Proc. Nat. Acad. Sci. USA
Vol. 73, No. 3, pp. 804–808, March 1976
Biophysics

Sequence-specific recognition of double helical nucleic acids by proteins

(base pairs/hydrogen bonding/recognition fidelity/ion binding)

NADRIAN C. SEEMAN, JOHN M. ROSENBERG*, AND ALEXANDER RICH

Department of Biology, Massachusetts Institute of Technology, Cambridge, Mass. 02139

Contributed by Alexander Rich, December 10, 1975

ABSTRACT **The base pairs in double helical nucleic acids have been compared to see how they can be recognized by proteins. We conclude that a single hydrogen bond is inadequate for uniquely identifying any particular base pair, as this leads to numerous degeneracies. However, using two hydrogen bonds, fidelity of base pair recognition may be achieved. We propose specific amino-acid side chain interactions involving two hydrogen bonds as a component of the recognition system for base pairs. In the major groove we suggest that asparagine or glutamine binds to adenine of the base pair, or arginine binds to guanine. In the minor groove, we suggest an interaction between asparagine or glutamine with guanine of the base pair. We also discuss the role that ions and other amino-acid side chains may play in recognition interactions.**

One of the important unsolved problems in molecular biology concerns the detailed molecular mechanism in the recognition of specific sequences of double helical nucleic acids by proteins. The problem which we are considering is the unique identification in a double helix of each of the four possible base pairs [A·U(T); U(T)·A; G·C; C·G] when compared with each of the other three while they are still in the double helical conformation.

There are only a few ways known by which polypeptides interact specifically with other molecules. These include hydrophobic interactions such as the base stacking interactions, and electrostatic interactions, of which the most important are hydrogen bonds. Because of the high specificity and directional character of hydrogen bonds, we believe they will play a major role in the recognition process. Here we indicate the similarities and differences of the four distinct Watson–Crick base pairs which could be probed by hydrogen bonding groups of proteins for specific recognition. It should be noted that base pair recognition is similar for both double helical DNA and RNA. The result of this analysis leads us to believe that a single hydrogen bond is unable to identify uniquely one base pair with a high degree of fidelity. We propose that specific systems involving pairs of hydrogen bonds between amino-acid side chains and base pairs in the double helix or base pairs and the backbone may be involved in the recognition system.

Base pair recognition

The geometry of Watson–Crick base pairs in the double helix was established from a study of fiber diffraction patterns (1). More recently, the crystallization of fragments of a double helix has yielded a high resolution structural analysis of G·C and A·U pairs in a double helical conformation (2, 3). Fig. 1 illustrates all of the discriminations which a protein must make in order to distinguish between individual base pairs. We will assume throughout that a protein can use the double helical backbone of the nucleic acid in order to establish a frame of reference from which to probe the base pair. A pair of ribose residues in an RNA double helix is shown with two different types of base pairs superimposed. The major groove of the double helix is at the top of each figure while the minor groove is at the bottom. Fig. 1a shows the comparison of the A·U and the U·A base pair. A methyl group would be attached to the 5 position of uracil to illustrate thymine. Fig. 1b superimposes the G·C and C·G pairs, Fig. 1c the A·U and C·G pairs, while Fig. 1d shows the A·U and G·C pairs. The pair U·A superimposed on G·C is also represented in Fig. 1c by simply rotating the pair about a vertical axis, while U·A superimposed on C·G is similarly represented in reverse order in Fig. 1d.

Potential sites for discrimination are labeled in the figure, where W stands for possible recognition sites in the major or wide groove and S for sites in the minor or small groove. These sites have been selected in the figure if the atom or atomic grouping is accessible for hydrogen bonding to the double helix when the probe approaches it. The position arrows point to the heavier atoms. In the amino group the position of the nitrogen atom is taken rather than the two hydrogens which are attached to it, even though recognition at this point must occur through interactions involving the hydrogen atoms. Because the molecule is organized as an antiparallel double helix, the ribose sugars in Fig. 1 are related to each other by a vertical 2-fold axis. The primed recognition sites are related to the unprimed ones of the same number by this same 2-fold axis. These recognition sites are in the same place in all four parts of the diagram. The six possible pairs of base pairs have been illustrated in groups of two for ease of comparison.

Major groove interactions

Six potential recognition sites are found in the major groove as indicated in Fig. 1. W1 is a position close to the imidazole N7 of a purine or C5 of the pyrimidine. W2 is the position occupied by either O4 or N4 of pyrimidines. W3 contains either O6 or N6 of purines. Positions W3′, W2′, and W1′ are related to these three through the dyad axis which relates the two riboses to each other. Only four of these six sites can be found in any one base pair.

Purines may be easily distinguished from pyrimidines by the contents of sites W1 and W1′. For pyrimidines this will be the C5 atom which has a nonpolar hydrogen atom bonded to it; in the purines the imidazole N7 atom is found at that site. A probe to site W1 or W1′ would be capable of discriminating a purine from a pyrimidine, since the purine N7 can accept a hydrogen bond while pyrimidine C5-H group is unable to form this bond and also protrudes about an ang-

Abbreviations: A, adenine; U, uracil; T, thymine; G, guanine; C, cytosine.

* Present address: Department of Chemistry, California Institute of Technology, Pasadena, Calif. 91109.

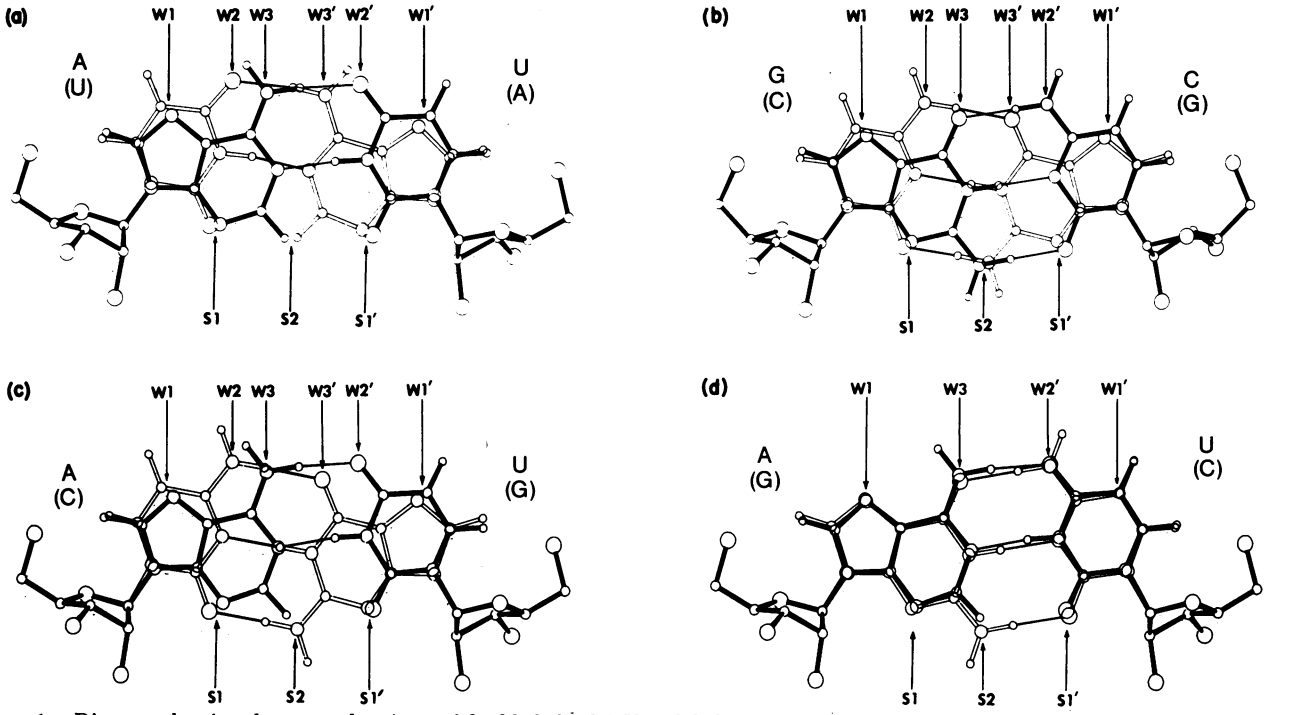

FIG. 1. Diagram showing the stereochemistry of double helical A·U and G·C base pairs. The geometry of the base pairs and the attached ribose residues were obtained from crystallographic analysis of double helical ApU (2) and GpC (3). The base pairs are superimposed upon each other with one base pair drawn with solid bonds and the other with outlined bonds. The upper letter at the side refers to the solid bases while the lower letter in parentheses refers to the outlined bases. However, both bases are drawn as attached to the same ribose residues in the antiparallel double helical conformation. W refers to a potential recognition site in the major or wide groove of the double helix; S refers to sites in the minor or small groove. The dyad axis between the two antiparallel ribose residues is vertical in the plane of the paper. (a) through (d) represent all of the possible base pair comparisons.

strom further into the wide groove. Sterically and electrostatically these sites are quite distinct for purines versus pyrimidines, thereby resulting in good discrimination between these alternatives. By this we mean that it is possible to imagine a conformation for the polypeptide chain such that it could bind to one and not to the other. A simple example in this case would be a hydrogen atom of a protein constrained to hydrogen bond to the N7 atom of the purine. This atom would be sterically incompatible with the C5 hydrogen atom of the pyrimidine, resulting in strong discrimination. Methylation of the 5 position of the pyrimidine such as in thymine or 5-methyl-cytosine would change the details of the interaction and thereby increase the strong discrimination which is already present. Table 1 summarizes the discrimination properties of all sites.

Sites W2 and W3 are geometrically distinct, being separated by 1.1 Å. If the base on the left is a pyrimidine, W2 will be occupied. For uracil (or thymine), W2 will be a carbonyl oxygen; for cytosine, an amino group will be present. If the base on the left is a purine, W3 will be occupied. An amino group is found in adenine and a carbonyl oxygen for guanine. Amino groups characteristically are hydrogen bond donors while the electron-rich carbonyl oxygen atoms are hydrogen bond acceptors. Thus, a combined steric (W2 or W3) and electrostatic (carbonyl or amino) identification system is possible for the central portion of the major groove. The hydrogen bonding recognition H atom of adenine N6 in Fig. 1a is very close to the position of uracil O4, while in Fig. 1b the cytosine N4 H atom lies in the same position as guanine O6. This further strengthens discrimination at these sites.

However, it is likely that such a system involving only

Table 1. Discrimination of Watson–Crick base pairs by single interactions

Sites	A·U / U·A	G·C / C·G	A·U / C·G	G·C / U·A	A·U / G·C	U·A / C·G
Outer major groove (W1/W1′)	+	+	+	+	0	0
Central major groove (W2, W3/W2′, W3′)	+	+	(0)	(0)	+	+
Outer minor groove (S1/S1′)	*	*	*	*	0	0
Central minor groove (S2)	0	(0)	+	+	+	+

The columns in this table refer to Figs. 1a, 1b, 1c, 1c(rev.), 1d, and 1d(rev.), respectively. This table applies to A·T as well as A·U pairs, except for the case of the outer major groove. In that case the two pyrimidines, cytosine and thymine, could be distinguished because of the thymine methyl group. However, the purines would still be degenerate. The symbols are defined as follows: +, indicates sites which could give strong discrimination between the alternatives listed. 0, indicates virtually identical sites resulting in potential ambiguities. (0), indicates only small steric differences, which might result in ambiguities if the interacting atom from the protein is free to move slightly. *, indicates that the hydrogen bonding properties of the site appear identical, but that discrimination could possibly occur through preferential ion binding.

sites W2 or W3 would fail because sites W2 and W3 are too close together to be discriminated reliably by a hydrogen bonding probe. Consider the right side of Fig. 1c, where a uracil and a guanine are to be discriminated by hydrogen bonding interactions at sites W2′ or W3′. It is not hard to position a hydrogen bonding H atom on an amino-acid side chain which could, with a movement of a fraction of an angstrom, donate a hydrogen bond to either uracil at W2′ or to guanine at W3′. A similar situation exists on the left side of Fig. 1c. A hydrogen bonding acceptor could with a small movement receive a hydrogen bond from either the amino N4 of cytosine or from the amino N6 of adenine. Thus, discrimination based solely upon the central major groove sites (W2 and W3) pictured in Fig. 1c would be quite poor, while that shown in Figs. 1a, 1b, and 1d would probably be satisfactory. (Table 1) The conclusion that we reach is that a single hydrogen bonding probe in the major groove would be insufficient to uniquely discriminate all base pairs. This is because small changes in the position of the protein hydrogen bond donor or acceptor would result in confused identification.

Minor groove interactions

The minor groove presents a very different geometric and electrostatic environment. Examination of Fig. 1 indicates that there are only three sites on this side of the base pair which contain functional groups. S1 and S1′ are symmetrically positioned by the vertical dyad axis which relates the sugar residues of the antiparallel chains, while S2 is located directly on this dyad axis. If the base on the left is a purine, S1 will contain an N3 atom while S1′ will contain O2 of the pyrimidine. Reversing the pair will reverse the occupants of S1 and S1′. Since both O2 and N3 atoms are electron rich, they can both act as hydrogen bond acceptors. It should be noted that the atomic centers of the superimposed purine nitrogen and pyrimidine oxygen are less than 0.5 Å away from each other so that little, if any, discrimination can be based on steric factors. In addition, the difference in hydrogen bonding energy to O2 and N3 is not likely to be great enough for polypeptide chains to easily differentiate between these atoms. Thus, it would be difficult or perhaps impossible to discriminate any base pair from another by hydrogen bonding solely to site S1.

The third site, S2, lies on the vertical dyad axis. In the case of the A·U and U·A base pairs, the nonpolar hydrogen atom bonded to C2 of adenine is located almost on the dyad axis. Since this hydrogen does not participate in directional interactions, there is no apparent mechanism for proteins to bind specifically to an A·U base pair rather than to one containing U·A from the minor groove by using either sites S1 or S2. Site S2 of the C·G or G·C base pair differs somewhat from this. Guanine N2 is almost on the dyad axis and its hydrogen atom which is not involved in the third Watson–Crick hydrogen bond projects into the minor groove. It would be easy to distinguish this hydrogen atom from the hydrogen attached to C2 of adenine both electrostatically through hydrogen bonding and sterically, since it protrudes further into the minor groove, as is seen in Figs. 1c and 1d.

Only mild discrimination is possible, however, between the G·C and C·G pairs, as will be noted from inspection of S2 in Fig. 1b. The protruding N-H bond of guanine N2 lies approximately 17° off of the dyad axis and thus the two N-H bonds being compared at S2 diverge by approximately 34°. Even though hydrogen bonds are directional, they distort rather easily. A hydrogen bond acceptor forming a linear hydrogen bond at the correct distance from the H atom on the left side of the dyad axis in Fig. 1b could not form a good hydrogen bond with the H atom on the right. On the other hand, if the acceptor were near the axis, it could interact reasonably well with either H atom. Thus a small motion of an amino-acid side chain could result in an inability to differentiate between a G·C versus a C·G base pair if it were bonding only on site S2.

Recognition in the minor groove is thus likely to be relatively insensitive to base pair reversals, as shown in Fig. 1a for A·U(T) versus U(T)·A, and perhaps 1b. An example of sequence conservation involving base pair reversal is seen in the constant G·C or C·G of the anticodon stem of tRNA (4). An interaction at site S2 will be capable of making the discriminations seen in Figs. 1c or 1d. Aspects of differences in the ability of proteins to detect base pairs in the major and minor groove have been discussed earlier (5), but the ambiguities noted here have not been pointed out.

In both grooves, one fundamental limitation in the discrimination of the individual base pairs by a single hydrogen bonding interaction arises from the difficulty in fixing the precise position of the hydrogen bond donor or acceptor. It should be noted that while small movements of amino-acid side chains are likely to occur, the situation may be quite different with hydrogen bond donors or acceptors involving the polypeptide backbone. These are more likely to be constrained in space and may therefore constitute a recognition system involving single hydrogen bonding interactions (6). In contrast to intermolecular interactions, some intramolecular interactions in the nucleic acids are subject to very precise stereochemical constraints such that a single hydrogen bonding interaction will be adequate to effect identification. An example of this is seen in the three-dimensional structure of yeast phenylalanine tRNA, where a G·C pair which is constant in all tRNA sequences is apparently fixed by an intramolecular hydrogen bond from the cytosine N4 to a neighboring phosphate oxygen atom (7, 8).

Two hydrogen bonds can be used for discrimination

We have suggested above that a single hydrogen bond is incapable of discriminating with great precision a particular base pair in a nucleic acid double helix. This information is summarized in Table 1. A row must contain all +'s for the site to afford complete discrimination, and none of them do. However, these degeneracies can be broken by the addition of a second hydrogen bonding probe at another site. These sites could be widely separated or involve opposite grooves of the double helix. Alternatively they can be close together and both utilize part of the same functional group. The use of two hydrogen bonding interactions in the same functional group provides a mechanism for fixing the position of the two bonds relative to each other with a much higher degree of precision than is possible with two independent hydrogen bonding interactions. Such a group would use the discrimination properties of two rows of Table 1 simultaneously. Furthermore, the (0) entries would convert into + entries since its geometric requirements are quite precise.

An interesting analogy can be made with the way polynucleotides are responsive to base sequences in double helical nucleic acids. For example, the polynucleotide double helix $(rA)_n \cdot (rU)_n$ can add with great specificity a third strand of poly(uridylic acid) which interacts with the double helix using two hydrogen bonding recognition sites (9, 10). Many highly specific polynucleotide interactions are found, all of which have the characteristic property of utilizing two hy-

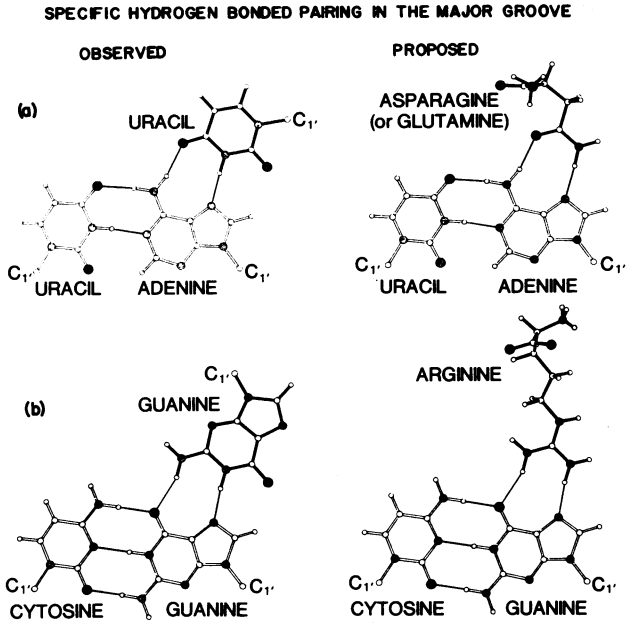

FIG. 2. Interactions between base pairs and other bases (observed) or amino-acid side chains (proposed). (a) The uracil binding to the U·A pair is seen in polynucleotides (11) as well as in single crystals of adenine and uracil derivatives (18). The conformation of the asparagine is taken from a neutron diffraction study (19). (b) The guanine binding to the C·G pair is seen in yeast phenylalanine tRNA (12, 13). The arginine conformation is taken from a neutron crystallographic analysis (20). Oxygen atoms have diagonal shading while nitrogen atoms are stippled.

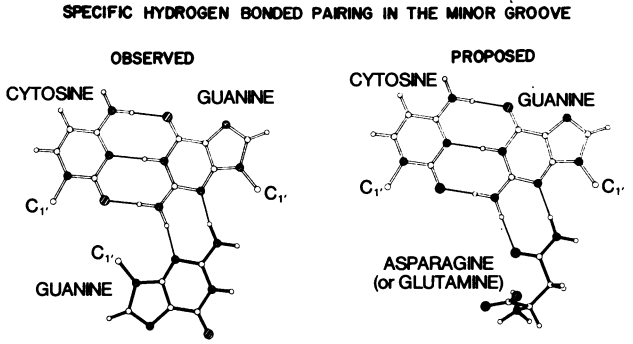

FIG. 3. Interactions between C·G base pairs and guanine (observed) or asparagine (proposed). The guanine interaction with the C·G pair is seen in the structure of the 9-ethyl guanine·1-methyl cytosine complex (16).

drogen bonds as the basis of specificity in the interaction (11). These analogies are useful in suggesting a system in which amino-acid side chains form similar specific pairs of hydrogen bonds with base pairs in the double helix. Examples are shown for the major groove in Fig. 2 and for the minor groove in Fig. 3. Fig. 2a shows the type of hydrogen bonding in which the uracil residue interacts specifically with the A·U pair in the triple-stranded complex (11). Sites W3′ and W1′ of adenine are used to form two specific hydrogen bonds with uracil O4 and N3-H. This may be regarded as a model of the proposed interaction in which amide side chains of amino acids asparagine or glutamine could form a similar pair of hydrogen bonds. In Fig. 2b a hydrogen bonding interaction is shown between guanine residues using sites W1′ and W3′ of the guanine in a C·G base pair. This arrangement is found in the three-dimensional structure of yeast $tRNA^{Phe}$ (12, 13). The guanine amino group N2 and N1-H serve as hydrogen bond donors to the guanine atoms O6 and N7 at sites W3′ and W1′. This may be a model of the proposed interaction in which the NH_2 groups of the guanidinium side chain of arginine are shown forming hydrogen bonds to the same sites. In Fig. 2b we have arbitrarily used the two guanidinium amino groups in this interaction, although it is clear that one amino and one imino group could also be used as the pair of donors.

The guanidinium group is used in a method of separating protein from nucleic acids (14). Furthermore, polymers of arginine have been reported which bind to DNA in a manner which seems to be a function of the G·C content (15). It is possible that these phenomena are related to the interactions described in Fig. 2b.

It is interesting to note that in the case of the asparagine interaction in Fig. 2a an additional amino group might form hydrogen bonds with uracil O4 and the amide oxygen atom. In a similar way in Fig. 2b an additional carboxyl group might further stabilize the interaction of arginine with the C·G base pair by having its oxygen atoms act as acceptors for the H atoms on cytosine N4 and the amino group of arginine.

Fig. 3 shows a way in which sites S1′ and S2 can be used for discriminating the C·G base pair in the minor groove. On the left we see how guanine interacts with the minor groove of the G·C pair as seen in the crystal structure of 9-ethyl guanine and 1-methyl cytosine (16). In this crystal structure the amino group N2 of guanine acts as a donor to site S1′ while N3 acts as an acceptor from site S2. An analog similar to this may be seen in the proposed interaction in which the amide of asparagine or glutamine forms a pair of hydrogen bonds with the same two sites of the guanine in the C·G base pair. An amino group might further stabilize this interaction in a manner analogous to that described for Fig. 2a.

There is discrimination in the specificity of the interaction of actinomycin D with double helical G·C pairs in DNA. In the structure of the complex between actinomycin D and deoxyguanosine, the specificity is determined by two hydrogen bonds from the peptide backbone of the antibiotic to guanine N3 and N2 (17). Even though these hydrogen bonds are made to the peptide backbone of the antibiotic, a donor–acceptor pair is used as suggested above for amide side chains.

The interactions with amino-acid side chains using pairs of hydrogen bonds are able to differentiate unambiguously A·U or G·C pairs in the major groove of the double helix and likewise G·C pairs in the minor groove. However, we have not been able to develop a similar system for differentiating the A·U base pair from the U·A base pair by an interaction in the minor groove.

Other types of recognition

In the above we have limited ourselves to hydrogen bonding interaction that occurs in the plane of the base pair. Probably other types of hydrogen bonding interactions occur that span adjoining base pairs. These interactions will be highly dependent on the conformation of the nucleic acid double helix, and thus may be specific for B-form DNA, for example, but not RNA or vice versa. An example of out-of-plane interactions is likely to be found in the interaction of the lysine side chain with A-T sequences of DNA in the B conformation (to be published). Furthermore, multiple hydrogen bonds from an amino-acid side chain need not involve the bases exclusively. Specificity may also derive from interactions which bridge between the bases and the double helical

backbone. Interactions of this type will be described elsewhere.

There is another type of interaction which may be of potential importance in determining specific base sequences. In the crystal structure of the double helical complex of adenylyl-3′,5′-uridine (ApU) a sodium ion is complexed to the two carbonyl oxygen atoms O2 of the uracil residues in the minor groove of the RNA double helix (2). This type of complex formation can only occur with the sequence ApU, since only this sequence brings the two carbonyl groups to the appropriate positions without any other interference. An ion complex is not seen in the double helical structure of GpC, probably because of a repulsive interaction with the amino group in the minor groove (3). It is possible that carboxyl or hydroxyl groups of amino acids could specifically chelate such a bound ion and create a sequence-determining mechanism. Ion binding of this type in the minor groove may be used to resolve the ambiguity of U·A or A·U base pairs in the minor groove.

In this analysis we have considered the role of hydrogen bonding rather than hydrophobic stacking interactions in the recognition process. Stacking interactions are somewhat sequence dependent. However, it is not obvious at present how the intercalation of planar amino-acid side chains can be used in a recognition system.

In this paper we have proposed a role for hydrogen bonding between proteins and base pairs in a nucleic acid double helix which could be used to distinguish base sequence. Our analysis has led us to the conclusion that single hydrogen bonding interactions are inadequate for the complete identification of base pairs, but that pairs of hydrogen bonded interactions may play a role in this process. It is hoped that the proposals set forth here will serve to stimulate experiments which may eventually reveal the mechanisms for protein–nucleic acid recognition.

This research was supported by research grants from the National Institutes of Health, the National Science Foundation, the National Aeronautics and Space Administration, and the American Cancer Society. N.C.S. is a fellow of the National Institutes of Health.

1. Fuller, W., Wilkins, M. H. F., Wilson, H. R. & Hamilton, L. D. (1965) *J. Mol. Biol.* **12,** 60–80 (*Appendix* by Arnott, S.).
2. Rosenberg, J. M., Seeman, N. C., Kim, J. J. P., Suddath, F. L., Nicholas, H. B. & Rich, A. (1973) *Nature* **243,** 150–154.
3. Day, R. O., Seeman, N. C., Rosenberg, J. M. & Rich, A. (1973) *Proc. Nat. Acad. Sci. USA* **70,** 849–853.
4. Kim, S. H., Sussman, J. L., Suddath, F. L., Quigley, G. J., McPherson, A., Wang, A. H. J., Seeman, N. C. & Rich, A. (1974) *Proc. Nat. Acad. Sci. USA* **71,** 4970–4974.
5. Adler, K., Beyreuther, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Muller-Hill, B., Pfahl, M. & Schmitz, A. (1972) *Nature* **237,** 322–327.
6. Gursky, G. V., Tumanyan, V. G., Zasedatelev, A. S., Zhuze, A. L., Grokhovsky, S. L. & Gottikh, B. P. (1975) *Mol. Biol. (Moscow)* **9,** 635–651.
7. Quigley, G. J., Seeman, N. C., Wang, A. H. J., Suddath, F. L. & Rich, A. (1975) *Nucleic Acids Research* **2,** 2329–2341.
8. Ladner, J. E., Jack, A., Robertus, J. D., Brown, R. S., Rhodes, D., Clark, B. F. C. & Klug, A. (1975) *Proc. Nat. Acad. Sci. USA* **72,** 4414–4418.
9. Felsenfeld, G., Davies, D. R. & Rich, A. (1957) *J. Am. Chem. Soc.* **79,** 2023–2024.
10. Felsenfeld, G. & Rich, A. (1957) *Biochim. Biophys. Acta* **26,** 457–468.
11. Davies, D. R. (1967) *Annu. Rev. Biochem.* **36,** 321–364.
12. Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H. J., Seeman, N. C. & Rich, A. (1974) *Science* **185,** 435–440.
13. Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C. & Klug, A. (1974) *Nature* **250,** 546–551.
14. Enea, V. & Zinder, N. (1975) *Science* **190,** 584–585.
15. Leng, M. & Felsenfeld, G. (1966) *Proc. Nat. Acad. Sci. USA* **56,** 1325–1332.
16. O'Brien, E. J. (1967) *Acta Crystallogr.* **23,** 92–106.
17. Jain, S. C. & Sobell, H. M. (1972) *J. Mol. Biol.* **68,** 1–20.
18. Chandross, R. & Rich, A. (1971) *Biopolymers* **10,** 1795–1807.
19. Verbist, J. J., Lehmann, M. S., Koetzle, T. F. & Hamilton, W. C. (1972) *Acta Crystallogr.* **B28,** 3006–3013.
20. Lehmann, M. S., Verbist, J. J., Hamilton, W. C. & Koetzle, T. F. (1973) *J. Chem. Soc. Perkin Trans. II,* 133–140.