*Proc. Natl. Acad. Sci. USA*
Vol. 92, pp. 12136–12140, December 1995
Biophysics

# How are close residues of protein structures distributed in primary sequence?

LUCIANO BROCCHIERI AND SAMUEL KARLIN

Department of Mathematics, Stanford University, Stanford, CA 94305-2125

*Contributed by Samuel Karlin, September 7, 1995*

ABSTRACT **Structurally neighboring residues are categorized according to their separation in the primary sequence as proximal (1–4 positions apart) and otherwise distal, which in turn is divided into near (5–20 positions), far (21–50 positions), very far (>50 positions), and interchain (from different chains of the same structure). These categories describe the linear distance histogram (LDH) for three-dimensional neighboring residue types. Among the main results are the following: (*i*) Nearest-neighbor hydrophobic residues tend to be increasingly distally separated in the linear sequence, thus most often connecting distinct secondary structure units. (*ii*) The LDHs of oppositely charged nearest-neighbors emphasize proximal positions with a subsidiary maximum for very far positions. (*iii*) Cysteine–cysteine structural interactions rarely involve proximal positions. (*iv*) The greatest numbers of interchain specific nearest-neighbors in protein structures are composed of oppositely charged residues. (*v*) The largest fraction of side-chain neighboring residues from $\beta$-strands involves near positions, emphasizing associations between consecutive strands. (*vi*) Exposed residue pairs are predominantly located in proximal linear positions, while buried residue pairs principally correspond to far or very far distal positions. The results are principally invariant to protein sizes, amino acid usages, linear distance normalizations, and over- and underrepresentations among nearest-neighbor types. Interpretations and hypotheses concerning the LDHs, particularly those of hydrophobic and charged pairings, are discussed with respect to protein stability and functionality. The pronounced occurrence of oppositely charged interchain contacts is consistent with many observations on protein complexes where multichain stabilization is facilitated by electrostatic interactions.**

Literature on sequential positions of structurally associated residues essentially starts with the study of Thornton (1) on the linear distribution of cysteines joined in disulfide bridges. Stickle *et al.* (2) survey "demographic trends" of residue pairs forming hydrogen bonds. Sippl (3) and Bauer and Beyer (4) plot $C^\alpha$–$C^\alpha$ distances for a prescribed separation in the primary sequence and propose energy potentials in residue atomic interactions as a function of interatomic distances and sequential separation. In this paper, we address the following problem. For all three-dimensional (3D) neighboring residue pairs (for example, nearest-neighbors or contact-neighbors as defined below), we investigate how they are distributed in the primary sequence. This is of interest in helping to understand how a tertiary fold is formed and in delineating chemical and physical constraints underlying the geometry of protein structures. Related questions include the following: To what extent do 3D nearest-neighbor pairings correspond to a single secondary structure, to distinct secondary structures, or to distinct domains? Are there significant differences in the linear-distance histogram (LDH) when both component residues of close structural pairs belong to $\alpha$-helices, $\beta$-strands, coils, or mixtures of these? How are these assessments influenced by solvent accessibility parameters?

## DATA AND METHODS

Our analyses are based on a data set of 172 nonhomologous well-resolved protein structures (see ref. 5 for criteria of selection and a complete list). This data set includes 109 monomers (23,560 residues), 49 dimers (18,319 residues), 1 trimer (349 residues), 12 tetramers (9023 residues), and 1 hexamer (1509 residues) for a total of 52,760 residues. Structural neighbors are defined based on $d_m$ distance, calculated as the minimum distance between side-chain atoms of the residue pair, or $D_m$ distance, calculated as the minimum distance between all atoms (side chain and backbone) of the residue pair (5). $d_m$ ($D_m$) nearest-neighbor pairings are defined as a residue and its $d_m$ distance ($D_m$ distance) closest residue. $d_m$ and $D_m$ contact-neighbors refer to residue pairs closer in $d_m$ distance and $D_m$ distance, respectively, than a given threshold (e.g., 4.0 Å). From our collection of protein structures, all $d_m$ distance and $D_m$ distance nearest-neighbor residue pairings and all 4.0 Å contact-neighbors were ascertained. An aggregate of 38,259 nearest-neighbor residue pairs for the $d_m$ distance and 36,628 for the $D_m$ distance, each composed of a reference residue and its $d_m$ or $D_m$ distance nearest-neighbor, were compiled (mutual nearest-neighbors were considered only once).

We refer to a 3D nearest-neighbor residue pairing as proximal if the component residues are at most four positions apart in the primary sequence and otherwise distal. For each type of nearest-neighbors the component residues are distributed in a LDH of different linear distance categories (proximal, 1–4 positions; near, 5–20 positions; far, 21–50 positions; very far, >50 positions; interchain, different chains of the same structure). The interval size differences are essentially balanced in that structural nearest-neighbor linear separations are expected to be more near than far. The proximal interval frequently includes 3D neighbors on the same side of $\alpha$-helices (separation of 3 or 4) or of $\beta$-strands (separation of 2). Since secondary structure units mostly do not exceed 20 residues, residues belonging to the same secondary structure generally fall in the proximal or near categories. Since the average exon length in higher eukaryotes is 40–45 residues (6, 7, 8), pairs within this distance will include residues from a common exon, module, or domain.

## RESULTS AND DISCUSSION

Contact-neighbor results highly correlate with those obtained using nearest-neighbors (Fig. 1). Therefore, the presentation below will concentrate mostly on nearest-neighbor residue pairs.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: 3D, three-dimensional; LDH, linear distance histogram.

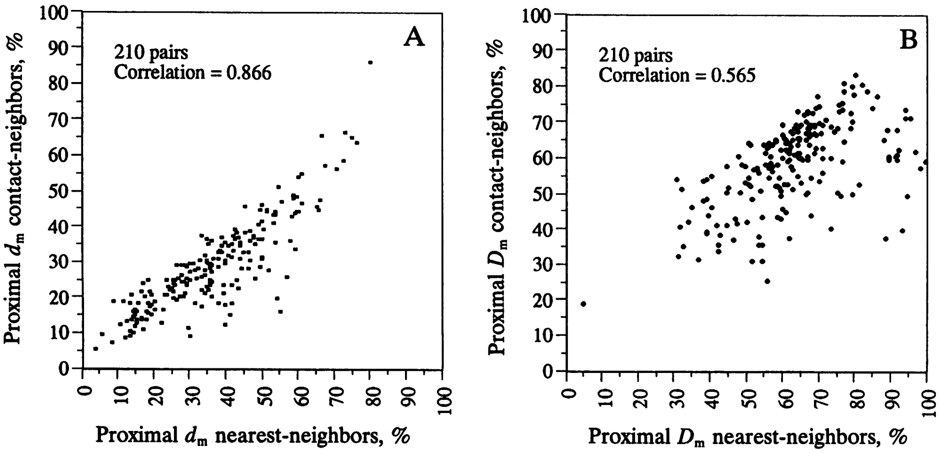

FIG. 1. Contact-neighbor pairs are here defined as pairings formed by a given residue and those residues with $d_m$ distance ($D_m$ distance) not exceeding 4 Å from that residue. (*A*) Plot of percentage proximal contact-neighbors vs. percentage proximal nearest-neighbors ($d_m$ distance) for all *XY* residue pairings is displayed. The very high correlation, 0.866, indicates that the results of the analyses on $d_m$ nearest-neighbors largely apply to $d_m$ contact-neighbors. (*B*) A strong significant correlation, 0.565, prevails for comparing $D_m$ nearest-neighbors with $D_m$ contact-neighbors.

**Numbers of Nearest-Neighbor Pairs.** Table 1 displays numbers and LDHs of selected $d_m$ distance and $D_m$ distance nearest-neighbor residue pairings. The following observations stand out: (*i*) Oppositely charged nearest-neighbor pairings EK, DK, DR, and ER (single-letter amino acid code), have highest or near-highest counts among nearest-neighbor pairs, reflecting on a preponderance of ionic attractions. (*ii*) Nearest-neighbor contacts among hydrophobics are abundant, featuring the pairings LV, AL, IL, AV, FL, LL, IV, FV, VV, and AI, each in excess of 300 occurrences, attesting to the importance of hydrophobic associations. (*iii*) Combinations of small hydrophilic residues S, D, and G show high counts, ≈400, of nearest-neighbor pairings, a contributing factor being their mutual occurrence in loops and turns and in common surface locations. (*iv*) There are 311 $d_m$ distance (306 $D_m$ distance) CC nearest-neighbor pairs (CC is also the most overrepresented pair) almost entirely accounted for by disulfide bridges (1, 5).

Table 1. Selected structural nearest-neighbor LDHs

| | $d_m$ distance LDH | | | | | | $D_m$ distance LDH | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pair | Total no. | 1–4 | 5–20 | 21–50 | >50 | Inter | Total no. | 1–4 | 5–20 | 21–50 | >50 | Inter |
| AF | 299 | 15.7 | 21.4 | 20.7 | 35.1 | 7.0 | 188 | 71.3 | 13.8 | 5.9 | 8.0 | 1.1 |
| AI | 331 | 13.6 | 20.8 | 26.9 | 36.3 | 2.4 | 293 | 61.4 | 13.0 | 11.9 | 12.6 | 1.0 |
| AM | 126 | 8.7 | 29.4 | 20.6 | 34.1 | 7.1 | 119 | 77.3 | 6.7 | 13.4 | 2.5 | 0.0 |
| II | 212 | 12.7 | 20.8 | 30.2 | 32.5 | 3.8 | 125 | 53.6 | 8.0 | 23.2 | 15.2 | 0.0 |
| LL | 394 | 14.7 | 21.3 | 25.4 | 34.8 | 3.8 | 311 | 66.6 | 9.0 | 10.9 | 12.9 | 0.6 |
| LV | 629 | 14.1 | 24.5 | 25.0 | 32.8 | 3.7 | 454 | 60.1 | 13.7 | 10.8 | 12.8 | 2.6 |
| FY | 163 | 13.5 | 27.0 | 21.5 | 27.0 | 11.0 | 89 | 42.7 | 19.1 | 18.0 | 14.6 | 5.6 |
| DK | 675 | 49.6 | 10.7 | 12.6 | 21.6 | 5.5 | 558 | 60.9 | 8.6 | 11.3 | 16.5 | 2.7 |
| EK | 729 | 43.3 | 16.6 | 13.6 | 21.9 | 4.5 | 563 | 58.4 | 10.8 | 7.6 | 18.5 | 4.6 |
| DR | 572 | 33.0 | 12.8 | 21.5 | 21.7 | 11.0 | 458 | 45.0 | 10.7 | 17.0 | 18.8 | 8.5 |
| ER | 566 | 35.0 | 16.6 | 14.1 | 23.9 | 10.4 | 428 | 40.7 | 14.7 | 12.1 | 22.0 | 10.5 |
| DH | 200 | 30.5 | 17.5 | 19.5 | 28.0 | 4.5 | 188 | 38.3 | 12.2 | 22.3 | 23.4 | 3.7 |
| EH | 166 | 39.2 | 13.9 | 16.9 | 26.5 | 3.6 | 145 | 44.8 | 11.7 | 18.6 | 22.1 | 2.8 |
| DS | 462 | 60.6 | 9.5 | 11.7 | 15.2 | 3.0 | 469 | 66.7 | 8.1 | 10.4 | 12.4 | 2.3 |
| ES | 351 | 51.6 | 13.4 | 14.8 | 15.4 | 4.8 | 342 | 64.9 | 9.6 | 12.3 | 10.2 | 2.9 |
| CC | 311 | 3.9 | 35.7 | 28.3 | 28.3 | 3.9 | 306 | 5.2 | 35.6 | 26.8 | 28.4 | 3.9 |
| GG | 339 | 80.2 | 2.9 | 5.3 | 8.3 | 3.2 | 186 | 59.1 | 11.8 | 10.2 | 16.1 | 2.7 |
| GP | 291 | 73.2 | 7.2 | 7.6 | 9.6 | 2.4 | 267 | 88.4 | 1.9 | 4.1 | 5.2 | 0.4 |
| AY | 298 | 20.5 | 18.5 | 27.5 | 28.9 | 4.7 | 175 | 65.1 | 8.6 | 9.7 | 15.4 | 1.1 |
| DY | 231 | 18.6 | 19.0 | 29.0 | 26.0 | 7.4 | 224 | 30.8 | 18.8 | 25.0 | 20.1 | 5.4 |
| EY | 236 | 26.3 | 16.9 | 22.9 | 27.5 | 6.4 | 200 | 39.5 | 16.0 | 18.5 | 20.5 | 5.5 |
| GY | 276 | 27.5 | 21.7 | 19.2 | 26.1 | 5.4 | 220 | 62.7 | 14.1 | 10.9 | 10.9 | 1.4 |
| IY | 259 | 19.3 | 20.1 | 29.3 | 26.6 | 4.6 | 140 | 39.3 | 25.7 | 12.9 | 20.0 | 2.1 |
| TY | 200 | 29.0 | 26.0 | 15.0 | 23.0 | 7.0 | 179 | 51.4 | 18.4 | 9.5 | 17.3 | 3.4 |
| DW | 82 | 25.6 | 30.5 | 20.7 | 15.9 | 7.3 | 68 | 32.4 | 20.6 | 17.6 | 13.2 | 16.2 |
| EW | 77 | 36.4 | 10.4 | 27.3 | 22.1 | 3.9 | 42 | 57.1 | 7.1 | 14.3 | 14.3 | 7.1 |
| TW | 75 | 36.0 | 22.7 | 10.7 | 28.0 | 2.7 | 76 | 76.3 | 13.2 | 3.9 | 6.6 | 0.0 |
| Total | 38,259 | 37.0 | 17.6 | 17.6 | 23.2 | 4.7 | 36,628 | 64.4 | 11.4 | 9.7 | 12.2 | 2.2 |

From a collection of 172 nonhomologous protein structures, all $d_m$ distance and $D_m$ distance nearest-neighbor residue pairings were determined. Total counts and percentages for the different linear distance categories 1–4 (proximal), 5–20 (near), 21–50 (far), >50 (very far), and connecting different polypeptide chains [interchain (Inter)] are displayed for selected structural nearest-neighbor residue pairings. A complete list of all pairs is available via anonymous ftp (gnomic.stanford.edu:/pub/linear_correspondence.data). In the counts of nearest-neighbors, mutual nearest-neighbors are considered only once.

The aggregate LDH over all nearest-neighbor pairings can be used as a reference standard. This gives the totals shown in Tables 1–3 (bottom rows).

Clearly there are more 3D residue pairs in proximal linear positions in the $D_m$ distance than in the $d_m$ distance measure, 64.4% against 37.0%, reflecting mainly proximal backbone–backbone interactions (2, 9). The $D_m$ distance distal categories are essentially of congruent proportions. The $d_m$ distance LDH shows a primary peak for proximal positions and a secondary peak for very far positions (U-shaped distribution).

**Hydrophobic and Aromatic Pairings.** There is an inverse relationship between the extent of 3D nearest-neighbors of proximal pairings and the degree of hydrophobicity. In fact, hydrophobic $d_m$ distance nearest-neighbor pairings display generally increasing counts in the LDH (Tables 1 and 2). Moreover, $d_m$ distance nearest-neighbor pairings involving aliphatic residues are distal in >80% of all cases. The linear distribution of proximal interactions among aliphatic residues reflects the usage of aliphatic residues in $\alpha$-helices, with a peak at separation distance ±4 in the primary sequence (data not shown). The preponderant contacts between hydrophobic amino acids of distal separation highlight their role in packing distinct secondary structure units of the protein.

$d_m$ distance nearest-neighbor interactions between aromatic residues also emphasize distal residues (e.g., FY 13.5% proximal, 75.5% distal), suggesting a role of the aromatic rings in connecting distinct secondary structure elements or a spatial affinity for functional purposes (5). Aromatic–aromatic nearest-neighbors are significantly more frequent in the $d_m$ distance than in the $D_m$ distance. Aromatic rings can contribute to hydrophobic interactions but they can also favorably interact with solvent and polar residues (10, 11).

**Nearest-Neighbor Charged Residues.** Oppositely charged residue pairs as structural $d_m$ distance nearest-neighbors are about equally frequent in proximal and near vs. far and very far linear positions. Nearest-neighbor pairings consisting of oppositely charged residues show a U-shaped LDH with a maximum count for proximal positions and a second maximum number among very far separations. A U-shaped distribution is also observed for the pairings DH and EH.

Among charged residues, arginine is paramount in distal interactions. Arginine is capable not only of charge–charge interactions but also of fomenting cation–aromatic interactions (12, 13), where interaction with the arginine delocalized charge may be decisive. The large size of these residues would also favor distal contacts. In this context, ionic and hydrogen bonds are known to stabilize amphipathic helices. Interhelical ionic interactions may also be relevant (14). The results showing that R is more likely to bind to D or E than K is related with the ability of the large guanidinium group to pair more easily with an anionic group. Most $D_m$ distance nearest-neighbors correspond to proximal linear positions, but the oppositely charged DR and ER tend to be more distal than the DK and EK nearest-neighbor pairings.

**Small Polar Residues.** Generally, $d_m$ distance nearest-neighbors among small polar residues emphasize proximal distances, presumably reflecting common exposed locations. With oppositely charged and hydrophobic pairings, the DS nearest-neighbor pairing shows the greatest numbers of nearest-neighbor occurrences in both the $d_m$ and $D_m$ measures. An asymmetric strong peak of DS nearest-neighbors for proximal distance +2 (S carboxyl to D; data not shown) is consistent with the arrangement of serine (S) at the N-cap of helices and aspartate (D) in the preceding turn (15–17). By contrast, for ES the $d_m$ proximal interactions tend to place E carboxyl to S, often near the N-cap of helices (stabilizing the helix) and S in the preceding turn (data not shown).

**Nearest-Neighbor Cysteine Pairs.** Structural cysteine doublets (CC) rarely involve proximal residues but are prevalent in near or far positions (cf. ref. 1). The $d_m$ and $D_m$ linear correspondences are almost always identical, as would be expected from interactions predominantly based on disulfide bonding. Correlations of linear CC di-residues (positions $i$, $i$ + 1) are significantly underrepresented (18) as contiguous cysteines apparently hinder conformational flexibility. On the other hand, noncontiguous linear cysteine doublets—i.e., the di-residue forms $CX_kC$ ($k$ = 2, 3, or 4)—are overrepresented. These include cysteines participating in metal ion coordination, but such forms rarely provide 3D nearest-neighbors.

**Small Residues.** The glycine (G) residue is often characterized as a "filler" or "hinge," conferring conformational flexibility and precision. Structural nearest-neighbor pairs involving G predominantly favor proximal (mostly contiguous) linear positions. Indeed, GG among all nearest-neighbor pairings register the highest percentage of proximal contacts (80%). To a substantial but lesser extent, the same applies to alanine (A), which serves in versatile ways in both core and surface positions and in different secondary structures. Because of its special cyclic side chain, the $d_m$ distance nearest-neighbors with proline are preponderantly associated to their immediately adjacent amino residue.

**Pairings of Residues Capable of Simultaneous Hydrophobic and Hydrophilic Interactions.** The linear distances of each of the $d_m$-distance nearest-neighbor pairings DH, EH, AY, DY, EY, GY, IY, TY, DW, EW, and TW depict a LDH approximately constant or U-shaped. Note that these pairings all involve at least one of the residues Y, W, or H. In many situations, Y can find a microenvironment that allows both hydrophobic (through its aromatic ring) and polar (through its hydroxyl group) interactions. Histidine is a prominent residue in a variety of proteinases, in many catalytic reactions, in metal coordination, or as a controllable element in conformational changes (19). W is mostly a hydrophobic residue (16), but it also entails hydrogen bonding potential through its imino nitrogen. All aromatics project $\pi$-electron clouds with a positive hydrogen atom periphery capable of generating electrostatic attractions in the case of aromatic–aromatic, cation–aromatic, and anion–aromatic interactions (e.g., see refs. 10 and 20–23). The residues Y, W, and H are distinguished in that for the $d_m$ distance all residue types are overrepresented as nearest-neighbors of at least one of the residues Y, W, and H (5).

Table 2. Structural nearest-neighbor LDHs for selected group types

| | $d_m$ distance LDH | | | | | | $D_m$ distance LDH | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pair | Total no. | 1–4 | 5–20 | 21–50 | >50 | Inter | Total no. | 1–4 | 5–20 | 21–50 | >50 | Inter |
| [DE][KR] | 2,542 | 40.8 | 14.2 | 15.2 | 22.2 | 7.6 | 2,007 | 52.3 | 11.0 | 11.8 | 18.7 | 6.2 |
| [LIVMF][LIVMF] | 4,287 | 15.7 | 23.5 | 25.0 | 31.5 | 4.3 | 2,890 | 55.7 | 15.3 | 14.3 | 13.3 | 1.4 |
| [YW][EDKR] | 1,234 | 26.3 | 19.4 | 23.4 | 24.2 | 6.6 | 877 | 40.8 | 15.3 | 18.0 | 19.4 | 6.5 |
| [YW][LIVMF] | 1,731 | 18.4 | 24.3 | 22.6 | 30.3 | 4.4 | 961 | 48.2 | 21.0 | 11.1 | 17.7 | 2.0 |
| [F][DEKR] | 482 | 32.4 | 18.7 | 17.4 | 25.9 | 5.6 | 529 | 67.7 | 8.9 | 8.3 | 13.8 | 1.3 |
| Total | 38,259 | 37.0 | 17.6 | 17.6 | 23.2 | 4.7 | 36,628 | 64.4 | 11.4 | 9.7 | 12.2 | 2.2 |

Structural nearest-neighbor LDHs are shown for selected amino acid groupings: negatively [DE] and positively [KR] charged, major hydrophobics (LIVMF), and aromatics distinguishing polar [YW] and nonpolar [F].

**Group LDHs.** Table 2 reports the LDH for selected nearest-neighbor residue groupings. The highest percentages of interchain nearest-neighbor pairs consist of oppositely charged residues and aromatics associated with charged residues. This presumably reflects salt bridge or hydrogen bond connections and possibly cationic and anionic–aromatic interactions. The counts of interchain contacts among hydrophobic nearest-neighbors is approximately the same as the counts of interchain charge–charge pairings, although the frequency of hydrophobic residues is more than twice that of charged residues. The LDH of oppositely charged nearest-neighbor side-chain pairings is bimodal (principal mode for proximal pairings and secondary mode for very far pairings), whereas hydrophobic pairings show an increasing LDH plot. The [YW] [EDKR] interactions have component residues in the LDH tallies that are approximately uniformly distributed.

**Protein Size Effects.** To confirm that protein size did not bias our results, the data set was divided into three sets of approximately equal size (in aggregate residue numbers)—namely, relatively small-sized structures (protein length, ≤317 aa; mostly <200 aa), medium-sized structures (318–478 aa), and large-sized structures (490–1544 aa). It is *a priori* conceivable that in small globular protein structures having relatively greater surface area per volume, pairings would tend to be more proximal and thus could influence the linear correspondence statistics. Correlations of percentage of proximal nearest-neighbors among these three size structure sets were calculated. The correlations were very high, all exceeding 0.8 (data not shown), indicating that protein sizes are not very influential.

**LDHs and Secondary Structures.** We investigate the nature of the linear separation of nearest-neighbors when both components belong to the same secondary structure type. The results are summarized in Table 3. Several contrasts stand out. (*i*) For residues of $d_m$ nearest-neighbor side chains conditioned so that both residues belong to α-helices, almost 50% involve proximal residues and almost 40% involve far or very far linear pairings. In sharp contrast, for $D_m$ nearest-neighbors (side-chain and/or backbone contacts) in α-helices, we observe that >90% of the component residues are linearly proximal and mostly entail backbone–backbone hydrogen bonds. (*ii*) Given β-strand/β-strand nearest-neighbors, the largest fraction of these occur in side-chain contacts with component residues near (i.e., 5–20 positions apart), providing evidence that linearly successive β-strands are often structurally associated. (*iii*) The greatest number of secondary structure associations are from nearest-neighbor coil–coil residues and these are mostly linearly proximal, generally in exposed locations. (*iv*) A relative paucity of nearest-neighbor pairings involve α-helix/β-strand associations, and in these cases the residues are rarely proximal.

**LDHs and Solvent Accessibility (Table 3).** What is the linear disposition of interacting residues with respect to solvent accessibility? (Assessment of solvent accessibility for a residue pair is taken as their average solvent accessibility using the procedure of ref. 24.) Exposed $d_m$ nearest-neighbor residue pairs lie predominantly in linearly proximal positions. In contrast, among buried $d_m$ nearest-neighbors, the linear distance histogram emphasizes far or very far distal residues. For all degrees of solvent accessibility with $D_m$ nearest-neighbors, the component residues are principally proximal.

**Interchain Nearest-Neighbors (Table 4).** Interchain contacts are dominated by side-chain interactions and emphasize oppositely charged residues, suggesting that electrostatic forces frequently mediate the formation of multimeric structures. Actually in both the $d_m$ and $D_m$ distances the top four pairings of interchain nearest-neighbors are DR (63 and 39 cases, respectively), ER (59 and 45 cases), DK (37 and 15 cases), and EK (33 and 26 cases). Electrostatic interactions facilitate important processes such as protein sorting, translocation, docking, localization, orientation, oligomerization, and binding to DNA and other protein molecules. In this context, charge effects are relatively long range, rapid, and localized (25–27). Protein–protein interfaces are made up of a mixture of hydrophobic and hydrophilic residues. Conceivably, charge interactions efficiently orient and position the appropriate protein surfaces on which hydrophobic forces propitiously act in stabilizing the interface bonds.

Evidence of interchain and protein complex formation facilitated by electrostatic interactions is increasing. Specific examples featuring interface charge associations include glutathione *S*-transferase (homodimer and heterodimer forms), D-xylose isomerase (homotetramer), and aspartate carbamoyltransferase (12 chains) (data not shown). Also, polymerization of single-stranded DNA binding protein RecA monomers appears to be modulated by electrostatic contacts (unpublished work). The importance of electrostatic interactions in mediating and stabilizing quaternary structures may be tested experimentally by subjecting these structures to a gradient of salt concentration and by mutation studies. Compare with refs. 28 and 29 on binding of the human growth hormone (hGH) and the hGH receptor. Table 4 also presents the corresponding results for contact-neighbors, which parallel those for nearest-neighbors.

**Linear Associations and Protein Folding.** Our results suggest that the tertiary structure—e.g., the pattern of associa-

Table 3. Nearest-neighbor LDHs for different secondary structures and solvent accessibilities

| | $d_m$ distance LDH | | | | | | $D_m$ distance LDH | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pair | Total no. | 1–4 | 5–20 | 21–50 | >50 | Inter | Total no. | 1–4 | 5–20 | 21–50 | >50 | Inter |
| | | | | | Secondary structure | | | | | | | |
| Helix–helix | 6,614 | 47.5 | 8.7 | 17.6 | 21.6 | 4.6 | 7,995 | 92.7 | 1.3 | 2.3 | 2.9 | 0.9 |
| Strand–strand | 4,818 | 18.8 | 30.2 | 21.3 | 25.7 | 4.0 | 5,724 | 25.6 | 29.2 | 20.1 | 22.5 | 2.6 |
| Coil–coil | 11,855 | 55.0 | 11.4 | 13.1 | 16.2 | 4.3 | 12,932 | 74.2 | 7.1 | 7.5 | 9.1 | 2.1 |
| Helix–strand | 2,247 | 4.9 | 34.1 | 21.4 | 36.7 | 2.9 | 530 | 17.5 | 27.4 | 18.7 | 32.5 | 4.0 |
| Helix–coil | 7,128 | 30.2 | 20.8 | 17.2 | 26.2 | 5.6 | 5,405 | 63.1 | 14.0 | 7.6 | 12.4 | 3.0 |
| Strand–coil | 5,597 | 23.4 | 19.7 | 22.8 | 28.5 | 5.6 | 4,042 | 39.6 | 14.6 | 18.8 | 23.5 | 3.5 |
| Total | 38,259 | 37.0 | 17.6 | 17.6 | 23.2 | 4.7 | 36,628 | 64.4 | 11.4 | 9.7 | 12.2 | 2.2 |
| | | | | | Solvent accessibility | | | | | | | |
| Buried | 11,117 | 14.3 | 22.0 | 24.3 | 34.0 | 5.4 | 8,796 | 46.5 | 15.9 | 15.0 | 19.5 | 3.2 |
| Half-buried | 14,767 | 28.9 | 20.0 | 19.4 | 25.4 | 6.3 | 17,608 | 61.2 | 13.1 | 10.4 | 12.8 | 2.5 |
| Exposed | 12,375 | 67.0 | 10.8 | 9.4 | 10.8 | 2.1 | 10,224 | 85.2 | 4.6 | 4.1 | 5.0 | 1.0 |
| Total | 38,259 | 37.0 | 17.6 | 17.6 | 23.2 | 4.7 | 36,628 | 64.4 | 11.4 | 9.7 | 12.2 | 2.2 |

LDH of $d_m$ and $D_m$ distance nearest-neighbors is shown relative to selected structural classes: pairs of residues of the same or different secondary structure type (α-helix, β-strand, coil) and pairs of residues with average percentage surface accessibilities of ≤7 (buried pairs), >7 and ≤40 (half-buried), and >40 (exposed).

Table 4. Numbers of most frequent interchain interactions

| Nearest-neighbors | | | | Contact-neighbors | | | |
|---|---|---|---|---|---|---|---|
| Pair | $d_m$ dist. | Pair | $D_m$ dist. | Pair | $d_m$ dist. | Pair | $D_m$ dist. |
| DR | 63 | ER | 45 | DR | 96 | DR | 112 |
| ER | 59 | DR | 39 | ER | 82 | ER | 98 |
| DK | 37 | EK | 26 | FL | 55 | LV | 71 |
| EK | 33 | DK | 15 | DK | 50 | FL | 69 |
| FL | 24 | LR | 14 | LV | 48 | LT | 68 |
| LV | 23 | GN | 13 | EK | 45 | LR | 63 |
| NS | 23 | CC | 12 | RY | 45 | EK | 53 |
| AF | 21 | LV | 12 | LY | 40 | LL | 53 |
| IL | 19 | QT | 12 | DY | 37 | DK | 53 |
| EN | 19 | QR | 12 | PR | 37 | GI | 51 |
| VV | 19 | DY | 12 | LL | 34 | RY | 51 |
| FY | 18 | DS | 11 | KY | 34 | GL | 49 |
| GN | 18 | EY | 11 | EN | 33 | IL | 49 |
| DY | 17 | DW | 11 | IL | 33 | DY | 47 |
| CC | 12 | CC | 12 | CC | 12 | CC | 13 |
| +− | 192 | +− | 125 | +− | 273 | +− | 316 |
| øø | 185 | øø | 40 | øø | 400 | øø | 570 |
| Total | 1781 | | 821 | | 3352 | | 5495 |

Counts of the most represented interchain nearest-neighbor and contact-neighbor pairings. $d_m$ dist. ($D_m$ dist.), pairings obtained by the $d_m$ distance measure ($D_m$ distance measure); +−, total count of oppositely charged interchain nearest-neighbors and contact-neighbors; øø, aggregate count of interchain nearest-neighbors and contact-neighbors involving the major hydrophobic residues L, I, V, M, and F.

tions between secondary structure elements of the protein—involves the close packing of hydrophobic side chains between rather than within secondary structure units. On the other hand, individual secondary structures are established by patterns of backbone hydrogen bonds and to some extent assisted by specific polar or ionic side-chain interactions. Packing density of α-helices or β-sheets has been extensively investigated in known structures. For example, α-helices can achieve close packing, intercalating the ridges formed by their side chains in a limited number of geometries (24). An optimal hydrophobic packing of secondary structure elements might by itself be sufficient to determine the native state conformation of the protein, as in a jigsaw puzzle model (19, 30). Alternatively, the folding process may involve first formation of the hydrophobic core of a flexible molten globule. Then, other interactions, involving extended polar (Y, W, H) or ionic (R) residues, may help orient the molten globule and maneuver various secondary structure elements until a favorable (native) conformation is attained.

In view of the pervasive overrepresentation of $d_m$ distance side-chain interactions with tyrosine (Y) by almost all residue types and the significant overrepresentations of tryptophan (W) and histidine (H) relative to many residue types, we proposed (5) that the residues Y, W, and H might perform as dynamic initiation and early intermediate foci of the protein fold. How does our analysis on linear sequence distances associated with structural closeness conform with the above hypothesis? The statistics that nearest-neighbors involving Y, W, or H show uniform or U-shaped LDH counts among the proximal, near, far, and very far categories is consistent with their capacity for versatile hydrophobic and hydrophilic interactions.

The predominance of $D_m$ distance nearest-neighbors in proximal positions among hydrophobic pairings seems to reflect the early formation of backbone–backbone hydrogen bonds in individual secondary structures. The relatively low frequency of side-chain nearest-neighbor proximal interactions and the preponderance of distal (especially far and very far) side-chain interactions among hydrophobic (core) residues, and among Y, W, H, and R residues, emphasize the role of these side chains in intersecondary structure packing.

We are happy to acknowledge helpful discussions and/or comments on the manuscript from B. E. Blaisdell, F. Cohen, M. Gerstein, J. H. Griffin, and Z.-Y. Zhu. This work was supported in part by National Institutes of Health Grants 2R01GM10452-31 and 5R01HG00335-07 and by National Science Foundation Grant DMS 9403553.

1. Thornton, J. M. (1981) *J. Mol. Biol.* **151,** 261–287.
2. Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992) *J. Mol. Biol.* **226,** 1143–1159.
3. Sippl, M. J. (1990) *J. Mol. Biol.* **213,** 859–883.
4. Bauer, A. & Beyer, A. (1994) *Proteins* **18,** 254–261.
5. Karlin, S., Zuker, M. & Brocchieri, L. (1994) *J. Mol. Biol.* **240,** 768–901.
6. Traut, T. W. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2944–2948.
7. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16,** 9893–9905.
8. Smith, M. W. (1988) *J. Mol. Evol.* **27,** 45–55.
9. McDonald, K. & Thornton, J. M. (1994) *J. Mol. Biol.* **238,** 777–793.
10. Levitt, M. & Perutz, M. F. (1988) *J. Mol. Biol.* **201,** 751–754.
11. Suzuki, S., Green, P. G., Bumgarner, R. E., Dasgupta, S., Goddart, W. A., III, & Blake, G. A. (1992) *Science* **257,** 942–945.
12. Flocco, M. M. & Mowbray, S. L. (1994) *J. Mol. Biol.* **235,** 709–717.
13. Brocchieri, L. & Karlin, S. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 9297–9301.
14. Zhou, N. E., Cyril, M. K. & Hodges, R. S. (1994) *J. Mol. Biol.* **237,** 500–512.
15. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648–1652.
16. Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties* (Freeman, New York).
17. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47,** 45–118.
18. Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991) *Annu. Rev. Biophys. Biophys. Chem.* **20,** 175–203.
19. Richardson, J. S. & Richardson, D. C. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation,* ed. Fasman, G. D. (Plenum, New York), pp. 1–98.
20. Burley, S. K. & Petsko, G. A. (1988) *Adv. Protein Chem.* **39,** 125–189.
21. Hunter, C. A., Singh, J. & Thornton, J. M. (1991) *J. Mol. Biol.* **218,** 837–846.
22. Perutz, M. F. (1993) *Philos. Trans. R. Soc. London A* **345,** 105–112.
23. Kumpf, R. A. & Dougherty, D. A. (1993) *Science* **261,** 1708–1710.
24. Richmond, T. J. & Richards, F. M. (1978) *J. Mol. Biol.* **119,** 537–555.
25. Perutz, M. F. (1994) *Protein Sci.* **3,** 1629–1637.
26. Karlin, S. (1995) *Curr. Opin. Struct. Biol.* **20,** 360–371.
27. Brendel, V. & Karlin, S. (1989) *Proc. Natl. Acad. Sci. USA* **85,** 6637–6641.
28. Wells, J. A. (1991) *Methods Enzymol.* **202,** 390–411.
29. Clackson, T. & Wells, J. A. (1995) *Science* **267,** 383–386.
30. Harrison, S. C. & Durbin, R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4028–4030.