Experimental observed conformation-dependent geometry and hidden strain in proteins

ARTICLE *in* PROTEIN SCIENCE · JULY 1996

Impact Factor: 2.85 · DOI: 10.1002/pro.5560050719 · Source: PubMed

1 AUTHOR:

Paul Andrew Karplus
Oregon State University

199 PUBLICATIONS **14,825** CITATIONS

SEE PROFILE

Protein Science (1996), 5:1406–1420. Cambridge University Press. Printed in the USA.

Experimentally observed conformation-dependent geometry and hidden strain in proteins

P. ANDREW KARPLUS

Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, New York 14853 and
Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331

(RECEIVED March 11, 1996; ACCEPTED May 1, 1996)

Abstract

A database has been compiled documenting the peptide conformations and geometries from 70 diverse proteins refined at 1.75 Å or better. Analysis of the well-ordered residues within the database shows ϕ,ψ-distributions that have more fine structure than is generally observed. Also, clear evidence is presented that the peptide covalent geometry depends on conformation, with the interpeptide N-Cα-C bond angle varying by nearly ± 5 degrees from its standard value. The observed deviations from standard peptide geometry are greatest near the edges of well-populated regions, consistent with strain occurring in these conformations. Minimization of such hidden strain could be an important factor in thermostability of proteins. These empirical data describing how equilibrium peptide geometry varies as a function of conformation confirm and extend quantum mechanics calculations, and have predictive value that will aid both theoretical and experimental analyses of protein structure.

Keywords: amino acids; conformational energetics; local geometry; peptide geometry; protein folding; protein stability; protein structure; Ramachandran plot

In a landmark paper, Ramachandran and Sasisekharan (1968) reviewed concepts put forth originally by Sasisekharan (1962) to show how two main-chain torsion angles, ϕ and ψ, are the key variables for describing protein conformation, and how simple spatial exclusion considerations place major limitations on the conformations accessible to polypeptides (Fig. 1). They also outlined how more realistic analyses using van der Waals potentials convert the simple allowed/disallowed distinction to a continuous function of conformational energy. The general agreement of these plots with the observed conformations in proteins (compare Fig. 1A,B) has provided strong evidence that local interactions within a single dipeptide are of primary importance to conformational preferences.

However, significant deviations between the ϕ,ψ-distributions observed in proteins and calculated energetics have spurred much effort to determine more accurately the energetics of dipeptide conformations. These calculations have been made with various empirical force fields and quantum mechanics, with and without solvation terms, generating both "rigid-geometry" energy maps by assuming a fixed peptide geometry and "relaxed-geometry," or "local geometry," energy maps by allowing the covalent peptide geometry to be minimized at each point of the ϕ,ψ-map (reviewed in Brooks & Case, 1993). Although the basic features of the plots remain largely the same as those based on early force fields, quantum mechanics calculations tend to show very low energies for the rarely observed γ and γ'-turn conformations (see Fig. 1 for conformational definitions used in this paper), and most local-geometry plots tend to blur the distinction between the well-populated and some rarely observed ϕ,ψ regions (e.g., Anderson & Hermans, 1988). A recent comparison of several methods concluded that none of the current force fields can be said to predict the experimental results well, and that the local-geometry methods exaggerate the ease with which the peptide geometry can be distorted (Roterman et al., 1989).

Although deviations from standard geometry are seen in small-molecule crystal structures (e.g., Sasisekharan, 1962; Winkler & Dunitz, 1971), and some conformational dependence has been seen (Chakrabarti & Dunitz, 1982), these data are limited and the general relevance of such deviations has remained uncertain. On the one hand, ab initio quantum mechanics calculations, which have proven reliable in calculating local-geometry trends for a variety of compounds (Schäfer et al., 1986), indicate that covalent dipeptide geometry is quite sensitive to conformation, with the interpeptide N-Cα-C bond angle varying ± 4 degrees from its standard value (Schäfer et al., 1984; Klimkowski et al., 1985). The variation is seen to be only weakly dependent on the amino acid residue type (Siam et al., 1989), and the trends have been used to test and refine parameter sets used in molecular mechanics force fields (Chuman et al., 1984; Momany et al., 1990).

Reprint requests to: P. Andrew Karplus, Section of Biochemistry, Molecular and Cell Biology, 223 Biotechnology Building, Cornell University, Ithaca, New York 14853; e-mail: pak4@cornell.edu.

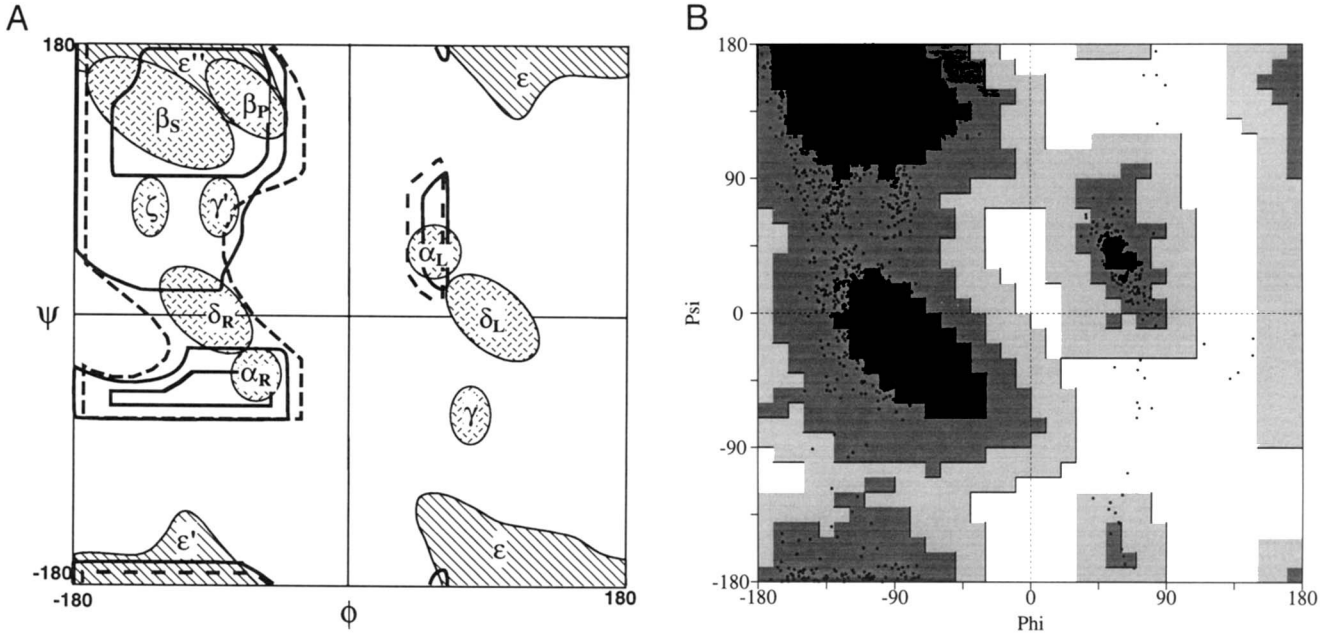

Fig. 1. Nomenclature and observed distribution of dipeptide conformations. **A:** Fully allowed and partly allowed regions for an Ala dipeptide with N-Cα-C angle = 110 degrees (thick solid lines) and the partly allowed regions for an N-Cα-C angle of 115 degrees (thick dashed lines) based on the hard sphere model (Ramachandran & Sasisekharan, 1968) are superimposed with a nomenclature for various regions of the plot that are used in this paper (shaded areas with central Greek letters). The regions are as follows: α_R, right-handed α-helix region; α_L, mirror image of α_R; β_S, region largely involved in β-sheet formation; β_P, region associated with extended polyproline-like helices (Adzhubei & Sternberg, 1993), but also observed in β-sheets; γ and γ', regions forming tight turns known as γ and inverse-γ turns (Milner-White, 1990); δ_R, right-handed region commonly referred to as the bridge region; δ_L, mirror image of δ_R region; ϵ, extensive region with $\phi > 0$, $\psi \approx \pm 180$ that is predominantly observed for Gly; ϵ' and ϵ'', mirror images of the two parts of the ϵ region, given distinct designations because ϵ'' overlaps heavily with the β_S and β_P regions that are observed commonly for other residues; ζ, a region newly defined in this paper that is largely associated with residues preceding Pro. The 3_{10}-helix conformation is located near the approach of the δ_R and α_R regions. Among the variety of nomenclatures used for regions of the ϕ,ψ-plot (e.g., Zimmerman et al., 1977; Richardson & Richardson, 1989; Rooman et al., 1992; Efimov, 1993; Thornton et al., 1995), this nomenclature is most similar to that of Efimov (1993). The regions are designated by simple shapes to imply that they do not have sharply defined borders useful for quantitative studies, but rather have been designed to encompass the most commonly observed conformations (compare with Fig. 1B) and serve to facilitate discussion. **B:** Plot for the non-Gly, non-Pro residues in the database that are not involved in *cis*-peptide bonds. This includes the group called "general residues" in this paper, plus Xpr residues. Background shading indicates the core (darkest), allowed (lighter), generously allowed (lightest), and disallowed (none) regions defined by Morris et al. (1992). The 32 residues with $\phi > 0$ but not in the α_L/δ_L region include three Thr, nine Ser, five Asn, six Asp, two His, one Cys, two Val, one Ile, one Phe, one Lys, and one Arg. Approximately one-third of the α_L/δ_L-residues are Asn. Asp and Asn also are highly represented in the ζ and γ' regions.

Recently, the trends also have been shown to be visible in high-resolution protein structures (Jiang et al., 1995). On the other hand, the assumption of fixed bond lengths and angles allows a great reduction in degrees of freedom of proteins and is still made in many theoretical and experimental studies of protein structure (e.g., Schaumann et al., 1990; Rice & Brünger, 1994; Maccallum et al., 1995a; Srinivasan & Rose, 1995). Furthermore, although standard macromolecular crystallographic refinement methods do generally allow variations in bond lengths and angles, a single set of ideal target bond lengths and angles is used for all conformations (e.g., Engh & Huber, 1991).

In order to refine our understanding of dipeptide energetics, it is important to obtain both accurate distributions of observed conformations and to determine the extent to which covalent geometry changes as a function of conformation. These results can, in turn, stimulate theoretical studies to provide explanations for these observations. In this paper, a database of high-resolution protein structures is described, and it is used to generate empirical distributions of observed conformations in which the effects of experimental error are minimized. Then, extending the work of Jiang et al. (1995), it is shown that this database provides unambiguous experimental evidence that the N-Cα-C bond angle is highly dependent on conformation. Other bond lengths and angles also show conformation dependence, but the local geometry trends are smaller and less well determined. Finally, the deviations in local geometry from the expected values indicate that conformational strain is not only present in residues residing in the "disallowed" regions of the ϕ,ψ plot, but is also found for many residues in the "allowed" regions.

Results and discussion

The database

Strategy for database generation

For the analysis of amino acid residue conformations, it is beneficial to have a *large* database of *independent* and *accurate*

ϕ,ψ values with which to work. The second two criteria tend to limit the amount of usable data, so the criteria for selecting database contents involves some compromise. Here, protein subunits were chosen that had less than 25% sequence identity with any other chosen subunit and that were crystallographically refined to *R*-factors ≤20% at ≤1.75 Å resolution (see Materials and methods). A 1.75-Å resolution cutoff was chosen rather the common 2-Å cutoff, because the effective resolution of a structure is often worse than the reported resolution, and because the apparent stereochemical quality of structures improves significantly between 2 and 1.75 Å resolution (Morris et al., 1992). Furthermore, conformations of residues with high *B*-factors are unreliable even at high resolution, so for the analyses here, a *B*-factor cutoff of 25 Å^2 was chosen as a compromise between ensuring well-defined conformations and having sufficient sampling. Residues meeting these criteria should all have well-defined electron density, and their coordinates should reflect the average atomic positions with accuracy better than 0.15 Å (Kuriyan et al., 1987; Fields et al., 1994). Similarly, stringent criteria have been used in other less extensive analyses of peptide conformations (Herzberg & Moult, 1991; Nicholson et al., 1992).

So that complete analysis of dipeptide geometry could be conducted, the parameters described in Figure 2 were calculated for the central residue of each dipeptide. These values include partial information for the peptide bond preceding the residue and full information for the peptide bond following the residue. Because the local environment of a single peptide is created by the conformational properties of two residues, a proper analysis of the conformational dependence of peptide geometry is a four-dimensional problem with variables ϕ_i,ψ_i, and ϕ_{i+1},ψ_{i+1}. However, the current database is not sufficiently large for such an analysis, and only ϕ_i,ψ_i are considered here.

Scope of the database

A total of 10,590 residues in the 70 proteins meet the main-chain *B*-factor criteria outlined above, and 7,600 residues qualify for χ_1 analyses (Table 1). Residues are represented in proportions typical of soluble proteins so that Cys, Trp, Met, and His have the fewest observations, with near 200 each (Table 1). All other amino acids have >350 observations. Residues of any type occurring before a Pro have unique conformational properties (MacArthur & Thornton, 1991) and have been grouped as a 21st amino acid type with the three-letter code Xpr. Also, because of the diversity of proteins chosen, all kinds of secondary structures are well-represented. Turns remain well-represented despite the *B*-factor cutoff, which tends to preferentially remove surface regions. The database includes 30 well-ordered *cis*-peptide bonds, including five non-prolyl ones (Table 1), but these have been excluded from the analyses presented in this paper.

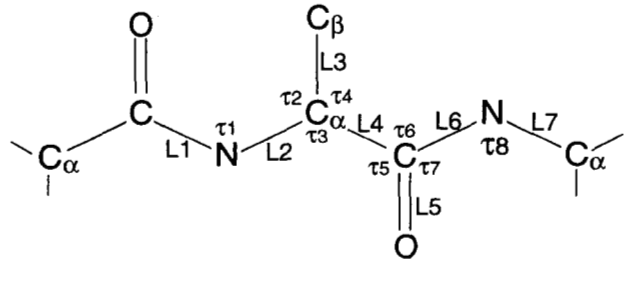

Fig. 2. Covalent geometric parameters stored in this database. A generic Ala-dipeptide is shown and the seven bond lengths (L1–L7) and eight bond angles (τ1–τ8) analyzed in this paper are indicated. Torsion angles discussed are ω_o and ω (the Cα-C-N-Cα angles for the peptide bonds preceding and following the central residue), ϕ (C_{i-1}-N_i-$C\alpha_i$-C_i), ψ (N_i-$C\alpha_i$-C_i-N_{i+1}), and χ_1 (N-C_α-C_β-C/O/S_γ).

High fidelity ϕ,ψ-distributions

Non-Pro, non-Gly residues

The observed ϕ,ψ-distribution for all non-Gly, non-Pro residues shows heavy population of the α_R/δ_R, β_S/β_P, and α_L/δ_L regions, with fairly sharply defined borders on the edges of these regions, especially as ϕ approaches 0 (Fig. 1B). It also shows discrete resolution of residues in the γ'-conformation, and a new region near (−120, 90) that is designated the ζ region. Thus, these regions are not simply general extensions of the β_S region. This distribution and the selected distributions shown in Figure 3 do not add fundamentally new insight, but they define more clearly the trends seen in other studies (e.g., Richardson & Richardson, 1989; Morris et al., 1992; Swindells et al., 1995) and strengthen the case that local energetics dominate the conformational preferences seen.

This distribution contrasts with the much broader distribution used by Laskowski et al. (1993) to set the current standard by which experimentalists assess what conformations are unusual in new protein structures (seen in the Fig. 1B shading). Their analysis used a database having a 2.5-Å resolution cutoff and no *B*-factor cutoff (Morris et al., 1992), so, as they noted, their distributions (and hence the definitions) reflect an unknown component of experimental error. For the database compiled here, 91.8%, 7.9%, 0.1%, and 0.2% of residues fall in regions Laskowski et al. (1993) defined as "core," "allowed," "generously allowed," and "disallowed," respectively. The plot (Fig. 1B) clearly indicates that much of the conformational space designated as "allowed" and "generously allowed," and even some of the "core" region is very rarely (or not at all) observed. Thus, the extents of these regions reflect considerable experimental error present in the database used to define them and are not optimal for evaluating new structures.

Interestingly, the higher accuracy database used here shows that more residues lie within the "disallowed" region than in the "generously allowed" region. These conformations are distributed along a strip with ϕ near +60 and in a region near $\phi = 135$, $\psi \approx -30$, and are mostly adopted by Asn, Asp, Ser, and Thr. For these short, hydrogen bonding residues, the side chains assume conformations that form intradipeptide hydrogen bonded rings with adjacent peptide O and NH atoms and presumably stabilize these strained conformations. Also, for the same reason (Karplus & Schulz, 1987), two-thirds of the general residues in the ϵ' region (especially with $\phi < -60$, $\psi \approx -175$) are Asn, Asp, Ser, and Thr residues (compare Fig. 1B with Fig. 3A,B,C). [Although His and Cys have atoms that theoretically could form intradipeptide hydrogen bonds, their distributions are more like the other amino acids, possibly due to lower conformational versatility (His) and lower polarity (Cys).]

Figures 3A, B, and C show selected subpopulations of the non-Gly, non-Pro residues that have distinct distributions. The non-β-branched, non-hydrogen bonding residues are well-distributed among the α_R, δ_R, β_S, β_P, and α_L regions, with only slight

Table 1. *Overview of the extent of the database*[a]

		Surviving cutoffs[b]				Secondary structures[c]										
Type	Total	Bmc	Bγ	Bsc	Xpr	H	h	G	g	E	e	T	t	B	S	–
Ala	1,069	944	NA	NA	35 (2)	486	29	38	17	171	38	58	29	6	22	50
Glu	570	465 (1)	323	276	15 (1)	215	15	19	3	97	14	35	9	3	20	35
Lys	749	629	445	316	24 (1)	253	21	26	6	127	30	62	27	15	26	36
Met	192	177	154	152	8	77	3	3	4	54	6	10	6	1	5	8
Gln	419	358	275	251	15 (1)	149	21	11	2	71	22	27	8	4	11	32
Arg	463	405 (1)	322	248	14 (1)	143	10	14	11	99	22	24	19	4	17	42
Leu	825	730	684	681	45 (2)	295	27	23	8	234	18	30	36	10	9	40
Phe	472	424	418	414	15 (1)	129	5	21	5	162	9	23	13	8	12	37
Tyr	471	425 (3)	417	402	23 (4)	99	10	13	9	169	8	31	26	11	12	37
Trp	202	178	177	175	5	48	2	13	4	67	10	7	11	3	4	9
Cys	201	178	174	174	15	45	4	12	1	57	9	13	6	5	6	20
His	242	213	197	191	13	72	14	11	3	50	10	12	12	1	7	21
Asn	594	518	417	408	23 (2)	112	49	15	8	87	36	74	29	8	40	60
Asp	714	602 (1)	452	436	28 (1)	169	47	31	14	91	35	76	37	3	42	57
Ser	782	654 (1)	528	528	34 (3)	165	46	21	4	170	29	76	33	5	49	56
Thr	763	641	582	578	28	153	31	10	5	218	35	44	19	15	38	73
Ile	560	521	512	510	34	191	8	15	3	229	15	14	12	5	9	20
Val	848	782	764	760	24	245	8	15	6	346	25	29	26	14	18	50
Gly	1068	868 (2)	NA	NA	31	141	66	30	14	139	47	189	76	12	68	86
Pro	518	438 (28)	417	417	11	68	10	32	9	41	25	101	8	10	39	95
Xpr	531	440 (26)	342	328	NA	29	70	3	29	70	30	7	84	13	30	75
Sum	12,253	10,590 (63)	7,600	7,245	440 (26)	3,284	496	376	165	2,749	473	942	526	156	484	939

[a] The number of occurrences are reported for each of the amino acids (excluding when they occur before proline) and separately for all residues occurring before proline (Xpr).

[b] Headings are as follows: Total, number of observations excluding chain termini and non-glycine residues missing a Cβ-atom; Bmc, number with average main-chain *B*-factor of residues $i-1$, i, and $i+1$ all ≤ 25 Å^2; Bγ, subset also having γ-atom *B*-factor ≤ 25 Å^2; Bsc, subset also having average side-chain *B*-factor ≤ 25 Å^2; Xpr, distribution of Xpr residues surviving main-chain *B*-factor criteria. The criteria are such that Bmc ≥ Bγ ≥ Bsc. After Bmc and Xpr entries, the numbers associated with *cis*-peptides (either before or after the residue) are given in parentheses. The five non-prolyl *cis*-peptides have sequence Glu-Tyr, Ser-Tyr, Pro-Tyr, Arg-Asp, and Gly-Gly.

[c] Distribution of the residues surviving the main-chain *B*-factor cutoff (Bmc) among the 11 possible secondary structure categories: H, α-helix; G, 3_{10}-helix; E, β-sheet; T, hydrogen bonded turn; B, isolated β-bridge; S, non-hydrogen bonded turn; lower-case letters h, g, e, t are assigned to the initial and terminal residues of the secondary structures represented by their upper-case letters; "–" indicates no defined secondary structure (Kabsch & Sander, 1983; Morris et al., 1992).

spreading into the δ_L and ϵ' regions (Fig. 3A). The only exceptions are three residues found near the small, classically allowed region (Fig. 1A) at $\phi,\psi = (60, -180)$ and a handful of residues in a region near $\phi,\psi = (-120, -120)$, bridging the ϵ' and the α_R regions. In all, 90% of these residues occupy only 10% of ϕ,ψ-space, close to the 7.5% value shown to be "fully-allowed" for Ala residues (Fig. 1A; Ramachandran & Sasisekharan, 1968). The β-branched residues strongly favor the α_R and β_S regions almost to the exclusion of δ_R, β_P, and α_L, so that 90% of Ile/Val residues are found in only about 6% of ϕ,ψ-space. The Xpr residues avoid the central β region near −120, +135 and are almost uniquely responsible for populating the ζ conformation near −120, 70.

Pro and Gly residues

For Pro, the ring structure constrains ϕ to be near −60 and this database shows that the limits of commonly observed conformations are $-90 < \phi < -45$ (Fig. 3D). The ϕ,ψ values are clearly correlated in both the α_R/δ_R and β_P regions, with values running at 45-degree angles intersecting $\phi,\psi = (-60, -30)$ and $(-65, 140)$, respectively. Based on a 90% cutoff, Pro residues are restricted to about 3% of ϕ,ψ-space.

For Gly, which can theoretically adopt many more conformations due to its lack of a β-carbon, observed conformations cluster in the α_R/δ_R, α_L/δ_L, and the ϵ, ϵ', and ϵ'' regions (Fig. 3E). As is commonly seen in proteins (e.g., Richardson & Richardson, 1989; Nicholson et al., 1992) and in small peptides (Görbitz & Etter, 1992), the classically allowed regions between $-120 < \psi < -60$ and $60 < \psi < 120$ are poorly populated, even though calculations show them to have the minimum energy (Hermans et al., 1992). The clustering is such that 90% of Gly residues fall within only 18% of the total conformational space, far short of the 45% fully allowed based on steric considerations alone (Ramachandran & Sasisekharan, 1968) and less than double the 10% seen for general residues in this analysis.

Another feature of the Gly distribution is that the observed conformations are distributed asymmetrically around the origin ($\phi,\psi = 0,0$). Most notable is the contrast between the highly populated δ_L-conformation and the rarely observed δ_R-conformation. The asymmetry could reflect significant nonlocal contributions to the energetics, but this need not be the case. It can also be explained as reflecting evolutionary selection based on the relative stabilities of the 20 residue types for the various conformations. For instance, assuming the δ_R-conformation is

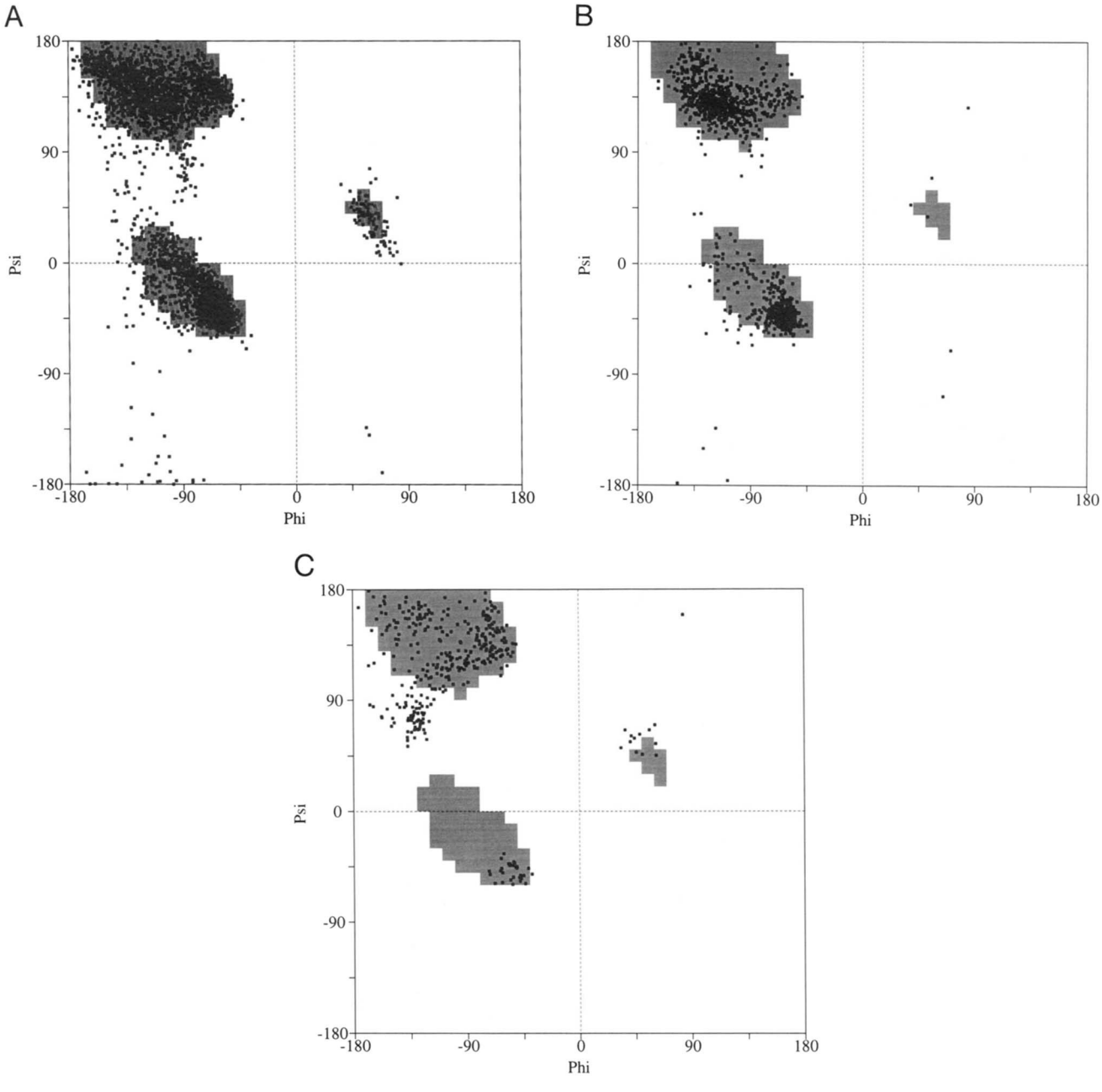

Fig. 3. Conformational distributions observed for selected residue types. **A:** The first 10 residue types from Table 1. Non-Pro residues with single γ-atoms, and not possibly involved in intradipeptide hydrogen bonding are grouped together here, although residues types within this group do have somewhat distinct distributions. **B:** The 1,303 Ile and Val (β-branched, non-hydrogen bonding residues). **C:** The 379 Xpr residues (non-Gly, non-Pro). **D:** The 410 Pro residues. **E:** The 866 Gly residues. For A–D, the background shading marks the most commonly observed (≥50 observations) conformational pixels for all non-Gly, non-Pro residues in this database. This allows differences in the distributions to be recognized easily. The local geometry maps (Figs. 4, 5, 7, and 8) also highlight these same regions. For E, the background shading is centrosymmetric, marking the most commonly observed pixels for Gly residues after centrosymmetric duplication (≥6 observations). Shaded regions include ~90% of the general and Gly residues, respectively. (*Continues on facing page.*)

equally stable for Gly and 15 other non-Pro, non-β-branched residues, but that the δ_L-conformation is uniquely stable for Gly, then only ~1/16th of all δ_R-residues will be Gly, whereas virtually all δ_L-residues will be Gly. This hypothesis is consistent with recent studies that show that the relative occurrences of residue types in a given conformation is correlated with their observed thermodynamic propensities (Muñoz & Serrano, 1994; Swindells et al., 1995). The existence of such selective pressure implies that the frequency distribution of any individual residue type should not be equated directly with the relative stabilities

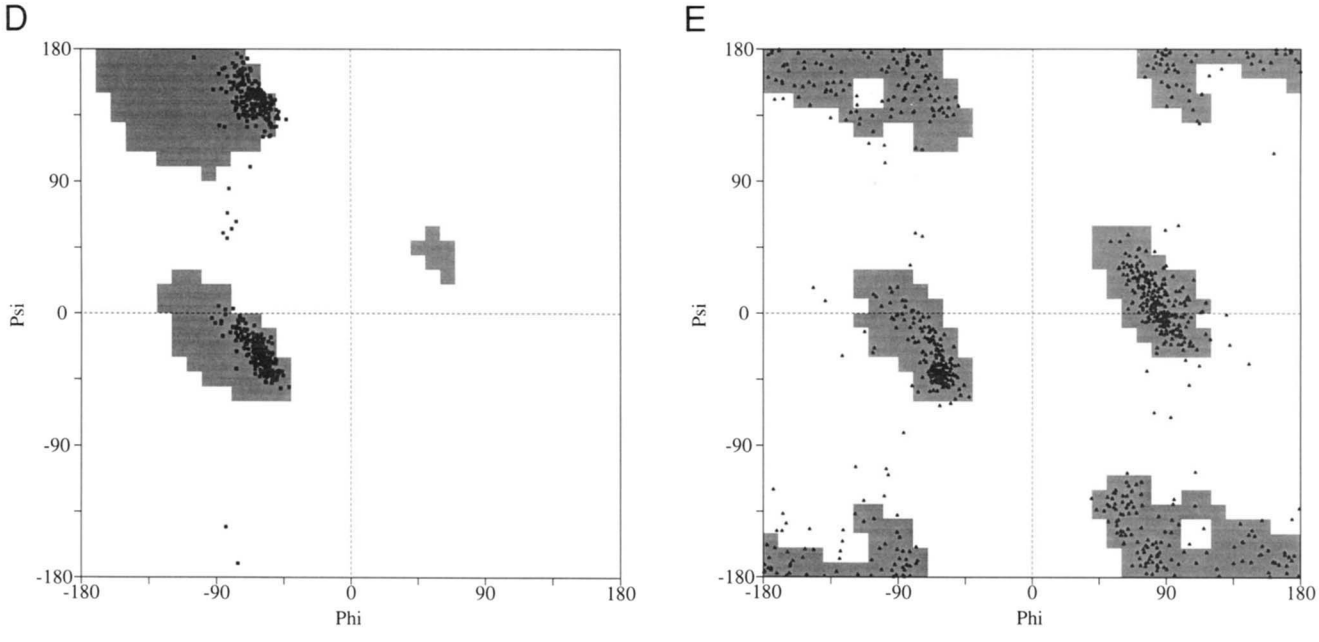

Fig. 3. *Continued.*

of the conformations for that residue, and thus should not be used directly to derive conformational entropies of residues in the denatured state (Stites & Pranata, 1995).

Implications

These high-fidelity ϕ,ψ-distributions show reliable details not well predicted by energetics calculations, and thus provide an important tool for refinement of such calculations. Also, because the proteins in the database are, at most, only distantly related, the observed population distributions should be generally applicable to soluble proteins and can be converted to probability maps that could improve conformational search strategies used in modern structure prediction methods. For instance, ~20% of the conformational space currently designated as α-helix and β-strand in the LINUS prediction method (Srinivasan & Rose, 1995) is rarely observed, so that ~50% of LINUS' helix and strand moves generate rarely observed conformations. Furthermore, analyses of this database can be extended to generate improved descriptions of other conformational properties of proteins (e.g., Efimov, 1993), such as more accurate correlations of main-chain conformation with the χ_1 torsion angle (Dunbrack & Karplus, 1993; also see below) and more accurate descriptions of turn conformations (Wilmot & Thornton, 1990).

Peptide geometry as a function of ϕ,ψ

Significant empirical variations occur

The average values of the bond lengths, bond angles, and ω-torsion angles in the database match the ideal values based on small-molecule crystal structures to within 0.01 Å for bond lengths and 1 degree for angles (Table 2). The averages observed here may provide slightly improved target values for use in refinement programs. However, what is more important than these small discrepancies is the observation that, for well-populated main-chain conformations, the average backbone bond lengths, bond angles, and the ω-torsion angles vary over ranges of about 0.015 Å, 4 degrees (and up to 8.8 degrees for $\tau 3$, the N-C_α-C angle), and 7 degrees, respectively (Table 2). The population sizes are such that the variations seen are 3 to $\geq$10 times the expected errors of the mean (Table 2). If less populated regions are included in the analysis, the absolute variations are even larger.

The variations tend to occur smoothly as a function of conformation, as is illustrated for the most variable parameters ω_o, ω (Fig. 4), and $\tau 3$ (Fig. 5). In the well-populated regions, the most notable feature of the ω_o distribution is a drop to near 177 degrees in the lower β_S region. The major trends for ω are a twisting from near 177 degrees in the upper β region to near 183 degrees in the lower β_S and γ' regions. Also, in the scattered conformations observed to the left of the α_R region (near $\phi,\psi = -120, -60$), the average ω_o and ω values rise above 185 degrees. Because the peptide planarity is likely to be strongly influenced by the conformations of the two contributing residues, the variations seen may underestimate the true conformational dependence of ω. The major trends for $\tau 3$ are a compression to values near 106 degrees in the lower β region and an extension to values near 114 degrees in the δ_R and α_L/δ_L regions (Fig. 5A). These observations validate the intuition of Sasisekharan (1962), who cautioned that variations in covalent geometry must be considered and later showed that the δ_R (and δ_L for Gly) region could become allowed by an opening of $\tau 3$ to 115 degrees (Fig. 1A). Separate averages over the α_R/δ_R and β_S/β_P regions as a whole show they have similar geometry, explaining why a previous analysis comparing the α and β classes of proteins noted little conformation dependence of peptide geometry (Marquart et al., 1983).

Table 2. *Expected and observed average values for peptide geometries*

Parameter[a]	E&H[b]	Average value (σ)[c]	Range[d]	Reference[e]
L1 (Å)	1.329	1.327 (0.018)	−0.009; +0.008	1.330 (0.005)
L2	1.458	1.467 (0.020)	−0.009; +0.005	1.465 (0.005)
L3	1.53	1.538 (0.023)	−0.008; +0.013	1.530 (0.005)
L4	1.525	1.526 (0.019)	−0.008; +0.009	1.525 (0.005)
L5	1.231	1.239 (0.018)	−0.006; +0.004	1.240 (0.005)
L6	1.329	1.327 (0.018)	−0.005; +0.011	1.330 (0.005)
L7	1.458	1.467 (0.020)	−0.006; +0.006	1.465 (0.005)
$\tau 1$ (°)	121.7	121.1 (2.5)	−1.3; +2.0	121 (1)
$\tau 2$	111.2	110.4 (2.8)	−1.3; +1.0	110 (1)
$\tau 3$	110.5	110.4 (3.3)	−4.7; +4.1	110 (1)
$\tau 4$	110.1	110.1 (2.9)	−1.8; +1.5	110 (1)
$\tau 5$	120.8	120.0 (2.4)	−1.6; +2.1	120 (1)
$\tau 6$	116.2	116.7 (2.5)	−2.0; +1.8	117 (1)
$\tau 7$	123.0	123.2 (2.5)	−1.2; +1.4	123 (1)
$\tau 8$	121.7	121.1 (2.6)	−1.2; +1.1	121 (1)
ω_o	180.0	179.5 (3.8)	−2.2; +3.8	180 (1)
ω	180.0	179.5 (3.8)	−2.6; +4.1	180 (1)
χ_1-g^+	NA	63.0 (10.0)	−7.0; +3.0	60 (5)
χ_1-t	NA	183.0 (12.0)	−4.0; +7.0	180 (5)
χ_1-g^-	NA	294.0 (10.0)	−6.0; +9.0	300 (5)

[a] Parameters are defined in Figure 2.

[b] Expected values taken from Engh and Huber (1991) for non-Gly, non-Pro residues.

[c] Mean and standard deviation found in this study for non-Gly, non-Pro, non-Xpr residues.

[d] Differences seen between the overall average value and the smallest and largest average values seen in individual ϕ,ψ regions containing 50 or more observations (or 25 or more for the χ_1-entries). For instance, for $\tau 3$, the range "−4.7; +4.1" means t3 ranges from 110.4 − 4.7 = 105.7 to 110.4 + 4.1 = 114.5 in the relevant conformations. Individual bins tend to have σ values similar to the overall σ value for the parameter, so the expected errors of the individual means are $\leq\sigma/\sqrt{50}$ for the main-chain parameters and $\leq\sigma/\sqrt{25}$ for the χ_1-entries.

[e] Base values (step sizes) used for calculating conformational strain energies and for the local geometry maps shown in this paper. For Pro residues, unique reference values are $\tau 2$ = 104 degrees, χ_1-g^+ = 20 degrees, and χ_1-g^- = 340 degrees.

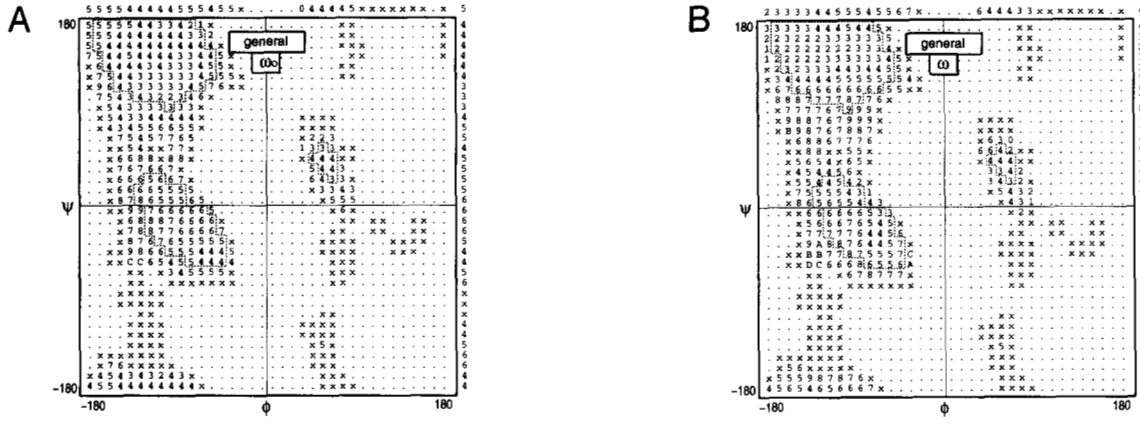

Fig. 4. Local geometry plots of peptide planarity. **A:** Average values of ω_o for the general residues (first 18 entries in Table 1). **B:** The same for ω. In this and all local geometry plots in this report, the observed average values are given for each ϕ,ψ pixel (in the central region), as a function of ϕ (along the top), as a function of ψ (along the right side), and the overall average (in the upper right-hand corner). The average values are reported with "5" representing the central reference value (in this case 180 degrees), "4," "3," "2," "1," "=," and "-" are given for successive steps (in this case, 1 degree) below the central value, and "6," "7,""8," "9," "A," "B," "C," "D," and "E" are given for successive steps above the central value. Values more than six steps below the reference value or more than nine steps above the central value maintain the "-" and "E" designations, respectively. The character "x" denotes a pixel with <5 observations and "." denotes a pixel with no observations. The central reference values and step sizes used for each parameter are given in Table 2. Dashed lines outline the ϕ,ψ pixels most commonly observed for general residues (≥50 observations). Trends of ω_o and ω seen in the β_S region may be influenced by long-range interactions such as interstrand hydrogen bonds because the averages shift somewhat if only residues not in secondary structure are analyzed (see the Electronic Appendix). The breakdown of the nonplanarity into out-of-plane bending and twisting components described by Winkler and Dunitz (1971) will require further analysis.

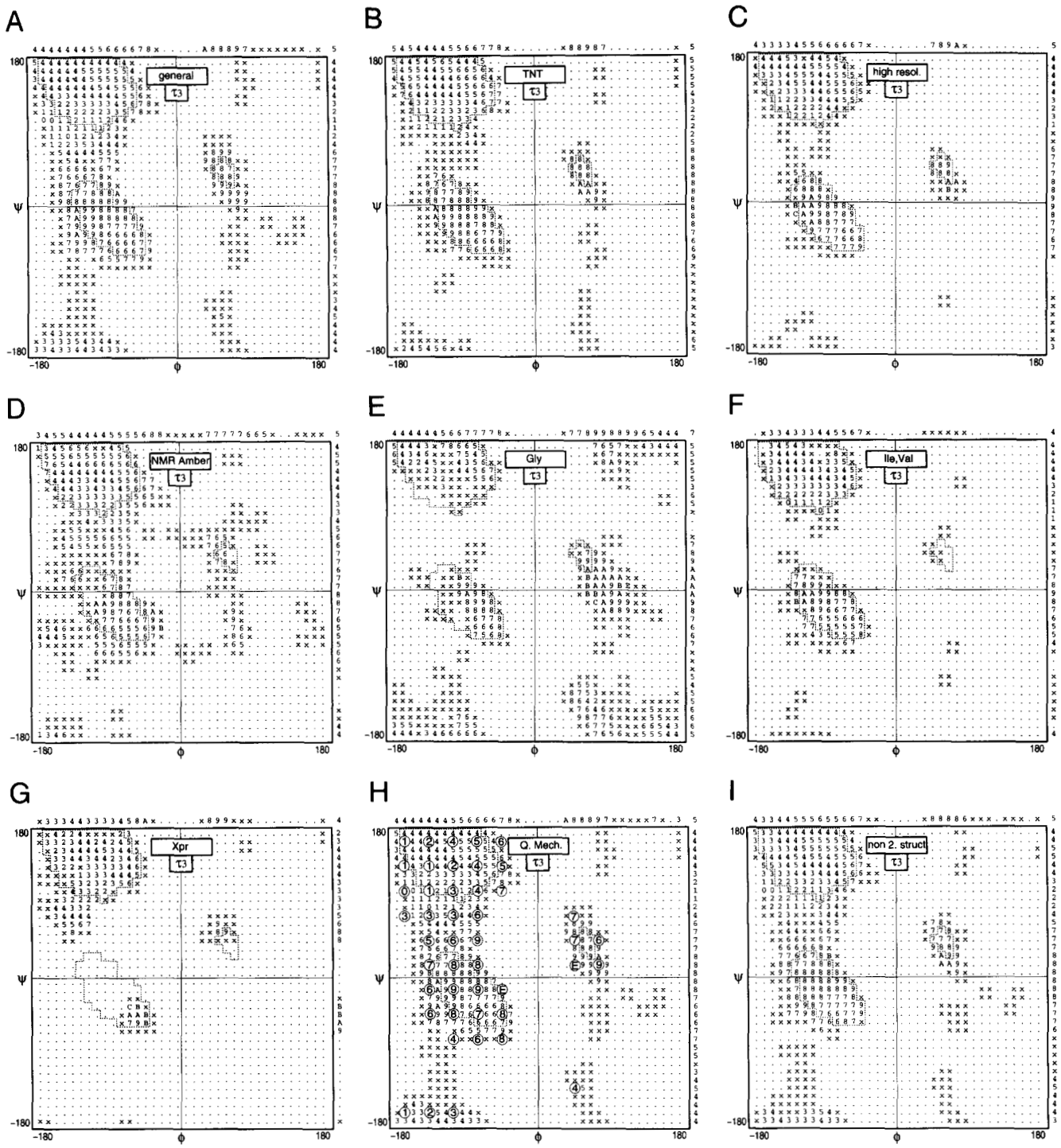

Fig. 5. Local geometry maps for $\tau 3$, the N-Cα-C angle. Symbols are as in Figure 4, but using a central reference value of 110 degrees and a step size of 1 degree. **A:** All 8,837 general residues in the database. **B:** The 1,873 general residues from structures refined with the TNT package. **C:** The 762 general residues from structures refined at ≤ 1.3 Å resolution. **D:** The 2,748 general residues from the multiple structures obtained for two proteins (see Materials and methods) by NMR restrained molecular dynamics using AMBER. **E:** The 866 Gly residues. **F:** The 1,303 Ile, Val residues. **G:** The 379 Xpr residues (non-Gly, non-Pro). **H:** Quantum mechanics results for discrete conformations of Ala (Schäfer & Cao, 1995). Quantum mechanics results are the larger circled numbers and, for reference, are superimposed on the results shown in A. **I:** The 2,215 general residues not centrally located in secondary structures (not designated H, G, T, E, or B).

The small variations seen in the bond lengths and most of the bond angles are important to document and understand, but they result in atomic shifts that are generally within the accuracy with which individual protein structures are determined. In contrast, the highly variable interpeptide bond angle $\tau 3$ has relatively large effects on structure because the peptide group magnifies the resulting atomic shifts. A $\tau 3$ angle of 115 degrees (or 105 degrees) compared with the standard ~110 degrees, mag-

nified by the 3.8-Å length of a peptide bond, corresponds to a 0.31 Å shift in the relative positions of 1–3 related C_α-atoms. Because the large variations in $\tau 3$ are the best defined and the most important to account for, they will be discussed here. Summaries of local geometry trends for all of the peptide parameters for various amino acid subgroups are provided in the Electronic Appendix.

The variations are not an artifact of refinement restraints

The observed variations in $\tau 3$ could be an artifact of crystallographic structure determination, resulting from the geometric restraints used during the refinement. Although the ideal bond length and angle restraints are the same for all peptides and thus should dampen rather than magnify any real deviations, repulsive and attractive nonbonded terms could cause conformational dependent deviations from ideal geometry. Two lines of evidence strongly indicate that the variations seen are real rather than restraint induced.

The first line of evidence makes use of the fact that the TNT refinement package uses no attractive nonbonded restraints, and repulsive restraints are only in effect when atoms separated by more than three bonds approach within 2.3 Å for N/O, O/O, and N/N pairs; 2.8 Å for C/N and C/O pairs; and 3.0 Å for C/C pairs (Tronrud et al., 1987). These distances are short enough that no repulsive terms come into play in any of the well-populated regions of the ϕ,ψ-map. The local geometry trends for the 13 TNT-refined proteins in the database (see Table 3) match those seen in the database as a whole (Fig. 5B), showing that the $\tau 3$ trends are not caused by nonbonded restraints.

The second line of evidence is that the $\tau 3$ variations for the eight database entries determined at ≤1.3 Å resolution (Table 3) are also similar (Fig. 5C). For these structures, the crystallographic data should dominate the refinement more strongly, diminishing any artifacts caused by restraints. In fact, the extent of variation seen in $\tau 3$ is actually somewhat larger than seen in the general database, indicating that the constraints used in refinement may cause an underestimation of the true extent of variation in the covalent geometry. However, the smaller number of observations will also contribute to greater variability. Interestingly, for these high-resolution structures, more significant deviations in bond lengths are also seen (see the Electronic Appendix), supporting the idea that a resolution near 1.5–1.75 Å is insufficient to fully resolve bond length deviations and thus underestimates these trends. However, with the increase in ultra high-resolution structures becoming available as a result of the use of synchrotron radiation (Dauter et al., 1995), it should be possible to compile a more extensive database for accurate analysis of these trends.

Local geometries in NMR-derived structures

Protein structures determined by NMR are derived by multiple approaches involving either fixed geometry, as in the program FANTOM (Schaumann et al., 1990), or restrained geometry with a muted van der Waals interaction (Nilges et al., 1988) or a full molecular dynamics force field (e.g., Gippert et al., 1990; Brünger, 1993). By definition, structures from fixed geometry methods show no local geometry trends. Also, structures (see Materials and methods) derived by molecular dynamics using a muted van der Waals interaction have peptide geometries clustered tightly around the target values (RMS deviations of <0.005 Å for bonds and <1 degree for angles) and show no conformation-dependent variation (data not shown). In contrast, structures (see Materials and methods) determined with a full force field, such as that of AMBER (Weiner et al., 1986), show a conformation dependence of $\tau 3$ that mimics, but does not duplicate the trends seen in the crystal structures (Fig. 5D). This shows that NMR data alone are not sufficient to reveal the variation in covalent geometry described above. This is not surprising, because the NMR-derived restraints leading to three-dimensional protein structures are rather weak in nature, specifying broad torsion angle and distance ranges. Based on this result, the variation seen in structures generated with the AMBER force field must then represent the effects of the force field itself. The deviations between the AMBER trends and the empirical trends [near (180, 180), in the lower β_S, and in the ζ and α_L regions] provide further evidence that the empirical results are not dictated by a force field used in refinement.

Variations in $\tau 3$ for various residue types

Analyses of small molecule structures have shown differences in the observed $\tau 3$ values for various residue types (Momany et al., 1975), most notably with Gly and Pro having high values near 112.5 degrees and 111.8 degrees, respectively (Engh & Huber, 1991) and Val/Ile having a low value near 108.5 degrees (Gould et al., 1985). This database shows similar trends, with overall average $\tau 3$ values for Gly of 112.1, Pro of 112.3, and Val/Ile of 109.3. However, the $\tau 3$-local geometry maps for Gly and Ile/Val (Fig. 5E,F) only show minor differences from the general residues (Fig. 5A), suggesting that the distinct average $\tau 3$ values are largely due to different conformational preferences: with Gly and Ile/Val largely populating regions with high

Table 3. *Codes for the 70 protein chains included in the database*[a]

119L_°	153L_°	1ADS_	1ARB_†	1BABB	1CBN_†	1CCR_	1CPCA	1CPCB	1CSEI†
1CUS_†	1DTS_	1ECA_	1ETBI	1FKB_	1FNC_°	1GCA_	1GDM_	1GOF_	1HFC_
1HTRP°	1IFC_†°	1ISUA°	1KNB_	1KNT_	1MDC_°	1OFV_	1PHP_	1PMY_	1POA_
1PPN_	1PTX_†	1ROPA	1S01_	1SBP_	1TCA_	1THV_°	1TRZB	1XNB_	2ALP_
2AYH_°	2BBKH°	2BOPA	2CBA_	2CCYA	2CPL_	2CTC_	2DRI_	2END_	2HBG_
2HMZA°	2IHL_	2MGE_	2MSBA	2RN2_	2SIL_°	2SN3_†	3CHY_	3CLA_	3DFR_
3EBX_	3SDHA	4GCR_	5P21_	7PTI_	7RSA_†	8ABP_	8RXNA†	8TLNE°	9RNT_

[a] The first four characters are the codes for the PDB (Bernstein et al., 1977) entry. The fifth character is the chain identifier, with "_" indicating that no chain identifier is required because there is only one chain in the file. "†" denotes structures refined at ≤1.3 Å resolution and "°" denotes structures refined by the TNT package.

(δ_L) and low (lower β_S) $\tau 3$, respectively. In contrast, Pro and Xpr residues show a distinct trend, with Pro having average $\tau 3$ values of 111.0 and 113.8 in β_P and α_R regions, respectively (see the Electronic Appendix) and Xpr having $\tau 3$ near 115 degrees in the α_R region (Fig. 5G). The α_R region is classically disallowed for Xpr residues (Schimmel & Flory, 1968; MacArthur & Thornton, 1991), but becomes energetically accessible if flexible-geometry is allowed (Hurley et al., 1992). Whereas the calculations predicted an implausible 8.5-degree change in the Cβ-Cα-C angle (Hurley et al., 1992), this analysis shows that the α_R region becomes sterically accessible through the noted distortion in $\tau 3$ and expansions of $\tau 4$ and $\tau 6$ by 3 degrees each (see the Electronic Appendix).

Comparisons with quantum mechanics

The local geometry trends seen here for $\tau 3$ confirm the relevance of quantum mechanical calculations, which have shown large conformation-dependent $\tau 3$ variations that are largely independent of residue type, and which tend to agree with values seen in high-resolution protein structures (Siam et al., 1989; Jiang et al., 1995; Schäfer & Cao, 1995). In particular, in the β_S and δ_R regions, the theoretical and experimental trends match well (Fig. 5H), but significant discrepancies of unknown origin occur at $\phi,\psi = (180, 180)$ and in the α_R and α_L regions. The former two deviations were also noted by Jiang et al. (1995) and were termed β-expansion and helix-compression. The quantum mechanical results have been used previously to refine the force fields of molecular mechanics packages (Momany et al., 1990), and the empirical results reported here provide more extensive data that can also be used for this purpose.

Variations seen are largely due to local interactions

The observed local geometry trends must reflect the sum of stresses caused by interactions between atoms of the dipeptide unit in question and interactions of that dipeptide with more distant residues in the structure. This analysis should emphasize trends due to local effects because, with the exception of residues in regular secondary structural elements, there is no strong correlation between the ϕ,ψ values of a residue and those of its neighbors, so that averages taken over many residues with a single conformation should average out any long-range effects. A separate analysis of residues that are not in helices, sheets, or turns shows similar $\tau 3$ trends (Fig. 5I), indicating that secondary structures do not perturb average $\tau 3$ values greatly. A small change in the α_R region from $\tau 3$ near 111 degrees for residues in α-helices to 112 degrees for residues not in α-helices is of interest. This shows that $\tau 3$ for "isolated" α_R-residues is closer the quantum mechanical result for an isolated dipeptide, and indicates longer-range contributions to the local geometry of α-helices. Also, for residues not involved in secondary structures, the α_R-conformation is somewhat less populated than the δ_R-conformation, suggesting that longer-range interactions in a helix stabilize the α_R conformation significantly. This is in agreement with other studies (Schäfer et al., 1993; Maccallum et al., 1995b), but is not accepted universally (Roterman et al., 1989).

Although the average local geometry trends reflect mostly intradipeptide interactions, it is likely that longer-range interactions cause significant real variation around the equilibrium values. The σ values, near 3 degrees for most bins, must reflect the sum of these real $\tau 3$ variations and coordinate error. An analysis of the best-resolved structures in the database ($\leq$1.3 Å resolution; Table 3) shows equivalent σ values, suggesting that real variations in $\tau 3$ contribute significantly to the observed spreads.

Implications for the rigid geometry approximation

Protein models incorporating the rigid geometry approximation are capable of representing protein conformations with a reasonable accuracy (e.g., Billeter et al., 1990; Vajda et al., 1993). However, given the extensive local geometry changes seen here, rigid geometry models cannot possibly portray the true energetics of intradipeptide nonbonded interactions. The δ_R region, which is generally considered unfavorable due to a very close approach [2.13 Å using rigid geometry (Ramachandran & Sasisekharan, 1968)] of the amide proton (NH_{i+1}) and peptide nitrogen (N_i), provides an illustrative example. The explanation for the frequent observation of this theoretically disfavored region is still a matter of debate (e.g., Roterman et al., 1989; Scully & Hermans, 1994), and the rigid geometry method ECEPP has sought to alleviate the discrepancy by "softening" the $NH_{i+1}\cdots N_i$ interaction potentials (Zimmerman et al., 1977). However, the observed changes in covalent geometry in this region are such that the $NH_{i+1}\cdots N_i$ separation (~2.36 Å) is actually very close to the 2.4 Å normal contact limit defined by Ramachandran and Sasisekharan (1968). Furthermore, Figure 6 shows how $\tau 3$ variations serve to maintain a special interaction geometry between NH_{i+1} and the π-orbital of N_i over a >60-degrees range in ϕ,ψ angles. The conservation of this geometry supports the idea that it may be a favorable interaction, which contributes to the stability of the δ_R region (Gieren & Dederer, 1978; Scarsdale et al., 1983).

Because the major variations occur for $\tau 3$, it is possible that an extension of rigid geometry methods to treat $\tau 3$ as a variable will improve their accuracy. Alternatively, for cases in which the number of variables must be kept to a minimum, the use of a table of conformation-dependent values of $\tau 3$, such as is available from this work, could account for the $\tau 3$ trends without an increase in the number of parameters.

Hidden strain visualized

Covalent strain

The shifts in covalent geometry seen for any given conformation are associated with an energetic cost that can be estimated by using bond stretching and angle bending force constants (see Materials and methods). The proper absolute values of such force constants can be debated, but because the purpose here is to define relative trends in energies, the absolute values are relatively unimportant. Also, because the covalent strains are based on the average observed geometries, they will be determined most accurately for well-populated regions. Finally, the presence of "covalent strain" energy does not necessarily imply that a conformation has high total energy, because shifts from standard geometry may be caused or offset by favorable electrostatic (or hydrogen bonding) interactions.

Among the regions well-populated (>50 residues) by general residues, the covalent strain is mostly $\leq$1 kcal/mol, but gets as high as 3 kcal/mol in three regions: the lower β_S region, near the lower left-hand portion of the δ_R region, and the $\phi =$

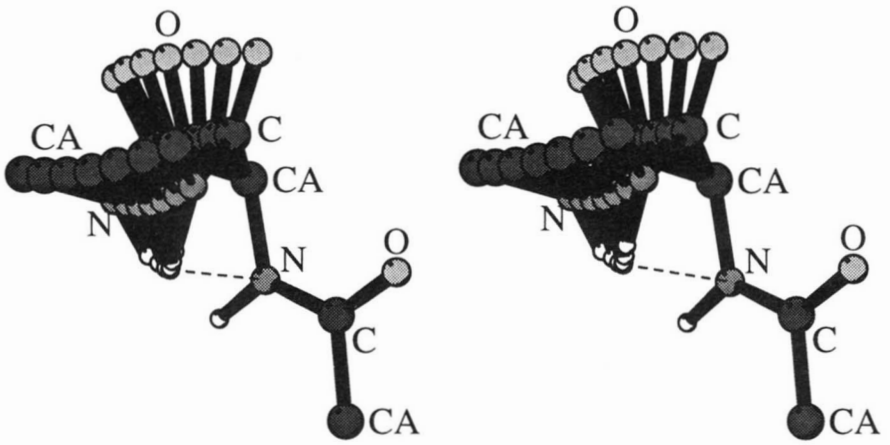

Fig. 6. Special intradipeptide interaction in the δ_R region. Seven dipeptide conformations are shown with the first peptides superimposed. The central conformation has $\phi,\psi = (-90,0)$ and the three conformations on each side have $\phi,\psi = (-100, 10)$, $(-110, 20)$, $(-120, 30)$, and $(-80, -10)$, $(-70, -20)$, $(-60, -30)$. These conformations span the diagonal line (defined by the equation $\psi = 90 - \phi$) that defines the strongly preferred conformations seen in the δ_R region (Fig. 1B). The structures shown all have a $\tau 3$ angle of 114 degrees. A dashed line indicates the close 2.36-Å approach of the NH_{i+1}-atom with the N_i-atom. Although the peptide dipoles are not well-aligned, the H-atom sits directly above the N_i-atom, where the nitrogen π-electron density should be highest. It has been hypothesized that the interaction involves an unconventional hydrogen bond to the π-electrons of the peptide bond (Gieren & Dederer, 1978; Scarsdale et al., 1983). Regardless of the name given to the interaction, it appears to be well conserved in the series of conformations shown, with the H···N distance increasing to only 2.49 in the end conformations. Furthermore, the observed shrinkage in the $\tau 3$ angle from 114 degrees to 111 degrees as one moves away from $(-90, 0)$ (see Fig. 5C) actually enhances the conservation of the H-position so that the H···N interaction distance increases to only 2.42 Å. The δ_L region involves an equivalent mirror-image interaction.

65–85-degree bin, corresponding to the δ_L region (Fig. 7A). In the β_S region, the main distortions are a 4-degree shrinkage in $\tau 3$ (Fig. 5A), and 3–4-degree twists of ω_o and ω (Fig. 4). These distortions appear related to an electrostatic attraction of the favorably aligned peptide dipoles, but optimization of interstrand hydrogen bonds may also play a role. Near the δ_R region, the main distortions are a 4–5-degree opening of $\tau 3$ (Fig. 5A), a 3-degree twist of ω_o (Fig. 4), and 2–3-degree compressions of $\tau 4$ and $\tau 5$. These distortions appear related to an electrostatically unfavorable approach of NH_i and NH_{i+1}. In the δ_L region, the main distortions are a 4-degree opening of $\tau 3$ (Fig. 5A), a 3-degree twist of ω (Fig. 4), and a 3-degree opening of $\tau 1$. These distortions all serve to relieve a close approach of O_{i-1} and C_β.

In the less-populated regions of ϕ,ψ-space, the apparent strain energies are often even higher, although these regions will have less accurately determined average properties so that the strain energies may be overestimated. Independent of the exact values seen, the higher strain near the edges of observed conformations is intuitively reasonable, because, at the borders between the observed and unobserved regions, the conformational energies should increase. Covalent strain maps for Ile/Val, Pro, and Xpr residues shows that these residue types have different strain maps related to their different conformational preferences (Fig. 7). Particularly noteworthy is the higher strain of all these residue types, especially Xpr, in the α_R and δ_R regions, and the higher strain seen for Ile/Val in the upper portions of the β_S and β_P regions.

Typically, researchers only consider residues located in "disallowed" regions as high energy conformations (Herzberg & Moult, 1991), but this analysis shows clearly that high-energy conformations are common near the edges of the well-populated regions. Also, because of the different conformational properties of certain classes of residues, even conformations central to regions observed commonly for general residues may have high energy. This view is consistent with results from experiments on staphylococcal nuclease, in which mutations to Gly are used to indirectly assess the strain energy associated with Cβ-atom interactions of residues in fringe conformations (Stites et al., 1994).

χ_1 torsional strain

Using this database, strain can also be seen in the deviation of torsion angles from the standard staggered values. As an example, local geometry maps of the average values for the three χ_1 rotamers of Ile/Val are presented in Figure 8. Because Ile/Val side chains are nonpolar, deviations from staggered conformations cannot be offset by electrostatic attraction and must be expensive energetically. The discrete distributions of the rotamers agree with previous studies (Dunbrack & Karplus, 1993), but what can be seen here is that the expected value of χ_1 for a given rotamer depends significantly on main-chain conformation, varying most dramatically for the *t* rotamer from near 175 degrees to near 205 degrees (Fig. 8B). The high values seen for the *t* rotamer in the lower δ_R region (near the 3_{10}-helix position) are intriguing because all three rotamers are about evenly populated in this limited region, implying that the other two rotamers are strained similarly. Also, for these β-branched side chains, another source of strain is due to conflicting optimal positions for the two γ-carbons. For instance, in the α_R region, unbranched, nonpolar side chains have average χ_1-rotamer values of 67 degrees, 184 degrees, and 290 degrees (see the Electronic Appendix). The deviations from staggered conformations suggest some strain for a single γ-atom, and the opposite directions of the deviations (−10 degrees for g^- and +4 degrees for *t*) must contribute additional strain by pushing the two

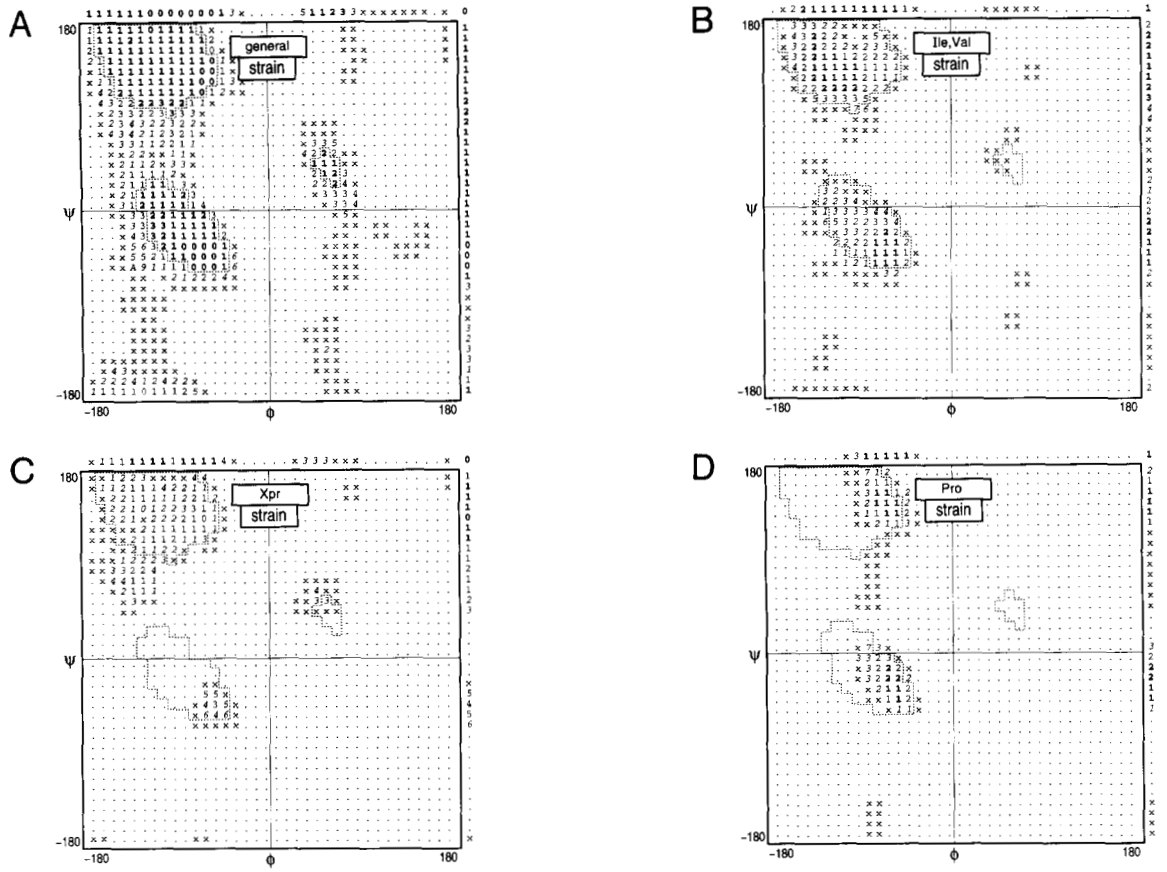

Fig. 7. Total covalent strain as a function of main-chain conformation. Strain is given to the nearest integer in units of kcal/mol, with "A" used for ≥ 10 kcal/mol. Pixels with ≥ 50 observations are shown in bold, pixels with ≤ 19 observations are shown in italics, "x" denotes a pixel with <5 observations; and "." denotes a pixel with no observations. **A:** Strain values for the average geometries seen for general residues. **B:** The same for Ile and Val residues. **C:** The same for Xpr residues. **D:** The same for Pro residues. The high strain for Xpr residues in the α_L region may be offset by local hydrogen bonding, because 8 of the 12 residues in that region are Ser, Thr, Asn, Asp, or Cys.

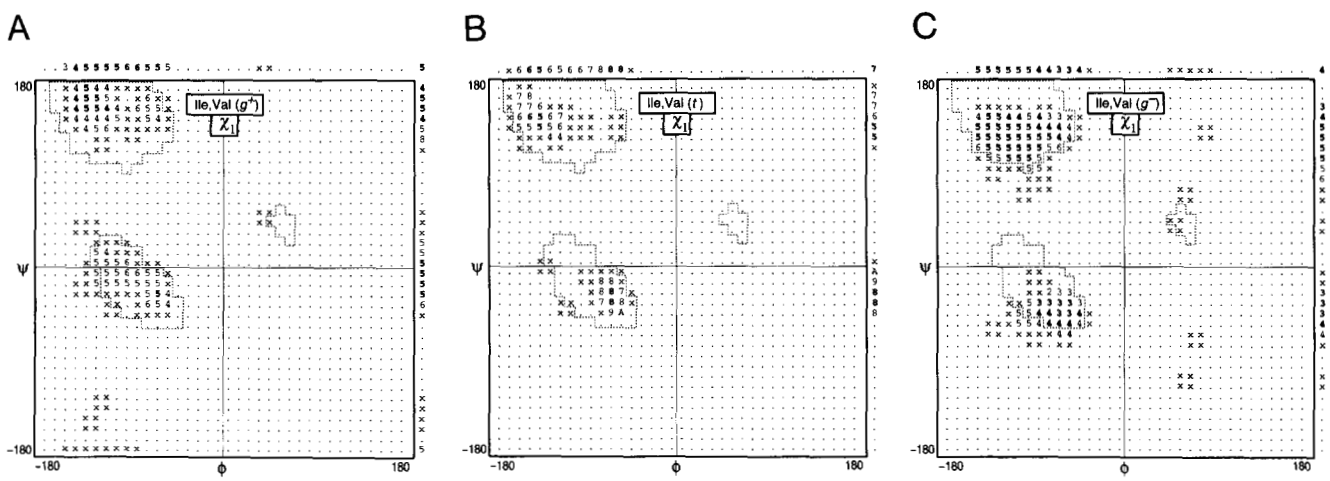

Fig. 8. Strain evidenced by deviation of χ_1 from staggered values. The local average χ_1 values are shown separately for the three rotamers of Ile and Val. The true χ_1 of Val is stored in the database, but 120 degrees has been added to the Val χ_1 values during statistical analyses to make it comparable to the χ_1 of Ile and Thr. **A:** The 212 residues found in the g^+ rotamer (central value 60 degrees). **B:** The 102 residues found in the t rotamer (central value 180 degrees). **C:** The 902 residues found in the g^- rotamer (central value 300 degrees). Symbols as in Figure 4, using a step size of 5 degrees. Bins with ≥ 20 observations are shown in bold.

γ-methyls apart in the predominant g^- rotamer seen for Ile/Val. Often, hydrophobicity (e.g., Blaber et al., 1994) and side-chain entropy factors (e.g., Creamer & Rose, 1994) are discussed as the primary factors relating to α_R-helix preferences (Chakrabartty & Baldwin, 1995). The results presented here suggest that steric strain, even within preferred rotamers, provides an additional reason that Ala appears to be the most stable residue in an α_R-helix, and why β-branched residues are less favorable. Generally, such steric factors may contribute significantly to the conformational preferences of amino acids, which in turn correlate with their secondary structural preferences (e.g., Muñoz & Serrano, 1994; Swindells et al., 1995).

Relation to protein stability and thermostability

The physical factors relating to protein thermostability have not yet been well-defined, because comparisons of thermostable proteins with their nonthermostable homologues often show few remarkable differences, leading to the view that thermostability results from the accumulation of numerous small differences, possibly including increased ion pairs, smaller loops, and a decrease in the surface-to-volume ratio (Russell et al., 1994; Adams & Kelly, 1995; Chan et al., 1995). The hidden strain documented here by analysis of main-chain geometry and side-chain torsion angle distortions is a potential contributing factor not explicitly considered in current discussions of thermostability. Such strain could be minimized through appropriate amino acid substitutions (for instance removing β-branched residues from 3_{10} and α-helices), or by subtle changes in structure, because strained conformations may differ by only a few degrees from unstrained conformations. In specific cases, the importance of this factor may be difficult to evaluate, because, when comparing single pairs of structures (even high-resolution structures), experimental errors seen in individual structures will complicate analyses. A lower level of hidden strain could contribute to the more favorable noncovalent energy seen in simulations of the hyperthermostable rubredoxin (Bradley et al., 1993), but, as long as energy-minimized structures do not match crystal structures accurately (e.g., Whitlow & Teeter, 1986), molecular mechanics calculations (e.g., Lazardis et al., 1995) may not capture the contribution of this kind of strain accurately.

Methods

Database construction

To select the protein chains, the World Wide Web (www/http: sander.embl.) was used to retrieve the June 1, 1995, listing of the set of representative Brookhaven Protein Data Bank (PDB) entries having less than 25% sequence identity (Hobohm et al., 1992; Hobohm & Sander, 1994). From this list, 70 protein subunit chains met the resolution and *R*-factor criteria and were selected (Table 3). For the analyses of NMR structures, five PDB entries were used: 9PCY and 1MYF as structures derived with a full AMBER force field, and 1IL8, 1BBN, and 6I1B as coordinates derived with a muted van der Waals potential.

Conformational information was extracted from each coordinate file by PROCHECK (Laskowski et al., 1993), which writes a ".rin" file containing sequence, secondary structure, torsion angles, and *B*-factors for the main chain (average), the γ-atom, and the side chain (average). Also, the "anglen" module of PROCHECK was modified to change the format of the output of bond lengths and angles written to the ".lan" file. Editing was required in some cases to remove alternate conformations, and one entry in which the *B*-factor slot contained atomic displacements. The individual ".rin" and modified ".lan" files were combined into a single master file. The geometric information contained in this file for the central amino acid residue includes the bond lengths L1-L7, the bond angles τ1-τ8, the torsion angles ϕ, ψ, ω, ω_o, χ_1, χ_2, and χ_3 (Fig. 2), as well as residue identification information, its secondary structure according to PROCHECK, and its average main-chain, γ-atom. and side-chain *B*-factors. To facilitate future analyses, a given record also contains the residue type, secondary structure, ϕ,ψ values, and main-chain *B*-factor for the preceding and following residues. The complete database and a program that reads the database and can be used to reproduce the analyses described in this paper is available by anonymous ftp from penelope.bio.cornell.edu (128.84.203.37).

Statistical analyses

For all analyses, mobile residues ($B \geq 25$ Å^2) and *cis*-peptides were excluded. Because three residues contribute atoms to any dipeptide, the average main-chain *B*-factors of all three residues were required to be $\leq$25 Å^2. Similarly, residues have been included in χ_1 analyses only if the γ-atom had a *B*-factor $\leq$25 Å^2. An ω-angle cutoff was also contemplated, but among residues fulfilling the main-chain *B*-factor criteria, the largest deviation in the ω angle from planarity was <25 degrees, so no cutoff was deemed necessary.

All statistical analyses of conformational properties were conducted in bins with a resolution of 10 degrees with an area of 20 degrees × 20 degrees. Thus, the bin centered at ϕ,ψ = (−65, −45) includes all residues with $-75 \leq \phi < -55$ and $-55 \leq \psi < -35$. This approach maintains a resolution comparable with the conformational accuracy, but incorporates a smoothing function that both accounts for the coordinate accuracy and serves to improve the statistics by increasing the numbers of observations within each bin.

Strain analysis

Covalent main-chain strain was calculated according to the equation:

$$\sum_{bonds} k_{bond}(L_{obs} - L_{base})^2 + \sum_{angles} k_{angles}(\tau_{obs} - \tau_{base})^2,$$

where k_{bond} = 1,480 kcal/(mol-Å^2) and k_{angle} = 216 kcal/(mol-rad^2). These values were based on the "tophcsd" parameter set of X-PLOR version 3.1 (Brünger, 1993), and were chosen such that the energy of deviations of 0.02 Å and 3 degrees (the typical standard deviations observed for bonds and angles in this database; see Table 2) would equal thermal energy at ambient temperature (~0.59 kcal/mol). The first sum was taken over five bond lengths (all but L3 and L7, because L3 differs for various residues and L7 is more associated with the next residue than the central one), and the second sum was taken over the eight bond angles (τ1–τ8) and the ω and ω_o peptide planarity angles. Although ω and ω_o are torsion angles, the same bending constant was used because ω and $\omega_o$have a similar observed standard deviations to the bond angles.

Supplementary material in Electronic Appendix

ASCII output files containing complete summary outputs for 11 separate analyses are found in the Electronic Appendix: sumary.general, general residues from all 70 proteins; sumary.gly, Gly residues from all 70 proteins; sumary.pro, Pro residues from all 70 proteins; sumary.xpr, Xpr residues from all 70 proteins; sumary.ile-val, Ile and Val residues from all 70 proteins; sumary.straight, Glu, Gln, Lys, Met, and Arg residues from all 70 proteins; sumary.nosecstr, general residues excluding those with H, G, E, B, or T secondary structures; sumary.hbonding, Ser, Asn, Asp, and Thr residues from all 70 proteins; sumary.highres, general residues from the subset of proteins refined at ≤1.3 Å resolution; sumary.tnt, general residues from the subset of proteins refined with the TNT program; sumary.amber, NMR structures determined with AMBER potential.

Acknowledgments

I thank all those crystallographers who have deposited their hard-earned coordinate sets in the PDB, Janet Thornton and her research group for making the code for PROCHECK publicly available, and Chris Sander and his research group for making publicly available useful results such as the updated representative PDB entries. I also thank Shing Ho for stimulating discussions, Lothar Schäfer for providing the exact values for use in Figure 5H, and the Department of Biochemistry and Biophysics at Oregon State University for their hospitality during my sabbatical, which allowed this work to be conducted. I am also grateful to Eduardo Padlan for calling to my attention that the ϕ,ψ-plot was first described by Sasisekharan alone at a 1960 conference on collagen (Sasisekharan, 1962), and only later together with Ramachandran. This work was supported in part by NIH grant 1P01 GM 48874.

References

Adams MWW, Kelly RM. 1995. Enzymes from microorganisms in extreme environments. *Chem Eng News 73*:32–42.

Adzhubei AA, Sternberg MJE. 1993. Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol 229*:472–493.

Anderson AG, Hermans J. 1988. Microfolding: Conformational probability map for the alanine dipeptide in water from molecular dynamics simulations. *Proteins Struct Funct Genet 3*:262–265.

Billeter M, Schaumann T, Braun W, Wüthrich K. 1990. Restrained energy refinement with two different algorithms and force fields of the structure of the α-amylase inhibitor tendamistat determined by NMR in solution. *Biopolymer 29*:695–706.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Blaber M, Zang X, Lindstrom J, Pepiot S, Baase W, Matthews B. 1994. Determination of α-helix propensity within the context of a folded protein: Sites 44 and 131 in bacteriophage T4 lysozyme. *J Mol Biol 235*: 600–624.

Bradley EA, Stewart DE, Adams MWW, Wampler JE. 1993. Investigations of the thermostability of rubredoxin models using molecular dynamics. *Protein Sci 2*:650–665.

Brooks CL, Case DA. 1993. Simulations of peptide conformational dynamics and thermodynamics. *Chem Rev 93*:2487–2502.

Brünger AT. 1993. *X-PLOR, version 3.1*. New Haven, Connecticut: Yale University.

Chakrabarti P, Dunitz JD. 1982. Structural characteristics of the carboxylic amide group. *Helv Chim Acta 65*:1555–1562.

Chakrabartty A, Baldwin RL. 1995. Stability of α-helices. *Adv Prot Chem 46*:141–176.

Chan MK, Muklund S, Kletzin A, Adams MWW, Rees DC. 1995. Structure of a hyperthermophilic tungstopterin enzyme aldehyde ferredoxin oxidoreductase. *Science 267*:1463–1469.

Chuman H, Momany FA, Schäfer L. 1984. Backbone conformations, bend structures, helix structures and other tests of an improved conformational energy program for peptides: ECEPP83. *Int J Pept Prot Res 24*:233–248.

Creamer TP, Rose GD. 1994. α-Helix-forming propensities in peptides and proteins. *Proteins Struct Funct Genet 19*:85–97.

Dauter Z, Lamzin VS, Wilson KS. 1995. Proteins at atomic resolution. *Curr Opin Struct Biol 5*:784–790.

Dunbrack RLJ, Karplus M. 1993. Backbone dependent rotamer library for proteins: Application to side chain prediction. *J Mol Biol 230*:543–574.

Efimov AV. 1993. Standard structures in proteins. *Prog Biophys Mol Biol 60*:201–239.

Engh RA, Huber R. 1991. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A 47*:392–400.

Fields BA, Bartsch HH, Bartunik HD, Cordes F, Guss JM, Freeman HC. 1994. Accuracy and precision if protein crystal structure analysis: Two independent refinements of the structure of poplar plastocyanin at 173 K. *Acta Crystallogr D 50*:709–730.

Gieren vA, Dederer B. 1978. Molecül- und Kristallstruktur von (S)-N,*N'*-di-*tert*-butyl-2-[*N*-(1-phenyl-ethyl)benzamido] malonamide, einem Nebenprodukt der Vierkomponenten-Kondensation (4CC). *Acta Crystallogr B 34*:533–539.

Gippert GP, Yip PF, Wright PE, Case DA. 1990. Computational methods for determining protein structures from NMR data. *Biochem Pharm 40*:15–22.

Görbitz CH, Etter MC. 1992. Hydrogen bond connectivity patterns and hydrophobic interactions in crystal structures of small, acyclic peptides. *Int J Pept Protein Res 39*:93–110.

Gould RO, Gray AM, Taylor P, Walkinshaw MD. 1985. Crystal environments and geometries of leucine, isoleucine, valine, and phenylalanine provide estimates of minimum nonbonded contact and preferred van der Waals interactions distances. *J Am Chem Soc 107*:5921–5927.

Hermans J, Anderson AG, Yun RH. 1992. Differential helix propensity of small apolar side chains studied by molecular dynamics simulations. *Biochemistry 31*:5646–5653.

Herzberg O, Moult J. 1991. Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins Struct Funct Genet 11*:223–229.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci 1*:409–417.

Hurley JH, Mason DA, Matthews BW. 1992. Flexible-geometry conformational energy maps for the amino acid preceding a proline. *Biopolymers 32*:1443–1446.

Jiang X, Cao M, Teppen B, Newton S, Schäfer L. 1995. Predictions of protein backbone structural parameters from first principles: Systematic comparisons of calculated N-Cα-C′ angles with high-resolution protein crystallographic results. *J Phys Chem 99*:10521–10525.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Karplus PA, Schulz GE. 1987. Refined structure of glutathione reductase at 1.54 Å resolution. *J Mol Biol 195*:701–729.

Klimkowski VJ, Schäfer L, Momany FA, Van Alsenoy C. 1985. Local geometry maps and conformational transitions between low energy conformers of *N*-acetyl-*N'*-methyl glycine amide: An ab initio study at the 4-21G level with gradient relaxed geometries. *J Mol Struct 124*:143–153.

Kuriyan J, Karplus M, Petsko GA. 1987. Estimation of uncertainties in X-ray refinement results by use of perturbed structures. *Proteins Struct Funct Genet 2*:1–12.

Laskowski RA, MacArthur MW, Moss DS, Thornton JM. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Crystallogr 26*:283–291.

Lazardis T, Archontis G, Karplus M. 1995. Enthalpic contribution to protein stability: Insights from atom-based calculations and statistical mechanics. *Adv Prot Chem 47*:231–307.

MacArthur MW, Thornton JM. 1991. Influence of proline residues on protein conformation. *J Mol Biol 218*:397–412.

Maccallum PH, Poet R, Milner-White EJ. 1995a. Coulombic attractions between partially charged main-chain atoms stabilize the right handed twist found in most β-strands. *J Mol Biol 248*:374–384.

Maccallum PH, Poet R, Milner-White EJ. 1995b. Coulombic interactions between partially charged main-chain atoms not hydrogen bonded to each other influence the conformations of α-helices and antiparallel β-sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includes all the coulombic interactions. *J Mol Biol 248*:361–373.

Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. 1983. The geometry of the reative site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallogr B 39*:480–490.

Milner-White EJ. 1990. Situations of gamma-turns in proteins: Their relation to alpha-helices, beta-sheets and ligand binding sites. *J Mol Biol 216*:385–397.

Momany FA, Klimkowski VJ, Schäfer L. 1990. On the use of conformationally dependent geometry trends from ab initio dipeptide studies to

refine potentials for the empirical force field CHARMM. *J Comp Chem 11*:654-662.

Momany FA, McGuire RF, Burgess AW, Scheraga HA. 1975. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, non-bonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J Phys Chem 22*:2361-2381.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. *Proteins Struct Funct Genet 12*:345-364.

Muñoz V, Serrano L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: Comparison with experimental scales. *Proteins Struct Funct Genet 20*:301-311.

Nicholson H, Tronrud DE, Becktel WJ, Matthews BW. 1992. Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme. *Biopolymers 32*:1431-1441.

Nilges M, Clore GM, Gronenborne AM. 1988. Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms circumventing problems associated with folding. *FEBS Lett 239*:129-136.

Ramachandran GN, Sasisekharan V. 1968. Conformation of polypeptides and proteins. *Adv Protein Chem 23*:283-438.

Rice LM, Brünger AT. 1994. Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins Struct Funct Genet 19*:277-290.

Richardson JS, Richardson DC. 1989. Principles and patterns of protein conformation. In: Fasman GD, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 1-98.

Rooman MJ, Kocher JPA, Wodak SJ. 1992. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry 31*:10226-10238.

Roterman IK, Lambert MH, Gibson KD, Scheraga HA. 1989. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. ϕ-ψ maps for *N*-acetyl alanine *N'*-methyl amide: Comparisons, contrasts and simple experimental tests. *J Biomol Struct Dynam 7*:421-452.

Russell RJM, Hough DW, Danson MJ, Taylor GL. 1994. The crystal structure of citrate synthase from the thermophilic archaeon, *Thermoplasma acidophylum*. *Structure 2*:1157-1167.

Sasisekharan V. 1962. Stereochemical criteria for polypeptide and protein structures. In: Ramanathan N, ed. *Collagen*. Madras, India: Wiley & Sons. pp 39-78.

Scarsdale JN, Van Alsenoy CV, Klimkowski VJ, Schäfer L, Momany FA. 1983. Ab initio studies of molecular geometries. 27. Optimized molecular structures and conformational analysis of Nα-acetyl-*N*-methylalaninamide and comparison with peptide crystal data and empirical calculations. *J Am Chem Soc 105*:3438-3445.

Schäfer L, Cao M. 1995. Predictions of protein backbone bond distances and angles from first principles. *J Mol Struct 333*:201-208.

Schäfer L, Ewbank JD, Klimkowski VJ, Siam K. 1986. Predictions of relative structural trends from ab intio derived standard geometry functions. *J Mol Struct 135*:141-158.

Schäfer L, Klimkowski VJ, Momany FA, Chuman H, Van Alsenoy C. 1984. Conformational transitions and geometry differences between low-energy conformers of *N*-acetyl-*N'*-methyl alanineamide: An ab initio study at the 4-21G level with gradient relaxed geometries. *Biopolymers 23*:2335-2347.

Schäfer L, Newton SQ, Cao M, Peeters A, Alsenoy CV, Wolinski K, Momany FA. 1993. Evaluation of the dipeptide approximation in peptide modelling by ab initio geometry optimizations of oligopeptides. *J Am Chem Soc 115*:272-280.

Schaumann T, Braun W, Wüthrich K. 1990. The program FANTOM for energy refinement of polypeptides and proteins using a Newton-Raphson minimizer in torsion angle space. *Biopolymer 29*:679-694.

Schimmel PR, Flory PJ. 1968. Conformational energies and configurational statistics of copolypeptides containing L-proline. *J Mol Biol 34*:105-120.

Scully J, Hermans J. 1994. Backbone flexibility and stability of reverse turn conformation in a model system. *J Mol Biol 235*:682-694.

Siam K, Kulp SQ, Ewbank JD, Schäfer L, van Alsenoy C. 1989. Ab initio studies of structural features not easily amenable to experiment. Part 64. Conformational analysis and local geometry maps of the model dipeptide *N*-acetyl *N'*-methyl serine amide. *J Mol Struct 184*:143-157.

Srinivasan R, Rose GD. 1995. LINUS: A hierarchical procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81-99.

Stites WE, Meeker AK, Shortle D. 1994. Evidence for strained interactions between side-chains and the polypeptide backbone. *J Mol Biol 235*:27-32.

Stites WE, Pranata J. 1995. Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone. *Proteins Struct Funct Genet 22*:132-140.

Swindells MB, MacArthur MW, Thornton JM. 1995. Intrinsic ϕ,ψ propensities of amino acids, derived from the coil regions of known structures. *Nature Struct Biol 2*:596-603.

Thornton J, Jones DT, MacArthur MW, Orengo CM, Swindells MB. 1995. Protein folds: Towards understanding folding from inspection of native structures. *Phil Trans Roy Soc Lond B 348*:71-79.

Tronrud DE, Ten Eyck LF, Matthews BW. 1987. An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr A 43*:489-501.

Vajda S, Jafri MS, Sezerman OU, DeLisi C. 1993. Necessary conditions for avoiding incorrect polypeptide folds in conformational search by energy minimization. *Biopolymers 33*:173-192.

Weiner SJ, Kollman PA, Nguyen DT, Case DA. 1986. An all atom force field for simulations of proteins and nucleic acids. *J Comp Chem 7*:230-252.

Whitlow M, Teeter MM. 1986. An empirical examination of potential energy minimization using the well-determined structure of the protein crambin. *J Am Chem Soc 108*:7163-7172.

Wilmot CM, Thornton JM. 1990. β-Turns and their distortions: A proposed new nomenclature. *Protein Eng 3*:479-493.

Winkler F, Dunitz J. 1971. The non-planar amide group. *J Mol Biol 59*: 169-182.

Zimmerman SS, Pottle MA, Nemethy G, Scheraga HA. 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromol 10*:1-9.