INCOHERENT RELATIONSHIP BETWEEN TORSION VALUES AND BOND LENGTHS IN ATOMIC RESOLUTION PROTEIN CRYSTAL STRUCTURES

Oliviero Carugo

*Department of General Chemistry, Pavia University, Viale Taramelli 12, 27100 Pavia, Italy
and
Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and
Biotechnology, Padrice 99, 34012 Trst, Italy*

Abstract

Deviations from planarity of the peptide Cα-C(=O)-N moiety are expected to be accompanied by the lengthening of the C-N and the shortening of the C=O bonds. Surprisingly, such a relationship cannot be detected by an analysis of the atomic resolution protein crystal structures deposited in the Protein Data Bank. This could be due to either an incorrect parameterisation of the refinement restraints or to the large mobility of the protein atoms, which could prevent their exact location.

Introduction

The analysis of known protein three-dimensional structures produced marvellous results during the last few decades, both in allowing the development of new technologies, like for example theoretical modelling and protein design,[1,2] and in improving our understanding of basic biological features, like for example the adaptation to high temperature and other extreme physical environments.[3,4] Given that most of the known information about protein three-dimensional structure has been provided by crystallographic methods, structural bioinformaticians routinely filter out from the Protein Data Bank[5] X-ray crystal structures with the characteristics desired for each particular analysis. The most usual filter is the crystallographic resolution which indicates how well defined the crystal structure is, i.e. how many experimental diffraction data have been measured relative to the number of variables that must be determined (positions of the atoms, atomic displacement parameters, etc.). The crystallographic resolution is of course not the only criterion to select the "best" protein three-dimensional structures to be analyzed in a bioinformatic project. The R factor as well as the free R factor are useful indicators of the "quality" of a crystal structure.[6-8] The diffraction data completeness in the highest resolution shell is a further possible

indicator of the structure "quality" as well as a critical assessment of the stereochemistry of the macromolecule[9] or the comparison between the number of solvent molecules expected and experimentally located.[10] Nevertheless, the crystallographic resolution is certainly the parameter most commonly used to select a three-dimensional protein structure subset from which to extract any trend to be related to any biological feature. The reason for this is very simple: contrary to other useful crystallographic parameters, the crystallographic resolution can be extracted very easily from the Protein Data Bank files.
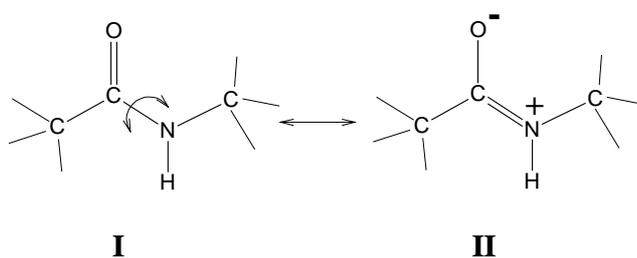

I **II**

Figure 1. Mesomeric equilibrium that determines the planarity of the peptide unit. The rotation around the C-N bond is impossible because of the electron delocalisation that confers C-N a partial double bond character. The ω torsion around C-N is thus constrained to 0 or 180° (180° in the figure). Deviations from these values imply deviations from planarity of the peptide unit with consequent decrease of importance of the limit formula II. As a consequence, the peptide C-O and C-N bonds are expected to shorten and lengthen, respectively.

For many years, the crystallographic resolution in protein crystallography has been limited to values, in the best cases, around 2 Å. Recent improvements in both diffraction data acquisition at synchrotrons and data manipulation allowed the development of the so called "atomic resolution" protein crystallography.[11] The resolution went down to values close to 1 Å, and in some cases even below.[12] The very large amount of experimental diffraction data available at this resolution allowed, in these cases, to treat a protein crystal structure nearly as chemists routinely do with small molecules. The quality of the crystallographic results is therefore expected to improve to such a level to allow very detailed interpretations of reaction mechanisms as well as structural trends. Some limitations, nevertheless, still persist, as shown below.

In the present communication, the planarity of the peptide backbone moiety, monitored by the ω torsion around C-N, is correlated with the peptide C-O and C-N bond lengths (Figure 1). It is expected that the electronic delocalisation within the

peptide unit decreases when the peptide deviates from planarity, with a consequent shortening of the C-O bond and a lengthening of the C-N bond.[13] The mesomeric limit formula II of Figure 1 is thus less representative of a non-planar peptide. Atomic resolution protein crystal structures, found in the Protein Data Bank[5] (Table 1), were analysed. Unexpectedly, deviations of the peptide unit from planarity are not associated with shortening and lengthening of the C-O and C-N bonds, respectively.

Table 1. PDB identification codes of the protein atomic resolution crystal structures examined in the present work. They have resolution of at least 1.3 Å, at least 99 percent of the atoms refined anisotropically, and at least 300 atoms. Given that the present work examines basic stereochemical features, the sequence similarity of these protein chains was disregarded as a criterion to define a representative data set.

1a6k	1a6m	1a6n	1a7s	1b0y	1b6g	1bkr	1bs9	1bx7	1bxo
1byi	1bz6	1bzp	1bzr	1c5e	1c75	1c9o	1cc7	1cku	1ctj
1cxq	1cy5	1cz9	1czb	1czp	1d4t	1dbf	1ejg	1euw	1irn
1iro	1ixg	1ixh	1lkk	1lks	1psr	1qj4	1ql0	1qlw	1qnj
1qow	1qt9	1qtw	1rb9	1rge	1rgg	1swu	2erl	2fdn	2igd
2nlr	2pvb	3chb	3lzt	3pyp	3sil	4lzt	7a3h	7fd1	8a3h

Results and discussion

The ω torsions and the C-O and C-N bond lengths were computed for all non-disordered residues (occupancy = 1) of the protein structures reported in Table 1. The first and the last residues within each chain were of course disregarded as well as residues surrounding disordered three-dimensional regions. For each residue the Δω parameter, defined as the minimal angular deviation from 0 or 180°, was then computed. These reference values of 0 and 180° correspond to the *cis* and the *trans* planar peptide conformation, respectively.

Figure 2 reports the distribution of the Δω values for solvent exposed and buried residues and for residues in various secondary structures. A residue was considered buried if its total solvent accessible area, computed with DSSP,[14] was less than 5 Å^2. The secondary structure types, also assigned with DSSP,[14] were simplified as helix (H), strand (B), turn (T), and coil (C), as described by Heringa and Argos.[15] Most of the residues have Δω values close to 0°, indicating a substantially planar peptide unit. The

distributions of $\Delta\omega$ values are statistically independent of the degree of burial within the protein core. Residues in helical backbone conformations show the greatest resistance from distorting planar peptide units. Only about 17-19% of them have $\Delta\omega$ values higher than 5°, while 35-45% of the residues in other backbone conformations have $\Delta\omega$ values over 5°. This clearly reflects the higher steric requirements of helical segments, whose ϕ and ψ torsions are known to be little variable.
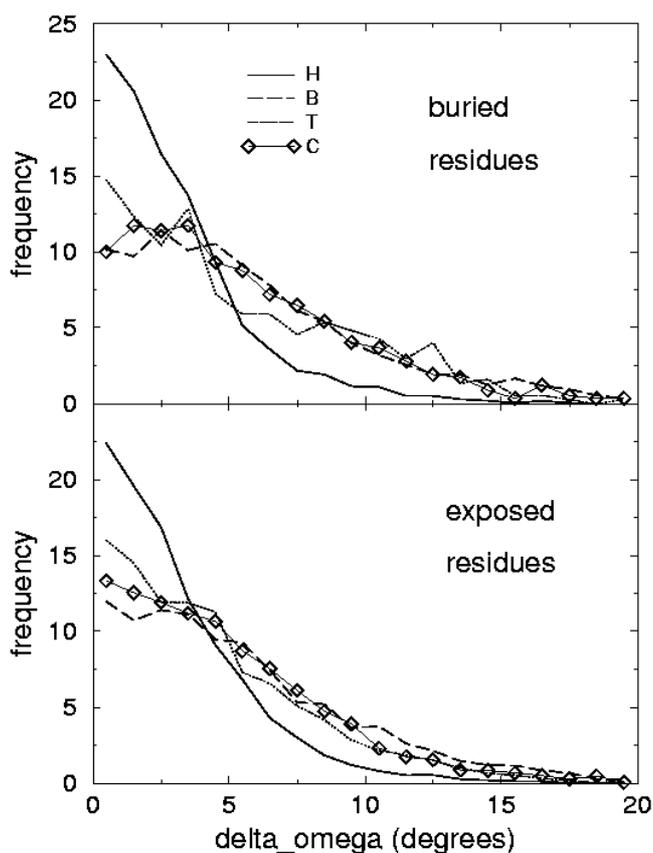

Figure 2. Distributions of the $\Delta\omega$ values for buried and solvent exposed residues associated with various secondary structure types (H = helix, B = strand, T = turn, C = coil).

The relationships between $\Delta\omega$ and the amide C-O and C-N bond lengths are depicted in Figure 3. Single observations are indicated with small circles while large black stars indicate mean bond length values for $\Delta\omega$ bins of 1 degree. The mean C-O bond lengths oscillate from a minimum of 1.232 Å to a maximum of 1.236 Å. Analogously, the mean C-N bond distances range from 1.327 to 1.332 Å. The standard

deviations of these mean bond distances are not reported in Figure 3, for clarity, but range from 0.001 to 0.004 Å. The variability of the mean C-N and C-O distances is therefore statistically insignificant. The linear-least-squares fit of the C-N and C-O bond lengths versus the $\Delta\omega$ values results in slopes statistically identical to 0.0. Analogous results were obtained by considering separately buried and solvent exposed residues as well as different secondary structural types. The correlation between $\Delta\omega$ values and C-O and C-N bond distances was also examined separately in each of the protein structures reported in Table 1. Only in 19% of the structures the C-O bonds were statistically shorter in the $\Delta\omega$ range 15-20° than in the 0-5° range. Only in 2% of the structures the C-N bond was statistically longer in the 15-20° $\Delta\omega$ range than in the 0-5° range. Similar results were obtained also by analyzing only 9 crystal structures determined at resolution better that 0.95 Å (1b0y, 1dy5, 1gci, 1nls, 1rb9, 2fdn, 2pvb, 3lzt, 3pyp) where, as expected, a modest, statistically significant correlation between the C-O and C-N bond distances has been recently observed.[16] For each these 9 ultra-high resolution structures, the slopes of the relationships of the C-O and C-N bond distances with the $\Delta\omega$ values were statistically identical to 0.0. It must therefore be concluded that distortions from planarity are not accompanied by bond length modifications.

The above results contrast observations on small molecule crystal structures. For example, $\Delta\omega$ values close to 15° are expected to be associated with C-N distances close to 1.38-1.40 Å,[13] much longer than those observed in the proteins of Table 1 (*ca.* 1.33 Å). Such an observation is particularly significant because there are not statistically relevant relationships between the rotameric state of the main-chain peptide bond and the residue atomic displacement parameter values (data not shown), *i.e.* between the rotameric state and the mobility and flexibility of the atom subset.

Amongst the possible reasons for this discrepancy between observations and expectations, two are worth noting. On one side, crystallographic refinements are rather different in small and large asymmetric unit crystallography. While the application of stereochemical restraints is very unusual in small molecule crystallography (with the exception of the hydrogen atom positions), it is nearly always necessary in protein crystallography. Even at very high resolution, like for the structures considered here, the experimental diffraction data are usually insufficient to allow refinements completely unrestrained. As a consequence, it might be possible that the weights applied to the

peptide planarity restraint and to the peptide bond distance restraints are inconsistent. Deviations of the bond distances from their target values might be penalized more than deviations from planarity.
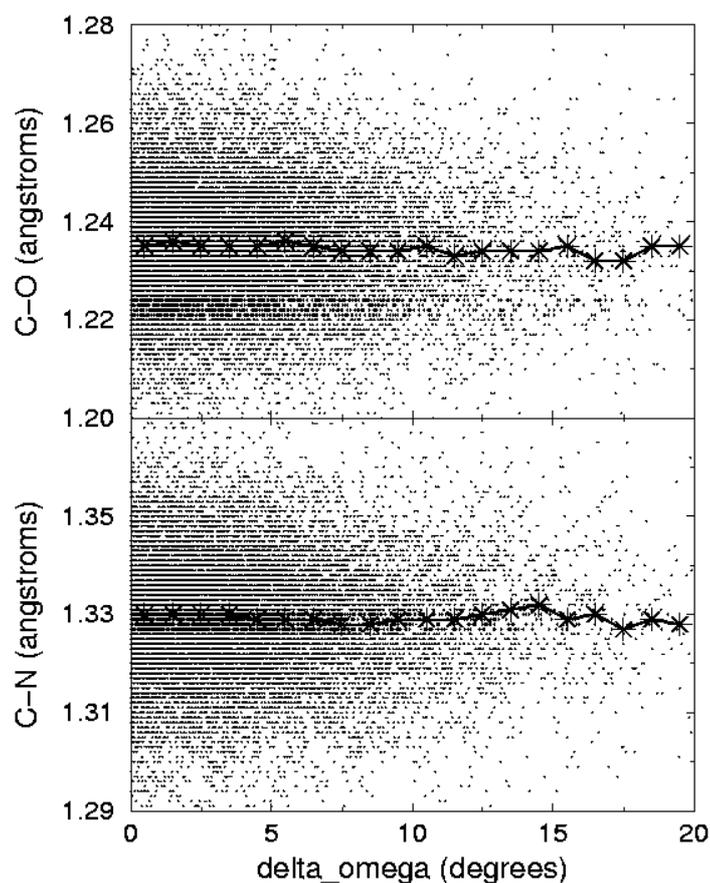

Figure 3. Relationships between the $\Delta\omega$ values and the peptide C-O and C-N bond lengths. Single occurrences are indicated with small circles while large black stars indicate the mean bond lengths for $\Delta\omega$ bins of 1 degree.

On the other side, the overall mobility of protein atoms is considerably high and could prevent the detection, even with diffraction data at very high resolution, of subtle features. One must remember that a quite usual atomic displacement parameter value of 20 Å^2 (in B units) corresponds to a mean square displacement of 0.25 Å^2 from the average atom position, a value not much minor than most of the covalent bonds. It might be possible that such an intrinsic flexibility of the protein molecules in their crystals will never allow very accurate, stereochemical characterizations. The fact that the $\Delta\omega$ values

are similar, on average, for both buried and exposed residues (Figure 2) actually supports this hypothesis.

Although it is not clear which are the ultimate reasons for the discrepancy between observations and expectations, it is nevertheless important for both crystallography and structural bioinformatics to consider that experimental results could be, at least in part, rather inaccurate. Crystallographic improvements must be achieved and caution in data mining still remains.

Experimental

All data were taken from the Protein Data Bank.[5] 60 crystal structures with resolution of at least 1.3 Å, at least 99 percent of the atoms refined anisotropically, and with at least 300 atoms were retained. Residue solvent accessible area values were determined with DSSP[14] (probe sphere radius = 1.4 Angstroms), as well as the secondary structural assignments and the omega angles.

References

1. D. Shortle, *Curr. Biol.* **1999**, *9*, R49–R51.
2. S. Lutz, S. J. Benkovic, *Curr. Opin. Biotechnol.* **2000**, *11*, 319–324.
3. A. Szilagyi, P. Zavodszky, *Structure* **2000**, *8*, 493–504.
4. G. Gianese, P. Argos, S. Pascarella, *Prot. Eng.* **2001**, *14*, 141–148.
5. H. M. Bermann, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucl. Acid. Res.* **2000**, *28*, 235–242.
6. A. Bacchi, V. Lamzin, K. S. Wilson, *Acta Cryst.* **1996**, *D52*, 641–647.
7. G. J. Kleywegt, A. T. Bruenger, *Structure* **1996**, *4*, 897–904.
8. A. T. Bruenger, *Nature* **1992**, *355*, 472–475.
9. R. A. Laskowski, M. W. MacArthur, J. M. Thornton, *Curr. Opin. Struct. Biol.* **1998**, *8*, 631–639.
10. O. Carugo, D. Bordo, *Acta Cryst.* **1999**, *D55*, 479–484.
11. G. M. Sheldrick, *Acta Cryst.* **1990**, *A46*, 467–471.
12. Z. Dauter, V. S. Lamzin, K. S. Wilson, *Curr. Op. Struct. Biol.* **1997**, *7*, 681–688.
13. H. Shao, X. Jiang, P. Gantzel, M. Goodman, *Chem. Biol.* **1994**, *1*, 231–234.
14. W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577–2637.
15. J. Heringa, P. Argos, *Proteins* **1999**, *37*, 30–43.
16. L. Esposito, L. Vitagliano, A. Zagari, L. Mazzarella, *Protein Eng.* **2000**, *13*, 825–828.

Povzetek

Pričakovali bi, da bo odmik od planarnosti vezi Cα-C(C=O)-N v peptidni skupini spremljalo podaljšanje C-N vezi in skrajšanje C=O vezi. Presentljivo pa z analizo proteinskih kristalnih struktur z atomsko ločljivostjo, ki so v zbirki proteinskih struktur PDB, take relacije ne moremo pokazati. Razlog za to je nepravilna parametrizacija omejitev, uporabljenih pri piljenju proteinskih struktur ali pa velika gibljivost proteinskih atomov, ki preprečuje določitev njihovega natančnega položaja.