Coordination geometry of nonbonded residues in globular proteins

Ivet Bahar[1,2] and Robert L Jernigan[1]

Background: Two opposite views have been advanced for the packing of sidechains in globular proteins. The first is the jigsaw puzzle model, in which the complementarity of size and shape is essential. The second, the nuts-and-bolts model, suggests that constraints induced by steric complementarity or pairwise specificity have little influence. Here, the angular distributions of sidechains around amino acids of different types are analyzed, in order to capture the preferred (if any) coordination loci in the neighborhood of a given type of amino acid.

Results: Some residue pairs select specific coordination states with probabilities about ten times higher than expected for random distributions. This selectivity becomes more pronounced at closer separations leading to an effective free energy of stabilization as large as –2 RT for some sidechain pairs. A list of the most probable coordination sites around each residue type is presented, along with their statistical weights.

Conclusions: These data provide guidance as to how to pack selectively the nonbonded sidechains in the neighborhood of a central residue for computer generation of unknown protein structures.

Addresses: [1]Molecular Structure Section, Laboratory of Mathematical Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, MSC 5677, Room B-116, Building 12B, Bethesda, Maryland 20892-5677, USA. [2]Chemical Engineering Department and Polymer Research Center, Bogazici University, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey.

Correspondence: Robert L Jernigan
e-mail: jernigan@lmmb.nci.nih.gov

Key words: coordination geometry, residue packing, specificity of interresidue interactions

Received: 05 Jun 1996
Revisions requested: 19 Jun 1996
Revisions received: 01 Jul 1996
Accepted: 03 Jul 1996

Published: 28 Aug 1996
Electronic identifier: 1359-0278-001-00357

Folding & Design 28 Aug 1996, 1:357–370

Introduction

The role of sidechain packing in the protein folding code, and the occurrence of a random, rather than an ordered, organization of sidechains in native structures, has been questioned in light of some conflicting experimental observations and theoretical analyses [1,2]. On the one hand, there is the experimental line of evidence confirming the tight packing of sidechains in the core [3,4], supported by several factors such as the small compressibility of proteins, the destabilizing effect of large cavity-creating perturbations [5], the selective role of sidechain fits in stabilizing different multimers of coiled coils [6], the critical influence of the precise coordination of sidechains (by a ligand such as Zn^{++} for example) in designing proteins [7], and the restricted mobility of sidechains revealed by dynamic simulations. On the other hand, some observations suggest that sidechain packing or specific residue–residue interactions are not critically important in defining particular folds: most importantly, proteins can tolerate a broad diversity of mutations [8–11]. Other observations are that in protein families with a common fold, such as globins, protein pairs can have only 16% sequence identity [12], and that distantly related proteins having similar three-dimensional structures may share as few as 12% of their sidechain–sidechain contacts [13].

Comparing lattice simulations of model chains with and without sidechains, Bromberg and Dill [1] propose that sidechain packing in proteins is more like the packing of nuts and bolts in a jar than like the matching of jigsaw puzzle pieces. This contrasts with the model originally proposed by Richards [14]. According to the model of Bromberg and Dill [1], nonbonded interactions are not restrained by steric complementarity or pairwise specificity. Sidechain degrees of freedom are instead suggested to be strongly coupled to those of the backbone; the sequences of hydrophobic (H) or polar (P) monomers contribute more to dictating the chain fold than do sidechain sizes and shapes. However, as was also recognized by these authors [1], the lattice model on which these inferences are based neglects internal sidechain degrees of freedom, the different sizes and shapes of amino acids, and all energetic interactions other than excluded volume. For an evaluation of the validity of such views, it should be informative to carefully analyze sidechain packing geometries and energetics from protein structures. This is the goal of the present work. Mainly, the angular position of the nearest nonbonded neighbors located within the first coordination shell of each type of amino acid will be analyzed (Fig. 1).

Packing effects in relation to the uniqueness of protein native structures have been thoroughly reviewed by Richards and Lim [3], who point out that a knowledge of packing interactions is necessary for “predicting the functionally relevant precise structures of new proteins as well as for achieving the elusive goal of designing useful and

Figure 1

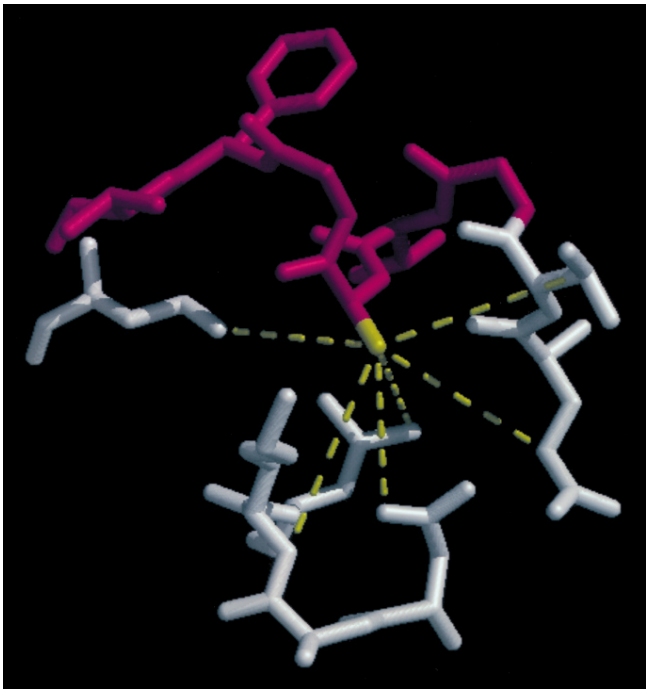


Illustration of the coordination of a given sidechain S_i by nonbonded neighbors S_j, $j \geq i+3$. The sidechain of the central residue (alanine), shown in yellow, is in close contact with six neighbors, shown by the yellow dotted lines, located in the first coordination shell ($r \leq 6.8$ Å). The portion of the protein shown in red refers to the nearest neighbors (up to $i \pm 2$) along the backbone. Here the central sidechain (Ala112 in the N-terminal domain of T4 lysozyme) is being coordinated by residues Lys83, Leu84, Asn81, Thr109, Glu108 and Leu118, the relative distances from the Ala112 C^{β} being 4.17 Å, 4.38 Å, 4.41 Å, 5.85 Å, 6.05 Å and 6.47 Å, respectively. We note that residues Lys83, Leu84, Asn81 (shown in the lower part) and Leu118 (upper part, left, shown in gray) are not sequentially near neighbors, while Thr109 and Glu108 are the third and fourth neighbors along the backbone. Such close neighbors have not been included in the distributions, their loci being predominantly determined by local constraints, rather than nonbonded preferences. We note that the near neighbors Phe114 and Thr115 are located at $r \geq 8.0$ Å.

stable proteins." If not the overall stability of the protein, then the structural uniqueness and, more importantly, the functionality are implied to correlate with the quality of packing. In parallel with those arguments, Lumb and Kim [15] recently called attention to the important role of specific interactions between buried polar groups in imparting structural uniqueness. In fact, they found a hydrogen bond between an asparagine pair, located on two neighboring helices at the hydrophobic interface between the monomers of a designed heterodimeric coiled coil, to be responsible for the choice of a unique structure. Variants in which leucine is substituted for asparagine lack structural uniqueness, in spite of their higher stability and ability to form heterotetramers. Thus, it is suggested that nonspecific hydrophobic interactions do contribute to protein stability in general, but that structural uniqueness is achieved by satisfying the hydrogen bond formation requirements of buried polar residues within a hydrophobic environment. This is a view that has been strongly supported by the values of residue–residue potential functions ([16,17]; I Bahar, RL Jernigan, unpublished data).

On the other hand, experiments of Lim and Sauer [18] with λ repressor reveal that the high tolerance to mutations, even at core positions, does not necessarily imply that no information is encoded in the specific sidechain packing of proteins in native structures. Even if the stabilities of overpacked or underpacked mutants remain comparable to that of the wild-type protein, the binding affinity of the mutants for DNA *in vivo* is shown to be significantly reduced [18]. Mutants whose core volumes deviate from that of the wild type by ± 3–6 methylene group equivalents are significantly destabilized and show no detectable DNA-binding affinity or antibody reactivity [5]. Other examples where function may be lost or reduced are where substitutions occur in active site regions or other functionally important areas. One interesting example of the latter is the cavity in Myb protein, which cannot be filled without loss of function [19]. Similarly, mutation of a buried valine to lysine in *Escherichia coli* enterotoxin is observed to cause loss of catalytic activity but no significant conformational change [20]. This residue is buried in the wild type, yet the bulkier and charged residue is substituted with minimal conformational change due to the presence of an internal cavity near the substitution region. This void volume is suggested to be required for accommodating residues displaced by the opening of the active site cleft in the wild-type protein, and the loss of activity may be explained by the reduced flexibility of this region. Likewise, the conformational flexibility afforded in cytochrome *c* by the leucine cluster in the heme pocket and the presence of an internal cavity are suggested to play a role in maintaining the electron transfer activity [21]. All these observations motivate a quantitative treatment of the coordination geometry of different types of sidechains in globular proteins.

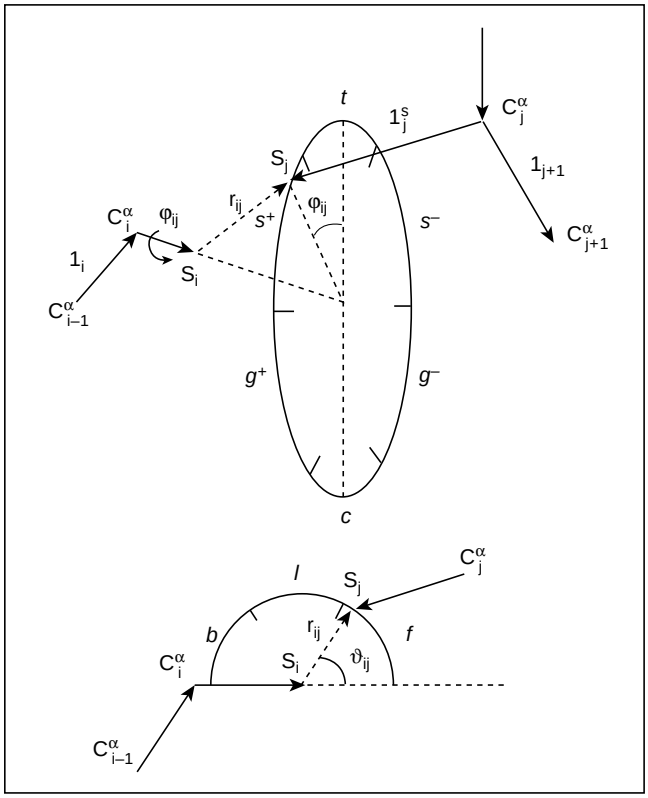
The preferred conformations of the backbone in proteins have been extensively explored. The geometries of the backbone in α-helices, β-sheets, and different types of turns are well established. The geometry of sidechains, on the other hand, has been characterized on a local scale, i.e. in terms of the most probable rotameric states accessible to sidechain bonds. Much less effort has been spent investigating and elucidating any three-dimensional coordination preferences of pairs of nonbonded sidechains. Pair distributions and effective contact potentials have been generally explored for sidechains as a function of radial separation [22], but with little attention to directional or orientational preferences that might characterize the specific interactions and the coordination geometries of par-

ticular residues. Indeed, the heart of the protein conformation problem might be considered to reside in the specific interactions between sidechains, which could be specific not only in energy but also in geometry.

The most detailed study, to date, of the packing of sidechains in proteins, considering both spatial and orientational distributions, has been performed by Singh and Thornton [23,24]. We also note that Vriend and Sander [25] made a detailed analysis of packing preferences of amino acid sidechains on the basis of atomic contacts. However, several issues remain to be clarified. Is the packing of sidechains random, devoid of any complementarity and directionality as suggested by Bromberg and Dill [1], or is there some preferred coordination geometry? If there is some preference, does packing geometry differ from one type of sidechain to another? Is it possible to identify a set of the most probable coordination loci around a given residue? How strong are these preferences? Do they significantly and usefully restrict conformational space, or is the spatial organization of nonbonded sidechains instead rather diffuse? Are there similarities between the coordination geometries in the neighborhood of different amino acids that permit reciprocal mutations without other significant changes?

Here, known protein structures will be analyzed using a virtual bond approximation in which each residue is represented by two interaction sites, one on the mainchain and the other on the sidechain. The sites on the mainchain are identified with the C^αs, whereas those on the sidechains are selected for each residue type on the basis of the specific structure and energy characteristics of the amino acid (I Bahar, RL Jernigan, unpublished data). An extension of the conventional representation of backbone geometry, in terms of bond torsion angles and bond angles, is applied to nonbonded neighbors (i.e. sidechain pairs separated by five or more residues along the backbone; Fig. 2). Mainly, three variables, r_{ij}, ϑ_{ij} and φ_{ij}, characterize the location of a sidechain S_j with respect to S_i. r_{ij} is the distance, ϑ_{ij} is the supplement of the angle between the virtual bond C^α_i–S_i and r_{ij}, and φ_{ij} is the torsional angle of the virtual bond C^α_i–S_i defined by the relative positions of the sites C^α_{i-1}, C^α_i, S_i and S_j. ϑ_{ij} and φ_{ij} will be referred to as the polar and azimuthal angles describing the coordination of a sidechain S_i by a sidechain S_j; the second subscript in the geometric variables will be omitted whenever the coordination of a given type of sidechain S_i by any other type of nonbonded sidechain is considered. Joint distributions of the geometric variables will be obtained for each residue type. Correlations between these variables are detailed for the first time. Overall, packing will be shown to possess some level of order that varies with the type of residue and cannot be viewed as a random combination of objects completely devoid of any structural and orientational preferences. However, the distributions will be shown to be relatively broad, in conformity with the tolerance of protein interiors to mutations, and with the adaptability of the overall fold to structural perturbations. This flexibility does not, however, rule out the existence of some statistically preferred loci for the coordinations of sidechains; in fact, some loci can exhibit as much as a 10-fold enhanced probability compared to that for a uniform distribution over the neighborhood of a given residue.

Figure 2



Schematic representation of the coordination between sidechain units S_i and S_j attached to the ith and jth C^αs respectively. r_{ij}, shown as a dashed line with an arrow at the end, is the separation vector pointing from S_i to S_j. The polar angle ϑ_{ij} represents the angle between r_{ij} and the extension of the sidechain bond vector l^s_i between C^α_i and S_i. φ_{ij} is the rotational angle about bond l^s_i defined by the relative positions of the four points C^α_{i-1}, C^α_i, S_i and S_j, assuming the value $\varphi_{ij} = 180°$ for the *trans* position. The three geometric variables r_{ij}, ϑ_{ij} and φ_{ij} characterize the coordination of S_i by S_j. The azimuthal angle space is divided into six states, *trans* (*t*), *skew*$^+$ (*s*$^+$), *gauche*$^+$ (*g*$^+$), *cis* (*c*), *gauche*$^-$ (*g*$^-$), and *skew*$^-$ (*s*$^-$), corresponding to successive intervals of width 60° in the range $0° \le \varphi_{ij} \le 360°$. The polar angles are divided into three successive intervals of width 60° in the full range $0° \le \vartheta_{ij} \le 180°$; these are referred to as the *front* (*f*), *lateral* (*l*), and *back* (*b*) positions.

Results and discussion

Coordination geometry around particular residues

In Figure 3, we show the distributions of azimuthal angles, $N(\varphi_i)$, obtained for all types of amino acids. Here, the results are compiled at 30° intervals for all types S_j of

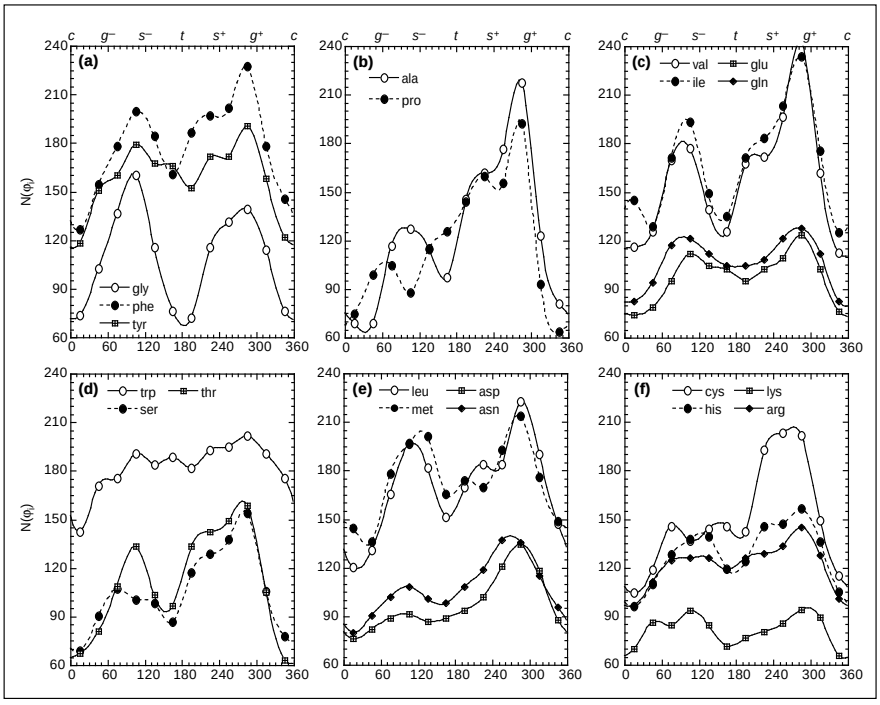
sidechains coordinating a given central residue of type S_i, i.e. $N(\varphi_i) = \Sigma_j N(\varphi_{ij})$, and rescaled to a common basis of 100 residues of each type, so that the absolute heights of the curves reflect the internal contact formation tendencies of the individual amino acids. The same ordinate scale is used in all figures to facilitate their comparison. We recall that $\varphi_i = 180°$ corresponds to the *trans* position of sidechain S_j with respect to bonds C^{α}_{i-1}–C^{α}_i and C^{α}_i–S_i and thus corresponds to a placement of S_j relatively close to residue i+1; and $\varphi_i = 0°$ or 360° is the *cis* position, which leads to interference with the residue i–1. These two regions are relatively inaccessible near S_i = glycine (as shown in Fig. 3a).

The relatively low occupancies of the region about $\varphi_i = 60°$, observed in some residues, originate in the steric interference of the backbone carbonyl and amide groups of the amino acids i and i+1, respectively. The access to the region near $\varphi_i = 0°$, on the other hand, is restricted by the backbone amide group $(NH)_i$. Examination of alanines, for example, reveals that the atoms O_i almost completely occupy the azimuthal angles $30° \le \varphi_i < 60°$ with respect to the sidechain S_i, their distance from the sidechain centroid (here C^{β}_i) being 3.3 Å, approximately. The amide nitrogens N_{i+1}, on the other hand, tend to occupy the range $60° \le \varphi_i < 90°$, their position also being remarkably close (3.1 Å) to C^{β}_i. It is not hard to visualize that this pair of atoms, on average, obstructs the accessibility of the region $\varphi_i \approx 60°$ near alanine. On the other hand, the backbone amide N_i also occupies a position very close to the side group ($r \approx 2.5$ Å, $\varphi_i \approx 15°$) invariably; this atom contributes to the exclusion of the region $0° \le \varphi_i < 15°$ for nonbonded neighbors approaching alanine. A similar interference of the three polar groups $(CO)_i$, $(NH)_i$, and $(NH)_{i+1}$ was verified to be operative in valine and isoleucine as well, whereas in leucine the positions of the three backbone atoms with respect to the sidechain show substantial variability, presumably because of the flexibility of the sidechain itself. Leucine may, therefore, accommodate a broad range of coordination angles.

Two other residues whose sidechain coordinations are influenced by backbone polar group interferences are serine and threonine (Fig 3d). The closest polar group to these sidechains is observed to be $(NH)_i$ most frequently. The separation between the atom N_i and the sidechain interaction centers of serine or threonine are generally closer than 3.3 Å. N_i is generally located at the *cis* position and therefore contributes to the exclusion of the region $-30° \le \varphi \le 30°$. In a few instances, the serine hydroxyl group is observed to form a hydrogen bond with $(CO)_{i-1}$. The curve for threonine lies slightly above that of serine, indicating that threonine experiences somewhat more nonbonded contacts.

The distribution of the polar angles ϑ_i have been analyzed in a similar way to that of the azimuthal angles above. Gaussian-like distributions are observed. The peak positions in the distribution curves are observed to depend on

Figure 3



Distribution of azimuthal angles $N(\varphi_i)$ for the coordination of any type of sidechain near ($2.0 \le r \le 6.8$ Å) a given sidechain S_i. Results are obtained at 30° intervals for **(a)** Gly, Phe and Tyr, **(b)** Ala and Pro, **(c)** Val, Ile, Glu and Gln, **(d)** Trp, Ser and Thr, **(e)** Leu, Met, Asp and Asn, **(f)** Cys, His, Lys and Arg. The biases arising from the differences in the natural occurrence of residues of different types are removed by expressing the results on the basis of 100 central residues of each type. The overall height of a given distribution curve thus reflects the number of internal contacts, or the extent of burial of the corresponding residue.

the types of residue S_i being coordinated. It shifts from about 75° (for S_i = Gly, Ala, Val, Ile, Cys, Pro, Ser, and Thr) to 90° (Phe, His, Asp, and Asn), to 95° (Leu, Met, Tyr, and Trp) or 110° (Glu, Gln, Lys, and Arg). The last case indicates that the relatively long polar and charged groups can more effectively penetrate within the regions of other residues and experience increased lateral interactions. The distributions are seen to be affected simply by the size and polarity of the virtual side bond, rather than by any specific energetics.

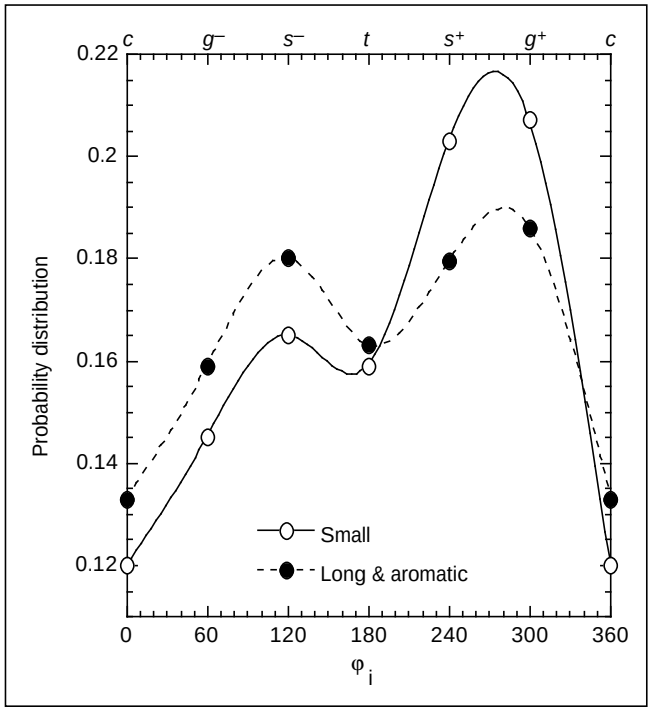
We note that at a given coordination site, the orientation of the sidechain S_j with respect to S_i is defined by the rotational state of the additional angle ζ_{ij} defined by the units C_i^α, S_i, S_j and C_j^α, in addition to the polar angles ϑ_{ij} and ϑ_{ji}. The distributions $P(\zeta_i) \equiv \Sigma_j P(\zeta_{ij})$ for polar and charged sidechains are observed to be relatively flat. Tyrosine, histidine and proline also exhibit a rather uniform distribution of χ_i values. Alanine, valine, isoleucine, leucine, methionine and tryptophan exhibit a preference for the *trans* state, indicating that these residues are often surrounded by residues pointing toward them; valine and isoleucine yield another peak near the *cis* state, presumably associated with their interaction in adjacent β-strands. Finally, two peaks at ± 120° rotations from the *trans* state are observed for cysteine, whose distribution is dominated by the geometry of disulfide bridges. The focus in the present study is on the identification of the most probable coordination loci in the neighborhood of different types of residues, which is defined by the two coordination angles ϑ_{ij} and φ_{ij} and consequently we will limit ourselves here to the detailed examination of these two variables.

A global view of coordination angle preferences

In order to acquire a more global view of the preferences of different types of sidechains, it is useful to divide the coordination space into a few broad regions. Conforming with the conventional classification of dihedral angles, the azimuthal angles may be described in terms of six states: *cis* (*c*, 0°), *trans* (*t*, 180°), *gauche*$^+$ (g^+, 300°), *gauche*$^-$ (g^-, 60°), *skew*$^+$ (s^+, 120°), and *skew*$^-$ (s^-, 240°), each of them covering a width of 60°. The polar angles, on the other hand, are classified as *front* (*f*), *lateral* (*l*), and *back* (*b*) positions of 60° intervals each, centered about the respective values 30°, 90°, and 150°. (See Fig. 2.)

Examination of the distributions in terms of these coordination states reveals the distinct behavior of two groups of residues, classified according to their size and shape characteristics. The first group includes the aromatic residues tryptophan, tyrosine, histidine, and phenylalanine, as well as those containing three or more rotatable bonds on the sidechain: lysine, arginine, glutamic acid, glutamine, and methionine. The second group contains only small size residues having either hydrophobic or polar sidechains: glycine, alanine, valine, isoleucine, leucine, serine, threonine, aspartic acid, and asparagine. Proline may also be included in this group, its azimuthal angle distribution resembling that of alanine (Fig. 3b). These will be referred to as long/aromatic sidechains and small sidechains. The azimuthal angle probability distributions obtained for the two groups are presented in Figure 4. Results are compiled for the six coordination states indicated on the upper abscissa scale and normalized in each group. Both curves exhibit a bimodal shape. However, the most probable coordination states are different for the two groups: small sidechains exhibit a strong preference for azimuthal angles, centered about a peak at φ_i = 270°, reflected by an enhancement in s^+ and g^+ coordination angles. Long and aromatic sidechains, on the other hand, have two nearly equal peaks on either side of 180°, the most likely coordination states being s^-, s^+, and g^+ in this case. The enhancement to the larger azimuthal angles for the group of small sidechains is observed to correlate partly with the formation of favorable backbone–backbone electrostatic interactions between the polar groups of the polypeptide segments near S_i and S_j. Long and aromatic sidechains, on the other hand, are more symmetric in this behavior because they are more decoupled from the interference of the backbone.

Figure 4



Probability distribution $P(\varphi_i)$ of azimuthal angles for two groups of residues. The group 'long and aromatic' refers to sidechains Arg, Lys, Glu, Gln, Trp, Tyr, His, Phe and Met. The group 'small' includes all the remaining residue types.

Residue-specific coordination states

Probability distributions $P(\vartheta_i, \varphi_i)$ for the joint occurrence of pairs of coordination angles (ϑ_i, φ_i) have been obtained for all types of sidechains S_i, on the basis of all nonbonded sidechains S_j located within the first coordination shell ($r \leq 6.8$ Å). The distributions are found to be relatively broad. A wide range of azimuthal angles is accessible in general, while the polar angles are generally confined to lateral positions. 2–4 maxima are discernible in each contour map obtained as a function of the variables ϑ_i and φ_i, which define the thermodynamically stable residue-specific coordination states in the neighborhood of the sidechain S_i. Also, some highly populated regions may be identified for each S_i. These may be viewed as additional possible sites for the association of nonbonded neighbors, considering the fact that completely coordinated residues can have as many as eight or more sidechains within $r \leq 6.8$ Å ([16,17]; I Bahar, RL Jernigan, unpublished data).

We have developed a list of the most probable coordination sites $\nu_i \equiv (\vartheta_i, \varphi_i)$ for each residue type. These constitute a set of potential sites for occupancy by nonbonded neighbors and thus may provide a guidance in computer simulations for selecting the loci of nonbonded sidechains in a given packing unit. Table 1 lists the seven most probable residue-specific coordination loci, ν_i^A–ν_i^G, for each sidechain S_i, along with the corresponding statistical weights. ω_i^X represents the probability of occupancy of the selected loci X of size $\Delta\vartheta_i = 60°$ and $\Delta\varphi_i = 60°$. Two or three states in each row represent local maxima, while the remainder are high-density regions. The last column gives the sum of the statistical weights of the seven loci for each residue type. This also represents the fraction of coordination states near S_i whose geometry conforms with one of the given residue-specific states. On average, 71% of the observed pairs are described with these seven residue-specific coordination states.

Coupling between φ_i, ϑ_i and r_i

The probability distributions $P(\vartheta_i, \varphi_i)$ have been tabulated for two distance ranges, $2.0 \leq r \leq 4.4$ Å and $4.4 < r \leq 6.8$ Å. These ranges were shown recently to be dominated by different types of interactions, mainly those of hydrophilic and hydrophobic nature, respectively (I Bahar, RL Jernigan, unpublished data). The average behavior over all sidechain pairs is shown in the contour maps of Figure 5a,a′. Parts (a) and (a′) refer to the broad and close distance ranges $2.0 \leq r \leq 6.8$ Å and $2.0 \leq r \leq 4.4$ Å, respectively. Two maxima are observed in part (a), encircled by the innermost contours. These are located at *gauche*$^+$ and *skew*$^-$ states, the former being more pronounced. These reflect somewhat the global preferences displayed in Figure 3. The state *gauche*$^+$

Table 1

Most probable coordination states ν_i and their statistical weights ω_i for all amino acids* ($r_{ij} \leq 6.8$ Å).

S_i	N_i†	ν_i^A	ω_i^A	ν_i^B	ω_i^B	ν_i^C	ω_i^C	ν_i^D	ω_i^D	ν_i^E	ω_i^E	ν_i^F	ω_i^F	ν_i^G	ω_i^G	$\Sigma\omega_i$‡
Gly	13086	ls^-	0.13	lg^+	0.11	ls^+	0.10	lg^-	0.10	fc	0.08	fg^+	0.07	fg^-	0.07	0.65
Ala	15509	lg^+	0.15	ls^+	0.13	ls^-	0.11	lt	0.10	fg^+	0.07	fc	0.07	fs^-	0.07	0.70
Val	17902	lg^+	0.15	ls^+	0.12	ls^-	0.11	lt	0.09	lg^-	0.09	fs^+	0.07	fg^+	0.07	0.71
Ile	15341	lg^+	0.14	ls^+	0.12	ls^-	0.12	lt	0.10	lg^-	0.09	fc	0.08	lc	0.07	0.72
Leu	23496	lg^+	0.13	ls^-	0.13	ls^+	0.11	lt	0.10	lg^-	0.09	lc	0.08	fs^-	0.07	0.71
Ser	9383	lg^+	0.14	ls^+	0.13	ls^-	0.10	lt	0.10	lg^-	0.09	lc	0.07	fg^-	0.06	0.70
Thr	9850	ls^+	0.14	lg^+	0.13	ls^-	0.12	lt	0.11	lg^-	0.08	lc	0.06	fg^+	0.06	0.70
Asp	7477	lg^+	0.15	ls^+	0.11	ls^-	0.11	lt	0.10	lc	0.10	lg^-	0.09	fg^+	0.06	0.72
Asn	6457	lg^+	0.14	ls^+	0.12	ls^-	0.11	lt	0.11	lg^-	0.09	lc	0.09	fc	0.06	0.71
Glu	7183	ls^-	0.13	lt	0.12	lg^+	0.12	lg^-	0.11	ls^+	0.10	lc	0.09	fg^-	0.05	0.72
Gln	5093	ls^-	0.13	lg^+	0.12	lt	0.11	ls^+	0.11	lg^-	0.11	lc	0.09	fg^-	0.06	0.74
Lys	6518	lg^-	0.14	ls^-	0.12	lg^+	0.11	lt	0.11	ls^+	0.10	lc	0.09	bg^+	0.06	0.74
Arg	6761	ls^-	0.12	ls^+	0.12	lt	0.11	lg^-	0.11	lg^+	0.11	lc	0.09	bg^+	0.05	0.70
Cys	5104	ls^+	0.15	lg^+	0.13	ls^-	0.10	lt	0.10	lg^-	0.09	lc	0.07	fg^-	0.07	0.72
Met	5507	ls^-	0.13	lg^+	0.12	ls^+	0.11	lt	0.10	lg^-	0.09	lc	0.08	fc	0.07	0.69
Phe	12103	lg^+	0.13	ls^-	0.12	ls^+	0.11	lt	0.11	lg^-	0.09	lc	0.08	fs^+	0.07	0.70
Tyr	9425	lg^+	0.13	ls^-	0.12	ls^+	0.12	lt	0.11	lg^-	0.11	lc	0.08	fs^-	0.06	0.73
Trp	4261	lg^+	0.13	ls^+	0.12	ls^-	0.11	lt	0.11	lg^-	0.10	lc	0.09	fs^-	0.06	0.73
His	4421	lg^+	0.13	ls^-	0.12	ls^+	0.12	lg^-	0.11	lt	0.11	lc	0.08	fs^-	0.06	0.72
Pro	7378	ls^+	0.13	lt	0.12	lg^+	0.11	ls^-	0.10	lg^-	0.08	bg^+	0.07	fc	0.06	0.67

*The notation $\nu_i = (\vartheta_i, \varphi_i)$ is used to describe the coordination states. †Total number of pairs coordinating S_i within $r_{ij} \leq 6.8$ Å, observed in the 302 PDB structures. ‡Sum of statistical weights of the listed seven residue-specific coordination loci for each S_i.

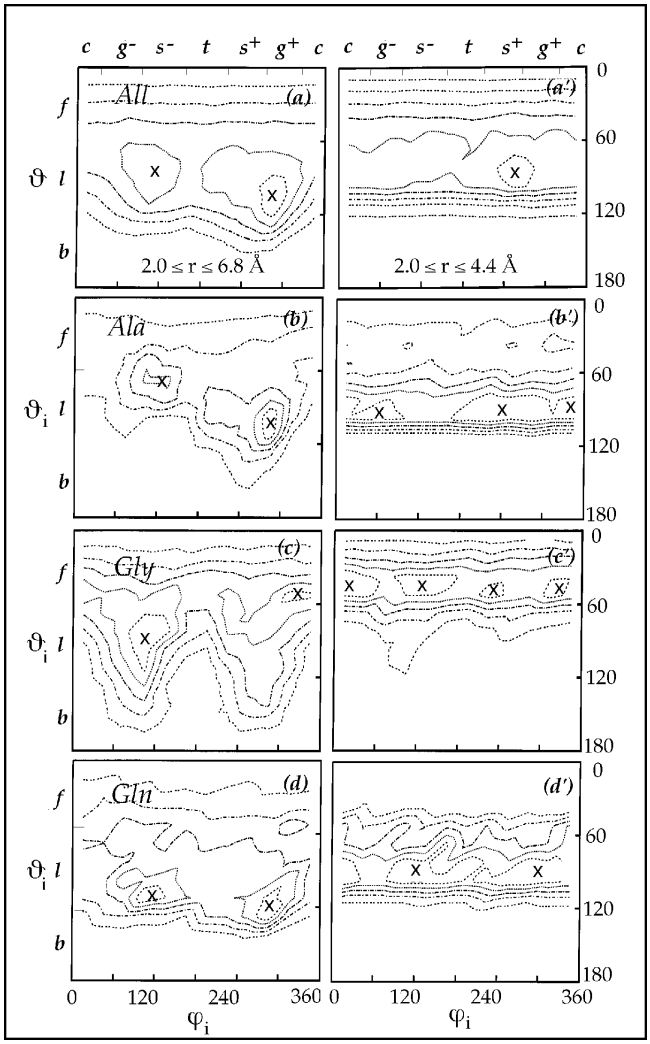
is preferred at the close distance range, as well, though it is slightly shifted towards the *skew*$^+$ region, whereas the loci near the *skew*$^-$ state are hardly distinguishable. Insofar as the polar angles are concerned, most pairs are concentrated in a band, which is rendered narrower and shifted to smaller values ($60 \leq \vartheta \leq 100°$, approximately) for the close distance regime. The polar angle range $\vartheta \leq 120°$ is completely excluded in part (a′).

In the contour maps (Fig. 5a,a′), specific effects are diminished because these represent the average behavior of all residues. The maps presented for alanine, glycine, and glutamine (Fig. 5b–d,b′–d′) illustrate the distinct behavior of these particular residues over the two distance ranges. In general, the changes in the coordination loci as one focuses on the close distance range are two: the shrinkage of the accessible polar angle range so as to favor front placements (smaller ϑ values) of nonbonded neighbors, and a perturbation in the most probable azimuthal angles.

One can identify in the contour maps of Figure 5a–d the residue-specific coordination sites that have been compiled in Table 1 for the broad distance range. By analogy, the six most probable coordination states for the close neighbors have been determined using maps of type (a′–d′) for all residues; these are presented in Table 2. The coordination types in the close distance range are more selective (or interactions are more specific) compared to those in the broad distance range, as may be verified from the larger weights ω_{ic} associated with some individual tabulated coordination states. It is interesting to note from the cumulative weights listed in the last column that the aromatic sidechains tryptophan, phenylalanine and tyrosine exhibit the highest selectivity in the close distance regime. This may be attributed to their large sizes and anisotropic shapes, which substantially restrict the coordination space and are also responsible for the relatively small number N_{ic} of occurrences reported in the second column.

It should be noted that sidechain pairs separated by at least four intervening residues have been included in the above distributions so as to eliminate the biases due to chain connectivity and secondary structure preferences. We note that in α-helices the residue S_j located on the succeeding turn with respect to S_i (j – i = 3 or 4) occupies the *gauche*$^-$ region of the azimuthal angle distributions, whereas those belonging to β-strands are located near the *trans* (j = i + 2) or *cis* (j = i – 2) positions. The most probable azimuthal angles taken up by the atoms O_i for different types of central sidechains S_i are *cis* and *gauche*$^-$ regions for all residues, and additionally *gauche*$^+$ for leucine and methionine, and *skew*$^-$ for phenylalanine and cysteine. For N_{i+1}, they are *gauche*$^-$ for all residues, *gauche*$^+$ for all residues except alanine, serine, threonine and aspartic acid, *cis* for aspartic acid, asparagine, phenylalanine, tyrosine and histidine, and *skew*$^+$ for aspartic acid and asparagine. Clearly, the analysis is complicated upon examination of the effect of individual atoms, and Tables 1 and 2 aim at providing a list of the regions left accessible at the end of the rather intricate interference of nearest bonded backbone or sidechain atoms.

Figure 5



Probability distributions of coordination angles ϑ and φ in the neighborhood of sidechains of different types. Contour maps **(a)** and **(a′)** represent the average behavior of all sidechains obtained for the respective distance ranges of $2.0 \leq r \leq 6.8$ Å and $2.0 \leq r \leq 4.4$ Å. Innermost contours refer to highest density regions and define the most stable coordination states. The peaks in the probability distributions are indicated by the symbol 'x'. Two maxima are observed in part **(a)** at *ls*$^-$ and *lg*$^+$ states. At the close distance range **(a′)**, the former is hardly distinguishable, while the latter is shifted to an *ls*$^+$ state. The range of accessible polar angles becomes narrower and shifts toward front positions in the close distance range. **(b–d)** Represent the same type of diagram obtained for Ala, Gly and Gln, respectively, in the broad distance range $2.0 \leq r \leq 6.8$ Å. Their close distance counterparts are presented in the maps **(b′–d′)**. See Table 3 for the list of the most probable residue-specific coordination loci for each type of sidechain, for the two distance ranges.

Examination of the association of specific pairs of sidechains

When specific pairs of sidechains are considered, the peaks in the probability distributions $P(\vartheta_{ij}, \varphi_{ij})$ become more pronounced, as expected, compared to those, $P(\vartheta_i, \varphi_i)$, averaged over all sidechains S_j near S_i. For illustrative purposes, a few pairs are presented in Figure 6a–f: Lys–Glu, Leu–Leu, Lys–Phe, Glu–Met, Thr–Thr, and Thr–Val. Here, the actual numbers $N(\vartheta_{ij}, \varphi_{ij})$ of observations of the various coordination geometries $(\vartheta_{ij}, \varphi_{ij})$ are shown in the form of three-dimensional plots. These are extracted from the complete set of 302 PDB structures, considering the broad distance range $2.0 \leq r \leq 6.8$ Å (Fig. 6a–d) and the close distance range $2.0 \leq r \leq 4.4$ Å (Fig. 6e,f). The projections of the distribution surfaces on the $(\vartheta_{ij}, \varphi_{ij})$ plane are shown in the form of five equally spaced contours. This permits the visualization of the most favorable regions, which are enclosed by the innermost contours.

The coordination preferences of small polar/charged (P) sidechains are generally induced by both the polarity of the sidechains and the constraints imposed by backbone connectivity and may be highly specific. For example, the most probable coordination state of the pair of sidechains Asp–Asn is observed to be *lg*$^+$. This state is enhanced by a factor of nine approximately relative to a random distribution of coordination states. The preference for specific coordination sites of P–P pairs is, however, considerably weakened with increasing size of the sidechains. Lys–Glu pairs, for example, exhibit a more diffuse distribution of coordination angles (as may be seen from Fig. 6a) due to the rotational flexibility of the sidechain bonds in these amino acids. The sidechains of lysine and glutamic acid have four and three rotatable bonds, respectively, which provide enough conformational freedom to accommodate a broad range of coordination angles.

The pair Leu–Leu (Fig. 6b) represents another case of a rather uniform sampling of a wide range of coordination angles. This behavior is explained by the adaptability to fit various coordination geometries with its branched, yet relatively small, structure. Leucine, being branched at C^γ, possesses a wide surface some distance from the backbone which can be coordinated from all sides, while for valine, for example, the C^γs are too close to the backbone to be as well coordinated. The number of Leu–Leu pairs in the distance range $2.0 \leq r \leq 6.8$ Å is strikingly large in the databank structures, which, together with its adaptability to various coordination geometries, suggests the considerable contribution of leucines to the stability of hydrophobic cores.

Lys–Phe pairs (Fig. 6c) illustrate amino–aromatic interactions. The polar angle is severely restricted here to the range $\vartheta_{ij} = 110° \pm 20°$, i.e. *lateral* and *back* positions with

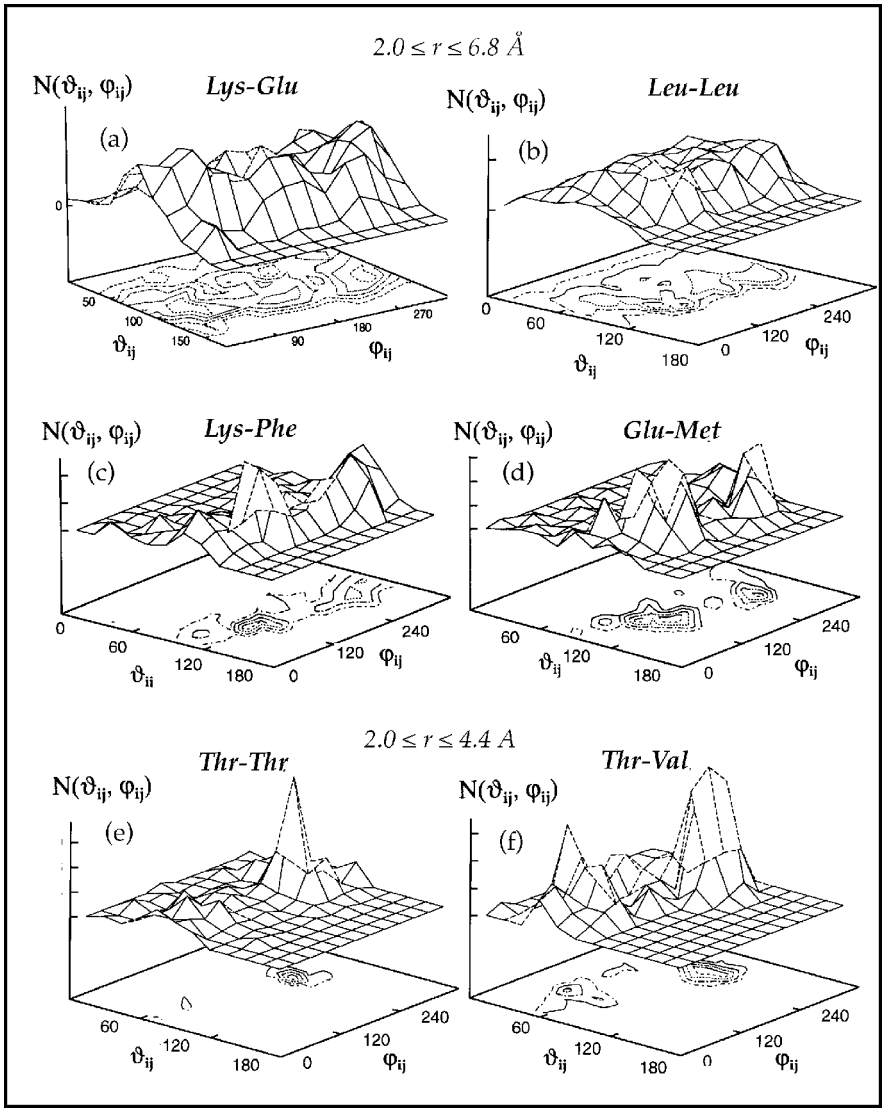
Table 2

Most probable coordination states* of close ($r \leq 4.4$ Å) neighbors of S_i, and their statistical weights.

S_i	N_{ic}	ν_{ic}^A	ω_{ic}^A	ν_{ic}^B	ω_{ic}^B	ν_{ic}^C	ω_{ic}^C	ν_{ic}^D	ω_{ic}^D	ν_{ic}^E	ω_{ic}^E	ν_{ic}^F	ω_{ic}^F	$\Sigma\omega_{ic}^X$
Gly	2459	fs$^-$	0.14	fc	0.14	fg$^+$	0.11	ls$^-$	0.11	fs$^+$	0.10	ls$^+$	0.09	0.68
Ala	3401	ls$^+$	0.15	lg$^+$	0.15	fs$^-$	0.11	fg$^+$	0.10	ls$^-$	0.10	lt	0.09	0.70
Val	2343	ls$^+$	0.14	fs$^+$	0.14	fg$^-$	0.11	fg$^+$	0.11	lg$^+$	0.10	fs$^-$	0.10	0.70
Ile	1874	ls$^+$	0.16	fs$^+$	0.15	ft	0.13	ls$^-$	0.13	fs$^-$	0.12	lg$^-$	0.07	0.76
Leu	2531	fs$^+$	0.14	fg$^+$	0.13	lt	0.13	ft	0.12	lc	0.12	lg$^+$	0.11	0.75
Ser	2148	lg$^+$	0.17	ls$^+$	0.16	lg$^-$	0.10	lt	0.09	fs$^-$	0.09	lc	0.08	0.68
Thr	1979	lg$^+$	0.16	ls$^+$	0.14	fs$^-$	0.11	fs$^+$	0.09	lg$^-$	0.09	fg$^+$	0.09	0.69
Asp	1778	lg$^+$	0.16	fg$^+$	0.12	ls$^+$	0.11	fs$^+$	0.10	lc	0.10	lt	0.09	0.69
Asn	1297	lg$^+$	0.17	ls$^+$	0.12	lt	0.10	lc	0.10	lg$^-$	0.09	fg$^+$	0.09	0.68
Glu	1667	lt	0.14	lg$^+$	0.13	ls$^-$	0.11	lg$^-$	0.10	ls$^+$	0.10	fg$^-$	0.09	0.67
Gln	974	lg$^+$	0.14	ls$^+$	0.13	lt	0.13	ls$^-$	0.12	lg$^-$	0.11	lc	0.09	0.72
Lys	1353	lg$^+$	0.15	lt	0.14	lg$^-$	0.14	ls$^-$	0.11	ls$^+$	0.11	lc	0.10	0.75
Arg	1594	ls$^+$	0.14	ls$^-$	0.13	lt	0.12	lg-	0.11	lg$^+$	0.11	lc	0.10	0.71
Cys	1439	ls$^+$	0.27	lg$^+$	0.16	ls$^-$	0.08	lg$^-$	0.08	lc	0.07	fs$^-$	0.07	0.73
Met	765	lg$^+$	0.14	ls$^+$	0.14	ls$^-$	0.10	lc	0.09	fs$^-$	0.09	lt	0.09	0.64
Phe	1042	lg$^+$	0.21	ls$^-$	0.17	ls$^+$	0.14	lt	0.13	lg$^-$	0.13	lc	0.11	0.89
Tyr	1045	lg$^+$	0.18	ls$^-$	0.16	lt	0.13	lg$^-$	0.12	ls$^+$	0.10	lc	0.10	0.79
Trp	475	lg$^+$	0.20	ls$^-$	0.19	ls$^+$	0.16	lt	0.14	lg$^-$	0.13	lc	0.12	0.94
His	777	lg$^+$	0.16	ls$^+$	0.15	lt	0.12	ls$^-$	0.11	lg$^-$	0.09	fs$^-$	0.07	0.71
Pro	923	lg$^+$	0.20	ls$^+$	0.17	fg$^+$	0.10	ls$^-$	0.10	fs$^+$	0.09	lt	0.08	0.74

*The notation is identical to that of Table 1, except for the use of the subscript c indicating close distance coordination.

Figure 6



Probability distributions of the coordination angles $P(\vartheta_{ij}, \varphi_{ij})$ for some specific pairs S_i–S_j of sidechains. (a) Lys–Glu, (b) Leu–Leu, (c) Lys–Phe, (d) Glu–Met, (e) Thr–Thr, and (f) Thr–Val. Distributions in parts (a–d) are obtained for the broad distance range $2.0 \le r \le 6.8$ Å; parts (e) and (f) refer to the close distance association for the pairs Thr–Thr and Thr–Val.

respect to the lysine sidechain are accessible only. The azimuthal angle, on the other hand, exhibits two preferred conformations centered about the *skew*$^-$ and *gauche*$^+$ states, leading to the most probable joint states *ls*$^-$ and *bg*$^+$. We note that a large number of amino–aromatic pairs were observed in the present study to be highly selective in parallel with Lys–Phe, confirming previous suggestions on the contribution of this type of interaction in stabilizing protein folds. Interestingly, a pair of H–P sidechains, as Glu–Met (Fig. 6d), may also exhibit strong preferences for specific coordination loci. The state *ls*$^-$ is enhanced here by a factor of about nine compared to a uniform distribution.

Figure 6e,f demonstrates the generally increased specificity of sidechain pairs in the close distance regime. The states *lg*$^+$ and *fs*$^-$ are the most likely coordination geometries of the pair Thr–Val, while a single sharp peak is observed for Thr–Thr pairs. In the latter case, the polarity of the sidechains, the interference of the polypeptide backbone, and the bulk of threonine sidechains restrict severely the coordination states accessible in the close distance range and lead to a strong preference for the state *ls*$^+$. The distinct distributions (Fig. 6a,e) signal the inadequacy of combining all hydrophilic residues into one unified group for coarse-grained simulations, as also pointed out in our recent review of residue–residue potentials [22].

The most probable coordination states have been sorted out and presented in Table 3 for the two distance ranges $4.4 < r \le 6.8$ Å and $2.0 \le r \le 4.4$ Å. The types S_i and S_j of the sidechain pairs exhibiting the most selective coordination geometry are listed, along with their most probable

coordination states $\nu_{ij}^{x} \equiv (\vartheta_{ij}^{x}, \varphi_{ij}^{x})$. The probability $p(\nu_{ij}^{x}|r \pm \Delta r)$ of that state relative to that expected for a uniform distribution ($p^{\circ} = 1/18$) is also presented; this is designated as p_{ij}/p°. It is useful for estimating an empirical free energy of stabilization $\Delta W(\nu_{ij}^{x}|r \pm \Delta r)$ for the coordination state 'x' of sidechains S_i and S_j located at $r \pm \Delta r$, using $\Delta W(\nu_{ij}^{x}|r \pm \Delta r) \equiv -RT \ln p(\nu_{ij}^{x}|r \pm \Delta r)/p^{\circ}$. $\Delta W(\nu_{ij}^{x}|r \pm \Delta r)$ values of all coordination states for all pairs are available as Supplementary material (published with this paper on the internet) for the three distance ranges. Here we give a summary of the pairs exhibiting the strongest attractions at specific loci. Precisely, the states whose frequency is enhanced by a factor of $p_{ij}/p^{\circ} \geq 6$ relative to that of a random coordination are listed for $2.0 < r \leq 6.8$ Å. We note that the corresponding stabilization energies are $\Delta W(\nu_{ij}^{x}|r \pm \Delta r) \leq -1.8$ RT. At shorter distances ($2.0 \leq r \leq 4.4$ Å) the fraction of coordination states satisfying the same stability criterion increases significantly due to the increased speci-

Table 3

Most stable coordination states* for specific pairs of sidechains.

$2.0 < r_{ij} \leq 6.8$ Å					$2.0 \leq r_{ij} \leq 4.4$ Å				
S_i	S_j	$N_{ij}(r)^{\dagger}$	ν_{ij}^{x}	$p_{ij}^{x}/p^{\circ\ddagger}$	S_i	S_j	$N_{ij}(r)^{\dagger}$	ν_{ij}^{x}	$p_{ij}^{x}/p^{\circ\ddagger}$
Cys	Cys	1028	ls^{+}	10.3	Cys	Cys	735	ls^{+}	13.0
Trp	Cys	107	lg^{+}	10.1	Thr	Val	125	lg^{+}	10.4
Glu	Met	140	ls^{-}	9.0	Asp	Asn	95	lg^{+}	9.3
Glu	Trp	138	ls^{-}	8.1	Ala	Phe	171	ls^{+}	8.9
Glu	Phe	278	ls^{-}	7.3	Thr	Thr	212	ls^{+}	8.8
His	Pro	173	ls^{+}	7.2	Asp	Ala	110	lt	8.1
Gln	Leu	457	ls^{-}	7.2	Tyr	Gly	121	lg^{-}	8.0
Pro	Phe	387	ls^{+}	6.9	Thr	Ser	164	ls^{+}	7.7
His	Leu	371	ls^{-}	6.8	Asp	His	95	lg^{+}	7.6
His	Trp	105	ls^{-}	6.8	Phe	Ala	171	lg^{+}	7.6
Arg	Phe	268	ls^{-}	6.8	Phe	Leu	150	ls^{-}	7.5
Tyr	Cys	211	lg^{-}	6.8	Asn	Thr	116	lg^{+}	7.7
Glu	Leu	490	ls^{-}	6.7	Ser	Arg	102	lg^{+}	7.4
Gln	Ile	276	ls^{-}	6.6	Ser	Asp	160	ls^{+}	7.4
Tyr	Lys	363	lg^{+}	6.5	Gly	Val	128	fc	7.3
His	Val	262	lg^{+}	6.4	Ile	Leu	292	fs^{+}	7.2
Pro	Asp	273	ls^{+}	6.4	Phe	Val	144	lg^{+}	7.2
Gln	Trp	101	ls^{-}	6.4	Asp	His	95	fs^{+}	7.2
His	Cys	113	lg^{-}	6.3	Tyr	Ala	130	lg^{+}	7.2
Thr	Phe	465	ls^{-}	6.3	Asn	Gly	135	ls^{+}	7.2
Cys	Asn	125	lt	6.3	Thr	Asp	135	lc	7.2
Ala	Pro	550	ls^{+}	6.3	Gly	Leu	158	fc	7.0
Asn	Val	380	lg^{+}	6.3	Gly	Val	128	ls^{-}	7.0
Asn	Lys	348	lg^{+}	6.2	Thr	Leu	113	fs^{-}	7.0
Cys	Tyr	211	lg^{+}	6.3					
Asp	Arg	685	lg^{+}	6.2	Asn	Lys	79	lg^{+}	10.9
Asp	Tyr	427	ls^{-}	6.2	Leu	Gln	75	lt	10.5
Glu	Val	411	ls^{-}	6.2	Trp	Leu	67	lg^{+}	10.8
Tyr	Met	298	ls^{-}	6.2	Trp	Leu	67	ls^{-}	11.8
Glu	Tyr	421	ls^{-}	6.2	Val	Trp	52	ls^{+}	11.8
Leu	Arg	485	lg^{+}	6.2	Pro	Phe	51	ls^{+}	12.0
Thr	Arg	407	lg^{+}	6.0	Trp	Ile	43	ls^{-}	10.0
Ser	Ile	548	lg^{+}	6.0	Arg	Phe	40	ls^{-}	12.6

*Coordination states (ϑ_{ij}, φ_{ij}) are indicated as ν_{ij}^{x}. †$N_{ij}(r)$ is the actual number of observations of the sidechain pair (S_i, S_j) in the indicated distance range for the set of 302 PDB structures. ‡p_{ij}^{x}/p° is the ratio of observed frequency to that of a uniform distribution over all coordination angles. Most probable states refer to $p_{ij}^{x}/p^{\circ} \geq 6$ for the broad distance range, and ≥ 7 for the close distance range. Only pairs with $N_{ij}(r) \geq 95$ have been included, except for a few cases in the close distance range distinguished by their $p_{ij}^{x}/p^{\circ} \geq 10$.

ficity at closer separations, whereas the absolute numbers of observations of particular pairs, $N_{ij}(r \pm \Delta r)$, are reduced because of the smaller volume of coordination. Results tabulated for this distance range have been limited to those having $p_{ij}/p^\circ \geq 7$ and $N_{ij}(r \pm \Delta r) \geq 95$, except for a few pairs which are distinguished by their high enhancement factors ($p_{ij}/p^\circ \geq 10$). Among them we note that the states *l*g^+ and *l*s^- account, alone, for the coordination of more than 63% of observed Trp–Leu pairs.

The pair exhibiting the strongest preference for a specific coordination geometry in the close distance is Cys–Cys whose specificity arises from disulfide bridge formation. This is followed by Thr–Val and Asp–Asn, the most probable coordination state being *l*g^+ in both cases. This directionality for Thr–Val is stabilized by an excess free energy of –2.3 RT with respect to a random association of the same sidechains at the same distance range. The most selective coordination geometry in the large distance range, on the other hand, is exhibited by Cys–Cys and Trp–Cys pairs. A noticeable feature in this distance range is that a large fraction of pairs involve either a polar or an aromatic sidechain, demonstrating the importance of size and polarity effects in determining the preferred coordination geometry. Residues that participate most likely in the most selective coordination states in the distance range $4.4 \leq r \leq 6.8$ Å are glutamic acid and tyrosine, apart from cysteine. These are followed by histidine and proline. In the close distance regime, threonine and aspartic acid are distinguished by their high frequency of selective coordination.

Conclusions

Major observations and relevance to experiments

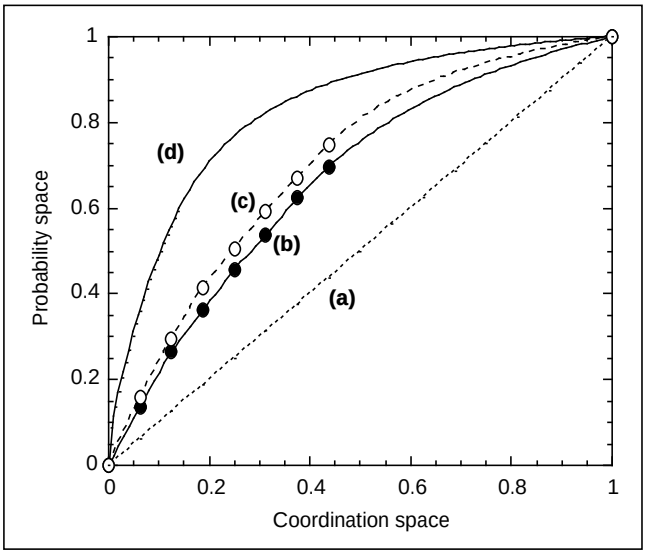
There exist some preferred coordination loci in the neighborhood of each type of sidechain

The distributions of the variables r_i, ϑ_i, and φ_i, which define the coordination geometry of all sidechains in the neighborhood of a given type of sidechain S_i, indicate that there are some well defined regions of the coordination space that are visited more often than others. The distributions are not random, in contrast to the implications of the nuts-and-bolts model, but biased towards some residue-specific coordination loci, which are compiled in Tables 1 and 2 for two (broad and close) distance ranges. An estimate of the departure from random association may be made by considering the curves in Figure 7. Here, the fraction of the coordination space is plotted against the occupancy probability. The dotted diagonal line, labeled as (a), corresponds to random association. Curves (b) and (c) refer to the coordination of sidechains S_i = glycine and glutamine, respectively, with all other sidechains, in the range $2.0 \leq r_{ij} \leq 6.8$ Å. Here, the ordinate is determined from the probabilities of the coordination states listed in Table 1 for these residues; each state covers 1/18 of the full conformational space spanned by (ϑ_i, φ_i), neglecting the normalization factor $\sin\vartheta_i$; the cumulative occupancy of the conformational space is given by the abscissa. Curve (d) illustrates the behavior of the specific pairs Thr–Thr in the close distance range. Here a finer division (12 × 12) of the coordination angles has been adopted to increase the accuracy. It is interesting to note that now 10% of the conformational space accounts for the coordination of half of the Thr–Thr pairs observed in the dataset. The departure of the curves (b–d) from the diagonal is strong evidence of the nonrandomness of the packing of sidechains.

Packing characteristics are residue specific

Figures 3 and 5 demonstrate that packing geometry near a given sidechain S_i varies with the type of sidechain. The azimuthal angle distributions have two sharp peaks for glycine, serine, and threonine; whereas histidine, arginine, leucine, methionine, etc. have more complex distributions (Fig. 3). The specificity of sidechain packing characteristics becomes more apparent from examination of the joint probability distributions for the coordination angles ϑ_i and φ_i which characterize those sidechains surrounding S_i. Comparison of the maps for alanine, glycine and glutamine in Figure 5 with the mean distribution of all residues shown in parts (a) and (a′) gives an indication of the variability of the probable coordination space with the residue type. Size and shape effects, in addition to polarity, play a major role in determining the coordination loci.

Figure 7



Probability space covered by the most probable coordination states shown against the fraction of the conformational space for **(a)** random association of sidechains, **(b)** association of all sidechains with S_i = Gly in the distance range $2.0 \leq r \leq 6.8$ Å, **(c)** association of all sidechains with S_i = Gln, for the same distances, and **(d)** Thr–Thr pairs in the close distance range $2.0 \leq r \leq 4.4$ Å. Departure from the dashed line (a) gives a direct measure of the specificity of interactions.

Three or more equally probable coordination states are identifiable in the neighborhood of each sidechain

The preference for particular coordination loci is not so pronounced as to preclude the adaptability of sidechain associations to local structural perturbations, in conformity with the experimental observations of tolerance to mutations [8–11]. Also, it is possible to perceive some common coordination states for all residues, provided that the definition of these states is made on a more flexible, broader basis. We note in particular that Leu–Leu contacts are the most frequent among all types of S_i–S_j contacts. The variability of this association over a wide range of coordination geometry is also remarkable (Fig. 6b), which conforms with experimental observations on the role of leucine clusters in maintaining conformational flexibility [18]. Thus, a compromise is achieved in native structures between (i) a sufficiently high packing density and nonrandom sidechain coordination and (ii) a certain degree of free volume or internal flexibility to optimize local interactions. The apparently ductile association of sidechains may, in fact, be a prerequisite for a specific function of the protein and may have been selected by nature so as to meet the requirements of specific biological activity, as suggested by several recent experiments [18–21].

Some general patterns may be discerned that dominate the association of all sidechains

One general feature emerging from the comparison of the maps obtained for the close and long distance regimes is the tendency to select smaller polar angles (*front* positions) as the interresidue distance becomes shorter. The azimuthal angles, on the other hand, exhibit two most probable coordination geometries. The first varies between the *skew*$^+$ and *gauche*$^+$ states, depending on the residue type, whereas the second is near the *skew*$^-$ state. The former is preferred by small sidechains, while the two states are almost equally probable in the case of large or aromatic sidechains. The regions *cis* and *trans* are generally disfavored by the steric interference of the backbone portions of the residue being coordinated. In the close distance regime, the coordination states are condensed to the region near the state *skew*$^+$.

Specificity becomes more pronounced when the coordination of particular pairs of amino acids is in question and is further enhanced in the close distance regime

The probability distribution surfaces displayed for various pairs of sidechains illustrate the specificity of sidechain pair coordination geometry. Well defined (ϑ_{ij}, φ_{ij}) peaks are observed that precisely characterize the coordination of the specific pairs. Size and shape are major factors affecting the choice of particular coordination loci. Large and anisotropic sidechains, such as the aromatic ones, are highly selective. Smaller amino acids also exhibit some preferred coordination geometries. Longer sidechains with two or more rotatable bonds are, on the contrary, decoupled from the backbone and may accommodate various coordination geometry so as to optimize specific interactions. This may result in a rather diffuse distribution of coordination geometries in spite of their polarities, unless the close distance regime is considered. A systematic analysis of the 20 × 20 pairs of sidechain associations yielded several coordination states, some of them listed in Table 3, with a stabilization energy ΔW_{ij} of about –2.0 RT in excess of that of random association at the same distance range.

Future prospects

Potential use of residue-specific coordination states in inverse protein folding calculations

Knowledge of residue-specific coordination loci is clearly useful for reducing conformational space and facilitates the search for stable folds in computer simulations. The loci presented in Tables 1 and 2 may serve as a guideline for selecting sidechain positions in packing units. For example, we recall that in the inverse protein folding calculations of Ponder and Richards [26], a few sequences only were shown to satisfy, without alteration of the overall structure, the steric restrictions and density requirements imposed on a given packing unit (a collection of 5–7 sidechains in close contact in the native state) by the protein backbone and by the surrounding sidechains. Sidechains in a packing unit were removed therein back to the C^β, and the resulting cavity was refilled. The size and shape of the cavity was shown to restrict severely the sequence space, reducing the latter by a factor of the order of 10^4–10^6. The requirement to match the native packing density effected a further reduction of the order of 10^3, leading to only approximately 50 possible sequences for a packing unit. A further discrimination is now possible on the basis of the stabilization energies $\Delta W(\nu_{ij}^x|r \pm \Delta r)$ associated with different coordination geometries ν_{ij}^x. In general, a hierarchical approach considering, sequentially, the distinct effects of broad-range ($r \leq 6.5$ Å) interresidue contact potentials, short-range ($r \leq 4.4$ Å) interresidue contact potentials [22], and finally the excess energies due to geometric preferences, $\Delta W(\nu_{ij}^x|r \pm \Delta r)$, ought to prove useful in rapid and more effective recognition of correct sequence/structure matches.

Generation of protein conformations in on-lattice or off-lattice simulations

$\Delta W(\nu_{ij}^x|r \pm \Delta r)$ may be interpreted as an excess or residual contribution, positive or negative, by the specific coordination geometry to the potential of mean force $W_{ij}(r \pm \Delta r)$ between S_i and S_j given that this pair of residues is located at separation $r \pm \Delta r$. The potential of mean force W_{ij} ($r \pm \Delta r$) associated with the interaction of each pair of sidechains S_i and S_j has been calculated in our previous work (I Bahar, RL Jernigan, unpublished data), along with the effective interresidue contact potentials for distance ranges comparable to those presently explored. The present results may be directly combined, by simple addi-

tion, with these contact potentials, for example, in order to estimate the potentials $W(\nu_{ij}^x; r \pm \Delta r)$ incorporating the joint dependence on the three variables r_{ij}, ϑ_{ij}, and φ_{ij}. The essential use of this approach would be to select among alternative nonbonded positions that appear to be equivalent from the point of view of simple distance-dependent interresidue potentials, but that do effectively prefer specific coordination angles. $\Delta W(\nu_{ij}^x | r \pm \Delta r)$ may be cast into a form amenable to lattice simulations, by integration over appropriate regions of the conformational space conforming with the geometry of the lattice, and lattice sites may be readily selected for optimal placement of nonbonded sidechains on the basis of these potentials. Work along this line is in progress.

Rapid evaluation of potentially disruptive mutations

One application of the potentials $\Delta W(\nu_{ij}^x | r \pm \Delta r)$ is for the quantitative evaluation of mutations at a given site, on the basis of the coordination preferences of substituted residues in the given neighborhood.

Materials and methods

Model and method

In the virtual bond model presently adopted, each residue type i is represented by two interaction sites: C_i^α on the mainchain, and S_i on the sidechain (Table 4). The sites on the sidechains are selected on the basis of the specific structure and energy characteristics of the amino acid (I Bahar, RL Jernigan, unpublished data). Sites are connected by virtual bonds: l_i is the backbone virtual bond vector connecting C_{i-1}^α to C_i^α, and l_i^s is the sidechain bond vector pointing from C_i^α to S_i. Backbone bond vectors have fixed lengths (3.81 Å), whereas sidechain bond vectors have lengths varying between 1.53 Å (Ala) and 5.64 Å (Trp). Table 4 lists the average lengths of sidechain bond vectors. These represent the average values over the different torsional states χ_1, χ_2, etc. of the atomic representation of sidechains whenever the position of S_i depends on one or more rotatable bond.

The geometry of a pair of interacting sidechains is characterized by six variables describing their relative positions and orientations. These may be selected as follows: three of them are the spherical components (r_{ij}, ϑ_{ij}, and φ_{ij}) of the separation vector r_{ij} between the centroids of the two sidechains S_i and S_j (as shown in Fig. 2), the counterparts of these variables, as viewed from the side of S_j, provide two additional variables (ϑ_{ji} and φ_{ji}) and finally, a hypothetical dihedral angle, say ζ_{ij}, defined by the sites C_i^α, S_i, S_j, and C_j^α completes the description. This representation is a straightforward extension of the conventional representation of the backbone geometry of polymers in terms of bond dihedral angles and bond angles to the characterization of nonbonded sites. ϑ_{ij} and φ_{ij} are found from the operations:

$$\vartheta_{ij} = \cos^{-1}\left[(l_i^s \cdot r_{ij}) / |l_i^s|\,|r_{ij}|\right]$$

$$\varphi_{ij} = sgn(n_i^s \times n_{ij}^r)\cos^{-1}(n_i^s \cdot n_{ij}^r) \qquad (1)$$

where

$$n_{ij}^r \equiv (l_i^s \times r_{ij}) / |(l_i^s \times r_{ij})| \qquad (2)$$

Accordingly, ϑ_{ij} approaches zero when the separation vector r_{ij} connecting the two sidechain positions S_i to S_j is collinear with l_i^s, and assumes small values, say $\vartheta_{ij} \leq 60°$ for the so-called *front* positioning of S_j with respect to S_i. $\vartheta_{ij} \approx 90° \pm 30°$ will be referred to as *lateral* positions with respect to l_i^s. Large values such as $\vartheta_{ij} \geq 150°$ are precluded by steric clashes between the side group S_j and the backbone attached to S_i. φ_{ij} assumes the value $\varphi_{ij} = 180°$ for *trans* placement of r_{ij} with respect to l_i, the values 0° or 360° for *cis*, and the respective values of 60° and 300° for *gauche*$^-$ and *gauche*$^+$ states. For glycine, in order to be able to define the angles ϑ_{ij} and φ_{ij}, a hypothetical sidechain bond vector is assumed along the bisector of the wide angle made by atoms C_{i-1}^α, C_i^α, and C_{i+1}^α, which is coplanar with the virtual bonds l_i and l_{i+1}, and pointing in the direction away from the mainchain.

Table 4

Atoms chosen for side group representations and the average lengths of sidechain virtual bonds C_i^α–S_i.

Residue type	Atoms defining side group interaction centers*	C_i^α–S_i bond length†
Gly	C^α	–
Ala	C^β	1.53 ± 0.02
Val	$C^{\gamma1}$, $C^{\gamma2}$	2.21 ± 0.05
Ile	$C^{\delta1}$	3.71 ± 0.34
Leu	$C^{\delta1}$, $C^{\delta2}$	3.27 ± 0.13
Ser	O^γ	2.41 ± 0.06
Thr	O^γ	2.40 ± 0.06
Asp	$O^{\gamma1}$, $O^{\gamma2}$	3.06 ± 0.09
Asn	$O^{\delta1}$, $N^{\delta2}$	3.08 ± 0.58
Glu	$O^{\epsilon1}$, $O^{\epsilon2}$	4.16 ± 0.40
Gln	$O^{\epsilon1}$, $N^{\epsilon3}$	4.14 ± 0.43
Lys	N^ζ	5.64 ± 0.62
Arg	N^ϵ, NH^1, NH^2	5.51 ± 0.51
Cys	S^γ	2.80 ± 0.07
Met	S^δ	3.79 ± 0.41
Phe	C^γ, $C^{\delta1}$, $C^{\delta2}$, $C^{\epsilon1}$, $C^{\epsilon2}$, C^ζ	3.79 ± 0.08
Tyr	C^γ, $C^{\delta1}$, $C^{\delta2}$, $C^{\epsilon1}$, $C^{\epsilon2}$, C^ζ, OH	4.15 ± 0.12
Trp	C^γ, $C^{\delta1}$, $C^{\delta2}$, $N^{\epsilon1}$, $C^{\epsilon2}$, $C^{\epsilon3}$, $C^{\zeta2}$, $C^{\zeta3}$	4.40 ± 0.20
His	C^γ, $N^{\delta1}$, $C^{\delta2}$, $C^{\epsilon1}$, $N^{\epsilon2}$	3.55 ± 0.09
Pro	C^β, C^γ, C^δ	1.87 ± 0.04

*Atom notations are those used in PDB entries. †Errors refer to the difference between the results obtained from two distinct sets of 150 known proteins (I Bahar, RL Jernigan, unpublished data).

The three variables (r_{ij}, ϑ_{ij}, and φ_{ij}) for a given sidechain i, compiled over all neighboring sidechains j, yield information on the coordination of residue i by other amino acids in general. We concentrate on the distributions of the variables (r_{ij}, ϑ_{ij}, and φ_{ij}) for the purpose of identifying some recurrent coordination geometries for particular amino acids of type i. ϑ_{ij} and φ_{ij} are referred to as the polar and azimuthal angles describing the coordination of a central residue i by a nearby residue j; the second subscript in these variables is omitted whenever the coordination of a given type of sidechain S_i by any other type of sidechain is considered.

Materials

Here, a total of 302 nonhomologous structures from the Brookhaven Protein Data Bank (PDB) [27,28] is examined. These consist of two sets, Set I and Set II, comprising 150 and 152 nonhomologous protein structures (the complete list of these two sets of PDB structures can be found in the Supplementary material published with this paper on the internet). We focus on the association of each type of amino acid S_i by each type of nonbonded neighbor S_j, totaling a set of 20×20 different coordination geometries. Volume elements of size $\Delta\vartheta_{ij} = 30°$ and $\Delta\varphi_{ij} = 60°$ are considered. We note that the coordination state of sidechain S_j near S_i, which is defined by the pair of spherical angles φ_{ij} and ϑ_{ij}, is different from that of S_i near S_j.

Calculations are in general performed for two distance ranges, referred to as the broad ($r_{ij} \leq 6.8$ Å) and close ($r_{ij} \leq 4.4$ Å) distance ranges; the occurrences in the range $r_{ij} \leq 2.0$ are excluded. The average internal coordination numbers, $q^*_i(r_c)$, of the 20 amino acids on the basis of a spherical volume of radius $r_c = 6.8$ Å lie in the range $2.76 < q^*_i(r_c) < 5.89$, the lower and upper bounds referring to lysine and leucine, respectively. Here the term 'internal' refers to the coordination by other residues in the same protein. The total coordination numbers (by other residues and by solvent), on the other hand, vary as $6.65 < q^*_i(r_c) \leq 8.31$, tryptophan and lysine being the two limiting cases. These are calculated using the methods described in our previous studies [16,17]. Nearest sequence neighbors along the chain are excluded in the evaluation of these coordination numbers. In the close distance range, $r_c = 4.4$ Å, the internal and total coordination numbers vary as 0.4 (Pro) $< q^*_i(r_c) < 1.5$ (Cys) and 0.9 (Trp) $< q^*_i(r_c) < 2.2$ (Gly), respectively. This indicates that single contacts between pairs, rather than a collection of (multibody) contacts, are predominantly observed within the close distance range.

Probabilities and free energy changes associated with specific coordination loci

The most probable coordination states for specific pairs are identified by the following procedure. The coordination space is divided into 6×3 grids or volume elements of size $\Delta\varphi_{ij} = 60°$ and $\Delta\vartheta_{ij} = 60°$ for a given pair of sidechains S_i and S_j. Each grid X represents a coordination state ν_{ij}^X characterized by two-letter symbols, using *c*, g^-, s^-, *t*, s^+ and g^+ for the azimuthal angles $0 \leq \varphi_{ij} \leq 360°$, and *f*, *l* and *b* for the polar angles $0° \leq \vartheta_{ij} \leq 180°$. The division of the coordination angles is shown in Figure 2. Two distance ranges $2.0 \leq r_{ij} < 4.4$ Å and $4.4 < r_{ij} \leq 6.8$ Å are considered. We examine all S_i–S_j pairs in the 302 PDB structures whose radial separation lies within the distance range $r \pm \Delta r$ of interest. These add up to $N_{ij}(r \pm \Delta r)$. The fraction of these pairs falling in each grid ν_{ij}^X is determined from equation 3:

$$p(\nu_{ij}^X | r \pm \Delta r) = N(\nu_{ij}^X | r \pm \Delta r) / N_{ij}(r \pm \Delta r) \qquad (3)$$

and compared to the frequency $p° = 1/18$ expected for a random distribution over all grids. The normalization factor $\sin\vartheta_{ij}$ is neglected, the absolute number of observations being of interest. The ratio of the observed and expected probabilities is used to estimate the effective free energy change:

$$\Delta W(\nu_{ij}^X | r \pm \Delta r) \equiv -RT \ln p(\nu_{ij}^X | r \pm \Delta r) / p° \qquad (4)$$

associated with the selection of the particular coordination geometry ν_{ij}^X given that the sidechains S_i and S_j are located at a distance $r \pm \Delta r$ from one another.

Supplementary material available

The complete list of the stabilization energies $\Delta W(\nu_{ij}^X | r \pm \Delta r)$ in RT units associated with the coordination state ν_{ij}^X of all pairs of sidechains S_i and S_j for three distance ranges is published with this paper on the internet. The list of most probable coordination loci for each type of residue pair, along with their statistical weights, is available. Also published on the internet is the list of proteins used in this study.

Acknowledgement

Partial support by NATO CRG Project #951240 is gratefully acknowledged.

References

1. Bromberg, S. & Dill, K.A. (1994). Side-chain entropy and packing in proteins. *Protein Sci.* **3**, 997–1009.
2. Behe, M.J., Lattman, E.E. & Rose, G.D. (1991). The protein-folding problem: the native fold determines packing, but does packing determine the native fold? *Proc. Natl. Acad. Sci. USA* **88**, 4195–4199.
3. Richards, F.M. & Lim, W.A. (1994). An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423–498.
4. Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure* **2**, 641–649.
5. Lim, W.A., Farruggio, D.C. & Sauer, R.T. (1992). Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry* **31**, 4324–4333.
6. Harbury, P.B., Zhang, T., Kim, P.S. & Alber, T. (1993). A switch between two-, three- and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407.
7. Handel, T.M., Williams, S.A. & Degrado, W.F. (1993). Metal ion-dependent modulation of the dynamics of a designed protein. *Science* **261**, 879–885.
8. Matthews, B.W. (1987). Genetic and structural analysis of protein stability problem. *Biochemistry* **26**, 6885–6888.
9. Matouschek, A., Kellis, J.T., Serrano, L. & Fersht, A.R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**, 122–126.
10. Matthews, B.W. (1993). Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.* **62**, 139–160.
11. Matthews, B.W. (1995). Studies on protein stability with T4 lysozyme. *Adv. Protein Chem.* **46**, 249–278.
12. Bashford, D., Chothia, C. & Lesk, A.M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.
13. Russell, R.B. & Barton, G.J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility. *J. Mol. Biol.* **244**, 332–350.
14. Richards, F.M. (1977). Areas, volume, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.
15. Lumb, K.J. & Kim, P.S. (1995). A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**, 8642–8648.
16. Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552.
17. Miyazawa, S. & Jernigan, R.L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
18. Lim, W.A. & Sauer, R.T. (1991). The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* **219**, 359–376.
19. Ogata, K., *et al.*, & Sarai, A. (1996). The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DNA recognition and *trans*-activation. *Nat. Struct. Biol.* **3**. 178–187.
20. Merritt, E.A., Sarfaty, S., Pizza, M., Domeninghini, M., Rappuoli, R. & Hol, W.G.J. (1995). Mutation of a buried residue causes loss of activity but no conformational change in the heat-labile enterotoxin of *Escherichia coli*. *Nat. Struct. Biol.* **2**, 269–272.
21. Lo, T.O., Murphy, M.E.P., Guilemette, J.G., Smith, M. & Brayer, G.D. (1995). Replacements in a conserved leucine cluster in the hydrophobic heme pocket of cytochrome c. *Protein Sci.* **4**, 198–208.
22. Jernigan, R.L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195–209.
23. Singh, J. & Thornton, J.M. (1990). SIRIUS. An automated method of analysis of preferred packing arrangements between protein groups. *J. Mol. Biol.* **211**, 595–615.
24. Singh, J. & Thornton, J.M. (1992). *Atlas of Protein Sidechain Interactions.* Oxford University Press, New York.
25. Vriend, G. & Sander, C. (1993). Quality control of protein folding models: directional atomic contact analysis. *J. Appl. Crystallogr.* **26**, 47–60.
26. Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.
27. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). In *Crystallographic Databases – Information Content Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds). p. 107. Data Commission of the International Union of Crystallography, Bonn, Cambridge and Chester.
28. Bernstein F.C., *et al.*, & Tasumi, M. (1977). The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.