research papers

Protein imperfections: separating intrinsic from extrinsic variation of torsion angles

Glenn L. Butterfoss,[a] Jane S. Richardson[b] and Jan Hermans[a]*

[a]Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599-7260, USA, and [b]Department of Biochemistry, Duke University Medical School, Durham, North Carolina 27710-3711, USA

Correspondence e-mail: hermans@med.unc.edu

In this paper, the variation of the values of dihedral angles in proteins is divided into two categories by analyzing distributions in a database of structures determined at a resolution of 1.8 Å or better [Lovell *et al.* (2003), *Proteins Struct. Funct. Genet.* **50**, 437–450]. The first analysis uses the torsion angle for the $C^\alpha - C^\beta$ bond (χ_1) of all Gln, Glu, Arg and Lys residues ('unbranched set'). Plateaued values at low B values imply a root-mean-square deviation (RMSD) of just 9° for χ_1 related to intrinsic structural differences between proteins. Extrapolation to high resolution gives a value of 11°, while over the entire database the RMSD is 13.4°. The assumption that the deviations arise from independent intrinsic and extrinsic sources gives ∼10° as the RMSD for χ_1 of these unbranched side chains arising from all disorder and error over the entire set. It is also found that the decrease in χ_1 deviation that is correlated with higher resolution structures is almost entirely a consequence of the higher percentage of low-B-value side chains in those structures and furthermore that the crystal temperature at which diffraction data are collected has a negligible effect on intrinsic deviation. Those intrinsic aspects of the distributions not related to statistical or other errors, data incompleteness or disorder correlate with energies of model compounds computed with high-level quantum mechanics. Mean side-chain torsion angles for specific rotamers correlate well with local energy minima of Ace-Leu-Nme, Ace-Ile-Nme and Ace-Met-Nme. Intrinsic RMSD values in examples with $B \leq 20$ Å^2 correlate inversely with calculated values for the relevant rotational energy barriers: from a low of 6.5° for χ_1 of some rotamers of Ile to a high of 14° for some Met χ_3 for fully tetrahedral angles and much higher for χ angles around bonds that are tetrahedral at one end and planar at the other (*e.g.* 30° for χ_2 of the *gauche*[−] rotamer of Phe). For the lower barrier Met χ_3 rotations there are relatively more well validated cases near eclipsed values and calculated torques from the rest of the protein structure either confine or force the C^ε atom into the strained position. These results can be used to evaluate the variability and accuracy of χ angles in crystal structures and also to decide whether to restrain side-chain angles in refinement as a function of the resolution and atomic B values, depending on whether one aims for a realistic distribution of values or a spread that is statistically suitable to the probable data-set errors.

1. Introduction

The premise underlying this study is that each dihedral angle of any given side chain in a protein has its own specific equilibrium value which is determined by the details of the packing of the rest of the protein around the side chain; the objective is to determine and analyze the distributions and means of these

equilibrium values over many instances of the same side chain in many proteins. In order to determine this (intrinsic) distribution of side-chain dihedral angles from high-resolution crystal structures, it will be necessary to eliminate any part of the variation that arises from errors or disorder in the structures. As for all experimental sciences, X-ray crystallography is beset by errors in the experimental data (especially in phase determination) that in turn introduce uncertainties in the atomic positions even after exhaustive refinement. Thermal motion, static disorder and irregularities of the crystal lattice further increase the width of the distribution. In protein crystals, side chains may occupy alternative low-energy conformations, the existence of which is often not explicitly modeled, while fluctuations over the many possible solvent structures are coupled with fluctuations in the positions of protein atoms near the molecular surface. Again for proteins, the effect of errors in interpretation of the electron-density map when the conformation of a side chain has been misassigned must be included.

The effect of the errors in the observed structure factors is reduced by increasing the ratio of independent observations to unknowns, *i.e.* by increasing the number of observations or by reducing the number of independent variables in the model used to represent the electron density, or both. The former is achieved by collecting data to higher resolution, while the latter can be achieved by using a simple model for the structure and by introducing restraining relations between variables, most often geometric restraints related to the presence of chemical bonds between atoms (Brünger *et al.*, 1987). Typically, fluctuations in atomic position are represented with a simple model such as (isotropic or anisotropic) Gaussian distributions, whose width is measured by the atomic B value(s). As a rule, the effect of errors and of positional uncertainty or disorder is to increase the width of these distributions, *i.e.* to increase the atomic B values. In addition, B values reflect errors, incompleteness and radial fall-off of the experimental data; high-resolution structures inherently have smaller overall B values than low-resolution structures.

These relationships suggest that information about the effect of statistical error, positional uncertainty and data incompleteness on the distribution of a particular type of torsion angle can be obtained by comparing distributions of torsion angles defined by atoms with high and with low B values in structures of low and of high resolution. Data used to derive the rotamer library of Lovell *et al.* (2000) were restricted to side chains containing atoms with B values all $\leq$40 Å^2 (thus restricting to the relatively well defined as well as the high-resolution instances) in order to lower the noise level.

For many purposes, molecular structures are more usefully described by internal coordinates (bonds, bond angles and dihedral angles) than directly by atomic positions. In a first approximation, the bond lengths and bond angles may be thought of as fixed and the structure described in terms of the values of the torsion angles for internal rotation about single bonds. The conformations of side chains in proteins of known structure, when characterized as multi-dimensional combinations of their torsion angles, distribute into mostly well separated clusters called rotamers (Ponder & Richards, 1987; Dunbrack & Karplus, 1993; Lovell *et al.*, 2000). The rotamers assume sets of torsion angles that correspond to low-energy conformations of small molecules. For example, torsion angles for aliphatic C—C bonds cluster near 'canonical' values of +60, −60 and 180°.

Theoretical estimates for the standard deviations of coordinates in protein crystal structures at resolutions of 1.5–2 Å are about 0.1–0.2 Å (Jensen, 1997; Tickle *et al.*, 1998). An error of 0.1 Å in an atomic coordinate corresponds to an error of up to 6° in a torsion angle dependent on the atom's position. Experimental error estimates from comparison of identical structures determined independently are much higher, in the range 0.5–0.8 Å (Kleywegt, 1999; Mowbray *et al.*, 1999).

The principal aim of this study has been to assess and eliminate the effects of errors and positional uncertainties on distributions of torsion angles and to obtain distributions that correspond only to the intrinsic variation of torsion angles across proteins of known structure, not only in terms of relative rotamer population but also in terms of mean rotamer-angle values and the extent of deviations of torsion angles from those mean rotamer values. Our approach makes use of two different extrapolations of distributions of torsion angles in a database of high-resolution X-ray structures of proteins (Lovell *et al.*, 2000, 2003), one to instances of low atomic B values and the other to structures of high resolution. We also investigate the correspondence between the average rotamer structures and the corresponding low-energy structures of molecules that are sufficiently small that the structures can be carefully optimized with accurate energy functions based on high-level quantum mechanics and justify deviations of the conformations assumed in proteins from the conformations of such isolated low-energy structures in terms of interactions with other parts of the protein as a result of non-bonded forces.

2. Methods

2.1. Model molecules and energy function

We have used accurate *ab initio* quantum-mechanics-based methods to calculate the energies reported in this paper using *Gaussian*94 and *Gaussian*98 (Frisch *et al.*, 1998). Three dipeptide model structures were used, namely for leucine, isoleucine and methionine. The conformation of the dipeptides was optimized at the HF/6-31G(d) level of theory, with backbone torsion angles φ and ψ kept fixed in either an α-type ($\varphi = -60$, $\psi = -40°$) or a β-type ($\varphi = -120$, $\psi = 140°$) conformation. *n*-Butane and ethyl methyl sulfide (EMS) were used to determine the torsional barriers. These calculations were performed at the MP2/6–311+G(d,p) levels of theory. The MP2 (full) option was specified and all calculations used the tight self-consistent field option.

2.2. Database of well ordered residues in high-resolution structures

The database of Lovell *et al.* (2003) was used to extract statistics of side-chain conformations in folded proteins. This is

based on 500 selected non-redundant protein structures of 1.8 Å or better resolution. Lovell *et al.* (2000) use a single-letter notation to describe conformation, **t** standing for *trans*, **m** for *gauche⁻* and **p** for *gauche⁺*, and describe a particular rotamer with several of these letters in series, each to describe a successive side-chain torsion; we have adopted this notation for this paper.

Subsets of the database were extracted: a set containing all Met, Glu, Gln, Arg and Lys (23 620 residues) and a set containing all Glu, Gln, Arg and Lys (21 476 residues). Mean values and mean-square deviations of torsion angles were computed for each side-chain rotamer of every residue type studied, but using only side chains with B values ≤ 20 Å^2 for every atom and only if more than 95 examples of that rotamer occurred in the database.

Side chains of methionine, leucine and isoleucine with specific backbone structure were selected from the entire database. Thus, for methionine a set of 680 α-helical examples was obtained by selecting all residues listed as having an 'α-helix' or 'α-helix ext' secondary structure as assigned by the *DSSP* program (Kabsch & Sander, 1983) and for which $-80 > \varphi > -40$ and $-60 > \psi > -20$, and a set of 562 residues in extended conformation ('beta') was obtained by selecting all residues that had backbone conformations within 40° of $(\varphi, \psi) = (-120, 140°)$ (both with disregard of B values).

Most of the 500 data-set structures could be assigned to one of the two ranges of data-collection temperature: either above freezing (near 300 K) or cooled with liquid nitrogen (near 100 K). Many of the PDB file headers listed temperature; some were given in publications and in a few cases the depositor was contacted. Those cases that could not be determined were omitted from Fig. 4.

2.3. Evaluation of non-bonded torsion potentials

The (mean) torque acting to strain a particular torsion angle i is properly defined as

$$\langle T_i \rangle = -\frac{\partial G}{\partial \chi_i}, \qquad (1)$$

where G represents the free energy. An approximate value of such a mean torque can in principle be evaluated in a molecular-dynamics simulation of the protein. Drawbacks of this approach are the need to equilibrate the system of protein and solvent before values can be considered to be representative and the general observation of not inconsiderable shifts in atomic position during the equilibration. By using the gradient of the energy, rather than the free energy, *i.e.* with

$$T_i = -\frac{\partial E}{\partial \chi_i}, \qquad (2)$$

a major contribution to $\partial G/\partial \chi_i$ can rapidly be evaluated given a structure with experimentally determined atomic coordinates. This approach has been applied here to torsion about the C$^\gamma$—S$^\delta$ bonds of methionine residues. It proved convenient to evaluate the torque vector **T** with a special-purpose

routine added to a molecular-mechanics program (Mann *et al.*, 2002), according to

$$\mathbf{T}_{\mathrm{C}^\gamma-\mathrm{S}^\beta} = (\mathbf{r}_{\mathrm{S}^\delta-\mathrm{C}^\varepsilon} \times \mathbf{F}_{\mathrm{C}^\varepsilon}) \cdot \mathbf{e}_{\mathrm{C}^\gamma-\mathrm{S}^\beta}, \qquad (3)$$

where **F** is the force exerted on C$^\varepsilon$ by surrounding atoms, **r** is the S$^\delta$—C$^\varepsilon$ bond vector and **e** is the unit vector along the C$^\gamma$—S$^\delta$ bond. The non-bonded terms of the potential energy and atomic forces were evaluated in terms of Lennard–Jones 6–12 potentials with parameters from the *CEDAR/GROMOS* force field (Hermans *et al.*, 1984), with use of the program's standard force routine, at the experimental value of the C$^\gamma$—S$^\delta$ torsion angle and at successive increments of this angle by 5°.

In the *CEDAR/GROMOS* force field, methyl groups have a net charge of zero and the necessary energy terms are those for interactions of the ε-methyl group with surrounding atoms (but excluding C$^\gamma$ and S$^\delta$ of the same residue). With use of the force field's 'united-atom' potentials for CH$_3$, CH$_2$ and CH groups, the positions of H atoms are not included in this calculation. Since well defined methionine side chains tend not to be exposed to solvent, interactions with solvent have been ignored.

3. Results

3.1. Variation of χ_1 of long unbranched side chains

As mentioned, for a single C—C bond with tetrahedral sp^3 geometry at each end (as in an aliphatic hydrocarbon) the canonical values of the torsion angle are -60, $+60$ and $180°$ (which are also the canonical values for a C—S bond) and the values of such torsion angles of any given residue in the database can be separated into clusters (rotamers), with the clusters' centers approximating a set of these canonical values (Ponder & Richards, 1987; Dunbrack & Karplus, 1993; Lovell *et al.*, 2000). In order to have a large data set to work with, we have chosen deviations of χ_1 from the rotameric mean values for all Glu, Gln, Arg and Lys side chains in the database: the 'unbranched' set. The deviations are evaluated separately for each rotamer cluster. Since the mean torsion angles of the clusters do not coincide exactly with canonical values (owing to interactions with local backbone or other parts of the protein), the deviation of each instance i of a torsion angle is evaluated relative to the mean torsion value of the rotamer cluster
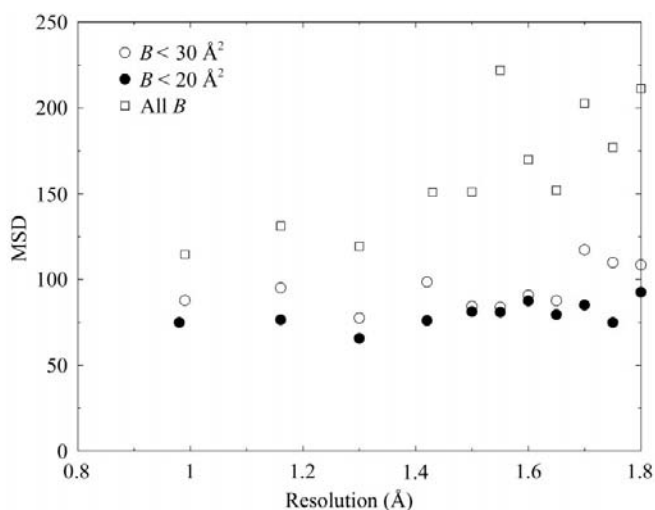
$$\delta\chi_i = \chi_i - \langle\chi_i\rangle_r, \qquad (4)$$

where the subscript r indicates the average over a particular rotamer cluster. The overall mean-square deviation of a given torsion angle $\langle(\delta\chi_i)^2\rangle$ is evaluated by averaging over the database and measures the empirical peak width or variability of χ_i in these data. For these tetrahedral geometry χ angles the distributions are nearly symmetric and unskewed, so the mean values correspond closely to the modal values used in Lovell *et al.* (2000) while allowing definition of mean-squared deviations.

Fig. 1 shows the variation of $\langle\delta\chi_1^2\rangle$ with resolution of the crystallographic structure for all examples in the unbranched

set and for those torsion angles defined by atoms with B values below 30 and below 20 Å^2. When torsion angles are considered regardless of the B values, the mean-square deviation (MSD) of χ_1 plateaus at a value near 120 deg^2 for an RMSD of 11°. At resolution worse than 1.5 Å the mean-square deviations increase and spread over an increasingly broad range. The number of dihedral angles used to compute $\langle \delta\chi_1^2 \rangle$ in each range of resolution is given in Table 1.

The mean-square deviations of χ_1 are better behaved as well as lower when the B values are limited, in spite of the smaller sample size. For the set of torsion angles with $B \leq$ 20 Å^2, $\langle \delta\chi_1^2 \rangle$ is, within statistical variation, independent of resolution of the structures and equal to 75 deg^2 for an RMSD of 8.7°.

Table 1
Distribution of Glu, Gln, Arg and Lys residues in the database into bins of decreasing resolution and RMSD of χ_1 from the rotameric mean.

Resolution range (Å)	No. angles	Mean resolution (Å)	RMSD (°)
$r < 1.10$	830	0.99	10.9
$1.10 \leq r < 1.22$	1034	1.16	11.4
$1.22 \leq r < 1.4$	801	1.30	11.2
$1.40 \leq r < 1.50$	1711	1.43	12.4
$r = 1.50$	2285	1.50	12.4
$1.50 < r < 1.60$	870	1.55	15.0
$r = 1.60$	3032	1.60	13.2
$1.60 < r < 1.70$	1827	1.65	12.5
$r = 1.70$	2960	1.70	14.4
$1.70 < r < 1.80$	1452	1.75	13.4
$r = 1.80$	4674	1.80	14.7
Entire data set	21476		13.4

Fig. 2 shows the correlation of $\langle \delta\chi_1^2 \rangle$ with B values. (Data are given in Table 2.) It can be seen that in this plot $\langle \delta\chi_1^2 \rangle$ plateaus at a value of 80 deg^2 for an RMSD of 9°. The results for structures of lower and higher resolution level off at the same value at low B and differ little at higher values of B. The value of $\langle \delta\chi_1^2 \rangle$ computed over this entire database is 180 deg^2. The MSD of χ_1 rises to quite large values for high B values, to almost half the MSD computed for a completely random distribution of χ_1 relative to three equally spaced canonical values, which equals 1200. Interestingly, one overall conclusion from Figs. 1 and 2 is that the increase in variance of χ_1 seen at lower resolution is entirely accounted for by the higher B values.

To give a broader view, Fig. 3 shows the MSD of χ_1 for all Met, Glu, Gln, Lys and Arg residues in each of a selection of structures, including some determined at lower resolution than the database. (The data of Fig. 1 for all B are included as filled circles.) A rapid rise in MSD for less well resolved structures can be noted. The wide spread of MSD values at lower resolutions may reflect differences in whether or not torsion-angle and other related restraints were imposed during crystallographic refinement and in whether rotamers were explicitly

Figure 1
Mean-square deviations (in deg^2) from mean canonical values of side-chain torsion angle χ_1 for Glu, Gln, Arg and Lys residues in the database, as a function of resolution, for all instances (squares), for atoms with $B \leq 30$ Å^2 (open circles) and for atoms with $B \leq 20$ Å^2 (filled circles).
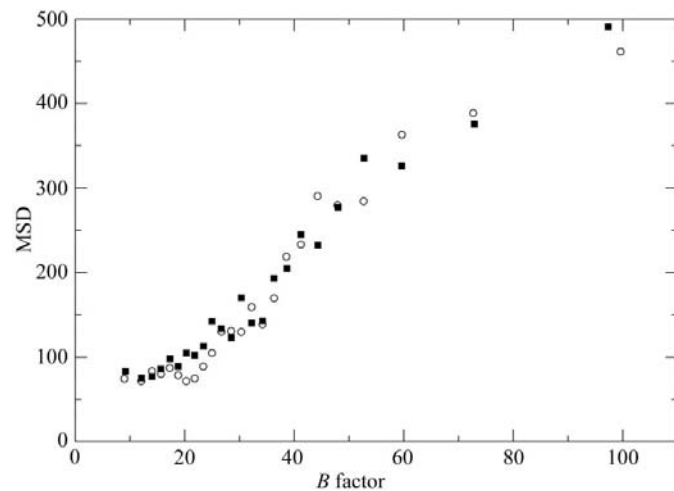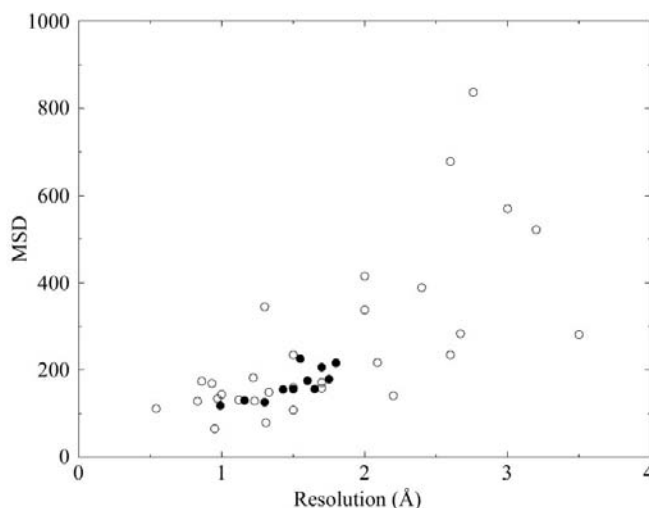

Figure 2
Mean-square deviations (in deg^2) from mean values of side-chain torsion angle χ_1 for all Glu, Gln, Arg and Lys residues in the database, as a function of B value (in Å^2), for the higher and lower resolution halves of the data set. Circles, resolution better than 1.63 Å. Squares, resolution between 1.63 and 1.80 Å.

Figure 3
Mean-square deviation of χ_1 (in deg^2) as a function of resolution. Filled circles, data from Fig. 1. Open circles, individual proteins (all Met, Glu, Gln, Lys, Arg residues).

Table 2
Distribution of Glu, Gln, Arg and Lys residues in the database into bins of increasing B value (in Å^2) and RMSD of χ_1 from the rotameric mean.

Range of B value	<1.63 Å resolution			1.63–1.83 Å resolution		
	Mean B value	No. dihedrals	RMSD of χ_1 (°)	Mean B value	No. dihedrals	RMSD of χ_1 (°)
$0.0 \leq B < 10.9$	9.0	630	8.6	9.2	368	9.1
$10.9 \leq B < 13.1$	12.1	664	8.4	12.1	335	8.7
$13.1 \leq B < 14.9$	14.0	598	9.1	14.0	397	8.8
$14.9 \leq B < 16.5$	15.7	627	8.9	15.6	379	9.3
$16.5 \leq B < 18.1$	17.3	604	9.3	17.3	390	9.9
$18.1 \leq B < 19.5$	18.8	571	8.8	18.8	428	9.4
$19.5 \leq B < 21.1$	20.3	577	8.5	20.3	424	10.2
$21.1 \leq B < 22.6$	21.8	530	8.6	21.8	473	10.1
$22.6 \leq B < 24.2$	23.4	539	9.4	23.4	459	10.6
$24.2 \leq B < 25.8$	25.0	524	10.2	25.0	477	11.9
$25.8 \leq B < 27.6$	26.7	496	11.4	26.7	508	11.5
$27.6 \leq B < 29.4$	28.5	500	11.4	28.5	494	11.1
$29.4 \leq B < 31.3$	30.3	490	11.4	30.3	511	13.0
$31.3 \leq B < 33.2$	32.2	503	12.6	32.2	502	11.8
$33.2 \leq B < 35.2$	34.2	474	11.8	34.2	526	12.0
$35.2 \leq B < 37.4$	36.3	470	13.0	36.3	527	13.9
$37.4 \leq B < 39.9$	38.5	449	14.8	38.7	551	14.3
$39.9 \leq B < 42.6$	41.2	487	15.3	41.2	513	15.7
$42.6 \leq B < 46.0$	44.2	433	17.1	44.3	568	15.2
$46.0 \leq B < 50.1$	47.9	390	16.7	48.0	609	16.6
$50.1 \leq B < 55.7$	52.6	386	16.9	52.7	617	18.3
$55.7 \leq B < 64.4$	59.6	355	19.1	59.6	644	18.1
$64.4 \leq B < 85.5$	72.7	349	19.7	72.9	652	19.4
$85.5 \leq B < 200.0$	99.6	246	21.5	97.3	366	22.2

used in side-chain building. Such differences matter increasingly at lower resolutions, where the experimental data are less definitively influential.

Fig. 4 shows the MSD of χ_1 of all Met, Glu, Gln, Lys and Arg residues as a function of the B value for χ_1 of the unbranched set, subdivided into crystals for which data were collected at room temperature and those collected under cryogenic conditions. Both sets plateau at nearly the same level (MSD $\simeq 80$ deg^2), with only a slight lowering (3 deg^2) for the cryogenic data, emphasizing that temperature does not affect the relation between B value and error in protein crystal structures.
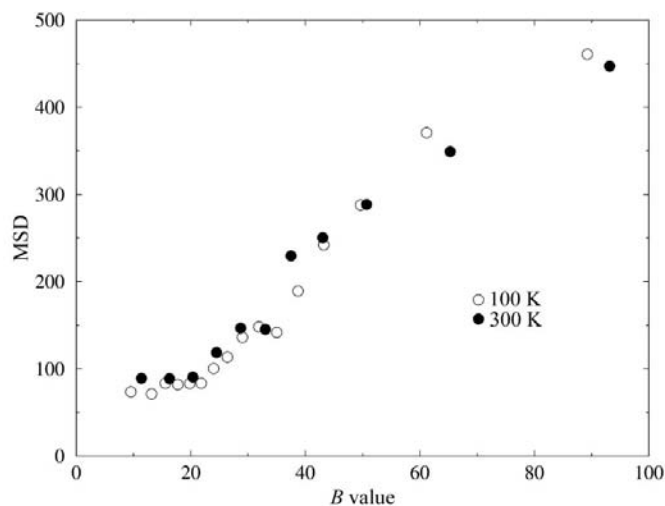
Figure 4
Effect of temperature, separated into structures with data collected near room temperature (filled circles) and at cryogenic temperatures (open circles). MSD in deg^2, B value in Å^2.

3.2. Separating statistical error and positional uncertainties from intrinsic variation

We make the not unreasonable assumption that the variation of χ arises from several sources, that the components from each source are independent and that each component produces a normal distribution of the deviation. Writing the deviation $\delta\chi$ as the sum of component deviations, the probability distribution of $\delta\chi$ is given by

$$\delta\chi = \delta_e\chi + \delta_s\chi + \ldots$$
$$P(\delta\chi) \simeq \exp\left(-\tfrac{1}{2}\delta\chi^2/\sigma^2\right) = \exp\left[-\tfrac{1}{2}\delta\chi^2/(\sigma_e^2 + \sigma_s^2 + \ldots)\right], \quad (5)$$

where the subscripts indicate different components and the variance σ^2 is equal to the sum of the variances of the components.

This analysis considers two separate sources of observed variation in the unbranched set: true structural differences between individual side chains (σ_s^2) and all sources of error combined, including among other factors positional uncertainty arising from the difficulty of fitting an average molecular structure to the data (σ_e^2). In principle, restriction to low B values (where the variance plateaus) should remove the extrinsic positional uncertainty. Assuming that this assumption applies to these data, the value of $80°$ obtained for the plateau at low B value can be equated with just the intrinsic structural variation. This then gives a value of $9°$ for the RMSD arising from actual structural differences, while the MSD for the whole set is 180 deg^2 (RMSD 13.4°), giving an MSD of 100 deg^2 (180 − 80) arising from accumulated errors for this set (RMSD 10°). This extrinsic variance would be even greater at resolutions lower than 1.8 Å.

Importantly, the level plots at low B in Figs. 1, 2 and 3 imply that by using only residues with no atomic B values above 20 Å^2, sets of torsion angles are obtained whose variation owing to statistical error and positional uncertainty is negligible compared with the intrinsic variation arising from structural differences.

3.3. Other side-chain torsion angles

The plateaued results described above show that the intrinsic variation of dihedral angles can be seen in isolation if examples are limited to atoms having B values no larger than 20 Å^2. Consequently, we have calculated the mean-square deviations of a set of the commoner rotamers of leucine, isoleucine, phenylalanine and methionine using only those with $B \leq 20$ Å^2. The results are given in Table 3. (A focus on non-polar side chains avoids the additional complications of electrostatics and hydrogen bonding.)

It can be seen that the $C^\alpha - C^\beta$ torsion angle (χ_1) for several conformations of Ile is more restricted than the χ_1 angle of the non-β-branched side chains (Gln, Glu, Lys, Arg, Leu, Met) or the $C^\beta - C^\gamma$ torsion angle χ_2 of Leu and Met. Torsion about the longer $C^\gamma - S^\delta$ bond of Met, χ_3, is less restricted than about the $C^\alpha - C^\beta$ and $C^\beta - C^\gamma$ bonds, giving higher variance, as can be seen in the Met and Lys χ_3 distributions compared in Fig. 5.

The greatest freedom is found for χ_2 of Phe, which is planar at C^γ, producing a much flatter χ_2 distribution that peaks near 90° rather than near 60° (see below). The MSD for χ_2 of Phe depends strongly on the conformation at the $C^\alpha - C^\beta$ bond because of differing interactions of the large rigid phenyl group with its own backbone at different values of χ_1. A lesser dependence of the variation of χ_2 on the value of χ_1 is seen for side chains that are tetrahedral at C^γ.

3.4. Correlation of mean rotamer and minimum-energy structures

Because of long-range interactions, especially those with the backbone, the coincidence of the observed rotamer mean torsion angles with canonical values for the bond type is not exact and the same is true of the torsion angles of minimum-energy structures of small molecules (except in the case of appreciable symmetry). In this section, we establish a correlation between the deviations from canonical values for, on the one hand, rotamers in proteins and, on the other hand, optimized conformations of small molecules. As a reference, the canonical staggered values of the torsion angle for single $C - C$ bonds (as in aliphatic hydrocarbons) are −60, 60 and 180° and these are also the canonical values for $C - S$ bonds in methionine.

The deviation for any one torsion angle of any one rotamer (subscript i) has been evaluated by averaging over all instances of that rotamer in the database, according to

$$\Delta\chi_{i,d} = \langle\chi_i\rangle_d - \chi_{i,0}, \qquad (6)$$

where the subscript d indicates averaging over the database and the subscript 0 indicates the corresponding canonical value of the torsion angle. [Note that the $\delta\chi_i$ of (4) is the

Table 3
Intrinsic RMSD of side-chain dihedrals in relatively well populated rotamers (B values 20 Å^2 or below).

| Residue	Rotamer	RMSD χ_1 (°)	RMSD χ_2 (°)	RMSD χ_3 (°)	No. instances
Unbranched†	m	8.6			3762
	p	9.3			537
	t	9.8			1997
	All	9.1	—	—	6296
Ile	mm	7.3	7.5	—	492
	mt	6.6	7.7	—	2185
	pt	6.5	7.0	—	380
	tt	8.4	7.6	—	208
Leu	mt	8.5	8.4	—	2979
	tp	9.3	7.7	—	1468
Met	mmm	9.7	9.3	10.4	195
	mtm	8.1	10.0	11.8	120
	mtp	7.3	9.3	10.8	157
	mtt	8.1	8.4	14.5	97
Phe	m‡	10.6	29.9	—	1530
	p	9.9	10.6	—	325
	t	10.3	19.7	—	980

† Gln, Glu, Lys and Arg.   ‡ Because of symmetry, χ_2 of Phe is distributed within a single interval, best considered as 0–180° because the preferred values are near 90°.

deviation of a single example from the observed rotamer mean, while $\Delta\chi_{i,d}$ is the deviation of that rotamer mean from the nearest canonical value.]

In a similar way, deviations from nearest canonical values of torsion angles of minimum-energy conformations of model compounds are defined with

$$\Delta\chi_{i,m} = \chi_{i,m} - \chi_{i,0}, \qquad (7)$$

where the subscript m indicates a particular torsion angle and deviation for a particular minimum-energy conformation. For the theoretical side of the comparison, we have used the quantum-mechanically optimized geometry of rotamers of the methionine dipeptide (Ace-Met-Nme) with fixed backbone
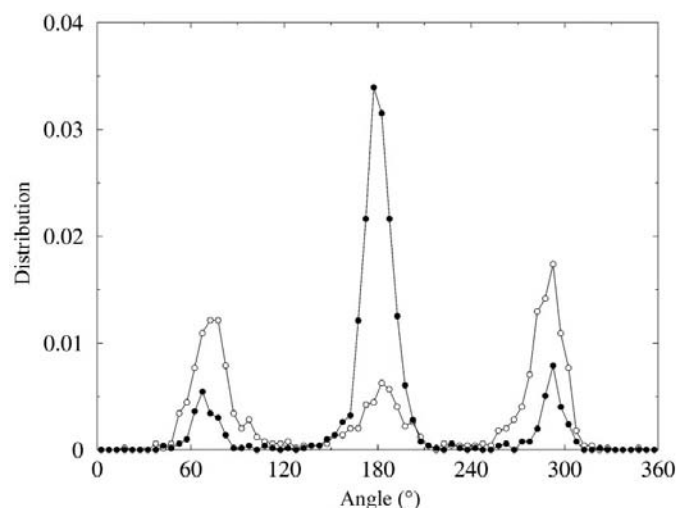
Figure 5
Distribution of χ_3 for lysine (filled circles) and methionine (open circles) side chains with $B \leq 20$ Å^2, showing the broader peaks for Met χ_3 and its preference for *gauche* rather than for *trans* values (Word *et al.*, 1999; Butterfoss & Hermans, 2003).

conformations ($\varphi = -120$, $\psi = 140°$ and $\varphi = -60$, $\psi = -40°$) representing β-sheet and α-helix sets in the database. (These dipeptide molecules contain a segment of protein-like backbone, but with only two additional single-bond torsional degrees of freedom.)

The comparison of observed *versus* theoretical geometries of the six most populated rotamers (**mmm**, **mtm**, **mtp**, **mtt** and **ttp** for both, **tpp** for the α-helix set and **ttm** for the β-sheet set) in each of these two sets is shown in Fig. 6 as a plot of the database $\Delta\chi_{i,d}$ (relative to nearest canonical values of 60, 180 or $-60°$) for a given rotamer as a function of deviation of the same rotamer in the structure of minimum energy, $\Delta\chi_{i,m}$. Fig. 7 shows similar results for leucine and isoleucine. The correlations are quite good.

3.5. Width of distributions and energy profile

The observation that the rotameric distributions cluster about minimum-energy structures of small molecules *in vacuo*, *i.e.* in the absence of the rest of the protein, makes it likely that the width of the distributions is somehow correlated with the energy required to deform the isolated structure locally from the energy minimum, which is to a first approximation determined by the second derivative of the energy U with respect to the torsion angle, $d^2U/d\chi^2$.

To illustrate this, we compare the distributions of Lys and Met χ_3 in the data set with the barriers for internal rotation of butane (C_2H_5–C_2H_5), 13.8 kJ mol^{-1} for the transition from *gauche* to *trans* and 23.0 kJ mol^{-1} for the transition from *gauche$^+$* to *gauche$^-$* (Allinger *et al.*, 1997), and the barriers for rotation of ethylmethylsulfide (EMS; $C_2H_5SCH_3$), 7.5 kJ mol^{-1} for the transition from *gauche* to *trans* and

19.2 kJ mol^{-1} for the transition from *gauche$^+$* to *gauche$^-$* (Butterfoss & Hermans, 2003). Because of the regular form of the dependence of U on χ, $d^2U/d\chi^2$ varies in the first instance as the barrier for rotation.

A comparison with the data reported in Fig. 5 and Table 3 shows that the lower barriers and hence greater ease of deformation of EMS relative to those of butane correlate with the wider distributions of the C—S torsion angle (χ_3 of methionine) compared with those of the C—C torsion angle (χ_3 of lysine) in comparably long unbranched side chains. Also, the preference of χ_3 of lysine for values near 180° but that of methionine for values near $\pm60°$ correlates with the energy difference between *gauche* and *trans* conformers of the models, the *trans* conformer being the more stable form for butane but the less stable form for EMS (Butterfoss & Hermans, 2003).

The large spread of values of χ_2 of phenylalanine, especially in the *gauche$^-$* conformation, correlates with a small value of $d^2U/d\chi^2$ and the minimal variation of the energy as this torsion angle is changed in an isolated model, Ace-Phe-Nme (data not shown). In general, torsion angles around a bond with tetrahedral geometry at one end (*e.g.* Phe C$^\beta$ or Glu C$^\gamma$) and planar geometry at the other end (*e.g.* Phe C$^\gamma$ or Glu C$^\delta$) have low rotational barriers and very broad distributions.

3.6. Relation between deviations of methionine χ_3 torsion angles and packing forces

Given the premise that the maxima in the distribution of torsion angles correspond to minima of the energy of the isolated side chain, any instance in which the mean value of the torsion angle does not coincide with such an energy minimum (or, rarely, maximum) is subject to an internal
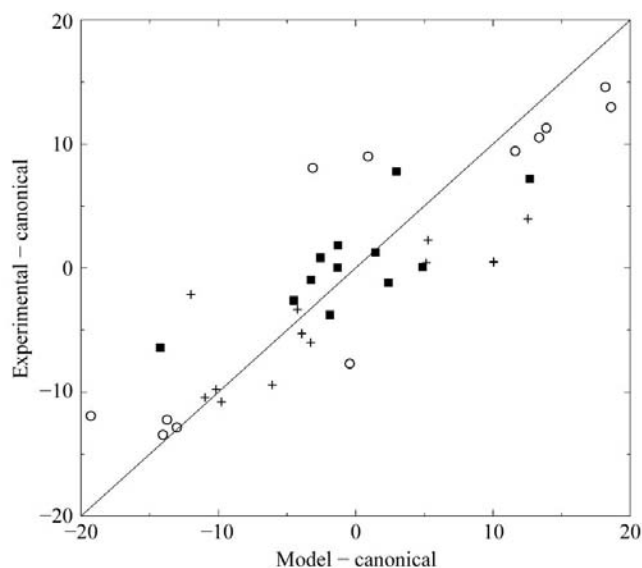
Figure 6
Correlation of 'deviation' (in degrees from canonical values of ±60 and 180°) of any of three dihedral angles in methionine side chains in the six most common conformers in the database in each of two sets having, respectively, α-helix and β-sheet backbone geometry, $\Delta\chi_{i,d}$ of (6), with the corresponding deviation for the minimum-energy structures of the dipeptide $\Delta\chi_{i,m}$ of (7). (Plus signs, χ_1; squares, χ_2; circles, χ_3; B values $\leq$ 30 Å^2.)
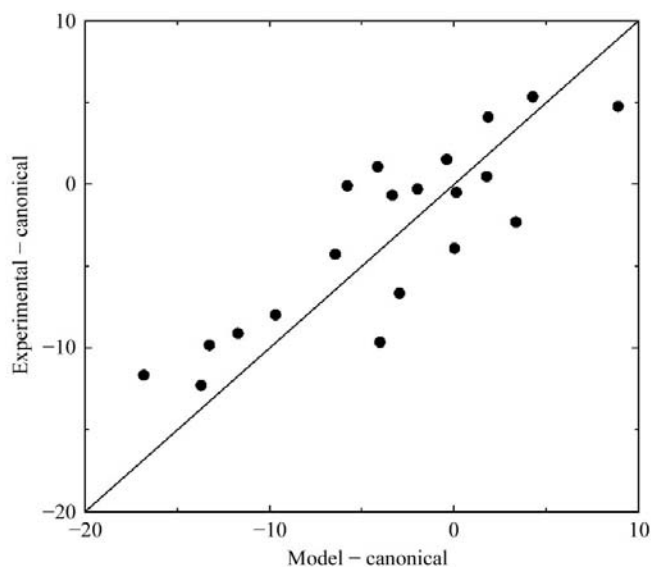
Figure 7
Correlation of 'deviation' (in degrees from canonical values of ±60 and 180°) of dihedral angles (χ_1 and χ_2) in leucine and isoleucine side chains in common conformers in the database, $\Delta\chi_{i,d}$ of (6), with the corresponding deviation for the minimum-energy structures of the dipeptides, $\Delta\chi_{i,m}$ of (7). (Leucine, **mt** and **tp**, α and β; isoleucine, **mt** and **mm**, α and β; **tt** and **pt**, only β.)

torque driving the conformation towards the nearest energy minimum. Since the net torque is zero, this internal torque must be opposed by an external torque of equal magnitude. In the case of χ_3 of methionine, the torque can be evaluated from the dependence on the $C^\gamma-S^\delta$ torsion angle of the interactions of the ε-methyl group with the remainder of the molecule. Here, energy and the corresponding torque of these interactions have been computed for conformations with all atoms held fixed except the ε-methyl and the energy and torque are evaluated in terms of pairwise interactions according to a Lennard–Jones potential of a molecular-mechanics force field. This calculation has been performed for 47 well defined methionine residues ($B \leq 20$) for which the torsion angle deviates significantly from the mean rotamer values, *i.e.* lies within 30° of the skewed conformations at ±120°.

Fig. 8(*a*) shows the dependence of the C^ε external non-bonded energy on χ_3 for a case in which the torsion angle observed for this residue is close to 60° away from a canonical value, *i.e.* in an eclipsed conformation near a maximum of the intrinsic energy, where the internal torque is small. The minimum of the external energy is here close to the experimental position. In contrast, Fig. 8(*b*) shows the dependence of the external energy for a case in which the residue's torsion angle is equal to 98° and the conformation is therefore roughly halfway between staggered and eclipsed. In this case the external torque is positive, driving the torsion angle to higher values, farther away from the canonical staggered value. In the first case, the rest of the protein can be said to confine the residue to the observed torsion angle, while in the second case the rest of the protein applies an external torque that forces the residue to the observed torsion angle.

Fig. 9 shows the all-atom contacts of the methionine residue from Fig. 8(*b*) with the surrounding protein. This high-resolution (1.5 Å) low-*B* side chain is very accurately located by clear electron density. The long-range protein environment prevents adoption of other rotamers. The local environment restrains the Met χ_3 angle in its unfavorable conformation, since any further rotation of the ε-methyl toward a *gauche* angle would produce a steric clash with the peptide backbone (its own CO and H$^\alpha$). Since those interactions are independent of backbone conformation, this rotamer (**mmp**) can never have a staggered χ_3. The modal χ_3 value for **mmp** is 103° (Lovell *et al.*, 2000), producing a pronounced shoulder in the Met χ_3 plot of Fig. 5. The fact that a fairly well populated rotamer cluster (3% of all Met) can occur so far from the 60° *gauche* value is yet another confirmation of the lower rotational barrier for Met χ_3. 18 of the 47 Met examples considered here are **mmp** and an additional eight cases are **tpm**, another backbone-constrained rotamer cluster first recognized here. The other half of the near-eclipsed cases, however, are generally in between two otherwise accessible staggered values and are confined by more distant parts of the protein.

Considering all 47 eclipsed Met χ_3 cases studied in detail, the results are as follows. In 13 instances the reported torsion angle is within 10° of an eclipsed value. Of these, the non-bonded energy profile has its minimum within 5° in nine instances, between 5 and 10° in one case and between 10 and

15° in three instances. The remaining 34 cases divide into three groups as follows: in 22 cases the external torque has the correct sign to compensate an internal torque corresponding to the deviation from optimal values, in eight cases the external torque has the wrong sign but the non-bonded energy has its minimum within 5°, while in four cases the torque has the wrong sign and the minimum lies farther than 5° away. Overall, side chains for which the agreement between packing forces and deviation from canonical values is unsatisfactory are relatively more exposed to solvent and in these cases agreement is expected to be poor because the calculations ignore interactions with solvent and forces resulting from crystal packing.

Methionine side chains tend to be buried in the protein interior where packing is quite tight and thus the observation that the computed non-bonded energy of the ε-methyl groups
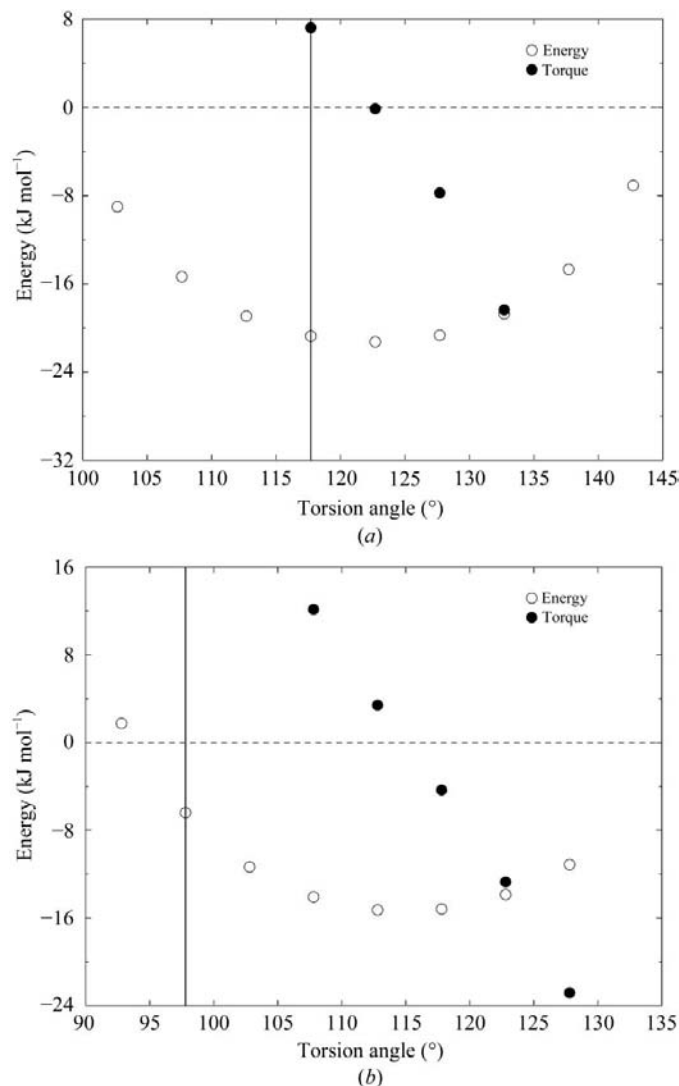
Figure 8
Examples of non-bonded energy and torque (in kJ mol^{-1} rad^{-1}) for χ_3 of well confined methionine residues. (*a*) A methionine in a near-eclipsed conformation (1bu7 Met112); (*b*) A side chain with a steep internal gradient (1a4i Met282). The vertical line indicates the experimental values, $\chi_3 = 118$ and 98°, respectively.

of many methionine side chains has a minimum within a 5° change of χ_3 is in theory unsurprising. In practice, the computed positions of the minima are inaccurate because of the rapid variation of non-bonded forces with interatomic distance and the inaccuracy of molecular-mechanics approximation, especially without explicit consideration of H atoms, while Met C^ε positions are quite unreliable at higher B values. However, the predominance of torques favoring reported large distortions away from the mean rotamer conformations is strong evidence that almost all the large deviations for these well defined cases are real and do not correspond to errors in interpretation of diffraction results.

To cross-validate these conclusions, the 47 examples were automatically analyzed to see whether any other conformation, either slightly shifted or very different, could yield a better combination of all-atom contact score and rotamer score. 35 of the 47 fit their surroundings best in the reported position; only five were strongly improvable and another seven marginal, in contrast to similar tests for C—C tetrahedral side-chain torsions where the majority of near-eclipsed cases are suspect. Of the seven cases with unsatisfactory torques or minima, six also fail the contact/rotamer test and are thus likely to be incorrect (unfortunately, only one had deposited structure factors). Both methods agree that low-B examples of eclipsed Met χ_3 are predominantly correct and are confined or forced into that position by the rest of the protein against the relatively low rotational barrier of the C—S bond.

4. Discussion

In this paper, we seek to make two main points. Firstly, the paper shows that the effects arising from experimental errors can effectively be eliminated to give only the intrinsic variation of structural features of proteins by using only residues with low B values and specifically that the intrinsic variation of side-chain torsion angles can be assessed by using only atoms with $B \leq 20 \text{ Å}^2$, essentially independently of the overall crystal resolution or the temperature at which the structure was solved. Secondly, the paper shows that model potential energies correlate with details of distributions of torsion angles and consequently that the energetics of highly unusual details can serve as a filter to discard erroneous interpretations of the diffraction data.

4.1. Relation to energies of model compounds

The conformations of protein molecules in solution correspond to minima of the free energy and, because the structures are highly organized, the conformations correspond closely to minima of the enthalpy. Mean torsion angles observed in protein structures for preferred conformations are shown here to correlate very well with minimum-energy conformations calculated for isolated residues. The specific conformation of an element within a protein, such as a given side chain, deviates from the minimum-energy conformation of the isolated element to the extent that the deviation causes a lower overall energy. One sees from the intrinsic χ MSD values in Table 3 that these individual deviations are almost all very small (e.g. an RMSD of 9° for unbranched χ_1).

The deviations reflect differences in the environment of the side chain for different protein architectures. Small deviations are more common than large ones and we have found the widths of the distributions of the dihedral angles of the rotamers to be closely related to the ease (expressed in terms of a small energy increase) of torsion of isolated simple model compounds. For instance, the different energy barriers calculated for C—S and C—C bonds agree with empirical distribution widths of CC—CH_3 and CS—CH_3 torsion angles for well ordered residues. Also, the observed strong preference for the staggered conformation of the former and the *gauche* conformation of the latter case (Fig. 5), which was earlier inferred to be owing to attractive van der Waals contacts of the H atoms (Word *et al.*, 1999), is reproduced by high-level quantum calculations (Butterfoss & Hermans, 2003).

4.2. Relevance to crystallographic structure refinement

The analysis presented here allows the distinction between two causes of the observed variation of the extent to which side-chain dihedral angles of protein structures deviate from canonical values, one extrinsic to the structure, arising from disorder and various errors in the experimental data or fitting, and the other intrinsic to the structure, arising from packing forces. For the C^α—C^β bond, the
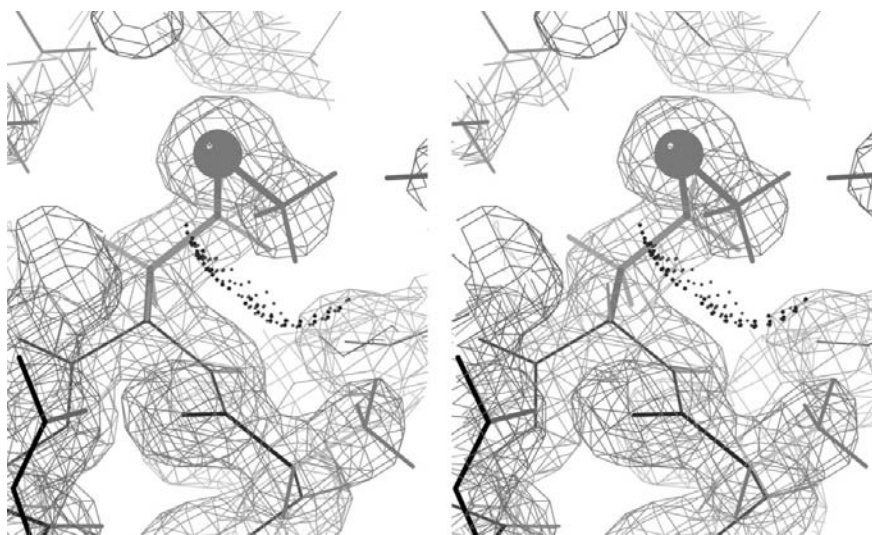
Figure 9
Stereoview of methionine residue 1a4i Met282 at 1.5 Å resolution (Allaire *et al.*, 1998), showing model $2F_o - F_c$ electron-density contours at 1.2σ and 3σ and all-atom contact dots between the side chain and its surroundings. C^ε is accurately positioned by clear electron density and is kept away from *gauche* χ_3 by contacts with its own CO group, in this case part of an α-helix.

component arising from disorder and error is found to become negligible with respect to the intrinsic component for well ordered side chains with B values below 20 Å^2. Furthermore, the contribution owing to thermal motion is very small and the effect of resolution on deviation, although large, is almost entirely caused by concomitant changes in B values.

The number of independent observations (structure factors) changes as the inverse third power of the resolution and thus the ratio of observations to unknowns (coordinates and thermal parameters) and the effect on both the statistical and fitting errors changes rapidly with resolution. From the data of Fig. 1 it can be estimated that the variance arising from all errors and disorder is of the same magnitude as the intrinsic variance for a resolution of 2 Å. This suggests that refinement of structures determined at poorer resolution than ~2 Å should proceed with restraints not only of bond distances, bond angles and planar groups, but also of torsion angles about single bonds. Care will be required to not overrestrain the torsion angles; this means that the intrinsic variation of individual types of torsion angles should be considered, with the data of Table 3 as a guide. Also, the stage or method of refinement must be capable of searching the multiple minima.

The database contains many instances of incorrect details in otherwise correctly determined protein model structures. In addition to overall measures such as resolution and free R value, validation of structure details typically checks a series of features such as combinations of successive backbone or side-chain torsion angles, non-bonded contact distances, bond angles, hydrogen-bond geometry, correlation with local electron density *etc.* (Laskowski *et al.*, 1993; Hooft *et al.*, 1996; Lovell *et al.*, 2003; Westbrook *et al.*, 2003). An important additional criterion is a comparison with details typically seen in structures of other proteins. Thus, a feature seen many times before is deemed acceptable, while a novel structural arrangement must be examined very carefully to ensure that it is a valid interesting exception and not just a qualitative error in fitting to the electron density.

It is understood that common details such as near-canonical C—C torsion angles, optimal non-bonded contacts between non-polar groups and linear hydrogen bonds correspond to energetically favorable arrangements and this is well illustrated by results presented in this paper. We believe a significant result of the work presented here and an earlier study (Butterfoss & Hermans, 2003) to be that they lead to a quantitative empirical relation between the prevalence of a particular structure feature and its intrinsic energy cost.

A specific example is offered by the distributions of the χ_3 torsional angles in methionine and lysine. The intrinsic variation of the torsion angle for the C^α—C^β bond corresponds to an RMSD of 9°; the energy of an eclipsed C^α—C^β bond ($\chi = \pm 120°$) is 13.8 kJ mol^{-1} above that of the staggered conformation. C—S bonds deviate on average farther from their canonical values; the energy of an eclipsed C—S bond ($\chi = \pm 120°$) is only 7.5 kJ mol^{-1} above that of the staggered conformation. A small but significant percentage of χ_3 torsional angles for the C—S bond in well ordered methionine residues deviate by a full 60° from the canonical values of 60,

180 and −60°, but for χ_3 of lysine such large deviations form a smaller percentage and are seldom supported by unambiguous electron density.

An energy penalty of 7.5 kJ mol^{-1} appears to be acceptable, but not one nearly twice as large. In fact, a more detailed analysis given elsewhere (Butterfoss & Hermans, 2003) has indicated that the statistical occurrence of a higher energy conformation varies exponentially with the energy penalty as $\exp(-\beta E)$, with $1/\beta \simeq 2.5$ kJ mol^{-1} (the Boltzmann form of a distribution). By extrapolation, an energy penalty of 7.5 kJ mol^{-1} is then likely to occur with a probability of 5%, while for a penalty of 13.8 kJ mol^{-1} the probability is only 0.4%. Details having a high intrinsic energy strongly suggest errors in the model (such as an error of 180° in fitting χ_1 of a valine side chain) unless compensated by several hydrogen bonds or other favorable interactions in the structure.

As more protein structures are solved, occasionally a novel structural feature will turn up. The absence of precedents should lead to a careful scrutiny of the validity of the particular interpretation of the electron density and it is here that consideration of energetic contributions favoring the unusual arrangement may turn out to be especially useful, firstly by using the intrinsic energy as a measure of an *a priori* probability and secondly by identifying factors that stabilize the higher energy conformation.

References

Allaire, M., Li, Y., MacKenzie, R. E. & Cygler, M. (1998). *Structure*, **6**, 173–182.

Allinger, N. L., Fermann, J. T., Allen, W. D. & Schaefer, H. F. III (1997). *J. Chem. Phys.* **106**, 5143–5150.

Brünger, A. T., Clore, G. M., Gronenborn, A. M. & Karplus, M. (1987). *Proc. Natl Acad. Sci. USA*, **83**, 3801–3805.

Butterfoss, G. & Hermans, J. (2003). *Protein Sci.* **12**, 2719–2731.

Dunbrack, R. L. & Karplus, M. (1993). *J. Mol. Biol.* **230**, 543–571.

Frisch, M. J. *et al.* (1998). *Gaussian*98. Gaussian, Inc., Pittsburgh, PA, USA.

Hermans, J., Berendsen, H. J. C., van Gunsteren, W. F. & Postma, J. P. M. (1984). *Biopolymers*, **23**, 1513–1518.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Proteins Struct. Funct. Genet.* **26**, 363–376.

Jensen, L. H. (1997). *Methods Enzymol.* **276**, 353–366.

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

Kleywegt, G. J. (1999). *Acta Cryst.* **D55**, 1878–1884.

Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins Struct. Funct. Genet.* **50**, 437–450.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *Proteins Struct. Funct. Genet.* **40**, 389–408.

Mann, G., Yun, R. H., Nyland, L., Prins, J., Board, J. & Hermans, J. (2002). *Computational Methods for Macromolecules: Challenges and Applications. Proceedings of the 3rd International Workshop on Algorithms for Macromolecular Modelling, New York, October 12–14, 2000*, edited by T. Schlick & H. H. Gan, pp. 129–145. Berlin: Springer-Verlag.

Mowbray, S. L., Helgstrand, C., Sigrell, J. A., Cameron, A. D. & Jones, T. A. (1999). *Acta Cryst.* D**55**, 1309–1319.

Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 243–252.

Westbrook, J., Feng, Z. K., Burkhardt, K. & Berman, H. M. (2003). *Methods Enzymol.* **374**, 370–385.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, P. K., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1711–1733.