Protein Science (1995), *4*:65-74. Cambridge University Press. Printed in the USA.

Hydrogen bonding motifs of protein side chains: Descriptions of binding of arginine and amide groups

LIAT SHIMONI AND JENNY P. GLUSKER

The Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111

(RECEIVED July 25, 1994; ACCEPTED October 19, 1994)

Abstract

The modes of hydrogen bonding of arginine, asparagine, and glutamine side chains and of urea have been examined in small-molecule crystal structures in the Cambridge Structural Database and in crystal structures of protein–nucleic acid and protein–protein complexes. Analysis of the hydrogen bonding patterns of each by graph-set theory shows three patterns of rings (R) with one or two hydrogen bond acceptors and two donors and with eight, nine, or six atoms in the ring, designated $R_2^2(8)$, $R_2^2(9)$, and $R_2^1(6)$. These three patterns are found for arginine-like groups and for urea, whereas only the first two patterns $R_2^2(8)$ and $R_2^2(9)$ are found for asparagine- and glutamine-like groups. In each case, the entire system is planar within 0.7 Å or less. On the other hand, in macromolecular crystal structures, the hydrogen bonding patterns in protein–nucleic acid complexes between the nucleic acid base and the protein are all $R_2^2(9)$, whereas hydrogen bonding between Watson-Crick-like pairs of nucleic acid bases is $R_2^2(8)$. These two hydrogen bonding arrangements [$R_2^2(9)$ and $R_2^2(8)$] are predetermined by the nature of the groups available for hydrogen bonding. The third motif identified, $R_2^1(6)$, involves hydrogen bonds that are less linear than in the other two motifs and is found in proteins.

Keywords: amide group binding; arginine binding; binding recognition; graph-set theory; protein–nucleic acid interaction

The interactions between proteins and nucleic acids provide a means of controlling processes such as transcription in biological systems. This type of recognition can be achieved in several ways, such as by a "zinc finger" motif (Pavletich & Pabo, 1991), which involves four highly conserved zinc-coordinating residues (Miller et al., 1985; Berg, 1988), or a helix-turn-helix motif (Anderson et al., 1981). Details of such protein–nucleic acid interactions are revealed by X-ray diffraction studies of complexes of proteins and portions of nucleic acids. The specificity of binding between an individual group on the protein and one on the nucleic acid is provided by protein side chains such as arginine, asparagine, glutamine, or histidine. These form hydrogen bonds to a purine, pyrimidine, or phosphate group in DNA.

We examine here the types of interactions formed by amino acid side chains that would be expected to form hydrogen bonds to nucleic acids. The side chains involved in such interactions are: arginine, with five possible proton donor groups; asparagine and glutamine, each with a proton donor and proton acceptor group; and cocrystals involving urea, which also contains proton acceptor and donor groups. The aim was to find out if there are specific binding motifs that can be identified in protein–nucleic acid interactions.

Seeman, Rich, and coworkers (Seeman et al., 1976) used the results of their X-ray analyses of small-molecule crystal structures of nucleic acid bases to study intermolecular interactions between bases and with solvent molecules. As a result, they proposed a recognition code of specific hydrogen bonding patterns between a double helical protein and nucleic acid base pairs. The formation of hydrogen bonds between Watson-Crick base pairs in a double helix is very specific, as illustrated in Figure 1. This specific hydrogen bonding determines how side chains in proteins can recognize the four different combinations in base pairs.

Hydrogen bonding motifs can be described simply by the graph-set theory, as suggested by M.C. Etter (Etter, 1990; Etter et al., 1990; Bernstein et al., 1994). The use of a graph-set notation to represent hydrogen bonds provides a new way to analyze and understand different hydrogen bonding systems. Graph-set theory deals with the identification of repeating patterns in the hydrogen bond networks rather than with the detailed geometrical parameters of these networks. By use of the graph-set notation, a description of very complicated networks can be reduced to that of a combination of four simple patterns: chains (**C**), rings (**R**), finite complexes (dimers) (**D**), and intramolecular (self) hydrogen bonds (**S**). These are the main descriptors of the hydrogen bond network. In addition, information on the number of hydrogen bond donors, **d** (subscript), and the number of hydrogen bond acceptors, **a** (superscript), is added to enhance the specification of each pattern, together with in-

Reprint requests to: Jenny P. Glusker, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, Pennsylvania 19111; e-mail: glusker@fccc.edu.

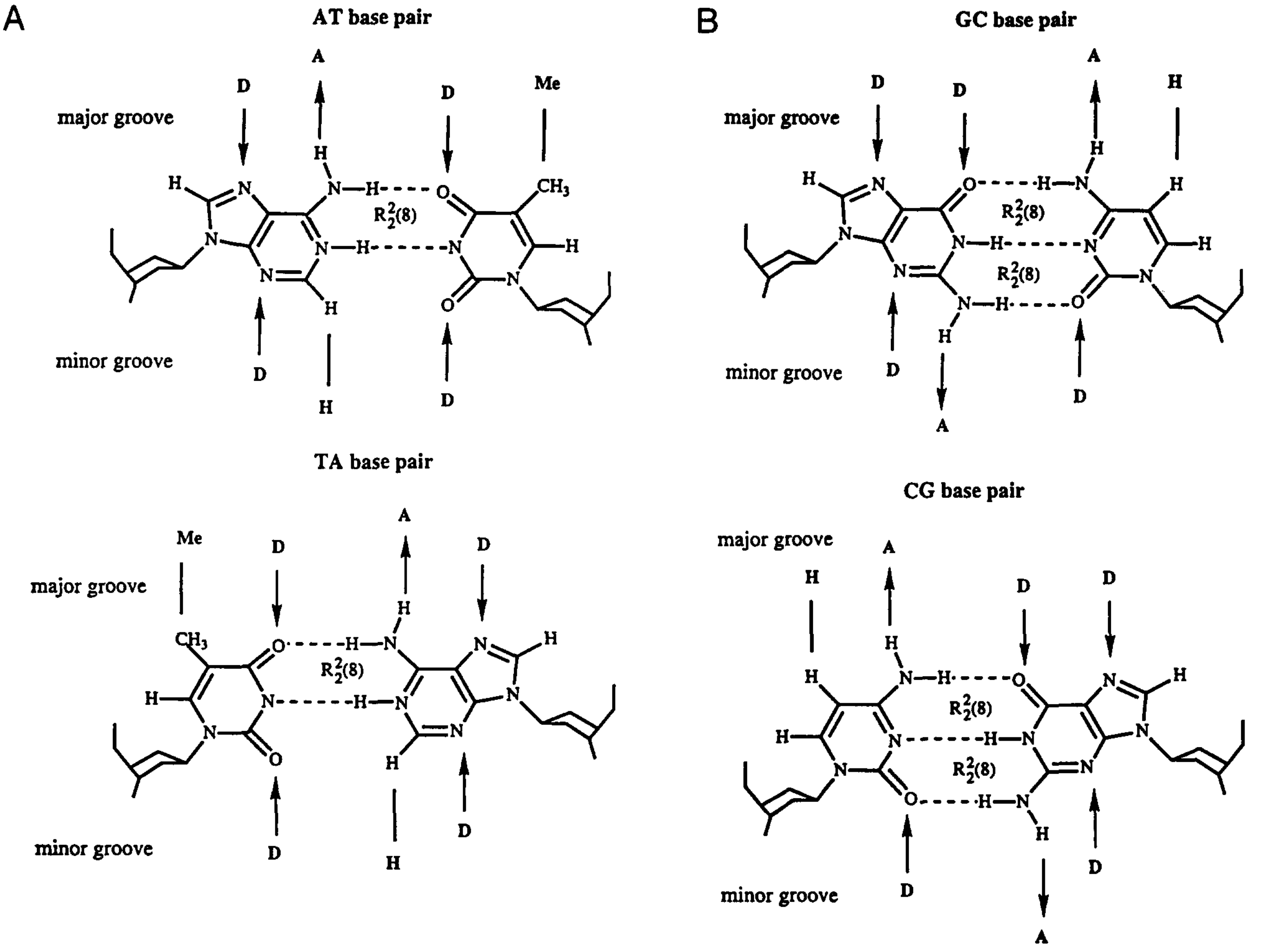

Fig. 1. Formation of hydrogen bonds between Watson-Crick base pairs in a double helix.

Major groove				Minor groove			
D	D	A	H	D	A	D	GC pair
D	A	D	Me*	D	H	D	AT pair
H	A	D	D	D	A	D	CG pair
Me*	D	A	D	D	H	D	TA pair

D, interaction with a hydrogen bond donor: D–H···A; A, interaction with a hydrogen bond acceptor: A···H–D; H, interaction with a C–H group; Me*, interaction with a methyl group. The graph-set notation $R_2^2(8)$ is shown to be formed between base pairs.

formation (in parentheses) on the number of atoms **n** in each pattern (the degree of the pattern). The graph-set notation is then $\mathbf{G}^a_d(n)$, where **G** represents one of the above-mentioned four possible patterns (**C**, **R**, **D**, or **S**).

The analysis of hydrogen bonding patterns in terms of these four simple categories means that different hydrogen bond networks may have the same graph-set notation, even though the arrangement of donor and acceptors may differ. The two (or more) systems are then described as "isographic systems." Such isographic systems may have different topologies, but they have the same pattern (i.e., either chain, ring, finite complex, or intramolecular hydrogen bond), the same number of donors and acceptors, and the same degree of the pattern (**n**). If a pattern can be identified in a hydrogen bond system, even with a relatively large H···X (X = F, O, N, S) bond distance, that interaction is referred to here as a hydrogen bond.

The motif in the binding recognition by hydrogen bonding between base pairs[1] in DNA can be described as $R_2^2(8)$. Possible

[1] We do not include here the formal graph-set assignments, which would first require a listing of all the different motifs in the first level, and then, in the higher levels, the different combination of the different types of hydrogen bonds, i.e., N_2, N_3,.... In this paper, a pattern may include one type of hydrogen bond and therefore should be referred to as a motif. In order to convey our idea more clearly, we refer to one type or a combination of types of hydrogen bonds as a pattern.

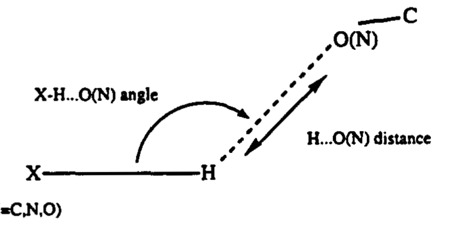

Fig. 2. Geometry of the hydrogen bond interaction that was studied.

recognition sites between proteins and bases in the major and the minor grooves of nucleic acids are diagrammed in Figure 1 (Seeman et al., 1976). The hydrogen bond interactions in the major groove of each of the four Watson-Crick bases can be summarized as follows. If D denotes a proton donor, A a proton acceptor, H a hydrogen atom, and Me a methyl group, then these groups on a protein will recognize individual nucleic acid bases as follows:

Guanine (G):	D	D
Cytosine (C):	H	A
Adenine (A):	D	A
Thymine (T):	Me	D

As can be seen in this scheme, the arrangements of donors, acceptors, and hydrophobic interactions in the major groove of DNA are unique for each of the four bases.

Of these protein-nucleic acid interactions, the more specific recognition mode involves the major groove rather than the minor groove of DNA. Each of the four sets of base pairs is differentiated by the nature of their possible hydrogen bonding in the major groove, but GC and CG are indistinguishable in the minor groove, as are AT and TA, so that interactions in the minor groove are less specific.

Major groove			Minor groove		
GC	DD	AH	GC	DA	D
CG	HA	DD	CG	DA	D
AT	DA	DMe	AT	DH	D
TA	MeD	DA	TA	DH	D

In this study, we consider the geometry and the planarity of the hydrogen bonding system in the different protein-nucleic acid binding motifs (see Fig. 2). Initially, the geometry of hydrogen bonding between small organic compounds (which were studied to high resolution) was analyzed by use of the Cambridge Structural Database (CSD) (Allen et al., 1979), and graph-set theory was used to describe the motifs that were found (Etter, 1990; Etter et al., 1990; Bernstein et al., 1994). Then the crystal structures of protein-nucleic acid complexes were examined in order to identify the hydrogen bonding motifs formed between the protein and the nucleic acid. The resolution of such macromolecular structure determinations is lower than that for small molecules but is usually sufficient for the overall motif to be identified correctly in spite of the lower percussion of the reported intermolecular geometry.

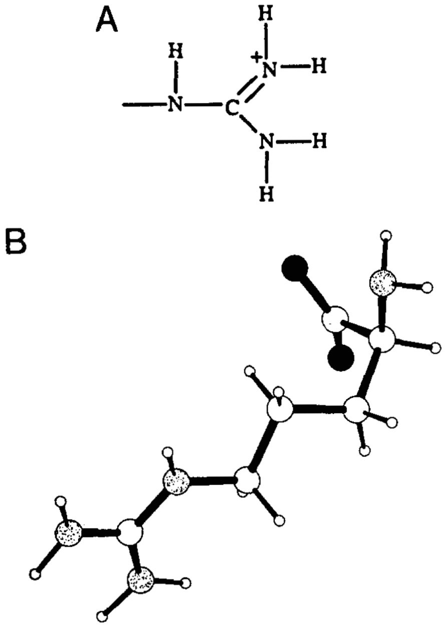

Fig. 3. **A:** Scheme of arginine side chain. **B:** An example for an organic compound containing an arginine side chain L-arginine dihydrate (ARGIND11). In this and all following figures, oxygen atoms are filled balls, nitrogen atoms are speckled balls, carbon atoms are open balls, and hydrogen atoms are small open balls.

Results and discussion

Arginine side-chain interactions

Arginine side chains are commonly used in protein-protein and protein-nucleic acid recognition (Mrabet et al., 1992). The arrangement of atoms found in the arginine side chain is shown in Figure 3A, and an example of the structure of an organic compound containing an arginine side chain is given in Figure 3B. In order to form two hydrogen bonds to an arginine group, an acceptor molecule will have to contain either one hydrogen bond acceptor (resulting in a bifurcated hydrogen bond)[2] or two acceptors (resulting in a bidentate hydrogen bond, that is, two hydrogen bonds). In an analysis using the CSD, 48 crystal structures[3] containing an arginine-like group were found. Of the 48 compounds that were studied in this group, only 27 had a bidentate binding motif. A diagram of the three patterns found that involve two hydrogen bonds to the same functional group is shown in Figure 4A, B, and C. For these, according to Etter et al. (1990), pattern #1 has the graph-set description $R_2^2(8)$,

[2] A bifurcated hydrogen bond involves two donors and one acceptor ($^{H}_{H}$O). A two-center hydrogen bond involves one donor and one acceptor (D-H$\cdots$Acceptor). A grouping $^{O}_{O}$H-O would be described as a three-center hydrogen bond.

[3] Refcodes: AGUAHP, ARGEPO10, ARGGLU10, ARGHCL10, ARGIND11, BGDUSM10, BURSUN, CAXNUK, CEJYAR, CESPAR, COVKUT, COXYET, CUWROB, DIYZEQ, DIVCEQ, DIYZIU, DUBSEY, DULVEL, DUNHID, DUXGIM, DUYCAB, FATZIJ, FEMPAO, GADWUD, GBOPSA10, GEKYOK, GELHIO10, GUABAC10, GUACET, GUAMPR, GUPRAC, HGUANS, JAMREU, JECYUL, KEMYUW, LARASC20, LARGPH01, MEGUHP10, MGLGUH10, NAGLYB10, NETSRN, SIHCAN, SITBIG, SITBOM, STIZOL, STOSEH10, VUZBAT.

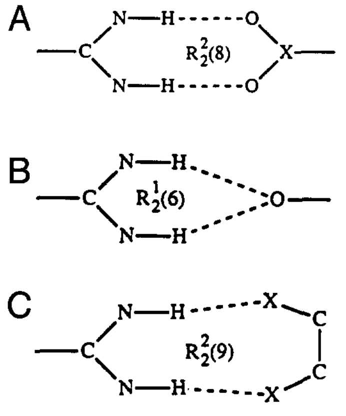

Fig. 4. Three patterns found to be formed in organic compounds with an arginine side chain. **A:** Pattern #1, with graph-set notation $R_2^2(8)$ (X = C, N, S). **B:** Pattern #2, with graph-set notation $R_2^1(6)$. **C:** Pattern #3, with graph-set notation $R_2^2(9)$ (X = O, N).

pattern #2 is $R_2^1(6)$, and pattern #3 is $R_2^2(9)$. Examples of these patterns as they appear in organic compounds are shown in Figure 5A and B. Mean and median H···O bond distances, N-H···O(N) angles, and deviations from the plane defined by the nonhydrogen atoms on the donor functional group are shown in Table 1. These give a measure of the precision of the structural data used.

In all the compounds found in the CSD that have the graph-set description $R_2^2(8)$ (Fig. 5A), the acceptor is an oxygen atom, generally two carboxylate oxygen atoms (CO_2^-) (in 10 compounds), or a sulfite (SO_3^{2-}) (3 compounds) or sulfonate (SO_2^{2-}) (1 compound). In the compounds that have the graph-set description $R_2^1(6)$, the acceptor is an oxygen atom, in half of the cases, a water oxygen (six compounds). Three other acceptors are a carbonyl group (C=O) and carboxylate (CO_2^-) and sulfate (SO_4^{2-}) groups (2, 3, and 1 compounds, respectively). Thus, in spite of the variety of these groups of participating atoms, the pattern of molecular recognition is constant. For the

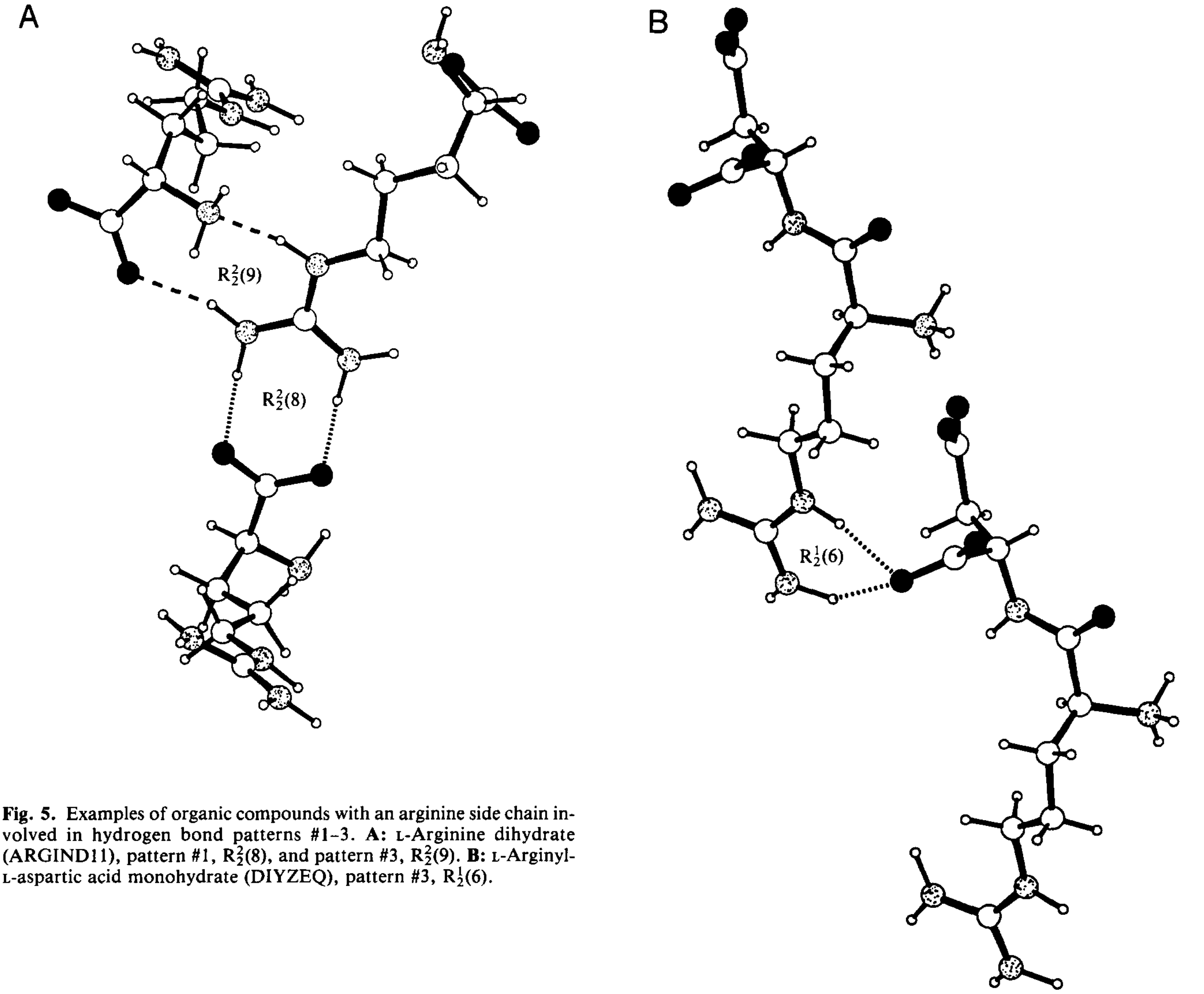

Fig. 5. Examples of organic compounds with an arginine side chain involved in hydrogen bond patterns #1-3. **A:** L-Arginine dihydrate (ARGIND11), pattern #1, $R_2^2(8)$, and pattern #3, $R_2^2(9)$. **B:** L-Arginyl-L-aspartic acid monohydrate (DIYZEQ), pattern #3, $R_2^1(6)$.

Table 1. *Numbers of entries containing a hydrogen bond pattern in compounds containing an arginine side chain*[a]

	Number of entries	Mean H···O(N) distance (Å)	Median H···O(N) distance (Å)	Mean N-H···O(N) angle (°)	Median N-H···O(N) angle (°)	Mean deviation of O from plane (Å)[b]	Median deviation (Å)
Pattern #1[c]							
$R^2_2(8)$	14[d]	1.9 (2)	1.9	165 (12)	168	0.40 (27)	0.13
Pattern #2[e]							
$R^1_2(6)$	12	2.1 (3)	2.0	149 (17)	153	0.39 (20)	0.39
Pattern #3[f]							
$R^2_2(9)$	1	1.9	1.9	163	163	0.58	0.58

[a] A complex may contain more than one pattern or the same pattern several times. Estimated standard deviations are given in parentheses for the last digits listed.

[b] The plane defined by the nonhydrogen atoms in Figure 3A.

[c] Refcodes: ARGIND11, CAXNUK, CUWROB, DIYZIU, GEKYOK, GBOPSA10, GUACET, GUPRAC, HGUANS, JAMREU, JGCYUL, NAGLYB10, VUZBAT.

[d] In seven cases the pattern was formed involving the two primary amines and in seven cases, one secondary amine and the adjacent primary amine.

[e] Refcodes: CEJYAR, DIYZEQ, DULVEL, GELHIO10, GUABAC10, GUACET, JECYUL, LARASC20, LARGPH01, MGLGUH10.

[f] Refcode: ARGIND11.

one compound that has the graph-set description $R^2_2(9)$, one acceptor is a nitrogen amide group (NH_2) and the second is a carboxylate oxygen atom (CO_2^-).

Compounds in this study that exhibit the $R^2_2(8)$ pattern and those that have the $R^1_2(6)$ pattern occur about the same number of times (14 and 12, respectively). In all, the two dominant patterns, $R^2_2(8)$ and $R^1_2(6)$, tend to have H···O distances between 1.9 and 2.1 Å, and both patterns show a tendency to be planar (within 0.39–0.40 Å) with N–H···O angles between 149 and 165°. The hydrogen bond donors are the two primary amine groups in arginine.

Asparagine and glutamine side-chain interaction

The second group of molecules (56 in all) that we studied had an amide functional group such as is found in asparagine or glutamine[4] in Figure 6A. An example of a compound with this functional group is shown in Figure 6B.

In asparagine and glutamine side chains, there are two hydrogen bond donors in the same group, NH_2, and a hydrogen bond acceptor, C=O, that can accept one or two hydrogen bonds. The two patterns that were found to repeat, #4 and #5, are shown in Figure 7A and B, and have the graph-set designations $R^2_2(8)$ and $R^2_2(9)$, respectively. Note that both patterns $R^2_2(8)$ and $R^2_2(9)$ are isographic with patterns found for the arginine side-chain interactions; pattern #4 is isographic with pattern #1 and pattern #5 is isographic with pattern #3 (see introduction for the definition of isographic systems). Examples of these patterns as they appear in organic compounds are shown in Figure 8A and B. The mean and median H···O(N) bond distances, N–H···O(N) angles, and deviations from the plane defined by the nonhydrogen atoms on the donor functional group are listed in Table 2. Of the 56 compounds that were studied in this group, only 30 had a binding pattern in which one acceptor and one donor were both on the same molecule. The $R^2_2(8)$ motif occurs more often in this group (25 compounds) than does the $R^2_2(9)$ motif (five compounds). In all of the compounds that have the graph-set description $R^2_2(9)$ (Fig. 8B), the acceptor groups are carbonyl oxygen atoms (C=O). The mean H···O(N) distance in these two motifs is about the same, near 2.0 Å, but there is a big difference in the deviation from the plane defined by the nonhydrogen atoms of the donor group between $R^2_2(8)$ and $R^2_2(9)$, 0.18 Å and 0.69 Å, $R^2_2(9)$ being the less planar of the

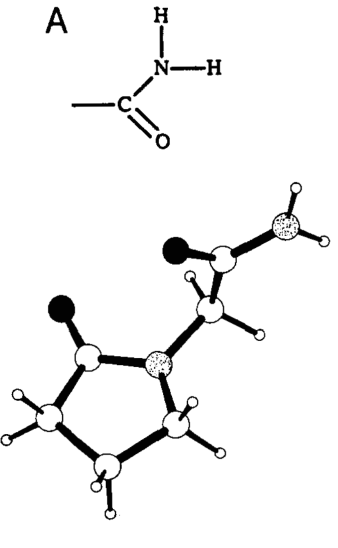

Fig. 6. **A:** Scheme of asparagine and glutamine side chain. **B:** An example of an organic compound containing an asparagine side chain, (2-oxo-1-pyrrolidinyl)-acetamide (BISMEV).

[4] Refcodes: ADIPAM10, ADPROP, AGLUAM10, ASPARM06, AZLMID01, BARKIO10, BAZFUD, BCPPGA, BELZIB, BHXPAM10, BIPHIR10, BISMEV, BISMEV01, BODCIG, BOHJIR, CANKEH, CBMURD, CERNIW, CIMJEN, CIRYUX, CLACAM03, COSDUJ, COXJII, COXJOO, COXKAB, CXMESX, CYANAC, CUCRIB, DAYREA01, DIBMEG, DIRCAI, DPPRAM, DUSMEJ, FACETA01, FECHIE, FESTIG, FOYZIC, FUDZIN, FUSMIP, GAKSEQ, GAXXIM, GEMZED, GIGZUR, GLLASP, GLUTAM01, GLUTAR10, JAHZEX10, JALHIN01, JATLUL, JAXDAN, JAXDER, JAXCUG, LASPZN, MALOAM, NOACTD, NPASPG, OHPHXD, PRHXAM, PSACAM, SAHWON, SAHXAA, SINMEH, SUCABT, SUCCAM10, VIMKEH, ZZZKAY01.

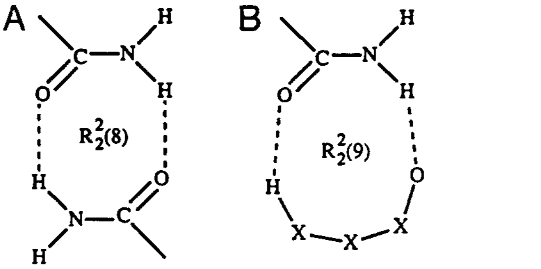

Fig. 7. Two patterns found to be formed in an organic compound containing an asparagine and glutamine side chain. **A:** Pattern #4, with graph-set notation $R_2^2(8)$. **B:** Pattern #5, with graph-set notation $R_2^2(9)$ (X = C, N, O).

two. For 22 hydrogen bonds of the total of 25 that exhibit an $R_2^2(9)$ pattern, only two acceptors on the second molecule were nitrogen instead of the amide oxygen, and in one bond the acceptor was carboxylate oxygen. A small difference was found in the N-H· · ·O(N) angle between the two patterns, patterns #4 and #5, 167° and 161°, respectively.

Urea

The last group of small molecules that we studied contains cocrystals with urea (25 in all).[5] A schematic notation is shown in Figure 9A, and an example for this type of compound is shown in Figure 9B. Urea can be considered to contain a combination of these two functional groups; two amide groups to give two hydrogen bonds to one acceptor group, similar to those formed by the arginine, and it has the amide oxygen atom to give patterns #4 and #5 found in asparagine and glutamine functional groups. Examples of patterns #1, #2, #4, and #5 as they appear in cocrystals containing urea are shown in Figure 10A, B, and C (not that pattern #1 and pattern #4 are isographs). Mean and median H· · ·O distances, N-H· · ·O(N) angles, and deviations from the plane defined by the nonhydrogen atoms on the urea are shown in Table 3.

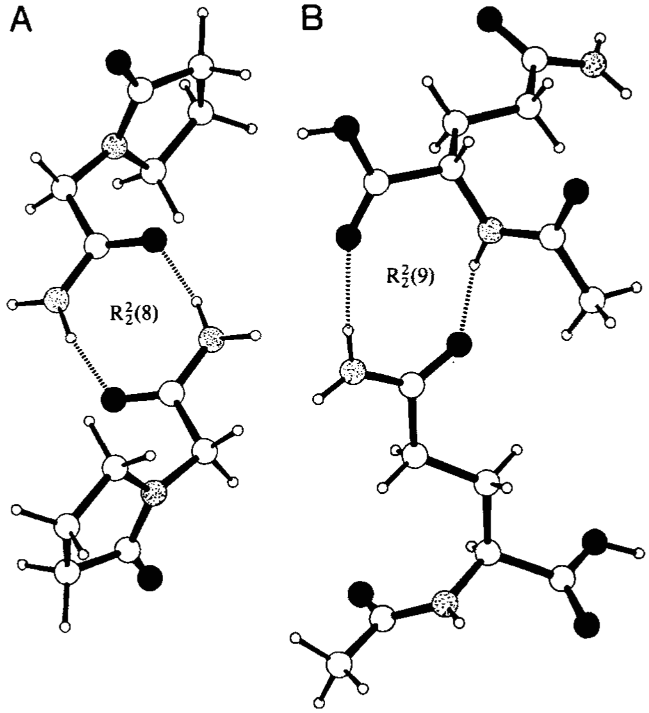

Fig. 8. Examples of organic compounds with an asparagine side chain involved in hydrogen bond patterns #4 and #5. **A:** (2-oxo-1-pyrrolidinyl)-acetamide (BISMAV), pattern #4, $R_2^2(8)$. **B:** N-acetyl-L-glutamine (AGLUAM10), pattern #5, $R_2^2(9)$.

From Table 3, we see that patterns $R_2^1(6)$ and $R_2^2(8)$ appear in large number of compounds (11 and 20, respectively) compared to $R_2^2(9)$, only once. The mean H· · ·O bond distance, N-H· · ·O(N) angle, and the deviation from the plane defined by the nonhydrogen atoms of the urea molecule are about the same for both patterns.

Protein-nucleic acid interaction

Over the last few years, an increasing number of crystal structures of proteins has become available, thus making possible this

[5] Refcodes: ACUFUR, ACURCU01, ACURCU10, ACURLB, ATURUO, ATURUO, BARBUR10, BORTAD, CABRUR10, CANURH, CEFHOK, CRBAMP02, CUFOUR01, FIBXET, JELSEY, SENMUT, SLCADC01, TCYURT10, URCASU, UREAMG, UROXAL, UROXAM, URSDUR03, URPRBN10, VEJXAJ.

Table 2. *Numbers of entries containing hydrogen bond patterns in compounds containing a glutamine side chain*[a]

	Number of entries	Mean H· · ·O(N) distance (Å)	Median H· · ·O(N) distance (Å)	Mean N-H· · ·O(N) angle (°)	Median N-H· · ·O(N) angle (°)	Mean deviation of O from plane (Å)[b]	Median deviation (Å)
Pattern #4[c]							
$R_2^2(8)$	25	2.0 (1)	2.0	167 (13)	171	0.17 (20)	0.07
Pattern #5[d]							
$R_2^2(9)$	5	2.0 (1)	2.0	161 (11)	158	0.69 (43)	0.74

[a] A complex may contain more than one pattern or the same pattern several times. Estimated standard deviations are given in parentheses for the last digits listed.

[b] The plane defined by the nonhydrogen atoms in Figure 6A.

[c] Refcodes: AZLMID01, BISMEV, BISMEV02, BOHJIR, CANKEH, CERNIW, CIMJEN, CIRYUX, CLACAM03, CXMESX, CYANAC, DPPRAM, FACETA01, GAXXIM, GIGZUR, JAHZEX10, JATLUL, JAXDAN, LASPZN, OHPHXD, PRHXAM, SUCABT, SUCCAM10, VIMKEH, ZZZKAY01.

[d] Refcodes: AGLUAM10, FESTIG, FOYZIC, FUSMIP, GLUTAR10.

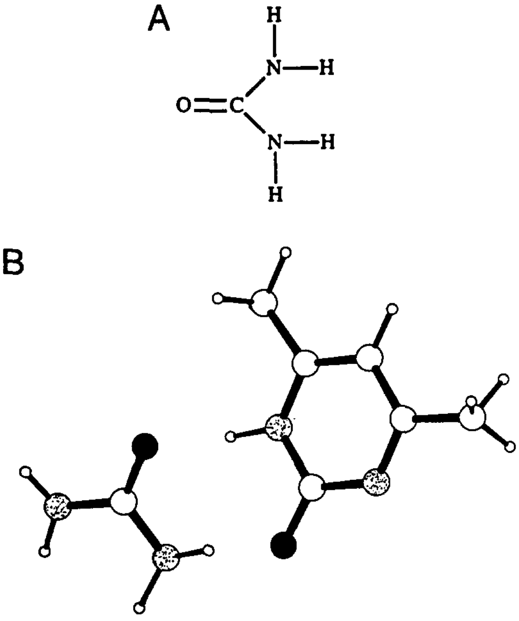

Fig. 9. **A:** Scheme of urea. **B:** An example of a cocrystal containing urea molecule 2,4-dimethylpyrimidin-2-one urea adduct (JELSEY).

study of the protein-nucleic acid binding interaction (Steitz, 1993). After studying the geometry of the binding recognition of the functional group of arginine, asparagine, and glutamine in small organic compounds, we looked at the binding recognition between protein and double-stranded nucleic acids that involve two hydrogen bonds to the same functional group (bidentate). Based on Figure 1, we can predict the only possible bidentate binding pattern for adenine is pattern #5 and for guanine is pattern #3, which is one of the two isographic systems found in the study of binding patterns found in small organic compounds. We examined protein-nucleic acid binding interactions, and the results are shown in Table 4.

In the binding recognition between protein and nucleic acid, the only two isographic patterns that were found were #3 and #5, both having the graph-set of $R_2^2(9)$. The arginine on the protein binds to guanine in the nucleic acid to give pattern #3 ($R_2^2(9)$), as shown in Kinemage 2, and the asparagine and glutamine on the protein bind to the adenine in the nucleic acid to give pattern #5 ($R_2^2(9)$), as shown in Kinemages 1 and 3. Reference to the *Atlas of Protein Side-Chain Interactions* (Singh & Thornton, 1992) indicates that, in protein crystal structures, a two-hydrogen bond binding interaction involving the secondary and primary amine hydrogen atoms of arginine in a protein appear to have preference over the binding interaction involving two primary amine hydrogen atoms. This difference, in the preference for different hydrogen bonding patterns between the systems for small organic compounds and those for proteins, should be noted. An examination of amide group interactions in the *Atlas of Protein Side-Chain Interactions* indicated that asparagine and glutamine functional groups show the same binding preferences in proteins as they do in small molecules.

Conclusion

This graph-set analysis has identified the types of hydrogen bond motifs found in protein–nucleic acid complexes and has compared these with experimental results in small molecules at higher resolution. In small organic compounds that contain an arginine or arginine-like group, the patterns $R_2^2(8)$ and $R_2^1(6)$ are preferred. In macromolecular crystal structures, the binding recognition isographic pattern between any two base pairs was found to be $R_2^2(8)$, and the binding recognition isographic pattern between proteins and nucleic acids was found to be $R_2^2(9)$. Although the types of hydrogen bonding systems appear to be similar in small molecules and macromolecular crystal structures, the relative frequencies are different and appear to be a function of the type of interaction involved. For example, we

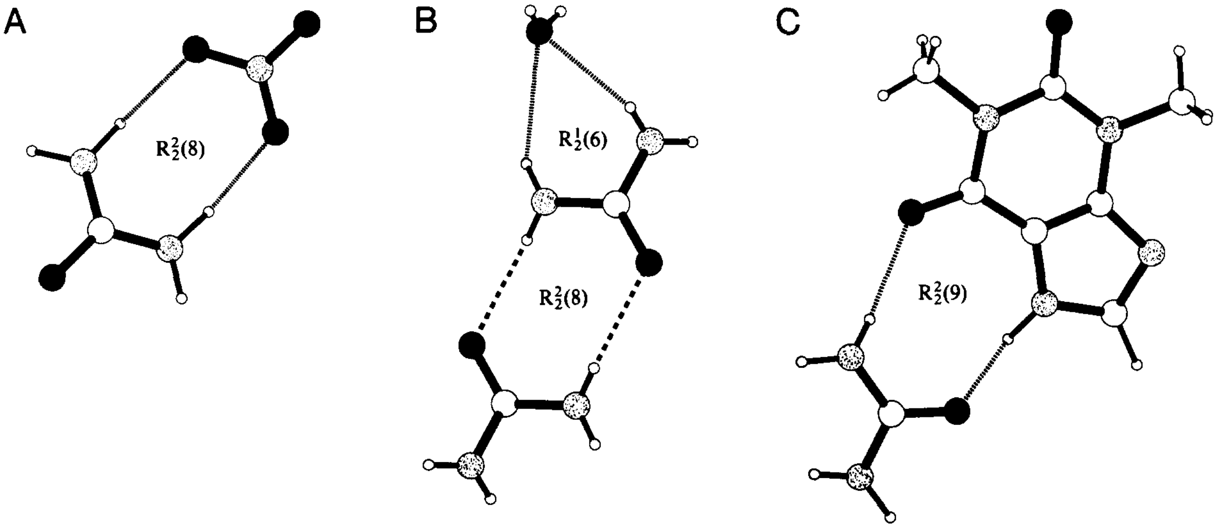

Fig. 10. Examples of organic cocrystals containing urea have a hydrogen bond pattern. **A:** Aqua-tetrakis (urea)-dioxo uranium (ATURNO), pattern #1, $R_2^2(8)$. **B:** Tetrakis (μ-acetato)-bisfurea-copper(ii) dehydrate (ACURCU01), patterns #2 and #4, $R_2^1(6)$ and $R_2^2(8)$, respectively. **C:** Theophylline urea (DUXZAX), pattern #5, $R_2^2(9)$.

Table 3. *Numbers of entries containing hydrogen bond patterns in cocrystals involving urea*[a]

	Number of entries	Mean H···O(N) distance (Å)	Median H···O(N) distance (Å)	Mean N-H···O(N) angle (°)	Median N-H···O(N) angle (°)	Mean deviation of O from plane (Å)[b]	Median deviation (Å)
Pattern #1[c]							
$R_2^2(8)$	6	2.1 (2)	2.1	158 (14)	163	0.58 (49)	0.49
Pattern #2[d]							
$R_2^1(6)$	11	2.2 (2)	2.2	149 (9)	151	0.36 (33)	0.20
Pattern #4[e]							
$R_2^2(8)$	14	2.1 (2)	2.0	160 (17)	163	0.24 (20)	0.19
Pattern #5[f]							
$R_2^2(9)$	1	2.0	–	174	–	0.11	–

[a] A complex may contain more than one pattern or the same pattern several times. Estimated standard deviations are given in parentheses for the last digits listed.

[b] The plane defined by the nonhydrogen atoms in Figure 9A.

[c] Refcodes: ACUFUR, ATURUO, FIBXET, JELSEY, URCASU, URSDUR03.

[d] Refcodes: ACURCU01, ACURCU10, ACURLB, BARBUR10, BORTAD, CANURH, CEJXOE, CRBAMP02, TCYURT10, UREAMG, VEJXAJ.

[e] Refcodes: ACURCU01, ACURCU10, ACURLB, ATURUO, BARBUR10, CUFOUR01, JELSEY, SENMUT, SLCADC01, UROXAL, UROXAM, URPRBNI0, URSPUR03, VEJXAJ.

[f] Refcode: DUXZAX.

Table 4. *Protein-nucleic acid complexes showing single amino acid-nucleotide recognition using bidentate hydrogen bonds*

DNA binding structure motif	Complex (PDB code for protein)	Protein-base pair graph-set pattern	Motif[a]	Amino acid-base interaction	Resolution	References
Helices of the dinucleotide fold	*Eco* RI-DNA (1R1E)	$R_2^2(9)$	#3	Arg 200-guanine	2.8 Å	McClarin et al., 1986; Rosenberg, 1990
Helix-turn-helix	λ-Repressor-DNA (1LMB)	$R_2^2(9)$	#5	Gln 44-adenine	2.5 Å	Jordan and Pabo, 1988
	λ Cro repressor-DNA (4CRO)	$R_2^2(9)$	#5	Gln 27-adenine	3.9 Å	Brennan et al., 1990
		$R_2^2(9)$	#3	Arg 38-guanine		
	434 Repressor-DNA (2OR1)	$R_2^2(9)$	#5	Gln 28-adenine	3.2 Å	Aggarwal et al., 1988
	434 Cro repressor-DNA (3CRO)	$R_2^2(9)$	#5	Gln 28-adenine	3.2 Å, 5.5 Å	Wolberger et al., 1988
	Trp repressor-DNA (1TRO)	$R_2^2(9)$	#3	Arg 69-guanine	2.4 Å	Otwinowski et al., 1988
	Homeodomain-DNA (1HDD)	$R_2^2(9)$	#5	Asn 51-adenine	2.8 Å	Kissinger et al., 1990
	CAP-DNA (1CGP)	$R_2^2(9)$	#3	Arg 180-guanine	3.0 Å	Schultz et al., 1991
Zinc fingers	Mouse zinc finger-DNA (1ZAA)	$R_2^2(9)$	#3	Arg 18-guanine	2.1 Å	Pavletich and Pabo, 1991
		$R_2^2(9)$	#3	Arg 24-guanine		
		$R_2^2(9)$	#3	Arg 46-guanine		
		$R_2^2(9)$	#3	Arg 74-guanine		
		$R_2^2(9)$	#3	Arg 80-guanine		
	Glucocorticoid receptor-DNA (1GLU)	$R_2^2(9)$	#3	Arg 466-guanine	2.9 Å	Luisi et al., 1991
	GLI-DNA	$R_2^2(9)$	#3	Arg 149-guanine	2.6 Å	Pavletich and Pabo, 1993
	Tramtrack-DNA (2DRP)	$R_2^2(9)$	#3	Arg 152-guanine	2.8 Å	Fairall et al., 1993
		$R_2^2(9)$	#3	Arg 152-guanine		
Leucine zipper	GCN4-DNA (1YSA)	$R_2^2(9)$	#3	Arg 243-guanine	2.9 Å	Ellenberger et al., 1992
β-Sheets	Arc-repressor-DNA (1PAR)	$R_2^2(9)$	#5	Gln 9-adenine	2.6 Å	Raumann et al., 1994
		$R_2^2(9)$	#5	Gln 9′-adenine		
		$R_2^2(9)$	#3	Arg 13′-guanine		

[a] The atom arrangement in motif #3, Figure 4C or motif #5, Figure 7B.

have shown that $R^2_2(9)$ is a common motif in protein–nucleic acid interactions.

The third pattern found in small organic compounds, $R^1_2(6)$, was found so far in protein–protein interactions, for example, in D-xylose isomerase (Carrell et al., 1994) crystal structure solved to a resolution of 1.6 Å. In this protein, arginine side chains are the main residues involved in protein–protein interactions. Eleven arginine residues are involved in the binding pattern $R^1_2(6)$ (see Kinemage 4), where the acceptor is mainly a water molecule (not a disordered one) and seven other arginine residues are involved in the binding pattern $R^2_2(8)$ with aspartic or glutamic acids as the acceptor group (see Kinemage 4).

The three graph-set patterns listed are the optimal binding patterns and they appear in many macromolecular complexes of proteins with nucleic acids when the balance of hydrogen bonding donors and acceptors is that normally found in biological systems. Based on the study of small organic molecules, the two hydrogen bonds to the same functional group system tend to give a planar motif. The asparagine and glutamine functional groups show the same preference in the binding recognition in small organic compounds and protein side chains, that is, a bidentate hydrogen bond involving two hydrogen bonds.

We are now investigating the energies of these systems, the extent to which there are deviations in macromolecular crystal structures from these initial rules, and the graph-set descriptions of such perturbed systems.

The role of arginine side-chain "indirect" interaction, i.e., with the backbone carbonyl oxygen, in the maintenance of protein tertiary structure was studied and well established by Pett and coworkers (Borders et al., 1994). The "structural" arginine, whether using the criteria of five or more hydrogen bonds to three or more backbone carbonyl oxygens, down to three or more hydrogen bonds to two or more backbone carbonyl oxygens (even without a single structure motif), is a complementary result to the one we found in our study on the arginine binding motif in a "direct" interaction, i.e., between arginine side chains and nucleic acids.

Materials and methods

Atomic parameters from crystal structures were extracted from the CSD (Allen et al., 1979). The geometry of the hydrogen bond interactions investigated in this analysis is illustrated in Figure 2. The CSD was searched for three types of functional groups: (1) those with a side-chain group as found in the amino acid arginine, (2) those with a side chain analogous to that in the amino acids asparagine and glutamine, and (3) cocrystals involving urea. The query information for the geometrical search was stored in an instruction file (*.que). The CSD program QUEST was run in order to search the database file and obtain a list of structures that matched the input requirements. The coordinates of the atoms in these crystal structures were then written on a data file (*.dat). The CSD program GSTAT was then run on this coordinate data file using as input an instruction file *.geo that caused selected geometrical quantities to be calculated and listed. The positions of hydrogen atoms were normalized to mean neutron distances and $N–H\cdots O(N)$ angles, based on values listed in a study by Taylor and Kennard (1984). In this way, values for the geometry of the hydrogen bonding, the angles, and the deviations from the plane associated with the selected fragment were obtained. In all the illustrations drawn with the program ICRVIEW (Erlebacher & Carrell, 1992), filled spheres represent an oxygen atom, the stippled sphere represents a nitrogen atom, and the open sphere represents a carbon atom.

Details of input files are given in Table S1 in the Electronic Appendix (SUPLEMNT directory, file Shimoni.TS1). A list of CSD reference codes (the identification code in the CSD files for each crystal structure reported) is given in Table S2 in the Electronic Appendix (\SUPLEMNT\Shimoni.TS2), together with bibliographical information and geometry data on structures used in this study.

Acknowledgments

We thank Prof. Joel Bernstein of Ben Gurion University for many helpful discussions. This work was supported by grants CA-10925, GM-44360, and CA-06927 from the National Institutes of Health, CN-10 from the American Cancer Society, and by an appropriation from the Commonwealth of Pennsylvania.

References

Aggarwal AK, Rodgers DW, Drottar M, Ptashne M, Harrison SC. 1988. Recognition of a DNA operator by the repressor of phage 434: A view at high resolution. *Science 242*:899–907.

Allen FH, Bellard S, Brice MD, Cartwright BA, Doubleday A, Higgs H, Hummelink T, Hummelink-Peters BG, Kennard O, Motherwell WDS, Rodgers JR, Watson DG. 1979. The Cambridge Crystallographic Data Center: Computer-based search, retrieval, analysis and display of information. *Acta Crystallogr B 35*:2331–2339.

Anderson WF, Ohlendorf DH, Takeda Y, Matthews BW. 1981. Structure of the cro repressor from bacteriophage λ and its interaction with DNA. *Nature 290*:754–758.

Berg JM. 1988. Proposed structure for the zinc-binding domain from transcription factor IIIA and related proteins. *Proc Natl Acad Sci USA 85*:99–102.

Bernstein J, Davis RE, Shimoni L, Chang NL. 1994. Graph-set analysis of hydrogen-bond patterns in organic crystals. A guide for the perplexed. *Angew Chem*. Forthcoming.

Borders CL Jr, Broadwater JA, Bekeny PA, Salmon JE, Lee AS, Eldridge AM, Pett VB. 1994. A structural role for arginine in proteins: Multiple hydrogen bonds to backbone carbonyl oxygens. *Protein Sci 3*:541–548.

Brennan RG, Roderick SL, Takeda Y, Matthews BW. 1990. Protein–DNA conformational changes in the crystal structure of a λ Cro–operator complex. *Proc Natl Acad Sci USA 87*:8165–8169.

Carrell HL, Hoier H, Glusker JP. 1994. Modes of binding substrates and their analogues to the enzyme D-xylose isomerase. *Acta Crystallogr D 50*:113–123.

Ellenberger TE, Brandi CJ, Struhl K, Harrison SC. 1992. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: Crystal structure of the protein–DNA complex. *Cell 71*:1223–1237.

Erlebacher J, Carrell HL. 1992. ICRVIEW. Philadelphia, Pennsylvania: Program from the Institute for Cancer Research.

Etter MC. 1990. Encoding and decoding hydrogen-bond patterns of organic compounds. *Acc Chem Res 23*:120–126.

Etter MC, MacDonald JM, Bernstein J. 1990. Graph-set analysis of hydrogen-bond patterns in organic crystals. *Acta Crystallogr B 46*:256–262.

Fairall L, Schwabe JWP, Chapman L, Finch JT, Rhodes D. 1993. The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature 366*:483–487.

Jordan SR, Pabo CO. 1988. Structure of the lambda complex at 2.5 Å resolution. Details of the repressor–operator interactions. *Science 242*:893–899.

Kissinger CR, Liu B, Martin-Blanco E, Kornberg TB, Pabo CO. 1990. Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: A framework for understanding homeodomain–DNA interactions. *Cell 63*:579–590.

Luisi BF, Xu WX, Otwinowski Z, Freedman LP, Yamamoto KR, Sigler PB. 1991. Crystallographic analysis of the interactions of the glucocorticoid receptor with DNA. *Nature 352*:497–505.

McClarin JA, Frederick CA, Wang BC, Greene P, Boyer HW, Grable J, Rosenberg JM. 1986. Structure of the DNA-EcoRI endonuclease recognition complex at 3 Å resolution. *Science 234*:1526-1541.

Miller J, McLachlan AD, Klug A. 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J 4*:1601-1614.

Mrabet NT, Van den Broeck A, Van den Brande I, Stanssen P, Laroche Y, Lambeir A, Matthijssens G, Jenkins J, Chiadmi M, van Tilbeurgh H, Rey F, Janin J, Quax WJ, Lasters I, De Maeyer M, Wodak SJ. 1992. Arginine residues as stabilizing elements in proteins. *Biochemistry 31*:2239-2253.

Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Marmorstein RQ, Luisi BF, Sigler PB. 1988. Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature 335*:321-329.

Pavletich NP, Pabo CO. 1991. Zinc finger-DNA recognition:Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science 252*:809-817.

Pavletich NP, Pabo CO. 1993. Crystal structure of a file-finger GL1-DNA complex: New perspectives on file fingers. *Science 261*:1701-1707.

Raumann BE, Rould MA, Pabo CO, Sauer RT. 1994. DNA recognition by β-sheets in the Arc repressor-operator crystal structure. *Nature 367*: 754-757.

Rosenberg JM. 1990. Refinement of EcoRI endonuclease crystal structure: A revised protein chain tracing. *Science 249*:1307-1309.

Schultz SC, Shields GC, Steitz TA. 1991. Crystal structure of a CAP-DNA complex. The DNA is bent by 90°. *Science 253*:1001-1007.

Seeman NC, Rosenberg JM, Rich A. 1976. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA 73*:804-808.

Singh J, Thornton JM. 1992. *Atlas of protein side-chain interactions*. Oxford, UK: Oxford University Press.

Steitz TA. 1993. *Structural studies of protein-nucleic acid interaction*. Cambridge UK: Cambridge University Press.

Taylor R, Kennard O. 1984. Hydrogen-bond geometry in organic crystals. *Acc Chem Res 17*:320-326.

Wolberger C, Dong Y, Ptashne M, Harrison SC. 1988. Structure of a phage 434 Cro/DNA complex. *Nature 335*:789-795.