Main-chain Bond Lengths and Bond Angles in Protein Structures

Roman A. Laskowski[†]**, David S. Moss**

Crystallography Department, Birkbeck College
Malet Street, London WC1E 7HX, England

and

Janet M. Thornton

Biomolecular Structure and Modelling Unit
Department of Biochemistry and Molecular Biology
University College, Gower Street, London WC1E 6BT, England

(Received 3 September 1992; accepted 19 February 1993)

The main-chain bond lengths and bond angles of protein structures are analysed as a function of resolution. Neither the means nor standard deviations of these parameters show any correlation with resolution over the resolution range investigated. This is as might be expected as bond lengths and bond angles are likely to be heavily influenced by the geometrical restraints applied during structure refinement. The size of this influence is then investigated by performing an analysis of variance on the mean values across the five most commonly used refinement methods. The differences in means are found to be highly statistically significant, suggesting that the different target values used by the different methods leave their imprint on the structures they refine. This has implications concerning the actual target values used during refinement and stresses the importance of the values being not only accurate but also consistent from one refinement method to another.

Keywords: protein structure; bond lengths; bond angles; refinement methods; stereochemical parameters

1. Introduction

An accurate knowledge of standard bond lengths and bond angles is of great importance in the determination and refinement of protein structures. "Ideal" or "target" values for these geometrical parameters are used to supplement the experimental data obtained from either X-ray crystallography or NMR studies; in effect this increases the number of experimental observations relative to the number of parameters being determined (the latter being the atomic co-ordinates and temperature factors).

The target values are typically obtained from crystallographic studies of small molecules, the data nowadays being taken from the Cambridge Structural Database, CSD (Allen *et al.*, 1979), which holds over 80,000 structures. Standard bond lengths and bond angles have been tabulated in many sources (e.g. see Kennard, 1968; Allen *et al.*, 1987). The most recent analysis of the structures in the CSD has produced an updated set of bond lengths and bond angles specifically for use in protein refinement (Engh & Huber, 1991).

One can perform similar analyses on proteins using the corresponding database of protein structures, namely the Brookhaven Databank (Bernstein *et al.*, 1977). However, in this case, not only are there limitations on the statistical analyses that can be performed but one also needs to be very careful about how the results are interpreted.

In the first place, the diffraction data obtained from protein crystals are much poorer than those obtained in small-molecule studies. The principal reason for this is that protein crystals diffract relatively weakly as they contain a large proportion, around 30 to 70%, of solvent (usually water). Data are typically obtained to around 1·7 to 2·5 Å resolu-

[†] Present address: Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England.

0022–2836/93/121049–19 $08.00/0

tion, whereas small-molecule data are typically in the range 0·8 to 1·0 Å, and can even reach the edge of the Cu sphere at 0·77 Å.

For proteins, therefore, at the resolutions achieved, the numbers of useful reflections obtained are generally of the order of the numbers of parameters being determined. Thus, when protein structures are refined it is necessary to apply restraints, such as target bond lengths and bond angles, to increase the number of observations over the number of parameters and so assist the refinement. However, the application of restraints introduces a bias which may be maintained through to the final structure. Hence, when analysing bond lengths and bond angles in the known protein structures the interpretation of the results is made difficult as the values obtained will be influenced by any biases that remain.

A second problem is the limitations on the statistical analyses that can be performed. Firstly, there are only around 600 proteins structures currently available, and very few have been solved independently (though a small number have been independently refined, see Hubbard & Blundell, 1987). So there are very few independently obtained structures that can be compared against one another. Secondly, published protein structures do not provide estimated standard deviations (e.s.d.'s†) for their atomic co-ordinates and *B*-values. So there is no indication of the size of the random errors in the cited figures, which the e.s.d.'s would provide. For small molecules, e.s.d.'s are routinely obtained during refinement from the least-squares covariance matrix (Cruickshank, 1965) being a part of the standard output of the SHELX refinement programs (Sheldrick, 1976, 1985, 1986; Robinson & Sheldrick, 1988). From the e.s.d.'s it is possible to estimate errors in bond lengths and bond angles as well as other geometrical properties.

For proteins, the calculation of e.s.d.'s is a much more difficult procedure. The larger size of the molecules means that the calculation of the least-squares covariance matrix has generally been considered too computationally intensive and requiring of too much computer memory to be routinely performed. It requires firstly setting up and then finding the inverse of the *full* normal equations matrix (i.e. not just a block-diagonal approximation) used in least-squares refinement; for proteins this can be several thousand elements square. Nowadays, with the advent of faster processors having larger amounts of RAM such calculations are becoming feasible (Laskowski, 1992). However, the calculations are further complicated by the fact that the e.s.d.'s of the atoms in the structure are not all independent, being biased by the restraints applied, and this affects the absolute values obtained (T. J. Tickle, R. A. Laskowski & D. S. Moss, work in progress).

The limitations just described prohibit, for example, the type of analysis performed, on small-molecule structures, by Taylor & Kennard (1983, 1985) in estimating average molecular dimensions or in detecting systematic errors in structures (Taylor & Kennard, 1986). They also prohibit one from addressing fundamental questions such as: what are the "true" bond lengths and angles in protein structures and the "true" deviations?

The analyses presented in this paper consider the main-chain bond lengths and bond angles of protein structures. As mentioned above, the most serious obstacle to a simple interpretation of the results is that of the likely biases introduced into the structures by the restraints applied during refinement. In least-squares refinement, bond length and bond angle restraints are applied as additional terms to the function being minimized. They are of the form:

$$\sum_{k=1}^{\text{Distances}} w_{dk}(d_{k0}-d_k)^2,$$

where d_k and d_{k0} are the actual and target distances, and w_{dk} is the weight applied to each restraint. The restraints do not *fix* the values (unlike constraints which *do*), but tend to "pull" the actual values towards the targets. Note that the expression above applies equally well to bond angles as these are commonly restrained by means of *distance* restraints between atoms 1 and 3 for an angle defined by atoms 1, 2, and 3. Angle restraints are often applied in this way to reduce the amount of computation involved.

In refinement methods that use energy minimization, similar terms appear in the expression for the overall energy. In molecular dynamics refinement, the restraints appear in the potential energy function that describes the forces between atoms.

As well as bond lengths and bond angles, other stereochemical parameters are sometimes used as additional restraints. These include deviations of atoms from a least-squares plane, torsion angles, occupancy factors, preservation of chirality, and the prevention of close contacts between non-bonded atoms (Hendrickson, 1985). Which parameters are restrained will differ from one refinement method to another.

In theory, the influence of the restraints should diminish as refinement proceeds and as the *R*-factor improves. This should be particularly true of high-resolution structures, where the quality of the data is high. Thus, by the end of the refinement process, the information from the experimental data (i.e. the observed structure factor amplitudes) should dominate the restraints applied. Nevertheless, there may still be a remnant bias in the final structures which will affect any statistical analyses.

Morris *et al.* (1992) analysed a number of geometrical parameters that are *not* usually restrained during refinement. The authors found that certain of these parameters correlate well with resolution, and reported some correlation with *R*-factor. It was concluded that some of the parameters examined can provide reasonable measures

[†] Abbreviation used: e.s.d.'s, estimated standard deviations.

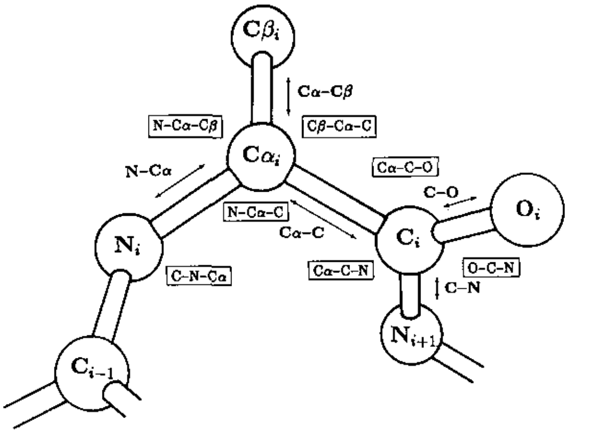


Figure 1. Main-chain bond lengths and bond angles (boxed).

of the stereochemical "quality" of a protein structure. The authors went on to use their results to propose a "stereochemical quality index", and to classify all the protein structures then available. Indeed, many of their observations have recently been incorporated into a program called PROCHECK, which provides a means of assessing the stereochemical quality of a given protein structure (Laskowski *et al.*, 1993).

In this paper we extend the analyses of Morris *et al.* (1992) to the consideration of bond lengths and bond angles, and then explore the extent of the biases introduced by restraints during refinement. The analysis presented here is limited to the five main-chain bond lengths and seven main-chain bond angles, shown in Figure 1. The means and standard deviations of each of the bond lengths and bond angles are computed as a function of resolution for all proteins in our data set. As expected, there is little variation of these properties with resolution, most probably because of the influence of the restraints. The only slight correlation is for the standard deviations of the bond angles, but this disappears when the data set is reduced to only the high-resolution structures, perhaps because bond angles are more likely to have been restrained amongst these structures.

The data set is further reduced to remove some of the outlying structures before comparing the mean figures with the small-molecule data of Engh & Huber (1991). The removal of the outliers is achieved by calculating a "variance score" for the bond angles of each structure and disposing of those whose score is a long way from the overall mean score.

We then investigate the influence of the restraints. This is done by grouping the structures according to their refinement method and performing an analysis of variance on the mean values from the five most commonly used methods (see Table 2 for references). For this analysis, the side-chain bond lengths and angles are considered along with the main-chain ones, giving a total of 46 different bond length types and 65 different types of bond angle (or, when broken down by individual residue types: 173 different bond-lengths and 234 different bond angles). Different refinement methods have different dictionaries of target values for these parameters. For example, the most usual parameters for X-PLOR come from the CHARMM parameter set (Brooks *et al.*, 1983); PROLSQ and TNT supply co-ordinate sets of ideal fragments, the latter from data tabulated by Bowen *et al.* (1958) and Vijayan (1976); EREF uses the Levitt potential-energy function (Levitt, 1974); and RESTRAIN supplies a dictionary of standard bond lengths and distances for bond-angle restraints.

While it is true that in most cases the supplied parameter dictionaries can be altered by individual users, we expect that most, if not all, users of a given refinement method have used the same dictionary of values. Thus, any significant differences in the mean values across the refinement methods can be attributed to their different target values and hence point to a remnant bias in the final structures.

We find that there are indeed detectable differences in the structures, implying that the targets do leave their mark. We discuss the implications of this finding for protein refinement.

2. Materials and Methods

(a) *Protein database*

For the analyses performed here, the protein structure co-ordinates were taken from the July 1991 release of the Brookhaven database (Bernstein *et al.*, 1977). Proteins for which only C$^\alpha$ co-ordinates are supplied and for which no resolution was given were excluded. This left a total of 523 proteins, listed in Table 1.

The initial analyses were performed on this complete data set. However, this set contains many poorly resolved structures, so the analyses were repeated on two reduced data sets aimed to exclude these structures.

The first reduction involved using the standard rule-of-thumb method for selecting well-resolved, well-refined structures: namely those having a resolution of 2·0 Å or better, and an R-factor no greater than 20%. This gave a reduced data set of 221 protein structures which can be identified by the highlighted Brookhaven codes (underlined or **bold-faced**) in the appropriate region of Table 1.

The second reduction involved removing all outlying structures: those having either a very large or a very small variability in their bond lengths and bond angles due to either very weak or very strong restraints applied during refinement, respectively. This variability was measured for each protein by taking the standard deviation, σ_i, of each of the 7 main-chain bond angles and calculating a variance "score", s, as follows:

$$s=\frac{\sqrt{\sum_{i=1}^{n}\sigma_i^2}}{n},$$

where n is the number of bond angle types (i.e. $n=7$). A similar score was also computed for the main-chain bond lengths (here $n=5$).

Outlying proteins were deemed to be those whose score was more than one standard deviation from the mean score for all 221 proteins. Very small scores indicated very

Table 1

Brookhaven codes of the 523 proteins used in the analysis

Resolution (Å)

R-factor	<1.0	1.0-1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
<0.11	1gma	1xy1 1xy2		6rxn 7rxn					
0.11		5rxn					7pcy		
0.12		4rxn		2sga			1sgc 3sgb		2lst
0.13						2alp 4hhb	2sec 3app 3rnt		1p01
0.14					2er7 2rhe		2apr 2rnt 2st1 3apr	1mbw 2gbp	1p02 2rsp 1ppd 3lym 2lhb 3mba 2lym 4er1
0.15		7rsa		1amt 4ins 5cyt		1l17 1s01 1l18 2mhr 1l19 3dfr 1l23 3lsm 1l24 4dfr 1l33	1l27 1l36 2aza 3cla 3er5 8pti	1l20 1l28 6pcy	2mba 2paz 5rsa 5tnc
0.16			256b	1tld 2prk 3b5c 4pti	1cdp 3est 9pap	1l03 6pti 1l22 1l30 1l32 1sgt 1snc 2hhb	1cho 1hne 1omd 2pcy 5er2 5pcy 8rsa	3pcy	1st2 2cts 3blm 5ebx
0.17		1cse 4ptp	3ebx	1ls1 1mbc	1pcy 1psg 2tmn 5tmn 6tmn 7pti	1bp2 1l15 1ctf 1l16 1fkf 1l29 1l06 1snm 1l07 2act 1l09 2ltn 1l10 3c2c 1l12 3tmn 1l13 4tmn 1l14 5cha	1gd1 1l31 1ubq 2cdv 2cga 4pep 9wga	1l21 1l34 1tmn 1tpa 2ca2 3csc	1ca2 1gp1 1hmq 1hms 2c2c 2i1b 4mbn 4tnc 5mbn 7wga
0.18				1pas 3grs 3ins	2wrp 451c 5cpv	1csc 1l26 1l01 1tgt 1l02 1tpo 1l04 2ccy 1l05 2csc 1l08 3ptb 1l11 8dfr	1l25 1ntp 1tgc 1tgs	2fb4 2gch 2ptc 3bcl 4csc 5cts	1gox 4mba 1i1b 6rsa 1rbp 7cpp 1rsm 2mcg 3er3 4er2
0.19		1ycc	1tpp	1ccr 1thb 2ovo 2ptn	1mba 2cpp 2utg 351c	2lsm 3cts 3ptn	1ton 2tga 7gch 9rsa	1rnt 3cpp 3rp2 3tpi 6cpp	1hoe 2fbj 1lyd 2mlt 1r69 3ca2 2cab 4i1b 2ci2 5cpp
0.20		5pti				2cyp 2tgt 3hhb	1srn 3wrp 4fxn 6cha	2tgp 3gch 4gch	2er6 3mcg 5er1 6ldh
0.21				4cpv	3tln			1ypi 3fxn 4fd1	
0.22					1tgn	1alc			2pka
0.23		1utg			1gcr	4cha		1fd2 2fd2	
0.24			1nxb						1hip
0.25				1rn3			2pab		1lzt 2sod
0.30+			1rdg						
None given			1eca 1ecd 1ecn 1eco 1mbd 1ppt	1crn 2sns 3rxn 5cpa	1mbo		1sn3 1tgb 2mb5 3cyt	1ovo	1acx 1lh5 2cna 2lyz 1fdx 1lh6 2lh1 2mhb 1fx1 1lh7 2lh2 3cpa 1hds 1lyz 2lh3 3fab 1lh1 1mbn 2lh4 3lyz 1lh2 1rei 2lh5 4lyz 1lh3 1rns 2lh6 5lyz 1lh4 2cha 2lh7 6lyz

Table 1 *(continued)*

Resolution (Å)

R-factor	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1 -3.2	3.5
<0.11									2tmv			
0.11					6cts							
0.13	1p05	1p09	1p08			1p04						
0.14	4sgb 5apr	1p03 2ccp	1p06 1p10 2gcr		5xia 6apr							
0.15	2sni 4ape	1ccp 2wgc 4pcy	1p07 4xia	1sdh 2cla 3gbp	4apr		5gch	2trm				
0.16		3ccp 4ccp	1cms 1trm 4tln	1pfk 2pfk 2sdh 4pfk	4icd 4mdh 8atc	3hla 4at1 5at1 6at1	2phh	1at1 8at1		4rhv		
0.17	1ldm 3bp2 8cpp	1tec 1wgc 2er9 3cln 4tpi	1cla 1dhf 1tlp 3icb 3pep 4hvp 5tln 7tln	2liv 3pfk	1pp2 2gd1 2or1 2ypi	2hla		2at1				1mcw
0.18	2tgd 3p2p 4cpp 6gch			1fcb	1pmb 2ins 3icd 4ts1 6acn		1cts	3at1 3hvp 7at1	4cts		2mcp	
0.19	3adk 4er4	2lz2	1cd4 1phh 1prc 2dhf 5dfr 5pep	2cro 6dfr 8adh	1cdt 8cat		5ldh	5csc 8ldh		2fnr 7api		
0.20	2tpi	1wrp	2fxb		1sbc				2plv	2ig2 9api		
0.21	5acn		2gn5	2lbp	1rbb 2est 7cat		3ts1 4fab	4gpd		2bp2	8api	
0.22	1hho		2ts1	2cd4	2kai		1mcp	2aat		2hmg 2mev	5hmg	
0.23						1tnf			3hmg	4hmg		
0.24					2abx 1lrd 2hfl 7dfr	1p2p				3hfm 3xia		
0.25					1wsy 3gap			4sbv		1hbs 2ldx	2gls	
0.26		1fnr			1lym					2ldb	1cn1	
0.27										2atc		
0.28								1ldb		2er0		
0.29					1cc5			3pgm			1r1a 7adh	
0.30			1fxb		3fxc 1cy3		1azu	1f19		1bmv 1llc	1pfc	3gpd
0.40+					1chg							
None given	2yhx		1cyc	1abp 3cna	155c 1est 1fdh 1mbs 1rhd 1sbt 1tim 2stv 3pgk 4cpa 7lyz 8lyz	2ssi	1hco 2hco	1ctx 1fc2 1pad 2dhb 2pad 2sbt 4pad 5pad 6pad	1coh 1etu 1fc1 1gpd 2tbv 5adh 6adh	1gcn 1pyp 1r08 1rmu 2r04 2r06 2r07 2rm2 2rmu 2rr1 2rs1 2rs3 2rs5 2taa 3ldh		1brd 1hkg

The codes are tabulated in ranges of resolution value and *R*-factor. The codes shown in **bold-face**, together with those underlined, correspond to the data set of 221 high-resolution structures. The underlined codes are the 35 structures removed to give the set of 186 "best" structures (see the text). The codes shown in **bold-face** thus correspond to the 186 "best" structures.

little variation in the main-chain angles, suggesting tight restraints had been applied during refinement. Large scores, on the other hand, suggested little or no refinement. Thus, the proteins that had been very tightly restrained were rejected as well as those that had been loosely restrained. The aim here was to exclude structures that might be more influenced by the geometrical restraints than by the experimental data.

The reduced set gave the 186 "best" structures. These are shown in **bold-face** in Table 1, while the removed proteins are shown underlined in the same Table.

In the analysis, dummy (i.e. zero occupancy) atoms and atoms with high temperature factors were not excluded, though strictly speaking they should have been. Nevertheless, because the analysis was concerned with main-chain, rather than side-chain, atoms their influence on the statistics would have been negligible: the zero occupancy atoms numbered only 6 in the whole data set, while only around 3% of the atoms in the whole data set, and 0·85% atoms in the reduced data set, had B-values larger than 50·0.

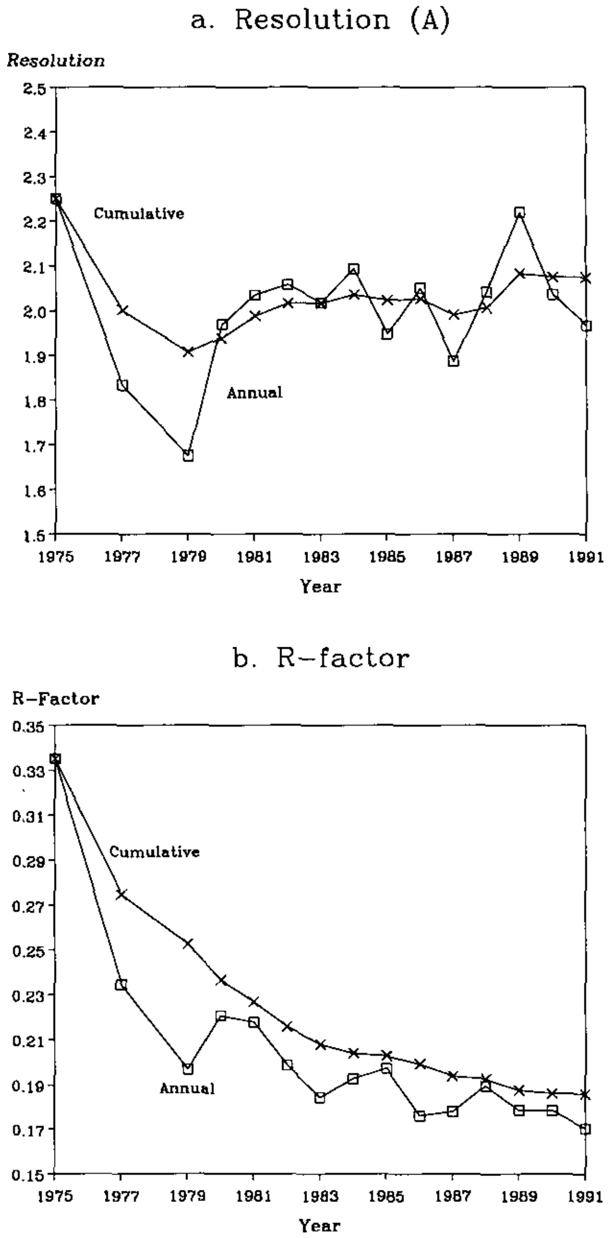


Figure 2. Year-by-year averages and cumulative averages for the resolution (a) and R-factor (b) values of proteins submitted to the Brookhaven Databank.

(b) *Testing the influence of refinement method*

To see whether the refinement method used biases a structure's bond lengths and bond angles, an *analysis of variance* was performed on the main-chain and side-chain bond lengths and bond angles from structures refined by the 5 most common methods, namely:

PROLSQ	Hendrickson/Konnert	198
TNT	Tronrud/Ten Eyck	59
EREF	Jack/Levitt	39
X-PLOR	Brünger *et al.*	23
RESTRAIN	Driessen/Moss	15.

The numbers on the right show the numbers of structures involved, and Table 2 lists the structures themselves and the references to the relevant literature on each refinement method.

For the analysis of variance the "model" used was a "one-way classification fixed-effects model" (see, for example, Milton & Arnold, 1986). It was "one-way" because we were interested in the influence of a single factor (i.e. the refinement method) on the observed variation in bond lengths and bond angles. The model was a "fixed-effects" model as the refinement methods considered had been deliberately selected, rather than chosen at random from many possible methods. Using this model, we made the null hypothesis that the mean value of a given parameter is *not* affected by the refinement method. That is, the mean values are the same:

Null hypothesis is H_0: $\mu_1=\mu_2=\ldots=\mu_5$.
Alternative hypothesis H_1: $\mu_i \neq \mu_j$ for some i and j (i.e. at least 2 of the means are not equal),

where $\mu_1,\mu_2,\ldots,\mu_5$ are the means of a given property (i.e. given bond length or bond angle) for the 5 methods.

Analysis of variance aims to determine how much of the variability in a given property is attributable to the "treatment" (in this case, the refinement method), and how much to the random fluctuations among the values *within* each treatment. Measures of these variabilities are provided by the sum of squares identity:

$$SS_{Tot}=SS_{Tr}+SS_E,$$

where SS_{Tot} is the total sum of squares, SS_{Tr} is the treatment sum of squares, and SS_E is the residual or error sum of squares. The 3 terms in the identity are calculated as follows:

$$SS_{Tot}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{..})^2,$$

$$SS_{Tr}=\sum_{i=1}^{k}n_i(\bar{X}_{i.}-\bar{X}_{..})^2,$$

$$SS_E=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2,$$

where X_{ij} is an individual measure, being the jth instance of a given bond length or angle for the ith refinement method, n_i is the number of the given bond lengths or bond angles for the ith refinement method, and k is the number of different refinement methods. $\bar{X}_{i.}=\Sigma_{j=1}^{n_i}X_{ij}/n_i=$ mean value for a given refinement method; $\bar{X}_{..}=\Sigma_{i=1}^{k}\Sigma_{j=1}^{n_i}X_{ij}/N=$ mean of all values; $N=\Sigma_{i=1}^{k}n_i$.

Table 2

Structures refined by the five most common refinement methods

A. Hendrickson/Konnert – PROLSQ									
1amt	1ldm	1phh	**1ubq**	2cyp	2mcp	2ypi	3rnt	4xia	**6rxn**
1azu	1lrd	1pmb	1utg	2dhf	2mev	**3app**	**3sgb**	5apr	7cat
1cc5	**1lyd**	1pp2	1wgc	2est	**2mhr**	**3apr**	3ts1	5cha	**7cpp**
1ccp	**1lz1**	**1ppd**	1wrp	**2fbj**	**2mlt**	**3b5c**	3wrp	5cpp	7dfr
1ccr	**1mbc**	1r1a	1wsy	2fxb	2paz	**3blm**	3xia	**5cpv**	**7gch**
1cdp	**1mbw**	**1r69**	**1ycc**	**2gbp**	**2pcy**	3bp2	4apr	5dfr	7pti
1cho	1mcp	1rdg	**256b**	2gd1	2phh	3c2c	4ccp	**5ebx**	**7rsa**
1cla	1mcw	**1rnt**	2aat	2gls	**2prk**	3ccp	4cha	5gch	**7rxn**
1cms	1nxb	**1rsm**	**2alp**	2gn5	2rsp	3cln	4cpp	**5pti**	**7wga**
1f19	**1p01**	**1s01**	**2apr**	2hfl	2sdh	**3cpp**	4cpv	**5rsa**	8atc
1fcb	**1p02**	1sbc	2atc	**2i1b**	**2sec**	**3ebx**	4fab	**5tnc**	8cat
1fxb	1p03	1sdh	2bp2	2lbp	**2sga**	3fxc	4gch	5xia	8cpp
1gd1	1p04	**1sgc**	2c2c	2ldb	2sni	3gch	4gpd	6apr	**8dfr**
1gox	1p05	**1sgt**	2ccp	2ldx	**2st1**	3gpd	**4ins**	6cha	8ldh
1hbs	1p06	**1snc**	2ccy	**2lhb**	2tmv	3icb	**4pep**	**6cpp**	**8rsa**
1hmq	1p07	1snm	**2cdv**	2liv	2trm	**3ins**	**4ptp**	6dfr	**9pap**
1hmz	1p08	**1st2**	**2ci2**	**2lym**	2ts1	**3lym**	4sbv	6gch	**9rsa**
1hne	1p09	1thb	2cla	2lz2	2utg	3mcg	4sgb	6ldh	**9wga**
1i1b	1p10	**1ton**	**2cpp**	**2lzt**	2wgc	**3pcy**	4tnc	**6pti**	
1ldb	1pfk	1trm	2cro	2mcg	**2wrp**	3pgm	4ts1	6rsa	

B. Tronrud/Ten Eyck – TNT									
1csc	**1l05**	**1l11**	**1l17**	**1l23**	**1l29**	**1l35**	3adk	**3tmn**	**5cts**
1fnr	**1l06**	**1l12**	**1l18**	**1l24**	**1l30**	**1psg**	**3bcl**	4csc	**5tmn**
1l01	**1l07**	**1l13**	**1l19**	**1l25**	**1l31**	1tlp	**3csc**	**4mba**	6cts
1l02	**1l08**	**1l14**	**1l20**	**1l26**	**1l32**	**2csc**	**3grs**	4mdh	**6tmn**
1l03	**1l09**	**1l15**	**1l21**	**1l27**	**1l33**	2fnr	**3lzm**	**4tmn**	7tln
1l04	**1l10**	**1l16**	**1l22**	**1l28**	**1l34**	**2mba**	**3mba**	5csc	

C. Jack/Levitt – EREF									
1cse	**1hoe**	**1tld**	**1tpp**	2ig2	**2ptc**	**3est**	**3tpi**	4tpi	8api
1cts	1llc	**1tmn**	**2cga**	2kai	2ptn	**3ptb**	4cts	5cyt	9api
1gp1	**1mba**	**1tpa**	**2cts**	**2ovo**	**2tgp**	3ptn	4mbn	**5mbn**	
1hho	**1tgs**	**1tpo**	**2hhb**	2pka	**3cla**	**3rp2**	**4pti**	7api	

D. Molecular dynamics – X-PLOR									
1at1	2at1	2hla	3gbp	3icd	4hmg	5acn	5hmg	6at1	8at1
1fd2	2cd4	2hmg	3hmg	4at1	4icd	5at1	6acn	7at1	**8pti**
1fkf	2fd2	3at1							

E. Driessen/Moss – RESTRAIN									
1gcr	2er0	**2er7**	2gcr	**3er5**	**4er2**	4er4	5er1	5er2	5pep
1rn3	**2er6**	2er9	3er3	**4er1**					

A, Hendrickson/Konnert, PROLSQ (Konnert, 1976; Hendrickson & Konnert, 1980; Hendrickson, 1985); B, Tronrud/Ten Eyck, TNT (Tronrud *et al.*, 1987); C, Jack/Levitt, EREF (Jack & Levitt, 1978); D, Molecular dynamics, X-PLOR (Brünger *et al.*, 1987); E, Driessen/Moss, RESTRAIN (Moss & Morffew, 1982; Haneef *et al.*, 1985; Driessen *et al.*, 1989). Codes shown underlined and in **bold-face** are as for Table 1.

If the refinement methods have an effect on the mean values, then SS_{Tr} should be large relative to SS_E. To test whether the difference in the 2 is significant an F-test is performed on the ratio MS_{Tr}/MS_E, where:

$$MS_{Tr}=\frac{SS_{Tr}}{(k-1)},$$

$$MS_E=\frac{SS_E}{(N-k)}.$$

The F-test is a right-tailed test using an F-distribution with $k-1$ and $N-k$ degrees of freedom. If the ratio MS_{Tr}/MS_E is greater than the tabulated value of $F_{k-1,N-k}$ at a given level of confidence, then the null hypothesis H_0 is rejected with the corresponding degree of confidence and one concludes that at least one of the methods has an influence on the geometrical property in question.

For the analysis, each different main-chain and side-chain bond length and bond angle was treated separately. Furthermore, the analyses were split by residue type as one might expect these to influence the geometrical properties to some extent. Thus, for example, the length of the N-C$^\alpha$ bond was considered separately for each of the 20 standard amino acids.

3. Results

(a) *Improvements in structure determination*

Firstly, we present some general observations concerning the 523 proteins in our data set. The structures provide evidence that the process of solving protein structures is continually improving. Figure 2 illustrates how the "quality" of the structures, as crudely measured by resolution and R-factor, has changed over time. While the average R-factor appears to have dropped, indicating a trend towards better structures, the average resolution has tended to get worse. This would seem to suggest that structures can nowadays be solved using poorer-resolution data.

Table 3

Numbers of protein structures in the Brookhaven database having no R*-factors, analysed by year of submission*

Year	Total structures submitted	Structures without R-factors	%-tage without R-factors
1972	1	1	100·0
1973	3	3	100·0
1974	1	1	100·0
1975	11	9	81·8
1976	14	14	100·0
1977	8	5	62·5
1978	1	1	100·0
1979	11	9	81·8
1980	12	5	41·7
1981	22	7	31·8
1982	43	24	55·8
1983	17	1	5·9
1984	27	5	18·5
1985	15	—	0·0
1986	16	—	0·0
1987	41	—	0·0
1988	79	12	15·2
1989	126	2	1·6
1990	69	1	1·4
1991	6	—	0·0
Total	523	100	19·1

Table 4

Numbers of proteins from the full data set of 523 proteins found in each of the resolution and R*-factor bands shown*

	Resolution (Å)																				
R- factor	0.8 -1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1 -3.2	3.5	Total
<0.11	3		2														1				6
0.11	1					1							1								3
0.12	1		1			2		1													5
0.13					2	3		1	1	1	1			1							10
0.14				2		4	2	8	2	2	3		2								25
0.15	1		3		11	6	3	4	2	3	2	3	1		1	1					41
0.16		1	4	3	8	7	1	4		2	3	4	3	4	1	2		1			48
0.17	2	1	2	6	20	7	6	10	3	5	8	2	4	1		1				1	79
0.18			3	3	14	4	6	10	4			1	5		1	3	1		1		56
0.19	1	1	4	4	3	4	5	10	2	1	6	3	2		1	2		2			51
0.20	1				3	4	3	4	1	1	1		1				1	2			22
0.21			1	1			3		1		1	1	3		2	1		1	1		16
0.22				1	1			1	1		1	1	1		1	1		2	1		12
0.23	1			1	1		2							1			1	1			8
0.24		1						1					4	1				2			9
0.25			1			1		2					2			1		2	1		10
0.26										1			1					1	1		4
0.27																		1			1
0.28																1		1			2
0.29													1			1			2		4
0.30		1											2		1	1		2	1	1	9
0.40+													1								1
None		6	4	1		4	1	32	1		2	2	12	1	2	9	7	15		2	101
Total	11	11	25	22	63	47	32	88	18	16	28	17	46	9	10	24	11	33	8	4	523

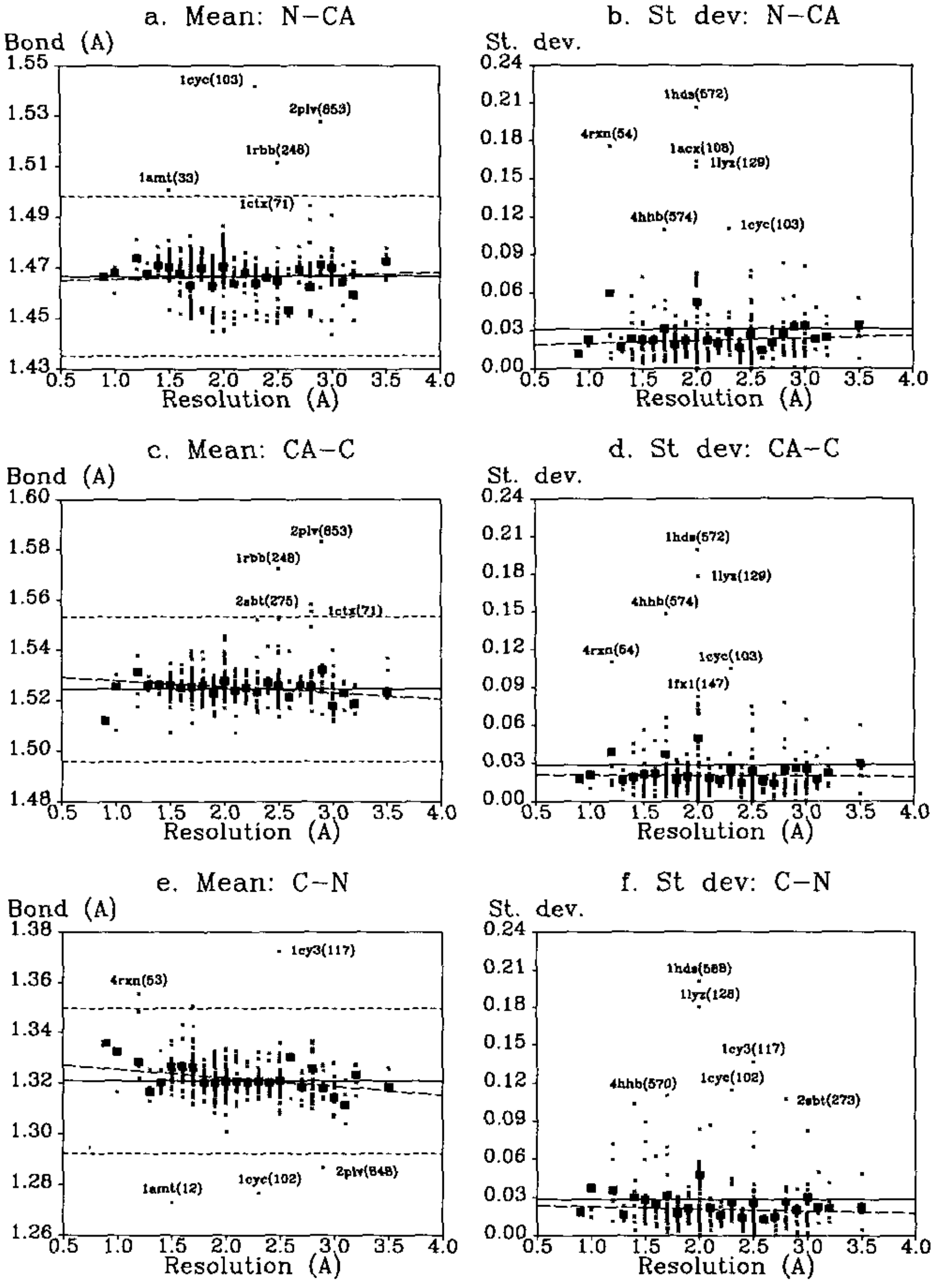


Fig. 3.

The numbers of structures that are submitted without quoted *R*-factors has decreased with time, as shown in Table 3. This suggests that fewer unrefined structures are submitted than in the early days of protein crystallography.

The 523 proteins in our data set are analysed by resolution and *R*-factor in Table 4. The Table shows that, for structures with quoted *R*-factors, there are two noticeable peaks in the distribution: one at 2·0 Å resolution with an *R*-factor between 0·17 and 0·19, and the other at 1·7 Å resolution with an *R*-factor between 0·15 and 0·18. These peaks probably say more about when the decision to publish a given structure is made than about any underlying physical principles!

(b) *Results for all 523 proteins*

The mean and standard deviation values of each of the main-chain bond lengths and angles were first analysed as a function of resolution. Figure 3 shows the results for the bond lengths, and Figure 4 for the bond angles.

For the main-chain bond lengths (Fig. 3) neither the mean values nor the standard deviations show any significant variation with resolution given the spread of the observed values. A number of outliers are clearly visible in the plots. In most cases these correspond to unrefined structures.

The main-chain bond angles, on the other hand, show a slight tendency for their standard deviation

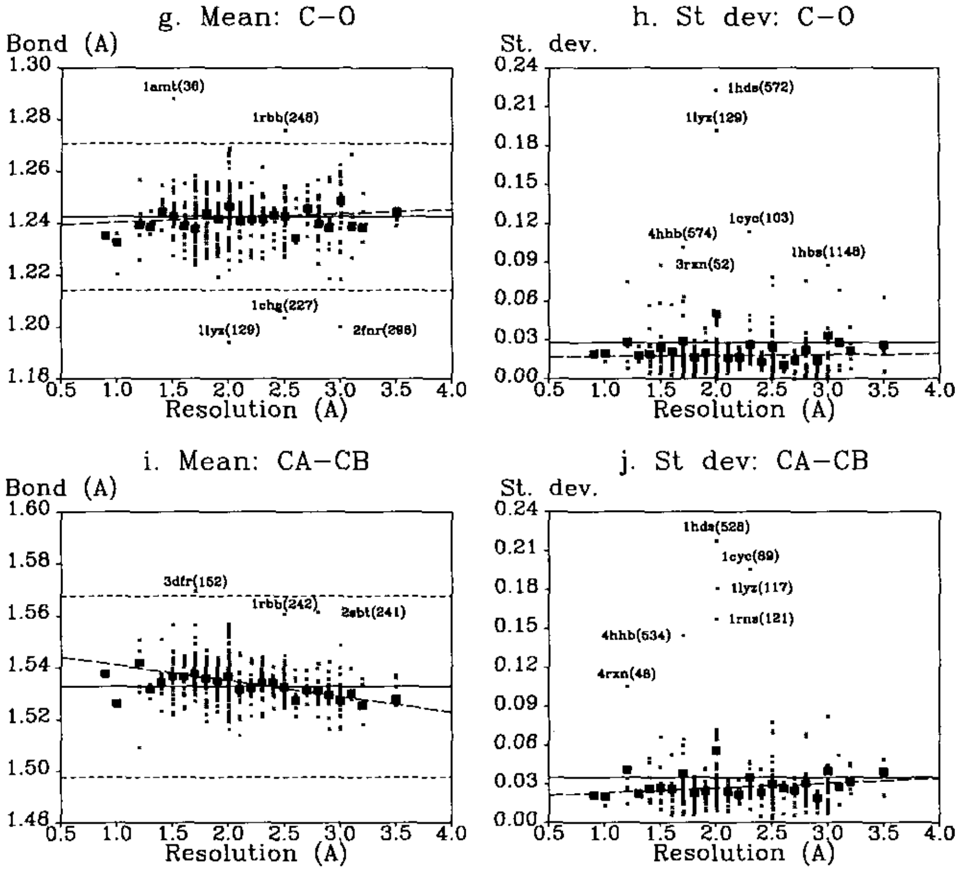


Figure 3. Protein-by-protein means and standard deviations of the 5 main-chain bond lengths, plotted as a function of resolution. The data come from 523 protein crystal structures in the Brookhaven Databank. In each plot, the continuous line indicates the overall mean, with the broken lines defining its standard deviation. The dashed line is a best-fit line to all data points. The dark squares are mean values for all proteins in each resolution bin (0·1 Å). Some of the outliers are identified by their 4-letter Brookhaven code, with the number of their bonds shown in parentheses.

to *decrease* with resolution (Fig. 4). This accords with the analyses of Morris *et al.* (1992) where the authors observed similar trends in the standard deviations of other geometrical properties.

The reason this holds for bond angles and not for bond lengths is possibly because these are less commonly, or less strongly, restrained than the bond lengths.

(c) *Results for 221 high-resolution structures*

The above analysis was repeated on the reduced set of 221 high-resolution structures (i.e. those having a resolution of 2·0 Å or better, and an *R*-factor no greater than 0·20). The proteins involved are shown as either underlined or in **boldface** in Table 1.

As expected the number of outliers on the plots was greatly reduced, though some still remained. The mean values of both the bond lengths and bond angles still exhibited no significant trends. What did change, however, was that the standard deviations of the bond angles tended to be more constant with resolution also. That is, the slight tendency for the standard deviations to *decrease* with resolution disappeared. This suggests that the bond angles in this reduced data set had been subject to stronger restraints.

Plotting the data as a function of *R*-factor revealed an interesting result. Some structures with low *R*-factors exhibited large standard deviations. An example is given in Figure 5 for the N–C$^{\alpha}$ bond where the structures involved appear as outliers. The two most striking outliers, **4rxn** and **4hhb**, were both in fact refined without any restraints. It is not surprising therefore that they should have such large standard deviations. Of the others, some may have attained low *R*-factors by allowing their geometry to relax to stereochemically implausible values.

(d) *Summed variance scores*

The summed variance "scores", described in Materials and Methods, give an idea of the overall variability in a given protein's bond lengths or bond angles. (Of course, this variability may not be the "true" variability within the protein molecule's *actual* structure. Rather, it reflects only the variability within the model of that structure, as solved using X-ray crystallography, and as deposited in the Protein Databank.)

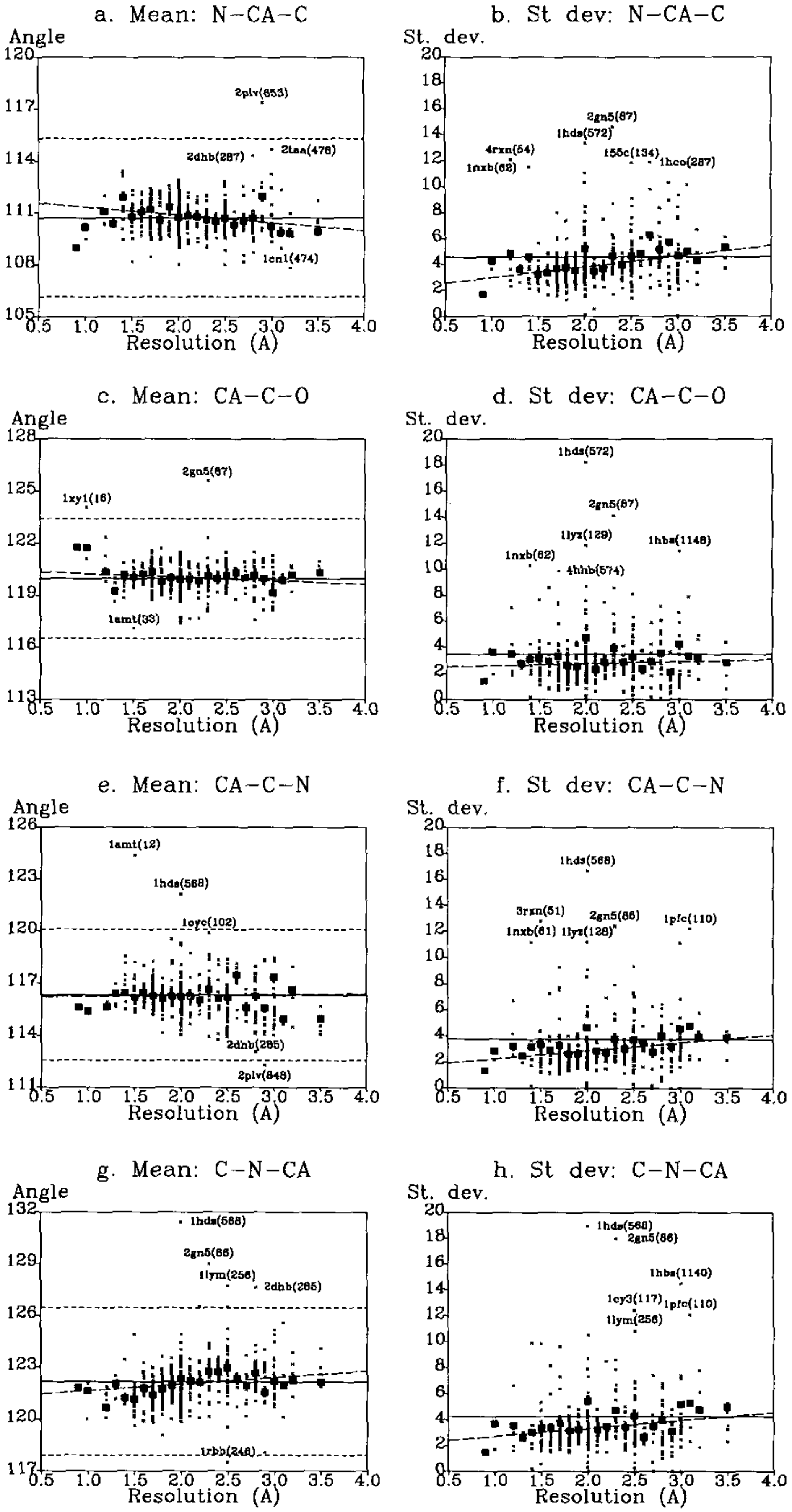
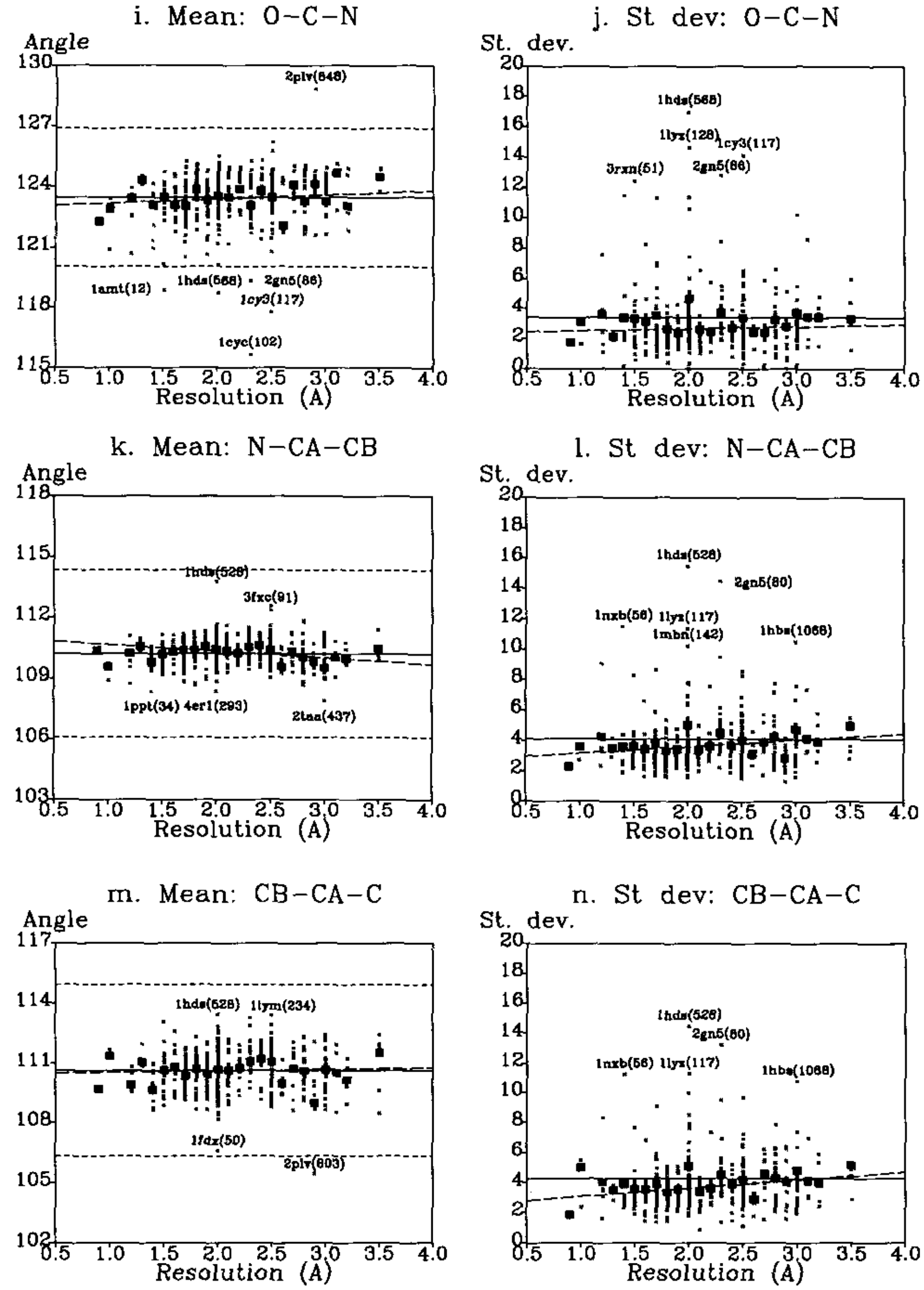

a. Mean: N–CA–C
Angle
120
117
114
111
108
105
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
2plv(653)
2dhb(287)
2taa(478)
1cn1(474)
b. St dev: N–CA–C
St. dev.
20
18
16
14
12
10
8
6
4
2
0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
2gn5(67)
1hds(572)
4rxn(54)
1nxb(62)
155c(134)
1hco(287)
c. Mean: CA–C–O
Angle
128
125
122
119
116
113
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
2gn5(67)
1xy1(16)
1amt(33)
d. St dev: CA–C–O
St. dev.
20
18
16
14
12
10
8
6
4
2
0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
1hds(572)
2gn5(67)
1lyz(129)
1hbs(1146)
1nxb(62)
4hhb(574)
e. Mean: CA–C–N
Angle
126
123
120
117
114
111
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
1amt(12)
1hds(568)
1cyc(102)
2dhb(285)
2plv(848)
f. St dev: CA–C–N
St. dev.
20
18
16
14
12
10
8
6
4
2
0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
1hds(568)
3rxn(51)
2gn5(66)
1pfc(110)
1nxb(61)
1lyz(128)
g. Mean: C–N–CA
Angle
132
129
126
123
120
117
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
1hds(568)
2gn5(66)
1lym(256)
2dhb(285)
1rbb(248)
h. St dev: C–N–CA
St. dev.
20
18
16
14
12
10
8
6
4
2
0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Resolution (A)
1hds(568)
2gn5(66)
1hbs(1140)
1cy3(117)
1pfc(110)
1lym(256)

Fig. 4.



i. Mean: O−C−N

j. St dev: O−C−N

k. Mean: N−CA−CB

l. St dev: N−CA−CB

m. Mean: CB−CA−C

n. St dev: CB−CA−C

Figure 4. Protein-by-protein means and standard deviations of the 7 main-chain bond angles, plotted as a function of resolution. *For a description of the graphs, see the legend to Fig. 3.*

The distributions of these summed variance scores for all 523 proteins are shown in Figure 6. The scores for the 221 high-resolution structures are highlighted. In the case of bond lengths, the distribution of the 221 high-resolution structures is very similar to that of all 523 proteins. For the bond angles, on the other hand, the higher resolution proteins tend to have lower scores. Again, this is possibly a result of the higher resolution, low R-factor structures having had their bond angles more commonly or more strongly restrained during refinement.

Also shown, for comparison, in Figure 6 are the scores for 13 structures solved by NMR techniques. As can be seen, these tend to have low scores. This suggests their parameters deviate relatively little from their target values and underlines the importance of geometrical restraints in the solution of NMR structures.

Within the 221 high-resolution structures, the scores obtained from bond lengths and bond angles are very well correlated. Figure 7 shows that proteins that have low scores on the bond lengths also have low scores on bond angles, and conversely proteins scoring high do so on both angles and lengths. Once more, the most striking outliers are **4rxn** and **4hhb**, which were both refined without restraints. For both structures, additional co-ordi-

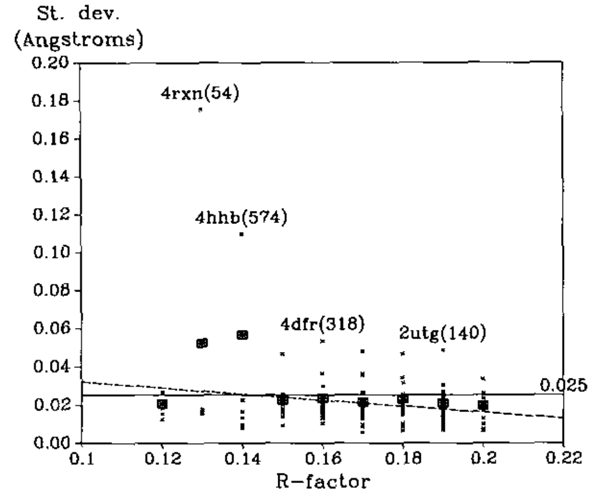


Figure 5. Protein-by-protein standard deviations of the N–C$^{\alpha}$ bond length, plotted as a function of R-factor. The data comprise only 221 high-resolution protein structures (i.e. resolution 2·0 Å or better, and R-factor no more than 20%). The continuous line shows the overall mean, and the broken line is a best-fit line to all data points. The squares give the mean values at each R-factor (grouped in bins of 0·01). Some of the outliers are identified by their 4-letter Brookhaven code, with the number in parentheses showing the number of their N–C$^{\alpha}$ bonds.

nate sets have been deposited which have undergone restrained refinement (**5rxn**, **6rxn**, and **7rxn**; **2hhb** and **3hhb**) and in all cases their scores tend towards the opposite extreme of having very tight restraints.

(e) *Mean values of the parameters*

From this set of 221 proteins, the summed variance scores on the bond angles were used to further refine the data set. Proteins whose scores fell more than one standard deviation either side of the mean score were discarded (see Figs 6 and 7). That is, proteins with very low scores were ejected as well as proteins with very high scores. The removed proteins are shown underlined in Table 1, with the remaining 186 proteins being shown in **bold-face**.

Using this data set of 186 "best" structures, the mean values of the bond lengths and bond angles were compared against the small-molecule data of Engh & Huber (1991). Some of the values had to be subdivided by residue type to correspond to the results of this latter study. Table 5 shows the comparison.

Unfortunately, it is not possible to calculate how statistically significant the differences in the values are, as the Engh & Huber paper gives only standard deviations and not standard errors in the mean values. Nevertheless, the larger differences must be significant; they are the same size as the standard deviations in the Engh & Huber paper which were quoted as being "several times larger than the standard deviation of the mean". The largest of the differences are between the proline C–N bond (0·018 Å) and between the C$^{\beta}$–C$^{\alpha}$–C bond angle in isoleucine, threonine and valine (2·23°).

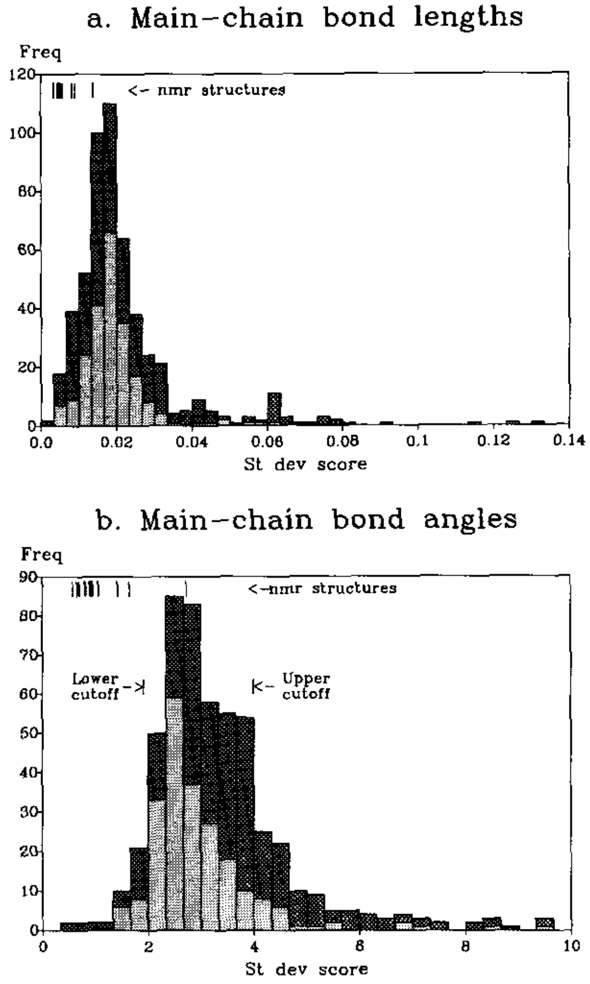


Figure 6. Histogram of summed variance "scores" for a, the 5 main-chain bond lengths, and b, the 7 main-chain bond angles. The lighter shaded bars represent the 221 high-resolution structures, while the darker bars represent the remainder of the full data set of 523 proteins. The lines at the top of each graph represent the scores for 13 structures solved using NMR techniques; these demonstrate how NMR structures tend to have lower standard deviations than structures solved by X-ray crystallography. The 13 structures, are: 1bds, 1cbh, 1il8, 1mhu, 1mrb, 1mrt, 2bus, 2mhu, 2mrt, 3ait, 4ait, 5hir, 6hir. Also shown are the upper and lower cutoffs used for defining the data set of 186 "best" structures (see the text).

(f) *Testing the influence of refinement method*

For the protein data, of course, the question that must be asked is to what extent have the values observed been influenced by the targets used during refinement.

To test this influence, an *analysis of variance* was performed (see Materials and Methods) on the bond lengths and bond angles from structures refined by the five most common methods. The structures are listed in Table 2. For this analysis, in addition to

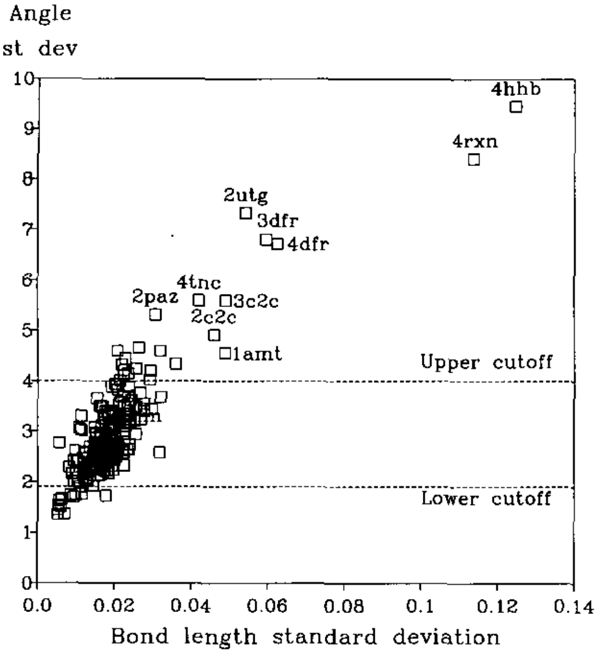


Figure 7. Correlation between the summed variance "scores" for protein bond lengths and bond angles. The results come from the data set of 221 high-resolution structures. The broken lines show the upper and lower cutoffs used for defining the data set of 186 "best" structures (see the text).

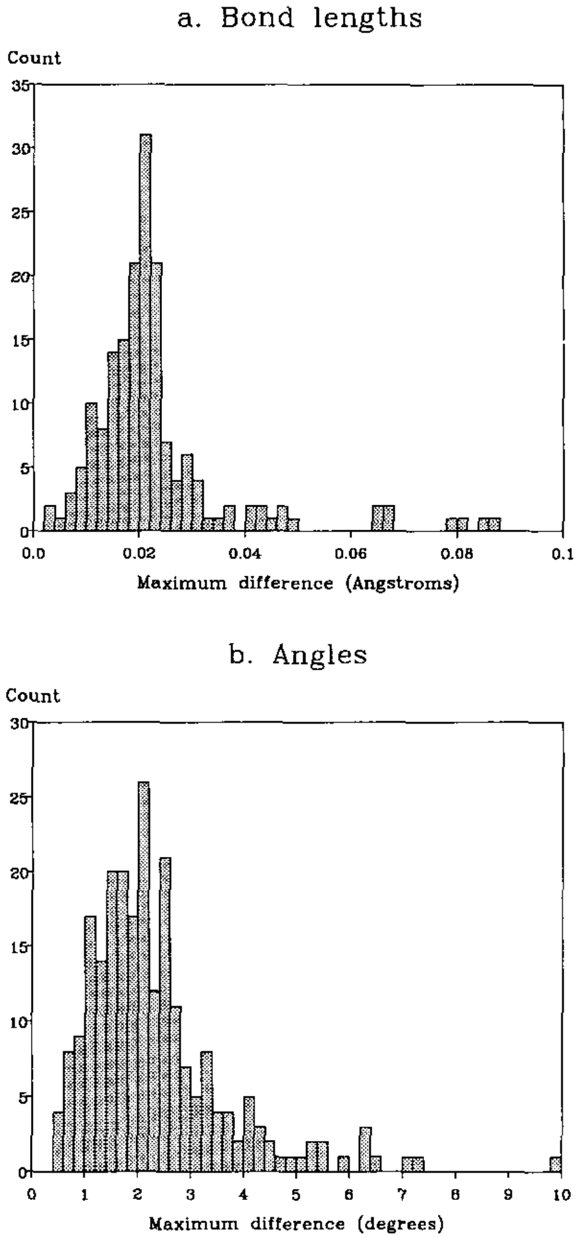


Figure 8. Maximum differences in mean values of bond lengths and angles for structures refined using 5 different refinement methods. The graph shows the distributions for a, the 173 different bond lengths and b, the 234 different bond angles.

the main-chain bond lengths and bond angles, all the side-chain ones were also considered, giving a total of 173 different bond-length types and 234 different bond angle types.

Table 6 gives an example of the results obtained. It shows the mean length of the N–C$^{\alpha}$ bond for seven different residue types as observed in structures refined by the five selected refinement methods. In each case, the difference across the refinement methods is statistically significant at the 99·0% confidence level. In other words, the *restraints applied during refinement have significantly influenced the final values.* In each case, the difference between the lowest and the highest mean value is about 0·02 Å.

From the Table, one can see that the bond lengths from the PROLSQ and RESTRAIN structures tend to be around 1·47 Å, those for the EREF structures around 1·465 Å, and those for the TNT and X-PLOR structures around 1·45 Å. The relative numbers of structures in the databank refined by each of these five different methods results in an overall mean N–C$^{\alpha}$ bond length of around 1·465 Å (Table 5). Yet, from the small-molecule data in Table 5, the mean length for the non-Gly and non-Pro residues should be slightly lower (1·458 Å), and that for the Gly residues should be lower still (1·451 Å).

In fact, virtually all bond lengths and bond angles were found to have significant differences in their mean values across the refinement methods at the 99·0% confidence level. The only two exceptions, out of the 407 different bond length and bond angle types, were the C$^{\epsilon2}$–C$^{\zeta}$–O$^{\eta}$ and C$^{\epsilon1}$–C$^{\zeta}$–O$^{\eta}$ bond angles of tyrosine (significant only at the 95·0% confidence level, and not significant, respectively).

So, the effect of the restraints applied during refinement is still detectable in the final structures and helps explain the significant differences between the mean values for proteins and those for small molecules (Table 5). This appears to be true despite *the supposition that the influence of the restraints* diminishes as the R-factor is reduced and the information from the experimental data (i.e. the observed structure factor amplitudes) takes over.

Table 5

Comparison between main-chain bond lengths and bond angles in proteins and small molecules

A. *Bond lengths*

				Engh & Huber		186 "best"		
Bond	X-PLOR labelling		X-PLOR target (Å)	Value (Å)[a]	σ	Value (Å)[b]	σ	Difference $a-b$(Å)
C–N	C–NH1	(except Pro)	1·33	1·329	0·014	1·323	0·020	0·006
	C–N	(Pro)	1·33	1·341	0·016	1·323	0·020	0·018
C–O	C–O		1·23	1·231	0·020	1·240	0·018	−0·009
C^{α}–C	CH1E–C	(except Gly)	1·52	1·525	0·021	1·525	0·018	0·000
	CH2G*–C	(Gly)	1·52	1·516	0·018	1·523	0·018	−0·007
C^{α}–C^{β}	CH1E–CH3E	(Ala)	1·52	1·521	0·033	1·530	0·018	−0·009
	CH1E–CH1E	(Ile,Thr,Val)	1·53	1·540	0·027	1·548	0·023	−0·008
	CH1E–CH2E	(the rest)	1·52	1·530	0·020	1·533	0·021	−0·003
N–C^{α}	NH1-CH1E	(except Gly,Pro)	1·45	1·458	0·019	1·466	0·020	−0·008
	NH1–CH2G*	(Gly)	1·45	1·451	0·016	1·465	0·022	−0·014
	N–CH1E	(Pro)	1·45	1·466	0·015	1·463	0·018	0·003

B. *Bond angles*

				Engh & Huber		186 "best"		
Bond	X-PLOR labelling		X-PLOR target (°)	Value (°)[a]	σ	Value (°)[b]	σ	Difference $a-b$(°)
C–N–C^{α}	C–NH1–CH1E	(except Gly,Pro)	120·0	121·7	1·8	121·69	3·03	0·01
	C–NH1–CH2G*	(Gly)	120·0	120·6	1·7	121·32	3·34	−0·72
	C–N–CH1E	(Pro)	120·0	122·6	5·0	122·16	2·84	0·44
C^{α}–C–N	CH1E–C–NH1	(except Gly,Pro)	117·5	116·2	2·0	116·31	2·46	−0·11
	CH2G*–C–NH1	(Gly)	117·5	116·4	2·1	116·34	2·75	0·06
	CH1E–C–N	(Pro)	117·5	116·9	1·5	116·39	2·61	0·51
C^{α}–C–O	CH1E–C–O	(except Gly)	121·5	120·8	1·7	120·06	2·49	0·74
	CH2G*–C–O	(Gly)	121·5	120·8	2·1	120·22	2·52	0·58
C^{β}–C^{α}–C	CH3E–CH1E–C	(Ala)	106·5	110·5	1·5	110·48	2·91	0·02
	CH1E–CH1E–C	(Ile,Thr,Val)	110·0	109·1	2·2	111·33	3·03	−2·23
	CH2E–CH1E–C	(the rest)	109·5	110·1	1·9	110·33	3·45	−0·23
N-C^{α}–C	NH1-CH1E–C	(except Gly,Pro)	111·6	111·2	2·8	110·77	3·29	0·43
	NH1-CH2G*-C	(Gly)	111·6	112·5	2·9	112·19	3·64	0·31
	N–CH1E–C	(Pro)	111·6	111·8	2·5	112·38	3·36	−0·58
N–C^{α}–C^{β}	NH1-CH1E-CH3E	(Ala)	108·5	110·4	1·5	110·52	2·66	−0·12
	NH1–CH1E–CH1E	(Ile,Thr,Val)	110·0	111·5	1·7	111·05	3·15	0·45
	N–CH1E–CH2E	(Pro)	104·0	103·0	1·1	104·91	2·09	−1·91
	NH1–CH1E–CH2E	(the rest)	110·0	110·5	1·7	110·61	3·17	−0·11
O–C–N	O–C–NH1	(except Pro)	121·0	123·0	1·6	123·40	2·42	−0·40
	O–C–N	(Pro)	121·0	122·0	1·4	123·00	2·31	−1·00

The protein data comes from the 186 "best" structures, and the small-molecule data comes from the analysis of Engh & Huber (1991). The atom-labelling follows that used in the X-PLOR dictionary, with some additional atoms (marked with an asterisk) as defined in the Engh & Huber paper.

To illustrate this remnant bias, we tried to see if we could actually detect which method had been used to refine a given structure simply by analysing the structure's bond lengths and bond angles. Using a very crude strategy we found that the refinement method could be detected with an overall accuracy of 95%. Structures refined by PROLSQ and RESTRAIN were correctly identified in 100% of the cases, while those refined by the other three methods were occasionally misidentified, most commonly as PROLSQ structures (see Table 7). This illustration shows that each refinement method does indeed leave its own imprint on the structures it refines by virtue of its unique dictionary of target values.

What is the size of the resultant differences across the refinement methods? Figure 8 gives an idea. It shows the maximum differences across the methods for each bond length and bond angle. The distribution of these differences shows that the bond lengths differ by around 0·02 Å across the methods, while the bond angles differ by around 2°. Curiously, these values are exactly the amounts by which bond lengths and angles are usually allowed to vary about their ideals during refinement (Hendrickson, 1985). One would have expected the differences to be *smaller* than these amounts. It should be stressed that these observed differences (0·02 Å for bonds and 2° for angles) should not be interpreted as target values for refinement. They merely reflect the current restraints.

The greatest variations are highlighted in Tables 8 and 9. These tables show that the two methods most often at odds with the others are EREF and X-PLOR. Both have parameter lists with few atom types, so do not always successfully cater for the local influences within certain of the amino acid residues.

Refinement not only influences the parameters, but also their standard deviations about their mean values. Indeed, by appropriate choice of weighting factors it is possible to obtain any desired deviation for a structure's bond lengths and angles. Thus, the

Table 6

Analysis of variance calculations for the mean N–C$^\alpha$ bond length for seven different residue types, obtained from five different refinement methods

Residue type	Refinement method	No. of examples	Mean length (Å)	Standard deviation (Å)	Max − min (Å)	$\frac{MS_{Tr}}{MS_E}$	Confidence level
	X-PLOR	1280	1·4522	0·0116			
	TNT	1373	1·4561	0·0200			
Ala	EREF	894	1·4654	0·0173	0·0204	258·39	99·0%
	PROLSQ	5590	1·4705	0·0251			
	RESTRAIN	355	1·4726	0·0170			
	X-PLOR	815	1·4504	0·0126			
	TNT	770	1·4532	0·0200			
Arg	EREF	330	1·4654	0·0157	0·0252	188·02	99·0%
	PROLSQ	2461	1·4704	0·0249			
	RESTRAIN	59	1·4756	0·0165			
	X-PLOR	1029	1·4503	0·0123			
	TNT	786	1·4508	0·0233			
Asn	EREF	594	1·4672	0·0171	0·0213	255·02	99·0%
	PROLSQ	3037	1·4712	0·0253			
	RESTRAIN	100	1·4716	0·0187			
	X-PLOR	1065	1·4509	0·0119			
	TNT	822	1·4520	0·0204			
Asp	EREF	514	1·4664	0·0169	0·0215	252·92	99·0%
	PROLSQ	3562	1·4722	0·0273			
	RESTRAIN	290	1·4723	0·0209			
	X-PLOR	293	1·4468	0·0112			
	TNT	137	1·4498	0·0184			
Cys	EREF	331	1·4658	0·0167	0·0242	112·23	99·0%
	PROLSQ	1392	1·4709	0·0223			
	RESTRAIN	49	1·4689	0·0128			
	TNT	1187	1·4511	0·0206			
	X-PLOR	1152	1·4499	0·0111			
Gly	EREF	1001	1·4658	0·0159	0·0216	207·51	99·0%
	PROLSQ	5492	1·4682	0·0304			
	RESTRAIN	483	1·4715	0·0207			
	TNT	455	1·4494	0·0198			
	X-PLOR	731	1·4569	0·0113			
Pro	RESTRAIN	184	1·4669	0·0229	0·0206	109·84	99·0%
	PROLSQ	2826	1·4675	0·0233			
	EREF	537	1·4700	0·0160			

For each residue type, the means and standard deviations of the bond length are shown in order of increasing value. Also shown is the difference between the maximum and minimum of these mean values, and the ratio MS_{Tr}/MS_E used in the F-test for assessing the significance of the differences in the means (see Materials and Methods). The rightmost column shows the confidence level at which the F-test indicates that the differences are statistically significant. For the F-test, the tabulated value of $F_{k-1,N-k}$ was used, where k=the number of refinement methods=5, and N=the total number of examples (taken to be ∞).

Table 7

Accuracy of detection of refinement method from an analysis of a structure's bond lengths and bond angles

Actual refinement method	Predicted refinement method					Total
	PROLSQ	TNT	EREF	X-PLOR	RESTRAIN	
PROLSQ	198 (100%)	–	–	–	–	198
TNT	7	52 (91·2%)	–	–	–	59
EREF	7	–	30 (78·9%)	–	1	38
X-PLOR	2	–	–	21 (91·3%)	–	23
RESTRAIN	–	–	–	–	15 (100%)	15
Total						333

The numbers down the diagonal give the numbers of structures whose refinement method was correctly detected, with the numbers in parentheses giving the percentage correct. The overall prediction success was 94·9%.

Table 8

Bond lengths having the largest discrepancies in their mean values across the different refinement methods

Bond and residue	Refinement method	No. of examples	Mean length (Å)	Standard deviation (Å)	Max −min (Å)
$C^{\delta 1}$–$N^{\epsilon 1}$ Trp	X-PLOR	194	1·3095	0·0075	0·118
	TNT	242	1·3715	0·0140	
	PROLSQ	829	1·3788	0·0265	
	EREF	182	1·4000	0·0203	
	RESTRAIN	68	1·4271	0·0180	
$C^{\delta 2}$–$N^{\epsilon 2}$ His	X-PLOR	442	1·3047	0·0080	0·087
	PROLSQ	1423	1·3724	0·0264	
	TNT	234	1·3799	0·0166	
	RESTRAIN	44	1·3838	0·0219	
	EREF	304	1·3915	0·0255	
$N^{\epsilon 1}$–$C^{\epsilon 2}$ Trp	EREF	182	1·2899	0·0351	0·086
	RESTRAIN	68	1·3129	0·0149	
	X-PLOR	194	1·3417	0·0091	
	PROLSQ	829	1·3644	0·0268	
	TNT	242	1·3756	0·0172	
$N^{\delta 1}$–$C^{\epsilon 1}$ His	X-PLOR	442	1·3115	0·0091	0·080
	PROLSQ	1423	1·3253	0·0238	
	TNT	234	1·3313	0·0168	
	RESTRAIN	44	1·3714	0·0330	
	EREF	304	1·3917	0·0264	
$C^{\epsilon 1}$–$N^{\epsilon 2}$ His	X-PLOR	442	1·3124	0·0092	0·079
	PROLSQ	1423	1·3186	0·0281	
	TNT	234	1·3339	0·0196	
	RESTRAIN	44	1·3644	0·0379	
	EREF	304	1·3910	0·0291	
C^{γ}–$C^{\delta 2}$ Trp	X-PLOR	194	1·3689	0·0105	0·068
	RESTRAIN	68	1·4046	0·0123	
	EREF	182	1·4065	0·0180	
	PROLSQ	829	1·4359	0·0164	
	TNT	242	1·4367	0·0180	

Table 9

Bond angles having the largest discrepancies in their mean values across the different refinement methods

Bond angle and residue	Refinement method	No. of examples	Mean value (°)	Standard deviation (°)	Max −min (°)
C^{γ}–S^{δ}–C^{ϵ} Met	X-PLOR	292	99·372	4·091	9·936
	EREF	193	99·419	2·782	
	TNT	385	100·068	4·228	
	PROLSQ	1220	100·563	4·155	
	RESTRAIN	20	109·308	2·709	
$C^{\delta 1}$–C^{γ}–$C^{\delta 2}$ Trp	EREF	182	105·335	0·681	7·228
	PROLSQ	829	105·988	1·154	
	TNT	242	106·368	0·809	
	RESTRAIN	68	107·979	0·988	
	X-PLOR	194	112·563	0·765	
C^{α}–C^{β}–C^{γ} Glu	RESTRAIN	93	110·254	3·003	7·011
	TNT	640	111·856	4·299	
	EREF	495	112·512	2·840	
	X-PLOR	964	114·442	5·124	
	PROLSQ	3224	117·265	6·413	
C^{β}–C^{γ}–$C^{\delta 1}$ Trp	X-PLOR	194	121·989	1·827	6·564
	RESTRAIN	68	125·654	2·913	
	TNT	242	126·462	1·433	
	PROLSQ	829	127·662	2·394	
	EREF	182	128·552	1·342	
C^{α}–C^{β}–$C^{\gamma 2}$ Val	X-PLOR	1090	109·978	2·967	6·334
	TNT	876	110·244	2·672	
	EREF	856	111·719	2·496	
	PROLSQ	4622	112·044	3·643	
	RESTRAIN	302	116·312	4·258	
C^{β}–C^{γ}–C^{δ} Glu	EREF	495	109·615	3·041	6·311
	RESTRAIN	93	110·908	2·471	
	TNT	640	112·638	3·649	
	X-PLOR	964	114·671	3·473	
	PROLSQ	3210	115·926	6·490	

analysis of known structures can only reflect what the refinement process has achieved rather than say anything about the true deviations in proteins.

4. Discussion

This paper has presented an analysis of the main-chain bond lengths and bond angles of protein structures in the July 1991 release of the Brookhaven database.

Three data sets were used. The first comprised all 523 proteins for which complete structures were available. The second was a reduced set of 221 high-resolution structures, being those with a resolution of 2·0 Å or better and an R-factor no greater than 0·20. The third was a further reduced set of 186 structures, obtained by removing structures having very high or very low "scores" of summed variance values for the main-chain bond angles.

The analyses showed that, unlike other stereochemical parameters, the standard deviations in the bond lengths and bond angles show little reliable variation with resolution. Thus, they are less useful as measures of "stereochemical quality" than the parameters identified by Morris *et al.* (1992). This was to be expected as bond lengths and bond angles are heavily influenced by the geometrical restraints applied during structure refinement.

The effect of the restraints was then investigated by looking at the variations in bond lengths and bond angles between proteins refined by the five most commonly used refinement methods. It was shown that the effects *are* statistically significant and, indeed, are worryingly large. The effects account for the differences observed between the mean values of these bond parameters as obtained from protein structures, and the corresponding mean values obtained from small-molecule structures.

These findings seem to support those of Engh & Huber (1991), who noted the significant differences between the parameter lists used by different refinement methods. The authors stated that these differences are larger than one would wish. The findings suggest that the dictionaries used for refinement should be modified to take into account accurate up-to-date small-molecule data such as tabulated by Engh & Huber (1991).

R.A.L. was supported by an SERC studentship during part of this work.

References

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr. sect B*, **35**, 2331–2339.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). Tables of bond lengths determined by X-ray and neutron diffraction. Part I. Bond lengths in organic compounds. *Chem. Soc. Perkin Trans. II*, S1–S19.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rogers, J. R., 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.

Bowen, H. J. M., Donohue, J., Jenkin, D. G., Kennard, O., Wheatley, P. J. & Whiffen, D. H. (1958). *Tables of Interatomic Distances and Configurations in Molecules and Ions* (Mitchell, A. D. & Cross, L. C., eds), The Chemical Society, London.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. *Science*, **235**, 458–460.

Cruickshank, D. W. J. (1965). Errors in least-squares methods. In *Computing Methods in Crystallography* (Rollett, J. S., ed.), pp. 112–116, Pergamon, Oxford.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures. *J. Appl. Crystallogr.* **22**, 510–516.

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

Haneef, I., Moss, D. S., Stanford, M. J. & Borkakoti, N. (1985). Restrained structure-factor least-squares refinement of protein structures using a vector processing computer. *Acta Crystallogr. sect. A*, **41**, 426–433.

Hendrickson, W. A. (1985). Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol.* **115**, 252–270.

Hendrickson, W. A. & Konnert, J. H. (1980). Incorporation of stereochemical information into crystallographic refinement. In *Computing in Crystallography* (Diamond, R., Rameseshan, S. & Venkatesan, K., eds), pp. 13.01–13.26, Indian Academy of Sciences, Bangalore, India.

Hubbard, T. J. P. & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171.

Jack, A. & Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Crystallogr. sect. A*, **34**, 931–935.

Kennard, O. (1968). Tables of bond lengths between carbon and other elements. In *International Tables for X-ray Crystallography*, pp. 275–276, Vol III, Kynoch Press, Birmingham.

Konnert, J. H. (1976). A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Crystallogr. sect. A*, **32**, 614–617.

Laskowski, R. A. (1992). *Prediction, analysis and determination of protein structure, including applications of parallel computing*, pp. 176–211, PhD thesis, University of London.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291.

Levitt, M. (1974). Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.* **82**, 393–420.

Milton, J. S. & Arnold, J. C. (1986). *Probability and Statistics in the Engineering and Computing Sciences.* McGraw-Hill, Singapore.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure co-ordinates. *Proteins*, **12**, 345–364.

Moss, D. S. & Morffew, A. J. (1982). RESTRAIN: a restrained least-squares refinement program for use in protein crystallography. *Comput. Chem.* **6**(1), 1–3.

Robinson, W. T. & Sheldrick, G. M. (1988). In *Crystallographic Computing 4: Techniques and New Technologies* (Isaacs, N. W. & Taylor, M. R., eds), pp. 366–377, International Union of Crystallography, Oxford Univ. Press.

Sheldrick, G. M. (1976). *SHELX76. Program for Crystal Structure Determination.* University of Cambridge, England (Computer program).

Sheldrick, G. M. (1985). In *Crystallographic Computing 3: Data Collection, Structure Determination, Proteins, and Databases* (Sheldrick, G. M., Krüger, C. & Goddard, R., eds), pp. 184–189, Clarendon Press, Oxford.

Sheldrick, G. M. (1986). *SHELX86. Program for Crystal Structure Determination.* University of Göttingen, Federal Republic of Germany (Computer program).

Taylor, R. & Kennard, O. (1983). The estimation of average molecular dimensions from crystallographic data. *Acta Crystallogr. sect. B*, **39**, 517–525.

Taylor, R. & Kennard, O. (1985). The estimation of average molecular dimensions. 2. Hypothesis testing with weighted and unweighted means. *Acta Crystallogr. sect. A*, **41**, 85–89.

Taylor, R. & Kennard, O. (1986). Accuracy of crystal structure error estimates. *Acta Crystallogr. sect. B*, **42**, 112–120.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr. sect. A*, **43**, 489–501.

Vijayan, M. (1976). In CRC *Handbook of Biochemistry and Molecular Biology*, 3rd edit., *Proteins*, vol 2 (Fasman, G. D., ed.), pp. 742–759, CRC Press, Cleveland.

Edited by R. Huber