Proc. Int. Symp. Biomol. Struct. Interactions, Suppl. J. Biosci.,
Vol. 8, Nos 3 & 4, August 1985, pp. 645-656. © Printed in India.

Recognition schemes for protein-nucleic acid interactions

GIRJESH GOVIL†, N. Y. KUMAR, M. RAVI KUMAR, R. V. HOSUR, KUNAL B. ROY* and H. TODD MILES**

Tata Institute of Fundamental Research, Bombay 400005, India
* All India Institute of Medical Sciences, New Delhi 110029, India
** National Institutes of Health, Bethesda, Maryland 20205, USA

Abstract. The molecular forces involved in protein-nucleic acid interaction are electrostatic, stacking and hydrogen-bonding. These interactions have a certain amount of specificity due to the directional nature of such interactions and the spatial contributions of the steric effects of different substituent groups. Quantum chemical calculations on these interactions have been reported which clearly bring out such features.

While the binding energies for electrostatic interactions are an order of magnitude higher, the differences in interaction energies for structures stabilised by hydrogen-bonding and stacking are relatively small. Thus, the molecular interactions alone cannot explain the highly specific nature of binding observed in certain segments of proteins and nucleic acids. It is therefore logical to assume that the sequence dependent three dimensional structures of these molecules help to place the functional groups in the correct geometry for a favourable interaction between the two molecules.

We have carried out 2D-FT nuclear magnetic resonance studies on the oligonucleotide d-GGATCCGGATCC. This oligonucleotide sequence has two binding sites for the restriction enzyme Bam H1. Our studies indicate that the conformation of this DNA fragment is predominantly B-type except near the binding sites where the ribose ring prefers a ^{3}E conformation. This interesting finding raises the general question about the presence of specificity in the inherent backbone structures of proteins and nucleic acids as opposed to specific intermolecular interactions which may induce conformational changes to facilitate such binding.

Keywords. Protein–nucleic acid interactions; nuclear magnetic resonance; quantum chemistry calculation.

Introduction

Specific protein–nucleic acid interactions play a key role in several stages of regulation and transfer of information in biological system (Helene and Lancelot, 1982). The interactions also have a structural role in the formation of organells such as ribosome, nucleosome, virus and a functional role in DNA repair. Several biological functions are performed through a highly specific sequence or structure dependent recognition of nucleic acids by proteins. However, unlike the Watson-Crick base pairing schemes for nucleic acid—nucleic acid interaction, the rules for protein-nucleic acid interactions are not yet clearly understood.

As is the case with other molecular systems, proteins and nucleic acids recognise each other through a set of relatively weak nonbonded interactions. Such intermolecular

† To whom correspondence should be addressed.

interactions include: electrostatic interactions of the charged basic amino acid residues with negatively charged phosphate groups in the nucleic acid backbone; stacking of aromatic acid residues with nucleic acid bases and base pairs; and hydrogen-bonding between the amino acid side chains or the peptide backbone with nucleic acid bases or base pairs. These interactions may be further stabilised by hydrophobic interactions, metal ions and water bridges.

In our laboratories, we have been approaching the problem of protein–nucleic acid interactions at two different levels. We have carried out quantum chemical calculations at a submolecular level to estimate the relative binding energies of the functional groups in proteins and nucleic acids with one another (Hosur, 1980; Hosur and Pohorille, 1981; Hosur *et al.*, 1981; Kumar and Govil, 1982,1984a, b, c). At an oligomer level, we have carried out 2D-NMR investigations to look for structural diversity in the protein binding regions of oligonucleotides (Hosur *et al.,* 1985). In this paper, we have summarised some of our findings.

Theoretical calculations

Methodology

For the purpose of theoretical calculations of intermolecular interaction energies, we have used a well known methodology (Claverie, 1978) based on second order perturbation theory. In this approach, the binding energy between two molecules (E_t) is expressed as a sum of contributions from electrostatic (E_e), polarization (E_p), dispersion (E_d) and repulsion (E_r) terms

$$E_t = E_e + E_p + E_d + E_r.$$

Simplified formulae have been developed for the various terms. The electrostatic interaction has been estimated by expressing the molecular charge distribution as obtained by the CNDO method (Pople and Beveridge, 1970) in terms of a multicentred multipole expansion and then expressing E_e as a sum of monopole-monopole, monopole–dipole and dipole-dipole interactions. The value of E_p is obtained through the use of bond polarizability (Le Ferre, 1965) while E_d and E_r are calculated using Kitaigorodskii type potential functions (Caillet and Claverie, 1975).

Interaction of basic amino acid residues

Positively charged amino acid residues Lys and Arg and under certain situations His can interact with the negatively charged phosphate group in the nucleic acid backbone through Coulombic interactions. We have calculated the interaction energies between the particular amino acid residues and DNA/RNA fragment-diribose triphosphate. The geometry of the RNA fragment was fixed in its usual A form, while both A and B conformations were used for calculations with the DNA fragment (Arnott and Hukins, 1972; Arnott, 1971). The relative geometries of the nucleotide and the peptide moieties were optimised by energy minimisation using suitable rotations and translations. The optimised geometries are shown in figure 1 and the binding energies are listed in table 1.

It may be noted that the binding energies are of the same order as that for ionic bonds in inorganic salts. Almost 90 % of the interaction energy arises from E_e. It is clear from

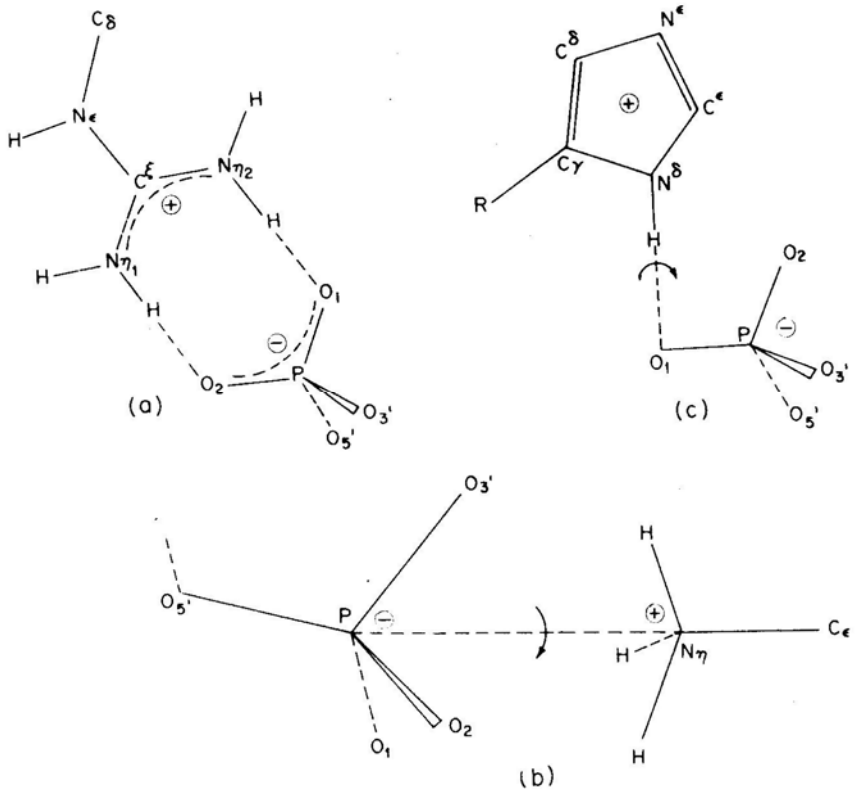

Figure 1. Minimum energy configuration of complexes of basic amino acids with the phosphate group (a) Arg^+, (b) Lys^+ and (c) His^+.

Table 1. Binding energies (in Kcal/mol) for basic amino acids.

	Arg^+	Lys^+	His^+
A-RNA	96·4	130·4	114·8
A-DNA	98·6	128·9	112·2
B-DNA	92·5	122·6	107·0

the results that a certain degree of structural specificity is present at the submolecular level. For example, Lys^+ and His^+ have a greater affinity for RNA rather than A-DNA or B-DNA. However, the interaction of Arg^+ is stronger for A-DNA compared with A-RNA. The relative affinities of basic peptides to nucleic acids follow the order $Lys^+ > His^+ > Arg^+$. Surprisingly, Arg^+ which can form two hydrogen bonds with the phosphate group has a lower binding energy than Lys^+.

In general, A-DNA forms more stable complexes with basic amino acid residues than the B-DNA. There can be two consequences of this effect. First, variations in

conformation of the DNA backbone can result in sites with differential affinity for proteins and variations in sequence dependent backbone geometries can be recognised by the basic amino acid residues. Secondly, the interaction with basic amino acid residue may induce a transition from B to A conformation of DNA.

Hydrogen bond interactions

Complexes involving two or more hydrogen bonds can lead to specific recognition between proteins and nucleic acids because of the directional nature of hydrogen bonds (Bruskov, 1975; Seeman *et al.,* 1976; Helene, 1977). Amino acid residues which can form a pair of hydrogen bonds with nucleic acid bases and base pairs are Asp, Glu, Asn and Gln. In addition, anionic forms of Asp and Glu and cationic form of Arg have been considered for such interactions. Even within the constraint of a minimum of two hydrogen bonds between the two moieties, the number of possible hydrogen bonding schemes is quite large. The possibilities are some what restricted for base pairs since several sites become inaccessible due to base pairing. We have made binding energy calculations for a number of possible hydrogen bond schemes. In each case, the geometry was optimised to give the highest binding energies. Some representative examples of hydrogen bonded configurations are shown in figure 2 and the binding energies are given in table 2.

In general, the charged residues Arg^+, Glu^- and Asp^- bind more strongly than the neutral residues. A large contribution to the binding energies in these cases comes from electrostatic interactions between atoms not directly involved in hydrogen bonds. Thus, these cases are in between the pure Coulombic interaction considered in the previous section and hydrogen bonding for neutral residues. The possible recognition schemes for charged residues are limited. For example, Gluu and Asp^- can interact only with G in single stranded nucleic acids while Arg^+ can form complexes with G and C in both single and double stranded nucleic acids. Thus, Glu^- and Asp^- can selectively recognise G in single stranded nucleic acids, while Arg^+ can recognise G–C base pairs in double stranded nucleic acids.

The hydrogen bond energies for neutral amino acid residues are significantly less. While all the four amino acids (Asn, Gin, Asp and Glu) can form a pair of hydrogen bonds with the four bases and the two base-pairs the binding energies vary significantly. The affinity of nucleic acid bases for the four amino acids follows the order

$$G > C > A > T\ (\ U\ )\ ,$$

and

$$G–C > A\text{-}T\ (U).$$

Similarly, the four amino acids have different affinities for the nucleic acid bases.

Stacking interactions

Aromatic amino acid residues (Trp, Tyr, Phe and His) can bind to nucleic acid bases and base pairs through stacking interactions. Due to a favourable contribution from hydrophobic effects, stacking assumes an even greater importance in aqueous solution.

Energy minimisation shows that a vertical separation of 3·2–3·4 Å, between the aromatic residues of proteins and nucleic acid bases leads to the most stable

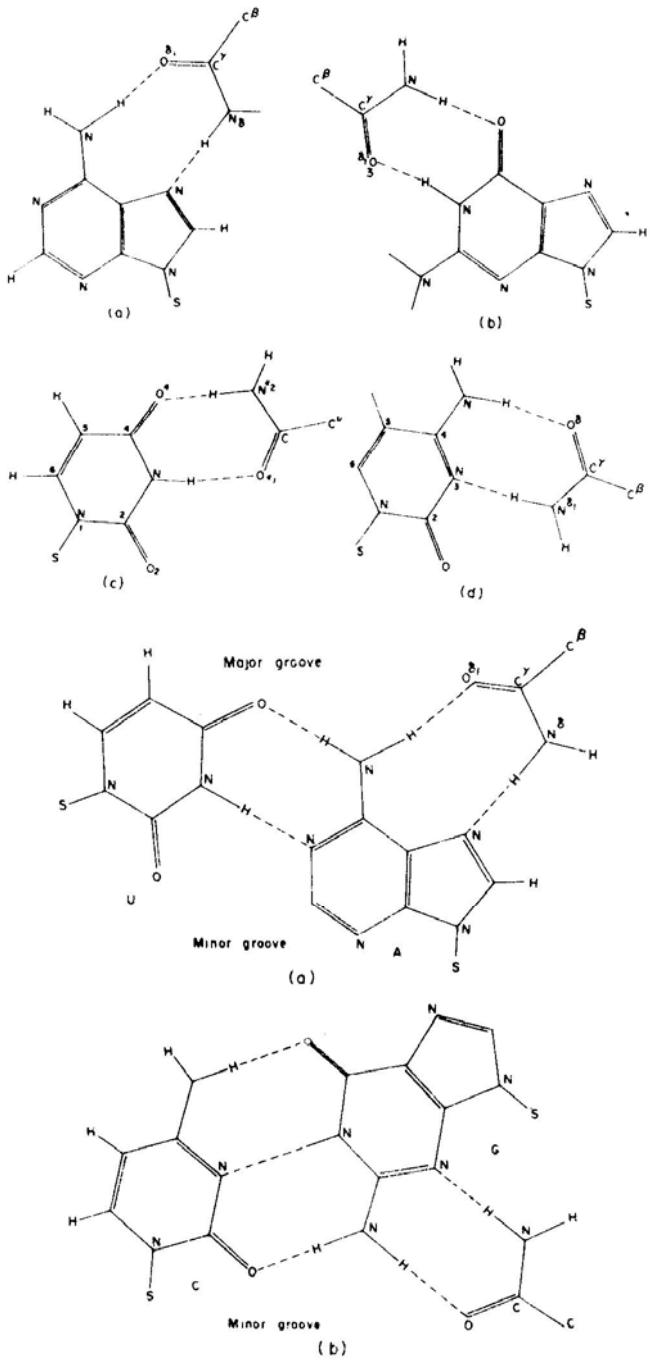

Figure 2. Typical hydrogen bonding schemes involving nucleic acid bases and protein side chains. In the calculations, the two hydrogen bonding moieties have been taken to be coplanar and the energy has been minimised with respect to the inter-moiety distance. (a), Individual bases; (b), base-pairs.

Table 2. Hydrogen bond energies (in Kcal/mol).

	Glu^-	Glu	Asp^-	Asp	Gln	Asn	Arg^+
A	—	9·2	—	9·7	7·9	7·9	—
G	24·6	11·0	23·1	11·2	14·6	15·1	38·1
U	—	6·6	—	6·9	7·5	7·2	—
T	—	6·6	—	6·8	7·4	7·1	—
C	—	9·8	—	10·7	11·3	11·7	36·1
A-U	—	8·0	—	8·6	5·5	5·3	—
A-T	—	8·1	—	8·7	5·6	5·5	—
G-C	—	9·1	—	9·8	5·4	5·7	39·0

configuration. A representative geometry of the optimised configuration is shown in figure 3 and the binding energies are given in table 3. The aromatic moieties of the two components overlap only partially supporting the selective "book mark hypothesis" (Brown, 1970; Gabbay *et al.*, 1972) which states that the nucleotide sequences can be recognised by amino acid side chains acting like book-marks. A common feature of the optimised configuration is that the hetero-atoms of the bases overlap with the aromatic part of the amino acids. The stacking geometry involving base pairs is markedly different from those of the component bases.

Comparison of the stacking energies reveals that G among the purines and C among the pyrimidines have higher binding energies. Thus, in actual protein-nucleic acid interactions, G–C rich regions will be preferred both in single and double stranded nucleic acids. For a particular base, the preference for binding by the amino acid generally follow the order His > Trp > Tyr > Phe.

Other interactions

Amino acid residues other than those listed above can interact with nucleic acid through relatively weak Van der Waal and hydrophobic interactions. However in such cases, an alternative mode of interaction involves hydrogen bonding between the nucleic acid bases and the peptide backbone (Hosur *et al.*, 1981; Hosur, 1980; Hosur and Pohorille, 1981).

Role of macromolecular conformation in protein-nucleic acid recognition

General methodology

Each of the mechanisms described in the previous section has a certain degree of specificity both in terms of sequence and in terms of structure. However, it is not clear from the above calculations whether the protein binding regions of nucleic acid and the respective regions in proteins are structured in a way to achieve a highly specific binding or whether such a shape is induced to optimise the interaction between the functional group in the two classes of macromolecules. This question can be answered only if the detailed three dimensional structure of the two classes of molecules can be investigated.

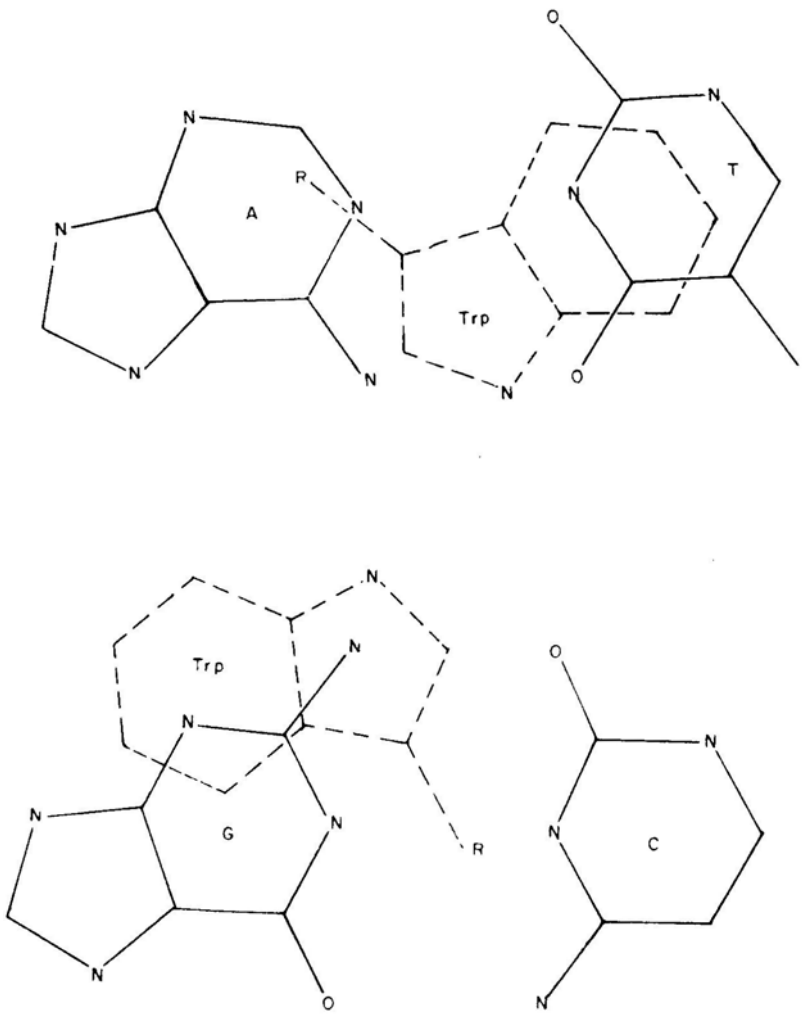

Figure 3. Stacking geometry of Trp with nucleic acid base-pairs. The geometry has been optimised for minimum energy both in terms of overlap and the distance between the two moieties. The planes of the two moieties have been maintained parallel to each other.

NMR has emerged as a very powerful technique for conformational studies in aqueous solutions (Govil and Hosur, 1982). With the advent of two dimensional (2D) Fourier Transform (FT) NMR techniques for elucidiating J coupling correlation (through the technique called COSY) (Aue *et al.*, 1976) and the dipolar coupling correlation (NOESY) (Anil Kumar *et al.*, 1980), it has become possible to determine the three dimensional structures of both proteins and nucleic acids (see Hosur *et al.*, 1985 for relevant references).

We have started investigations on a series of oligonucleotides with different sequences and composition containing protein binding regions, with the following

Table 3. Stacking energies (in Kcal/mol) for aromatic amino acids and nucleic acid bases.

	His	Trp	Tyr	Phe
G	12·5	11·5	9·4	9·1
A	8·7	8·2	7·1	6·3
C	12·8	10·7	8·7	8·3
T	8·8	8·3	7·1	7·5
U	9·4	8·2	6·7	6·8
G-C	14·3	14·5	12·6	10·8
A-T	12·4	14·2	12·1	10·0
A-U	12·1	13·7	11·9	9·8

aims:

(i) devise strategies for resonance assignments and identification of the gross conformational structure;
(ii) identify the influence of the sequence on the local structure of the molecules; and
(iii) correlate structure and protein binding properties.

Strategies for assignment

The resonance assignments in oligonucleotides can be achieved in two steps (i) identify spin systems within the individual nucleotide units from COSY spectrum and (ii) attach the spin systems to particular nucleotides along the chain through the use of NOESY. The second procedure requires a knowledge of intrastrand-interesidue short contacts, which in turn depend on the structure of the molecule. Calculations of interatomic distances show that the assignment strategies for right handed DNA (A and B forms) and the Z-DNA are quite distinct (figure 4). The success of a particular assignment strategy would thus itself pin down the gross conformational details of the molecule. However, more detailed conformational information can be obtained from a careful analysis of intramolecular NOE.

The distances between the base H8 proton in purines or the H6 proton in pyrimidines and the sugar H1′ protons depend on the glycosidic torsion angle χ. The intra-residue H8/H6 . . . . H1′ distance is a minimum for the *syn* domains of the glycosidic torsion angle (for $\chi = 245°$, the distance is around 2·1 A°). Thus, a strong cross peak between the intraresidue H1′ and H8/H6 proton is expected if the base conformation is *syn.* The values of the actual distances between H8/H6 and H2′ or H2″ depend both on the sugar geometry and the glycosidic dihedral angle. The minimum value of these distances occurs around $\chi = 150°$. For ^{2}E conformation, both H2′ and H2″distances from the base proton are less than 2A°. Thus in the high anti domain and a ^{2}E sugar conformation, strong and equally intense cross peaks are expected between H2′, H2″ and H8/H6 protons. In the normal anti domain, the distance H8/H6 to H2′ is shorter than the corresponding distance to H2″, irrespective of the sugar geometry. Thus, such a conformation will produce cross peaks in NOESY spectrimi with unequal intensities.

Information about the sugar geometry can be obtained from the scalar coupling

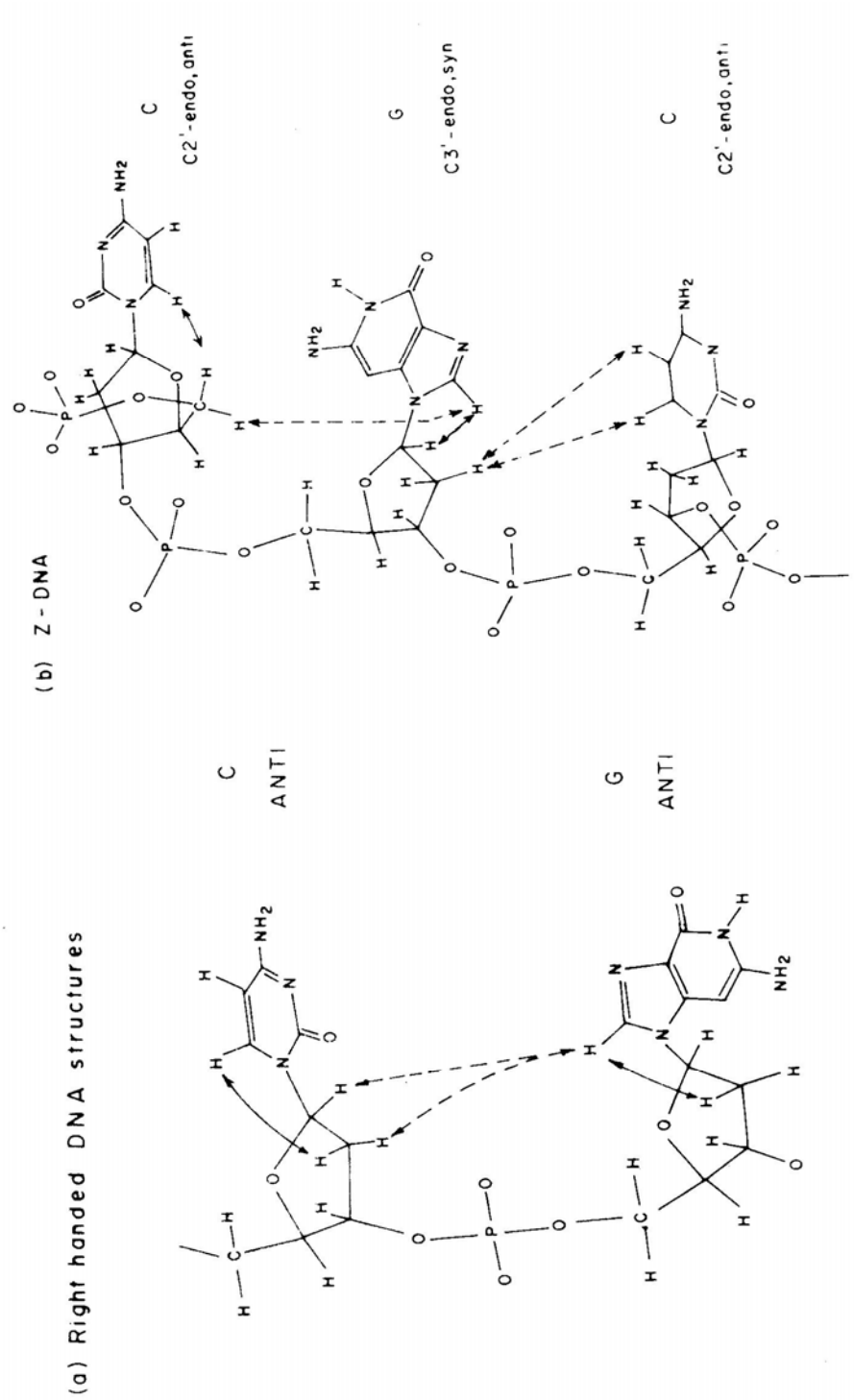

Figure 4. Strategies for sequential resonance assignments from NOESY spectra of oligonucleotides (a) right handed helices (b) left handed Z-type structure. The dotted line indicates short distances in the structure which can be used to jump from one nucleotide to the other within the strand. The solid lines indicate short intra residue distances which can be used for estimating the glycosidic bond angle and for checking intrastrand assignments in conjunction with COSY.

constants J involving H1′ proton and H2′, H2″ protons. For a ^{2}E conformation, both these coupling constants are in the range 6–8 Hz. For a ^{3}E conformation, one of the coupling constants is large (~ 10 Hz), while the other one is small (~ 1 Hz). Thus, qualitative information about the sugar conformation can be obtained from the relative intensities of the cross peaks H1′ · · H2′ and H1′ . . H2″ in the COSY spectra. More accurate values of J couplings and hence the structure can be obtained from the 2D-J resolved spectroscopy.

Structure of d-GGATCCGGATCC

Based on above arguments, we have been able to work out detailed solution conformation of d-GGATCCGGATCC (table 4). We find that the strategy given in figure 4a proves successful for the sequential assignments in this molecule showing that the molecule adopts a right handed conformation. The fact that the deoxyribose conformation for most of the rings is ^{2}E, shows that the oligonucleotide adopts a predominantly B-DNA conformation. Evidence for double helical base-paired structure is obtained from the observation of four additional cross peaks in the NOESY spectrum (figure 5) corresponding to interstrand NOE's between adenine H2 and H1′ protons of thymine and cytosine across the strand.

One of the most interesting findings in the above structure is the fact that sugars attached to G1 and G7 adopt a ^{3}E conformation while a regular B-DNA structure with ^{2}E pucker is adopted by the rest of the molecule. The reason for this local variation in the conformation is not clear. Possibly, by assuming this geometry, the 5′ residues may assume a more favourable interaction with the next residue. However, these are the very

Table 4. Structural information from relative intensities in selected regions of COSY and NOESY spectra.

Base	NOSEY		Inference	COSY		Inference
	NOE from base to H2′	H2″		J-coupling with H1′ H2″	H2′	Sugar Pucker
G	w			s		3′-endo
G	w	o*		s	s	2′-endo
A	s	s*	High Anti	s	s	2′-endo
T	w			s	s	2′-endo
C	s	s*	High Anti	s	s	2′-endo
C	w	s*	High Anti	s		2′-endo[a]
G	w	w*	Anti	s		3′-endo
G	o	o*	Anti	s	s	2′-endo
A	s	s*	High Anti	s	s	2′-endo
T	s	s*	High Anti	s	s	2′-endo
C	s	w*	Anti	s	s	2′-endo
C				s		2′-endo[a]

w = Weak; s = strong; o = overlapping; * = used in sequential assignment.

[a] Equivalence of H2′, H2″ protons in C6 and CI2 did not allow establishment of sugar geometry from COSY spectrum. The sugar conformations in these are thus determined from 2D J-resolved experiment.

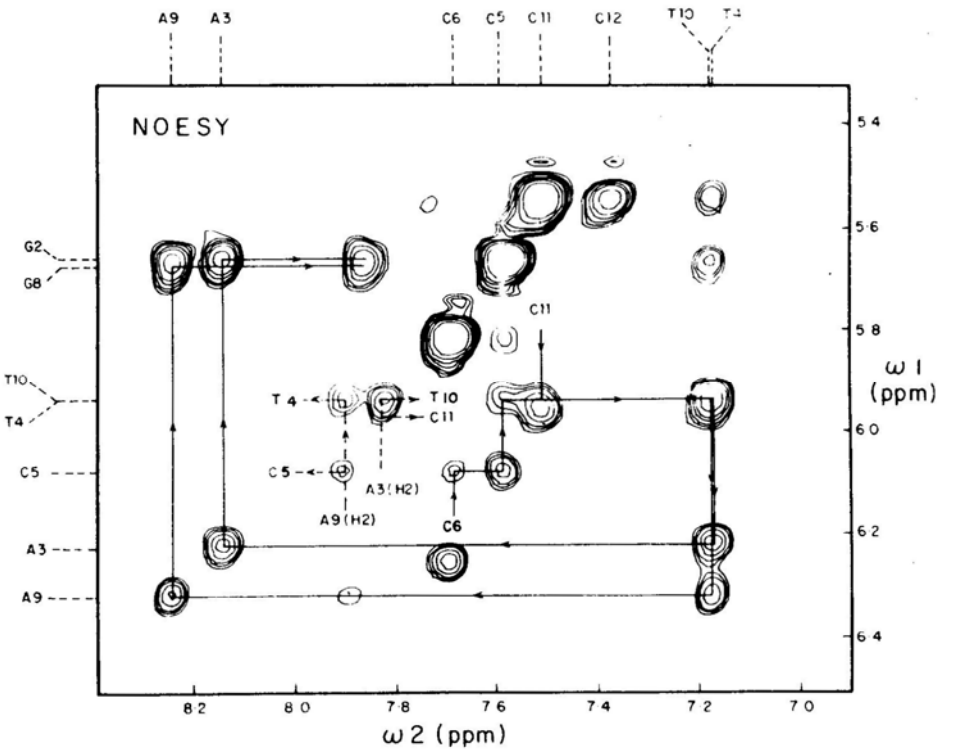

Figure 5. Portion of NOESY spectrum d-GGATCCGGATCC at 25°C and pH = 7·2. Interstrand NOE connectivities between H2 protons of adenines (A3 and A9) and H1′ protons of thymines (T4 and T10) and cytosines (C5 and C11) are indicated by dotted lines. The thick vertical lines are the sequential connectivities from base (H8/H6) protons of one nucleotide to H1′ protons of the previous nucleotide. Thick horizontal lines connect $(H8/H6)_n - (Hl')_n$ cross peak and the sequential $(H8/H6)_{n+1} - (H1')_n$ cross peak.

sites which are recognised by the endonuclease Bam H1 which cleaves the molecule between G1-G2 and G7-G8. Since GC base pairs are also known to provide a stronger binding energy for aromatic amino acid residues (which are involved in recognition in the case of Bam H 1), one may conclude that both structural and binding energies work hand in hand in this case.

We are presently in the process of solving some more protein binding nucleic acid structures to check if this is a general feature.

Acknowledgements

The authors are grateful to the Computer Facility and the 500 MHz FT NMR National Facility at the Tata Institute of Fundamental Research for their help in this work.

References

Arnott, S. (1971) *Prog. Biophys. Mol. Biol.* **21,** 265.

Arnott, S. and Hukins, P. W. L. (1972) *Biochem. J.* **130,** 453.

Aue, P., Bartholdi, E. and Ernst, R. R. (1976) *J. Chem. Phys.* **64,** 2229.

Brown (1970) *Biochim. Biophys, Acta.* **213,** 282.

Bruskov, V. I. (1975) *Molek. Biol.* **9,** 304.

Caillet, J. and Claverie, P. (1975) *Acta. Crystalogr.* **31A,** 448.

Claverie, P. (1978) in *Intermodular Interactions; From Diatomics to Biopolymers,* (ed. B. Pullman) (New York: Wiley) p. 69.

Gabbay, E. J., Dastefano, R. and Sanford, K. (1972) *Biochem. Biophys. Res. Commun.*, **46,** 155.

Govil, G. and Hosur, R. V. (1982) *Conformation of Biological Molecules,* (New York: Springer Verlag).

Hosur, R. V. (1980) *Curr. Sci.,* **49,** 928.

Hosur, R. V., Kumar, N. V. and Govil, G. (1981) *Int. J. Q. Chem.,* **20,** 23.

Hosur, R. V. and Pohorille, A., (1981) *Int. J. Q. Chem.,* **20,** 33.

Hosur, R. V., Ravi Kumar, M., Roy, K. B., Tan Zu-Kun, Miles, H. T. and Govil, G. (1985) in *Magnetic Resonance in Biology and Medicine* (eds G. Govil, C. L. Khetrapal and A. Saran) (New Delhi: Tata McGraw-Hill).

Helene, C. (1977) *FEBS Lett.,* **74,** 10.

Helene, C. and Lancelot, G. (1982) *Prog. Biophys. Mol. Biol.,* **39,** 168.

Kumar, A., Wagner, G., Ernst, R. R. and Wuthrich, K. (1980) *Biochem. Biophys. Res. Commun.,* **96,** 1156.

Kumar, N. V. and Govil, G. (1982) in *Conformation in Biology,* (eds R. Srinivasan and R. H. Sarma) (New York: Adenine Press) p. 313.

Kumar, N. V. and Govil, G. (1984a) *Biopolymers* **23,** 1979.

Kumar, N. V. and Govil, G. (1984b) *Biopolymer*, **23,** 1995.

Kumar, N. V. and Govil, G. (1984c) *Biopolymer* **23,** 2009.

Le Fevre, R. J. W. (1965) *Adv. Phys. Org. Chem.* **3,** 1.

Pople, J. A. and Beveridge, D. L. (1970) *Approximate Molecular Orbital Theory* (New York: McGraw-Hill).

Scheek, R. M., Zuiderweg, E. R. P., Boelens, R., Van Gunsteress, W. F. and Kaptein, R. (1985) in *Magnetic Resonance in Biology and Medicine* (eds G. Govil, C. L. Khetrapal and A. Saran), (New Delhi: Tata McGraw Hill).

Seemann, N. C., Rosenberg, J. M. and Rich, A. (1976) *Proc. Natl. Acad. Sci. USA,* **73,** 804.