz and Serrano, 1994; Swindells *et al*., 1995). Interestingly, the preferences of residues to occur in E_I strands also reflect similar characteristics. This strongly reinforces the earlier reports (Swindells *et al*., 1995) that strand formation is determined by the intrinsic preferences of amino acid residues (Dinner *et al*., 1999). In contrast, as is well known, the loops prefer polar residues.

One interesting feature seen from the amino acid propensities shown in Table II is that prolyl residues show a very high preference to occur in isolated E-strands, that is shared only by the PPII strands, in which case the reason is obvious.

Preference for prolines in E_I strands

The enhanced preference for proline to occur in E_I strands led us to investigate the existence of polyproline type II strands (PPII) (Soman and Ramakrishnan, 1983; Adzhubei *et al*., 1987a–c; Ananthanarayanan *et al*., 1987; Adzhubei and Sternberg, 1993; Stapley and Creamer, 1999; Vlasov *et al*., 2001) which resemble the E_I strands in that the participating residues of the former also possess extended conformation. The PPII regions were recognized as a contiguous stretch of (ϕ,ψ) values in the polyproline region (see Materials and methods) and did not depend upon the occurrence or otherwise of proline. The search yielded a total of only 56 examples of PPII strands. When the PPII strands were weeded out of the dataset, it was found that these represented only a very small fraction of the extended segments. The recalculated values of the propensities, shown in Table II, after the removal of such strands, still show a striking preference for proline to go into E_I strands over the aligned β-strands.

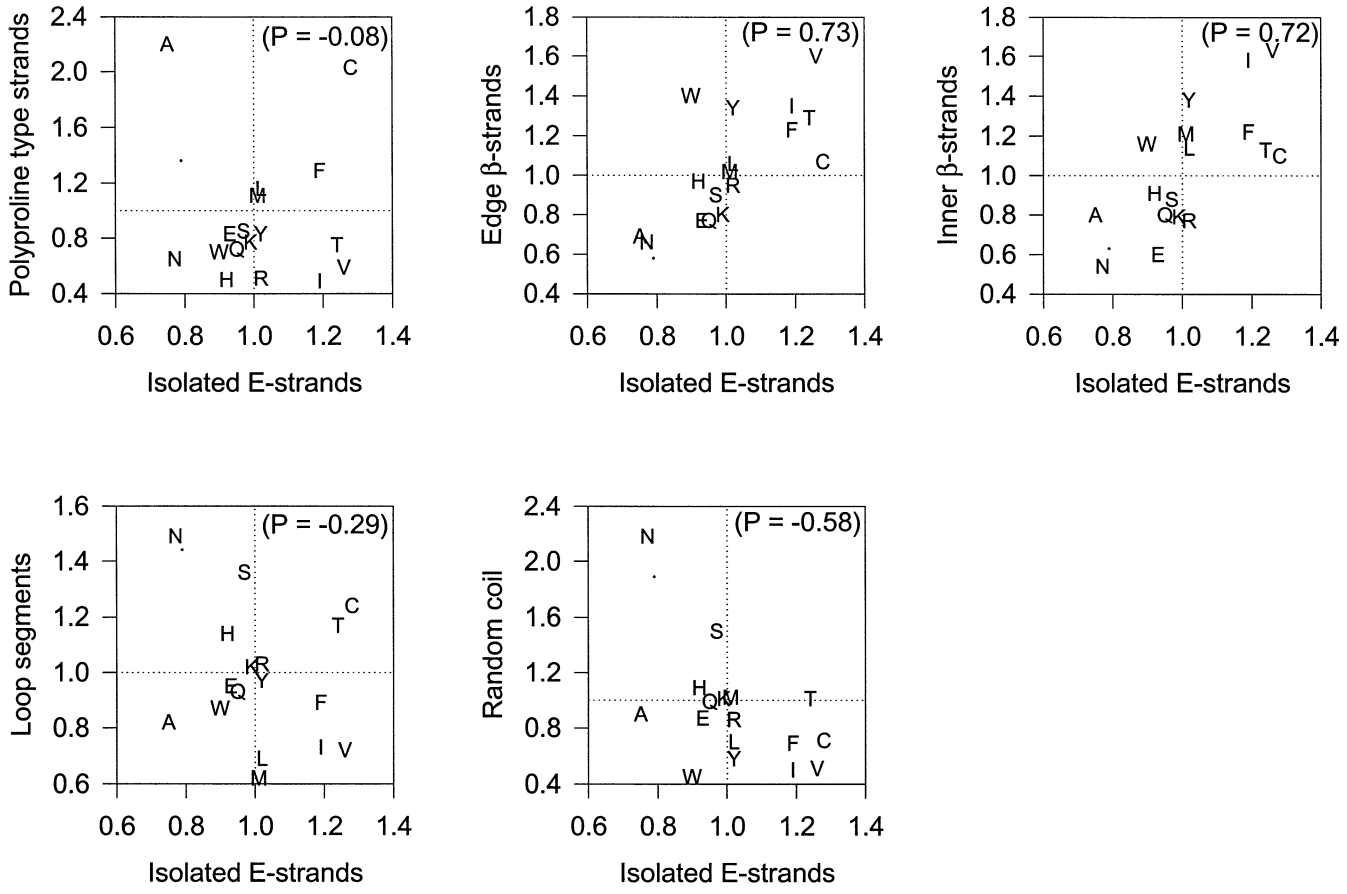

Fig. 3. Pairwise comparison of propensities of various amino acid residues to occur in extended segments and loops. The horizontal axis represents E_I strands and the vertical axis represents the segment indicated. The points are represented as the single-letter code of the amino acids. Propensities of some of the non-Gly, non-Pro amino acids are not plotted here or in Figure 4 as we are unable to find the appropriate data in the literature.

Table III. Correlation coeffiecients between pairs of propensities of amino acid residues to occur in the various extended segments and loops

	Isolated E-strand	Polyproline type helix	Edge β-strand	Inner β-strand	Loops
Isolated E-strand	1.00	0.80	0.43	0.28	–0.09
Polyproline type helix		1.00	-0.06	–0.17	0.23
Edge β-strand			1.00	0.92	–0.60
Inner β-strand				1.00	–0.67
Loops					1.00

Close to 42% ($N = 216$) of the 518 segments classified as E_I strands contain at least one proline residue in its sequence. Also, these proline residues are interspersed in the sequence with no specific preference for any particular position within the sequence. These observations lead to two interconnected features that can be conceived as the cause of the high preference for proline in E_I strands. First, the lack of an amide hydrogen in the backbone of proline makes it an unsuitable candidate for inclusion into any of the standard secondary structures in which backbone hydrogen bonding plays a crucial role, as in α-helices and β-sheets (Richardson and Richardson, 1988; Aurora and Rose, 1998; Gunasekaran *et al*., 1998). Second, proline possesses an intrinsic feature of influencing the backbone torsion angles of the residue preceding it to adopt an extended conformation (Gibrat *et al*., 1991; MacArthur and Thornton, 1991; Hurley *et al*., 1992). These unique characteristics of proline seem to be the reason for its preferred existence in E_I than either the β_E or β_B strands.

Comparison of propensities of occurrence between various kinds of segments

Since E_I strands are not part of β-sheets, secondary structure recognition algorithms usually classify these as loops. Thus, in order to assess their relationship with the PPII, β_E and β_B strands and loops, we calculated Pearson's correlation coefficient (P value) (Minor and Kim, 1994a) between the various pairs of amino acid propensities. The P value was calculated using the equation

$$P^2 = \{\Sigma(x_i - x_{av})(y_i - y_{av})/[\Sigma(x_i - x_{av})^2\Sigma(y_i - y_{av})^2]^{\frac{1}{2}}\}^2$$

where x_i and y_i pairs correspond to the amino acid propensities; i represents the index of summations and is the number of amino acid types considered and x_{av} and y_{av} represent mean x and y values, respectively. The P values are listed in Table III. In order to avoid the bias made by the two special residues, the highly flexible Gly and the rigid Pro, they were eliminated from the dataset and the coefficients were recalculated (Swindells *et al*., 1995). The plots describing these correlations are shown in Figure 3. As we have a reasonably large number of residues in our dataset, the reliability of propensity values is expected to be unaffected by the exclusion of prolyl and glycyl residues from the calculations.

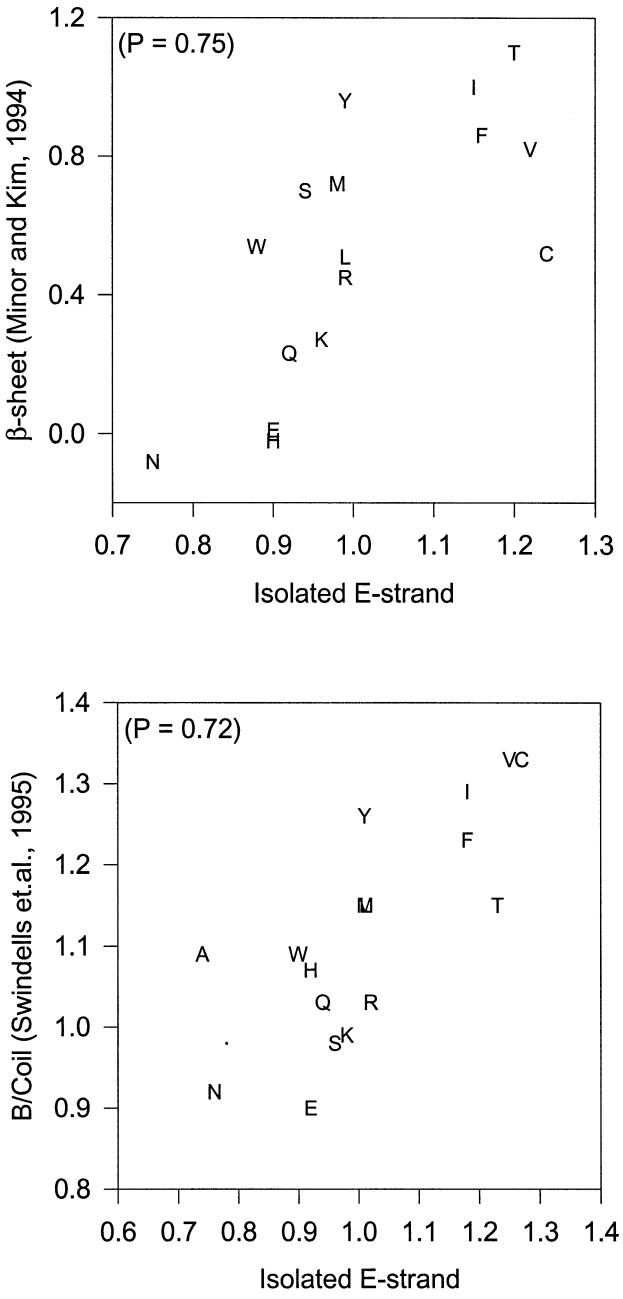

Fig. 4. Pairwise comparison of propensities of amino acid residues to occur in E_I strands with two of the reported scales from the literature. The top panel shows the comparison with the scale from Minor and Kim (Minor and Kim, 1994a) and the bottom panel shows that from Swindells *et al.* (Swindells *et al*., 1995).

From Table III, it can be seen that E_I strands seem to show a very good correlation with the PPII strands (P = 0.80). However, from the propensity values shown in Table II it can be seen that the trends of amino acid preferences are not very similar. The high correlation shown in Table III for this pair was found to be due to the very high values of propensity for Pro. On removal of this residue (and also Gly for uniformity with other pairs) from the calculation of the correlation coefficient, the value was found to fall drastically to P = –0.08 (shown in Figure 3). On the other hand, the correlation between E_I and either of the aligned β-strands (β_E or β_B) improves on the removal of Pro and Gly. From a low P value (P = 0.43 for β_E and P = 0.28 for β_B with E_I strands) when Pro and Gly are included, E_I strands show a good correlation with both the β_E (P = 0.73) and the β_B (P = 0.72) strands. The simultaneous good correlation between E_I and β_E and E_I and β_B is not surprising since it can be seen from Table III that there is an extremely high correlation (P = 0.92) between β_E and β_B strands. Moreover, this high correlation does not change on the removal of Gly and Pro residues (data not shown). This shows that E_I strands are similar to the β_E or β_B strands with the exception of the enhanced preference for Pro in E_I strands. In contrast to the earlier case (that between the E_I and PPII strands), where the Pro boosted the correlation, in the latter case the correlation between the E_I and either the β_E or β_B was hidden owing to the enhanced preference for Pro in E_I strands.

On the other hand, E_I strands show a negative correlation (P = –0.09) with the loop segments. There does not seem to be any drastic change in this value even after the removal of Gly and Pro (P = –0.29). The fact that the residue preferences of E_I strands show a strong correlation with the β_E and β_B strands and simultaneously show a negative correlation with loops (in the same way as the β_E and β_B strands; data not shown) induces us to propose that the E_I strands resemble the β-sheet forming β-strands in terms of the residue preferences (except for Pro) and structure.

Comparison of propensities with other scales

Since the discussion above leads us to believe that E_I strands are similar to the other aligned β-strands, at the level of residue preferences (except for Pro), we compared our propensities with other scales reported in the literature as done, for example, by Finkelstein (Finkelstein, 1995). Since the experimental scales were all derived by host–guest studies by measuring the $\Delta\Delta G$ for replacement of one residue with another, the results are reported on a scale relative to one of the amino acid residues, usually alanine or glycine. Also, most of these scales also give an abnormal value of $\Delta\Delta G$ for proline. For these reasons, we eliminated all the three residues from our calculations.

The propensities of various amino acids for E_I strands correlate best with the β-sheet propensities derived by Minor and Kim (Minor and Kim, 1994b) (P = 0.75). The comparison of propensities is shown in Figure 4. The same authors also demonstrated the context dependence of amino acid preferences by analysing edge and interior positions (Minor and Kim, 1994a, 1996) but our propensities show only a very weak correlation with this scale (P = 0.26). We also compared our data with two other scales (Kim and Berg, 1993; Smith *et al*., 1994), but both showed very low correlations of –0.53 and –0.20, respectively.

Two theoretically derived scales, which describe the intrinsic propensity of an amino acid to take up a particular structure (Munoz and Serrano, 1994; Swindells *et al*., 1995), were also used to compare the propensities we derived for E_I strands. Our results matched well with the propensities for 'B/Coil' of Swindells *et al.* with P = 0.72 (shown in Figure 4) whereas it showed a low correlation with that of Munoz and Serrano (P = 0.46).

Accessibilities of the isolated extended segments

The degree of solvent accessibility of an isolated extended strand was calculated as the ratio of its total accessible surface area (ASA) (Lee and Richards, 1971) as it occurs in the protein to the sum of the ASA of each of the constituent residues as it occurs in an extended conformation (Miller *et al*., 1987). It is

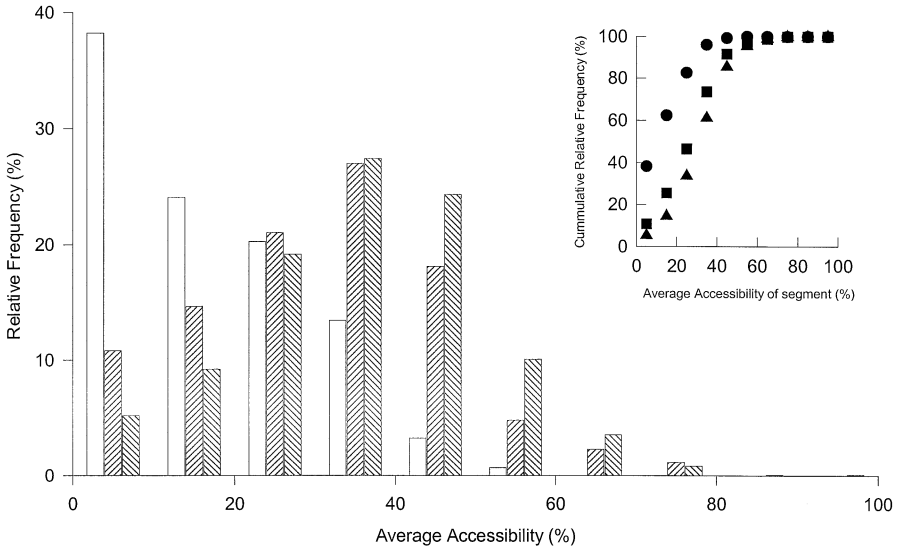

Fig. 5. Distribution of the average accessibility (%) the β_B (left bar), E_I (middle bar) and loop segments (right bar). The inset shows the cumulative frequencies of β_B (circles), E_I (squares) and loops (triangles) versus the average accessibility of the segments. The similarity of trends between the E_I and loop segments is evident.

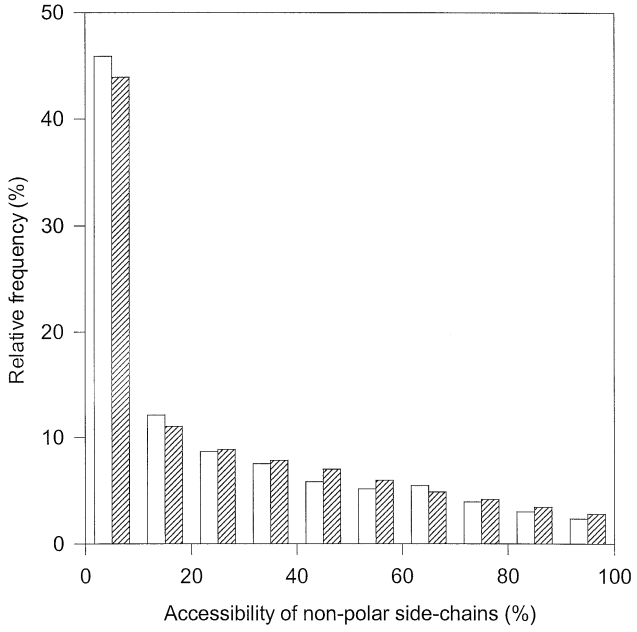

Fig. 6. Distribution of the accessibilities of the side chains of non-polar residues participating in E_I strands (left bar) and loops (right bar).

observed that just over 90% of the 518 segments of β_I strands have accessibilities in the range 0–50% with about 27% in the range 30–40%. Only very few of the segments (13.5%) have low accessibilities (<10%), indicating that most of the E_I strands tend to be exposed to the solvent.

In a bid to compare the accessibility profiles of E_I strands with both the traditional (aligned) β-strands and the loops, Figure 5 shows a comparison of the profiles of each of these segments. It can be seen immediately that most of the aligned β-strand segments have very low accessibility values. Close to 55% of the segments have accessibilities in the range 0–10% with the population at successive intervals progressively falling. This can also be seen from the inset in Figure 5, which shows the cumulative frequency against the accessibility intervals, where the curve corresponding to the aligned β-strands reaches a plateau fairly rapidly.

On the other hand, the behaviour of loops is very similar to that of E_I strands. The peaks of the frequency distribution (25.3% of the loops) for these two kinds of segments almost coincide, in the interval between 30 and 40%. A small point of difference is that the distribution of accessibilities for the loops extends over the next interval between 40 and 50% and also with close to 24% of the loop segments. Nevertheless, the curves of the cumulative frequencies of these two segments (shown in the inset) almost coincide, indicating that the E_I strands are as much exposed as loops in protein structures.

From the amino acid preferences of E_I strands we see that most of the preferred residues are non-polar in nature. However, we also see that these segments of E_I strands are exposed to the solvent like the loops. To resolve this dichotomy, we analysed the side-chain accessibilities of the non-polar residues in both the E_I strands and loops. The results are shown in Figure 6. It can be seen that the behaviour of non-polar side chains is almost identical in both of these kinds of segments. Close to 45% of the non-polar side chains of E_I strands are buried from the solvent indicated by the first peak in Figure 6, indicating that the high accessibility is contributed by polar side chains and main-chain atoms.

Segments flanking isolated E-strands in protein structures

The structural environment of E_I strands was analysed by identifying the first occurrence of a secondary structural element before and after these segments. We searched for patterns of the form $S_aXXX–E_I$ strand–$XXXS_b$, where S corresponds to a residue in one of the secondary structures and X is a residue which could be part of a regular secondary structure or loop. Table IV shows the frequencies of occurrence of various secondary structural segments at positions S_a and S_b in the vicinity of the isolated E-strands.

One of the questions which could be asked about the E_I strands is whether these segments are extensions of the aligned β-strands which do not have a neighbouring segment to be in register. It can be seen from Table IV that such examples are

Table IV. Frequency of occurrence of structural elements at positions S_a and S_b on either side of the E_I strand obtained by search patterns of the form $S_aXXX–E_I–XXXS_b$

S_b	S_a							
	α-Helix	3_{10}-Helix	Isolated E-strand	Polyproline type helix	Edge β-strand	Inner β-strand	Loops	Total (preceding)
α-Helix	49	6	10	3	9	10	41	128
3_{10}-Helix	1	0	3	1	1	0	3	9
Isolated E-strand	12	1	8	2	8	1	16	48
Polyproline type helix	4	0	0	1	1	0	2	8
Edge β-strand	6	1	5	1	8	2	19	42
Inner β-strand	4	0	6	0	4	0	15	29
Loops	57	1	17	1	27	12	82	197
Total (succeeding)	133	9	49	9	58	25	178	

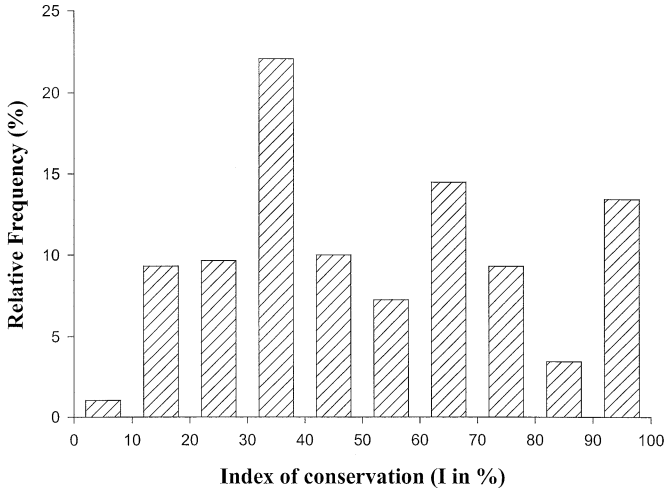

Fig. 7. Distribution of the index of conservation of isolated E-strands (E_I) in families of homologous proteins.

very few. There are only a total of 71 examples of E_I strands which may be the N-terminal extensions of aligned β-strands and 83 examples of C-terminal extensions. About 25% (N = 128) of E_I segments are flanked by α-helices at the N-terminus and about 26% (N = 133) at the C-terminus. In contrast, most of the examples of E_I strands are flanked by loop segments on one or both sides. Close to 38% (N = 197) of the examples have a loop segment on the N-terminal side while 34% (N = 178) have a similar structure on the C-terminal side. Also, there are 82 examples where they are flanked by loops on both sides.

Hence it appears that a stretch of extended conformations is the best type of structure to provide the maximum end-to-end distance for a given number of residues. These E_I strands may supplement the long loops that connect secondary structural elements which are spatially well separated.

Hydrogen bonds to the backbone groups of E_I strands

Since E_I strands are not part of β-sheets they lack the periodic hydrogen bonding ladder that characterizes a β-sheet. Hence, the backbone carbonyls and amides of isolated E-strands would have to be satisfied with hydrogen bonds from other protein atoms or the solvent. The 2564 residues participating in E_I strands were analysed for hydrogen bonds to or from their backbone polar groups. Of the 2564 residues, 263 were proline residues with only the carbonyl oxygen available as an acceptor for hydrogen bonds. Of these, only 94 examples were hydrogen bonded and the other 169 examples were not. Among the non-prolyl 2301 examples we have four possibilities: a residue could be hydrogen bonded through the amide nitrogen, carbonyl oxygen, both or neither. It was found that 380 examples were hydrogen bonded through the amide nitrogen, 391 through the carbonyl oxygen and 656 through both and 874 examples had no hydrogen bonds to the backbone polar groups.

It can be seen from above that close to 41% of the 2564 residues in E_I strands are not hydrogen bonded. Given the solvent-exposed nature of these segments, the hydrogen bonding potential of these polar groups could be satisfied through the surrounding water molecules.

Interaction of E_I strands with the adjacent molecules in the crystals

Crystallographic symmetry-related molecules of all the protein structures in our dataset with at least one potential isolated E-strand have been generated as outlined in the Materials and methods section. We investigated the interaction of E_I strands with all the adjacent molecules in the crystal lattice. In the cases with more than one molecule in the asymmetric unit interaction between an E_I strand and the other chains within the asymmetric unit was also studied.

Of the 518 putative isolated E-strands identified in our analysis only 34 are involved in any prominent interaction with the adjacent molecules in the crystals. At least two hydrogen bonds involving the main-chain carbonyl or amide at the strands and polar groups from the adjacent molecules could be identified in these 34 examples. Eighteen of these examples result from interaction between two molecules in the asymmetric unit of the crystal structure. Some of these examples correspond to the β-sheet formation with β-strands coming from different tertiary structures such as seen in the structure of pea lectin. Other examples correspond to interactions between the main-chain carbonyl or amide in the strand with the side-chain polar atoms from a neighbouring molecule.

Based on these observations it is clear that 484 (= 518 – 34) E-strands in the dataset deemed as isolated by considering a copy of the tertiary structure remain isolated even if the adjacent molecules in the crystals are considered.

Conservation of E_I strands in families of homologous proteins

In order to assess the extent to which the E_I strands are conserved in homologous proteins, an analysis was carried out on a database of families of aligned homologous protein structures (HOMSTRAD) (Mizuguchi *et al.*, 1998).

Considering 97 families of the database that had more than three members, one structure from each family was chosen at random to function as the reference structure. Isolated E_I strands present in that structure were identified and their index of conservation was computed amongst the members of that family. The index of conservation (I) of a E_I strand from the reference structure was calculated as the percentage ratio of the number of members of the family in which at least 90% of the length of the segment from the reference structure is structurally conserved to the total number of members in that family. The results are shown in Figure 7. It can be seen that about 41% of the 290 examples of E_I strands analysed are conserved with a very high value of the index (ranging from 60 to 100%). However, the majority of the examples have low indices of conservation. Thus, the data seem to suggest that these segments are indeed variable in structure, resembling the loop segments of proteins.

Conclusions

Isolated E-strands commonly occur in proteins. In spite of the lack of regular hydrogen bonding partners they seem to form stable stretches which are potentially stabilized by the surrounding water molecules and the side chains of polar residues in the protein. It has also been shown that almost all of these isolated E-strands remain isolated even in the context of quaternary structure and interaction of a protein molecule with neighbouring copies in the crystal lattice. In terms of the residue preferences, except for the abundance of proline, they show good similarity to the β-strands (that are part of sheets), supporting the fact that strand formation is determined by the intrinsic preferences of certain types of residues. On the other hand, they have their other characteristics similar to loops. They seem to be as exposed to the solvent as loops and the hydrophobic groups present in these strands behave in a similar fashion to those in the loops, being buried from the solvent. These extended structures seem to be supplementing the loops in efficiently traversing long distances in the protein with a minimal number of residues. Finally, these observations indicate that isolated E-strands occupy an individual existence with its characteristics shared partly with that of the β-sheet forming β-strands and partly with the loops.

Acknowledgement

One of us (N.S.) was supported by a senior research fellowship from the Wellcome Trust, London.

References

Adzhubei,A.A. and Sternberg,M.J.E. (1993) *J. Mol. Biol.*, **229**, 472–493.
Adzhubei,A.A., Eisenmenger,F., Tumanyan,V.G., Zinke,M., Brodzinski,S. and Esipova,N.G. (1987a) *Biofizika (Moscow, Engl. Ed.)* **32**, 159–162
Adzhubei,A.A., Eisenmenger,F., Tumanyan,V.G., Zinke,M., Brodzinski,S. and Esipova,N.G. (1987b) *J. Biomol. Struct. Dyn.*, **5**, 689–704.
Adzhubei,A.A., Eisenmenger,F., Tumanyan,V.G., Zinke,M., Brodzinski,S. and Esipova,N.G. (1987c) *Biochem. Biophys. Res. Commun.*, **146**, 934–938.
Ananthanarayanan,V.S., Soman,K.V. and Ramakrishnan,C. (1987) *J. Mol. Biol.*, **198**, 705–709.
Aurora,R. and Rose,G.D. (1998) *Protein Sci.*, **7**, 21–38
Baker,E.N. and Hubbard,R.E. (1984) *Prog. Biophys. Mol. Biol.*, **44**, 97–179.
Barlow,D.J. and Thornton,J.M. (1988) *J. Mol. Biol.*, **201**, 601–619.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Chothia,C. (1976) *J. Mol. Biol.*, **105**, 1–14.
Chou,P.Y. and Fasman,G.D. (1974) *Biochemistry*, **13**, 211–222.
Creamer,T.P. and Rose,G.D. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 5937–5941.
Creamer,T.P. and Rose,G.D. (1994) *Proteins*, **19**, 85–97.
Creamer,T.P. and Rose,G.D. (1995) *Protein Sci.*, **4**, 1305–1314.
Dinner,A.R., Lazardis,T. and Karplus,M. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 9068–9073.
Evans,S.V. (1993) *J. Mol. Graph.*, **11**, 134–138.
Finkelstein,A.V. (1995) *Protein Eng.*, **8**, 207–209.
Gibrat,J.F., Robson,B. and Garnier,J. (1991) *Biochemistry*, **30**, 1578–1586.
Gunasekaran,K., Nagarajaram,H.A., Ramakrishnan,C. and Balaram,P. (1998) *J. Mol. Biol.*, **275**, 917–932.
Henrick, K. and Thornton,J.M. (1998) *Trends Biochem. Sci.*, **23**, 358–361.
Hurley,J.H., Mason,D.A. and Matthews,B.W. (1992) *Biopolymers*, **32**, 1443–1446.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kim,C.A. and Berg,J.M. (1993) *Nature*, **362**, 267–270.
Kumar,S. and Bansal,M. (1998) *Biophys. J.*, **75**, 1935–1944.
Kunishima,N., Fukuyama,K., Matsubara,H., Hatanaka,H., Shibano,Y. and Amachi,T. (1994) *J. Mol. Biol.*, **235**, 331–344.
Lee,B. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.
Lifson,S. and Sander,C. (1979) *Nature*, **282**, 109–111.
Lifson,S. and Sander,C. (1980) *J. Mol. Biol.*, **139**, 627–639.
MacArthur,M.W. and Thornton,J.M. (1991) *J. Mol. Biol.*, **218**, 397–412.
Martin,A.C.R., Toda,K., Stirk,H.J. and Thornton,J.M. (1995) *Protein Eng.*, **8**, 1093–1101.
Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) *J. Mol. Biol.*, **196**, 641–656.
Minor,D.L.,Jr and Kim,P.S. (1994a) *Nature*, **371**, 264–267.
Minor,D.L.,Jr and Kim,P.S. (1994b) *Nature*, **367**, 660–663.
Minor,D.L.,Jr and Kim,P.S. (1996) *Nature*, **380**, 730–734.
Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington J.P. (1998) *Protein Sci.*, **7**, 2469–2471.
Munoz,V. and Serrano,L. (1994) *Proteins*, **20**, 301–311.
Otzen,D.E. and Fersht,A.R. (1995) *Biochemistry*, **34**, 5718–5724.
Overington,J., Johnson,M.S, Sali,A. and Blundell,T.L. (1990) *Proc. R. Soc. London, Ser. B*, **241**, 132–145.
Pauling,L. and Corey,R.B. (1951) *Proc. Natl Acad. Sci. USA*, **37**, 251–256.
Pauling,L., Corey,R.B. and Branson,H.R. (1951) *Proc. Natl Acad. Sci. USA*, **37**, 205–211.
Presta,L.G. and Rose,G.D. (1988) *Science*, **240**, 1632–1641.
Ramakrishnan,C. and Soman,K.V. (1982) *Int. J. Pept. Protein Res.*, **20**, 218–237.
Richardson,J.S. and Richardson,D.C. (1988) *Science*, **240**, 1648–1652.
Richardson,J.S., Getzoff,E.D. and Richardson,D.C. (1978) *Proc. Natl Acad. Sci. USA*, **75**, 2574–2578.
Rose,G.D. and Wolfenden,R. (1993) *Annu. Rev. Biophys Biomol. Struct.*, **22**, 381–415.
Smith,C.K. and Regan,L. (1995) *Science*, **270**, 980–982.
Smith,C.K., Withka,J.M. and Regan,L. (1994) *Biochemistry*, **33**, 5510–5517.
Soman,K.V. and Ramakrishnan,C. (1983) *J. Mol. Biol.*, **170**, 1045–1048.
Soman,K.V. and Ramakrishnan,C. (1986) *Int. J. Biol. Macromol.*, **8**, 89–95.
Stapley,B.J. and Creamer,T.P. (1999) *Protein Sci.*, **8**, 587–595.
Stapley,B.J. and Doig,A.J. (1997) *J. Mol. Biol.*, **272**, 456–464.
Sternberg,M.J.E. and Thornton,J.M. (1977) *J. Mol. Biol.*, **110**, 285–296.
Stickle,D.F., Presta,L.G., Dill,K.A. and Rose,G.D. (1992) *J. Mol. Biol.*, **226**, 1143–1159.
Street,A.G. and Mayo,S.L. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 9074–9076.
Swindells,M.B., MacArthur,M.W. and Thornton,J.M. (1995) *Nat. Struct. Biol.*, **2**, 596–603.
Venkatachalam,C.M. (1968) *Biopolymers*, **6**, 1425–1436.
Vlasov,P.K., Kilosanidze,G.T., Ukrainskaya,D.L., Kuzmin,A.V., Tumanyan,V.G. and Esipova,N.G. (2001) *Biofizika (Moscow, Engl. Ed.)* **46**, 573–576.
Wouters,M.A. and Curmi,P.M.G. (1995) *Proteins*, **22**, 119–131.

Received April 8, 2002; revised March 31, 2003; accepted April 8, 2003