Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2579–2583, April 1988
Biophysics

Some rules for predicting the base-sequence dependence of DNA conformation

W. L. PETICOLAS, Y. WANG, AND G. A. THOMAS

Department of Chemistry and Institute of Molecular Biology, University of Oregon, Eugene, OR 97403

Communicated by Peter H. von Hippel, December 18, 1987

ABSTRACT **Two tables have been constructed showing the crystal and solution conformations of short sequences of DNA. Each of these DNAs has been found to be in one of three different conformations—the A, B, or Z form—depending upon the base sequence and the environmental conditions. A set of rules is presented showing the tendency of certain base pairs to direct the DNA conformation into the A, B, or Z genus in saturated salt solutions and in crystals. These rules are based on a consideration of nearest-neighbor interactions that are interpreted in terms of 10 different two-letter code words made from the letters denoting the bases guanine (G), cytosine (C), adenine (A), and thymine (T). One table discusses the effect on DNA conformation of 3 strong words that tend to direct a DNA oligomer into either the A, B, or Z genus in crystals or in aqueous solutions containing a high salt concentration (6 M). The second table discusses the remaining 7 code words that appear to have a much weaker effect on conformation. The sequences that are most likely to lead to A–Z, B–Z, and A–B junctions are discussed, as is the possible biological significance of these rules.**

The biological activity of DNA is carried out in association with numerous proteins in order to perform various cellular functions. Protein–nucleic acid interactions are often discussed in terms of two main characteristics—the overall topology of the two partners and the interactions between the nucleotides and the main chain or side chains of the protein (1). The primary specificity of the DNA–protein interaction is undoubtedly governed by the base-sequence-dependent array of hydrogen-bonding donor and acceptor patterns that occur along the major and minor grooves of the DNA double helix (2–4). However, it is reasonable to assume that variations in the conformation of DNA may occur along the double-helical axis as it performs its functions in the cell and that these variations in conformation may play a role in the regulation of its biological function. Although the usual conformation of DNA is the canonical B form, it may be that, through interaction with other components in the cell, particularly proteins, partial or total change from this form may occur in regions of DNA. Even a partial change from the canonical B conformation toward the A genus or Z genus would result in a deformation of the hydrogen-bonding donor and acceptor patterns along the major and minor grooves. This effect could be used to aid DNA-binding proteins in locating base-sequence-specific binding sites. As we will show, the possible conformational variations of DNA appear to be determined by specific base sequences as well as the environment.

Use of Raman Spectroscopy to Obtain Statistical Data on DNA Conformation in Crystals and in Solutions

It has long been known that DNA can exist in at least two canonical right-handed, double-helical forms, the A genus and the B genus (5, 6). The first crystallographic structure established for DNA was the left-handed "Z-DNA" (7). The characterization of the conformation of short sequences of DNA in crystals and in concentrated salt and/or alcohol solutions by Raman spectroscopy has been carried out in a number of different laboratories including our own (for recent reviews, see refs. 8–10). The exact structures for the canonical B, Z, and A forms of DNA have been determined by x-ray diffraction of crystals of d(CGCGAATTCGCG) (11, 12), d(CGCGCG) (7), and d(GGTATACC) (13), respectively. By comparing the Raman spectrum of a crystal of a DNA in an unknown conformation with the standard Raman spectra obtained from DNA crystals of known conformation, it is possible to assign the unknown DNA conformation to the A, B, or Z genus by means of conformational Raman marker bands. We have found that, although it is often difficult to grow DNA crystals large enough for x-ray diffraction or conventional Raman spectroscopy, it is usually easy to find conditions for the growing of well-formed microcrystals that vary in size from 10 to 100 μm on a side. The use of a laser Raman microscope allows us to easily obtain the Raman spectrum from microcrystals and determine the conformational genus to which each one belongs (14–16). In this way we have been able to rapidly acquire a statistically significant data base relating base sequence to conformation. Using this data base, it is now possible to suggest some rules for predicting the conformation of DNA under various environmental conditions in terms of its base sequence.

Effect of Nearest-Neighbor Interactions in Directing DNA Conformation

Raman, NMR, and other studies have established that every DNA oligomer containing only unmodified bases exists in the B genus in dilute aqueous solution at pH 7 and 0.1–1.0 M NaCl (14–22). To induce an oligonucleotide into a conformation that is different from the B genus, it is necessary to decrease the activity of the water in the DNA. This is easily done by increasing the salt concentration, adding a heavy-metal cation, adding an alcohol, or crystallizing the DNA from an alcohol/water mixture. Under these conditions many but not all DNA sequences will go into conformations belonging to the A or the Z genus. In order to devise some rules that appear useful in predicting the conformation of oligonucleotides under partially dehydrating conditions, we make the simplifying assumption that the conformational tendencies of DNA sequences are directed to a large extent by nearest-neighbor interactions. By the nearest-neighbor approximation, oligonucleotides may be considered to be made up of sequential two-letter conformation code words in the 5′→3′ direction. There may also be a cooperative effect so that if one of these two-letter code words is repeated sequentially, the effect on conformation may be greater than when the code word is repeated nonsequentially along the double helix. If a sequence contains two or more different code words that tend to direct the double helix into different

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

conformations, it may not be possible at the present time to predict unequivocally the resulting conformation because our data base is not large enough to treat all such cases.

In any double-helical oligodeoxynucleotide with a length of $2n$ bases, there are 4^{2n} different sequences in a single chain. Of these, 4^n will be self-complementary. For a two-base sequence (i.e., $n = 1$), there are $4^2 = 16$ two-letter combinations in one strand. Four of these—AT, TA, CG, and GC—are self-complementary and hence are the same in both chains. The total base-stacking energy is twice that of one pair. The remaining 12 two-base sequences always occur in six complementary pairs. Consequently, in addition to the four words listed above, one only needs to consider six non-self-complementary two-letter words and/or their complements. For example, an AC in one chain will always be paired with a GT in the other. This word may be listed as AC, or GT, or (AC, GT). The stacking energy of such a word is the sum of the stacking energies of the two nearest-neighbor pairs in opposing strands. Consider an oligomer, $d(CGCGAATTCGCG)_2$. It may be considered to be made up of the following combination of code words: four (CG, CG), two (GC, GC), two (AA, TT), one (TC, GA), one (AT, AT), and one (TC, GA).

It would be of great importance to discuss the experimental results in the framework of a larger theoretical treatment. In order to do this, it would be necessary to have the free energies of interaction of the nearest neighbors together with any cooperative effects that might occur. A large component of the free energy of interaction is the base-stacking energy between neighbors along the chain. Base-stacking energies have been calculated by several authors (23, 24). Aida and Nagata (24) have made *ab initio* calculations of base stacking for both the A and the B forms. From these results one can calculate the energy difference, ΔE in kcal/mol, for the B→A transition of an oligonucleotide. ΔE is simply a summation of all of the nearest-neighbor stacking-energy differences in going from B to A. Although the entropy of the base interactions is missing from these calculations, it is of interest to compare the observed conformation with that calculated from the following assumption: when ΔE is negative, the oligomer helix conformation will go from the B to the A form in high salt or its crystal, whereas if ΔE is positive, the B form will remain under these conditions. It is found that most of the oligomers obey this rule. Those that do not are marked with an asterisk. In Tables 1 and 2, we give the calculated total change in base-stacking energy in going from the B to the A conformation for those oligomers that are in either the A or the B form. For self-complementary oligomers, this energy is twice that given by Aida and Nagata (24), who only considered one chain. For non-self-complementary oligomers the total stacking energy of both chains has been calculated and listed. Unfortunately, there are no available calculations of base-stacking energies for the Z form of DNA. Consequently there are no values for ΔE given for the Z-forming oligonucleotides in either table.

Beginning with (AA, TT) the two-letter code words are listed in Tables 1 and 2. Also listed in these tables are the crystalline and solution conformations of oligonucleotides that contain different arrangements of the code words. The patterns that emerge illustrate the tendency of these two-letter words to either (*i*) prevent an oligonucleotide from leaving the B conformational genus, (*ii*) direct an oligomer into the A or the Z conformation, or (*iii*) remain relatively neutral. Neutral words are those that appear to offer only little apparent help or hindrance to a sequence undergoing a B→Z or B→A transition. The conclusions reached by inspection of Table 1 may be summarized by stating that sequences containing repetitions of the word (AA, TT) tend not to leave the B genus, whereas (CC, GG) and (CG, CG) tend to direct the DNAs into the A form and Z form, respectively. In Table 2, it is seen that the seven remaining two-letter code words tend to favor the B genus but can be taken into the A or Z form if the (CC, GG) or (CG, CG) words are present in a dominating amount. It appears that CG will be most effective in driving an oligomer into the Z form if the oligomer contains only alternating pyrimidine-purine sequence.

Table 1. Effect of strong two-base code words on the conformation of short oligonucleotides

Obs. conf.	Example (comment or ref.) Crystal	Solution†	ΔE‡
	AA,TT (ΔE = *1.01 kcal/mol*)		
B	AAAAATTTTT (16)	AAAAATTTTT (20)	6.04
B	CGCGAATTCGCG (12)	CGCGAATTCGCG (25)	8.26
B	GGAATTCC (§)		−3.02*
B		$(dA)_n$·$(dT)_n$ (¶, 20)	1.01
B	GGTTAACC (§)		−2.24
	CC,GG (ΔE = *−1.68 kcal/mol*)		
A	CCGG (26, 27)		0.0
A	CCCCGGGG (28)		−6.88
A	GGGGCCCC (29)		−13.28
A		$(dG)_n$·$(dC)_n$ (21)	−1.72
A	GGCCGGCC (30)		−9.36
A	GGATGGGAG (31)	GGATGGGAG (§, ‖)	−5.88
A	GGTATACC (13)	GGTATACC (15)	−5.18
A		CCTATAGG (§)	−4.62
A		GGTTAACC (§)	−2.24
B*	GGATATCC (15)	GGATATCC (15)	−5.96*
	CG,CG (ΔE = *3.44 kcal/mol*)		
Z	CG (15)		
Z	CGCG (32, 33)	CGCG (22)	
Z	CGCGCG (7, 33)	CFCFCF (14, 15)	
Z	CGCGTACGCG (15)	CGCGTACGCG (15)	
Z	CGCATGCG (34)	CGCATGCG (14, 15)	
Z	TGCGCGCA (14, 15)	TGCGCGCA (14, 15)	
Z	ACGCGCGT (14, 15)		
Z	CGTACGTACG (††, 35)		
Z		CGCGTATACGCG (36)	
Z		CGCGCGTATACGCGCG (37)	
Z		CGTGCGCACG (15)	
Z		CGCACGTGCG (15)	
Z	CACGTG (‡‡)		
B*	CGTGCGCACG (15)		5.60
B*	CGCACGTGCG (15)		5.60
B*	CGTGCACG (14, 15)	CGTGCACG (14, 15)	−5.12*
B*	CGTACG (§)		6.06
B*	CGCGAATTCGCG (12)	CGCGAATTCGCG (25)	8.26

Asterisks indicate exceptions to the "negative ΔE favors A" rule (see text). Obs. conf., observed conformation.

†All solution conformations were obtained in 4–6 M NaCl except as noted (footnote ‖).

‡Base-stacking-energy difference between A and B forms in kcal/mol. Negative ΔE favors A form.

§Unpublished results.

¶Fiber or film at 75% relative humidity.

‖Exists in B form in saturated NaCl solution but goes into the non-B conformation in a solution formed by adding 20% ethanol to saturated NaCl solution.

††Crystalline complex between hexamminecobalt(II) and d(CGTACGTACG).

‡‡Subirana, J. A., Ninth International Biophysics Congress, Aug. 23–28, 1987, Jerusalem, p. 28 (abstr.).

Table 2. Effect of weak two-base code words on the conformation of short oligonucleotides

Obs. conf.	Example (comment or ref.) Crystal	Solution†	ΔE‡
	TA,TA (ΔE = *1.12 kcal/mol*)		
A	GGTATACC (13)	GGTATACC (15)	−5.18
A		$d(A\text{-}T)_n$ (§, 38)	−0.92
B	CGTACG (¶)		
Z	CGCGTACGCG (15)	CGCGTACGCG (15)	
Z		$d(A\text{-}T)_n$ (‖, 39)	
Z	CGTACGTACG (††, 35)		
Z		CGCGTATACGCG (36)	
	AT,AT (ΔE = *−2.04 kcal/mol*)		
B	GGATATCC (15)	GGATATCC (15)	−5.96*
B		CGCGATATCGCG (36)	−5.32*
A	GGATGGGAG (31)	GGATGGGAG (¶, ‡‡)	−5.88
	GC,GC (ΔE = *−2.56 kcal/mol*)		
B		GCGC (22)	−2.48*
B	GCTATAGC (15)	GCTATAGC (15)	−7.10*
B	GCATATGC (15)	GCATATGC (15)	−5.74*
	CA,TG (ΔE = *−1.57 kcal/mol*)		
B	CGTGCGCACG (15)		5.60
B	CGCACGTGCG (15)		5.60
Z		CGTGCGCACG (15)	
Z		CGCACGTGCG (15)	
Z	TGCGCGCA (14, 15)	TGCGCGGCA (14, 15)	
Z	CACGCGTG (14, 15)	CACGCGTG (14, 15)	
	AC,GT (ΔE = *−0.97 kcal/mol*)		
Z	ACGCGCGT (14, 15)	ACGCGCGT (14, 15)	
A	ACCGGCCGGT (14, 15)		−4.90
	TC,GA (ΔE = *−0.22 kcal/mol*)		
A	GGATGGGAG (31)	GGATGGGAG (¶, ‡‡)	−5.88
	AG,CT (ΔE = *−0.69 kcal/mol*)		
A	GGATGGGAG (31)	GGATGGGAG (¶, ‡‡)	−5.88

Asterisks indicate exceptions to the "negative ΔE favors A" rule (see text). Obs. conf., observed conformation.

†All solution conformations were obtained in 4–6 M NaCl except as noted (footnote ‡‡).

‡Base-stacking-energy difference between A and B forms in kcal/mol. Negative ΔE favors A form.

§Fiber or film at 75% relative humidity.

¶Unpublished results.

‖Z form is observed in presence of Ni(II).

††Crystalline complex between hexamminecobalt(II) and d(CGTACGTACG).

‡‡Exists in B form in saturated NaCl solution but goes into the non-B conformation in a solution formed by adding 20% ethanol to saturated NaCl solution.

Repeated Sequences of GG or CC Tend to Facilitate the B→A Transition

The two-letter code word GG or CC tends to direct an oligonucleotide into the A genus under dehydrating conditions such as in a crystal or a saturated salt solution. This is predicted by the value of ΔE, −1.68 kcal/mol, for this non-self-complementary pair. For example, d(CCGG) crystallizes in the A form (26, 27). The oligomers d(GGCCGGCC) (30, 40), d(CCCCGGGG) (28), and d(GGGGCCCC) (29) crystallize in the A form due to the presence of sequential CC and GG sequences. Poly(dG)·poly(dC) (21) goes into the A genus in saturated salt solution. The sequence d(GGTATACC) crystallizes in the A form and goes into the A form in saturated salt solution (13, 15). Changing the order of GG and CC in this oligomer to obtain d(CCTATAGG) (unpublished data) results in a sequence that is also in the A form in solution. The sequence d(GGATGGGAG) crystallizes in the A form (31) and goes into the A form in saturated salt solution with 20% ethanol (15). The tendency to go into the A genus is undoubtedly helped by the sequences GG and GGG. Finally it should be noted that d(CCGG) crystallizes in the A form but is in neither the A nor the B form in concentrated salt solutions (22). Thus this sequence appears to have no preference for either the A or the B form in aqueous solution at high salt concentration. This unusual experimental observation is in remarkable agreement with the calculated value of ΔE = 0.0. From both theory and experiment it appears that, for this sequence, the base-stacking energies for both the A and the B form are very nearly identical.

Repeated CG Sequences Tend to Induce the B→Z Transition in DNA Containing Other Alternating Pyrimidine-Purine Sequences, but Exact Sequence Is Important

The two-letter code word CG tends to direct sequences into the Z genus, especially when it appears consecutively or in a sequence with other alternating pyrimidine-purine sequences. At pH 7, d(CG) crystallizes in the Z form (15), as do the alternating sequences $d(CG)_n$ for $n > 1$. This is illustrated by crystal and/or concentrated salt solution structures of d(CGCG) (22, 32, 33), d(CGCGCG) (7, 33), and d(CGCGCGCG) (14, 15, 34). In addition to the simple alternating CG frequencies, the following sequences, which contain other unmodified bases, are directed by CG pairs into the Z genus in the crystal and in saturated salt solution (also see Table 1): d(CGCATGCG) (14, 15, 19), d(ACGCGCGT) (15), d(TGCGCGCA) (14, 15), and d(CACGCGTG) (14, 15). The sequence d(ACGCGCGT) (15) is obviously directed into the Z genus by the interior sequence (CGCGCG). The oligomer d(CGTACGTACG) crystallizes in the Z form with the help of the Co(III) cation (35). The presence of the three CG sequences and the alternating pyrimidine-purine, TA, leads to the Z form. The three decamers d(CGCGTACGCG), d(CGCACGTGCG), and d(CGTGCGCACG) all go into the Z genus in saturated salt solution (15), but only d(CGCGTACGCG) crystallizes in the Z form (15). The fact that the other two decamers go into the Z genus in solution but crystallize in the B form is surprising and shows the importance of crystal packing forces that are hard to predict. One possible explanation is the fact that a completed turn of a B-type double helix is 10 bases, whereas the completed turn of a Z-type helix is 12 bases, so that decamers may tend to crystallize in the B form more easily than octamers or dodecamers. The oligomer d(CGTGCACG) does not go into the Z genus in either the crystal or ethanol/salt solutions (14, 15). That d(CGCATGCG) goes into the Z form, while d(CGTGCACG) does not, may be due to the inversion of the CA and TG sequences that occurs in the interior of these oligomers (15).

The two self-complementary hexamers d(CGTACG) (unpublished results) and d(CACGTG)† have recently been shown to crystallize in the B form and Z form, respectively. The former does not go into the Z form under saturated salt conditions (unpublished results); the conformation of the d(CACGTG) sequence in saturated salt solution has not yet been reported. From these results we conclude that a CG word in the interior of an alternating pyrimidine-purine sequence may be more Z-directing than isolated CG words at

†Subirana, J. A., Ninth International Biophysics Congress, Aug. 23–28, 1987, Jerusalem, p. 28 (abstr.).

the ends. This conclusion is supported by the data on the octamers discussed above, since the octamers that have two CG sequences in the interior go into the Z form more readily than those with terminal CG sequences. Note that if the sequence TACG is added to the hexamer d(CGTACG), the sequence d(CGTACGTACG), discussed above, is obtained and this sequence only crystallized in the Z form with the aid of Co(III) (35).

Repeated Sequences of AA or TT Tend to Hold a DNA Oligomer in the B Form

The code word (AA, TT) tends to prevent an oligomer from going out of the B genus. This is in agreement with the ΔE value of +1.01 for this pair. Uninterrupted sequences of adenine or thymine tend to constrain a DNA oligomer to remain in a B genus under any conditions. As will be discussed below, it has been reported that such sequences also tend to bend the DNA while retaining the B conformation, showing that there may be a long-range conformational effect induced by such sequences. However, both NMR and Raman measurements indicate that poly(dA)·poly(dT) is in the B genus in aqueous solution (17, 20). Most oligomers containing repeating sequences of T and of A cannot be induced into the A or Z form. For example, d(AAAAATTTTT) crystallizes in the B genus and remains in the B genus in saturated salt solution and in crystals (16, 20). Films or fibers of poly(dA)·poly(dT) will not go into the A genus at low (75%) humidity or in saturated salt solutions (20). [The B genus may be taken to include the proposed heteronomous conformation in which one strand adopts a conformation with some or all of the furanose rings having the C3′-endo ring pucker usually characteristic of the A genus, but with most of the other torsional angles and the 10-base-per-turn repeat of the B genus remaining essentially the same (20, 41)]. The only known oligonucleotide 4–12 bases in length that contains AA (or TT) and that exists in any form other than the B genus is the octamer d(GGTTAACC), which shows a slight amount of the A form in saturated salt solution but which crystallizes in the B form. AA and TT in d(CGCGAATTCGCG) (11, 12) appear to be responsible for the B conformation of this oligomer in the crystal since, as discussed above, $d(CG)_n$ always crystallizes in the Z form. One must conclude that the presence of even short stretches of $(AA)_n$ or $(TT)_n$ in a DNA will tend to keep it from adopting a Z conformation, whereas longer stretches will prevent the formation of the A form (20).

Code Words TA, AT, GC, CA, AC, GA, and GA and Their Complements Appear To Be Relatively Neutral to Conformational Transitions

The code words TA and AT appear to be neutral and will go into either the A or Z genus depending upon the other base sequences present in the oligomer. Fibers and films of the alternating sequence $d(TA)_n$ can be induced into the A genus at low (75%) humidity (20, 38). It has also been reported that poly[d(A-T)] can be forced into the Z genus upon complexation with Ni(II) (39). There appears to be a difference in the behavior of short $(AT)_n$ and $(TA)_n$ sequences when they occur in the interior of an oligonucleotide. The former has a somewhat greater tendency to keep the oligomer in the B genus, whereas the latter more readily supports both the A and the Z forms. This may be seen from Table 2, where various examples of TA and AT in the A and Z form are given. Neither d(TATATATATA) nor d(ATATATATAT) will go into the A genus in saturated salt solutions (unpublished results). It may be concluded that the TA and AT words exist in sequences undergoing the B→Z or B→A transition but that TA sequences are more readily converted to the A and Z forms than AT sequences. These conclusions, which are based on experimental evidence, are not in agreement with the stacking-energy calculations, since ΔE = −2.04 kcal/mol for AT and +1.12 kcal/mol for TA. This disagreement with stacking energies is strikingly shown in Table 2, where all but one of the sequences listed under AT are calculated to be in the A form but are found in the B form. These sequences have an asterisk by their ΔE values. The only sequence listed under AT that goes into the A form contains only one AT pair.

Experimental data indicate that the code word GC tends to stabilize the B genus and destabilize the A genus. These experimental observations are not in agreement with the stacking-energy calculations (24) where GC has a large negative value of ΔE = −2.56 kcal/mol. (See Table 2 under GC, where again all of the energy values have an asterisk.) Other examples include the oligomers $d(GC)_n$ ($n < 5$), which will not go into the Z form (42). Replacing the GG . . . CC with GC . . . GC converts the d(GGTATACC) sequence into the d(GCTATAGC). Unlike the former sequence, which goes readily into the A form, this latter sequence stays in the B form both in the crystal and in high salt (15). However the calculated ΔE value for this latter sequence is −2.89 kcal/mol, which indicates that this sequence should go into the A form. Our conclusion is that either (*i*) the stacking-energy calculation for the (GC, GC) word is also in error or (*ii*) entropy effects make an important contribution to the free energy of base stacking for this word.

Few data are available for the four two-letter code words (CA, TG), (AC, GT), (AG, CT), and (TC, GA). The ΔE values for these pairs are given in Table 2. With the exception of (CA, TG) they are small and negative. These pairs are occasionally predominant in oligomers found in the A and Z forms.

Junctions May Be Formed When Strong Code-Word Sequences Are Linked Together

It appears that conformational junctions are likely to occur in oligonucleotides in concentrated salt solutions and possibly in crystals when a sequence containing one set of strong words is linked to a sequence containing another set of strong words. Thus candidates for B–A junctions are sequences such as GGGGGTTTTT, GGGGGAAAAA, GGCCGGTTAATT, etc. and their complements. Similarly, candidates for the B–Z junction are CGCGCGTTTTT, CGCGCGTTAATT, etc. and their complements, while candidates for the A–Z junction are GGGGGGCGCGCG, GGCCGGCGCGCG, etc. and their complements. We have begun synthesizing and examining such sequences with Raman spectroscopy. The first sequence we have examined is GGGGGTTTTT. This sequence forms a wobble-paired double helix of the B genus in solution at both low and high salt concentration. However, this sequence forms a Crick–Watson paired duplex with its complement, AAAAACCCCC, that shows all of the characteristic Raman bands of the B form in low-salt solutions. In saturated salt solution, the Crick–Watson duplex produces a Raman spectrum in which the Raman bands belonging to adenine and thymine show clearly only the characteristic B-form frequencies and a total absence of vibrational frequencies characteristic of the A form (38). However, the Raman bands belonging to guanine show the presence of the A conformation as well as B. This indicates an example of an A–B conformational junction in an oligodeoxynucleotide in which the TTTTT portion of the decamer is held entirely in the B form while a transition to the A form occurs at the end of the GGGGG sequence. The length of DNA required to support the B–Z junction is not known.

Possible Biological Significance of Conformational Transitions and Junctions and Predictions Based on Sequence

The question arises as to the biological significance of the B→A and B→Z transitions and possible conformational junctions that may occur in double-helical DNA. Biological significance has been ascribed to both the A and the Z conformation (43, 44). Certain proteins are able to recognize the sequence GGATGGAG in RNA (which is usually in the A form) as well as in DNA (43). Hence, it has been suggested that this DNA sequence may be in the A form in native DNA (43). Similarly, the B→Z transition is often discussed as involving DNA-binding proteins (24, 44). Recently, McLean *et al.* (45) have shown that consecutive A·T pairs can adopt a left-handed DNA conformation in a supercoiled plasmid. Since such sequences as $(AT)_n$ and $(TA)_n$ do not go into the Z form (unpublished data), the importance of supercoiling as an environmental effect is evident. Obviously more work is needed to determine whether or not the A or the Z form actually occurs in DNA *in vivo*. However, as discussed in the introduction, even a partial transition from the B form toward either the A or the Z form would modify the hydrogen-bonding patterns found along the major and minor grooves in the double helix (1–4). This sequence-dependent deformation of the B form could aid the recognition of specific sites on DNA by proteins.

The base-stacking energies of Aida and Nagata (24) are quite useful for predicting the conformation of many of the sequences. However, it seems clear that there is much experimental data from our laboratory and elsewhere that indicate that sequences with high GC and AT content do not appear to go into the A form in either the crystalline state or in concentrated salt solution. The A form would be predicted from the large negative change in stacking energies. It appears either that these calculated base-stacking energies are in error or that other factors such as entropic effects play a dominant role. It is not unreasonable to suppose that entropic effects vary with nearest-neighbor pairs along the double helix. Different neighboring pairs will bind water differently along the major or minor groove, and this could lead to substantial changes in the entropy of nearest-neighbor interactions.

We wish to thank Prof. Peter von Hippel for his careful reading of the manuscript and his helpful comments. We acknowledge financial support from the National Institutes of Health (Grants GM15547 and GM33825).

1. Saenger, W. (1984) *Principles of Nucleic Acid Structure* (Springer, New York), pp. 385–431.
2. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 804–808.
3. Woodbury, C. P., Jr., & von Hippel, P. H. (1981) in *The Restriction Enzymes*, ed. Chirikjian, J. (Elsevier, Amsterdam), Vol. 1, pp. 181–207.
4. Berg, O. G. & von Hippel, P. H. (1987) *J. Mol. Biol.* **193**, 723–750.
5. Franklin, R. E. & Gosling, R. G. (1953) *Nature (London)* **171**, 740–741.
6. Watson, J. D. & Crick, F. H. C. (1953) *Nature (London)* **171**, 964–967.
7. Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G. & Rich, A. (1979) *Nature (London)* **282**, 680–686.
8. Thomas, G. J., Jr., Benevides, J. M. & Prescott, B. (1986) in *Proceedings of the Fourth Conversation in the Discipline Biomolecular Stereodynamics*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Guilderland, NY), pp. 227–253.
9. Peticolas, W. L., Kubasek, W. L., Thomas, G. A. & Tsuboi, M. (1987) in *Biological Applications of Raman Spectroscopy*, ed. Spiro, T. (Wiley, New York), Vol. 1, pp. 81–133.
10. Tsuboi, M., Nishimura, Y., Hirakawa, A. Y. & Peticolas, W. L. (1987) in *Biological Applications of Raman Spectrometry*, ed. Spiro, T. (Wiley, New York), pp. 109–179.
11. Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. (1980) *Nature (London)* **287**, 755–758.
12. Drew, H. R. & Dickerson, R. E. (1981) *J. Mol. Biol.* **151**, 535–556.
13. Shakked, Z., Rabinovich, D., Cruse, W. B. T., Egert, E., Kennard, O., Sala, G., Salisbury, S. A. & Viswamitra, M. A. (1981) *Proc. R. Soc. London Ser. B* **213**, 479–487.
14. Wang, Y., Thomas, G. A. & Peticolas, W. L. (1987) *Biochemistry* **26**, 5178–5186.
15. Wang, Y., Thomas, G. A. & Peticolas, W. L. (1987) *J. Biomol. Struct. Dyn.* **5**, 249–274.
16. Patapoff, T. W., Thomas, G. A., Wang, Y. & Peticolas, W. L. (1988) *Biopolymers*, in press.
17. Behling, W. R. & Kearns, D. R. (1986) *Biochemistry* **25**, 3335–3346.
18. Benevides, J. M., Wang, A. H.-J., Rich, A., Kyogoku, Y., van der Marel, G. A., van Boom, J. H. & Thomas, G. J., Jr. (1986) *Biochemistry* **25**, 41–50.
19. Benevides, J. M., Wang, A. H.-J., van der Marel, G. A., van Boom, J. H., Rich, A. & Thomas, G. J., Jr. (1984) *Nucleic Acids Res.* **12**, 5913–5925.
20. Taillendier, E., Ridoux, J., Liquier, J., Leupin, W., Denny, W. A., Wang, Y., Thomas, G. A. & Peticolas, W. L. (1987) *Biochemistry* **26**, 3361–3368.
21. Nishimura, Y., Torigoe, C. & Tsuboi, M. (1986) *Nucleic Acids Res.* **14**, 2737–2748.
22. Thomas, G. A. & Peticolas, W. L. (1983) *Biochemistry* **23**, 3202–3207.
23. Ornstein, R. L., Rein, R., Breen, D. L. & MacElroy, R. D. (1978) *Biopolymers* **17**, 2341–2360.
24. Aida, M. & Nagata, C. (1986) *Int. J. Quantum Chem.* **29**, 1253–1271.
25. Kubasek, W. L., Wang, Y., Thomas, G. A., Patapoff, T. W., Schoenwaelder, K. H., van der Sande, J. H. & Peticolas, W. L. (1986) *Biochemistry* **25**, 7440–7445.
26. Conner, B. N., Takano, T., Tanaka, S., Itakura, K. & Dickerson, R. E. (1980) *Nature (London)* **295**, 294–299.
27. Conner, B. N., Yoon, C., Dickerson, J. L. & Dickerson, R. E. (1984) *J. Mol. Biol.* **174**, 663–695.
28. Haran, T. E., Shakked, Z., Wang, A. H.-J. & Rich, A. (1987) *J. Biomol. Struct. Dyn.* **5**, 199–217.
29. McCall, M., Brown, T. & Kennard, O. (1985) *J. Mol. Biol.* **183**, 385–396.
30. Wang, A. H.-J., Fujii, S., van Boom, J. H. & Rich, A. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3968–3972.
31. McCall, M., Brown, T., Hunter, W. N. & Kennard, O. (1986) *Nature (London)* **322**, 661–664.
32. Drew, H., Takano, T., Tanaka, S., Itakura, K. & Dickerson, R. E. (1980) *Nature (London)* **286**, 567–573.
33. Crawford, J. K., Kolpak, F. J., Wang, A. H.-J., Quigley, G. J., van Boom, J. H., van der Marel, G. A. & Rich, A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4016–4020.
34. Fujii, S., Wang, A. H.-J., van de Marel, G., van Boom, J. H. & Rich, A. (1985) *Biopolymers* **24**, 243–250.
35. Brennan, R. G., Westhof, E. & Sundaralingam, M. J. (1986) *J. Biomol. Struct. Dyn.* **3**, 649–665.
36. Nishimura, Y., Torigoe, C., Datahira, M., Tate, S., Tanaka, K., Tsuboi, J., Matsuzaki, J., Hotoda, H., Sekine, M. & Hata, T. (1986) *Nucleic Acids Symp. Ser.* **17**, 195–198.
37. Patel, D. J., Kozlowlki, S. A., Hare, R., Reid, B., Ikuta, S., Lander, N. & Itakura, K. (1986) *Biochemistry* **24**, 926–935.
38. Thomas, G. J., Jr., & Benevides, J. M. (1985) *Biopolymers* **24**, 1101–1105.
39. Taillandier, E., Liquier, J. & Taboury, J. A. (1985) in *Advances in Infrared and Raman Spectroscopy*, eds. Clark, R. J. H. & Hester, R. E. (Heyden, New York), Vol. 12, pp. 65–114.
40. Wang, A. H.-J., Hakoshima, T., van der Marel, G., van Boom, J. H. & Rich, A. (1984) *Cell* **37**, 321–331.
41. Arnott, S., Chanrasekaran, R., Hall, I. H. & Puigjaner, L. C. (1983) *Nucleic Acids Res.* **11**, 4141–4155.
42. Quadrifoglio, F., Marzini, G., Yathindra, N. & Cred, R. (1983) *Jerusalem Symp. Quantum Chem. Biochem.* **16**, 61–74.
43. Miller, J., McLachian, A. D. & Klug, A. (1985) *EMBO J.* **4**, 1609–1614.
44. Rich, A., Nordheim, A. & Wang, A. H.-J. (1984) *Annu. Rev. Biochem.* **53**, 791–846.
45. McLean, M. J., Blaho, J. A., Kilpatrick, M. W. & Wells, R. D. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 5884–5888.