Nucleic Acids Research, 1993, Vol. 21, No. 6 1369–1380

A study into the effects of protein binding on nucleotide conformation

Stuart L.Moodie and Janet M.Thornton
Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, UK

Received January 5, 1993; Revised and Accepted February 18, 1993

ABSTRACT

In this study, we examine the effects of binding to protein upon nucleotide conformation, by the comparison of X-ray crystal structures of free and protein-bound nucleotides. A dataset of structurally non-homologous protein-nucleotide complexes was derived from the Brookhaven Protein Data Bank by a novel protocol of dual sequential and structural alignments, and a dataset of native nucleotide structures was obtained from the Cambridge Structural Database. The nucleotide torsion angles and sugar puckers, which describe nucleotide conformation, were analysed in both datasets and compared. Differences between them are described and discussed. Overall, the nucleotides were found to bind in low energy conformations, not significantly different from their 'free' conformations except that they adopted an extended conformation in preference to the 'closed' structure predominantly observed by free nucleotide. The archetypal conformation of a protein-bound nucleotide is derived from these observations.

INTRODUCTION

One of the most important properties of proteins is their ability to specifically recognise other molecules, which is essential if they are to function as enzymes or regulatory proteins. Factors affecting protein recognition have significant regulatory effects in cellular metabolism, whether it be affecting an enzyme's affinity for its substrate or altering the specificity of a DNA binding protein for a particular base sequence. Understanding these processes, therefore, is highly desirable and of great practical use, particularly in fields such as drug design and protein engineering.

To make an effective study of such molecular recognition, most information can be gained by studying a particular class of molecule, that is well characterised in its unbound state, but for which there are a large number of protein-bound structures elucidated at the atomic level. One such class of molecule is the nucleotides, for which there is a large number of the protein structures solved, and an unbound species that has been studied extensively. The study of nucleotide binding to protein has importance in its own right, since as a class these molecules are components of nucleic acids, and are very important enzyme cofactors and substrates. Indeed, the phosphorylation of ADP and the hydrolysis of ATP provides the primary method of energy transfer in the cell.

An important first step in the study of protein–nucleotide recognition is to consider how the conformation of free nucleotide differs from that observed in a protein bound nucleotide. Since there are now at least 65 structures of nucleotides bound to protein, and 336 structures of free nucleotides, we decided to carry out such a study by comparing the conformations of the protein-bound nucleotides with those of free nucleotide. Recently, most attention has been directed towards nucleic acid conformation and the only previous study on all protein-bound nucleotide conformations was limited by the small number, and variable quality of solved structures available at the time (1). In general, studies of nucleotide conformation in protein have been confined to specific protein–nucleotide complexes, usually when the structure of the complex has been solved.

For the free nucleotides there is a large body of experimental and theoretical work concerning their conformational preferences (1). The preferred orientations of the main torsion angles describing nucleotide conformation have been determined in these studies, and are generally regarded as the lowest energy states for a particular bond. Information such as this led Yathindra and Sundaralingam (2) to propose the concept of the 'rigid' nucleotide. This suggested that in general the conformation of a nucleotide was quite constrained, with a sugar pucker of $C_{2'}$-*endo* or $C_{3'}$-*endo*, and the torsion angles χ (describing the glycosidic bond), γ ($O_{5'}$-$C_{5'}$-$C_{4'}$-$C_{3'}$), β (P-$O_{5'}$-$C_{5'}$-$C_{4'}$), at *anti*, $+sc$, and *ap*, respectively (figure 1). This conformation may be referred to as a 'closed' conformation, since the $+sc$ rotation of the γ angle forms a relatively compact structure by positioning $O_{5'}$ and its attached phosphate over the furanose ring.

In this study we examine the conformations of the currently determined protein-bound nucleotides, and compare them to the crystal structures of unbound nucleotides. In this way we assess how protein binding affects the conformation of a nucleotide, and, in particular, how far it deviates from the stable conformations found for free nucleotide.

DATA AND METHODS

Generation of the free nucleotide dataset

The dataset of unbound nucleotides was generated from the nucleotide crystal structures in the Cambridge Structural Database (CSD) (3) in a two step process. First, all the nucleotides in the

main database were extracted and used to generate a smaller database, which was used in all subsequent analysis (this reduced the duration of searches in the next stage, because the number of non-nucleotide molecules in the search was significantly reduced). Second, the sub-database was searched for appropriate nucleotide molecules, from which specific torsion angles were calculated. The first stage was performed by the CSD program QUEST89, which generated the sub-database by searching on a pre-defined fragment (figure 2a), that extracted all molecules in the main database containing a dideoxyribose sugar, irrespective of its substituent groups. This reduced the dataset from some 10 000 entries to around 700.

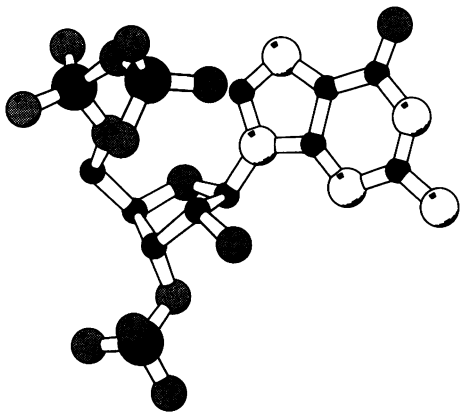


Figure 1. A molecule of 3'-phospho-guanosine-5'-diphosphate in the 'rigid' conformation described by Yathindra and Sundaralingam (2). The figure shows how the 5'-diphosphate group is oriented over the sugar to produce a 'closed' conformation. Carbon atoms are shown as small dark spheres; oxygen, larger light grey spheres; nitrogen, large white spheres; phosphorus, large dark grey spheres.

The second stage was carried out using another CSD program GSTAT89. Queries were performed by searching for molecule fragments in the sub-database. These fragments were also used as a framework on which the GSTAT89 could define and calculate specific torsion angles. To obtain all the torsion angles in the unbound dataset seven basic fragments were necessary (figure 2b–h). Certain flags were set during these searches, in order to exclude the selection of fragments containing spurious linkages between defined atoms, e.g. a cyclic bond formed between $C_{3'}$ and $C_{2'}$ through an oxygen atom.

Outliers from the main distributions were examined visually using QUEST89 and those molecules that differed structurally, from the 'standard' isomeric form shown in figure 3, were eliminated from the dataset. Particularly useful in this task was the δ torsion angle, which is characteristic of the chiralities of the carbon atoms within the furanose ring. The region characteristic of the standard isomer is that between 70° and 160°; angles outside this region were rejected. In general, the most extreme outliers in any of the distributions were from non-standard isomers.

Generation of the bound nucleotide dataset

The bound nucleotide dataset was derived from those nucleotides complexed with proteins in the Brookhaven Protein Data Bank (PDB) (4). Such complexes were identified by examining all the data bank and manually accepting those proteins containing nucleotides. This gave an initial dataset of 65 protein–nucleotide

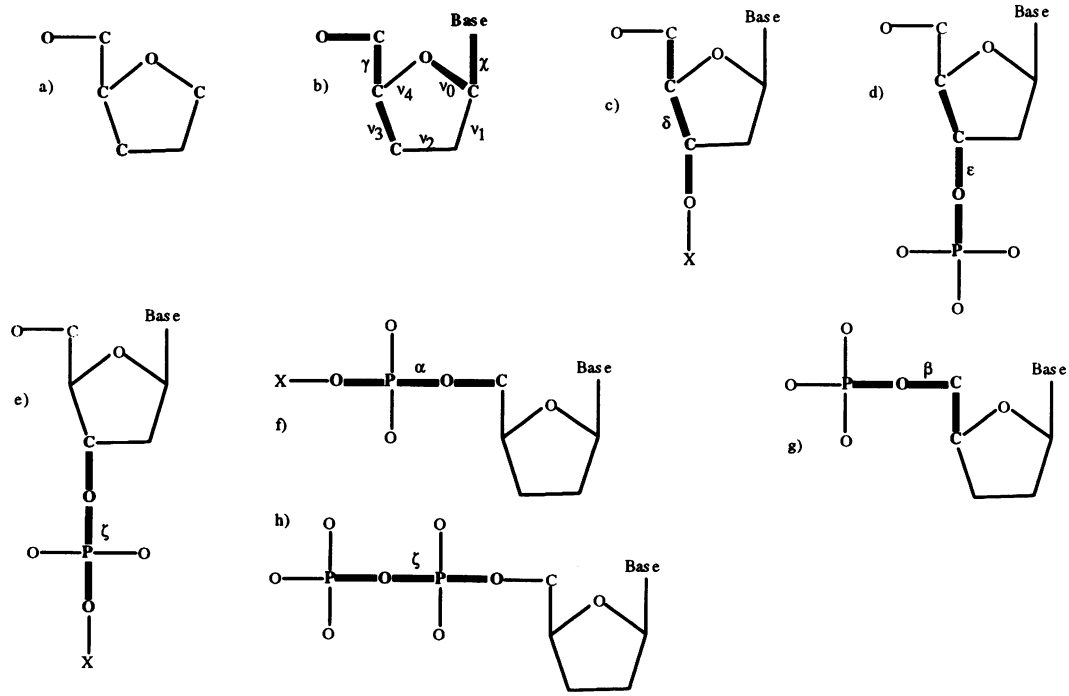


Figure 2. The fragments used in searches of the Cambridge Structural Database. (a) The ribose template used in QUEST89 to obtain the sub-database. (b–h) The seven fragment types used in searches of the sub-database with GSTAT89. Thick bonds and bold type denote the torsion angles defined by each fragment. 'X' in fragments c, e and f can be any atom except hydrogen.

complexes that were then refined to remove all structurally homologous protein–nucleotide complexes. This was done using the protocol and programs of Orengo *et al.* (5), and in particular SSAP, which compares the tertiary structure of two proteins and scores their similarity.

First, the sequences of the proteins in the dataset were compared with each other using a standard sequence comparison algorithm, and grouped into 'families' of proteins with greater than 30 per cent homology. At this stage the initial 65 structures were divided into 32 such families. The 'best' structures, i.e. those with the highest resolution, were then used in another set of pair-wise comparisons (structures that had identical resolution were selected on the basis of the lowest R-factor value),this time directly comparing tertiary structures using the SSAP algorithm. Comparisons with SSAP scores (a normalised, logarithmic score (5)) of greater than 80 were considered to be structurally homologous. This approach was especially important for nucleotide binding proteins, many of which contain similar nucleotide binding motifs, such as the Rossmann fold (6, 7). Although these motifs have little sequential similarity, they are structurally very similar and are the result of divergent evolution, and, therefore, not independent examples of protein–nucleotide binding. An example of this can be seen in this dataset where malate dehydrogenase (4MDH (8)) and lactate dehydrogenase (1LDM (9)) had a sequence identity below 30 per cent, but on structural comparison were put into the same homology family (both contain Rossmann folds).

In this way the 32 sequentially homologous families were merged into 26 structurally homologous ones. Some families contained proteins bound to different types of nucleotide, and these were also included in the final dataset. Hence, the final dataset contained 38 independent protein–nucleotide complexes (table 1) from the original 65 structures. Torsion angles were calculated directly from the PDB files using a program written by SLM.

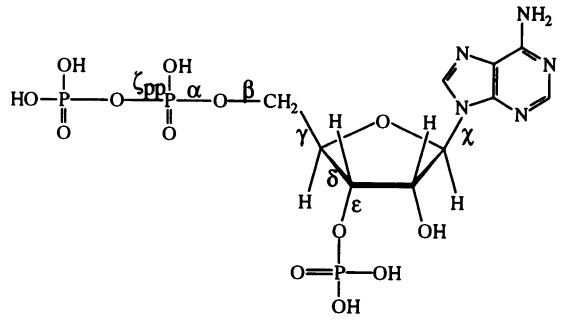


Figure 3. A standard isomer of 3′-phospho-adenosine-5′-diphosphate with torsion angles labelled. The base is in the *endo* position, i.e. on the same side of the ribose ring as the 5′ carbon atom, and the hydroxyl groups on the 2′ and 3′ carbon atoms are in the *exo* position. Changes in the chirality of each of these atoms were found in the initial CSD derived dataset.

Torsion angle and furanose pucker calculations

Torsion angles were defined according to the current IUPAC-IUB nomenclature (10) (table 2), except in the case of the ζ angle.

Table 1. The protein-nucleotide complexes from the Brookhaven Data bank used to generate the bound nucleotide dataset.

Family	Code	Protein Name	Number of Family Members	Bound Nucleotide	Resolution/ Å	Family	Code	Protein Name	Number of Family	Bound Nucleotide	Resolution/ Å
1	1AK3	Adenylate Kinase	1	AMP	1.9	12	3DFR	Dihydrofolate Reductase	4	NADP	1.7
2	2CSC	Citrate Synthase	5	Acetyl CoA	1.7	13	3GAP	Catabolite Gene Activator Protein	1	cAMP	2.5
2	2CTS	Citrate Synthase	5	CoA	2.0	14	4AT1	Aspartate Carbamoyl Transferase	4	ATP	2.6
2	6CTS	Citrate Synthase	5	Citryl-thioether CoA	2.2	14	5AT1	Aspartate Carbamoyl Transferase	4	CTP	2.6
3	1Q21	c-H-Ras Protein	3	GDP	2.2	15	5ADH	Alcohol Dehydrogenase	2	ADP-Ribose	2.9
4	1GOX	Glycolate Oxidase	2	FMN	2.0	15	6ADH	Alcohol Dehydrogenase	2	NAD	2.9
5	1FNR	Ferredoxin Reductase	2	FAD	1.7[4]	16	8RSA	Ribonuclease A	3	N-Acetyl dThy[2]	1.8
5	2FNR	Ferredoxin Reductase	2	2′-Phospho-5′-AMP	3.0	16	9RSA	Ribonuclease A	3	N-Acetyl dUrd[3]	1.8
6	2FCR	Flavodoxin	7	FMN	1.8	16	6RSA	Ribonuclease A	3	Uridine Vanadate	2.0
7	1GD1	GPD[1]	3	NAD	1.8	17	7CAT	Catalase	2	NADP	2.5
8	1LDM	Lactate Dehydrogenase	5	NAD	2.1	18	2SNS	Staphylococcal Nuclease	3	pdTp	1.5
8	5LDH	Lactate Dehydrogenase	5	S-Lac-NAD	2.7	19	2TSC	Thymidylate Synthase	1	dUMP	1.97
9	1PFK	Phosphofructokinase	2	ADP	2.4	20	1COX	Cholesterol Oxidase	1	FAD	1.8
10	1PHH	p-Hydroxybenzoate Hydroxylase	2	FAD	2.3	21	1FBP	Fructose-1,6-Bisphosphate	1	AMP	2.5
10	2PHH	p-Hydroxybenzoate Hydroxylase	2	ADP-Ribose	2.7	22	2SAR	Ribonuclease SA	1	3′-Guanylic Acid	1.8
11	2RNT	Ribonuclease T_1	4	G-2′-p-5′-G	1.8	23	3GRS	Glutathione Reductase	3	FAD	1.54
11	7RNT	Ribonuclease T_1	4	2′-Adenylic Acid	1.9	24	3PGK	Phosphoglycerate Kinase	1	ATP	2.5
11	1RNT	Ribonuclease T_1	4	2′-Guanylic Acid	1.9	25	3TS1	Tyrosyl-Transfer RNA Synthetase	1	Tyrosine Adenylate	2.7
11	5RNT	Ribonuclease T_1	4	pGp	3.2	26	9ICD	Isocitrate Dehydrogenase	1	NADP	2.5

Each structurally-homologous family of proteins is indicated by a number, together with the number of proteins within each family. Protein codes are those used by the Brookhaven Data bank; standard nucleotide abbreviations are used except were specified.

[1]Glyceraldehyde-3-Phosphate Dehydrogenase

[2]N-Acetyl Deoxythymidine

[3]N-Acetyl Deoxyuridine

[4]The ligand coordinates used here are a corrected and better resolved version of the FAD molecule in the original complex.

Table 2. Summary of the unbound and bound nucleotide conformations for the standard nucleotide torsion angles (described in the text)

		Preferred Orientations	
Torsion Angle	**Angle Definition**	**Unbound Dataset**	**Bound Dataset**
α	O-P-$O_{5'}$-$C_{5'}$	*-sc, +sc*	*+sc, -sc*
β	P-$O_{5'}$-$C_{5'}$-$C_{4'}$	140° via *ap* to -120°	130° via *ap* to -110°
γ	$O_{5'}$-$C_{5'}$-$C_{4'}$-$C_{3'}$	*+sc, ap, -sc*	*+sc, ap/-sc*
δ	$C_{5'}$-$C_{4'}$-$C_{3'}$-$O_{3'}$	(70° to 90°)/(130° to 160°)	(70° to 90°)/(130° to 160°)
ε	$C_{4'}$-$C_{3'}$-$O_{3'}$-P	−90° to −160°	-80° to -150°
ζ_{pp}	P-O-P-$O_{5'}$	*+ac, ap/-sc, -ac*	free rotation
χ_A	$O_{4'}$-$C_{1'}$-N_9-C_4	*anti, syn*	*anti*
χ_G	$O_{4'}$-$C_{1'}$-N_9-C_4	*anti, syn*	*anti, syn*
χ_{pyr}	$O_{4'}$-$C_{1'}$-N_1-C_2	*anti*	*anti*
P	(see text)	$C_{2'}$-*endo*, $C_{3'}$-*endo*	$C_{2'}$-*endo*, $C_{3'}$-*endo*

The preferences are for nucleotides containing all base types with the exception of the χ torsion angles, which were subdivided into those nucleotides containing adenine (χ_A), guanine (χ_G) and pyrimidine bases (χ_{pyr}). *P* is the pseudorotation angle. Preferred orientations are listed from left to right, those on the left being the most favoured.

In order to signify the different nature of the P-O bond in a pyrophosphate group from that in a phosphodiester linkage, which is not directly addressed by the IUPAC-IUB standard, a subscript of pp was used to signify the former. Their orientations were described using the Klyne−Prelog system (10),.which divides torsion angle orientations into the six equal sectors of *sp*, *+sc*, *+ac*, *ap*, *−ac* and *−sc*.

The pucker of the furanose ring was described by its pseudorotation phase angle, as derived by Altona and Sundaralingam (11). This uses the five endocyclic torsion angles of the furanose ring to produce an angle using the equation:

$$\tan P = \frac{(V_4 + \nu_1) - (\nu_3 + \nu_0)}{2 \cdot \nu_2 \cdot (\sin 36° + \sin 72°)}$$

$$\text{if } V_2 < 0 \text{ then } P = P + 180°$$

that reflects the conformation of its pucker. These angles correspond to a more descriptive nomenclature, that describes whether an atom is above or below the plane of the other atoms in the ring. Thus, $C_{2'}$-*endo* would describe a pucker with $C_{2'}$ on the same side of the ring as $C_{5'}$ and above the other atoms in the plane of the furanose ring.

These angles were represented in the classic manner using the 'pie' plots to represent a given angle by a radial line, augmented by a histogram indicating the population of observations in 10° slices (figure 4). For torsion angles this angle ranged from −180° to 180° with sectors being appropriately shaded according to the 'staggered' orientations of *−sc*, *+sc*, and *ap*. Pseudorotation phase angles, on the other hand, ranged from 0° to 360° and were also appropriately shaded at the favoured $C_{2'}$-*endo* and $C_{3'}$-*endo* regions.

RESULTS AND DISCUSSION

The unbound dataset

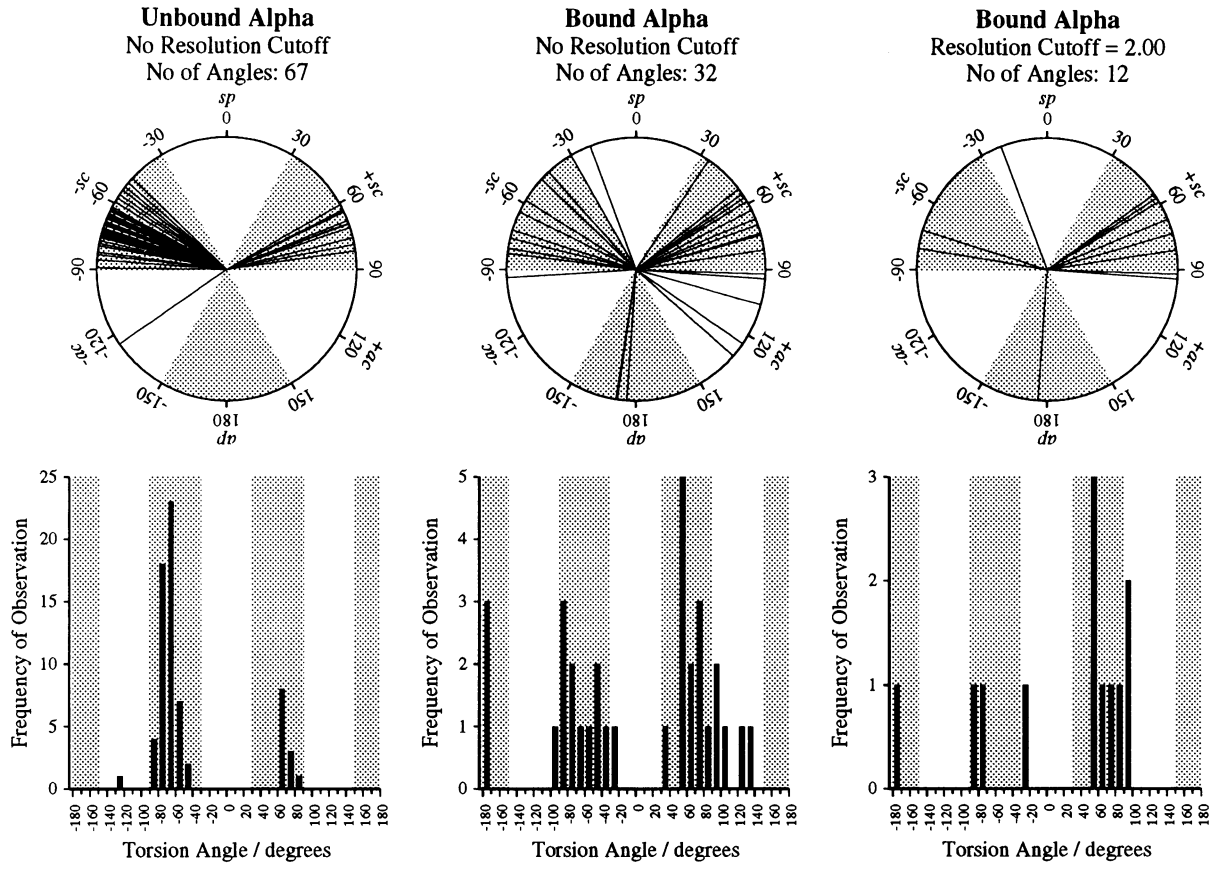
The dataset of non-protein-bound nucleotides contained 336 structures, which were a mixture of deoxyribo- and ribonucleosides, nucleotides, dinucleotides and paired dinucleotides. Although somewhat heterogeneous, it conformed well to the preferred torsion angle orientations reviewed (1) and reported previously (2, 12). Data for the torsion angle ζ ($C_{3'}$-$O_{3'}$-P-$O_{5'}$) were not used here, since there were no examples of this bond in the bound dataset. However, its preferred distribution was observed to be *−sc*, as would be expected from previous studies (1). In view of this agreement with previous work, little analysis was performed on the dataset itself, except where relevant for comparison with the protein bound dataset. A summary of the conclusions drawn from these data is provided in table 2, with pie-plots shown in figure 4(a−h).

Conformations of protein bound nucleotide compared to unbound nucleotide

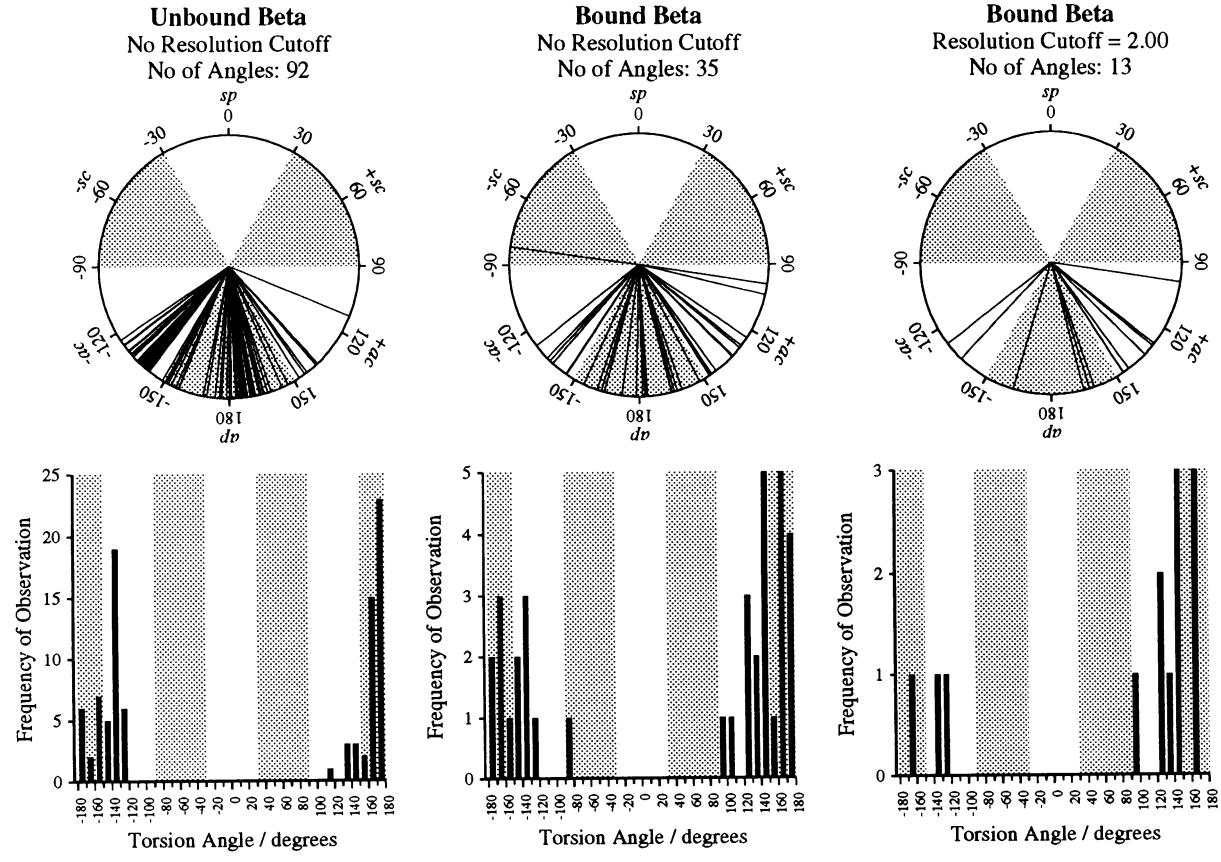
The preferences of the bound dataset are shown together with those of the unbound dataset in figure 4, with the preferences for both summarised in table 2.

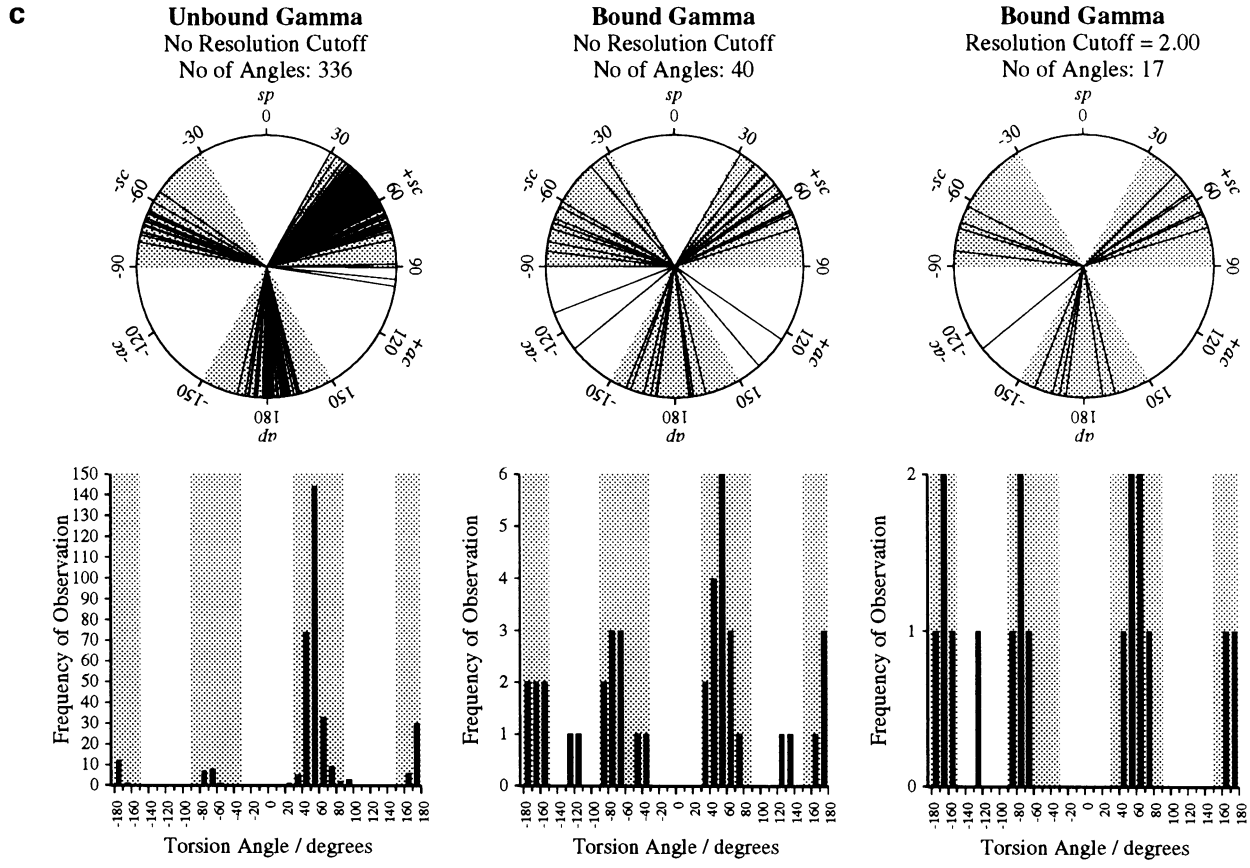
The α angle. The most preferred orientation of α in the bound nucleotide dataset is *+sc* (figure 4a, table 2). This is preferred over *−sc*, with weaker preferences observed for *ap* and *+ac*,

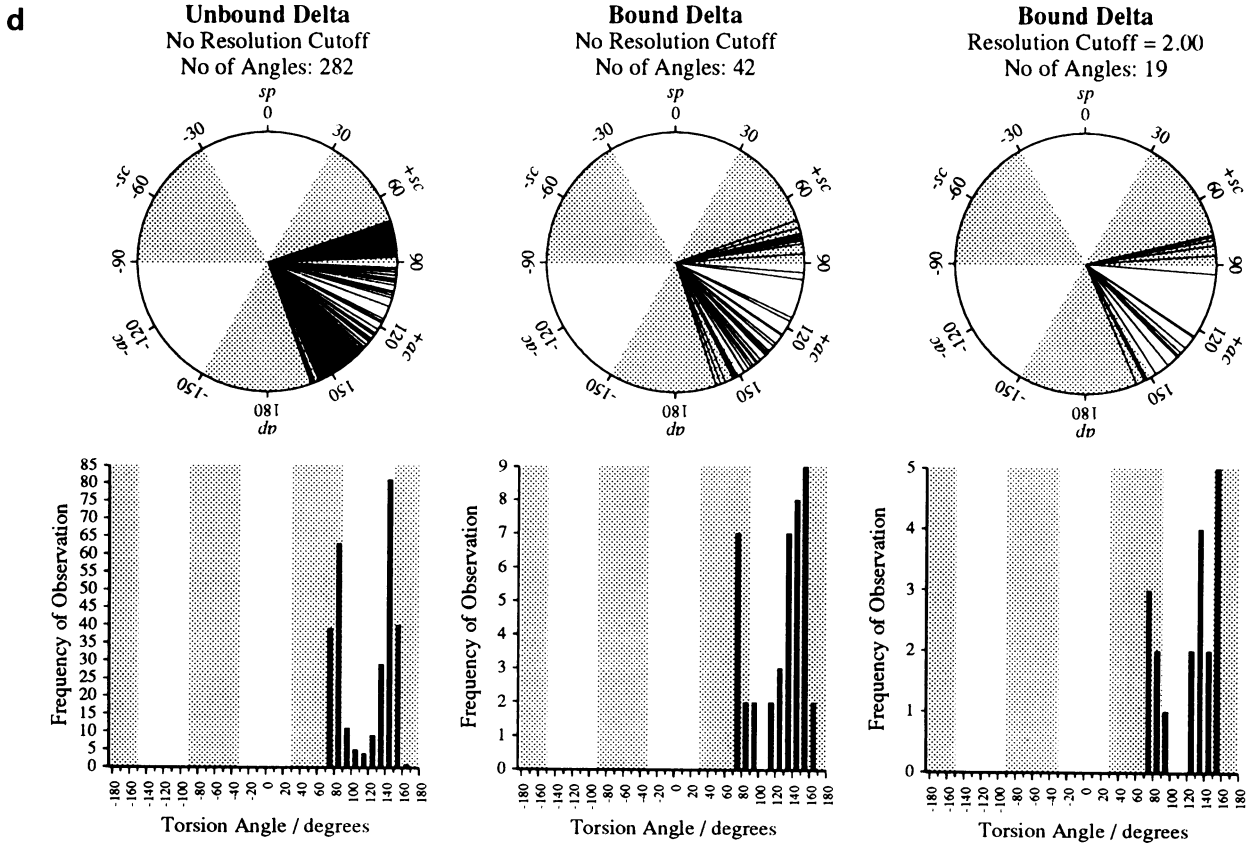
a


Unbound Alpha
No Resolution Cutoff
No of Angles: 67
Bound Alpha
No Resolution Cutoff
No of Angles: 32
Bound Alpha
Resolution Cutoff = 2.00
No of Angles: 12
sp
dp
+sc
-sc
+ac
-ac
Frequency of Observation
Torsion Angle / degrees

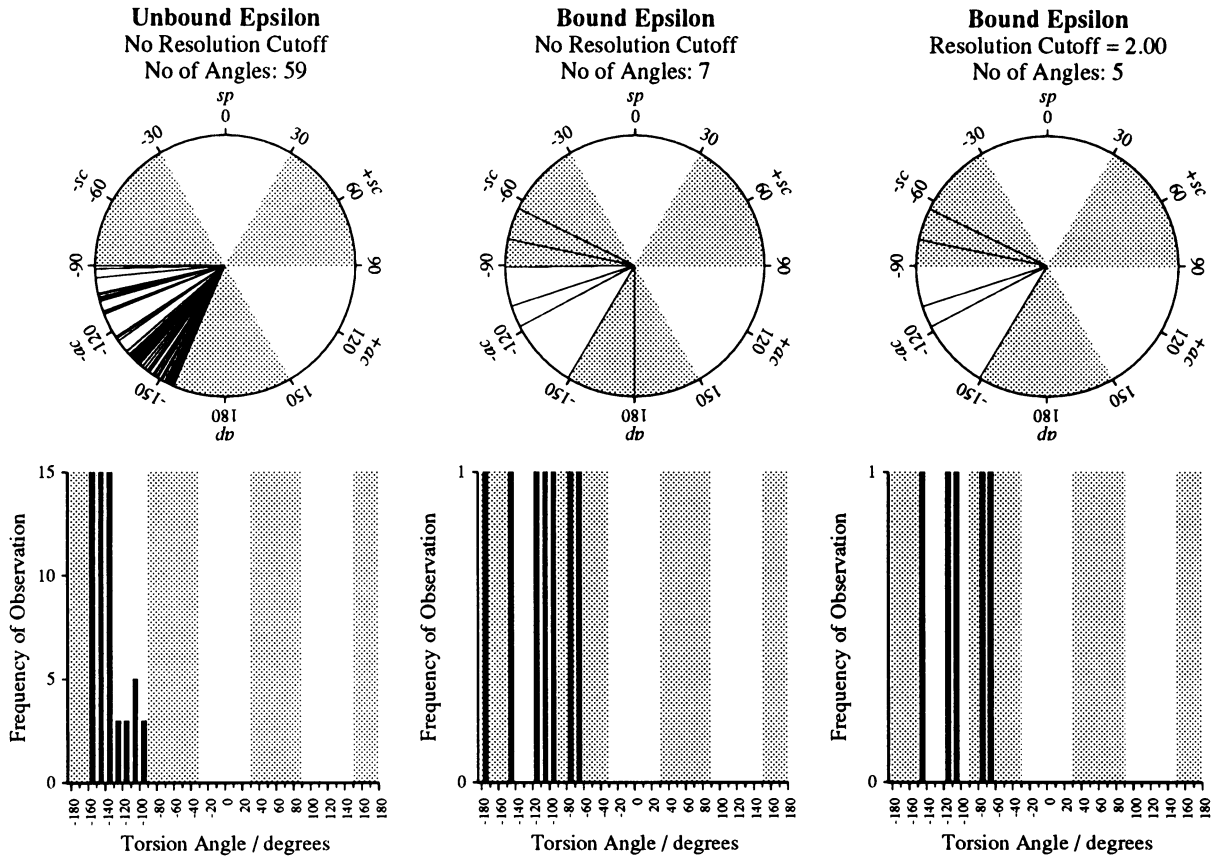
b


Unbound Beta
No Resolution Cutoff
No of Angles: 92
Bound Beta
No Resolution Cutoff
No of Angles: 35
Bound Beta
Resolution Cutoff = 2.00
No of Angles: 13
sp
dp
+sc
-sc
+ac
-ac
Frequency of Observation
Torsion Angle / degrees

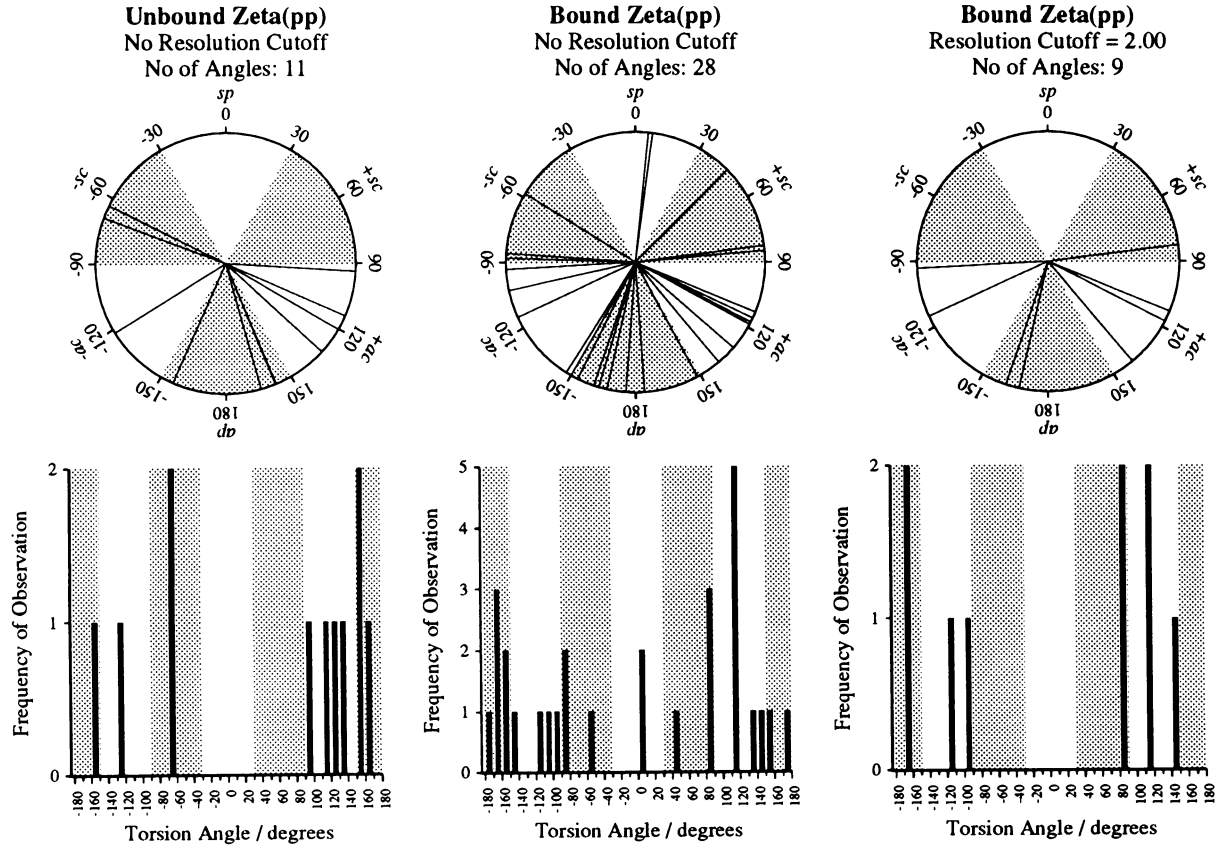

c
Unbound Gamma
No Resolution Cutoff
No of Angles: 336
Bound Gamma
No Resolution Cutoff
No of Angles: 40
Bound Gamma
Resolution Cutoff = 2.00
No of Angles: 17
sp
ap
+sc
-sc
+ac
-ac
Frequency of Observation
Torsion Angle / degrees


d
Unbound Delta
No Resolution Cutoff
No of Angles: 282
Bound Delta
No Resolution Cutoff
No of Angles: 42
Bound Delta
Resolution Cutoff = 2.00
No of Angles: 19
sp
ap
+sc
-sc
+ac
-ac
Frequency of Observation
Torsion Angle / degrees

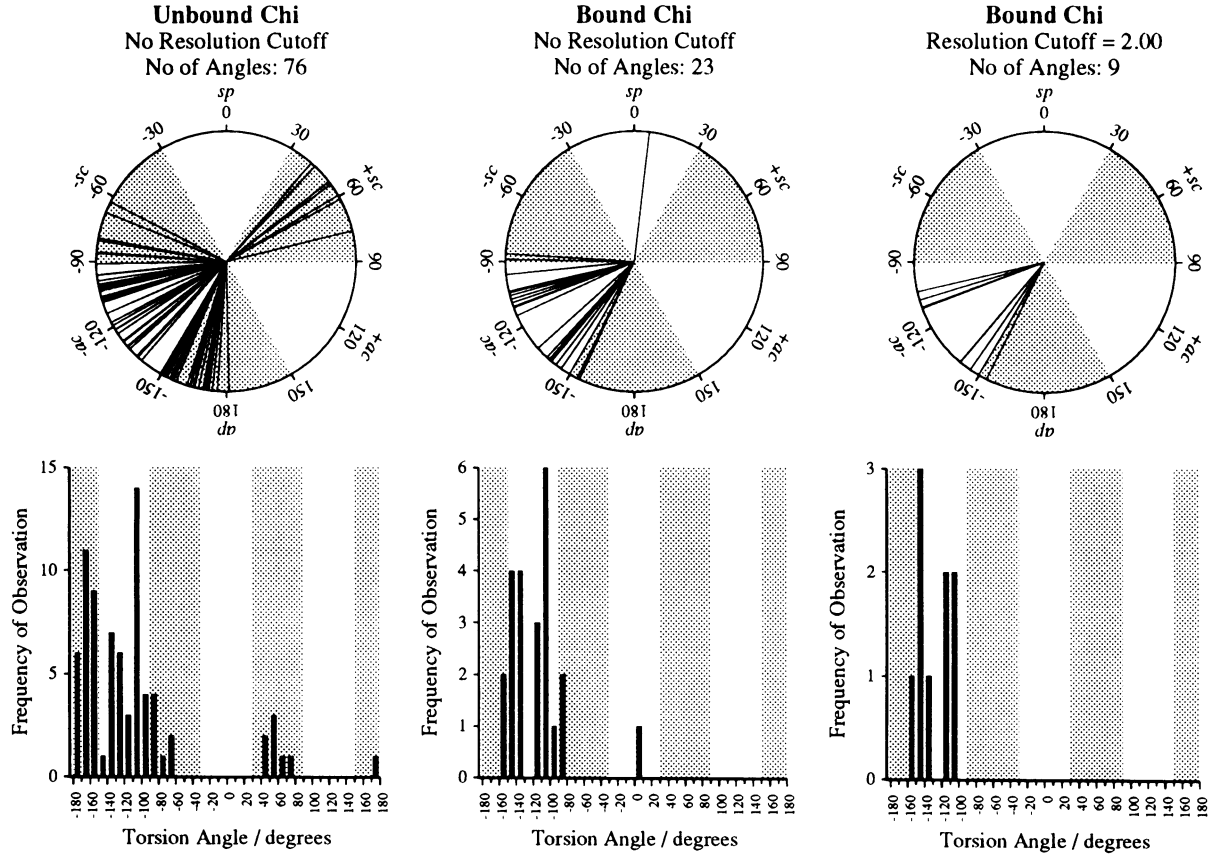
e


Unbound Epsilon
No Resolution Cutoff
No of Angles: 59
Bound Epsilon
No Resolution Cutoff
No of Angles: 7
Bound Epsilon
Resolution Cutoff = 2.00
No of Angles: 5
sp
ap
+sc
+ac
-sc
-ac
Frequency of Observation
Torsion Angle / degrees

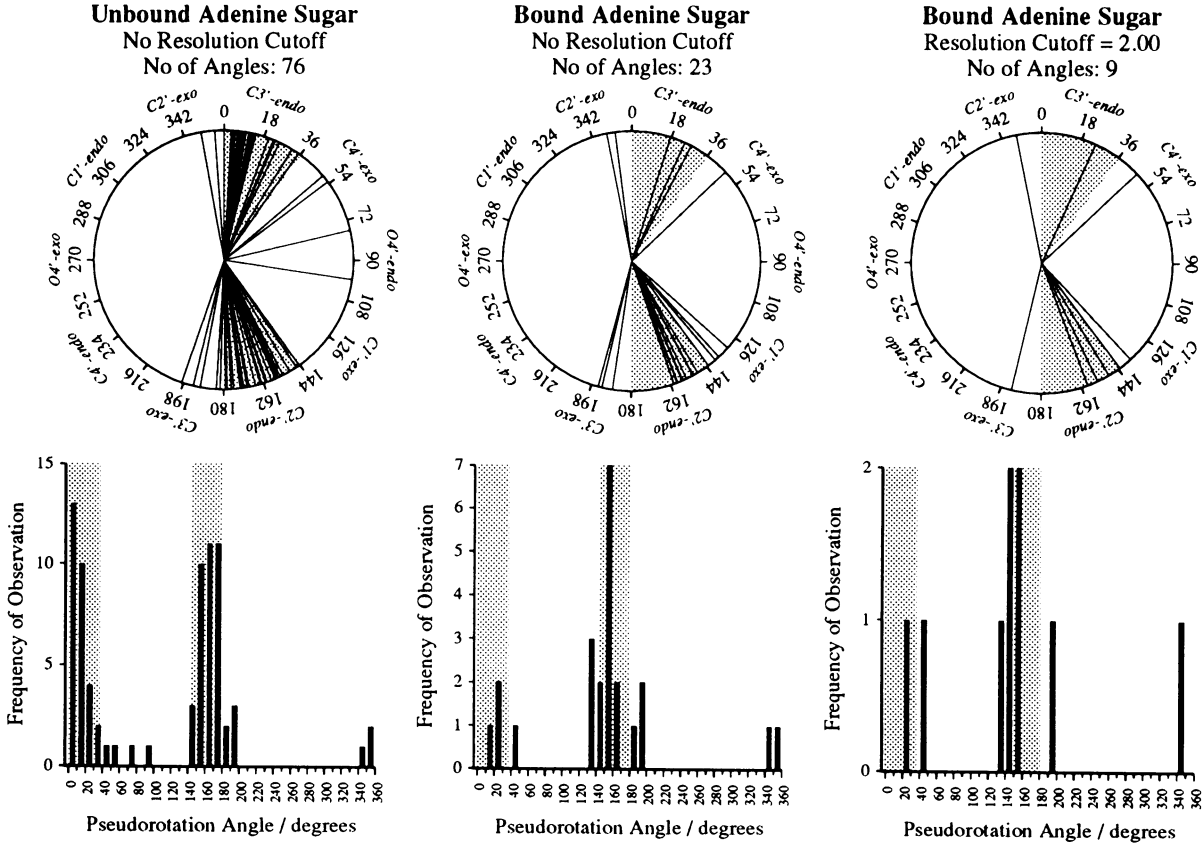
f


Unbound Zeta(pp)
No Resolution Cutoff
No of Angles: 11
Bound Zeta(pp)
No Resolution Cutoff
No of Angles: 28
Bound Zeta(pp)
Resolution Cutoff = 2.00
No of Angles: 9
sp
ap
+sc
+ac
-sc
-ac
Frequency of Observation
Torsion Angle / degrees

g


Unbound Chi
No Resolution Cutoff
No of Angles: 76
Bound Chi
No Resolution Cutoff
No of Angles: 23
Bound Chi
Resolution Cutoff = 2.00
No of Angles: 9
sp
ap
+sc
-sc
Frequency of Observation
Torsion Angle / degrees

h


Unbound Adenine Sugar
No Resolution Cutoff
No of Angles: 76
Bound Adenine Sugar
No Resolution Cutoff
No of Angles: 23
Bound Adenine Sugar
Resolution Cutoff = 2.00
No of Angles: 9
C2'-exo
C3'-endo
C4'-exo
O4'-endo
C1'-exo
C2'-endo
C3'-exo
C4'-endo
O4'-exo
C1'-endo
Frequency of Observation
Pseudorotation Angle / degrees

although these are only found, with the exception of one angle (2CSC (13)) at *ap*, among structures of less than 2Å resolution. This may suggest that the $+sc$ distribution is in fact broader than that observed in the unbound dataset, extending far into $+ac$.

The unbound preference for $-sc$ over the other staggered orientations can be explained by a combination of the *gauche* effect (14–17), and steric constraints. The former describes how the lone-pair electrons of the oxygen in the $O_{5'}$-P bond are partially donated to the polar $P^{\delta+}$-$O^{\delta-}$ bond when they are anti-parallel, i.e. at $+sc$ and $-sc$ (figure 5). This means an orientation, such as *ap*, where they are not anti-parallel is energetically disfavoured. The preference for $-sc$ over $+sc$, for the two orientations favoured by the *gauche* effect, can be attributed to steric effects on the group linked to the phosphate, since at an orientation of $-sc$ any such groups would be more distant from both base and sugar (with β and γ at their preferred orientations of *ap* and $+sc$) than at $+sc$. The loss of this preference for $-sc$ in the bound dataset probably reflects the significant reduction in the preference of γ for $+sc$ (see below and table 2), since with γ at $-sc$, a $+sc$ orientation appears to be sterically favoured for α. Figure 6 shows how an orientation of $-sc$, $-sc$ for α and γ is noticeably less preferred than one of $+sc$, $-sc$. With γ at $+sc$ or *ap*, this distinction does not seem to apply. Of those angles at $+ac$ and *ap*, all, except tyrosine adenylate in 3TS1 (18), are from coenzyme structures of NAD, FAD, or coenzyme A. In these cases it would seem that the accommodation of such large groups attached at the 5′ phosphate compensates for any extra energy necessary to accommodate a less favourable *ap* orientation at α. The outlier in the unbound dataset at $+ac$ is due to the only crystal structure of NAD (19), which may indicate that the α angle in such coenzyme structures may be less constrained than in 'standard' nucleotides.

The β angle. The observed preferences of the bound and unbound datasets for β are shown in figure 4b, and summarised in table 2. Clearly there was no significant difference between both datasets, a broad range centred on *ap* being almost exclusively favoured in both. In the bound dataset this broad range appears to be shifted anti-clockwise by 10°, but this is the only difference apart from an additional number of outliers in the bound dataset. Two of these, 2FNR (20) and 5RNT (21), were low resolution structures, with relatively low electron densities around the nucleotide, while the other, 2RNT, was of a higher resolution, but the GpG to which it was bound was less defined at its 3′ end (22).

The orientation of β is primarily influenced by steric hindrance, especially when γ is at $+sc$, since this puts $O_{5'}$ above the furanose ring making any *syn* orientation very sterically unfavourable (23–27). Such hindrance may not be so acute as γ approaches *ap*, since the one outlier in the unbound dataset at 112° for β (figure 4b), also has γ at *ap*. It is noteworthy that all the outliers in the bound dataset also have γ angles at *anti*.

The γ angle. Data from the bound dataset are shown in figure 4c, from which it is readily apparent that the predominance of $+sc$ is very much reduced (these observations are summarised in table 2). In addition to the three staggered conformations two angles were observed at $+ac$, 4AT1 (28) and 5LDH (29), and two at $-ac$, 2FNR (20) and 8RSA (30). None of these structures are resolved to better than 2 Å, except 8RSA, which was covalently bound to a histidine residue in the nucleotide binding site of the protein.

The unbound orientation of γ at $+sc$ is attributed mainly to electrostatic interactions between the base and $O_{5'}$, which stabilises the orientation of this atom over the furanose ring (1) (figure 1). The bound data here (figure 4c) do indicate a preference for $+sc$. They also suggest that the number of nucleotides in an extended conformation (i.e. *ap*, $-ac$, or $-sc$), significantly exceed those in a 'closed' one. As a general observation, the distributions of the three staggered conformations within the bound dataset, especially at $-sc$ and *ap*, do not appear to be as tightly clustered as in the unbound dataset. This may indicate that the $C_{4'}$-$C_{5'}$ bond is less constrained in a nucleotide that is bound to protein. The outliers at $+ac$ and $-ac$, noted above, may be extreme examples of this.

The δ angle. The distributions of the bound and unbound datasets can be seen in figure 4d and are summarised in table 2. These show that there is no significant difference between the datasets for δ (figure 4d). The $C_{4'}$-$C_{3'}$ bond described by δ is part of the furanose ring in a nucleotide, which places a significant constraint on its rotation. Theoretical energy calculations (31) agree with the orientation range observed in both datasets of between 70° and 160° (figure 4d, table 2).

The ϵ angle. The preferred orientations of ϵ in the bound and unbound datasets are shown in figure 4e and summarised in table 2. The summary of the preferred bound orientation in the table has been a rather liberal interpretation, since there are only six observations. However, none of the angles in the bound dataset deviated significantly from the region preferred in the unbound dataset. Two angles were lost at higher resolution, of which one lay outside the preferred unbound orientation at 180°, 5RNT (22), and had an especially low resolution of 3.1 Å. Of the four higher resolution structures, there were two, 2SNS (32) and 2SAR (33), which lay outside the preferred region of the unbound dataset.

The orientation of ϵ in the unbound dataset is principally influenced, in a similar manner to b, by steric hindrance. The clashing groups in this case are the 3′-phosphate, and the sugar, which are brought into proximity with each other in the positive hemisphere of the Klyne–Prelog cycle and at $-sc$ orientations. In a previous study the distribution of ϵ angles for a tRNA molecule (34) was broader than that observed in the unbound dataset here, ranging from 170° through 180° to −65°. This suggests that where it is necessary to stabilise the structure, a broader range of orientations at ϵ is permitted. The outliers in the bound dataset, noted above, may therefore be indicative of a broader range for ϵ due to the stabilising effects of the protein environment.

Figure 4(a–h). 'Pie plots' and histograms showing the distributions of the torsion (a–g) and pseudorotation phase angles (h) of the bound and unbound nucleotide datasets. The unbound data are displayed in the left plot, bound data in the central and right plots. Observations from all the structures in the bound dataset are shown in the central plot, while only those resolved to less than 2 Å are shown in the right plot. Observations were obtained from nucleotides of all base types except for those of plots g and h, which were restricted to those nucleotides containing adenine only.

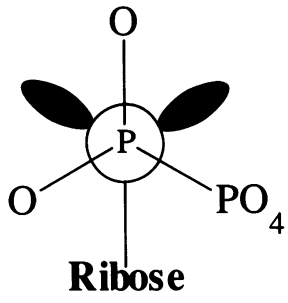


Figure 5. A Newman projection showing the orientation of the P-O pyrophosphate bond, defined by α, in a $+sc$ orientation favoured by the *gauche* effect. The phosphate group is in the anti-parallel orientation required for the partial donation of electrons from the lone-pair orbital.

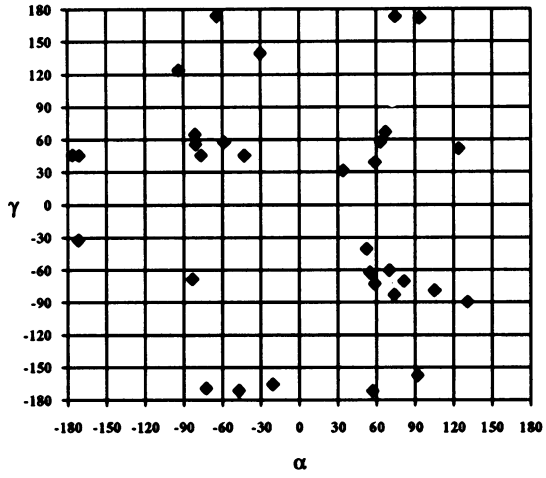


Figure 6. A scatter plot showing the distribution of γ angles in relation to α angles. In particular note the preference of α for $+sc$ (30° to 60°) when γ is at $-sc$ (−30° to −60°) over its preference for $-sc$ when γ is at the same orientation.

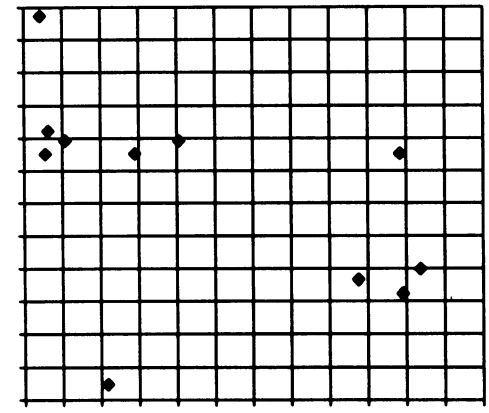


Figure 7. A scatter plot showing the relationship between ζ and γ. Note how γ does not occupy $+sc$ (30° to 60°) whenever ζ is at *sp* or $+sc$ (−30° to 90°).

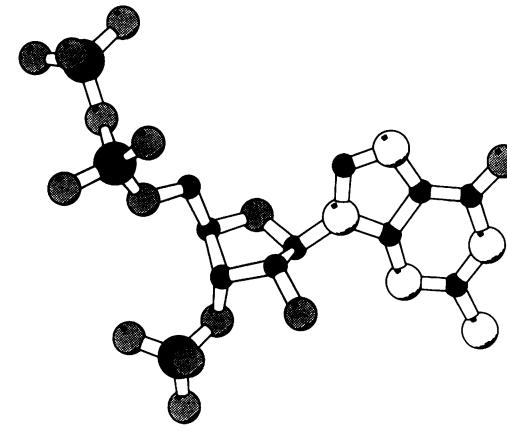


Figure 8. A molecule of 3′-phospho-guanosine-5′-diphosphate in the idealised extended conformation described in the text. The figure shows how the pyrophosphate group extends away from the base and sugar providing a greater potential for interaction with the protein environment. Atom representations are as described for figure 1.

The ζ_{pp} angle. The distributions of the angles in the bound and unbound datasets can be seen in figure 4f, and are summarised in table 2. The most striking feature of the two datasets is the disperse nature of their distributions, which agrees with previously observed data for unbound nucleotides (19) in that there is little constraint on the orientation of this bond in a pyrophosphate group. Such freedom is permitted by the unusually large distance between the two phosphates of the pyrophosphate linkage, caused by relatively long P-O bond lengths and a P-O-P angle significantly more obtuse than the ideal tetrahedral angle of 109°. Thus, eclipsed conformations can be seen in the unbound structure at $+ac$ and $-ac$. However, the fact that none are seen at *sp* (figure 4f) may be due to steric hindrance between the β-phosphate of the pyrophosphate group and the base of the nucleotide, which clash severely if γ is at its preferred $+sc$ orientation. This is supported by the bound data, where structures containing ζ_{pp} *at sp* and $+sc$ (not observed in the unbound data) were observed in all cases to have γ oriented at either $-sc$ or *ap* (figure 7).

The χ angle. The χ angle is known to be correlated with base type, *syn* being very much disfavoured in pyrimidine nucleosides and nucleotides (1). In the bound dataset, only the adenine nucleotides were numerous enough to enable a meaningful comparison with the unbound dataset for both angles and so only these were considered in detail (figure 4g, table 2). The protein bound nucleotides show a much greater preference for *anti* than the unbound adenine nucleotides; the small preference for *syn* in the latter being lost when bound to protein. There is one outlier, 1FBP (35), that adopts an orientation of *sp*. This orientation was not observed in any of the structures of the unbound dataset and may suggest it is an orientation made favourable by protein binding. However, the structure is of a relatively low resolution for our purposes, 2.5 Å. Of the other bases, there were five examples of guanine angles, two at *syn*, and three at *anti*, and a total of five pyrimidine angles, all at *anti* (data not shown).

The bases of nucleotides and nucleosides prefer an *anti* or high-*anti* (−90° to −60°) orientation when unbound, since this is sterically more favourable than a *syn* orientation that places the more bulky parts of the base over the furanose ring (1). For the pyrimidines, this bulky group is the oxygen at position 2 of the ring, while for the purines it is the second, six membered ring of the molecule. In addition to these repulsive influences the bases

also favour *anti* when unbound, because this facilitates the attractive interactions between $O_{5'}$ and the base when γ is at $+sc$ (see above). The almost exclusive loss of angles at *syn* for the adenine nucleotides and the narrowing of its distribution about $-ac$, suggests that an extended structure is most preferred when they complex with protein. With only five observations it is not possible to determine whether this suggestion holds for the pyrimidine bases, however, with two observations at $+sc$, it is quite clear that this is unlikely to be the case for guanine bases.

Guanine nucleosides and nucleotides have a significantly greater preference for *syn* than the other bases, since this orientation is stabilised by hydrogen bonding between an amino group at position two of the base and $O_{5'}$, when γ is at $+sc$ (1). These stabilising interactions were observed in the protein dataset for ribonuclease T_1, when complexed with 2'-guanylic acid (1RNT (36)), and guanylyl-2',5'-guanosine (2RNT (22)). In both cases, however, the $O_{5'}$ atom interacting with guanine, was not attached to a phosphate and would presumably be less constrained by any favourable phosphate interactions with the protein. In the three structures containing a 5' guanine nucleotide, all were observed at *anti*, corresponding to the adenine observations above.

The pseudorotation angle. The distributions for the pseudorotation angle of the sugar observed in the bound and unbound datasets are shown in figure 4h, and summarised in table 2. The $C_{2'}$-*endo* pucker is clearly more preferred over $C_{3'}$-*endo* in the bound dataset, with the latter being very much reduced relative to the former. Theoretical and experimental studies have shown that the $C_{2'}$-endo pucker is generally more stable in deoxyribose nucleosides and nucleotides, but is energetically equivalent with $C_{3'}$-*endo* in those molecules containing ribose (37). Thus, the preference for $C_{2'}$-*endo* in the unbound dataset reflects the mixture of deoxyribo- and ribo-nucleosides and nucleotides in the unbound dataset. This makes the preference for $C_{2'}$-endo more distinct in the unbound dataset, since it contained only four deoxyribose structures. An explanation for this behaviour is not clear.

CONCLUDING REMARKS

In examining the data drawn from these two datasets it is important to bear in mind that the data were derived from X-ray crystal structures. By definition a crystallised molecule is in a somewhat different state to that of a molecule in solution, especially if one considers the unusual conditions or co-crystallising molecules sometimes used to induce crystallisation (38). Despite this, the structures determined by this method are usually those with low potential energy, and, as we assume here, are preferred conformational states. The validity of this assumption for the unbound data has been verified by the large number of spectroscopic and theoretical studies that have been performed on these molecules in solution (1). However, equivalent solution structures of protein–nucleotide complexes do not exist, since proteins that bind nucleotides are generally too large for 2D-NMR determination. An *in situ* NMR study of dTpdA in the active site of staphylococcal nuclease has been carried out and although it attributed some conformational changes in the binding site to crystal packing, the main torsion angles of the nucleotide did concur with those observed in the bound dataset (39).

Another, important factor in the bound dataset was the resolution of crystallographic data. In the bound dataset this varied from as low as 3.2 Å for 5RNT to 1.5 Å for 2SNS, but in the unbound dataset was better than 1 Å for all structures. Although structures of lower resolution have a greater degree of inaccuracy in their stereochemistry (40), and are therefore less reliable in these studies, the bound dataset was not large enough to enable the extra data provided by such structures to be discarded. However, these inaccuracies could not be ignored, and a resolution cut-off was used in all analyses. The value of the resolution, however, is only an average for the structure as a whole, of which the nucleotide is but a small part. Thus, a high resolution structure may contain a bound nucleotide that is poorly defined. This is the case in 2RNT (22), where a structure of 1.8 Å resolution is much less well resolved at the GpG molecule in the complex. However, since there will inevitably be errors associated with the torsion angle data, we have restricted our discussion to broad areas of torsion angle space rather than discussing specific values.

The most obvious conclusion that can be derived from these data is how little conformational change is undergone by a nucleotide when it binds to protein. The major change occurs at the $C_{5'}$-$C_{4'}$ bond, defined by γ, which loses its unbound preference for $+sc$ and a 'closed' conformation, in favour of the other staggered orientations of $-sc$ and *ap*. This change in preference was first noted by Saenger (1), in the 16 protein–nucleotide complex structures solved at the time (1984). Despite the low resolution or poor quality of several of the structures Saenger observed that only two complexed nucleotides had a γ torsion angle oriented at $+sc$. In contrast, our data suggest that $+sc$ is still the most favoured of the three staggered orientations ($+sc$, $-sc$, *ap*), and in agreement with Saenger, the 'closed' conformation it forms is significantly less preferred than the extended conformations generated by orientations of $-sc$ or *ap*.

Correlated to this change in orientation of γ, appears to be a shift in the preference of α from a preferred orientation of $-sc$ in the unbound dataset to one of $+sc$ when protein bound. Similarly, the steric constraints which act on the orientations of the β and ζ_{pp} angles may be reduced, to such an extent in the latter case as to permit its free rotation. The prevalence of these orientations, rarely observed in the unbound dataset is, we suggest, a direct correlation with the change in preferred orientation of the $C_{5'}$-$C_{4'}$ bond, described by γ, and not directly due to the influence of the protein environment. However, there appears to be a broadening of the $+sc$ distribution at a, that may be promoted by protein binding of the larger nucleotide coenzymes.

The extended conformation formed by the rotation of $C_{5'}$-$C_{4'}$ to $-sc$ or *ap* was referred to by Saenger as an 'open' conformation. The reorientation of γ opens up the base and ribose to the protein environment, enabling the formation of a greater number of favourable inter-molecular contacts (1). Our data concur with this conclusion, but in addition we also observe an almost exclusive preference for *anti* in the adenine nucleotides, which may be applicable to guanine nucleotides as well. A narrowing of this distribution also suggests that adenine may adopt an orientation that projects optimally into the protein environment. In a similar manner to Yathindra and Sundaralingam, who defined the conformation of a 'rigid' nucleotide (2) (figure 1), we may define the optimal conformation of a protein bound nucleotide as follows: α, $+sc$; β, *ap*; γ, $-sc$ or *ap*; δ, 130° to 160°; ϵ $+ac$; χ, *anti* ($-ac$); sugar pucker, $C_{2'}$-*endo* (figure 8).

These conclusions provide a guide to those concerned with the experimental structure determination of protein nucleotide complexes, or prediction of the favoured, and possibly forbidden, nucleotide conformations. More generally, however, our data show that the nucleotides do not vary significantly from their most stable unbound conformations, and that they prefer to bind in low energy extended conformations, presumably to maximise their contact with the protein environment. Similar extended conformations have been observed in the binding of peptides to protein (41–49). Therefore, the adoption of extended conformations, which combine a low-energy 'internal' conformation with maximal recognition potential may be a recurring feature of protein-ligand recognition and binding.

REFERENCES

1. Saenger, W. (1984) Principles of Nucleic Acid Structure. Springer-Verlaf, New York.
2. Yathindra, N. and Sundaralingam, M. (1973) *Biopolymers*, **12**, 297–314.
3. Allen, F. H., Kennard, O. and Taylor, R. (1983) *Acc. Chem. Res.*, **16**, 146–153.
4. Bernstein, F. C., Koetzle, T. F., William, G. J. B., Meyer, E. J., Jr, Brice, M. D., Rogers, J. R., Kennard, O. and Shimanouchi, T. (1977) *J. Mol. Biol.*, **112**, 532–542.
5. Orengo, C. A., Flores, T., Taylor, W. R. and Thornton, J. M. (Submitted to Protein Engineering).
6. Rao, S. T. and Rossmann, M. G. (1973) *J. Mol. Biol.*, **76**, 241–256.
7. Rossmann, M. G., Liljas, A., Branden, C.-I. and Banaszak, L. J. (1975) In (ed.), The Enzymes. Academic Press, New, York, XI, 61–102.
8. Birktoft, J. J., Rhodes, G. and Banaszak, L. J. (1989) *Biochemistry*, **28**, 6065–6081.
9. Abad-Zapatero, C., Griffith, J. P., Sussman, J. L. and Rossmann, M. G. (1987) *J. Mol. Biol.*, **198**, 445–467.
10. IUPAC-IUB Joint Commission on Biochemical Nomenclature. (1983) *Eur. J. Biochem.*, **131**, 9–15.
11. Altona, C. and Sundaralingam, M. (1972) *J. Amer. Chem. Soc.*, **94**, 8205–8212.
12. de Leeuw, H. P. M., Haasnoot, C. A. G., Altona, C. (1980) *Isr. J. Chem.*, **20**, 108–126.
13. Karpusas, M., Holland, D. and Remington, S. J. (1991) *Biochemistry*, **30**, 6024–6031.
14. Lemieux, R. U. (1971) *Pure Appl. Cem.*, **25**, 527–548.
15. Wolfe, S. (1972) *Acc. Chem. Res.*, **5**, 527–548.
16. Radom, L., Hehre, W. J. and Pople, J. A. (1972) *J. Amer. Chem. Soc.*, **94**, 2371–2381.
17. Brunck, T. K. and Weinhold, F. (1979) *J. Amer. Chem. Soc.*, **101**, 1700–1709.
18. Brick, P., Bhat, T. N. and Blow, D. M. (1988) *J. Mol. Biol.*, **208**, 83–98.
19. Saenger, W., Reddy, B. S., Muhlegger, K. and Weimann, G. (1977) *Nature*, **267**, 225–229.
20. Karplus, P. A., Daniels, M. J. and Herriot, J. R. (1991) *Science*, **251**, 60–66.
21. Lenz, A., Heinemann, U., Maslowska, M. and Saenger, W. (1991) *Acta Crystallogr.*, **B47**, 512–527.
22. Koepke, J., Maslowska, M., Heinemann, U. and Saenger, W. (1989) *J. Mol. Biol.*, **206**, 475–488.
23. Broyde, S. B., Wartell, R. M., Stellman, S. D., Hingerty, B. and Langridge, R. (1975) *Biopolymers*, **11**, 25–56.
24. Pullman, B. and Saran, A. (1976) *Prog. Nucleic Acid Res. Mol. Biol.*, **18**, 215–322.
25. Lakshminarayanan, A. V. and Sasisekharan, V. (1969) *Biopolymers*, **8**, 489–503.
26. Saran, A. and Govil, G. (1971) *J. Theor. Biol.*, **33**, 407–418.
27. Pullman, B., Perahia, D. and Saran, A. (1972) *Biochim. Biophys. Acta*, **269**, 1–14.
28. Stevens, R. C., Gouaux, J. E. and Lipscomb, W. N. (1990) *Biochemistry*, **29**, 7691–7701.
29. Grau, U. M., Trommer, W. E. and Rossmann, M. G. (1981) *J. Mol. Biol.*, **151**, 289–307.
30. Nachman, J., Miller, M., Gilliand, G. L., Carty, R., Pincus, M. and Wlodawer, A. (1990) *Biochemistry*, **29**, 928–937.
31. Levitt, M. and Warshel, A. (1978) *J. Amer. Chem. Soc.*, **100**, 2607–2613.
32. Cotton, F. A., Hazen, E. E. and Legg, M. J. (1979) *Proc. Natl. Acad. Sci., USA*, **76**, 2551–2555.
33. Sevcik, J., Dodson, E. J., Dodson, G. G. (1991) *Acta Crystallogr.*, **B47**, 240–253.
34. Holbrook, S. R., Sussman, J. L., Warrant, R. W. and Kim, S.-H. (1978) *J. Mol. Biol.*, **123**, 631–660.
35. Ke, H., Zhang, Y. and Lipscomb, W. N. (1990) *Proc. Natl. Acad. Sci., USA*, **87**, 5243–5247.
36. Arni, R., Heinemann, U., Maslowska, M., Tokuoka, R. and Saenger, W. (1987) *Acta Crystallogr.*, **B43**, 548–554.
37. Olson, W. K. and Sussman, J. L. (1982) *J. Amer. Chem. Soc.*, **104**, 270–278.
38. Bernstein, J. (1992) In Domenicano, A. and Hargittai, I. (ed.), Accurate Molecular Structures: Their determination and importance. Oxford University Press, New York, 469–497.
39. Weber, D. J., Mullen, G. P. and Mildvan, A. S. (1991) *Biochemistry*, **30**, 7425–7437.
40. Morris, A. L. and MacArthur, M. W. (1992) *Proteins*, **12**, 345–364.
41. Sali, A., Veerapandian, B., Cooper, J. B., Moss, D. S., Hofmann, T. and Blundell, T. L. (1992) *Proteins*, **12**, 158–170.
42. Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. and Davies, D. (1987) *J. Mol. Biol.*, **196**, 877–900.
43. James, M. and Sielecki, A. R. (1986) *Nature*, **319**, 33–38.
44. James, M. and Sielecki, A. R. (1983) *J. Mol. Biol.*, **163**, 299–361.
45. Jaskolski, M., Miller, M., Rao, J., Leis, J. and Wlodawer, A. (1990) *Biochemistry*, **29**, 5889–5898.
46. Baudys, M., Foundling, S., Pavlik, M., Blundell, T. and Kostka, V. (1988) *FEBS Letts.*, **235**, 271–274.
47. Pearl, L. and Blundell, T. (1984) *FEBS Letts.*, 96–101.
48. Foundling, S. I., Cooper, J., Watson, F. E., Cleasby, A., Pearl, L. H., Sibanda, B. L., Hemmings, A., Wood, S. P., Blundell, T. L., Valler, M. J., Norey, C. G., Kay, J., Boger, J., Dunn, B. M., Leckie, B. J., Jones, D. M., Atrash, B., Hallett, A. and Szelke, M. (1987) *Nature*, **327**, 349–352.
49. Blundell, T. L., Jenkins, J. A., Sewell, B. T., Pearl, L. H., Cooper, J. B., Tickle, I. J., Veerapandian, B. and Wood, S. P. (1990) *J. Mol. Biol.*, **211**, 919–941.