Use of a 3D Structure Data Base for Understanding Sequence-dependent Conformational Aspects of DNA

Masashi Suzuki1,2*, Naoki Amano1, Jun Kakinuma1,2 and Masaru Tateno1,3

1AIST-NIBHT, CREST Centre of Structural Biology
Higashi 1-1, Tsukuba
305 Japan
2Graduate School of Human and Environmental Sciences
University of Tokyo
Komaba 3-8-1, Meguro, Tokyo
153 Japan
3Computational Science Laboratory, Institute of Physical and Chemical Research
RIKEN, Hirosawa 2-1, Wako
351-01 Japan

*Corresponding author

The roll-twist-slide correlation in the DNA crystal structures that are collected in the Nucleic Acid Data Base is analyzed in order to obtain a general understanding of the effects of the nucleotide sequence on the 3D structure of a dinucleotide step. It is concluded that the differences between the pyrimidine bases and the purine bases in terms of their physical shapes are the major factors that determine the stereochemical characteristics of the steps through base to backbone and base to base interactions. The characteristics are further modulated by the differences between the A:T and G:C base-pairs, which can be explained by enhancement of the purine-pyrimidine asymmetry in the A:T base-pair.

Keywords: DNA conformation; crystal structure; TATA box; DNA-protein interaction; DNA bending

Introduction

A protein structure is stabilized by side-chain–side-chain interactions, which can be made between residues a long way from each other in the linear polypeptide chain. In contrast, distant contacts are not frequently observed in the double-stranded DNA structure. As a consequence, DNA molecules are more flexible. The flexibility allows DNA, for example, to bind around a protein surface (Barber & Zhurkin, 1990; Suzuki & Yagi, 1995; Juo et al., 1996; Suzuki et al., 1996b), or to adopt a particular crystal packing (Yanagi et al., 1991; Dickerson, 1992; Dickerson et al., 1994; Timsit & Moras, 1995; Grzeskowiak, 1996). Therefore, to understand the stereochemical characteristics of DNA, it is essential to crystallize the same DNA fragment in as many conditions as possible and to make comparisons with other determined conformations, i.e. by using data bases (Calladine & Drew, 1984; Dickerson, 1992; Heinemann et al., 1992; Suzuki & Yagi, 1995; Young et al., 1995; Suzuki et al., 1996a; Olsson & Zhurkin, 1996; Olson, 1996; EL Hassan & Calladine, 1997).

In these data bases, the conformational changes found in protein complexes are larger than those caused by crystal packing, since the protein-binding DNAs have been designed in evolution to produce various superstructures by the selection of appropriate sequences, while most of the DNA molecules crystallized in the absence of a protein are packed more or less straight. Therefore, the greatly distorted DNA structures in complexes with TBP (Kim, Y. et al., 1993; Kim, J. L. et al., 1993; Juo et al., 1996), CAP (Schultz et al., 1991; Parkinson et al., 1996a,b) and E2 (Hegde et al., 1992) gave us a good insight into this subject (Barber & Zhurkin, 1990; Suzuki et al., 1996b; Guzikevich-Guerstein & Shakked, 1996).

The number of DNA crystal structures so far determined is large enough to make a statistical survey of the conformations of the ten independent types of dinucleotide steps, and thus these steps are the focus of this analysis. In general, a data base-oriented analysis proceeds, first by collecting crystal structures, second by defining the parameters that will be used for characterizing the structures, third by calculating the parameters using an appropriate algorithm, and finally by interpreting the calculated values in order to develop a concept that can explain various characteristics of the 3D structures. Here the Nucleic Acid Data Base (NDB; Berman et al., 1992) is used at the first step, the Cambridge code parameters (Dickerson, 1989) at the second step, and
the Babcock-Olson program (Babcock et al., 1994) at the third step.

The study of DNA conformations in this way is not totally novel (Dickerson, 1992; Bhattacharyya & Bansal, 1992; Sponer & Kypr, 1993; Suzuki & Yagi, 1995; Young et al., 1995; Gorin et al., 1995; Suzuki et al., 1996a; Olson, 1996; EL Hassan & Calladine, 1997). However, the final step of interpreting the statistics, though most important, is very difficult. The already published interpretations do not seem to provide a simple and consistent concept that can cover the overall features of sequence-dependence in DNA conformations.

There are at least two novel and important aspects in this work. First, the correlation of the three parameters, roll, helical twist (here referred to as twist) and slide, is concentrated on for the reasons described below, and is analyzed in a systematic way. Second, a simplified 3D model of a dinucleotide step is used extensively to provide a consistent explanation. Through this process of simplification, by keeping important stereochemical characteristics of the dinucleotide conformation intact and by leaving less important ones out of the model, essential features will become clear. Thus, as discussed here in our model the shape of the base-pairs is simplified to a trapezoid but not rectangle in order to introduce the sequence-dependent conformational characteristics (Figure 1). For the systematic understanding of the subject, in addition to the new findings, some conclusions that have been made by other groups and ourselves are repeated. It is the simplicity and the consistency in the explanation that is important here.

The movement of a base-pair relative to the neighboring base-pair can be described by using six parameters: roll, twist, tilt, slide, shift and rise (Dickerson, 1989). Including as many as six parameters, however, makes it difficult to find a meaningful pattern in the statistical values.

We have learned from previous analysis (Suzuki & Gerstein, 1995; Suzuki et al., 1996a) that when there are many parameters that are not totally independent, it is important to separate the key parameters that drive the conformational change from other less important ones, which follow the change. Here, the six parameters are correlated through the sugar-phosphate backbones that connect the two base-pairs.

The rise distance scarcely changes from 3.4 Å in any dinucleotide step (Suzuki et al., 1996a). Thus, a dinucleotide conformation changes by keeping the hydrophobic interaction between the two base-pairs intact (Mazur et al., 1989; Friedman & Honig, 1995; Claverie et al., 1996; Guckian et al., 1996). Two parameters, shift and tilt, are different from the others and less important (see Materials and Methods). It is therefore the remaining three parameters, roll, twist and slide, that need to be analyzed in detail.

Results and Discussion

Y/R differences in the averaged parameters

Some important characteristics of dinucleotides become clear when the average values for the three parameters (Figure 2a) are plotted in three dimensions (Figure 3). First, steps in A-DNA and those in B-DNA are separated from each other in the plot. The former are clustered together, while the latter diverge. In what follows, unless specified otherwise, B-steps are discussed.

Second, the B-steps of each of the three dinucleotide types, RY, YY/R and YR (here Y is the abbreviation for pyrimidines, T and C, and R for purines, A and G), are clustered into a single domain and the three domains do not intersect each other. Thus, sequence dependent conformational characteristics do exist in B-DNA. Here, the differences between the pyrimidines and purines are more important than those between the two pyrimidines or those between the two purines.

Third, the standard B-conformation is represented best by the YY steps, and the YR and RY
steps are located on opposite sides of the YY domain in the plot. Thus an alternating Y/R sequence has a zig-zag type of conformation (Klug et al., 1979). The Y and R bases affect the dinucleotide conformation in opposite ways if positioned at the same end, and each base affects the conformation in opposite ways if positioned at opposite ends.

Fourth, although, in general, steps in complexes with proteins behave essentially in the same way as those of the same types in B-DNA crystallized in the absence of a protein, the former have a lower helical twist, a larger roll, and a more negative slide in comparison with the latter (Figure 2b).

**Separation of the YR steps from the RY steps**

The Y/R-dependent conformational characteristics can be confirmed by analyzing the distribution of individual entries of each type in the plot (Figure 4). The isoprobability ellipsoid that covers between 59.1% and 69.8% of the entries was generated for each Y/R type in the plot of each pair of the parameters by the method of Mahalanobis (1936) with two degrees of freedom. The Mahalanobis’ ellipsoid of YR and that of RY are positioned on the sides opposite that of YY in the roll-twist, slide-twist and slide-roll plots (Figure 4a to c).

The size of the original Mahalanobis’ ellipsoid is dependent on the number of the examples, since the major and minor axes of the Mahalanobis’ ellipsoid are proportional to the standard deviations (SD) along these axes, and since SD values are expected to be roughly inversely proportional to the square-root of the number of the examples. Thus, for visual comparison of two or more ellipsoids the size of each ellipsoid was compensated by dividing the axes by the square-root of the ratio of the average of the number of examples in the groups compared to the number of the examples in the reference group (here referred to as the compensated Mahalanobis’ ellipsoids).

In order to evaluate the separation of two ellipsoids, the percentage of the examples of one group (YR etc.) found inside the ellipsoid of another group (RY etc.) that was compensated corresponding to the number of the former examples (YR etc.), and the percentage of the examples of the latter group (RY etc.) found inside the ellipsoid of the former group (YR etc.) compensated corresponding to the number of the latter examples (RY etc.) were averaged. The resultant value is referred to as the inclusion index of the two groups. The inclusion indices of YR and RY for the three plots are 19.6, 27.1 and 30.3, having an average of 25.7, while the original ellipsoids cover roughly 60% of the examples (Figure 4d).

**Differences between the YR and RY conformations**

We now interpret the differences between the Y/R types by using a simple model of a dinucleotide
Since the Y/R differences are more important than the A/G or T/C differences, the important characteristics are likely to be those of the overall physical shapes of the base-pairs but not the positioning of the chemical groups. The A:T and G:C base-pairs are fairly symmetrical on the minor groove side, but are asymmetric towards the major groove (Figure 1). The pyrimidines are pushed further into the major groove than the purines. Thus in what follows a base-pair will be simplified to a trapezoidal shape.

Two rods that mimic the sugar-phosphate backbones are attached to the minor groove side of the trapezoid in our model (views looking into the major groove are shown in Figure 5, and those into the minor groove in Figure 6). This asymmetry is needed to create the major groove–minor groove asymmetric characteristics of DNA, which are reflected in the plus/minus asymmetry of the slide and roll parameters (Figure 2a).

The direction in rolling of YR steps is positive on the average, i.e. around the major groove (Figure 5a), while that of RY steps is more on the negative side (Figure 5b). The YR steps possess a positive slide on the average (Figure 5c), while the RY steps have a negative one (Figure 5d). The YR steps have a larger twist (Figure 5e), and the RY a smaller one (Figure 5f). In short, the YR steps have a tendency to produce more positive values in the three parameters, while the RY steps produce more negative values. Therefore, the axis connecting YR to RY runs from the all-positive pole to the all-negative pole in the 3D plot (Figure 3).

The above differences between YR and RY can be explained by focussing on the pyrimidine-purine asymmetry, i.e. the edges of the pyrimidines. The positive roll value can be adopted by a YR step, probably because each Y base in the step can find a space on the major groove side for rolling into (Figure 5a). However, positive rolling of an RY step is more difficult (Figure 5b). To avoid collision of the Y bases against the R bases and probably to the sugar-phosphate backbone, an RY step tends to adopt a slightly negative roll value.

It is expected that YR and RY steps adopt slide or twist values that would move the edges of the two Y bases further from the nearest sugar-phosphate backbones (Figure 5c to f). Since the Y bases are found at different ends in YR and RY, the two types would adopt values shifted in opposite senses. The observations are consistent with this assumption. A YY step has characteristics intermediate between those of YR and RY, since it has two Y bases on both sides of the phosphate group.

**Effects of the right-handed helicity on the Y/R differences**

In general, the base 5'-terminal to the sugar-phosphate backbone is more distant from the backbone than the base 3'-terminal, because of the right-handed helicity of DNA (compare the Y bases in Figure 5a and b). As a consequence, restrictions on a YR step, where the two Y bases are 5’-terminal to the backbones, are expected to be
Figure 4. Separation of the YR, YY and RY groups in the parameters. a to c, That in the slide-twist (a), roll-twist (b), and slide-roll (c) plots. The compensated Mahalanobis' ellipsoids are shown. The size of the original Mahalanobis' ellipsoids is dependent on the number of the entries. Thus, for visual inspection the size of the ellipsoids is compensated so that it corresponds to the average of the number of the examples in the three groups (see Materials and Methods). YR, YY and RY entries are plotted in blue, red and green, respectively. d, The inclusion index values. The inclusion index indicates the degree to which the two groups compared are close to each other (see Materials and Methods). For the recognition of well-separated groups the values, 50 or smaller, are highlighted in bold. Note that the maximum values become between 59.1 and 69.8 instead of 100 (the values for YR versus YR, YY versus YY and RY versus RY, see Materials and Methods).
smaller than on an RY step. Furthermore, in an RY step the two base-pairs are stacked more closely (Figure 5b) than in a YR step (Figure 5: see also Figure 9 of Suzuki et al. 1995), creating more stability (Friedman & Honig, 1995). The size of the compensated ellipsoid of YR is markedly larger than that of RY, suggesting more flexibility of YR (Figure 4).

Another consequence of the right-handed helicity is that sliding in the negative direction is expected to be easier than in the opposite direction (Figure 6e). The former involves the approach of the bases 5'-terminal to the backbones, while the latter involves the bases 3'-terminal. Therefore, if forced by the proteins, any Y/R type of step tends to slide in general in the negative direction, unless the step becomes totally helically untwisted and thus loses its helical nature.

**Correlation of the parameters**

In the previous subsection the differences between the YR and RY groups in individual parameters were discussed, i.e. the "static" conformational differences. We now discuss the correlation of the three parameters in order to understand the more "dynamic" aspects of the steps. The correlation can be identified by plotting the parameters of individual examples of the same Y/R type (Figure 7: see also a review by Dickerson, 1997), and by comparing the average values of the same Y/R type but of different sequences (Figure 8).

First, helical untwisting is associated with positive rolling of a step (Suzuki & Yagi, 1995; Gorin et al., 1995). The finer detail in the roll-untwist correlation depends on the Y/R type, but the correlation is clear for each individual type (Gorin et al., 1995).

Second, helical untwisting is associated also with negative sliding among the RY steps (Suzuki et al., 1996a). The YR steps are more spread out along the slide axis, which reduces the correlation. All the B-steps are distributed along essentially a
single correlation curve, although the three Y/R types are distributed in different places along it.

Third, sliding of the RY steps in the negative direction is correlated with unrolling of a negative roll.

<table>
<thead>
<tr>
<th>Roll-Twist</th>
<th>A-D</th>
<th>B-D</th>
<th>C-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>YR</td>
<td>-0.42</td>
<td>-0.34</td>
<td>-0.45</td>
</tr>
<tr>
<td>YY</td>
<td>-0.44</td>
<td>-0.44</td>
<td>-0.42</td>
</tr>
<tr>
<td>RY</td>
<td>-0.33</td>
<td>-0.33</td>
<td>-0.42</td>
</tr>
<tr>
<td>Slide-Twist</td>
<td>0.33</td>
<td>0.35</td>
<td>0.36</td>
</tr>
<tr>
<td>Slide-Roll</td>
<td>0.24</td>
<td>0.44</td>
<td>0.46</td>
</tr>
</tbody>
</table>

**Figure 7.** Correlation of the parameters in the YR, YY and RY groups of B-DNA crystallized in the absence or presence of a protein. a, Correlation of roll and twist (upper), slide and twist (middle), and slide and roll (lower). The first column (A, all; D, all) shows the values of the linear correlation coefficient calculated using all the examples in B-DNA, the second column (A ≤ 33°, D ≤ 1.3 Å) that using the steps that have an angle value of 33° or smaller, and a distance value of 1.3 Å or smaller, the third column (A ≤ 33°, D ≤ 1.15 Å) an angle value of 33° or smaller, and a distance value of 1.15 Å or smaller, and the fourth column (A ≤ 33°, D ≤ 1.0 Å) an angle value of 33° or smaller, and a distance value of 1.0 Å or smaller. The values of the correlation coefficient, 0.4 or larger and −0.4 or smaller, are underlined. These indicate linear correlation between the two parameters. The angle and distance parameters are the evaluation of disorder of a base-pair (Babcock et al., 1994). Base-pairs in the standard B-DNA produce the two parameters close to 0. Thus these can be used for excluding broken base-pairs, which are mainly found at the ends of the DNA double strands in the crystals, from the statistics. Since there is no clear border between broken base-pairs and largely distorted but still maintained base-pairs, the examples used for the calculation were limited stepwise by increasing the cut-off value in distance. The angle value of 33° was also used, since this value seemed to separate largely propeller twisting base-pairs in the A-tracks from more broken base-pairs.

The above correlation can be explained as follows by the possible movement of the step conformation (Figure 6). When helically untwisted, the two base-pairs in a dinucleotide become distant from each other, since the length of the sugar-phosphate backbone, 5.5 Å, is longer than the stacking distance, 3.4 Å; here, the backbone length is defined as the average of the distances between pairs of the C1′, C2′, C3′, C4′ and O4′ atoms in the two sugar rings neighboring in the same strand (here referred to as the L length). The L length scarcely changes, the SD value is 0.2 Å (Figure 9b).

To regain the hydrophobic interactions the step can adopt tactics of either sliding or rolling. Positive rolling is effective, because it can make the two base-pair edges on the major groove side approach each other. Sliding in any direction appears to be appropriate, but sliding in the negative direction is easier for any step, as has been described, whereas sliding in the positive direction is difficult except for the loosely stacking YR steps, resulting in a looser distribution of the YR steps. In order to slide an RY step in the negative direction, unrolling of a negative roll is necessary to create a smooth interface between the two tightly stacked base-pairs, thereby avoiding the possible collision on the minor groove side (Figure 5d).

**Y/R differences in the correlation**

The precise rolling is likely to involve contact of the edges of the two base-pairs and thus is dependent on the Y/R type of the step (Figure 6). In contrast, the slide/twist motion takes place in a 2D plane, and thus the slide/twist characteristics are likely to result more from base-backbone interactions. As a consequence, sliding and twisting produce correlation less dependent on the type (Figures 4 and 8). A nearly single correlation curve follows the YR-RY axis in the twist-slide plot (Figures 7 and 8), but the RY and YR curves run in the direction approximately perpendicular to the
YR-RY axis in the roll-untwist plot (Figures 4 and 8).

The RY steps have more restrictions in terms of base-base and base-backbone interactions, and show the clearest correlation between the parameters (Figure 7a). The YR steps have preference in terms of the directions for change, although the restrictions are the smallest, and show a poorer correlation. The YY steps have the two Y bases on both ends and thus do not have any preference, and show the poorest correlation. This poorest correlation arises partly from a separation of the TY steps from the CY steps (Figures 8 and 10a and d). This will be discussed further in the next subsection.

A:T to G:C differences in the conformation

We now discuss differences in the conformations of the same Y/R type, which arise from the differences between the A:T and G:C base-pairs. In the roll-twist plot (Figure 8) the YR steps are distributed along the correlation curve so that TA is located on the high twist side, while among the RY steps AT is on the low twist side. The YY steps, TC, TT, CC and CT, are distributed in this order from the high twist side to the low twist side, separating the TY steps from the CY steps (Figure 10a). With regard to slide, AT lies on the negative side among the RY steps, but the three types of YR steps do not show significant difference.

By the combination of the differences between the three Y/R types in the roll-untwist correlation and the different positioning of A/T-composed and G/C-composed steps in the correlation of each type, AT, TT and TA differ more in helical twisting (Figure 10b and e), while GC, CC and CG differ more in rolling (Figure 10c and e).

Enhancing the Y/R differences by methyl groups

The sequence-dependent characteristics of RY steps in the slide and twist parameters can be explained by focussing on the two methyl groups of the T bases in AT, which would further slide the step in the negative direction and helically untwist to avoid the possible collision with the sugar-phosphate backbones 5’-terminal to the T base (Figure 5d and f). The GT step has characteristics intermediate between those of AT and GC. The sequence-dependent characteristics of YR steps in twisting can be explained in a similar way but in the opposite sense because of the possible collision of the T bases in TA with the sugar-phosphate backbones 3’-terminal (Figure 5c and e). In short, the Y/R-dependent conformational characteristics are expected to be clearer in A/T-rich sequences than in G/C-rich sequences (Hunter, 1993; Suzuki et al., 1996a).

The TA step produces positive rolling smaller than the CG step probably in order to minimize possible collision of the methyl groups with the neighboring bases and the nearby backbones. It is more difficult to explain rolling of the GC step in the negative direction in contrast to the almost no rolling of GT and AT (Figure 2a), since AT is expected to produce a higher negative roll for the
same reason as that given for TA. This can be explained as the result of correlation of the roll parameter to the twist parameter but, if this is the case, it leads to a conclusion that the roll parameter is less dependent on the sequence than the twist parameter. Alternatively, this could be due to the highly selected sequences of B-DNA crystals that have GC steps. Namely, the nucleotide sequences adjacent to GC will have some effects and GC preferentially occurs in sequences such as CGCG mainly in B-DNA.

Among the YY steps, the TY steps and the CY steps have different characteristics (Figure 10a). The TY steps have a higher twist, almost no roll and almost no slide, while the CY steps have a lower twist, a larger roll and a higher positive slide. The T base at the 5'-end could change the conformation in order to create more distance from the sugar-phosphate backbone, resulting in a higher twist (Figure 10f), and similarly a more positive slide in order to avoid approach of the methyl group to the sugar-phosphate backbone. A very small roll can be created either in correlation with untwisting or in order to avoid possible collision between the methyl group in T and the neighboring base-pair. The T base at the 3'-end would change the conformation in the opposite direction (Figure 10f).

It is more difficult to separate the YY steps into YT and YC steps. This is slightly puzzling, as the difference at the 3'-end is expected to be more significant than that at the 5'-end, since the former is closer to the backbone than the latter. An explanation for the above observation can be given if the effect of either type of Y base at the 3'-end of a YY step is sufficiently significant, i.e. the conformation changes up to the limit, even if the C base is positioned there.

**B-to-A and B-to-Å transition**

Dinucleotides in complexes with proteins have a slightly higher roll, a lower twist and a more negative slide than in B-DNA (Figure 2b). When B-steps are forced further in the above directions, they converge towards A-steps (Figure 3). To avoid the possible collision between the sugar-phosphate backbones and the 5'-terminal bases (Figure 6e), which will be induced by the large negative sliding, and to produce a longer backbone, the mode of the sugar puckering is altered. Consequently, the slide parameter clearly distinguishes steps in A-DNA from those in B-DNA (Calladine & Drew, 1984). This is also expected from the energy calculations made by Mazur et al. (1989), who concluded that in the absence of the backbones the base-pairs would have intrinsic bi-stability that would produce A and B conformations alternatively. The large negative slide would decrease the stacking between the two base-pairs and thereby relax the Y/R differences.

Another type of step, which is characterized by having a highest roll, a smallest twist and a large positive slide (Figure 12e), is found in complexes with the TATA box-binding protein (Suzuki & Yagi, 1995; Guzikevich-Guerstein & Shakked, 1996); here the symbol Å is used. Guzikevich-Guerstein & Shakked (1996) pointed out that the Å-steps have the same backbone conformation as that of A-steps, but the values of the γ angle, the rotation angle around the connection between the sugar and the base, are different. The result is to rotate the base away from the original partner position in A-DNA and to make a new pairing with another base positioned two bases towards the 3'-terminus along the other nucleotide strand.

If one of the two strands in B-DNA is fixed (Figure 11b) and the other strand moves relative to
the fixed strand towards its own 5' terminus (Figure 11a), the DNA conformation will change to $A$, creating a negative slide of the steps, while, if the strand moves to its 3' terminus (Figure 11c), the conformation will change to $AÄ$, creating a positive slide. Thus, when the conformation changes from $A$ to $AÄ$, it might adopt the $B$-conformation at an intermediate stage, as has been suggested by Lebrun et al. (1997).

Comparison with the explanation given by the other groups

Some years ago, Calladine (1982) tried to explain DNA conformations by another model. In this, it was assumed that propeller twisting of the bases in a pair around the hydrogen bonds was the major determinant of the DNA conformation. The pyrimidines are narrower in the pairing direction than the purines. Thus the steric hindrance caused by the propeller twisting to the neighboring base-pair would be different, depending on the Y/R arrangement in the dinucleotide.

The Calladine (1982) model is, however, inconsistent with the energy calculations made by Sarai et al. (1988), which produced very similar roll and twist values for DNA with and without introducing propeller twisting. EL Hassan & Calladine (1996) have found a correlation between the slide parameter and the propeller twist angle. However, the fact that YR, YY and RY share the same slide-propeller twist correlation curve was not expected from the original argument (see Figure 3.6 of Calladine & Drew (1992), which suggests that no slide is necessary for propeller twisting of YY/RR, and Figures 3.10 and 3.11 of the same work, which suggest that sliding is necessary for propeller twisting of YR and RY to the same degree but in the opposite directions). It is difficult to find out whether the slide fixes propeller twisting or vice
versa as El Hassan & Calladine (1996) stated. The original Watson-Crick type of hydrogen bonds are energetically most stable when a base-pair has no distortion. Thus it is likely that distortion such as propeller twisting occurs as a consequence of the conformational change caused by more fundamental factors.

The electropolarization of the base-pairs has been analyzed but the conclusions are not totally consistent (Stewart, 1970; Ornstein et al., 1978; Zhurkin et al., 1980; Kollman et al., 1981; Poltev et al., 1981; Weiner et al., 1984; Hunter & Sanders 1990; Pearlman & Kim, 1990; Sponer et al., 1996a). Hunter (1993) has concluded that the electrostatic interactions between the neighboring base-pairs would stabilize the stacking in vacuum, while Friedman & Honig (1995) have concluded otherwise in aqueous solution mainly because the stacking would prevent favorable base-solvent electrostatic interactions.

Given the above contradicting reports, it is difficult to interpret our observations in the light of electrostatic interactions. Sponer et al. (1996b) have calculated the stacking energy between the two neighboring bases of different sequences by changing the roll, rise and displacement parameters. It will be useful if this direction of research is extended to understand the stacking of base-pairs, closely examining the ranges of the three parameters reported here.

Recently, Hunter & Lu (1997) have reported, by comparing energy calculations and the crystal structures, that the twist parameter is likely to be fixed by the constraints of the backbone, which are missing in the calculations. This is essentially consistent with our findings, that the length of the backbone is kept nearly constant and that the DNA conformation needs to be understood without totally omitting the backbones.

In order to explain the sequence-dependent conformational differences among the dinucleotides of the same Y/R type, possible collision of the methyl group of the A base to the base neighboring in the same strand and that to the sugar-phosphate backbone were focussed on in our discussion. Gorin et al. (1995) have provided another equally good explanation by focussing on possible collision between the exocyclic groups of the bases at the positions closest to the pairing partners on the sides of the minor and major grooves, i.e. positions 2 and 4 in the Y bases and positions 2 and 6 in the R bases. More study is necessary to judge which chemical groups are most important for understanding the dinucleotide conformation.

Role of data bases in structural biology

The number of DNA structures determined by X-ray crystallography might not increase tenfold over the next ten years. Even if that happens, however, the amount of information that can be obtained from the structures might not increase proportionally. It has been argued by many scientists that trinucleotides could be the structural units of DNA (Yanagi et al., 1991), but it is unlikely that a close examination of that hypothesis using crystal structures would be possible in the near future because of the limitation in the number of crystal structures. Apparently we need to seek another type of approach.

Data bases that collect the atomic co-ordinates or the amino acid/nucleotide sequences, nevertheless, have a very important function in biology. Entries in these data bases are independent of the history in which they were created, and are self-contained,
and the data bases have a standard format that facilitates direct comparison, and thus these can be shared by the international community. In one sense, data bases can be compared with equations in physics, which provide a common language between physicists.

Materials and Methods

Crystal structures

The crystal co-ordinates were taken from the Nucleic Acid Data Base (Berman et al., 1992). The classification of the structures to A-DNA and B-DNA written in NDB was used in this study. DNA oligomers of the same sequence that were crystallised in the same form (such as the Dickerson-Drew dodecamers determined by slightly different methods or determined in different environments) are represented by one of the structures.

The NDB code names of the structures studied are ADH006, 007, 008, 0012, 0014, 0020, 0023, 0030, 0038, 0039, 0041, 0047, ADI009, ADJ022, 049, ADL045, 046, BDJ017, 019, 025, 031, 036, 039, 051, 052, BDL006, 007, 015, 020, 026, 038, 042, 047, PDE009, PDR001, 004, 006, 008, 009, 010, 011, 012, 013, 015, PDRCO1, PDT002, 003, 004, 006, 007, 008, 009, 011 and PDV001.

Many DNAs studied here have reverse symmetric sequences. The two halves are related by crystallographic 2-fold symmetry in some examples, such as the two halves of the E2 binding site (PDV001). Such identical steps related by symmetry were treated as two independent examples. This is because some other structures such as the Gal4 binding site (PDT003) were determined without assuming such a crystallographic 2-fold symmetry, and thus two identical steps in the binding site must be treated as independent examples, although the two are, in fact, almost the same. Since whether the structures become crystallographically symmetric or pseudo-symmetric is, in a sense, a matter of assumption, it was felt better to include all the symmetric or pseudo-symmetric DNA structures into the statistics rather than only half the structures. Some structures such as in the tramtrack crystal (PDT0011) contain two molecules of the complex per asymmetric unit and the two structures were treated as independent examples.

**Figure 12.** Calculation of the backbone length using a model. a, A view looking along the helix axis of a dinucleotide model. The line connecting A1 and B1, and that connecting A2 and B2 show the central lines of the base-pairs. The r distance is the distance between the center of the base-pair and the major groove edge, and the l distance that between the base-pair center and the center of the sugar ring, while θ is the angle between the two l directions. The values of l, r and θ used for the calculation are 6.97 Å, 2.73 Å and 134°, respectively. The positive direction of helical twisting and that of sliding are shown by arrows. If the slide and twist parameters had no correlation, the L length between the two sugar rings neighboring in the same strand would be:

\[ L = \sqrt{108.85 - 97.29 \cos(180T/\pi)} \]

where T is the roll angle, and:

\[ L = \sqrt{S^2 + 3.37S + 30.14} \]

where S is the slide distance. b, The roll movement was introduced, first by the rotation of each base-pair around the center line (A1B1, A2B2 in a) by half the roll angle, and second by separating the base-pairs along the helix axis (by rising) so that the minimal distance between the two base-pairs in this direction would become 3.4 Å. If rolling had no correlation with twisting or sliding, the L length would be:

\[ L = \sqrt{30.14 + 127.56 \sin^2(90R/\pi) + 99.79 \sin(90R/\pi) - 1.77 \sin(90R/\pi)} \]

where R is the roll angle when it has a positive value, and:

\[ L = \sqrt{30.14 + 8.72 \sin^2(90R/\pi) + 25.66 \sin(90R/\pi) - 1.77 \sin(90R/\pi)} \]

when it has a negative value.
Calculation of the dinucleotide step parameters

The six parameters were calculated by using a computer program (Babcock et al., 1994). Altogether 930 dinucleotide steps were collected: 66 TG, 154 CG, 122 TA, 92 TT, 70 TC, 80 CC, 112 GT, 76 GC and 94 AT. Some unusual steps found in the complexes were excluded from the statistics; a non-Watson-Crick G:C base-pair in PDT003, two nicked steps in PDR006, and three unpaired base-pairs in PDT002.

A dinucleotide step of a reverse symmetric sequence such as GC provides two examples (note that the partner of GC is again GC, and the two conformations are usually different, while for example, that of GT is not GT but AC and therefore provides only one example). For the reverse symmetric sequences the averaged tilt angle becomes zero. Even when the sequence is not reverse symmetric the two parameters do not change much.

Evaluation of the separation of the groups in the plots

For each plot of the parameters (the slide-roll plot, etc.) of each dinucleotide group (the YR group, etc.) the isoprobability ellipsoid was calculated by using Mahalanobis' generalized distance (Mahalanobis, 1936) with two degrees of freedom (referred to as the Mahalanobis' ellipsoid). Between 59.1% and 69.8% of the entries of each group were found inside its ellipsoid.

The size of the Mahalanobis' ellipsoid is dependent on the number of the examples, since the deviations of the two parameters calculated along the opposite strands cancel, and thus the average value becomes zero. Even when the sequence is not reverse symmetric the two parameters do not change much.

Thus for the comparison of YR and RY, which have 169 and 154 examples, respectively, the axes of the Mahalanobis' ellipsoid of YR were divided by \( \sqrt{169} \), and those of RY by \( \sqrt{154} \).

For the quantitative characterization of separation of two groups, the size of the Mahalanobis' ellipsoid of one group was compensated corresponding to the number of the examples in the other group, and the percentage of the latter examples found inside the compensated ellipsoid was calculated. The calculation was carried out twice in each direction independently, and the two values were averaged; the average is referred to as the inclusion index of the two groups. Thus, for the characterization of separation of YR and RY in the roll-twist plot, first the size of the ellipsoid of YR was adjusted so that it would correspond to the example number of 154 by dividing the major and minor axes by \( \sqrt{154} \), and the percentage of RY steps found inside the compensated ellipsoid was calculated as 27.3%. Second the ellipsoid of RY was adjusted so that it would correspond to the example number of 169 by dividing the axes by \( \sqrt{169} \), and the percentage of YR steps found inside was calculated as 11.8%. The inclusion index was calculated by averaging the two values as 19.6.

Theoretical calculation of the backbone length

The base-pairs were simplified to a rectangular shape (Figure 12), in which the length between the base-pair center and the major groove edge, \( r \), was 2.73 Å, that between the base-pair center and the sugar group, \( l \), 6.97 Å, and the angle between the two \( l \) directions, 0°, 134°.

The distance between the two sugar rings neighboring in the same strand, \( l \), was calculated by changing the slide distance or helical twist angle independently. Rolling movement was introduced, first by rotation of each base-pair by half the rolling angle around the base-pair center line, and second by increasing the rise distance so that the shortest distance between the two base-pairs along the helix axis became 3.4 Å. The second step was necessary in order to avoid possible collision between the two base-pairs.

The 5 length calculated is:

\[
\sqrt{108.85 - 97.29 \cos(180T/\pi)}
\]

upon changing the twist angle, \( T \), and:

\[
\sqrt{S^2 + 3.37S + 30.14}
\]

upon changing the slide distance, \( S \), while:

\[
\sqrt{30.14 + 127.56 \sin^2(90R/\pi) + 99.79 \sin(90R/\pi) - 1.77 \sin(90R/\pi)}
\]

upon changing the roll angle, \( R \), in the positive direction, and:

\[
\sqrt{30.14 + 8.72 \sin^2(90R/\pi) + 25.66 \sin(90R/\pi) - 1.77 \sin(90R/\pi)}
\]

in the negative direction.
Acknowledgements

We thank Dr. J. T. Finch for his critical reading of the manuscript. This work was supported by the CREST (Core Research for Evolutional Science and Technology) Project of the Japan Science and Technology Corporation.

References


Edited by B. Honig

(Received 4 April 1997; received in revised form 15 September 1997; accepted 15 September 1997)