UNDERSTANDING DNA-PROTEIN HYBRIDIZATION-DEPENDENT INTERACTIONS

By
LAURA O'GRADY

A thesis submitted to the
Graduate School – New Brunswick
Rutgers, the State University of New Jersey
in partial fulfillment of the requirements
for the Degree of
Master of Science

Written under the direction of
Dr. Wilma K. Olson
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey
October 2012

ABSTRACT OF THE THESIS

Understanding DNA-Protein Hybridization-Dependent Interactions

By LAURA O'GRADY

Thesis Director
Dr. Wilma K. Olson

The study of DNA-protein interactions is important since these associations are critical to the operation of living cells. DNA-protein contacts show distinct microenvironments. Hybridization is a notable atomic feature that contributes to the distribution of charge within a protein. In order to study a manageable amount of discreet data, a dataset of 499 non-redundant structures was generated. These structures were chosen from DNA-protein structures of high resolution (> 2.5Å), available in the Protein Data Bank up to June 2011. This research describes differences in DNA-protein interactions based upon atomic hybridization within the set of non-redundant protein structures generated.

Although generalizations can be made in terms of hybridization-dependent DNA-protein interactions, the data in this research show hydrogen bond donor-acceptor relationships, as well as electrostatic (positive-negative) attractions, are the primary features regulating close contacts in the major grooves. In general, atoms with a more electropositive environment show a greater number of close contacts with sp^2 hybridized atoms, and conversely atoms with a more electronegative environment show a greater number close contacts with sp^3 hybridized atoms. However, for proteins comprised of amino acids with delocalized electrons, it is the charge (positive or negative) on the

functional group, as well as the structure (cyclic or aliphatic) that have the greatest impact on the number and type of close contacts.

Acknowledgements

I would like to thank Dr. Wilma Olson for her guidance and support though out this research project. She possesses a wealth of information, and always found time to patiently answer my questions and provide advice. She tirelessly reviewed this research paper, always providing helpful guidance. Her support and direction throughout this research were paramount, and I can not thank her enough.

I would also like to express my sincere thanks to Mauricio Esguerra and Andrew Colasanti, without their help and support this research would not have been possible. Their patient instruction is tremendously appreciated.

Dedication

To my husband Tom, for always being there for me, and my son Rob for making my life worth living.

Table of Contents

Listing of Tables

Listing of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 1
INTRODUCTION

The basic unit of a living entity is the cell. Although most of the mass of a cell is accounted for by water, there is also important dry matter that consists primarily of proteins and nucleotides. A nucleic acid is a polymer composed of nucleotides. Each nucleotide contains a base (purine or pyrimidine), a sugar and a phosphate. There are two kinds or nucleotides: those containing the sugar ribose, known as ribonucleic acids (RNA), and those containing the sugar deoxyribose, known as deoxyribonucleic acid (DNA) [1]. It is well known that DNA and RNA govern the function and development of all living creatures [2]. DNA represents the blueprint of life since it is the carrier of genetic information.

DNA is a long polymer consisting of purine and pyrimidine bases held together by alternating sugar and phosphate residues that create a backbone. DNA exists as a two-stranded double helix. The currently accepted double-helical structure of DNA was first suggested by Watson and Crick [3]. DNA exists in many possible conformations. In crystals DNA exists in three main structural forms; A-DNA, B-DNA and Z-DNA. The shape of A-DNA and B-DNA is known as a right-handed helix. The rails of the helical staircases are formed by anti-parallel sugar-phosphate strands and the rungs by purine-pyrimidine base pairs. B-DNA is found in fully hydrated DNA and it is this form that is thought to be most commonly found in vivo [4]. In B-DNA, there are two grooves on the surface of the helix. In B-DNA, the major groove is on average, approximately 50% wider than the minor groove [5]. Figure 1.1 provides a simplified illustration of the major and minor groove in B-DNA. Although the biological significance and in-vivo occurrence of Z-DNA are controversial, this structure could produce some tensional

strain relief from the B-DNA conformation. Z-DNA is a left-handed helix. Noteworthy, Z-DNA was the first crystal structure of DNA longer than two base pairs to be solved [4].

DNA function is intimately linked to interaction with proteins. In B-DNA, protein-base interactions occur mainly in the major groove where the bases are most accessible [6]. Furthermore, the manner with which proteins interact with DNA is limited. Interaction between proteins and DNA may be dependent on the DNA base sequence or it may be non-specific. Sequence-specific interactions primarily include hydrophobic, hydrophilic and electrostatic interactions. In addition, DNA-protein interactions are controlled by the bonding abilities of the chemical groups on the protein surfaces. The geometric structure of the DNA is an important factor in sequence recognition by proteins. Even non-specific interactions require contacts with complementary atoms. Therefore, interaction between proteins and DNA requires harmonized interaction between the chemical groups, as well as complementary three-dimensional structures.

Proteins are macromolecules composed of amino acids. They are formed when amino acids bind together by peptide bonds. A peptide bond forms between the carboxyl and amino groups on adjacent amino-acid residues. Figure 1.2 illustrates the formation of a peptide bond. The backbone of the polypeptide is given by the repeated sequence of three atoms in the chain: the amide nitrogen, the alpha carbon, and the carbonyl carbon. Rotations in the chain take place about the bonds in the backbone. The peptide bond is usually inflexible. The side chains of the polypeptide are noted by R_n in Figure 1.2.

The primary structure of proteins consists of the sequence of residues in the polypeptide chain. Secondary structures arise from interactions between atoms within

the polypeptide chain, mainly hydrogen bonds between backbone atoms. Tertiary structures involve the packing of secondary structures with each other. Motifs and domains are combinations of secondary structures.

Figure 1. 1: Basic illustration of major and minor grooves of B-DNA

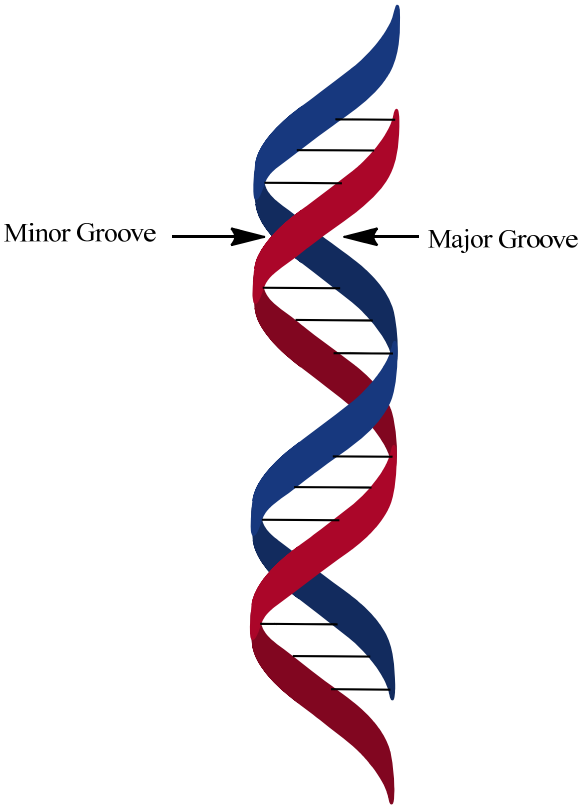
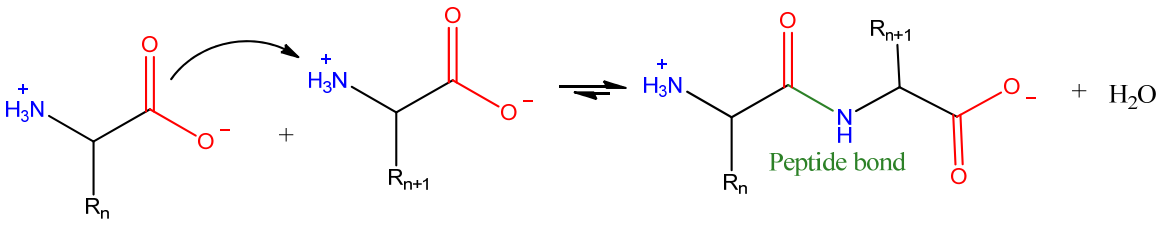


Figure 1. 2: Formation of a peptide bond



1.1 DNA-Protein Interaction

As previously stated, DNA-protein interactions can be non-specific, or the binding can be specific to a particular DNA base sequence. In addition, interactions can occur between external atoms and atoms in the backbone or the bases of DNA. Although a large number of contacts arises between proteins and the DNA backbone, particularly the phosphates, presumably as a means of stabilization, the interaction between proteins and the DNA bases are also extremely important for biological function.

Non-specific interactions are typically ionic in nature. These ionic interactions typically occur as salt bridges between protein side chains and the DNA backbone. There are also DNA sequence-specific interactions. These interactions have been shown to occur mostly through van der Waals contacts and hydrogen bonding [7].

For B-DNA, the major groove provides a much better environment for sequence-specific interaction than the minor groove for two reasons. First, the major groove is wider and therefore offers more accessible surface area for proteins to interact with. Second, the pattern of possible hydrogen bonds that can be formed with the edges of the base-pairs is more discriminatory in the major groove than in the minor groove [8].

Although this research will focus on protein-DNA interactions in the major groove, minor-groove interactions will not be ignored, since this is a source for many non-specific interactions. As noted in thesis work performed by Yun Li [9], there are ligands that are able to recognize the DNA minor groove without degeneracy. Dervan *et al.* [10, 11] prepared polyamides that formed four-ring pairs able to recognize G-C, C-G, T-A and A-T base pairs with specificity in the minor groove. Therefore, this research will also secondarily investigate hybridization effects of interaction in the minor groove.

1.1.1 DNA-Protein Structural Discrimination

Some of the most important electrostatic interactions between proteins and DNA are believed to be hydrogen bonds. Seeman *et. al*. suggested that two hydrogen bonds uniquely identify recognition between amino-acid side chains and DNA base pairs [8]. Furthermore, the authors postulated that recognition in the major groove is far more sensitive to base-pair sequence discrimination than that in the minor groove [8]. In addition to direct hydrogen bonding, the ability of amino acids to interact with DNA also depends upon hydrophobic interactions as well as DNA and protein structural parameters. Thus there are two generally recognized methods for DNA sequence recognition. These include direct interactions via hydrogen bonding, and indirect recognition through sequence-dependent structural features and mediation by water molecules.

Seeman *et. al*. describe a scheme of hydrogen bonding in which the bonds form in the base-pair plane and the amino-acid structure harmonizes with the stacking of base pairs within the DNA secondary structure [8]. It should be noted that there is significant flexibility in the base pairs and the sugar-phosphate backbone.

Proteins recognize specific DNA sequences through DNA binding domains within their polypeptide chains. These domains have particular structural features that allow the protein to recognize the specific DNA sequences and orient properly. Studies have shown a limited number of protein structures that bind to specific DNA sequences [12]. Two of the more common structural motifs include the helix-turn-helix motif and the zinc finger motif.

1.1.2 Electrostatic Hydrophobic and Hydrophilic Interactions

Proteins are composed of chains of amino acids linked by peptide bonds. Although all proteins are composed of the same 20 amino acids, they are structurally diverse due to the varying nature of the amino-acid side chains. Amino acids are classified in many different ways, for example by polarity, structure or biochemical properties. Polarity is the most commonly used method for classification, and it contributes substantially to protein interactions. Furthermore, amino-acid atoms are classified as positively charged, negatively charged, hydrogen-bond donors or acceptors and hydrophobic.

Hydrogen bonding is believed to be among the most important type of hydrophilic interaction between DNA and a ligand. Hydrophobic interactions are typically formed between non-polar R groups on aliphatic or aromatic side chains of certain amino acids. These non-polar groups tend to aggregate between themselves and away from more polar or charged surfaces.

1.1.3 DNA Electrostatic Attraction and Hydrogen-Bonding Sites

As noted, the most important type of electrostatic attraction is hydrogen bonding. Seeman *et al.* compared the hydrogen-bonding sites of all four Watson-Crick base pairs (A-T, T-A, G-C, and C-G) [8]. The Watson-Crick base pairs with the hydrogen bonds associated with base pair atoms are illustrated in Figure 1.3. The hydrogen bonds between base pairs are denoted by the dashed lines in this figure. For the purpose of this discussion, the naming convention for the commonly recognized hydrogen-bonding sites on the free edge of these base pairs, utilized by Yun Li in his PhD dissertation, will be adopted [9]. This naming convention for the hydrogen-bonding sites along with the IUPAC numbering of base atoms is illustrated in Figure 1.4. In addition to potential hydrogen bonding with

protein groups, there is important hydrogen bonding between the base pairs that accounts for the double helix structure of DNA.

In the major groove of the DNA helix the potential hydrogen-bonding sites are denoted with the letter W (wide). Refer to Figure 1.4 for an illustration of these hydrogen-bonding sites. The W1 site contains the imidazole N7 atom on a purine base or the C5 atom on a pyrimidine base. The W2 site contains the exocyclic group attached to the C6 atom of a purine or the C4 atom on a pyrimidine base. In the minor groove, hydrogen-bonding sites are denoted with the letter S (small). The S1 site corresponds to the N3 atom on a purine base or the carbonyl oxygen on the C2 atom of a pyrimidine. The S2 site corresponds to the exocyclic group at the C2 position on a purine. For adenine the S2 site is occupied by an H atom and for guanine it is occupied by an NH_2 group. Each W and S site on a base will have a complementary site denoted with a prime on the corresponding base that forms the pair. Although this convention was formulated for hydrogen bonding, it will be utilized in the general study of interactions undertaken in this research. For the purpose of discussion in this paper, Figure 1.4 notes major-groove purines as W1 and W2, and pyrimidines as W1′ and W2′. In the minor groove, purine binding sites are denoted with an S1 and S2, and pyrimidine binding sites with an S1′.

Figure 1. 3: Purine and pyrimidine base pair structures

The orientation of the figures bellow shows x as the vertical axis, y as the horizontal axis and z is perpendicular to the plane of the page
Dashed lines represent hydrogen bonds

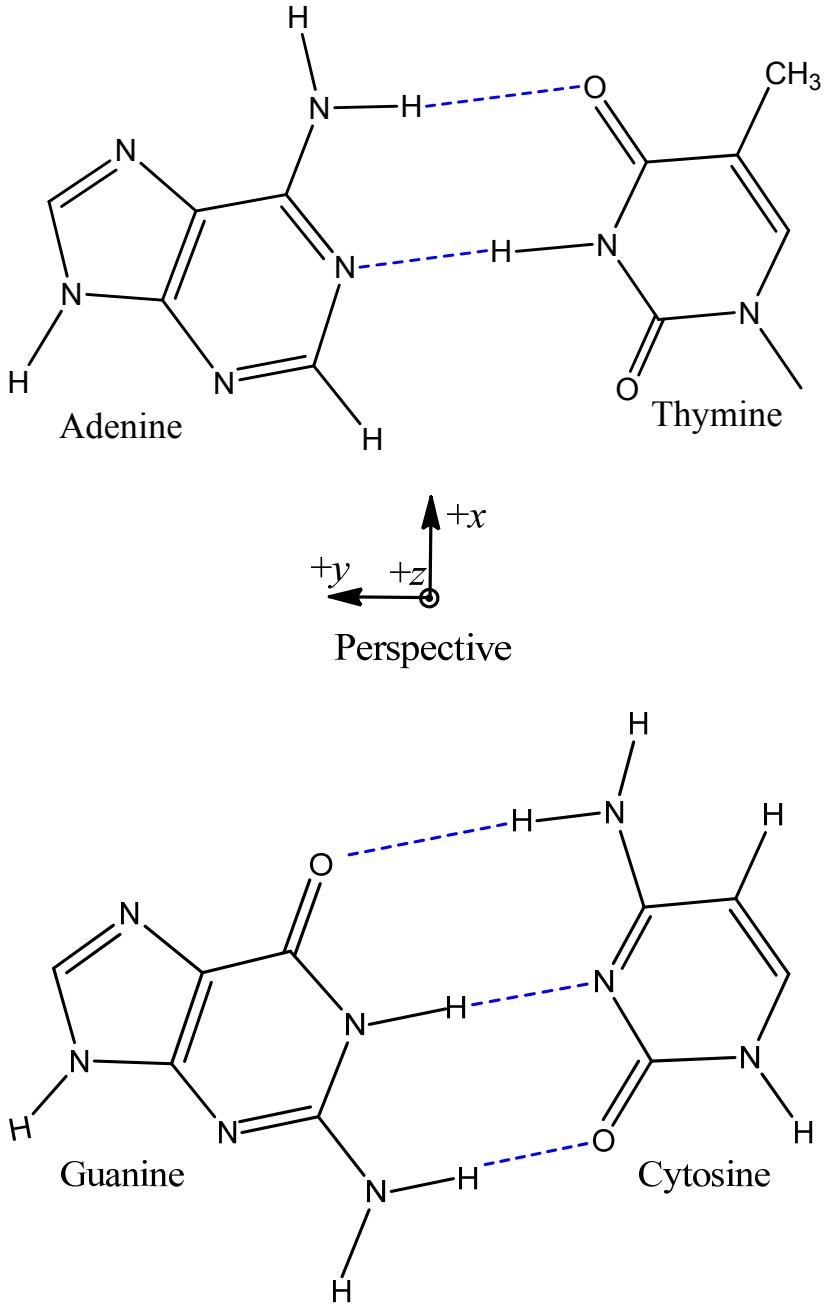
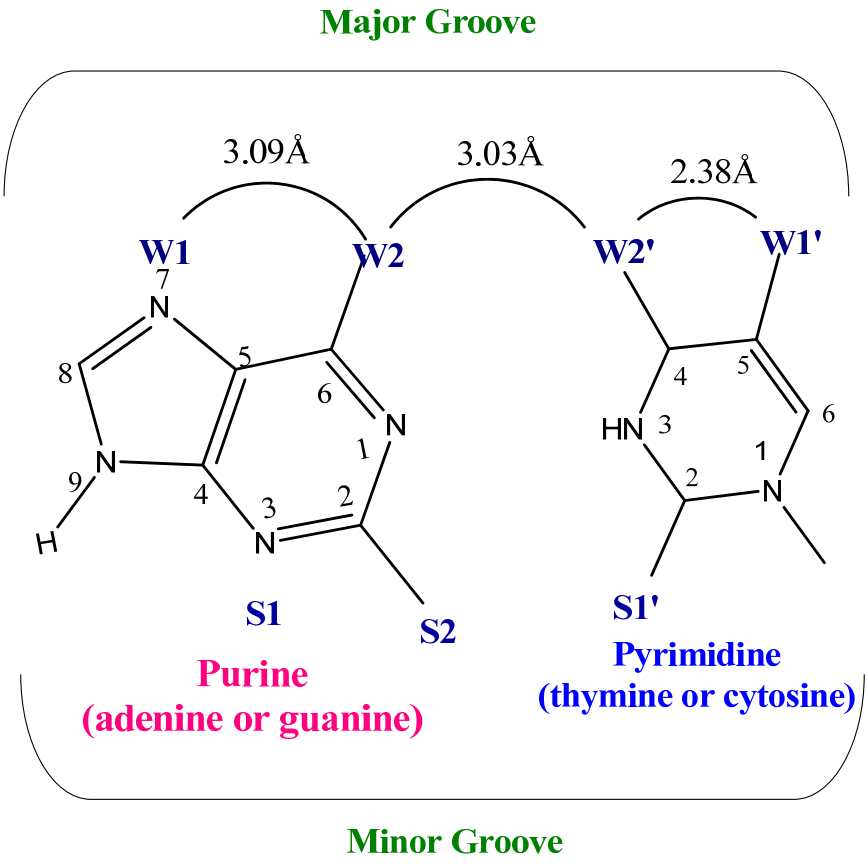


Figure 1. 4: IUPAC numbering and hydrogen bonding for generic base pair



Legend: W = Major groove binding sites
S = Minor groove binding sites

1.1.3.1 DNA Major-groove Hydrogen-bonding Sites

Both sterically and electrostatically, major-groove hydrogen-bonding sites (W1, W1′ and W2, W2′) are quite different for each of the Watson-Crick bases (adenine, guanine, cytosine and thymine). In the major groove, the W1′ site for pyrimidines (thymine, cytosine) has non-polar atoms in this location (C-Me, C-H respectively), whereas the corresponding W1 site on purines (adenine, guanine) has a polar imidazole nitrogen atom (N7). In addition, the W1 site of the purine adenine has an adjacent NH_2 group, but the W1 site of guanine has an adjacent carbonyl. It is noteworthy that the W1 site on guanine is surrounded by a more negatively charged environment than the W1 site of adenine. For guanine there is an imidazole nitrogen in the W1 position and a carbonyl in the W2 position. However, for adenine there is an imidazole nitrogen in the W1 position, but an NH_2 in the W2 position. The W1′ sites on the pyrimidines, thymine and cytosine, are very different sterically. Thymine has a bulky methyl group attached to the C5 atom, whereas cytosine has a small hydrogen atom.

The W2 site (or complimentary W2′) on each base is occupied by either a carbonyl or an amine group. For the purines, adenine has an NH_2 group in the W2 position, whereas guanine has the W2 position occupied by the carbonyl O. For the pyrimidines, thymine has the carbonyl in the W2′ site, whereas cytosine has the NH_2 group in this position. Furthermore, it should be noted that the carbonyl is considered a hydrogen bond acceptor and the NH_2 a hydrogen bond donor. In terms of size, the carbonyl group is also very different from the bulky NH_2.

As described in the preceding paragraphs, amino-acid atoms within proteins are sterically and electrostatically different. Therefore, interactions between the DNA bases

and the amino acids on proteins are governed by the steric and electrostatic environments on both bodies. According to Seeman's model, it is postulated that arginine (ARG) recognizes guanine in the major groove and asparagine (ASN) and glutamine (GLN) recognize adenine in the major groove [8]. However, other studies have shown large numbers of other types of interactions. Mandel-Gutfreund *et al*. found that lysine (LYS) interacts with guanine and aspartic acid (ASP) and glutamine acid (GLU) interact with cytosine to a large extent in the major groove [13]. These additional interactions cannot be completely accounted for by hydrogen bonding and electrostatic attraction between DNA base atoms and protein atoms, and this finding supports the idea that recognition in the major groove depends on structural as well as electrochemical compatibility of the species.

1.1.3.2 DNA Minor-groove Hydrogen-bonding Sites

Both sterically and electrostatically, minor-groove hydrogen-bonding sites (S1 and S2) are also quite different. In the minor groove, the S1 sites for the purine adenine as well as the purine guanine both contain the N3 atom. The S2 site, however, differs between these purine bases. Adenine has an H atom at the S2 location and guanine an NH_2. Sterically these two positions are different. In addition, the S2 position of guanine is occupied by a hydrogen bond donor. In fact, the NH_2 group forms an H-bond with its corresponding base pair partner, cytosine. In contrast, there is no corresponding S2′ position of the pyrimidines, thymine and cytosine. Furthermore, for these two pyrimidines the S1' position is associated with the same functional group. Both have a carbonyl at this position. However, it should be noted that for cytosine the carbonyl is in association with its Watson-Crick base-pair partner through H-bonding, but for thymine there is no H-bonding with its Watson-Crick partner at this location.

1.2 DNA-Protein Structural Recognition

DNA is a double helix in which two polynucleotide strands are wrapped around each other. These two strands are complementary and anti-parallel. In vivo, DNA does not exist as a free acid, but forms a complex with a variety of proteins. For many structures, the objective of a DNA-protein interaction is to achieve the most favorable conformational and electrostatic arrangement.

Within the DNA anti-parallel strands there is a great deal of flexibility. According to Calladine and Drew, the interaction of base-pair dimers as they stack upon each other is a major factor governing the contour of DNA [14]. Base pairs stack upon each other in order to optimize favorable internal interactions and avoid unfavorable ones (both steric and electrostatic). Furthermore, the edges of DNA base pairs are important points of close contact with solvent and protein molecules. The 3DNA software package developed by the Olson group provides a standard methodology for mathematically calculating the parameters needed to describe the DNA structure. The 3DNA software provides an accurate description of a DNA base-pair step defined by 18 rigid-body parameters [15].

1.3 Polarity and Hybridization

Polarity arises from unequal charge density. In the case of atoms, it results from the unequal distribution of electrons surrounding the positively charge nucleus. In the case of atomic bonds, charge density is determined by the nature of the bond. An extreme case of polarity is illustrated in an ionic bond. However, covalent bonds show varying degrees of polarity. Polarity in a molecule results from having atoms with different electronegativities.

In the Molecular Orbital Theory, the atomic orbitals of bonding atoms combine to form molecular orbitals [16]. It is within these orbitals that the electrons move about the molecule as a whole, and therefore are an important factor in the distribution of charge in a molecule. In addition, in order to explain the bonding that occurs, particularly among carbon, nitrogen and oxygen atoms, and to account for molecular geometry, the hybridization theory was introduced by Linus Pauling [17].

Hybridization refers to the internal linear combination of atomic orbitals. When the s and p orbitals of an atom combine, their individual wave functions combine to form a new hybrid wave function. For a carbon atom to form four identical single bonds, one s orbital must combine with the three p orbitals to form four new hybridized sp^3 orbitals. Single bonds on carbon atoms arise from sp^3 hybridization. When one s orbital and two p orbitals combine, three new hybridized sp^2 orbitals are formed. As a result of this hybridization, one p orbital on the atom remains unhybridized. Thus, an sp^2 hybridized orbital and an unhybridized p orbital from each of the two bonded atoms characterize a double bond. The double bond arises from one sigma bond (arising from the overlap of two sp^2 hybridized orbitals from the bonded atoms) and one pi bond (resulting from the overlap of two unhybridized p orbitals from the bonded atoms). Figure 1.5 illustrates the orbitals in a double bond.

Hybridized orbitals are strongly directional, and as such are oriented in specific spatial directions around the nucleus. As previously illustrated, the number bonds or attachments (non-bonded electron pairs) to a particular atom determine hybridization. Hybridization is often manifested in the geometry of the molecule. Table 1.1 describes the geometry and bond angles of hybridized species.

Table 1. 1: Geometry of Hybridized Species

Attachments	Hybridization	Geometry	Angles
6	sp^3d^2	Octahedral	90°, 180°
5	sp^3d	Trigonal bipyramidal	90°, 120°, 180°
4	sp^3	Tetrahedral	109.5°
3	sp^2	Trigonal planar	120°
2	sp	Linear	180°

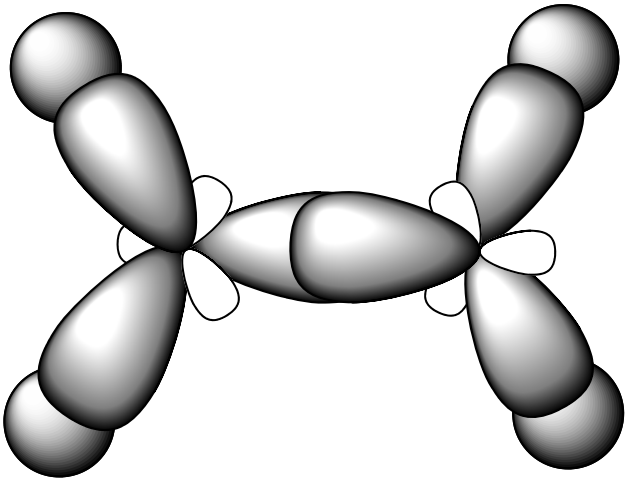
Source: http://chemistry.boisestate.edu/people/richardbanks/inorganic/bonding%20and%20hybridization/bonding_hybridization.htm

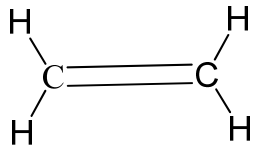
Aromatic compounds are cyclic compounds that contain a ring of alternating single and double bonds. Since these rings are regular in shape, and the atoms in the rings are at equal distance, it would indicate that these bonds are of a bond order between single and double, rather than strictly alternating. Aromatic rings of this type are therefore described as conjugated, with a bond order of 1.5 and pi electrons delocalized around the ring. Each atom contains three sp^2 hybridized orbitals; however the unhybridized p orbital perpendicular to the plane of the ring contains a delocalized electron, rather than one localized in a double bond. Figure 1.6 illustrates the resonance structures of a six sided aromatic ring such as benzene. Figure 1.7 illustrates the delocalized structure of the six-sided aromatic ring. Other hydrocarbons can also form conjugated rings. Aromatic hydrocarbons that consist of conjugated rings are known as annulenes. Conjugation can also be extended to charged species such as cyclopentadienes and heteroaromatic specifies such as pyridine. Figure 1.8 provides an example of additional types of conjugated species.

There are a few amino-acid side chains that contain aromatic rings with conjugated atoms. They include phenylalanine, tyrosine, histidine and tryptophan. Proteins containing these side chains will be studied in the context of sp^2 hybridized atoms, since these atoms do contain sp^2 hybridized orbitals, but the conjugated nature of the rings will also be investigated separately.

Figure 1. 5: Hybridized orbitals in a double bond

Each of the two carbon atoms in the double bond has three sp^2 hybridized orbitals. The chemical structure and orbital diagram for ethene are diplayed below.





Each carbon in the double bond also has an unhybridized p orbital that contributes to a pi bond.

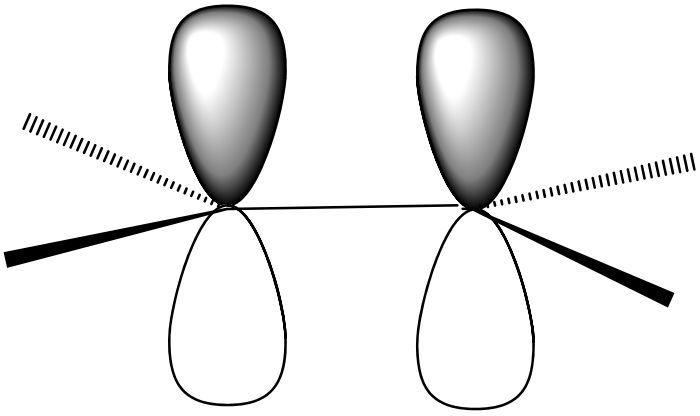


Figure 1. 6: Resonance structure of 6-annulene

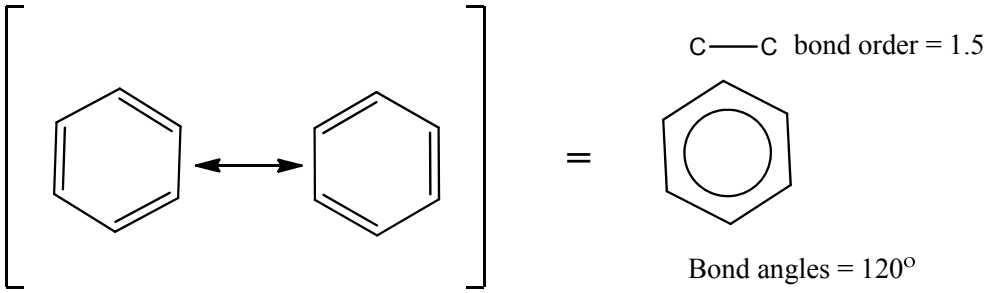


Figure 1. 7: Conjugated double bonds in 6-annulene

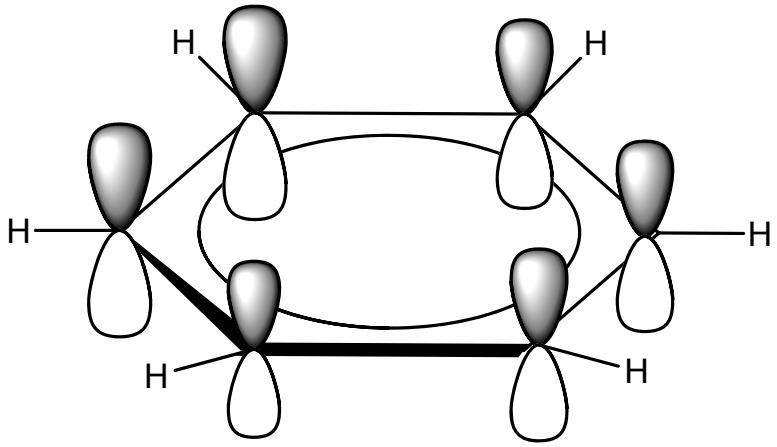
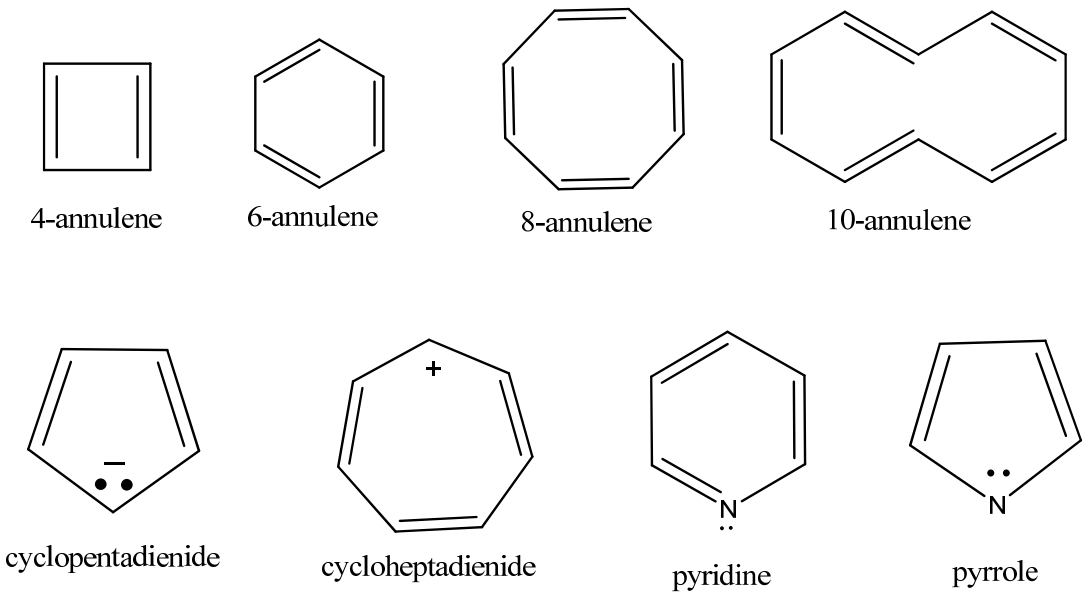


Figure 1. 8: Conjugated species



1.4 DNA-Protein Interactions and Hybridization

A variety of protein-base contacts in the major and minor grooves of DNA were studied by Yun Li in his PhD Dissertation [9]. The interactions studied included those between the Watson-Crick bases in DNA and the commonly classified amino-acid atom types (positive, negative, hydrogen-bond donors, hydrogen-bond acceptors and hydrophobic). However, interactions based upon amino-acid atom hybridization were not investigated. These interactions are therefore the topic of this research.

1.5 Amino-acid Structure, Function and Hybridization Properties

1.5.1 Structure and Function

As previously described, proteins are macromolecules composed of amino acids. Hence the structure and function of these proteins that interact with the bases in DNA are linked to structures of amino acids from which they are composed.

All amino acids have at least one amine group and one acid group. Although all amino acids have these common structural features, their properties differ mainly as a result of different chemical side chains, or "R" groups. The common amine/acid portion of the structure is generally referred to as the backbone, and the "R" group as the side chain. The interactions between proteins and DNA components are largely the result of structural properties of the two species. Therefore an understanding of the structural properties of amino-acid protein components is critical to understanding the types of interactions they are involved in. The structures of the 20 common amino acids are illustrated in Figure 1.9.

In terms of structurally related properties, amino-acid side chains may be polar or non-polar. The polarity depends on the side-chain groups determined by the recognized

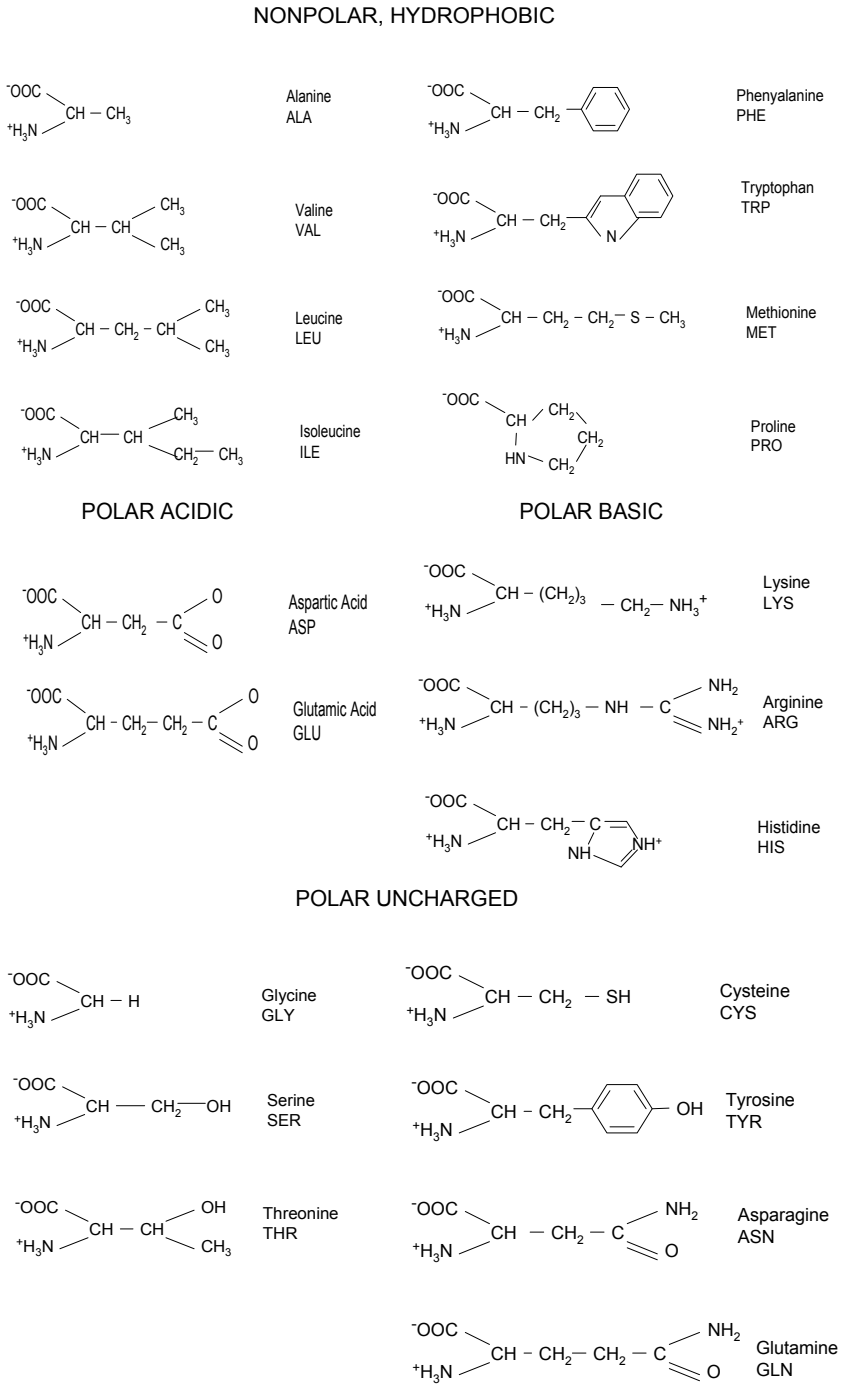
polarity rankings (Amide > Acid > Alcohol > Amine > Ether > Alkane). In addition, side chains may be acidic or basic. The acidity depends on the net effect of all acidic and basic units.

Size is another feature that should be considered when reviewing protein and base interactions. The amino acid and base atoms interact in the major or minor grooves in a way that minimizes the free energy of the system, and is dependent on both molecular and conformational considerations. Therefore, proteins with bulky amino-acid side chains will interact differently than those with smaller ones.

Moreover, the conformation of the polypeptide chain composing a protein is a very important feature in the interaction with the nucleotide bases of DNA. These polypeptide chains are not rigid bodies, but rather can be viewed as rigid peptide units that rotate about each other with two degrees of torsional freedom. Each rigid peptide unit contains a carbon atom C_A attached a side chain (R), and a carbon atom C attached to an oxygen as part of the carbonyl. In order to illustrate these degrees of torsional freedom, the chain is typically divided into peptide units that span one C_A to the next C_A. Each unit can rotate around two bonds: the C_A-C bond (ψ) and the N-C_A bond (φ). Figure 1.10 illustrates the peptide chain and degrees of freedom about the main chain. Polypeptides adopt conformations based upon allowed combinations of ψ and φ angles. It should be noted that many combinations are not allowed due to steric collisions. Although there are only two degrees of freedom in the main chain of polypeptides, side chains longer than alanine can in themselves adopt different conformations due to rotations about the side-chain atoms.

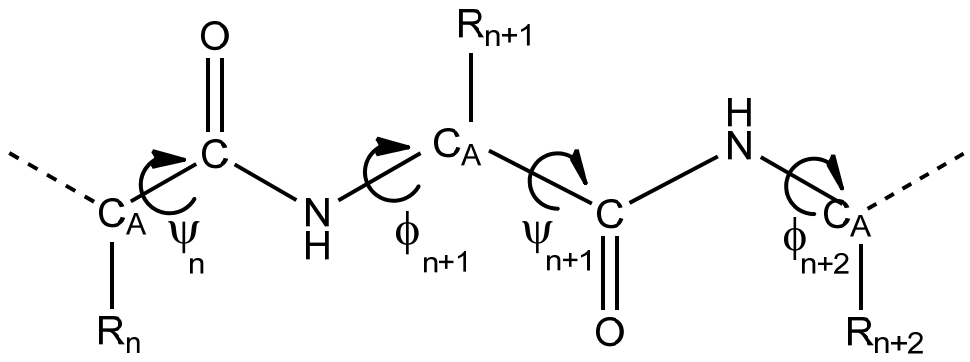
Figure 1. 9: Representation of 20 common amino acids

These are categorized according to structurally related side chain properties



Source: http://biotech.matcmadison.edu/resources/proteins/labManual/chapter_2.htm

Figure 1.10: Degrees of freedom in a polypeptide



1.5.2 Hybridization Features in Proteins

Proteins which are composed of amino acids also have a varying number of double and single bonds within their structures. Hybridization within the protein will affect the conformation of the molecule as well as the charge distribution. The features of the hybridized bond were presented in Section 1.3 and are illustrated in Figure 1.5.

Amino acids contain sp^2 and sp^3 hybridized atoms in varying proportions in their backbone and side chains. Figures 1.11, 1.12 and 1.13 illustrate the 20 common amino acids and display the hybridization assignments (sp^2 or sp^3) for the amino-acid atoms within the database utilized for this research. Each amino acid has at least one double bond (two sp^2 hybridized atoms) as a result of the carbonyl in the backbone. Therefore sp^2 hybridized atoms are common in the backbone of amino acids. Proteins synthesized when amino acids combine through a peptide bond also contain sp^2 hybridized atoms in the polypeptide chain. The sp^2 hybridized double bond atoms in the polypeptide chain are shown in Figure 1.15.

Although the C - N bond in a polypeptide chain appears to be a single bond, with an additional lone pair of electrons on the nitrogen, there is also partial double bond nature to the C - N bond [18]. The peptide bond therefore contains delocalized electrons. This delocalization is consistent with the geometry of the peptide bond in a polypeptide. If the peptide bond were purely a single bond, it should be sp^3 hybridized, and therefore exhibit a tetrahedral arrangement. However, in a peptide the entire arrangement seems to flatten out to a planar arrangement. In 1951, Linus Pauling described the helical configuration of polypeptide chains in proteins by postulating the planar structure of the peptide bond based on a study of structural factors for two helical configurations [19].

The planar arrangement (bond angles closer to 120°) in the peptide bond is more indicative of trigonal planar structure (three sp^2 orbitals plus a p orbital), as described in Table 1.1. Furthermore, quantum mechanical calculations support the overlap of electrons between the N, C and carbonyl O atoms. Figure 1.15 illustrates the peptide bond with the p orbital overlap. Because only one assignment per atom can easily be made in the database, the backbone nitrogen (N) has been assigned a hybridization of sp^2.

As shown in Figures 1.11, 1.12 and 1.13, among the 20 amino acids, four have sp^2 hybridized atoms in an aromatic side chain (R). These amino acids are phenylalanine, tryptophan, histidine and tyrosine. Phenylalanine and tryptophan are non-polar hydrophobic, histidine is polar basic and tyrosine is polar uncharged. The aromatic side chains of these four amino acids contain delocalized electrons. Although the nitrogen (N_{E1}) atom in tryptophan has 3 bonds and a lone electron pair, it is part of the delocalized ring structure with a planar arrangement as shown when viewing the structure in the Jmol viewer [20]. Therefore the N_{E1} atom of tryptophan has been assigned sp^2 hybridization, along with all of the carbon atoms in the double ring structure. Histidine also has a similar atomic arrangement. Both nitrogen atoms have been assigned a hybridization of sp^2 due to the delocalized electrons in the ring.

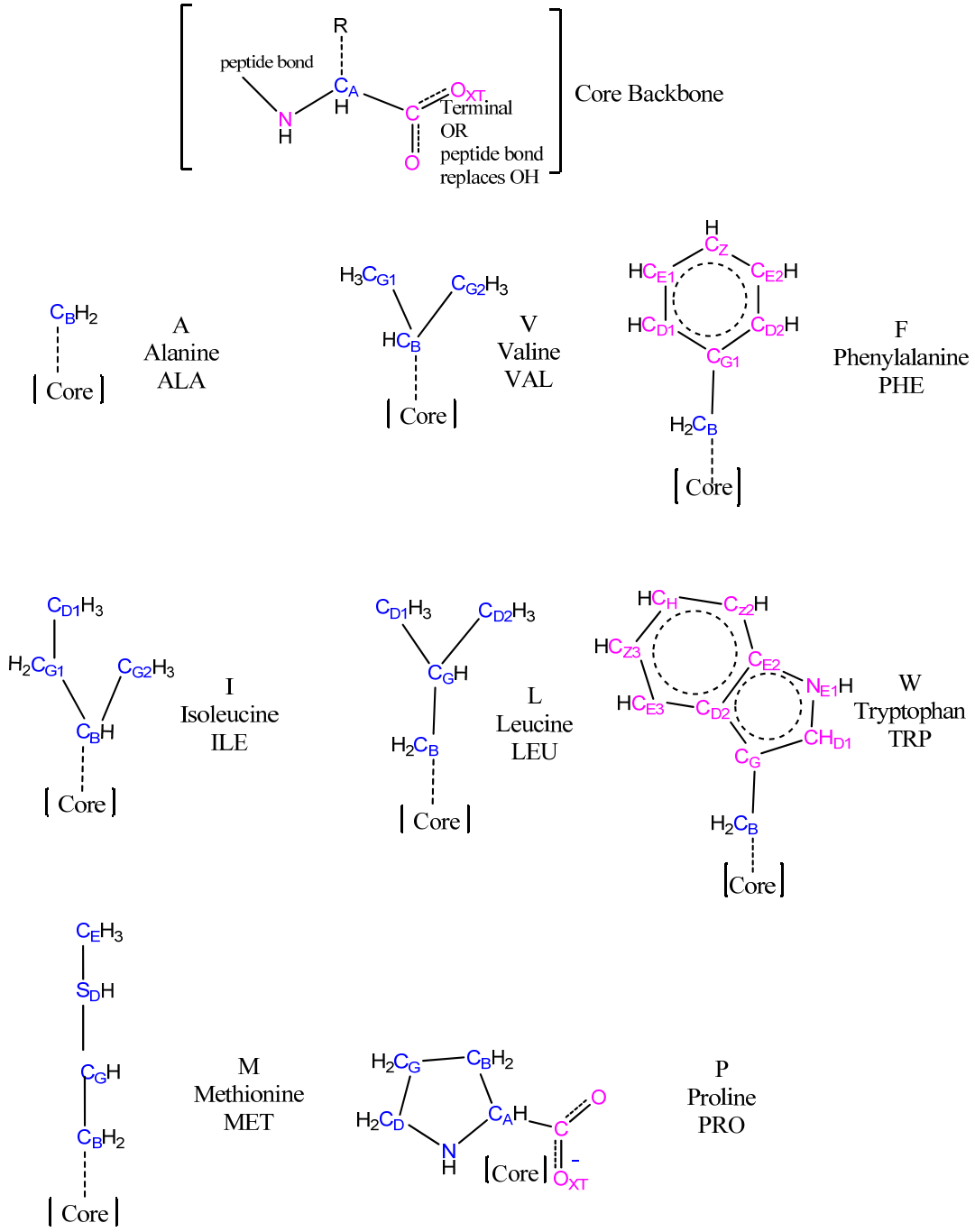
Aspartic and glutamic acids also contain sp^2 hybridized atoms in their side chains. They are acidic amino acids that contain a carbonyl in their side chains, and therefore sp^2 hybridized atoms in their side chains. For these acids, both oxygen atoms have been assigned a hybridization of sp^2 since it is hypothesized that there is delocalization of charge and is impossible to determine the extent of sp^2 versus sp^3 hybridization. They are

therefore both treated as delocalized sp^2 hybrids. Refer to Figure 1.13 for the hybridization descriptions of aspartic and glutamic acid.

Arginine is the most basic amino acid, and exists in its cationic form with a positively charged guanidinium group in neutral, acidic and basic environments. An important feature of arginine is the delocalization of electrons present in the guanidinium group. Figure 1.13 shows the delocalization of electrons in the guanidinium group, and Figure 1.14 shows the resonance forms of arginine. Since the nitrogen atoms on the side chain can not be distinguished from each other, they are both treated as sp^2 hybridized with delocalization of charge. As a result of its structure and the presence of the positively charged guanidinium group, arginine acts a very good hydrogen bond donor. Arginine can actually donate several hydrogen bonds.

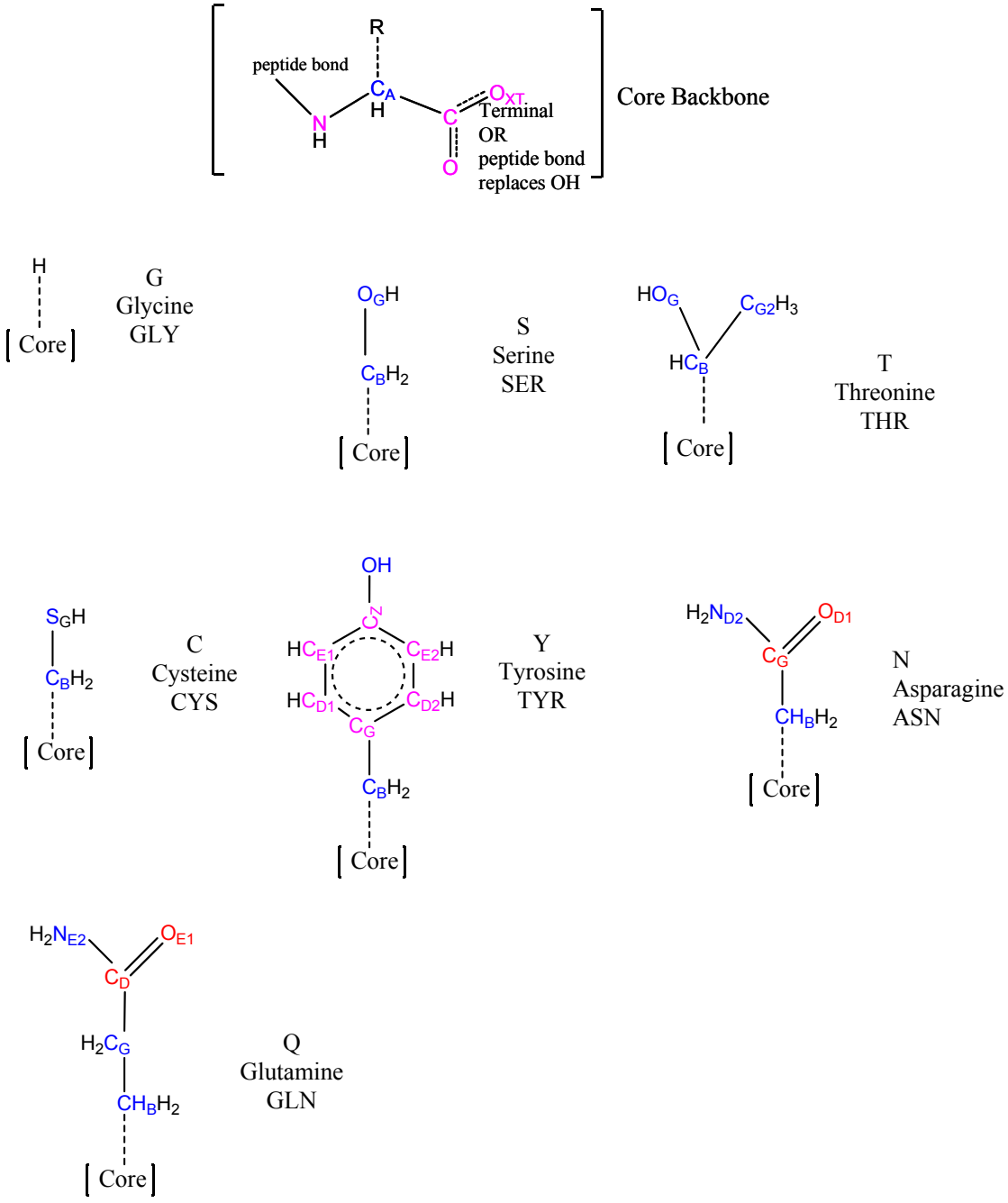
As shown in Figure 1.12, of the eight polar uncharged amino acids, asparagine and glutamine are amides and tyrosine has an aromatic ring containing sp^2 hybridized atoms as previously described. Asparagine and glutamine have potential of delocalized electrons as a result of the amide side chain. However, as a result of inspection of specific interactions, as described in Section 3 of this research paper, an assignment of sp^3 was made for the NH_2 atoms for both amino acids.

Figure 1.11: Description of hybridization in nonpolar hydrophobic amino acids



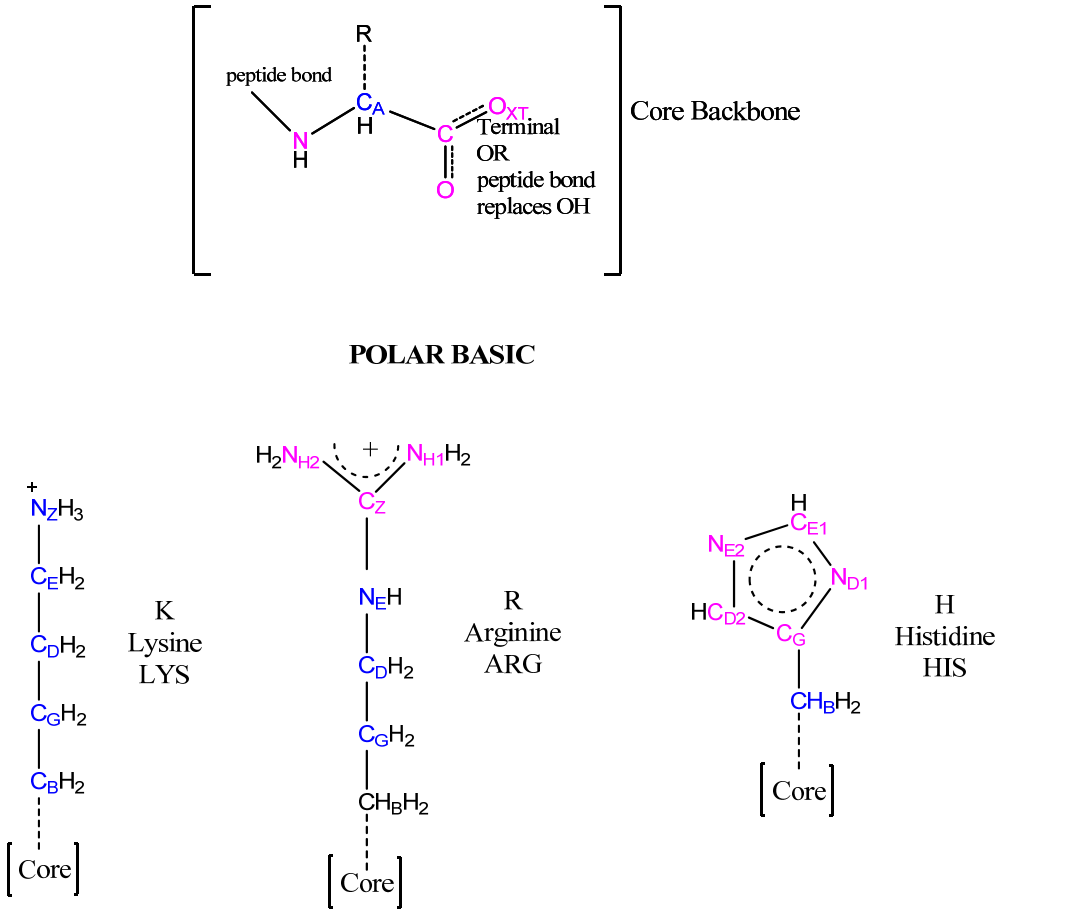
Legend: Red = sp^2 hybridized atom Blue = sp^3 hybridized atom
Fuschia = sp^2 with delocalized p electrons

Figure 1.12: Description of hybridization in polar uncharged amino acids



Legend: Red = sp^2 hybridized atom Blue = sp^3 hybridized atom
Fuschia = sp^2 with delocalized p electrons

Figure 1.13: Description of hybridization in polar basic and polar acidic amino acids



Legend: Red = sp^2 hybridized atom Blue = sp^3 hybridized atom
Fuschia = sp^2 with delocalized p electrons

Figure 1.14: Resonance structures of arginine

These structures show the guanidinium functional group with its positive charge delocalized among the nitrogen atoms

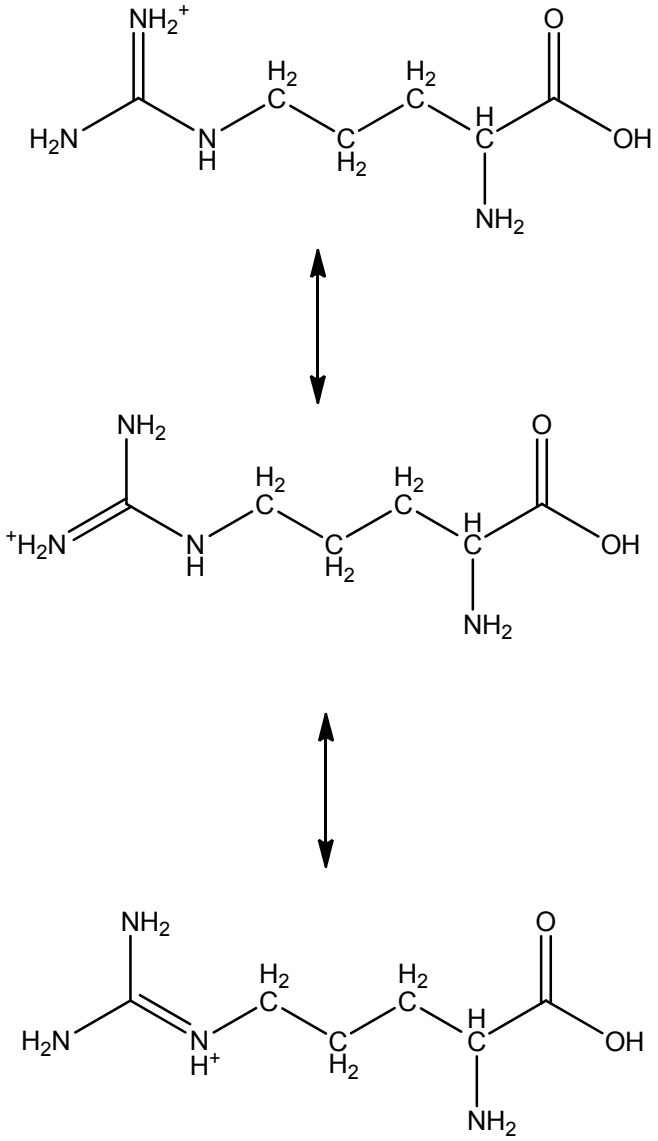
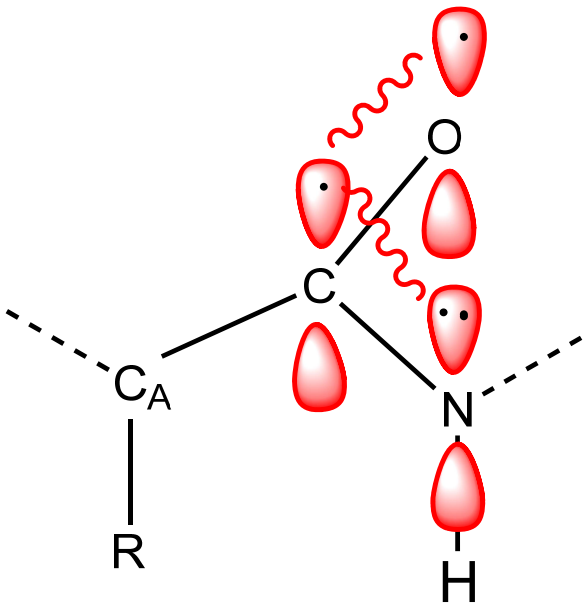


Figure 1.15: Orbital diagram of peptide bond



1.6 Protein Structure

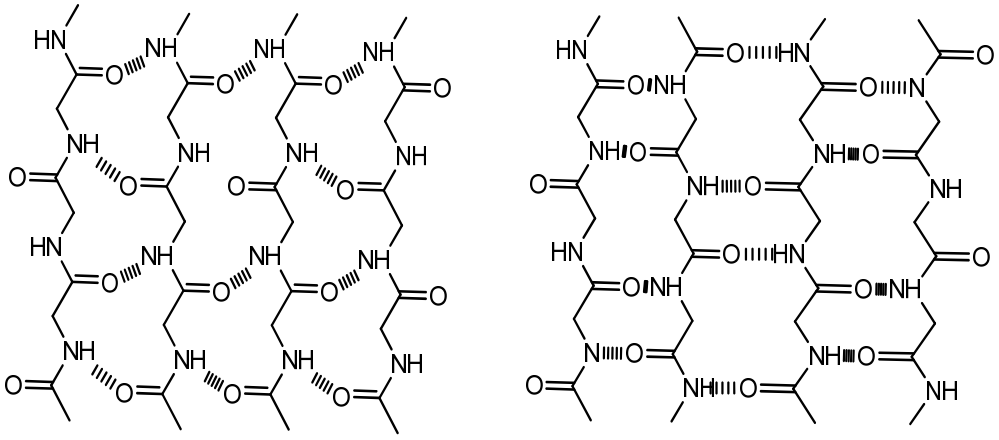
As described in Section 1.5, proteins are composed of amino acids connected by peptide bonds. The polypeptide chain folds into a three-dimensional structure, known as a conformation. The conformation of a protein is determined by non-covalent interactions between the groups. Three common types of non-covalent interactions are hydrogen bonds, hydrophobic interactions and electrostatic interactions [21].

Secondary protein structures arise when a polypeptide chain folds as a result of non-covalent interactions. Two types of stable secondary structures are alpha helices and beta sheets. The alpha helix is the classic protein structure, and is built from a continuous strand of amino acids. The right-handed alpha helix is one of the most frequently encountered secondary structures in proteins. Variations of the right-handed alpha helix, as well as left handed helices, are encountered less frequently. A simplified illustration of a right-handed alpha helix is presented in Figure 1.16. Beta sheets are most common in globular proteins. Beta sheets are built from several regions of a polypeptide chain. There are two possible ways beta strands interact to form a pleated sheet. The strands can run in the same direction, where the amine groups are aligned with amine groups on the adjacent strands and the carboxyl groups are also aligned with carboxyl groups on adjacent strands. This is referred to as a parallel sheet. The strands can also run in alternating directions with amino terminals on one strand aligned with carboxyl terminals on the adjacent strands. This is referred to as an anti-parallel arrangement. Figure 1.17 shows a parallel and an anti-parallel arrangement of beta strands in a beta sheet side by side.

Most proteins are made up from combinations of alpha helices and beta sheets connected by loop regions. Loop regions that connect two adjacent anti-parallel beta strands are called hairpin loops. Motifs are combinations of secondary structures [22]. For example, the simplest motif with a specific function consists of two alpha helices joined by a loop. One such motif is the helix-loop-helix motif. Tertiary structures arise when motifs combine to form globular structures. These tertiary structures are known as domains. When several polypeptide chains from different proteins interact in a specific way, this is known as a quaternary structure.

Figure 1.16: Right-handed alpha helix

Figure 1.17: Beta-sheet diagrams



CHAPTER 2
METHODS

This investigation builds upon research performed by Yun Li, a PhD graduate from the Olson group [9]. It will look at how the hybridization of amino-acid atoms within proteins affects the interaction of these atoms with specific atoms in the bases of a Watson-Crick DNA base pair. This research also expands the relational database initially established by Yun Li. This database contains DNA-protein structures extracted from the Protein Data Bank (PDB) [23]. The established database was updated with new data available in the PDB up to 15-Jun-2011. The revised relational database referred to as NAPID (Nucleic Acid-Protein Interaction Database) contains tables that describe molecular contacts in DNA-protein complexes.

The data used in the analyses contained in this paper are based on a set of non-redundant structures derived from data contained in the PDB. The set of non-redundant structures initially identified by Yun Li in his dissertation, dated February 2006, has been expanded upon using the updated data available in the PDB and extracted into the NAPID database up to 15-Jun-2011.

Furthermore, the files in the Nucleic Acid Database are designated as PDB files since they follow the standard format of the PDB. This section will briefly describe the NAPID structure and function, as well as the methods utilized for the establishment of a new database specifically for the study of hybridization effects on the DNA-protein interaction.

2.1 Overview of NAPID

Within NAPID, the tables allocate the data in terms of entities that may be either an actual object or a relationship such as a molecular contact. The entities belong to categories, and entities in the same category have similar attributes. The database contains the following categories: atom, residue, chemical group, chain, complex, secondary structural unit, contact. The primary categories under review in this research are the atom, chemical group, residue and contact. There are four tables in the atom category: nucleic acid atoms, protein atoms, ligand atoms and water atoms. There are three tables in the residue category: nucleotides, amino acids and ligands. There are also three tables in the chemical group category: sugars, phosphates and bases. In the contacts groups, these contacts can be between nucleic acid atoms and atoms of proteins, ligands or water molecules.

2.1.1 Atomic Contacts

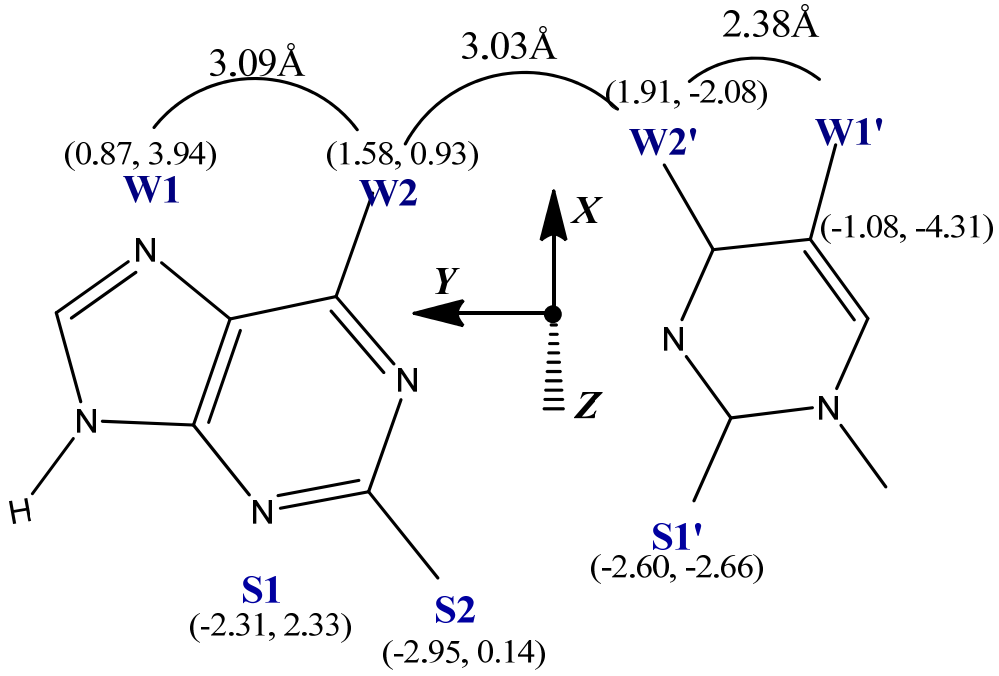
Atomic contacts are an important component of NAPID for the purpose of this research. The molecular interactions involved in DNA sequence recognition are non-covalent interactions, among which hydrogen bonding interactions are extremely important. It is believed that specificity is determined by compatibility between the DNA and protein secondary structures. The assumption made in this research is that all hydrogen bonds are formed in the base reference frame.

Because of the enormous complexity of pair-wise distance calculations in large molecular structures, in order to aid in the retrieval of pair-wise distance information, atomic distances are stored in NAPID tables, which can be searched relatively quickly. Atomic distances between nucleic acids, proteins and ligands are readily available in

NAPID by searching three tables: naproteinContacts, naligandContacts and nawaterContacts. For the study of DNA-protein atomic distances in this research, only the naproteinContacts is utilized. The interactions of DNA with water and ligand molecules, although an important and interesting study, are not the focus of this paper.

Each contact in this database is classified as either a phosphate, major-groove or minor-groove contact. Base or sugar contacts can occur in either the major or minor groove. In a standard anti arrangement, a base contact is defined as a major-groove contact if the local coordinates of the protein atom relative to the base satisfy the following criteria: $x > 0$ or $y + 0.75\ x - 3.5 > 0$. Figure 2.1 illustrates the hydrogen bonding sites in the standard base-pair reference frame. This plane is defined as the x-y plane. For the purpose of this research, sugar contacts were not studied, the focus of this paper is nucleic acid-base atom contacts.

Figure 2. 1: Generic base pair hydrogen bonding sites



2.1.2 Nucleic Acid Binding Proteins Stored in NAPID

Although this paper discusses atomic interactions in terms of DNA base and amino-acid atoms, it is important not to lose sight of the fact that the DNA bases are part of a larger helix and the amino acids are part of a larger highly complex protein. Proteins play a critical role in cell function and there are large varieties of proteins. Protein structure is an important factor that determines its function. A factor that makes proteins so valuable is their ability to bind to other molecules. DNA-binding proteins have domains with affinity for sites on the DNA helix. In addition, because the chemical versatility of the 20 amino acids is not unlimited, metal ions are sometimes incorporated as an intrinsic part of the protein structure. Iron, zinc, magnesium and calcium are often included in biological proteins.

Secondary protein structural elements play an important part in building the framework for biological interactions. The alpha (α) helix is the classic element of protein structure, with the second major element being the beta (β) sheet. These secondary structures then combine utilizing loop regions to build up larger more complex structures. These more complex structures are referred to as motifs. In addition, DNA-binding proteins can incorporate metals such as zinc, calcium and iron.

2.1.3 Non-Redundant Protein Set

In order to achieve a non-biased set of interactions, 499 high resolution (better than 2.5Å) non-redundant protein-DNA complexes were selected. The criteria used for selection of these structures were established based on a study of the original data set of 234 non-redundant structures created by Yun Li in his PhD dissertation. The list of original 234 non-redundant structures was expanded upon using a combination of manual

and automated methods, and 265 additional non-redundant structures were selected from those available in the PDB up to June 2011. Both sequence and structure similarities were studied as part of the evaluation process.

2.1.3.1 Classification Methods Within the PDB

Protein classification schemes available within the PDB web site (http://www.rcsb.org) were utilized as part of the evaluation process. The PDB web site provides a categorization of available structures using the widely employed methods: CATH (Classification, Architecture, Topology, Homologous superfamily), SCOP (Structural Classification of Proteins) and Pfam.

Within the CATH classification system there are four major levels: Class, Architecture, Topology (fold family) and Homologous superfamily [24]. Class is determined according to the secondary structure composition and packing within the structure. Three major classes are recognized: mainly-alpha, mainly-beta and alpha-beta. Architecture describes the overall shape of the structure based on the orientations of the secondary structures. Topology or fold represents the overall shape and connectivity of the secondary structures in the domain core. Homologous structures are believed to share a common ancestor.

The SCOP classification system categorizes structures based on class (domain), family, superfamily and fold [25]. Class refers to general architecture of the protein domain. Fold describes similar arrangements of secondary structures but without confirmation of ancestry. Superfamly includes both structural and functional similarity in order to infer a common ancestry, but not necessarily detectable sequence similarity.

Family includes some sequence similarity (usually >30%) and a clear evolutionary relationship.

Pfam provides a description of structures based upon domains [26]. It is a manually generated database that categorizes domains into families. Where possible, a Pfam family corresponds to a single structural domain.

All of the above classification systems were utilized in order to develop the non-redundant list. In addition, tools available within the PDB were utilized to classify the newly acquired non-redundant list in a manner consistent with data previously obtained by Yun Li.

2.1.3.2 PDB Query Interface

In order to generate a more manageable number of structures from the enormous number available in the PDB, the query interface available within the PDB web site was initially utilized. Incorporated into the query interface is the ability to select a subset of structures from which similar sequences have been largely removed. The query interface available within the PDB web site was employed in order to select high-resolution structures (better than 2.5Å), containing both a protein and DNA component, and also to eliminate major sequence similarity and non-redundant structures already selected by Yun Lin (deposited into the PDB after Dec 2004).

The PDB web site comparison tool within the PDB web site eliminates redundancy using sequence clustering [27]. The process employed for removing similar sequences is based on pre-calculated clusters of protein chains of at least 20 amino acids. These chains can be clustered by 100%, 95%, 90%, 70%, 50%, 40%, and 30% sequence similarity. For the purpose of this research 90% similarity was selected. Within each

cluster, the chains are then sorted (ranked by quality and methodology). From each cluster, the highest ranked chains (smallest rank #) are selected in order to generate a non-redundant set of chains. The query generated 555 non-redundant high resolution (better than 2.5Å) structures containing DNA and protein components that were deposited in the PDB database after Dec 2004.

2.1.3.3 Manual Evaluation Methodology

Results from the 555 structures (generated via the automated query capability described in the previous section) were then tabulated in an excel file and clustered based on CATH, SCOP and Pfam descriptions available within the PDB. Within each cluster, DNA and protein strands were sequentially compared. Identical sequences were then manually inspected within the PDB by comparing space groups and utilizing Jmol and "simple viewer" software applications for viewing chemical structures in 3D [20, 28] If redundancy was still questionable, the PyMOL utility was used to overlay the structures and make spatial comparisons. PyMOL is a molecular viewing system for visualizing 3D images of small molecules and biological macromolecules [29]. For the purpose of this research, structures containing a mutant DNA, as well as single-stranded DNA were eliminated.

Appendix 1A provides the NDB file numbers and classifications of the protein-DNA complexes studied. Within the database, the DNA-binding proteins were organized based on function, sequence homology and architecture. Appendix 1B presents the available PDB classification, as well as the CATH, SCOP and Pfam information for each structure studied. Because many structures have not been assigned data in all four classification systems, data from all these systems are presented graphically. The protein

families represented in NAPID are described in Figure 2.2 a-d. For clarity, the CATH, SCOP and Pfam classification pie charts do not show counts < 5. For the PDB general classification pie chart, counts < 10 are not shown. Classifications with these small counts (< 5 or < 10) are presented together within a category called "other".

Figure 2.2a: Protein families in the NAPID database–PDB classification

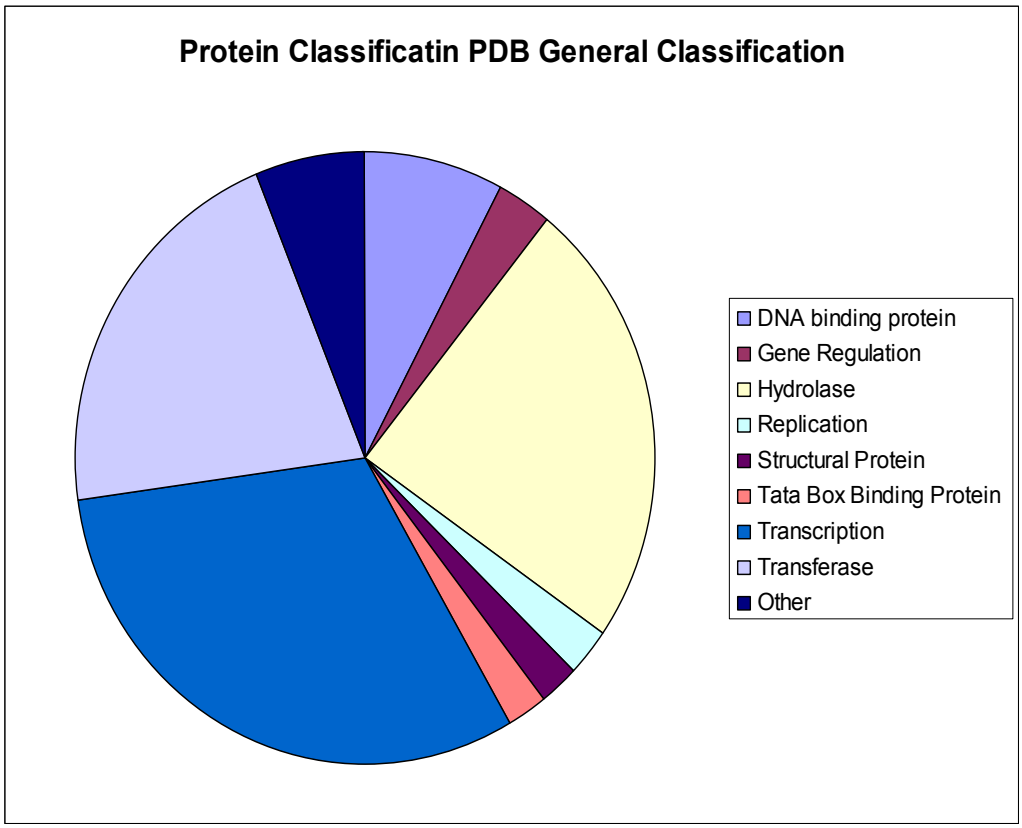


Figure 2.3b: Protein families in the NAPID database–CATH classification

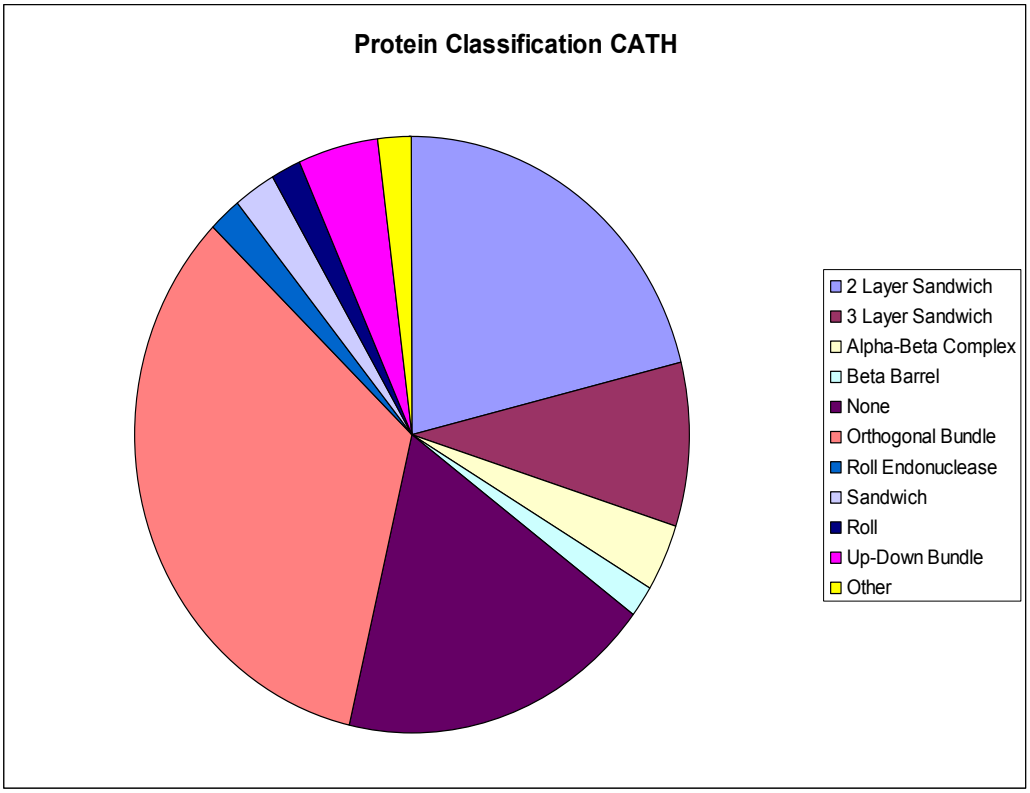


Figure 2.4c: Protein families in the NAPID database – SCOP classification

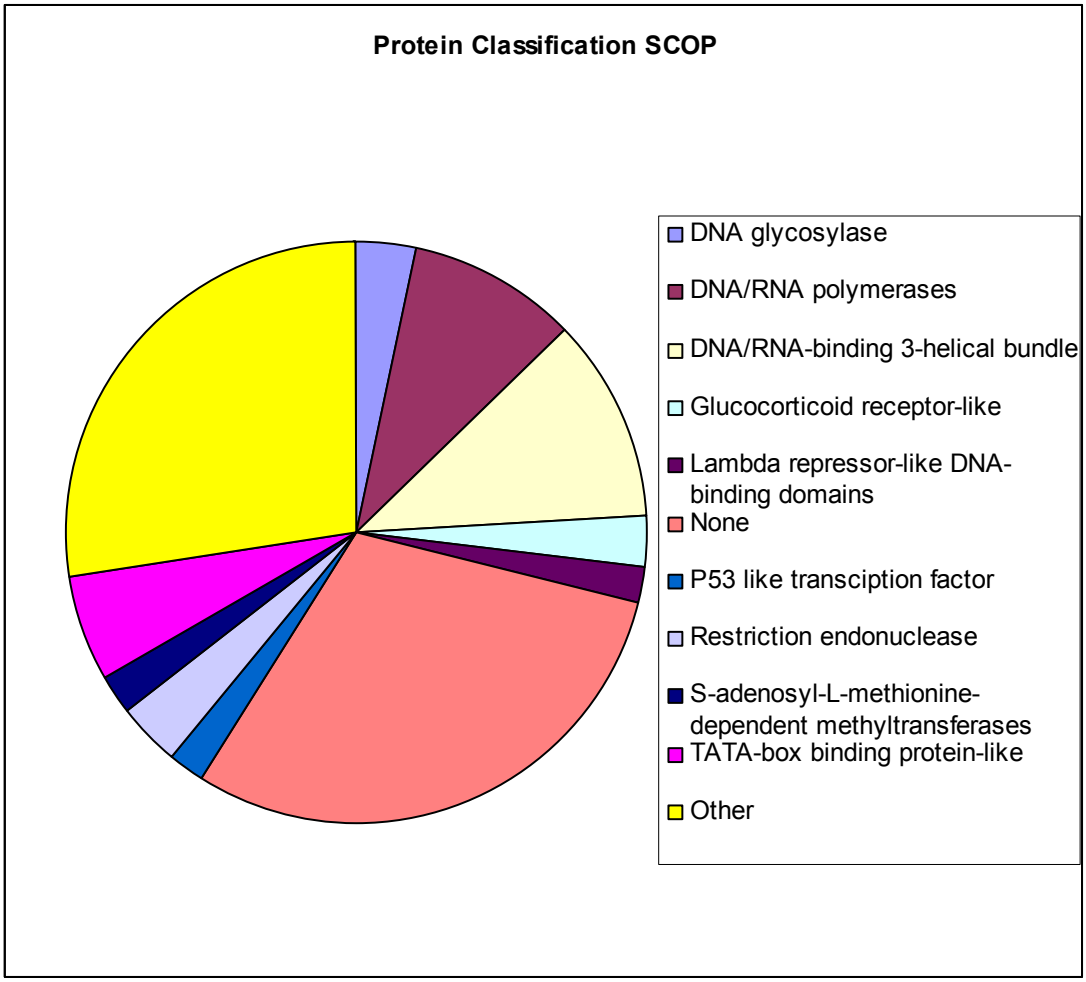
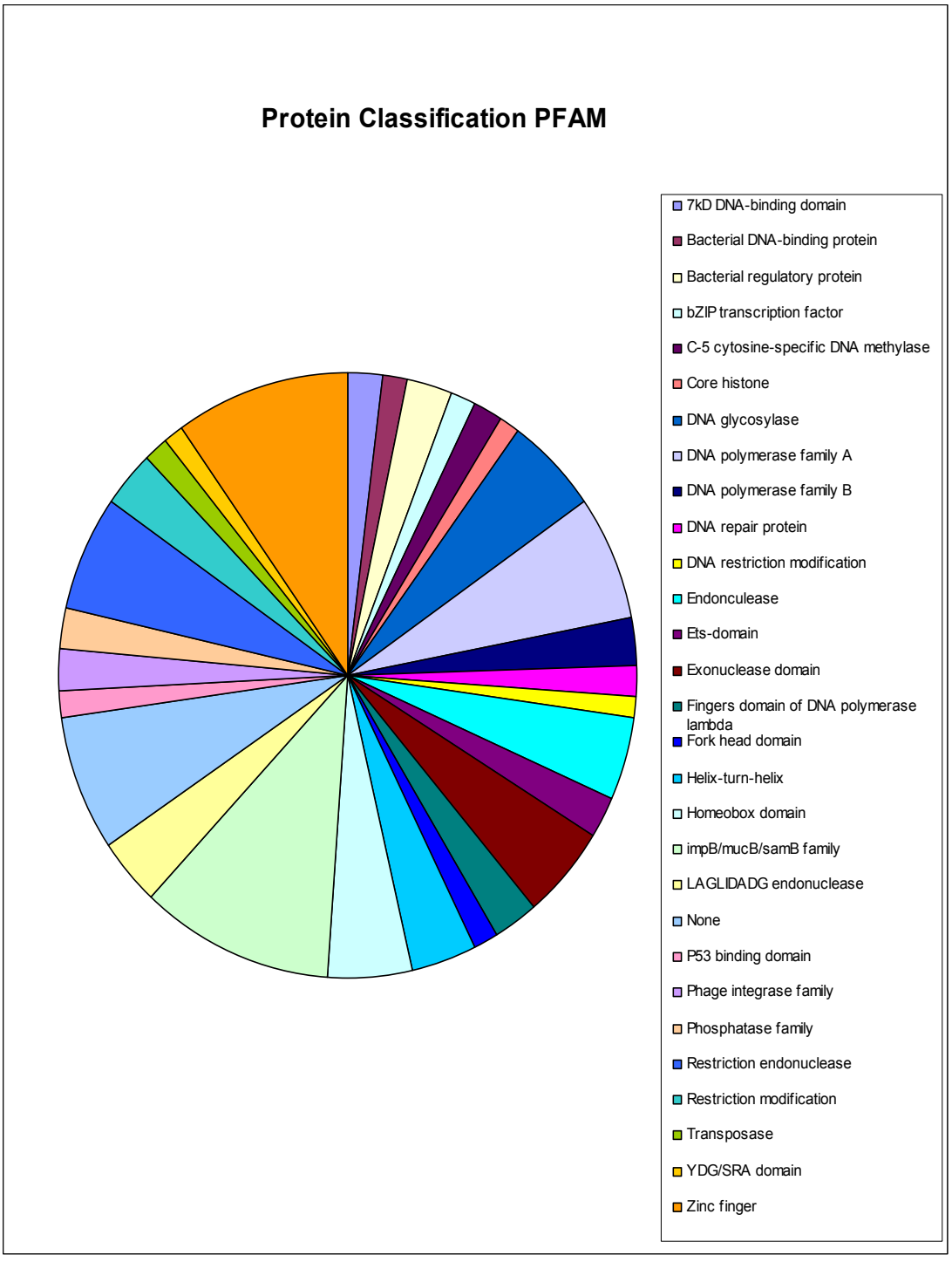


Figure 2.5d: Protein families in the NAPID database – Pfam classification



2.2 Generation, Storage and Analysis of Hybridization Information

In order to study the consequence of orbital hybridization in amino-acid atoms, a new database was created to specifically store, search and analyze this type information. This database is referred to as the Hybrizationtest Database (HTD). It is a relational database with the same structural attributes as NAPID.

2.2.1 Hybridization Data Tables

In order to study the relationship between amino-acid atom hybridization and nucleic acid atom contacts, a table describing the hybridization of each atom in the 20 basic amino acids is necessary. The 20 amino acids studied are illustrated in Figure 1.9.

The table which contains the hybridization assigned to each atom in the 20 amino acids stored in the HTD is designated as the AAHybridization table. This table was generated by first extracting the structural features associated with the 20 unique amino acids to be studied into a dataset. The following parameters were extracted into this dataset from the ProteinAtom table of NAPID: resName (standard name of amino acid), name (atom name in standard IUPAC designation), and groupType (location, backbone or side chain). The use of Structured Query Language (SQL) was utilized to retrieve this information. Appendix 2 describes the SQL script utilized to retrieve these data.

After the standard amino-acid dataset was created as a text file, each atom was assigned a hybridization designation. Atoms that contained a double bond were assigned sp^2 hybridization and atoms where all bonds were single bonds were assigned sp^3 hybridization. Since only one hybridization assignment per atom can be made in the AAHybridization table, atoms containing delocalized electrons have been designated as sp^2.

In order to make comparisons to data in NAPID, a table to store the hybridization information was created using an additional SQL script. The SQL script is included in Appendix 3, and the output to this query in Appendix 4. The data from the query output was stored as a flat text file and was then loaded into the table in the shared database. The command used to load the data is included in Appendix 5A.

2.2.2 Hybridization Comparisons

The data stored in the AAHybridization table was then compared to the data derived by Yun Li in NAPID. Distances of $\leq$ 3.4Å are considered close contacts, therefore this distance was used as a point of reference for investigation. For the purpose of the research, close contacts are considered two atoms that come within a distance of 3.4 Å of each other. A study of sp^2 versus sp^3 hybridized amino-acid atoms in contact with each Watson-Crick nucleic acid base (Adenine, Thymine, Guanine and Cytosine) was undertaken. Although major-groove contacts are more abundant than minor groove contacts, contacts in both grooves were investigated in this study. Appendix 5B is an example of a SQL query that was developed in order to study sp^2 guanine interactions in the major groove with contact distances $\leq$ 3.4Å.

Within the NAPID database, two important tables were utilized to generate the data within the HTD. These tables are the naProteinContacts and the naAtoms tables. The naProteinContacts table stores contacts between nucleic acid atoms and protein atoms. The naAtoms table stores nucleic acid atom information and the proteinAtoms table stores amino-acid atom information.

The critical table within HTD used to generate close contact information is the AAHybridization table. As previously noted, the AAHybridization table is the table that stores hybridization information for each amino-acid atom.

CHAPTER 3
HYBRIDIZATION CONTACT ANALYSES IN THE MAJOR GROOVE

The functioning of biological systems is highly dependent on the interaction of proteins with DNA in the cell. Therefore it is important to study the various types of DNA-protein interactions. As previously discussed, these interactions can be specific or non-specific in nature. The non-specific interactions are mostly independent of base sequence [30]. Furthermore, they are typically ionic in nature. In contrast, some proteins such as transcription factors interact with DNA in a sequence-specific manner [31].

The contacts between atoms within proteins and the atoms of the DNA bases, particularly in the major groove, are an important factor in DNA recognition [9]. Furthermore, the DNA-protein contacts show distinct microenvironments. Yun Li in his PhD research studied interaction of DNA with the following types of amino-acid atoms: positively charged, negatively charged, hydrogen-bond donors, hydrogen-bond acceptors and hydrophobic atoms. The results were presented in his dissertation [9].

The effect of hybridization on protein-base interactions has not previously been studied and is of potential importance. As described in Section 2, a comparison of sp^2 versus sp^3 hybridized amino-acid atoms in contact with DNA nucleic acid atoms was undertaken in this research.

Among the 499 non-redundant structures, it is important to note that each of the four nucleic acid bases studied, adenine, thymine, guanine and cytosine, are present in almost equal proportion. Within the dataset studied there are 5400 adenine bases, 5433 thymine bases, 5399 guanine bases, and 5375 cytosine bases. Since the denominator for all counts discussed in the subsequent sections is approximately the same for all bases, for simplicity it has been omitted specifically from calculations. Since the bases are

present in almost equal proportions within the research database, comparisons will be representative of those taking into account overall number of bases in the database.

3.1 General Overview of sp^2 versus sp^3 Hybridized Atom Interactions in the Major Groove of DNA

Figure 3.1 shows DNA major-groove contact counts (≤3.4Å) by distance for each of the four bases within the dataset studied (499 representative non-redundant protein-DNA structures); adenine, thymine, guanine and cytosine. Blue represents sp^2 hybridized atom contacts and purple represents sp^3 hybridized atom contacts. Data in this section illustrate the effect of amino-acid atom hybridization on interaction with the nucleic acid bases in the major groove on a general level. Specific atomic details will be explored in later sections.

The plot in Figure 3.1 illustrates that within this data sample, the guanine-cytosine base pair shows a higher absolute number of close interactions (≤ 3.4 Å) than the adenine-thymine base pair (4552 and 2320, respectively). Furthermore, for the guanine-cytosine base pair, close interactions (≤ 3.4 Å) are dominated by contacts with sp^2 hybridized amino-acid atoms. For guanine there are 1854 close contacts with sp^2 hybridized atoms and 1346 close contacts with sp^3 hybridized atoms. For cytosine, within this data set there are 1057 close contacts with sp^2 hybridized atoms and 295 close contacts with sp^3 hybridized atoms. Although guanine shows a higher absolute number of close contacts, the relative dominance of sp^2 over sp^3 close contacts is much greater for cytosine than guanine. This is supported by data from Yun Li's dissertation showing that negatively charged atoms from GLU and ASP, which have sp^2 hybridization, accumulate to a greater extent on cytosine bases [9].

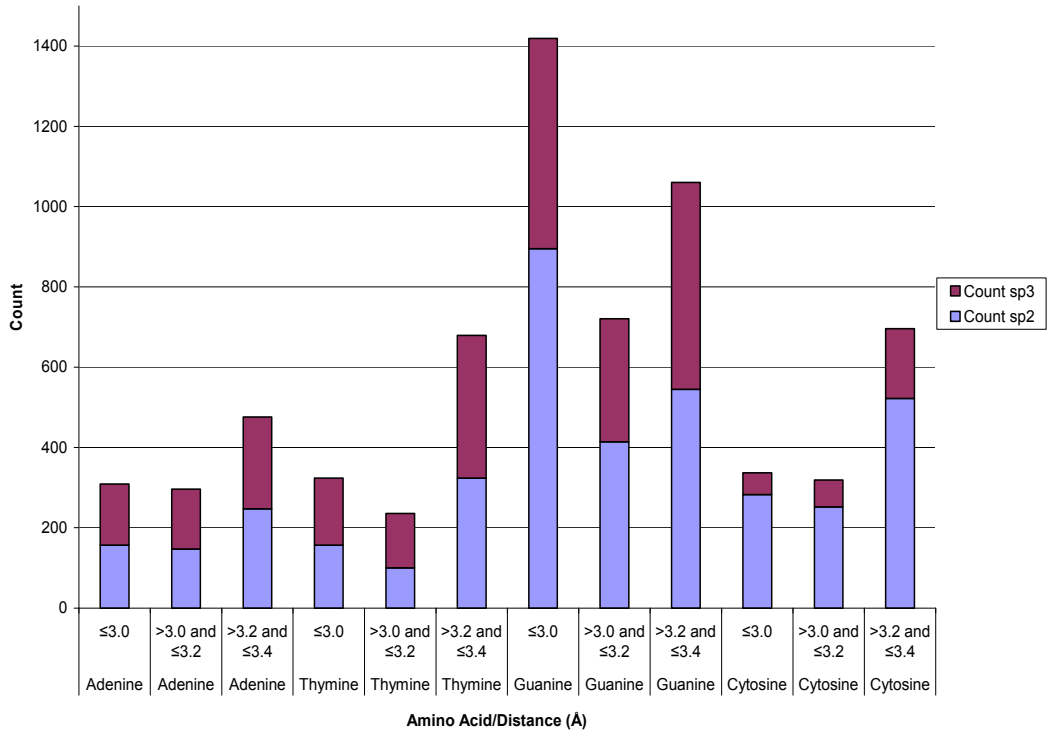
The information in Figure 3.1 also shows that of the bases studied, guanine is most intimately contacted (≤3.0 Å) with amino-acid atoms, with a preference for sp^2 hybridized atoms over sp^3 hybridized atoms (895 and 524, respectively). In contrast to guanine, its Watson-Crick base-pair partner, cytosine, appears to experience a much lower number of very close contacts (≤3.0 Å). These very close contacts are indicative of ionic interaction rather than a hydrogen bond. The atoms of these bases appearing in very close proximity (≤3.0 Å) to protein atoms may actually be the result of ionic bonding between the DNA backbone and protein atoms, and so simultaneously are presenting as being very close to base atoms. Although this research does not specifically sort out these features, it is a topic that is worth further investigation.

For guanine, as the distance is increased (between > 3.0Å and ≤ 3.4Å) the number of contacts becomes more equal between sp^2 and sp^3. Cytosine shows a higher relative number of sp^2 contacts over sp^3, over all contact distances studied. This is consistent with previous findings for the non-redundant list studied by Yun Li [9].

As illustrated in Figure 3.1, for the adenine-thymine Watson-Crick base pair, over all the close contact distances ≤ 3.4Å studied in this research, thymine shows higher numbers than adenine (1239 and 1081 contacts, respectively). These counts include both sp^2 and sp^3 hybridized amino-acid atoms, and are driven mostly by the higher number of contacts at moderately close distance >3.2Å and ≤3.4Å rather than the very close distance ≤3.0 Å. At very close distance (≤3.0 Å), thymine appears to have only a slightly higher number of very close contacts than adenine within the 499 structures studied (324 and 309, respectively). At distances >3.2 and ≤ 3.4Å, the number of contacts increases further for thymine, particularly those involving an sp^3 hybridized amino-acid atom (581

total, 324 sp^2 and 355 sp^3). Therefore, it appears that the steric hindrance imposed by the methyl group on adenine is less pronounced at the somewhat higher distances studied, allowing even more interactions with the carbonyl on thymine.

Figure 3. 1: Major groove close contact counts for sp^2 and sp^3 hybridized amino-acid atoms



This graph shows DNA major-groove close contact counts by distance for each of the bases — adenine, thymine, guanine and cytosine within the dataset studied (499 representative non-redundant protein-DNA structures). Blue represents sp^2 hybridized atom contacts and purple represents sp^3 contacts.
Overall base numbers within research dataset: adenine = 5400, thymine = 5433, guanine = 5399, cytosine = 5375

Source data Appendix 6

3.2 Detailed Comparison of Amino-acid Hybridization Contacts with Specific Nucleic Acid Atoms Types in the Major Groove

As noted in Section 1, the W1 and W2 sites are the major points of hydrogen-bonding interaction in the major groove. These sites contain the primary H-bond donor and H-bond acceptor groups (carbonyl and amine, respectively). Refer to Figure 1.3 and Figure 1.4 for an illustration of the base pair structures and hydrogen-bonding sites.

For adenine, the imidazole N7 atom represents the W1 site, and the N6 atom of an amine group the W2 site. Thymine has the W1′ site occupied by a C5M atom of a methyl group, and the W2′ position occupied by the O4 atom of a carbonyl. The W1′ site occupied by the C5M atom of the methyl group does not represent a site of H-bond interaction but rather a potential site for hydrophobic interaction.

For the G-C Watson Crick base pair, the guanine W1 site is occupied by an imidazole N7 atom similar to adenine; however the W2 site is occupied by an O6 atom of a carbonyl. For cytosine the W1' site is occupied by a hydrogen atom, as compared to the bulky methyl group in thymine. Therefore the cytosine W2′ location is much less obstructed in the C-G base pair than the W2′ site on thymine in the A-T pair, since there is no methyl group hindering the W1′ location for cytosine.

It is of interest to study close interactions ≤ 3.4 Å for the various atoms types in each base pair, since these potentially represent hydrophilic, hydrophobic and hydrogen bonding type interactions. The location and type of atom within each base cumulatively result in a greatly different number of close contacts, as shown in Figures 3.2 and 3.3. Furthermore, there are differences between contact numbers for sp^2 versus sp^3 hybridized atoms in each location studied.

3.2.1 Review of Major-groove contacts by Base Atom for Adenine-Thymine

Figure 3.2 shows the counts of close contacts (≤3.4Å) by atom type for the Watson-Crick A-T base pair in the major groove for the non-redundant 499 samples chosen. Blue represents sp^2 close contacts and purple represents sp^3 close contacts. The N6 in the W2 position and N7 in the W1 position of adenine show the largest number of close contacts for this base, with N6 having only slightly more close contacts than N7 (435 and 427, respectively, However, N6 shows a much higher number of sp^2 close (≤3.4Å) interactions than N7 (326 and 129, respectively). This is consistent with the premise that the amine group in the W2 position would interact largely as an H-bond donor with electron rich sp^2 hybridized orbitals (e.g. carbonyl atoms) that would conversely act as H-bond acceptors. The N7 position is dominated by close contacts with sp^3 hybridized atoms. This is likely due to the fact that the electron rich double bond between the N7 and C8 atoms is not as attracted to other electron rich sp^2 hybridized amino-acid atoms.

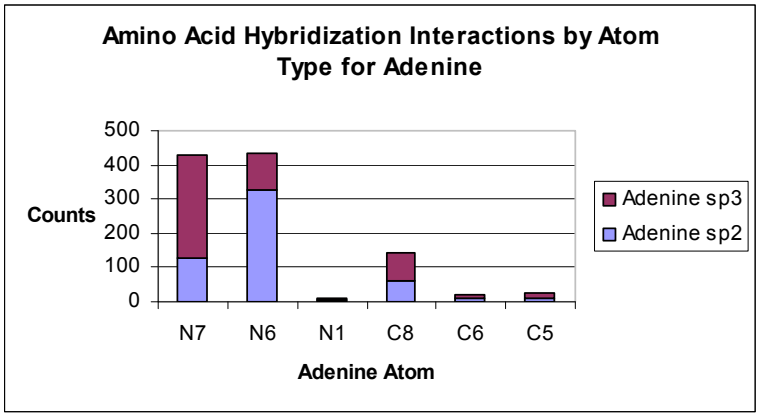
The highly exposed C8 at the edge of the major groove in adenine also shows some close interactions, primarily with sp^3 bonded atoms. Of the 141 total close contacts involving C8, 81 are with sp^2 hybridized atoms and 60 are with sp^3 hybridized atoms. This is consistent with the C8 interacting closely with the NH atoms of the peptide bond in proteins. The C5, C6 and N1 atoms of adenine do not show many close interactions due to their inaccessible location buried within the major groove (24, 22 and 12 total close contacts, respectively).

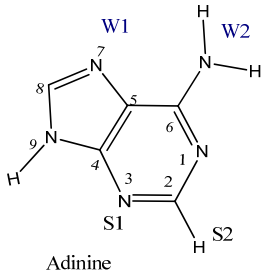
For the thymine base, the highest number of close (≤ 3.4 Å), interactions occur at the W2' site occupied by an O4 carbonyl atom (661 close contacts). These close

interactions are dominated by close contacts with sp^3 hybridized atoms (469 close contacts). This would be consistent with a side-chain amine in a protein interacting with the O4 atom of the carbonyl group on the base. Once again, this is a typical H-bond donor/acceptor relationship.

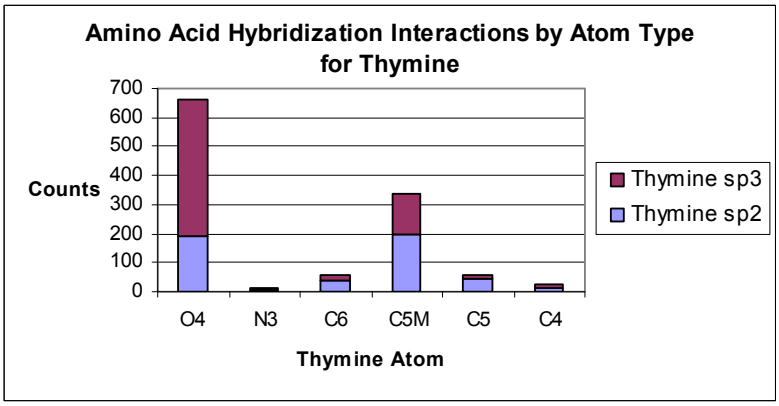
As shown in Figure 3.2 the C5M atom in thymine also shows a high relative number of close contacts (339 total close contacts). The C5M atom has a more equal number of sp^2 and sp^3 interactions (195 and 144, respectively). This non-polar group is slightly dominated by close interactions with sp^2 atoms of the amino acids. However, there are also are a significant number of sp^3 contacts. Since the methyl group is non-polar there is a lack of repulsion with the electron rich sp^2 hybridized atoms in the double bonds of the interacting amino acids as well as non-polar sp^3 hybridized atoms. As previously discussed, close interactions with the non-polar methyl group are hydrophobic in nature.

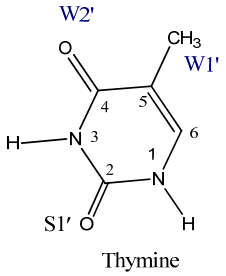
Figure 3. 2: Major groove close contact counts by base atom type for A-T





Adinine





Thymine

Overall A-T base numbers within research dataset: adenine = 5400, thymine = 5433
Source Data Appendix 7

3.2.2 Review of Major-groove contacts by Base Atom for Guanine-Cytosine

Figure 3.3 shows the counts of close contacts (≤ 3.4 Å) by atom type for the G-C base pair in the major groove for the non-redundant 499 structures under review. Blue represents sp^2 close contacts and purple represents sp^3 close contacts. As shown, the O6 atom in the W2 location of guanine has the largest overall number of close contacts (≤ 3.4 Å). The number of close interactions to the O6 atom of guanine is relatively similar for sp^2 and sp^3 atoms. Of the 1491 total close contacts with the O6 atom in guanine, there 783 with sp^2 hybridized atoms and 708 with sp^3 hybridized atoms. This is consistent with the premise that the negatively charged carbonyl group on guanine would interact largely with a terminal side-chain amine (sp^3 hybridized) of proteins, as well as with the NH atoms within the peptide bonds of proteins (sp^2 hybridized). Once again this is a typical H-bond donor/acceptor relationship.

The W1 location on guanine which is occupied by the N7 atom also has a high number of close contacts (≤ 3.4 Å). Of the 1275 total close contacts with the N7 atom in guanine, there 809 with sp^2 hybridized atoms and 466 with sp^3 hybridized atoms. The N7 is rich in electrons from the sp^2 hybridized double bond it makes with the C8 atom Similar to the O6, it can closely interact with a terminal amine (sp^3 hybridized), as well as the NH atoms within the peptide bond of proteins (sp^2 hybridized).

The C6, C5 and N1 atoms on guanine show the fewest close interactions with amino-acid atoms (104, 49 and 12 total close contacts, respectively). This is understandable as a result of the steric factors associated with these atoms, since they are not as easily accessible within the groove. The C8 shows a slightly higher number of interactions with amino-acid atoms than the C6, C5 or N1 atoms since it is located on the

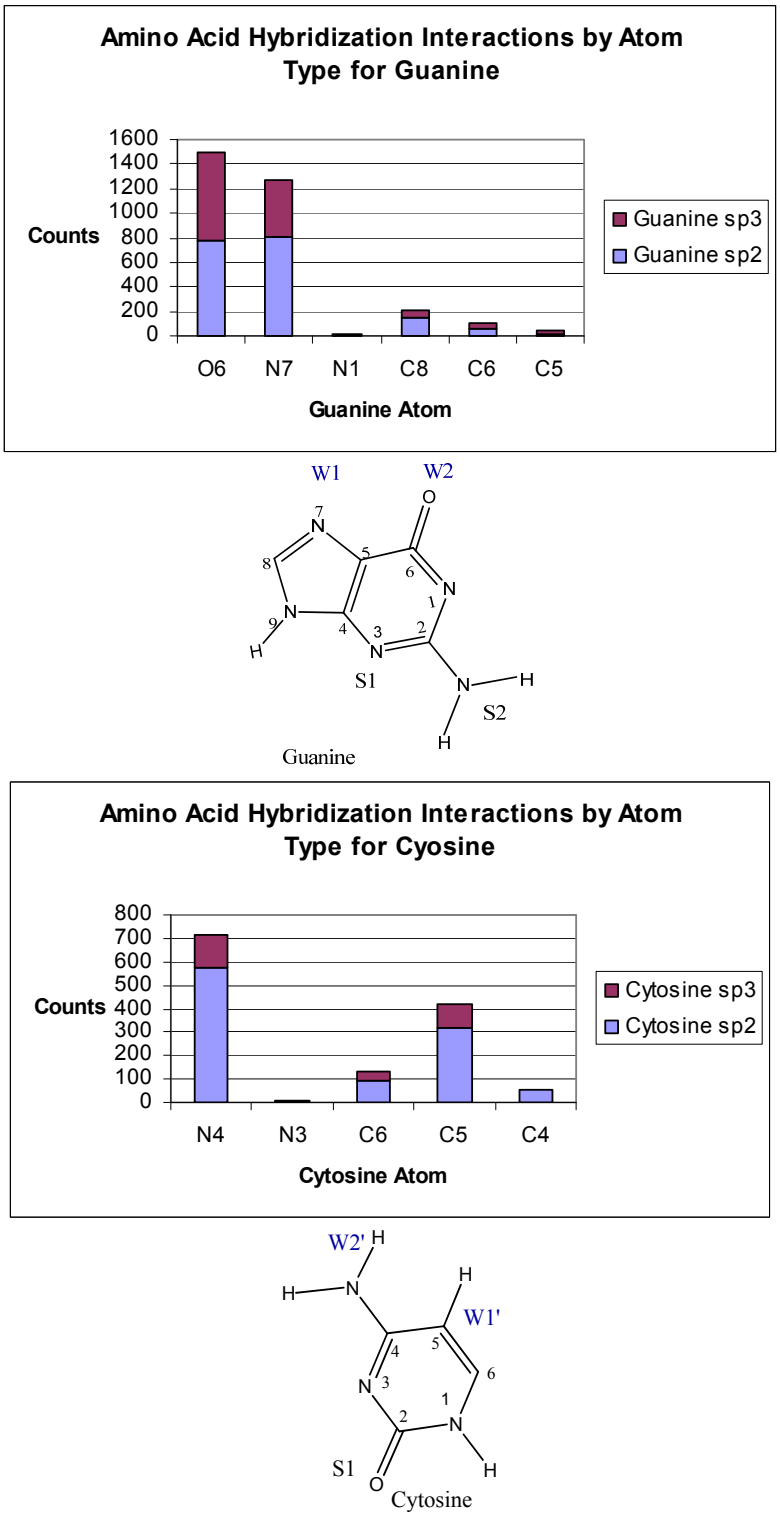
outside edge of the major groove and would be more accessible (214 total close contacts). Similar to A-T interactions, these graphs are also in accord with the idea that the W1 (N7) and W2 (O6) sites are the primary sites of protein-DNA interaction in the major groove.

For cytosine, the highest number of close interactions (≤ 3.4 Å) occurs at the W2' site occupied by the N4 atom in the amine group. There are 712 close contacts with the N4 atom. These close contacts are significantly dominated by interactions with the sp^2 hybridized atoms. There are 572 close contacts between N4 and sp^2 hybridized atoms and 140 between N4 and sp^3 hybridized atoms. This would be consistent with a carbonyl within a protein interacting with the amine group on cytosine.

The C5 aromatic carbon on cytosine also has a large number of contacts, but not as great as the N4. There are 422 close contacts involving the C5 atom. This finding is consistent with the accessible location of the C5 atom in the major groove. The C5 atom on cytosine is more accessible than the C5 atom on thymine, as a result of the absence of the bulky methyl group for cytosine.

The N3 atom has the fewest number of close interactions for atoms in the major groove (4 total contacts). Understandably this is due to its inaccessible location. The C4 atom also has relatively few close contacts (57 total contacts). This is potentially due the steric constraints imposed by the amine group, which could prevent atoms from coming in close proximity. Finally, the C6 atom on cytosine has slightly more close contacts than C4 (129 total close contacts). This is reasonable due to its more accessible location on the outer edge.

Figure 3. 3: Major groove close contact counts by base atom type for G-C



Overall G-C base numbers within research dataset: guanine = 5399, cytosine = 5375

Source Data Appendix 7

3.3 Amino-Acid Backbone versus Side-chain Contacts in the Major Groove

Figure 3.4 compares close contact (≤ 3.4 Å) counts for atoms defined as being located on the backbone or side chain of amino acids with each of the four bases studied in this research. Refer to Figures 1.11, 1.12 and 1.13 for an illustration of the core backbone and side chains for each of the 20 amino acids studied. Blue represents close contacts between each of the bases noted and amino-acid backbone atoms, and purple represents close contacts with side-chain atoms. Notably, across the four bases, close-contacts are significantly dominated by side-chain interactions. Of the 6781 total close contacts (among the 499 protein-DNA structures studied) only 723 of these close contacts involve a backbone atom, as compared to 6058 close contacts that involve a side-chain atom.

Data in Figure 3.4 show that the A-T Watson-Crick base pair has fewer overall close contacts than the G-C base-pair (2233 and 4548, respectively). Moreover, this figure illustrates that the G-C Watson-Crick base pair shows a higher degree of backbone close interactions (≤ 3.4 Å) than the A-T pair (549 and 174, respectively).

3.3.1 Review of Adenine-Thymine Close Contacts in the Major Groove

Both adenine and thymine show a low number of close contacts (≤ 3.4 Å) with backbone atoms, but adenine shows a slightly lower number of backbone contacts than thymine, as can been seen in Figure 3.4. Among the 499 structures studied, adenine only showed a total of 82 close atomic contacts (≤ 3.4Å) with backbone atoms. This figure also shows that adenine has a higher relative number of close contacts with backbone atoms for sp^2 over sp^3 hybridized atoms. Of these 82 close contacts, 69 involve an sp^2 hybridized atom and only 13 involve an sp^3 hybridized atom. This is consistent with the

fact that there is a carbonyl in the backbone of all amino acids that could presumably interact through hydrogen bonding with the amine in the W2 position of adenine. The relatively low numbers for backbone interactions would appear to indicate that there are also conformational factors influencing close interaction of amino-acid backbone atoms with adenine.

As seen in Figure 3.4, thymine has a slightly higher number of overall close contacts with backbone atoms than adenine. Thymine showed 92 close atomic contacts with backbone atoms. For thymine close interactions with sp^2 backbone atoms are favored over close interactions with sp^3 backbone atoms (63 and 29, respectively).

As noted in Figure 3.4, most close contacts ($\leq$ 3.4 Å) experienced by adenine are with side-chain atoms. This figure illustrates that among the 499 structures studied in this research, adenine experiences a total of 997 close atomic contacts with side-chain atoms. Therefore for adenine it is apparent that there is a greater affinity for side chain than backbone atoms. Adenine interacts to a slightly greater extent with sp^3 amino-acid side-chain atoms than sp^2 side-chain atoms. As described in Figure 3.4, adenine shows a total of 516 close contacts with sp^3 side-chain atoms versus 481 close contacts with sp^2 side-chain atoms.

For thymine there is a greater domination of sp^3 versus sp^2 hybridized side-chain atom close contacts than for adenine. As described in Figure 3.4, thymine shows a total of 628 close contacts with sp^3 side-chain atoms versus 434 close contacts with sp^2 side-chain atoms. This is due to the fact that the carbonyl O4 oxygen is a strong hydrogen bond acceptor that can interact with terminal amino groups in the side chain of amino acids within proteins. These data also indicate the importance of conformational factors,

since the side-chain amino groups may be better able to fit into the DNA major groove than the backbone groups.

3.3.2 Review of Guanine-Cytosine Close Contacts in the Major Groove

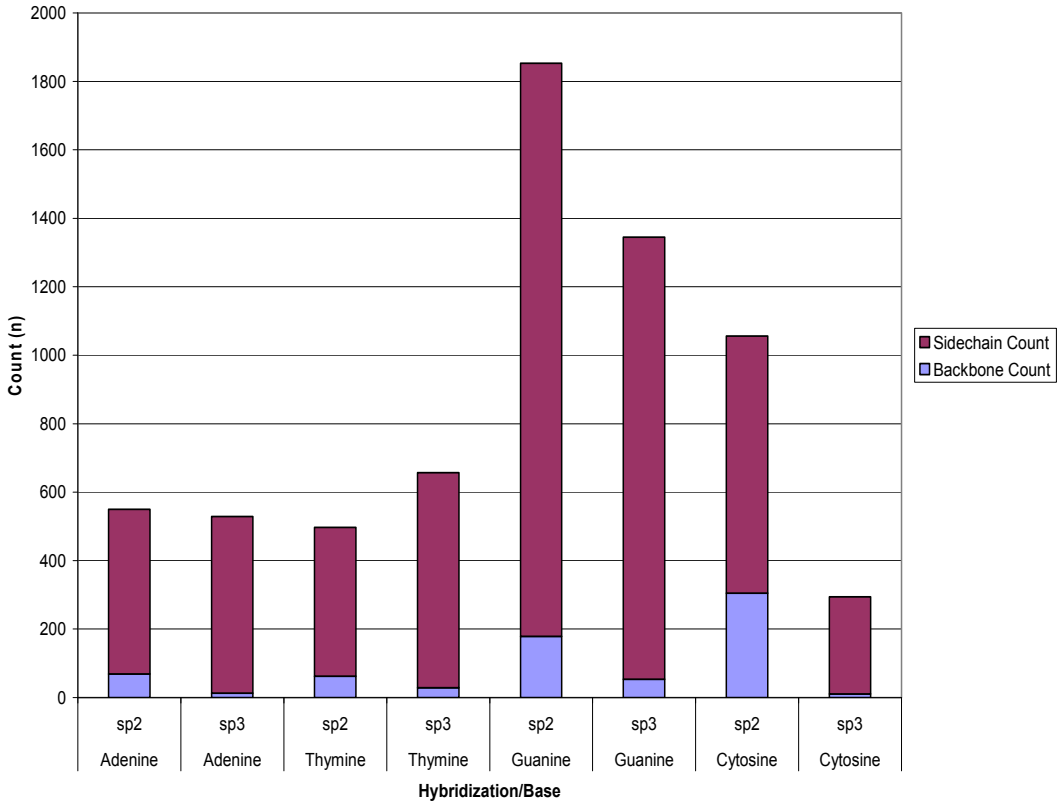
As shown in Figure 3.4, within the dataset studied, the G-C base pair exhibits a higher number of close contacts (≤ 3.4Å) with amino-acid backbone atoms, than the A-T base pair in the major groove. Although guanine shows a higher comparative number of close contacts with backbone atoms than adenine or thymine, the number of close contacts with backbone atoms is much lower than with side-chain atoms. Among the 499 protein-DNA structures studied, guanine showed a total of 233 close atomic contacts (≤ 3.4Å) with backbone atoms, in contrast to 2965 close atomic contacts with side-chain atoms. Of the 233 close contacts with backbone atoms, 179 involve an sp^2 hybridized atom and 54 involve an sp^3 hybridized atom.

As described in Figure 3.4, in the major groove, cytosine is the only base studied that demonstrated a moderate absolute number of close contacts with backbone atoms. Within the dataset studied, cytosine exhibits 316 close contacts with backbone atoms. Of the 316 close contacts with backbone atoms, 305 involve an sp^2 hybridized atom and 11 involve an sp^3 hybridized atom. Of the 1034 close contacts with side-chain atoms, 751 involve an sp^2 hybridized atom and 283 involve an sp^3 hybridized atom.

Of the four bases studied, guanine exhibits the highest number of close contacts (≤ 3.4Å) with side-chain atoms. For guanine, of the 2965 close contacts with side-chain atoms, 1674 involve an sp^2 hybridized atom and 1291 involve an sp^3 hybridized atom. Cytosine shows a higher number of close contacts with side-chain atoms than adenine or thymine. However, the number of close side-chain contacts is not as great as guanine. In

addition, cytosine shows a greater number of close contacts with sp^2 hybridized side-chain atoms than sp^3 hybridized side chain atoms (751 and 283, respectively).

Figure 3. 4: Major groove close contact counts backbone versus side chain by hybridization



This graph compares close contact (≤ 3.4 Å) counts for backbone and side-chain amino-acid atoms with each of the four bases studied in this research. Blue represents amino-acid backbone atom close contacts and purple represents side-chain atom close contacts with each of the bases.

Overall base numbers within research dataset: adenine = 5400, thymine = 5433, guanine = 5399, cytosine = 5375

Source Data Appendix 8

3.3.3 Hybridized Atom Interactions by Amino-acid Type in the Major Groove

Figures 3.5 through 3.22 describe protein-DNA interactions by specific amino acids contained in the DNA-protein structures studied in the major groove. It is interesting to note the differences in the degree of interaction among the 499 structures included in this research. Refer to Figures 1.11, 1.12 and 1.13 for the amino acid atom naming convention utilized in the discussion text in Sections 3.3.1.1 through 3.3.3.8.

3.3.3.1 Sp^2 Hybridized Atom Contacts with Adenine in the Major Groove

Sp^2 Hybridized Side-Chain Atom Contacts

Figure 3.5 displays close interaction (≤ 3.4 Å) by an sp^2 hybridized amino-acid side-chain atom with an adenine atom, classified by amino-acid type. There are a total of 481 close contacts involving sp^2 hybridized atoms with adenine. As shown in this figure, the polar amino acids, arginine, asparagine and glutamine show the highest number of sp^2 hybridized-atom close contacts with adenine in the dataset studied (153, 152 and 87, respectively). This is consistent with the postulate by Seeman *et al.* that asparagine and glutamine recognize adenine in the major groove [8]. Although aparagine and glutamine show a high number of close contacts with adenine in the major groove, arginine, glutamic acid, histidine and aspartic acid also show close contacts with adenine in the major groove. Moreover, the highest numbers of close contacts for adenine in the major groove occur with arginine. These data are also consistent with the work of Madel-Gutfreud *et al.,* which showed large numbers of other types of close interactions in the major groove beside those occurring with asparagine and glutamine [13].

As shown in Figure 1.12, arginine has been assigned two sp^2 hybridized nitrogen atoms within the AAhybridization database. Because arginine has two high potential sp^2

hybridized interaction sites, there is a high likelihood for arginine to exhibit a great many close contacts. Among the structures studied, there is almost an equal split in the number of close contacts for the N_{H1} and N_{H2} atoms of arginine(62 and 79, respectively). These sp^2 hybridized atoms on arginine closely interact with the N6 and the N7 of adenine. For the N6 atom of adenine, there are 24 close contacts with the N_{H1} atom and 38 with the N_{H2} atom. For N7 atom of adenine, there are 24 close contacts with N_{H1} and 20 with N_{H2}. The N_{H1} and N_{H2} atoms have been assigned sp^2 hybridization within this database, and the similar number of close interactions with N7 and N6 would indicate that these two nitrogen atoms act similarly, and therefore it is appropriate treat them as such. Furthermore, although the assignment of the N_{H1} and N_{H2} labels is arbitrary, these data provide support for the premise of delocalized bonds within arginine. It appears that these nitrogen atoms have both sp^2 and sp^3 character, since they interact with hydrogen bond donors (N6) and acceptors (N7). However, as previously noted, only one assignment per atom could be made for comparison purpose within the database, and therefore sp^2 was selected due to the delocalization of electrons. It is interesting that although there is only small absolute number of close contacts between the N_{H1} and N_{H2} atoms of arginine and the N6 atom of adenine, these close interactions indicate that although the guanidinium group exhibits an overall positive charge, there are still some close contacts formed between this group and hydrogen bond donors such as the N6 atom of adenine.

Asparagine and glutamine are amides, as illustrated in Figure 1.12. For the interaction of adenine with asparagine, 134 of the 152 close contacts involve the O_{D1} carbonyl atom and 18 involve the C_G atom. For the interaction of adenine with glutamine,

84 of the 87 close contacts involve the O_{E1} carbonyl atom and only 3 involve the C_D atom. Asparagine and glutamine are very similar in structure. Asparagine has one additional CH_2 group in its backbone. As shown in Figure 3.2, the N6 atom on adenine shows the highest number of interactions with sp^2 hybridized amino-acid atoms. For asparagine there are 108 close contacts between its O_{D1} atom and the N6 atom on adenine, and only 20 close contacts between its O_{D1} atom and the N7 atom on adenine. For glutamine there are 66 close contacts between its O_{E1} atom and the N6 atom of adenine, and 13 close contacts between its O_{E1} and the N7 atom of adenine. Clearly the sp^2 hybridized oxygen on both of these amides prefers the N6 amine atom in the W2 location of adenine.

It is interesting to note that although phenylalanine, tryptophan and tyrosine have a comparatively high number of sp^2 hybridized side-chain atoms within the heterocyclic ring(s); these sp^2 hybridized atoms do not come in close contact with adenine in the major groove. Of these, only phenylalanine and tyrosine show a small number of close contacts of sp^2 hybridized side-chain atoms with adenine in the major groove (4 close contacts each). This could potentially be the result of the bulky size of these rings, and therefore there may be some sort of size constraint. In addition, these heterocyclic rings may not serve as good proton acceptors. For phenylalanine, only the C_{E1} and C_{E2} atoms show any close interactions (3 and 1, respectively). Similarly for tyrosine the C_{E1} and C_{E2} atoms both show 2 close contacts. All of these, are side-group atoms.

Sp^2 Hybridized Backbone Atom Contacts

As previously noted, among the 499 structures studied, adenine shows relatively few close contacts ($\leq$ 3.4 Å) with backbone amino acids. Conformational factors may play an

important role in backbone-base interactions. Adenine showed only 69 close interactions with sp^2 hybridized backbone atoms in this dataset. Figure 3.6 illustrates that threonine and asparagine show the largest comparative number of sp^2 hybridized backbone atom close contacts with adenine (15 and 14, respectively). For threonine, all of these contacts involve the carbonyl O_G atom. For asparagine 12 of the 14 close contacts involve the carbonyl O_{D1} atom.

Other than threonine and asparagine, there does not seem to be a preference among other amino acids that show sp^2 hybridized backbone atom close interactions with adenine. All the other amino acids that interact with adenine have very low numbers of close contacts with their sp^2 hybridized backbone atoms (<10 close contacts each). In addition, there is great diversity among the amino acids that show backbone interactions with adenine. This is even further indication that conformational factors are important contributors to the protein-DNA interactions.

Figure 3. 5: Sp^2 hybridized side-chain close contacts with adenine by amino acid in the major groove

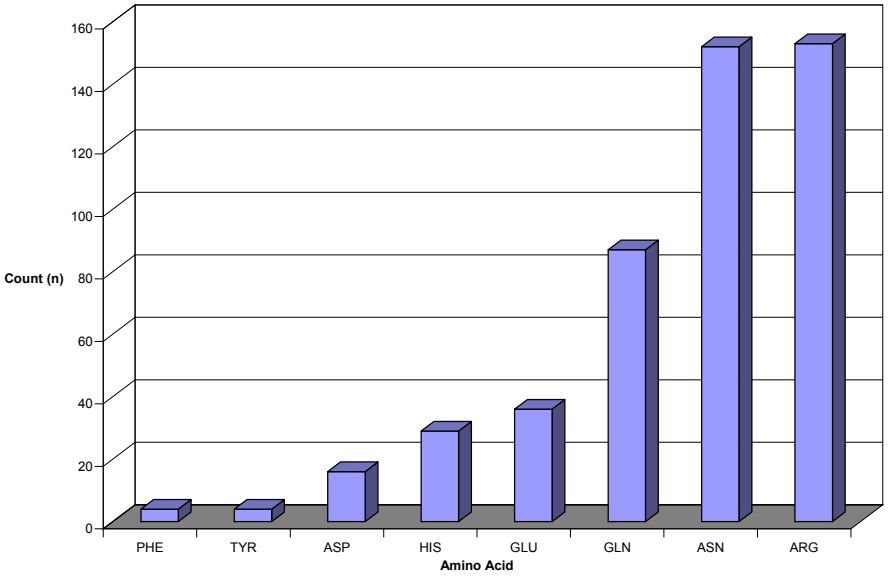
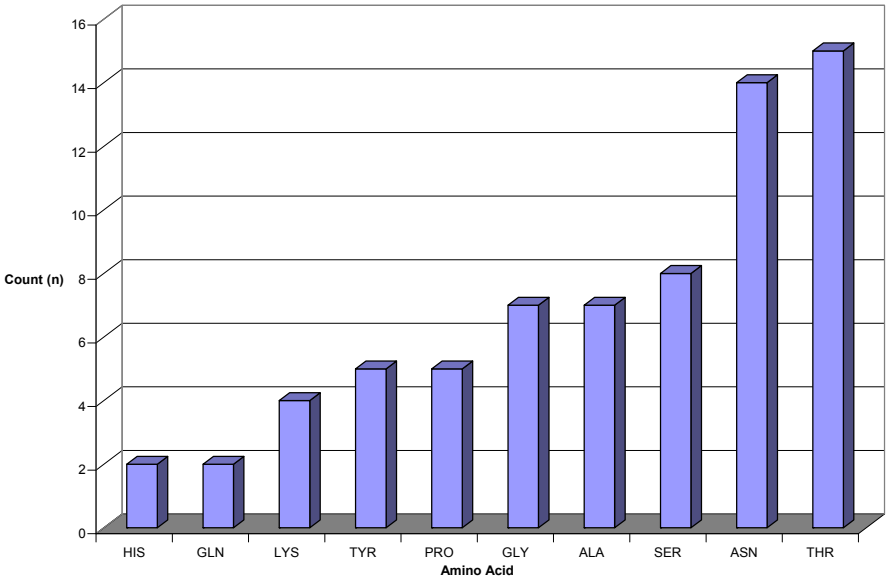


Figure 3. 6: Sp^2 hybridized backbone close contacts with adenine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^2 hybridized amino-acid side-chain and backbone atoms with adenine, classified by amino-acid type, within the dataset studied.
Overall adenine base numbers within research dataset: adenine = 5400

Source Data Appendix 9

3.3.3.2 Sp^3 Hybridized Atom Contacts with Adenine in the Major Groove

Sp^3 Hybridized Side-chain Atom Contacts

As shown in Figure 3.4, for sp^3 hybridized amino-acid interactions with adenine, the side-chain interactions dominate the backbone atom interactions. Figure 3.8 displays close contacts ($\leq$ 3.4 Å) of hybridized sp^3 amino-acid side-chain atoms with adenine atoms, classified by amino-acid type. This chart illustrates that a wide range amino acids have sp^3 hybridized side-chain atoms that interact closely with adenine. Fifteen amino acids demonstrate sp^3 hybridized side-chain atom close contacts with adenine in the major groove within the dataset studied.

Polar uncharged asparagine, glutamine, threonine and serine, as well as polar basic lysine show a comparatively high number of close contacts (132, 108, 62, 53 and 79 close contacts, respectively). As shown in Figure 3.2 most of the close contacts with sp^3 hybridized atoms occur with the highly electronegative N7 atom on adenine (298 close contacts); however, there is also a relatively significant number of contacts with N6 (109 close contacts). Due to the large number of close contacts, detailed information is tabulated. Table 3.1 describes the sp^3 contacts for each of the five amino acids noted previously.

Table 3. 1: Specific amino-acid atom sp^3 side-chain atom close contacts with adenine in the major groove

Amino Acid	Atom	Count
Asparagine	C_B	3
	N_{D2}	129
Glutamine	C_G	2
	N_{E2}	106
Lysine	C_G	2
	C_E	12
	C_D	13
	N_Z	52
Threonine	C_{G2}	5
	C_B	9
	O_{G1}	48
Serine	C_B	15
	O_G	38

Asparagine and glutamine represent an interesting study in terms of atomic hybridization. The nitrogen atom in these two amides has the potential for delocalizion, and therefore has both sp^3 and sp^2 character. Kemnitz and Loewen have estimated that the overall amide resonance hybrid is represented by resonance structure "A", 62% and by resonance structure "B", 28%, as described in Figure 3.7 [31]. This would equate to the potential for more sp^3 character for these two amino acids. Therefore the N_{D2} of asparagine and the N_{E2} of glutamine have been assigned sp^3 hybridization in this research database.

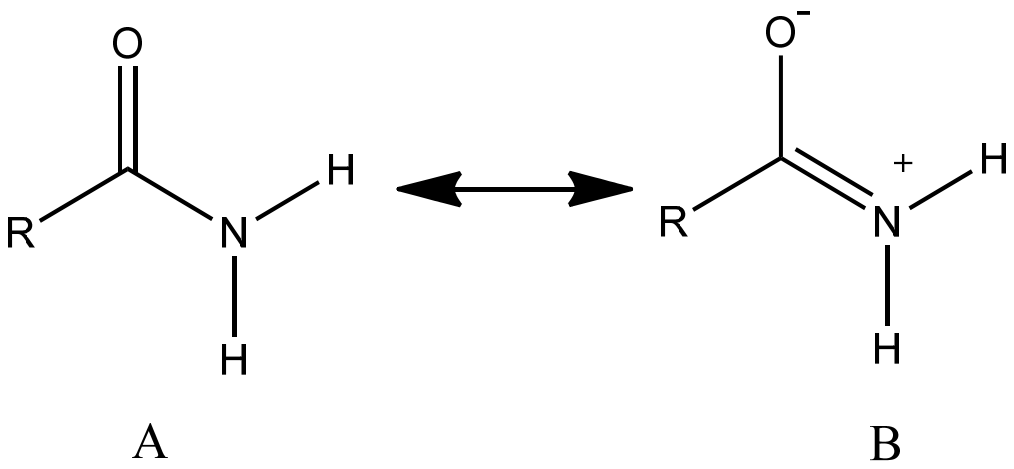
A review of close interactions between asparagine and adenine in this dataset reveals a comparatively high number of close interactions with electronegative N7 of adenine. Of the 129 total close contacts between adenine and sp^3 hybridized side chain atom, 81 are between N_{D2} of asparagine and N7 of adenine, and 22 are between N_{D2} and the N6 (good hydrogen-bond donor) of adenine. From these data it appears that the N_{D2} atom of asparagine within the majority of DNA-protein structures studied, does not act as

a hydrogen bond donor since there are far fewer interactions with N6 (hydrogen bond donor) than N7. A study of close interactions between glutamine and adenine from this dataset reveal and even stronger preference for interaction between N_{E2} and N7 of adenine. Of the 106 close contacts between sp^3 hybridized atoms on glutamine and adenine, 83 are between the N_{E2} and the N7 and only 6 between the N_{E2} and N6. These data provide support for the sp^3 hybridization assignments for N_{D2} and N_{E2}, since there is a strong preference for these atoms to interact closely with electronegative atoms on the nucleic acid bases in this dataset.

Sp^3 Hybridized Backbone Atom Contacts

Figure 3.9 displays close contacts (≤ 3.4 Å) involving sp^3 hybridized amino-acid backbone atoms with an atom on the adenine base in the major groove, classified by amino-acid type. This chart shows that close interactions involving backbone sp^3 hybridized atoms are almost non-existent. There are only 13 close contacts ≤ 3.4Å between sp^3 hybridized backbone atoms and adenine in the major groove. Within the dataset, only serine, glycine, histidine and glutamine show a small number of sp^3 backbone atom close interactions with adenine (6, 5, 1 and 1, respectively). As with the sp^2 interactions, there does not seem to be a clear electrostatic rationale for the discrimination and therefore steric and conformational factors may be acting to bring about the interactions. The small size of glycine and serine may make for an easier fit into the groove. Once again this is a good indication that there are some conformational aspects that are favorable for interaction in the major groove.

Figure 3. 7: Amide Resonance



A B

Figure 3. 8: Sp^3 hybridized side-chain close contacts with adenine by amino acid in the major groove

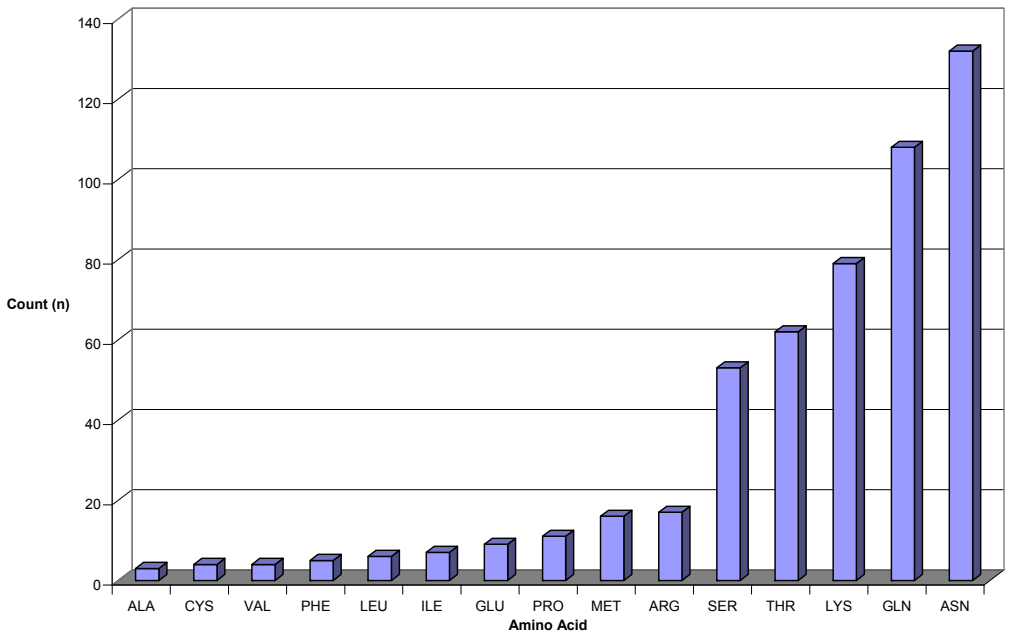
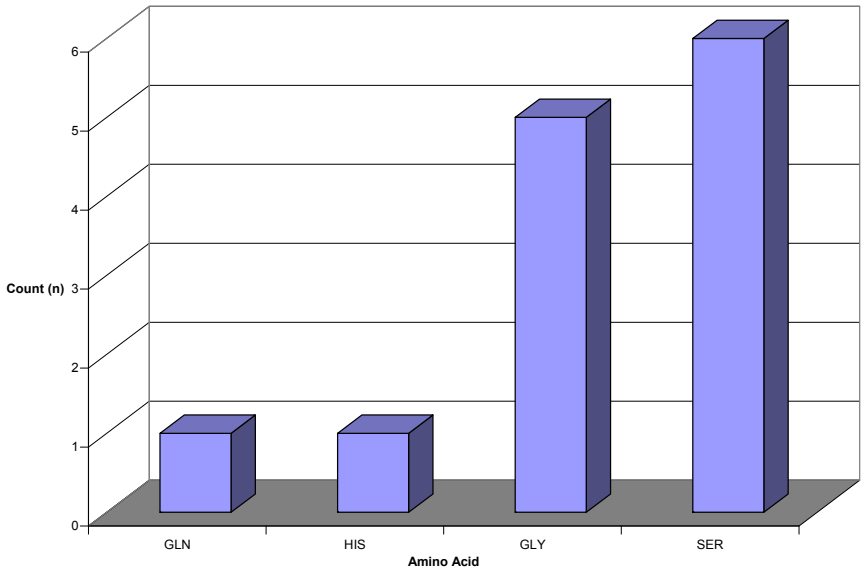


Figure 3. 9: Sp^3 hybridized backbone close contacts with adenine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^3 hybridized amino-acid side-chain and backbone atoms with adenine, classified by amino-acid type, within the dataset studied
Overall adenine base numbers within research dataset: adenine = 5400

Source data Appendix 9

3.3.3.3 Sp^2 Hybridized Atom Contacts with Thymine in the Major Groove

Sp^2 Hybridized Side-chain Aton Contacts

Figure 3.10 describes sp^2 hybridized amino-acid side-chain atom close contacts with thymine by amino-acid type in the major grove. As shown in this figure, the polar basic arginine shows the highest number of close interactions involving sp^2 hybridized side-chain atoms (201 close contacts ≤ 3.4 Å). Arginine represents an interesting amino acid for this study since it contains NH_2^+ and NH_2 atoms on its side chain. As noted in Section 1.5.2, since these nitrogen atoms are conjugated and cannot be distinguished from each other, both have been assigned a hybridization of sp^2. They are designated as N_{H1} and N_{H2} for this study. Therefore, arginine has a high potential for sp^2 hybridized side-chain atom contacts.

Within this dataset, the most common close interaction occurs between the sp^2 hybridized N_{H1} and N_{H2} atoms of arginine and thymine. These atoms interact closely with both the C5M group at the W1′ location of thymine and the O4 at the W2′ location. Because these atoms have delocalized electrons, arginine can take part in hydrogen-bonding interactions, as well as hydrophobic interactions, depending on the protein it is acting within. The number of close interactions (≤ 3.4 Å) with thymine is similar for the N_{H1} and N_{H2} atoms of arginine (82 and 100, respectively). As noted in Figure 3.2, the C5M is the dominant location on thymine for close contacts with sp^2 hybridized side-chain protein atoms. Specifically, 23 sp^2 hybridized side-chain atom close contacts arise from interaction between the N_{H1} and C5M atoms and 34 arise from interaction between the N_{H2} and C5M atoms. In addition, 33 sp^2 hybridized side-chain atom close contacts occur between the N_{H1} atom on arginine and the O4 atom on thymine, and 43 between the

N_{H2} atom on arginine and the O4 atom on thymine. These data support delocalization of electrons between the N_{H1} and N_{H2} with both atoms showing similar numbers of close interactions with both C5M and O4 of thymine.

From Figure 3.10, it can be seen that except for the sp^2 hybridized amino-acid side-chain atoms of arginine, there is not a dominant amino-acid type that interacts with thymine. In the major groove, other than the polar basic arginine, the additional amino-acid types that form close interactions between thymine and sp^2 hybridized side-chain atoms include, polar uncharged, polar acidic and non-polar hydrophobic. Therefore, beyond the electrostatic attraction, there appear to be other factors that govern these interactions in the major groove. Basic arginine, with its delocalized amine groups, appears to have a very advantageous structure, both sterically and electronically, for close interactions with the thymine base.

Sp^2 Hybridized Backbone Atom Contacts

As shown in Figure 3.11, thymine interacts to a very limited extent with sp^2 hybridized backbone atoms. Among the 499 structures studied, there are only 63 close contacts that include sp^2 hybridized backbone atoms. Although alanine shows the highest number of sp^2 hybridized close interactions, there does not seem to be a strong preference for one type of amino acid with thymine. These interactions include polar uncharged, polar basic, polar acid and non-polar amino acids. Seventeen amino acids have sp^2 backbone atoms that show a small number of close contacts with thymine. Therefore it is clear there is little specificity thymine shows very little among the sp^2 hybridized backbone atoms

Figure 3. 10: Sp^2 hybridized side-chain close contacts with thymine by amino-acid in the major groove

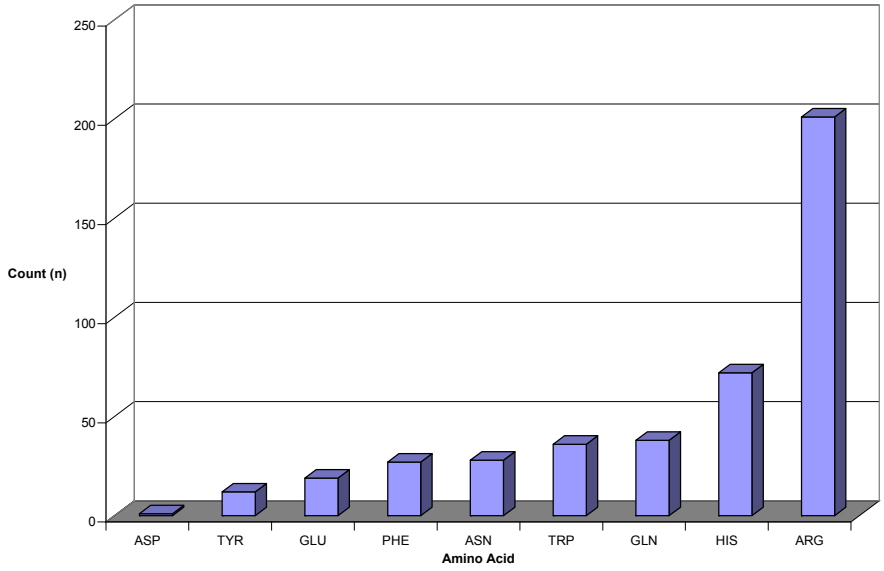
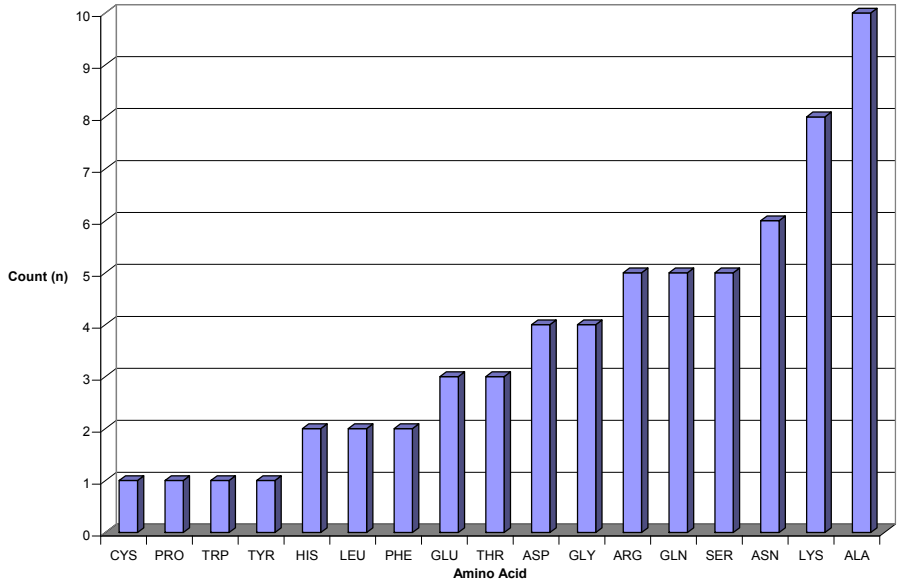


Figure 3. 11: Sp^2 hybridized backbone close contacts with thymine by amino-acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^2 hybridized amino-acid side-chain and backbone atoms with thymine, classified by amino-acid type, within the dataset studied
Overall thymine base numbers within research dataset: thymine = 5433

Source data Appendix 10

3.3.3.4 Sp^3 Hybridized Atom Contacts with Thymine in the Major Groove

Sp^3 Hybridized Side-chain Atom Contacts

Figure 3.12 describes sp^3 hybridized amino-acid side-chain atom close contacts with thymine by amino-acid type in the major groove. For sp^3 hybridized atoms closely interacting ($\leq$ 3.4 Å) with thymine, the side-chain counts are far greater than the backbone counts. This is potentially due to the W2' carbonyl on thymine closely interacting with sp^3 hybridized amine atoms. From Figure 3.2 it can be seen that the greatest number of close contacts with sp^3 hybridized atoms arise from interactions with the O4 atom in the W2′ position on thymine.

Threonine, lysine and glutamine show a very high number of sp^3 hybridized side-chain atom close contacts with thymine in the major groove, within this dataset (105, 96 and 95, respectively). For threonine, 67 of the 105 close contacts involve the O_{G1} atom with thymine. Of the 67 interactions, 30 are between the O_{G1} atom on threonine and the O4 atom on thymine and 27 are between the O_{G1} atom on threonine and the C5M on thymine. The dominant close interaction with lysine is between the NH_3^+ (N_Z) and thymine. There are 62 close contacts between the N_Z atom on the side chain and thymine, and 56 of these involve close interaction with the O4 atom on thymine. For glutamine, there are 77 close interactions between the N_{E2} atom and thymine, 65 of which involve close interaction with the O4 atom on thymine.

Arginine also shows relatively high number of close interaction between sp^3 hybridized side-chain atoms and thymine. For arginine, there are 86 close interactions between sp^3 side-chain atoms and thymine, 37 of which are with the N_E atom and 30 of

which are with the C_G atom. Of the 37 close interactions between the N_E atom and thymine, 16 are with the O4 atom and 19 are with C5M.

Sp^3 Hybridized Backbone Atom Contacts

As shown in Figure 3.13 thymine interacts to a very limited extent with sp^3 hybridized amino-acid backbone atoms in the major groove. Among the 499 structures studied, there are only 29 close contacts (≤ 3.4 Å) involving sp^3 hybridized backbone atoms from amino acids. Glycine shows the highest number of close contacts with thymine involving sp^3 hybridized backbone atoms (18 close contacts). Although this is the largest comparative number, it is still a very small absolute number of contacts. The small size of the glycine probably aids its ability to allow backbone atoms to interact closely in the major groove. Asparagine also shows a small number of close contacts involving sp^3 hybridized backbone atoms (4 close contacts).

Figure 3. 12: Sp^3 hybridized side-chain close contacts with thymine by amino acid in the major groove

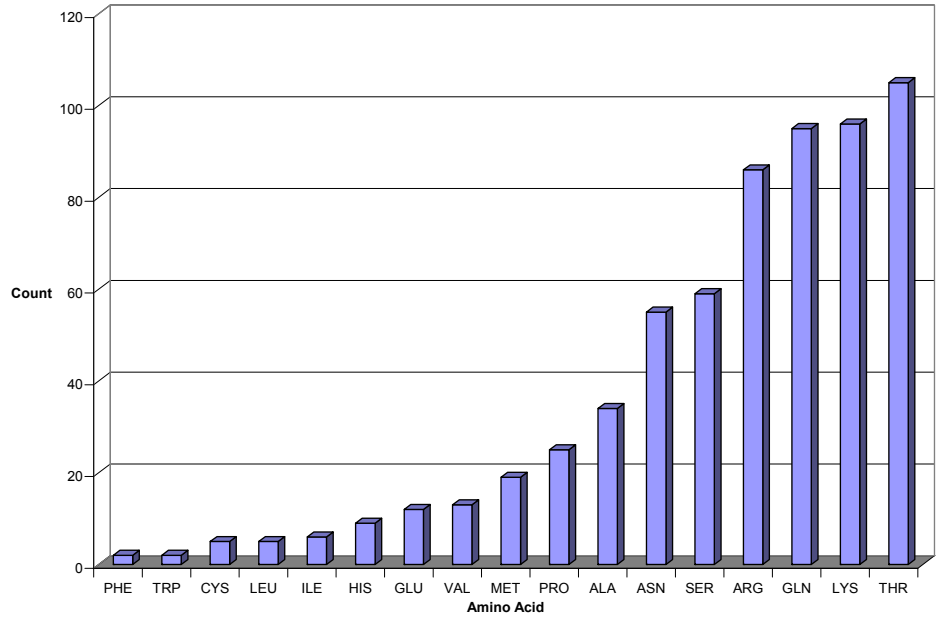
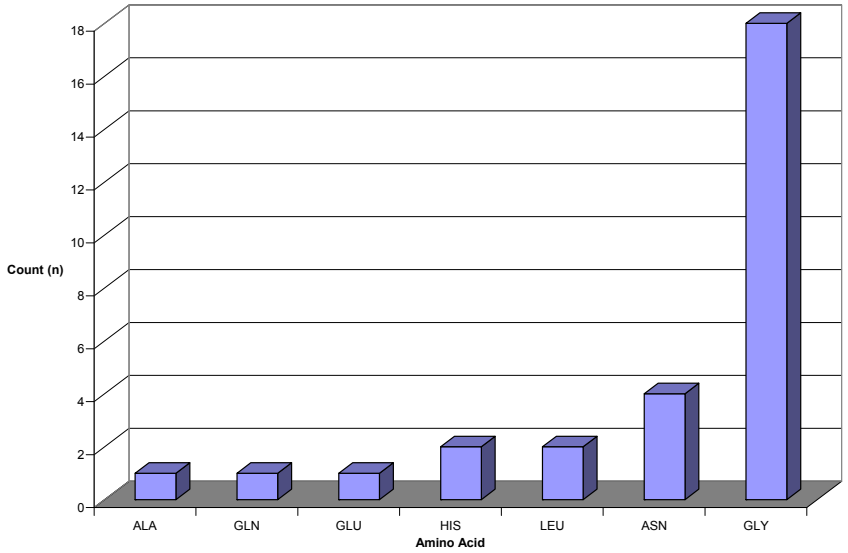


Figure 3. 13: Sp^3 hybridized backbone close contacts with thymine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^3 hybridized amino-acid side-chain and backbone atoms with thymine, classified by amino-acid type, within the dataset studied
Overall thymine base numbers within research dataset: thymine = 5433

Source data Appendix 10

3.3.3.5 Sp^2 Hybridized Atom Contacts with Guanine in the Major Groove

Sp^2 Hybridized Side-chain Atom Contacts

Figure 3.14 shows specific amino-acid sp^2 hybridized side-chain atom close contacts in the major groove with guanine among the 499 non-redundant structures studied. Within this dataset, arginine shows a very high number of sp^2 hybridized side-chain atom close contacts with guanine. This is consistent with the Seeman model in which arginine recognizes guanine [8]. For guanine, the distribution of sp^2 side-chain interactions is dominated by the extremely large number of close interactions ($\leq$ 3.4 Å) with arginine (1323 out of 1674). Once again, this is not surprising since arginine has two potential sp^2 hybridized nitrogen atoms that are part of the guanidinium group. Of the 1323 close interactions that involve arginine and guanine, 512 involve the N_{H1} atom of arginine and 737 involve the N_{H2} atom of arginine. As previously noted, the arginine side chain contains delocalized electrons, and therefore the nitrogen atoms have both been assigned an sp^2 hybridization, since only one assignment per atom can be made in the database. As described in the subsequent paragraph, once again it is discovered that these atoms act equivalently, and therefore cannot readily be distinguished from one another. Although not addressed in this research, a look at specific DNA-protein structures and their interaction with arginine, is an interesting subject for further investigation and may further elucidate the nature of the bonds.

A detailed look at specific interactions that involve the O6 atom on guanine shows that both N_{H1} and N_{H2} interact closely with this atom. Refer to Figure 1.13 for an illustration of arginine's structure and Figure 1.14 for an illustration of its resonance structures. There are 214 interactions between the N_{H1} atom of arginine and the O6 atom

of guanine and 315 close contacts between the N_{H2} atom of arginine and the O6 atom of guanine. These close contacts potentially involve a hydrogen bond between the N_{H2} and the O6. As shown in Figure 1.12a, the cationic guanidinium of arginine has the ability to donate hydrogen atoms. In addition to interactions with the O6 atom of guanine, N_{H1} and N_{H2} atoms also show a large number of close contacts with the N7 atom of guanine. Among the 499 structures studied, there are 266 close interactions between N_{H1} and the N7 atom of guanine and 319 interactions between N_{H2} and the N7 atom of guanine. Since the N7 atom of guanine is electronegative, the high number of close interactions with both N_{H1} and N_{H2} also indicate an attraction to NH_2^+, and therefore add additional support for the sp^2 hybridization assignment.

Histidine shows the second highest number of sp^2 hybridized close side-chain interactions with guanine (258 close contacts). Within the dataset studied, histidine is the only amino acid other than arginine that has >100 close contacts with guanine. Histidine contains a cyclic component, the imidazole ring, in its side chain. The imidizole ring contains delocalized electrons and has a high potential for sp^2 hybridized atom interactions. For histidine, within the dataset studied there are 50 close contacts with N_{D1}, 91 with N_{E2}, 93 with C_{E1} and 14 with C_{D2}. It appears the C_{E1} and N_{E2} atoms toward the end of the side chain, have the highest number of close interactions. This may indicate that the conformation of the imidizole may play a large part in the interaction of histidine with DNA bases within the major groove. Among the other amino-acid interactions discovered, all others show a very small number of close sp^2 hybridized side-chain atom interactions with guanine.

Sp^2 Hybridized Backbone Atom Contacts

As shown in Figure 3.15, guanine interacts to a limited extent with sp^2 hybridized amino-acid backbone atoms in the major groove. Among the 499 structures studied, there are only 179 close contacts (≤ 3.4 Å) involving sp^2 hybridized backbone atoms from amino acids with guanine. Not surprisingly, glycine and serine show the largest number of sp^2 hybridized backbone atoms closely interacting with guanine (53 and 41, respectively). Their small size makes them favorable for close backbone interactions with the bases.

Glycine shows the largest number of sp^2 hybridized backbone close contacts with guanine (50 close contacts <3.4 Å). These close contacts involve the N atom in the peptide bonds of proteins containing glycine, as well as the carbonyl oxygen atom. Among the 53 total close contacts with glycine sp^2 hybridized backbone atoms, 37 are between the N in glycine and guanine and 16 close contacts are between the carbonyl oxygen atom and glycine.

Serine also shows a comparatively large number of sp^2 hybridized backbone close contacts with guanine (41 close contacts ≤ 3.4 Å) in the major groove. These close contacts involve the N atom in the peptide bond, as well as the carbonyl oxygen atom. Among the 41 total contacts with serine sp^2 hybridized backbone atoms, 12 are between the N atom in serine and guanine and 27 close contacts are between the carbonyl oxygen atom and serine. Within this dataset, there were 2 close interactions involving the sp^2 hybridized atoms of the carbonyl carbon in the backbone of serine and the guanine base.

Figure 3. 14: Sp^2 hybridized side-chain close contacts with guanine by amino acid in the major groove

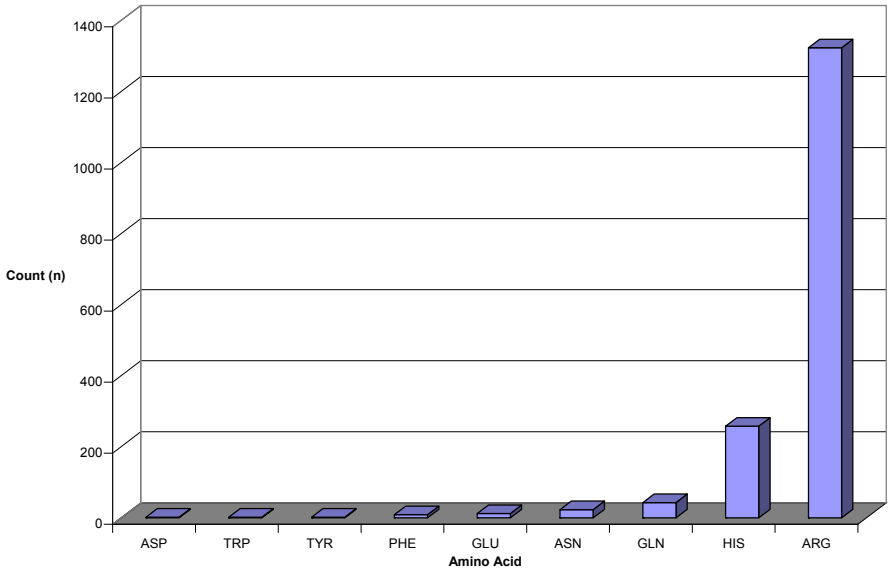
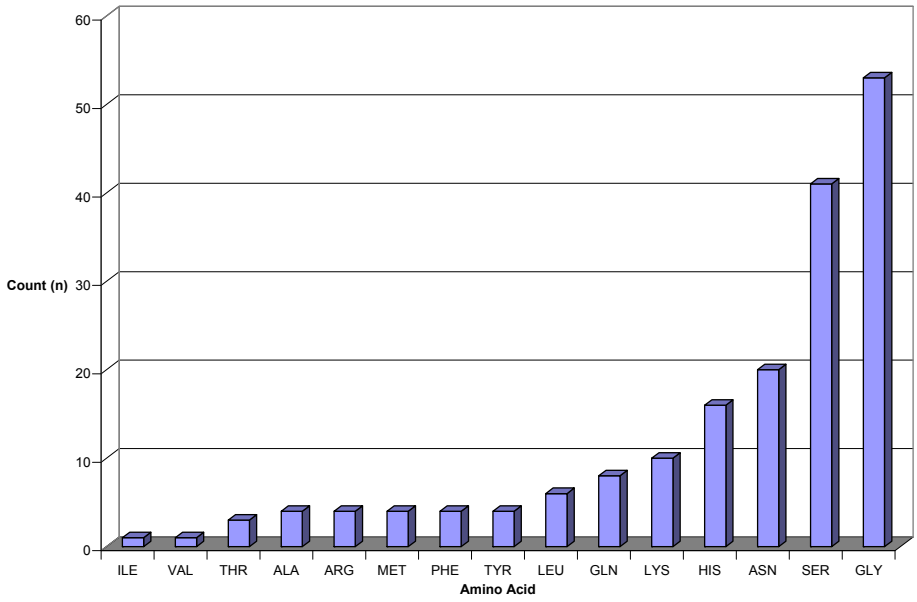


Figure 3. 15: Sp^2 hybridized backbone close contacts with guanine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^2 hybridized amino-acid side-chain and backbone atoms with guanine, classified by amino-acid type, within the dataset studied
Overall guanine base numbers within research dataset: guanine = 5399

Source data Appendix 11

3.3.3.6 Sp^3 Hybridized Atom Contacts with Guanine in the Major Groove

Sp^3 Hybridized Side-chain Atom Contacts

Among the 499 protein-DNA structures studied, guanine shows the highest comparative number of sp^3 hybridized side-chain atom close contacts (≤ 3.4 Å) of the four bases studied. There are a total of 1291 close contacts between sp^3 hybridized side-chain atoms and the guanine base. As shown in Figure 3.16, these sp^3 hybridized side-chain atom close contacts are dominated by interaction with lysine.

From Figure 3.16 it can be seen that lysine shows the highest comparative number of sp^3 hybridized close side-chain atom contacts (≤ 3.4 Å) with guanine. There are 686 close contacts between guanine and lysine sp^3 hybridized side-chain atoms. This is consistent with Mandel-Gutfreund's findings that lysine interacts to a large extent with guanine [13]. Out of the total 686 sp^3 hybridized close contacts with lysine, 555 involve the N_Z atom within the NH_3^+ group. Once again these interactions are dominated by close contacts between with the electron rich N7 and O6 atoms on guanine, in contact with the N_Z atom of the positively charged NH_3^+ group on lysine. There are 211 close contacts between the N_Z atom of lysine and the N7 atom of guanine, as well as 261 close contacts between the N_Z atom of lysine and the O6 atom on guanine. There are also a comparatively large number of close contacts between guanine and the C_E atom of lysine (109 close contacts). There are 57 close contacts between the C_E atom and the O6 atom and 47 close contacts between the C_E atom and N7 atom.

Within this dataset, arginine shows the second highest number of sp^3 hybridized side-chain atom close contacts, with a total of 289 close contacts with guanine. The arginine sp^3 hybridized close interactions are dominated by contacts with the N_E atom in

the side chain. Out of the total 289 sp^3 hybridized close contacts with arginine, 176 involve the N_E atom on its side chain. As expected, the N_E atom interacts to a large extent with the electron rich N7 and the carbonyl O6 atoms on guanine. As shown in Figure 3.3, in the major groove the primary amino-acid sp^3 close contacts are with the N7 and O6 atoms on guanine. There are 67 close contacts between the N_E atom and N7 on guanine, and 96 close contacts between the N_E and the O6 atom on guanine. These results lend support to the N_E atom acting as a hydrogen bond donor. As shown in Figure 1.13, although the N_E atom of arginine has been assigned a hybridization of sp^3, due to the potential for delocalization, it may have some sp^2 character. As previously discussed, the guanidinium functional group can act as a hydrogen donor.

Serine and asparagine also show a moderate number of close sp^3 hybridized side-chain atom close interactions with guanine (11 and 85 close contacts, respectively). The OH group on serine can act as a hydrogen bond donor. In addition, the relatively small size of serine makes it both electronically and sterically favorable for interactions in the major groove. For serine, of the 119 close contacts with sp^3 hybridized side-chain atoms, there are 84 that involve its O_G atom. For asparagine, the N_{D2} atom within the amine provides a good hydrogen bond donor for interactions with O6 atom of guanine. For asparagine, of the 85 close contacts with sp^3 hybridized side-chain atoms, there are 76 between the N_{D2} atom of asparagine and the O6 atom of guanine. Refer to Figure 1.12 for the structures of serine and asparagine.

Sp^3 Hybridized Backbone Atom Contacts

As previously noted, although guanine shows the highest number of overall close contacts, of the four bases studied, within this dataset there are only a small absolute

number of sp^3 hybridized backbone atom close contacts with guanine. Figure 3.4 shows that guanine experiences a total of 54 close contacts ($\leq$ 3.4 Å) with sp^3 hybridized backbone atoms.

As described in Figure 3.17, only glycine, serine, lysine and leucine show sp^3 hybridized backbone atom close contacts with guanine (37, 11, 4 and 2, respectively). All of the close interactions are with the sp^3 hybridized C_A atom in the backbone of these amino acids. Because of the very small size of glycine, it more easily fits into the groove so that its backbone can come into close contact with the base-pair atoms.

Figure 3. 16: Sp^3 hybridized side-chain close contacts with guanine by amino acid in the major groove

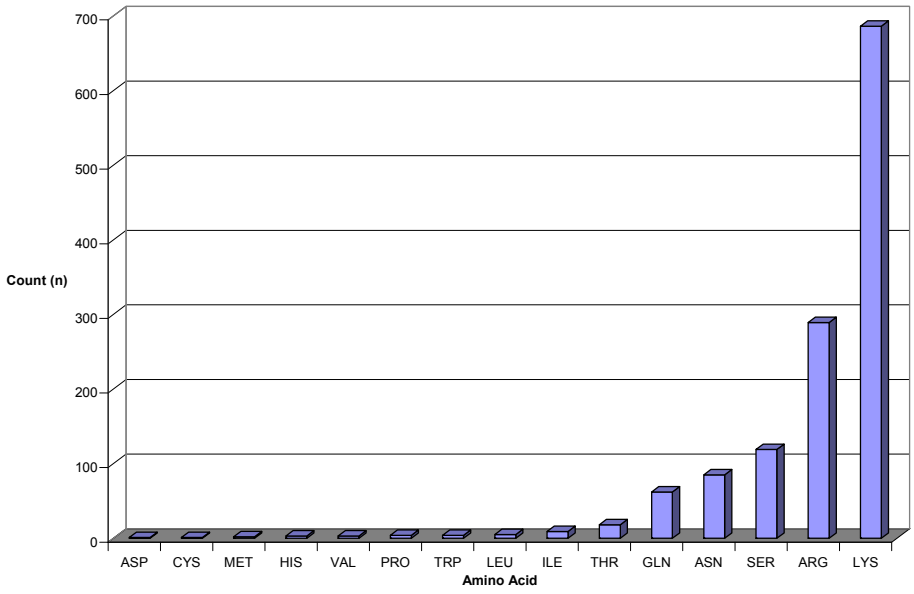
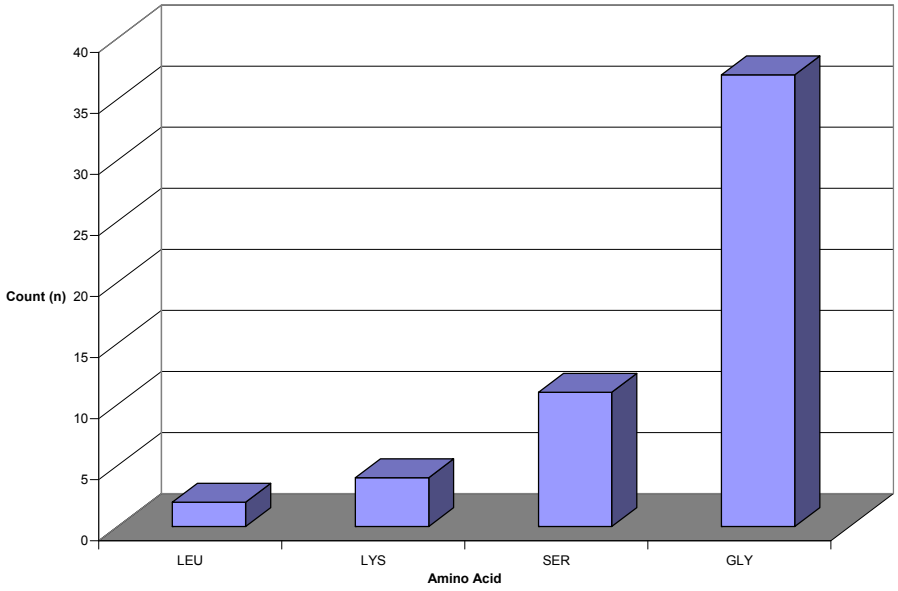


Figure 3. 17: Sp^3 hybridized backbone close contacts with guanine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^2 hybridized amino-acid side-chain and backbone atoms with adenine, classified by amino-acid type, within the dataset studied
Overall guanine base numbers within research dataset: guanine = 5399

Source data Appendix 11

3.3.3.7 Sp^2 Hybridized Atom Contacts with Cytosine in the Major Groove

Sp^2 Hybridized Side-chain Atom Contacts

As shown in Figure 3.4 cytosine experiences a somewhat higher number of close contacts (≤ 3.4Å) with sp^2 hybridized side-chain atoms than adenine and thymine, but much fewer than its Watson-Crick partner guanine (751 close contacts). As described in Figure 3.18, cytosine shows the greatest number of sp^2 hybridized side-chain atom close contacts (≤ 3.4Å) with glutamic acid, arginine and aspartic acid.

Figure 3.18 illustrates that 230 of the 678 total sp^2 hybridized side-chain atom close contacts are between cytosine and glutamic acid. This is consistent with the findings of Madel-Gutfruend *et al*. which indicated that cytosine interacts with glutamic acid [13]. The close interactions with cytosine are predominantly with the carbonyl (O_{E1} and O_{E2}) on glutaminc acid (107 and 103 contacts, respectively out of the total 230). As described in Section 1.5.2, because of the possible delocalized electrons within the carbonyl, both O_{E1} and O_{E2} have been designated as sp^2 hybridized. Close interactions with cytosine are dominated by the carbonyl oxygen (O_{E1} and O_{E1}) on the glutamic acid side chain and the N4 atom of the NH_2 group in the W2 position of cytosine, as well as the electron rich C5 atom of cytosine. The O_{E1} atom on the glutamic acid side chain makes 67 of its 99 sp^2 hybridized close contacts with the N4 atom and 35 with the C5 atom on cytosine. The O_{E2} atom on the glutamic acid side chain makes 34 sp^2 hybridized close contacts with the N4 atom and 45 with the C5 atom on cytosine.

As shown in Figure 3.18, aspartic acid and arginine also have a relatively high comparative number of close sp^2 hybridized side-chain atom interactions with cytosine atoms (203 and 159 close contacts, respectively). For aspartic acid these interactions are

primarily between the O_{D1} and O_{D2} atoms (66 and 86 respectively) and cytosine. Because of the possible delocalized electrons within the carbonyl, both the O_{D1} and O_{D2} atoms have been designated as sp^2 hybridized. Close interactions with cytosine are dominated by the carbonyl oxygen (O_{D1} and O_{D1}) on the aspartic acid side chain and the N4 atom of the NH_2 group at the W2 position of cytosine, as well as the electron rich C5 atom. The carbonyl (O_{D1}) atom on the aspartic acid side chain makes 46 of its 66 sp^2 hybridized close contacts with the N4 atom and 17 with the C5 atom on cytosine. The carbonyl (O_{D2}) atom on aspartic acid side chain has 71 of its 86 sp^2 hybridized close contacts with the N4 atom and 15 with the C5 atom on cytosine.

As noted above, arginine sp^2 hybridized side-chain atoms also show a comparatively high number of close interactions with cytosine in the major groove. These close interactions occur predominantly with the N_{H1} and N_{H2} atoms on arginine (64 and 105, respectively out of the 203). As described in Section 1.5.2, because of the possible delocalized electrons within the amine, both N_{H1} and N_{H2} have been designated as sp^2 hybridized. Close interactions with cytosine are dominated by the amine (N_{H1} and N_{H2}) on the arginine side chain and the N4 atom at the W2 position of cytosine. From these data, there does not seem to be clear favoritism for a particular atom on cytosine. Table 3.2 shows the number of close contacts between the N_{H1} and N_{H2} atoms on arginine and the various cytosine atoms in the major groove.

Table 3. 2: Cytosine interactions with arginine in the major groove

Arginine Atom	Cytosine Atom	Count
N_{H1}	N4	34
	C4	9
	C5	14
	C6	5
N_{H2}	N4	32
	C4	3
	C5	37
	C6	28

Sp^2 Hybridized Backbone Atom Contacts

As described in Figure 3.4, among the four bases studied, cytosine shows the highest comparative number of close contacts (≤ 3.4Å) in the major groove with sp^2 hybridized amino-acid backbone atoms within this dataset. As noted in Section 1.5.2, the backbone of all proteins contains two important sp^2 hybridized atoms. These are the carbonyl oxygen (O) and the nitrogen of the peptide bond (N). Figure 3.19 shows that the major sp^2 hybridized backbone atom close interactions with cytosine within this research data set involve asparagine, alanine, threonine, arginine, lysine, glycine and serine (63, 39, 31, 31, 28, 27 and 26, respectively). All of these amino acids show the carbonyl oxygen in the backbone being the primary sp^2 hybridized backbone atom involved in the interaction with cytosine. For asparagine, alanine, threonine, lysine, arginine, serine and glycine there are respectively 62, 37, 31, 28, 22, 26 and 23 close contacts with cytosine involving the carbonyl O of the backbone. However, there is a small variation within this group in terms of the cytosine atom that the carbonyl oxygen closely interacts with.

All of the closely interacting amino acids described in the preceding paragraph show close contacts to a comparatively large extent with the N4 atom in the W2 position on cytosine. However, there are differences in the extent to which the amino acids interact with the C4, and C5 atoms on cytosine. Of the amino acids that interact with

cytosine there is only a limited amount of sp^2 hybridized backbone atom close contacts with the C6 atom on cytosine. Figure 3.20 shows specific close contact counts among the primary four cytosine atoms interacting with proteins in the major groove (C4, C5, C6 and N4). Figure 3.21 describes the proportion of interactions with these atoms. Each of the closely interacting amino acids is included in this chart, along with the percent each cytosine atom (C4, C5, C6 and N4) contributes to the total number of close contacts with given amino acid.

From Figure 3.20, asparagine and alanine show the highest absolute number of close contacts between sp^2 hybridized backbone atoms and the N4 atom on cytosine (33 and 27 close contacts, respectively). From Figure 3.20, serine shows the highest absolute number of close contacts between sp^2 hybridized backbone atoms and the C4 atom on cytosine (12 close contacts). As shown in Figure 3.21, serine shows the highest percentage of its total close contacts with cytosine with the C4 atom. Conversely, serine shows no interaction with the C5 atom of cytosine. Asparagine shows the highest number of close interactions between the sp^2 hybridized backbone atoms and the C5 atom of cytosine (18 close contacts).

Figure 3. 18: Sp^2 hybridized side-chain close contacts with cytosine by amino acid in the major groove

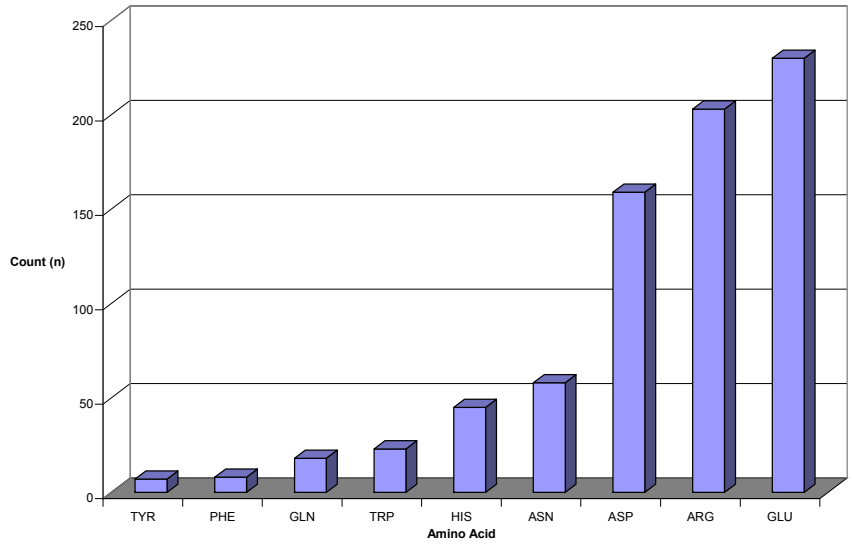
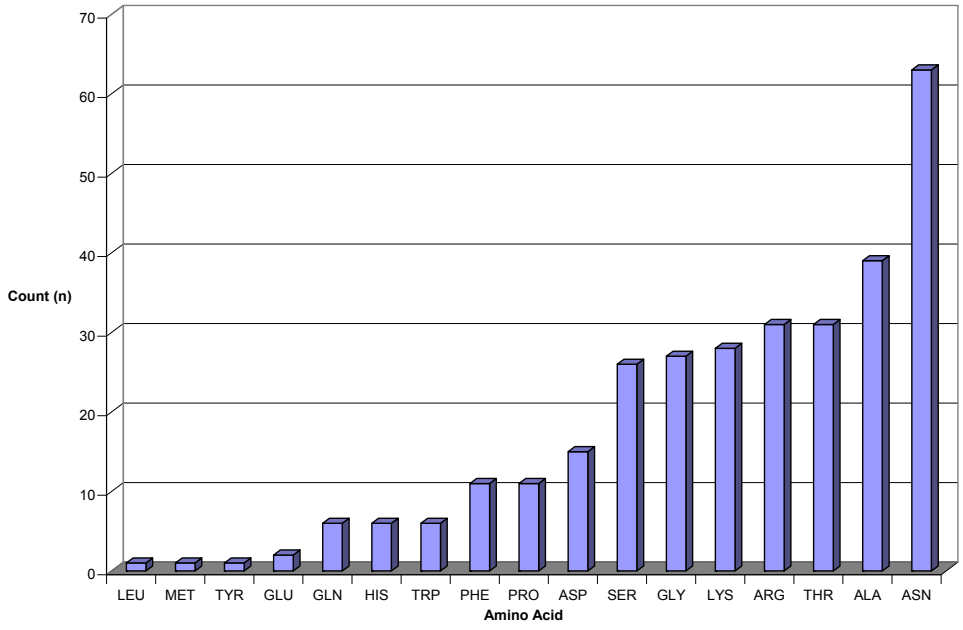


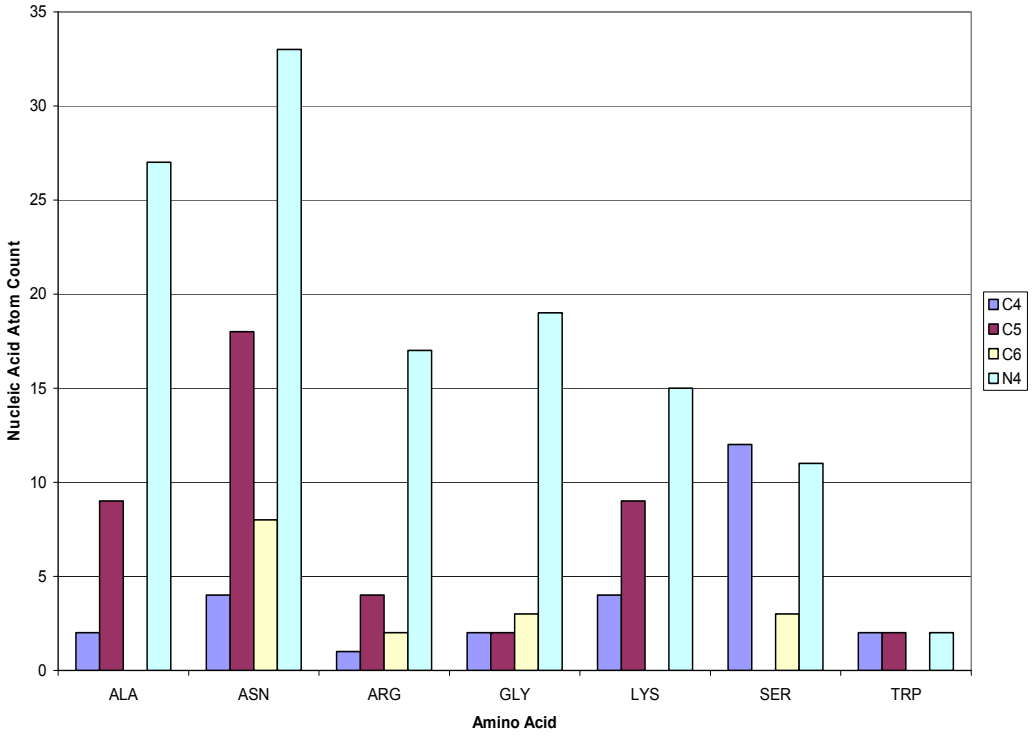
Figure 3. 19: Sp^2 hybridized backbone close contacts with cytosine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^2 hybridized amino-acid side-chain and backbone atoms with cytosine, classified by amino-acid type, within the dataset studied
Overall cytosine base numbers within research dataset: cytosine = 5375

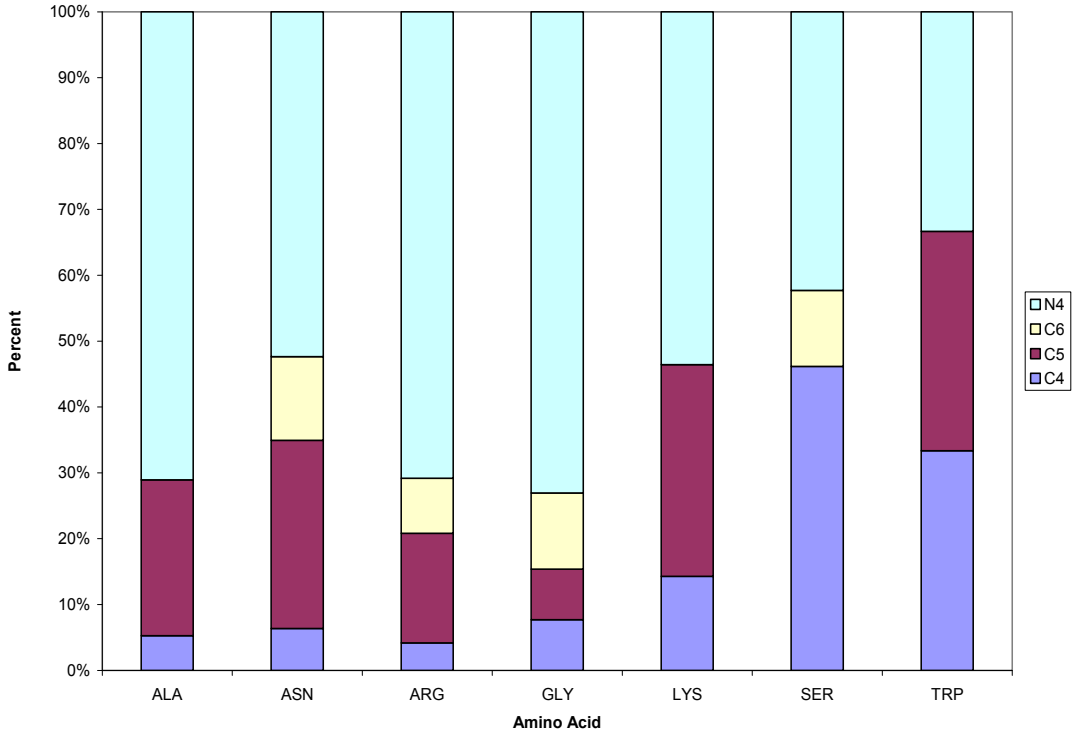
Source data Appendix 12

Figure 3. 20: Sp^2 hybridized backbone atom close contacts with specific cytosine atoms in the major groove



Source data Appendix 12

Figure 3. 21: Proportion of Sp^2 hybridized backbone atoms in close contact with cytosine atoms in the major groove



Source data Appendix 12

3.3.3.8 Sp^3 Hybridized Atom Contacts with Cytosine in the Major Groove

Sp^3 Hybridized Side-chain Atom Contacts

As shown in Figure 3.4, in the major groove cytosine exhibits the lowest comparative count of close contacts (≤ 3.4Å) with sp^3 hybridized amino-acid side-chain atoms among the four bases within the dataset examined (283 close contacts). As described in Figure 3.22, cytosine exhibits low specificity for sp^3 hybridized side chain close contacts with the amino acids present in this research data set. Furthermore, cytosine shows the highest numbers of close contacts with the sp^3 hybridized side-chain atoms of threonine, serine, asparagine and arginine side chains.

As shown in Figure 3.22, threonine and serine exhibit 56 and 50 close contacts, respectively, between sp^3 side-chain atoms and cytosine. These data show that serine and threonine interact closely with cytosine almost exclusively through the oxygen atom of OH group. For serine this is the O_G atom (42 close contacts) and for threonine it is the O_{G1} atom (52 close contacts). For threonine there are 33 close contacts between its O_{G1} atom and the N4 atom of cytosine, and 19 close contacts between its O_{G1} atom and the C5 atom of cytosine. For serine there is less dominance toward one site of interaction. For serine there are 20 close contacts between its O_G atom and the N4 atom of cytosine, and 18 close contacts between its O_G atom and the C5 atom of cytosine.

Asparagine and arginine also show a modest number of close contacts between sp^3 hybridized side-chain atoms and cytosine (47 each). For asparagine, as noted in Section 1.5.2, the N_{D2} atom in its side chain has the potential for delocalization. For asparagine, 43 of the 47 close contacts involve the N_{D2} atom. Moreover, the largest comparative number of these close contacts occurs between the N_{D2} atom of asparagine

and the C5 atom on cytosine. There are 23 out of 43 close contacts between the N_{D2} atom of asparagine and C5 of cytosine. There are comparatively few (11) close contacts between the N_{D2} atom of asparagine and N4 atom of cytosine.

For arginine, the sp^3 hybridized side-chain atom close contacts involve the N_E and C_D atoms (21 and 23, respectively). The N_E atom shows close contact primarily with the C5 atom on cytosine (15 of the 21 close contacts). The C_D atom interacts closely with C5 and N4 to an equal extent (9 close contacts each). For arginine there are also 4 close contacts between its C_D atom and the C4 atom on cytosine, as well as 1 close contact between its C_D atom on arginine and C6 on cytosine.

Sp^3 Hybridized Backbone Atom Contacts

As shown in Figure 3.23, in the major groove cytosine exhibits an extremely low number of close interactions ($\leq$ 3.4Å) with sp^3 hybridized backbone atoms on proteins in the dataset studied. For the 499 structured investigated, there were only 11 total close contacts with sp^3 hybridized backbone atoms. Arginine shows the highest comparative number of close interaction involving sp^3 hybridized backbone atoms with cytosine, as described in Figure 3.22 (6 close contacts). Glycine exhibits 3 close contacts involving sp^3 hybridized backbone atoms. Two of these close contacts are between the C_A atom on its backbone and cytosine. Glutamine only has 2 close contact between sp^3 backbone hybridized atoms and cytosine.

Figure 3. 22: Sp^3 hybridized side-chain close contacts with cytosine by amino acid in the major groove

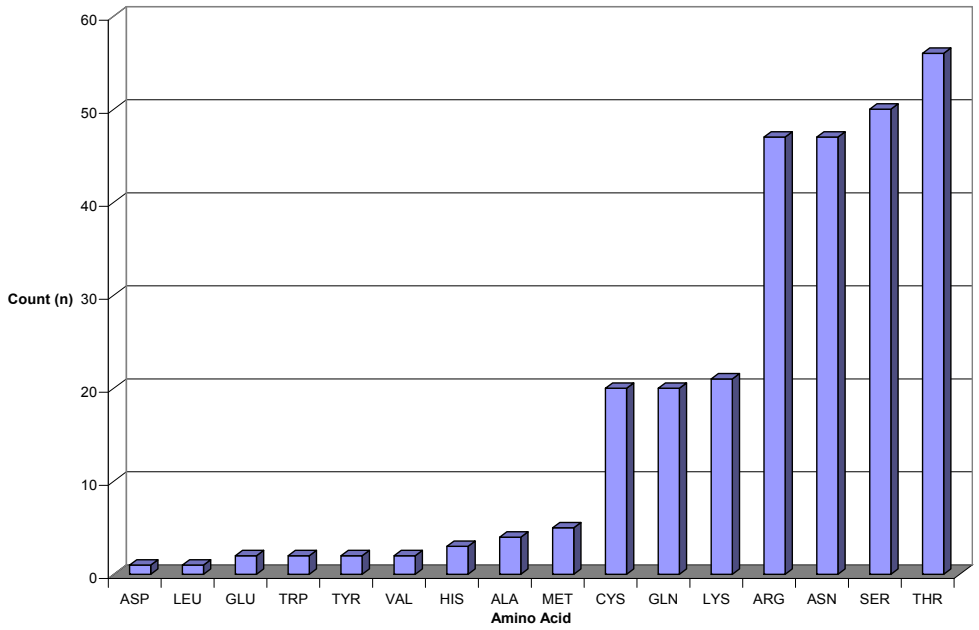
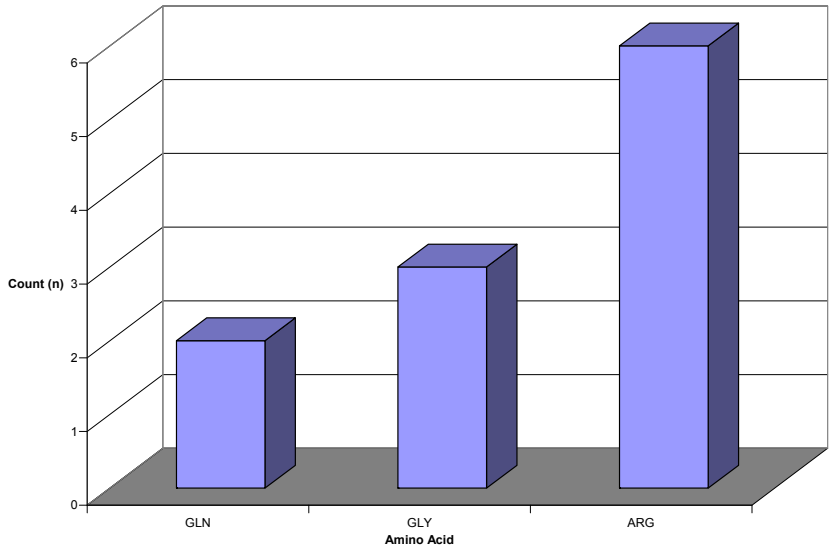


Figure 3. 23: Sp^3 hybridized backbone close contacts with cytosine by amino acid in the major groove



These graphs displays close interactions (≤ 3.4 Å) by sp^3 hybridized amino-acid side-chain and backbone atoms with cytosine, classified by amino-acid type, within the dataset studied
Overall cytosine base numbers within research dataset: cytosine = 5375

Source data Appendix 12

3.4 Major Groove Close Contacts Summary Discussion

The following summary represents information for DNA-protein close contacts (≤3.4Å) obtained from the dataset (499 representative structures) studied in this research**.**

3.4.1 General Comparison of Close Contacts

In the major groove, as shown in Figure 3.1, within this dataset, the G-C base pair exhibits a much higher absolute number of close contacts (≤ 3.4Å) than the A-T base pair. In addition, the G-C base pair shows a greater imbalance between sp^2 and sp^3 hybridized atom close-contacts compared to the A-T base pair. As can be seen in Figure 1.1, for the G-C base pair there are 3 potential hydrogen bonds between the bases, but for the A-T base pair, there are only 2 potential hydrogen-bonding sites between the bases. In the minor groove, the adenine S2 site is occupied by a CH group, in contrast to the guanine S2 position occupied by NH_2. Although the S sites are located in the minor groove, it could be postulated that the differences in hydrogen bonding could cause distinct differences in conformational flexibility, and this could in turn contribute to differences in environment in the major groove.

As observed in Figure 3.2 and Figure 3.3, for both base-pair sets (A-T and G-C), the carbonyl (O4 atom of thymine, O6 atom of guanine) and amine (N6 atom of adenine, N4 atom of cytosine) functional groups contribute significantly to the number of close contacts experienced in the major groove. These findings are not surprising since these atoms are potential hydrogen-bonding sites. The O6 of guanine exhibits a greater number of close contacts than the O4 of thymine. This includes overall close contacts as well as close contacts with sp^2 hybridized side-chain atoms. By far the greatest number of close contacts with sp^2 hybridized side-chain atoms for both guanine and thymine are with

arginine. As discussed, arginine has a unique structure in which the atoms of the guanidinium group have delocalized electrons, therefore both amines exhibit both sp^2 and sp^3 character, almost equally. These nitrogen atoms of the guanidinium group, exhibit distinct differences with regard to close contacts with the N6 atom of adenine verses the N4 atom of cytosine. The N4 atom of cytosine shows a greater overall number of close contacts than the N6 of adenine. The N4 atom on the smaller pyrimidine is closer to the outer edge of the major groove. Moreover, the N6 atom of adenine, although it is a hydrogen bond donor, is in a more negative environment than the N4 of cytosine, since it is close proximity to the electronegative N7.

The N7 atoms of adenine and guanine also exhibit a relative high number of close contacts in the major groove. The electronegative N7 atom offers a good site for electrostatic interactions. Although the N7 atom in the W1 position of both purine bases offers a large comparative contribution to close contact numbers, the absolute number of counts is much smaller for adenine than guanine. This is reasonable since the bulky amine in the W2 position of adenine may hinder close contact with the N7 atom of adenine in the major groove.

As can be seen in Figure 3.2, for adenine, the N7 atom in the W1 position and the N6 atom in the W2 position are the dominant sites of close contacts ($\leq$ 3.4Å). For thymine, the dominant site for close interactions is the O4 atom in the W2′ position. There is also a relatively high number of close contacts with C5M in the W1′ position of thymine, but not as high as with the O4 atom. Not surprisingly, the neutral C5M appears to be a less attractive site for interaction with proteins than the electronegative carbonyl.

In the major groove, even with the close proximity of the bulky C5M, the O atom in the W1′ still dominates the number of close contacts.

In Figure 3.3 it is observed that for the G-C pair, the O6 atom in the W2 position on guanine dominates the close contacts over the N7 atom in the W1 position of this base. From these data it is apparent that the carbonyl oxygen offers a very favorable binding site for protein atoms. For cytosine, the N4 atom in the W2′ position shows the highest degree of close interactions for this base, with a higher number of close contacts than the C4 atom in the W1′ position.

3.4.2 Hybridization Specific Close Contacts

As previously stated, the G-C base pair shows a much higher number of close contacts overall than the A-T base pair. Furthermore, guanine and cytosine show a large dominance of close contacts that involve sp^2 hybridized atoms compared to sp^3 hybridized atoms. For Adenine, close contacts with sp^2 hybridized atoms dominate over those with sp^3 hybridized atoms, but to a much smaller extent than for guanine or cytosine. For thymine, there is a small dominance of close contacts with sp^3 hybridized atoms over close contacts with sp^2 hybridized atoms.

Guanine shows the highest overall number of close contacts (≤ 3.4Å) in the major groove, of the four bases studied within the 499 non-redundant structures. For guanine, the number of close contacts with the sp^2 hybridized atoms of arginine lead over the number of close contacts with all the other amino acids in the non-redundant dataset utilized. Details of the specific interactions with Guanine in the major groove are described in Section 3.3.3.5. As shown in Figures 1.13 and 1.14, arginine has two potential sp^2 hybridized nitrogen atoms on its side chain. As described, there is

delocalization of electrons between the N_{H1} and N_{H2} atoms that are part of the guanidinium group of arginine. As a result, these nitrogen atoms have both been defined as sp^2 within the database since the degree of sp^2 versus sp^3 character varies depending on protein. Moreover, the extremely large numbers of close interactions between these atoms and the electronegative O6 and N7 atoms of guanine give good evidence of the delocalization of electrons and the overall positive charge distributed between the terminal amine groups of arginine.

Although there is a higher number of close interactions between guanine and sp^2 hybridized atoms in the major groove, guanine also exhibits the highest comparative number of close contacts in the major groove with sp^3 hybridized atoms. Furthermore, lysine shows the highest comparative number of sp^3 close contacts with guanine in the major groove within this dataset. As shown in Figure 1.13, the terminal amine of lysine retains a positive charge. Therefore, this group interacts strongly with the electronegative O6 and N7 atoms on guanine.

For cytosine, the N4 atom in the W2′ position provides the primary close interaction site for sp^2 hybridized atoms. This is not surprising since it is a good hydrogen bond donor. Furthermore, glutamic and aspartic acid show a high relative number of close interactions between its sp^2 hybridized side-chain atoms and cytosine. The terminal carbonyl oxygen atoms of glutamic and aspartic acid (O_{E1}/O_{E2} and O_{D1}/O_{D2}, respectively) show a high degree of close interaction with the N4 atom of cytosine. This is consistent with the hydrogen bond donor-receptor relationships. Arginine also shows a relatively high number of close interactions with cytosine in the major groove. These close contacts involve the terminal N_{H1} and N_{H2} atoms of arginine and the N4 atom of cytosine.

However, the data do not show a high degree of specificity for a particular cytosine atom in the major groove.

For the A-T base pair, adenine shows only a slightly higher number of close contacts with sp^2 hybridized atoms than sp^3 hybridized atoms. Furthermore, as shown in Figure 3.2, the sp^3 close contacts are dominated by interaction with the N7 atom of adenine. This is reasonable since the N7 atom is electronegative, and therefore it interacts strongly with the sp^3 hybridized N_{D2} atom of asparagine, as discussed in Section 3.3.3.2. The majority of close interactions with sp^2 hybridized atoms and adenine involve arginine and asparagine. For asparagine, the electronegative O_{D1} atom acts a good hydrogen bond acceptor with the N6 atom of adenine acting as the donor. Although there is potential for delocalization of electrons among the atoms of this amide, as shown by Kemnitz and Loewen, the dominant resonance structure for this amide is the one in which the carbonyl has a double bond [32]. Hence, it is postulated that there is more sp^2 hybridized character for the O_{D1} atom of asparagine.

The delocalization of electrons between the N_{H1} and N_{H2} atoms of arginine and the overall positive charge on the terminal side chain of this protein make it a very good species for closely interacting with hydrogen-bonding acceptors (in a donor acceptor relationships) and electronegative sites (in electrostatic attractions) in the Watson-Crick base pairs studied. Within this dataset, the N_{H1} and N_{H2} atoms of arginine interact to an almost equal extent with the N6 and N7 atoms of adenine, once again affirming that these atoms act similarly as a result of delocalization.

Thymine is the only Watson-Crick base studied that showed a higher relative number of close contacts with sp^3 hybridized atoms than sp^2 hybridized atoms. This is

logical since the main hydrogen-bonding site on thymine is the electronegative O4 atom. In the major groove, the neutral methyl group on thymine does not act as a strong hydrogen bond donor or acceptor, although possibly this bulky groove prevents some close interactions with the O4 atom from occurring. Among the 499 non-redundant structures in the dataset studied, a wide variety of amino acids were shown to closely interact with thymine in the major groove. The results of this research show that are 17 different amino acids that exhibit close interactions involving sp^3 atoms and thymine in the major groove. Furthermore, Figure 3.2 shows that the majority of these close interactions are with the O4 atom of thymine. Although close interactions with thymine involving sp^3 hybridized atoms dominate over those involving sp^2 hybridized atoms, there were also a significant absolute number of close contacts involving sp^2 hybridized atoms. The majority of these close interactions are with arginine, as a result of the delocalization electrons between the N_{H1} and N_{H2} atoms and overall positive charge on the terminal side chain.

3.4.3 Backbone and Side-chain Atom Close Contacts

As previously described in this paper, close contacts with side-chain atoms far outnumber close contacts with backbone atoms. Cytosine shows the highest relative proportion of close interactions ($\leq$ 3.4Å) with backbone amino acids of the four bases studied. Furthermore, within this dataset, it is the sp^2 hybridized atoms of the carbonyl in the backbone of amino acids that predominantly interact closely at the W2′ position of cytosine. The conformational aspects of the G-C pair therefore appear to make the amine of cytosine a high target for close interaction with the carbonyl in the backbones of proteins. Guanine also shows a comparatively higher number of close backbone

interactions compared to adenine or thymine, but not as high as cytosine. However, in contrast to cytosine, the close contacts with guanine include sp^2 hybridized atoms and a small percentage of sp^3 hybridized atoms from the peptide backbone.

Adenine and thymine also show a dominance of close contacts with sp^2 rather than sp^3 hybridized backbone atoms, but thymine has a slightly higher number of close contacts with sp^3 hybridized atoms than adenine. Furthermore, most close contacts for adenine occur with arginine and asparagine. The absolute number of close contacts with arginine is much lower with adenine than guanine. The side-chain carbonyl group of asparagine, with its sp^2 hybridized atoms, shows the highest number of close contacts with adenine. In addition, the sp^3 hybridized N_{D2} atom on asparagine interacts closely with the N7 atom in the W1 position of adenine. These observations are consistent with the prediction by Seeman *et. al*., that asparagine recognizes adenine and arginine recognizes guanine in the major groove [8].

Among the 499 non-redundant structures, arginine shows the highest relative number of close side-chain interactions in the major groove, both in terms of sp^2 and sp^3 hybridized atoms. In the major groove, arginine shows a particularly high number of close contacts between its sp^2 hybridized side chain atoms and guanine. Arginine is a very versatile amino acid due to the delocalization of electrons between its N_{H1} and N_{H2} atoms. Furthermore, this study shows that interactions are dependent on the particular protein-base interaction. As expected, based on the electropositive charge of the guanidinium group, both the N_{H1} and N_{H2} interact with hydrogen-bond acceptors, such as the O4 atom of thymine and the O6 atom of guanine. It is interesting to also note that although there is only a small absolute number of close contacts between the N_{H1} and N_{H2}

atoms of arginine and the N6 atom of adenine. This observation indicates that although the guanidinium group exhibits an overall positive charge, there are still some close contacts formed between this group and hydrogen bond donors such as the N6 of adenine.

CHAPTER 4
HYBRIDIZATION CONTACT ANALYSES IN THE MINOR GROOVE

Although the major groove provides the primary location for protein-DNA interactions, a sizable number of close interactions also occur in the minor groove. As previously noted, the differences in the number of hydrogen-bonding sites between the two sets of Watson-Crick base pairs may play a role in DNA structure and conformation and subsequently result in differences in DNA-protein interactions. The structural flexibility of DNA is therefore potentially affected by interactions that occur in both the major and the minor grooves.

Reviewing close contacts in the minor groove and comparing them to the observations made for the major groove will expand on the body of knowledge and contribute to a better understanding of the structure and function of DNA-protein interactions. This chapter will provide the results of studying the close interactions ($\leq$ 3.4Å) between amino-acid atoms and base atoms in the minor groove, among the 499 non-redundant structures in the NAPID database. In particular, observation of differences between sp^2 and sp^3 hybridized amino-acid atoms will be discussed.

As stated in Chapter 3, for hybridization contacts in the major groove, among the 499 non-redundant structures, each of the four nucleic acid bases studied, adenine, thymine, guanine and cytosine, are present in almost equal proportion. Within the dataset studied there are 5400 adenine bases, 5433 thymine bases, 5399 guanine bases, and 5375 cytosine bases. Since the denominator for all counts discussed in the subsequent sections is approximately the same for all bases, for simplicity it has been omitted specifically from minor groove calculations. Since the bases are present in almost equal proportions

within the research database, all comparisons within the minor groove will be representative of those taking into account overall number of bases in the database.

4.1 General Overview of sp^2 versus sp^3 Hybridized Atom Interactions in the Minor Groove of DNA

Figure 4.1 shows DNA minor-groove close contact counts by distance for each of the four bases; adenine, thymine, guanine and cytosine within the dataset studied (499 representative protein-DNA structures). Blue represents sp^2 hybridized atom close contacts and purple represents sp^3 hybridized atom close contacts. Data in this section elucidate the effect of amino-acid atom hybridization on interaction with the nucleic acid bases in the minor groove on a general level. Specific atomic details will then be explored in later sections.

Comparison of Figure 4.1 to Figure 3.1 shows that among the 499 non-redundant structures there is less than half the number of close contacts (≤ 3.4Å) between bases and proteins in the minor groove as the major groove (2718 and 6872, respectively). However, it is noteworthy that there are still a relatively large absolute number of interactions in the minor groove. As shown, by comparing the two figures, for the minor groove there is a somewhat more uniform distribution of contacts between the four bases, than in the major groove. These data indicate less specificity in the minor groove than the major groove.

Within the major groove, there is a significantly higher number of close contacts between proteins and the G-C base pair than proteins and the A-T base pair (G-C 4552 and A-T 2320 close contacts). However, in the minor groove, they are not as disparate, with the A-T base pair actually showing a higher number of close contacts with proteins than the G-C base pair (1560 and 1158, respectively). This is consistent with the fact that there is a lack of a hydrogen bond donor in the S2′ position of adenine in the minor groove. This leaves the oxygen in the S1′ position of thymine more available for close

interactions with incoming proteins. Additionally, there are important differences in the electronic structure of the minor groove edges of the A-T versus G-C base pairs.

Comparison of Figure 3.1 and Figure 4.1 shows that adenine has a higher number of close contacts (≤ 3.4Å) in the major than in the minor groove (1081 and 670 respectively). Furthermore, for adenine in the major as well as the minor groove, close contacts with sp^2 hybridized amino acid atoms are slightly higher in number than those with sp^3 hybridized amino acid atoms. In the minor groove there are 372 close contacts between adenine and sp^2 hybridized atoms, and 298 close contacts between adenine and sp^3 hybridized atoms. In addition, in the minor groove there are significantly more interactions with adenine at distances >3.2 and ≤ 3.4Å than at the closer distances, as shown in Figure 4.1. In the major groove, there is a more even distribution across the distance ranges studied for adenine, as shown in Figure 3.1.

Thymine has more close contacts (≤ 3.4Å) than adenine in the minor groove among the 499 non-redundant structures. Although, thymine shows a larger number of close contacts in the major groove than in the minor groove, these numbers are not that disparate. In the major groove thymine exhibits 1239 close contacts and in the minor groove thymine exhibits 1025 close contacts. Thymine shows a similar number of close contacts between sp^2 and sp^3 hybridized atoms in the minor groove as well as the major groove. In the minor groove there are 472 sp^2 close contacts with thymine and 418 sp^3 close contacts with thymine. Thymine shows the highest number of very close contacts (≤ 3.0Å) with amino-acid atoms of the four bases studied in the minor groove. Within the dataset studied, thymine exhibits 247 very close contacts (≤ 3.0Å) with protein atoms in

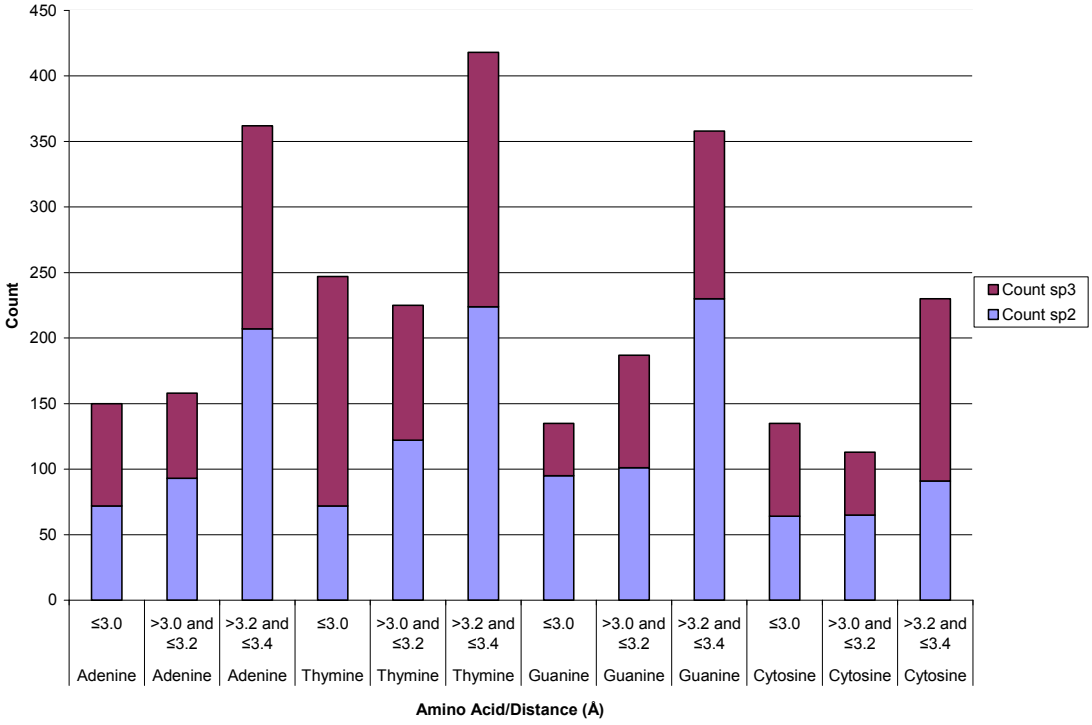
the minor groove. As previously noted, contacts at distances this close are likely ionic in nature.

Comparison of Figure 3.1 and Figure 4.1 shows much fewer very close (≤ 3.0Å) contacts for guanine in the minor groove than the major groove (135 versus 1419, respectively). Guanine exhibits the largest difference in very close interactions between the major and minor grooves of the four bases studied. In the major groove, across the three ranges studied, guanine shows the highest comparative number at very close distances (≤ 3.0Å). However, in the minor groove, across the three ranges studied, guanine shows a low number of contacts at very close distances (≤ 3.0Å), and the number is identical to that of its Watson-Crick partner cytosine. This low number of close contacts in the minor groove could be due to the fact that there is a strong internal hydrogen bond established between guanine and cytosine. This is the hydrogen bond between the amine in the S2 position of guanine and the carbonyl in the S1′ position of cytosine. For guanine, in the minor groove, sp^2 hybridized atom close contacts (≤3.4Å) dominate over sp^3 hybridized atom close contacts (426 and 254, respectively). This is also the case in the major groove (1854 and 1346, respectively).

Cytosine shows fewer very close contacts (< 3.0 Å) in the minor groove than major groove, but the difference is not as disparate as for guanine (135 and 337, respectively). In both the minor and major groove, for cytosine, sp^3 hybridized atom close interactions (≤ 3.4 Å) outnumber sp^2 hybridized atom close interactions. In the minor groove, there are 258 close contacts between sp^3 hybridized atoms and cytosine and 220 close contacts between sp^2 hybridized atoms and cytosine. Conversely, in the major groove, for cytosine, sp^2 hybridized atom close contacts, far outnumber sp^3 hybridized

atom close interactions. In the major groove, there are 1057 close contacts between sp^2 hybridized atoms and cytosine and 295 close contacts between sp^3 hybridized atoms and cytosine.

Figure 4. 1: Minor groove close contact counts for sp^2 and sp^3 hybridized amino-acid atoms



This graph shows DNA minor-groove close contact counts by distance for each of the four bases; adenine, thymine, guanine and cytosine within the dataset studied (499 representative protein-DNA structures). Blue represents sp^2 hybridized atom contacts and purple represents sp^3 contacts.
Overall base numbers within research dataset: adenine – 5400, thymine – 5433, guanine – 5399, cytosine - 5375

Source Data Appendix 13

4.2 Detailed Comparison of Amino-acid Hybridization on Contacts with Specific Nucleic Acid Atoms Types in the Minor Groove

As can be seen in Figures 1.3 and 1.4, within the minor groove, the S1' and S2 sites differ in terms of hydrogen bonding for the A-T and G-C base pairs. For adenine there is no hydrogen bond donor/acceptor in the S2 location. The thymine atom of the A-T Watson Crick base pair has the S1′ site occupied by the O2 atom of a carbonyl group. The O2 atom can act as hydrogen bond acceptor for hydrogen bond donors within proteins. In the minor groove, an important distinction between the A-T and G-C base pairs is that there is no available hydrogen bonding between atoms at the S2 and S1′ locations for A-T base pair, whereas for the G-C pair there is a potential hydrogen bond between these atoms.

4.2.1 Review of Minor Groove Contacts by Base Atom for Adenine-Thymine

Figure 4.2 shows the counts by atom type for the Watson-Crick A-T base pair in the minor groove for the non-redundant 499 structures chosen. Blue represents close contacts (≤3.4Å) with sp^2 hybridized atoms and purple represents close contacts with sp^3 hybridized atoms. As can be seen in this figure, within the minor groove, the N3 atom on adenine shows the largest number of close interactions (≤ 3.4Å) with protein atoms. The N3 atom of adenine exhibits 484 close contacts in the minor groove. Furthermore, there are a slightly higher absolute number of contacts between the N3 atom and sp^2 hybridized atoms (246 close contacts) over sp^3 hybridized atoms (238 close contacts). It is not surprising since many of the terminal amines present in proteins contain sp^3 hybridized atoms. However, as noted in Section 1.5.2, due to the potential delocalization of elections in proteins (backbone, side chain and peptide bond), the assignment of some sp^2

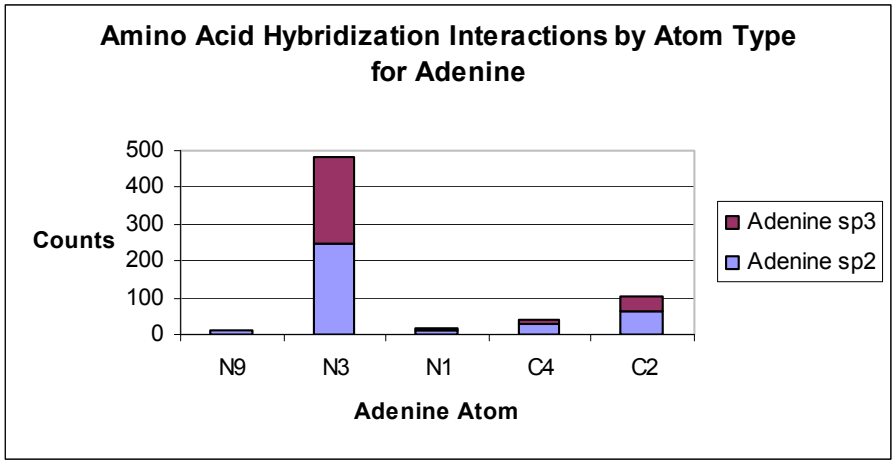
hybridized atoms was made because a dual assignment could not be given within the database. Within this dataset, there appears to be a significant number of close interactions between the N3 atom of adenine and sp^2 as well as sp^3 hybridized atoms within proteins.

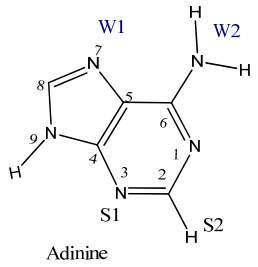
The C2 atom shows the second highest comparative number of close contacts (≤ 3.4Å) with amino-acid atoms (101 close contacts) in the minor groove. This atom shows a somewhat higher affinity for sp^2 over sp^3 hybridized amino-acid atoms (65 and 36, respectively). Although the C2 atom exhibits sp^2 hybridization, it is not as electron rich as the N3 atom of adenine, since the N3 atom has a non-bonded pair associated with it. The other atoms of adenine present in the minor groove show comparatively low numbers of close contacts with amino-acid atoms (< 50 close contacts each).

Comparison of Figures 3.2 and 4.2, shows that there are a similar number of close contacts for thymine in the minor groove and the major groove, but these close contacts have very different distributions. Interesting, thymine has two carbonyl oxygen atoms in its structure, one in the major groove and one in the minor groove. The carbonyl oxygen in the S1′ position in the minor groove shows a higher affinity for close interactions than the carbonyl oxygen in the W2′ position in the major groove (878 and 661, respectively). This is not unexpected, since the carbonyl oxygen in the S1′ site is not under the influence of any other hydrogen-bonding forces. In the major groove, the W2′ oxygen of thymine is involved in a hydrogen bond with the W2 position of adenine. In addition, for thymine, the O4 atom in the major groove is sheltered next to a large methyl group, whereas the O2 atom in the minor groove is next to a hydrogen atom on the outer

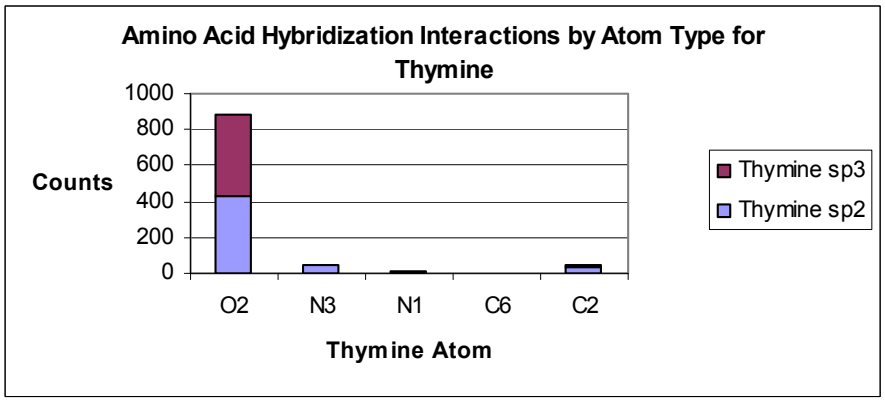
edge. The other atoms of thymine present in the minor groove show a comparatively low number of close contacts with amino-acid atoms (< 50 close contacts each).

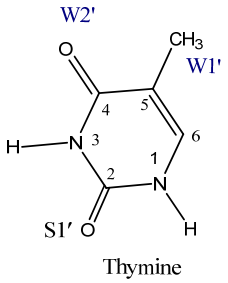
Figure 4. 2: Minor groove close contact counts by base atom type for A-T





Adinine





Thymine

Overall A-T base numbers within research dataset: adenine – 5400, thymine – 5433

Source Data Appendix 14

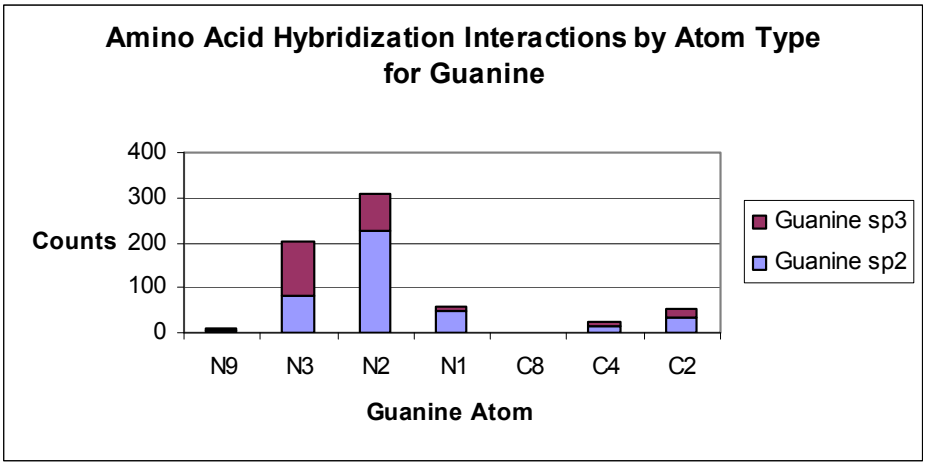
4.2.2 Review of Minor Groove Contacts by Base Atom for Guanine-Cytosine

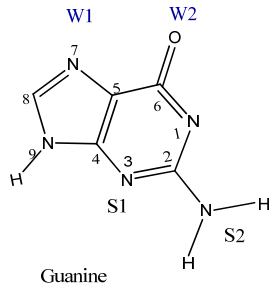
Figure 4.3 shows the close contact (≤3.4Å) counts by atom type for the G-C base pair in the minor groove for the non-redundant 499 structures under review. Blue represents close contacts with sp^2 hybridized atoms and purple represents close contacts with sp^3 hybridized atoms. As shown, the O2 atom in the S1′ location of cytosine has the largest absolute number of overall close contacts (≤ 3.4 Å) among the individual atoms of the G-C base pair (397 close contacts). For the O2 atom, close contacts with sp^3 hybridized amino-acid atoms dominate over close contacts with sp^2 hybridized atoms (243 and 154, respectively). This is consistent with the premise that the negatively charged carbonyl group on cytosine would interact strongly with sp^3 hybridized atoms, such as those associated with a terminal amine group of a protein. Once again this is a typical H-bond donor/acceptor relationship. The other atoms on cytosine in the minor groove show comparatively low numbers of close contacts (< 50 close contacts each).

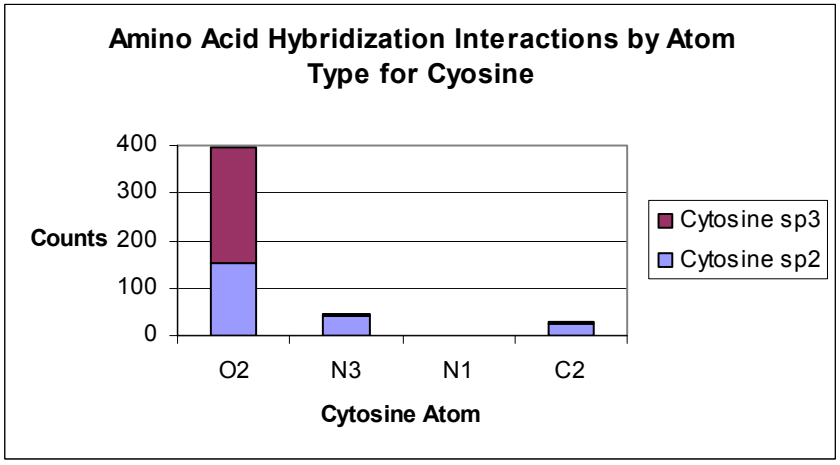
For guanine, the N2 atom in the S2 location shows a higher number of close contacts than the N3 atom in the S1 location (307 and 204, respectively). The N2 atom in the S2 position is more likely to interact with sp^2 hybridized atoms. As shown in Figure 4.3 there are 226 close contacts with sp^2 hybridized amino-acid atoms and 81 close contacts with sp^3 hybridized atoms. The N3 atom in the S1 location also shows a relatively high number of close contacts in the minor groove. However, for the N3 atom, close contacts with sp^3 hybridized atoms outnumber those with sp^2 hybridized atoms (120 and 84, respectively). This is understandable, since the electron rich double bond would more likely be attracted to sp^3 hybridized atoms than sp^2 hybridized atoms in a double bond. Although not technically in the minor groove, based on the mathematical formula

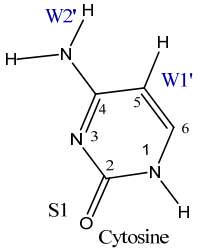
used in this research, the N1 atom of guanine also shows comparatively moderate number of close interactions with sp^2 hybridized atoms and has been designated as minor groove within some structures (57 close contacts observed).

Figure 4. 3: Minor groove close contact counts by base atom type for G-C









Overall G-C base numbers within research dataset: guanine – 5399, cytosine - 5375
Source Data Appendix 14

4.3 Amino-acid Minor Groove Backbone versus Side-chain Contacts in the Minor Groove

Figure 4.4 compares close contact (≤ 3.4 Å) counts for backbone and side-chain amino-acid atoms within the minor groove for each of the four bases studied in this research. Blue represents amino-acid backbone atom close contacts and purple represents side-chain atom close contacts. Across the four bases, close-contacts are dominated by amino-acid side-chain interactions in the minor groove. Of the 2814 total close contacts (among the 499 protein-DNA structures studied) within the minor groove, only 432 of these close contacts involve an amino-acid backbone atom, as compared to 2382 close contacts that involve an amino-acid side-chain atom.

Figure 4.4 shows that in the minor groove the A-T Watson-Crick base pair has a higher number of close contacts overall than the G-C base-pair, for both side-chain and backbone contacts (1660 and 1154, respectively). Comparison of the A-T versus the G-C base pairs in Figure 4.4, there are a similar number of backbone close contacts with amino acids between the two sets of base pairs, in the minor groove (245 and 187, respectively). Comparison of Figure 3.4 and Figure 4.4 indicates that these base pairs act differently in the minor groove than in the major groove. In the major groove the G-C base pair exhibits a higher number of overall close contacts. Furthermore, in the major groove, the G-C base pair shows a much higher degree of close interaction with backbone atoms than the A-T base pair (549 and 174 close contacts, respectively). In the minor groove the A-T base pair shows a higher number of close interactions with backbone atoms than the G-C base pair (245 and 187 close contacts, respectively).

4.3.1 Review of Adenine-Thymine Close Contacts in the Minor Groove

As shown in Figure 4.4, both adenine and thymine show a low number of close contacts (≤ 3.4 Å) with backbone amino-acid atoms, but adenine shows an even lower number of backbone atom contacts than thymine. Among the 499 structures studied, adenine exhibits a total of 94 close atomic contacts (≤ 3.4Å) with amino-acid backbone atoms, whereas thymine shows a total of 151 close contacts with backbone atoms. As previously noted, for the A-T base pair, both adenine and thymine show a higher number of close contacts with backbone atoms in the minor groove than in the major groove (245 and 174, respectively).

Adenine shows a higher number of close contacts with sp^2 hybridized backbone atoms than sp^3 hybridized backbone atoms, as shown in Figure 4.4. Of these 94 close contacts, 73 involve sp^2 hybridized atoms and 21 involve sp^3 hybridized atoms. This is not surprising; since all amino acids in proteins have a carboxyl oxygen atom in the backbone exhibiting sp^2 hybridization that could presumably interact through hydrogen bonding. In addition, all proteins have nitrogen with sp^2 hybridization contained in the peptide bond. Although the N3 atom in the S1 position of adenine is electron rich, it still appears to be a good site for close interaction with sp^2 hybridized atoms. From Figure 4.2 it can be seen that the N3 atom in the S1 position of adenine is the primary interaction site for amino acid side chains with this base.

As noted in Figure 4.4, thymine has a somewhat higher number of overall close contacts with backbone amino-acid atom in the minor groove than adenine, within the dataset studied. Thymine exhibits 151 close atomic contacts with amino-acid backbone atoms. For thymine the sp^2 hybridized atom close interactions are also favored over sp^3

hybridized atom close interactions (110 and 41, respectively). From Figure 4.2 it can be seen that the O2 atom in the S1' position of thymine is the primary interaction site for this base. From this figure it can be seen that the O2 atom interacts with both sp^3 and sp^2 hybridized atoms almost equally. In addition, there are also small numbers of close contacts between the N3, N1, C6 and C2 atoms of adenine and sp^2 hybridized atoms. These close contacts are described further in Section 4.3.3.3.

It should be noted that most close contacts (≤ 3.4 Å) experienced by adenine and thymine in the minor groove are with side-chain atoms. Figure 4.4 illustrates that among the 499 structures studied in this research, adenine experiences a total of 574 close atomic contacts with amino-acid side-chain atoms. Therefore in the minor groove, as in the major groove, there is far greater interaction with side-chain than backbone atoms. Adenine interacts closely to a slightly greater extent with sp^2 amino-acid side-chain atoms than sp^3 side-chain atoms. As described in Figure 4.4, adenine shows a total of 298 sp^2 close contacts versus 276 sp^3 close contacts with amino acids.

For thymine there is a slight dominance of sp^3 versus sp^2 side-chain atom close contacts (≤ 3.4 Å) in the minor groove. As described in Figure 4.4, thymine shows a total of 430 sp^3 close contacts versus 411 sp^2 close contacts with amino acids. Although the O2 atom in the S1′ position of thymine is a strong hydrogen bond acceptor that can interact with terminal amino groups of proteins, as described in Section 1.5.2, for some proteins that contain amino acids such as arginine, there is a potential for delocalization of elections. Therefore these side-chain atoms have been designated as sp^2 even though they have some sp^3 character.

4.3.2 Review of Guanine-Cytosine Close Contacts in the Minor Groove

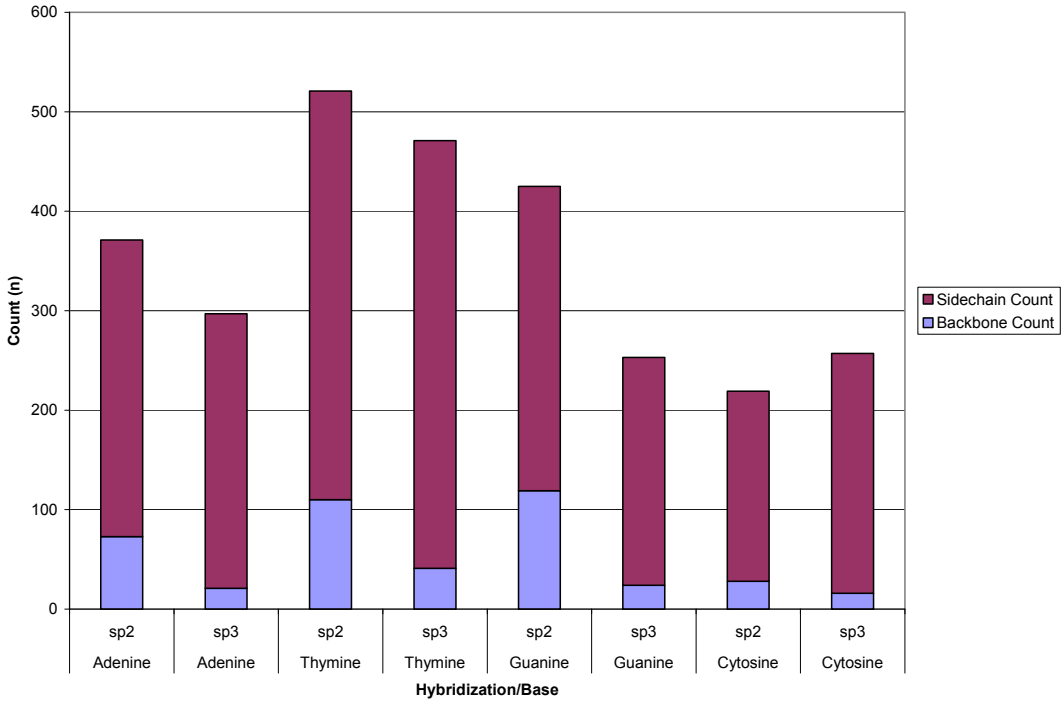
In the minor groove, the G-C Watson-Crick base pair shows a lower number of overall close contacts (≤ 3.4 Å) than the A-T base pair for the 499 non-redundant structures studied (1154 and 1660, respectively). This is opposite to how these pairs compare in the major groove. As previously noted, in the minor groove there is one less potential hydrogen-bond site for the A-T base pair. Since a hydrogen bond is not present between the S2 site and the S1′ site, these two sites more available to form hydrogen bonds with external entities. For the G-C base pair, the amine in the S2 position and the carbonyl in the S1′ site form a hydrogen bond in the minor groove. Therefore, there is internal competition with external entities for hydrogen bonding.

Among the 499 non-redundant structures, guanine shows a higher number of close contacts (≤ 3.4 Å) than its Watson-Crick partner cytosine in the minor groove. Figure 4.4 shows that for guanine, both sp^2 and sp^3 hybridized amino-acid atom close contacts have side-chain counts that are much higher than backbone counts. Among the 499 protein-DNA structures studied, guanine has only 143 close atomic contacts (≤ 3.4Å) with amino-acid backbone atoms, in contrast to 535 close atomic contacts with amino-acid side-chain atoms. Of the 143 close contacts with backbone atoms, 119 involve an sp^2 hybridized atom and 24 involve an sp^3 hybridized atom.

As shown in Figure 4.4, cytosine, like guanine, demonstrates a higher number of contacts with sp^2 hybridized side-chain atoms. However, cytosine interacts closely to a somewhat higher extent, with sp^3 hybridized side chains than sp^2 hybridized atoms (241 and 191, respectively). Once again, the O2 atom in the S1′ position of cytosine appears to interact closely with both sp^2 and sp^3 hybridized atoms. These contacts are dominated by

side-chain atoms, and furthermore, the side-chain close contacts show a preference for sp^3 hybridized atoms. Cytosine experiences very few close contacts with backbone atoms of proteins. For cytosine, there are 432 close contacts with side-chain atoms and 44 with backbone atoms.

Figure 4. 4: Minor groove close contact counts backbone versus side chain by hybridization



This figure compares close contact (≤ 3.4 Å) counts for backbone and side-chain amino-acid atoms within the minor groove for each of the four bases studied in this research. Blue represents amino acid backbone atom close contacts and purple represents side-chain atom close contacts.

Overall base numbers within research dataset: adenine – 5400, thymine – 5433, guanine – 5399, cytosine - 5375

Source Data Appendix 15

4.3.3 Hybridized Atom Interactions by Amino acid Type in the Minor Groove

Figures 4.5 through 4.22 describe protein-DNA interactions by specific amino acids in the minor groove. It is interesting to note the differences in the degree of interaction among the 499 structures included in this research, and also how the extent of interaction differs in the major and minor grooves. Refer to Figures 1.11, 1.12 and 1.13 for the amino-acid atom naming convention utilized in the discussion text in Sections 4.3.3.1 through 4.3.3.8.

4.3.3.1 Sp^2 Hybridized Atom Contacts with Adenine in the Minor Groove

Sp^2 Hybridized Side-Chain Atom Contacts

Figure 4.5 displays close interactions (≤ 3.4 Å) in the minor groove between sp^2 hybridized amino-acid side-chain atoms and adenine, classified by amino acid type. Polar basic arginine shows the highest number of sp^2 hybridized side-chain atom contacts with adenine (195 close contacts out of 298 total close contacts) in the minor groove. These close contacts are primarily the result of interaction between adenine and the terminal amines of arginine. Since arginine has two potential interaction sites with sp^2 hybridized atoms on its side chain (N_{H1} and N_{H2}), it has a high potential for close contacts. Of the 195 close contacts involving the arginine sp^2 side-chain atoms, 89 are with the N_{H1} atom 94 are with the N_{H2} atom and 12 are with the C_Z atom. Furthermore, as shown in Figure 4.2, the electronegative N3 atom on adenine shows the highest number of interactions with sp^2 hybridized amino-acid atoms. For arginine there are 69 close interactions between the N3 atom on adenine and the N_{H1} atom on arginine and 75 between N_{H2} atoms on the arginine side chain. Furthermore, there are 15 close contacts between the C2 atom of adenine and the N_{H1} atom and 18 between the C2 atom of

adenine and the N_{H2} atom. The almost equal number of close interactions for the N_{H1} and N_{H2} atoms would seem to indicate that both atoms act similarly in terms of interactions and therefore delocalization seems likely. Furthermore, the large number of close contacts with the electronegative N3 atom of adenine, adds support for the structure of arginine with a positively charged guanidinium group as shown in Figures 1.12 and 1.12a.

Phenylalanine also shows a comparatively large number of close contacts (≤ 3.4Å) with adenine involving side-chain atoms in the minor groove (61 close contacts). This group has a heterocyclic ring on its side chain. Interestingly, in the major groove this amino acid does not show as high a degree of close interaction as in the minor groove. The smaller number of close contacts in the major groove could be due to competition with the hydrogen bonding between the W2 position on adenine and the W2′ on thymine. In the minor groove, the atoms in heterocyclic ring of phenylalanine are not all equal in terms of close interaction adenine. For phenylalanine, there are 2 C_{D1}, 10 C_{E1}, 4 C_{E2} and 45 C_Z atom close contacts with adenine. These data suggest that phenylalanine is partially intercalated in the structure and interactions occur at locations where there is some sort of DNA deformity.

Sp^2 Hybridized Backbone Atom Contacts

As previously noted, among the 499 structures studied, adenine makes relatively few close contacts (≤ 3.4 Å) with backbone amino acid atoms in the minor groove. Of the 94 close contacts between backbone atoms and adenine, 73 involve sp^2 hybridized backbone atoms in the protein. Figure 4.6 illustrates that arginine shows by far the largest comparative number of sp^2 hybridized backbone atom close contacts with adenine in the

minor groove (50 close contacts). As seen with many other close interactions in the major and minor grooves, steric and conformational factors may play an important role in backbone-base interactions. For the sp^2 hybridized backbone atoms of arginine in contact with adenine, the counts are almost equally divided between the carbonyl oxygen (26 close contacts) and the nitrogen of the peptide backbone (23 close contacts) within this dataset.

Other than arginine, there does not seem to be a preference among other amino acids that present sp^2 hybridized backbone atoms in close contact with adenine. All the other amino acids that interact with adenine have very low numbers of close contacts with their sp^2 hybridized backbone atoms ($\leq$ 10 close contacts each). In addition, there is great diversity among the amino acids that show backbone atom interactions with adenine. It is interesting to note the difference in amino acid preference for close contacts between the major and the minor groove. Whereas arginine shows the highest comparative number of sp^2 hybridized close backbone atom contacts in the minor groove, threonine shows the highest comparative number of close backbone atom contacts in the major groove.

Figure 4. 5: Sp^2 hybridized side-chain close contacts with adenine by amino acid in the minor groove

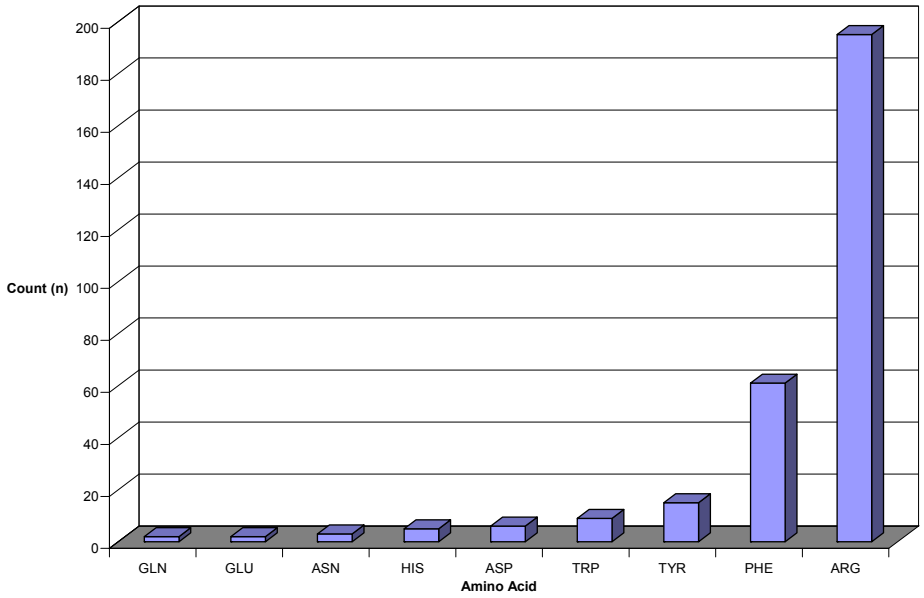
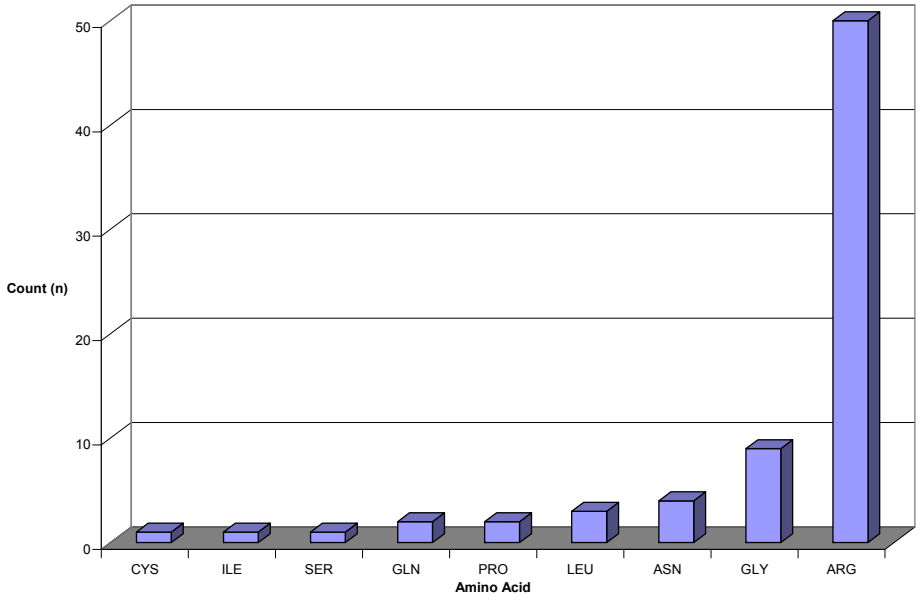


Figure 4. 6: Sp^2 hybridized backbone close contacts with adenine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^2 hybridized amino-acid side-chain and backbone atoms with adenine, classified by amino acid type in the minor groove.
Overall adenine base numbers within research dataset: adenine = 5400

Source Data Appendix 16

4.3.3.2 Sp^3 Hybridized Atom Contacts with Adenine in the Minor Groove

Sp^3 Hybridized Side-chain Atom Contacts

As shown in Figure 4.4, for sp^3 hybridized amino-acid interactions with adenine, the side-chain atom interactions dominate the backbone interactions in the minor groove. Figure 4.7 displays close contacts (≤ 3.4 Å) of a hybridized sp^3 amino-acid side-chain atom with adenine atoms, classified by amino acid type for the 499 non-redundant structures. This chart illustrates that a wide range amino acids have sp^3 hybridized side-chain atoms that interact closely with adenine. Polar uncharged asparagine and polar basic lysine show the highest number of close contacts (68 and 58 close contacts, respectively). Interestingly, in the minor groove glutamine shows a much lower number of close contacts than asparagine (11 versus 68 close contacts, respectively) even though these two amino acids differ by only one carbon in the side chain.

Within this research dataset, in the minor groove, close contacts (≤ 3.4 Å) are dominated by the interactions of the sp^3 hybridized nitrogen atom on asparagine and lysine and the electronegative N3 atom of adenine. As shown in Figure 4.2 most of the close contacts with amino-acid sp^3 hybridized atoms occur with this highly electronegative N3 atom on adenine. For asparagine the NH_2 group in its side chain acts a good proton donor. Each of the 68 close contacts experienced between asparagine sp^3 hybridized atoms and adenine involves the N_{D2} atom on asparagine. Furthermore, 66 of these close contacts are between the N_{D2} atom on asparagine and the N3 atom on adenine. For lysine the NH_3^+ group acts as a proton donor and there is a comparatively large number of close contacts that involve the N_Z atom of this group. However there are also a significant number of close contacts with the C_E atom of lysine. There are 35 close

contacts between the N_Z atom on lysine and adenine and 20 close contacts between the C_E atom of lysine and adenine.

Sp^3 Hybridized Backbone Atom Contacts

Figure 4.8 displays close contact (≤ 3.4 Å) of sp^3 hybridized amino-acid backbone atoms with an atom on the adenine base, classified by amino acid type. This chart shows that there are very few close interactions in the minor groove involving backbone sp^3 hybridized atoms (21 close contacts ≤3.4Å). Although glycine exhibits the highest number of close interactions (16 close contacts) comparatively, this is still a small absolute number of close contacts. As with the sp^2 interactions, there does not seem to be a clear electrostatic rationale for the discrimination and therefore steric and conformational factors may be acting to bring about the interactions. The small size of glycine may make it easier to fit into the groove. Once again this is a good indication that there are some conformational aspects that are favorable for interaction in the minor groove.

Figure 4. 7: Sp^3 hybridized side-chain close contacts with adenine by amino acid in the minor groove

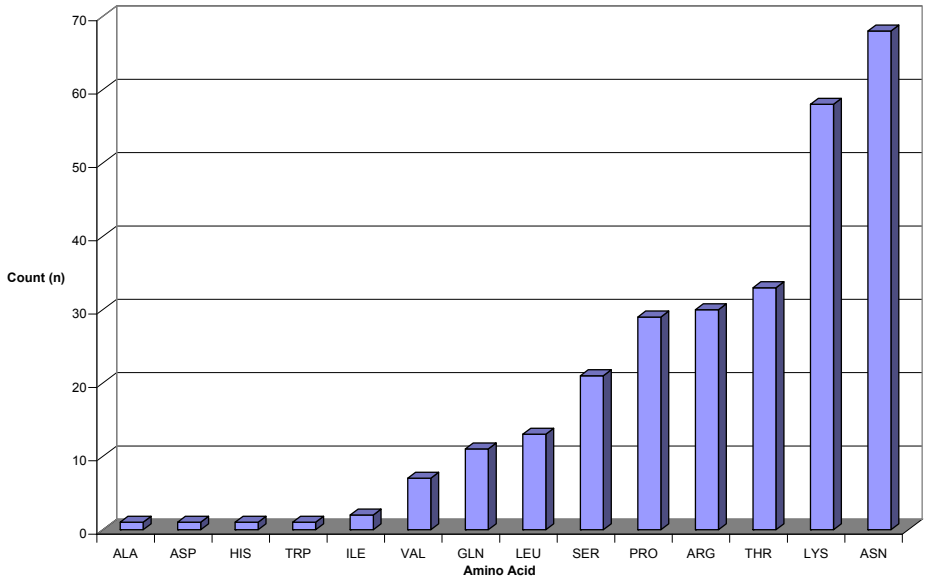
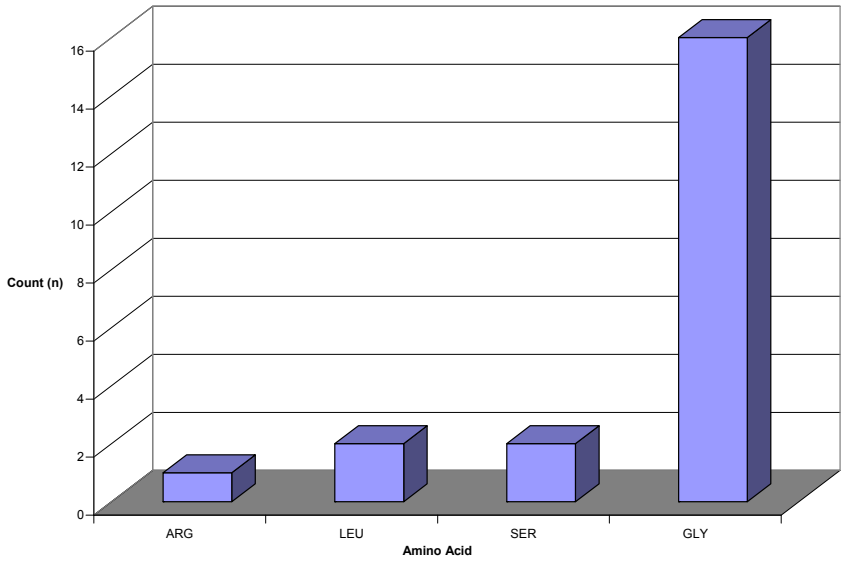


Figure 4. 8: Sp^3 hybridized backbone close contacts with adenine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^3 hybridized amino-acid side-chain and backbone atoms with adenine, classified by amino acid type in the minor groove.
Overall adenine base numbers within research dataset: adenine = 5400

Source Data Appendix 16

4.3.3.3 Sp^2 Hybridized Atom Contacts with Thymine in the Minor Groove

Sp^2 Hybridized Side-chain Atom Contacts

Figure 4.9 describes sp^2 hybridized amino-acid side-chain atom interactions with thymine in the minor groove by amino acid type among the 499 non-redundant structures studied. As shown in this figure, the polar basic arginine shows the highest number of close contacts involving sp^2 hybridized side-chain atoms (266 close contacts ≤ 3.4 Å). The most common close interaction for arginine occurs between the sp^2 hybridized nitrogen atoms of arginine and thymine. Specifically, 109 of the 266 total sp^2 arginine side-chain atom close contacts with thymine arise from interaction between the sp^2 hybridized N_{H1} atom of arginine and 100 arise from interaction with the sp^2 hybridized N_{H2} atom. Furthermore, these close contacts are primarily with the O2 atom in the S1′ location of thymine. As noted in Figure 4.2, the O2 atom on thymine is the dominant location for close contacts in the minor groove for the structures within this dataset. Examination of these close contacts shows that there are 106 close contacts between the O2 atom on thymine and the N_{H1} atom on arginine, and 91 close contacts are between the O2 atom on thymine and N_{H2} on arginine. There are also 57 close interactions between the C_Z atom on arginine and thymine.

Phenylalanine also makes a significant contribution to the number of close sp^2 hybridized side-chain atom interactions with thymine in the minor groove. As shown in Figure 4.7 there are 87 close contacts between the sp^2 side-chain atoms on phenylalanine and thymine. Although all of the heterocyclic carbons on the side chain interact with thymine, the number of close contacts varies for each. The C_Z atom exhibits the highest number of close sp^2 contacts with thymine (58 close contacts) for this dataset. There are

56 close contacts between the C_Z atom of phenylalanine and the O2 atom in the S1′ position of thymine and 2 with the C2 atom of thymine. Once again the large cyclic ring shows a preference for interaction with thymine in the minor groove. Since there is no hydrogen bonding between the S2 and S1′ sites in the minor groove for the A-T base pair, this seems to allow for a greater number of close contacts with the large cyclic ring of phenylalanine that is not seen in the major groove. These data again imply intercalation of phenylalanine into the DNA structure.

Sp^2 Hybridized Backbone Atom Contacts

As shown in Figure 4.10, thymine interacts closely to a very limited extent with sp^2 hybridized amino-acid backbone atoms present in the research dataset. However, these interactions occur with a wide range of amino acids, indicating low specificity in the minor groove. Among the 499 structures studied, there exist only 110 close contacts that include sp^2 hybridized backbone atoms. Although glycine and arginine show the highest number of sp^2 hybridized close interactions (34 and 20, respectively), there does not seem to be a strong preference for one type of amino acid with thymine. These interactions involve 17 different amino-acid species and include polar uncharged, polar basic, non-polar and hydrophobic amino acids.

Figure 4. 9: Sp^2 hybridized side-chain close contacts with thymine by amino acid in the minor groove

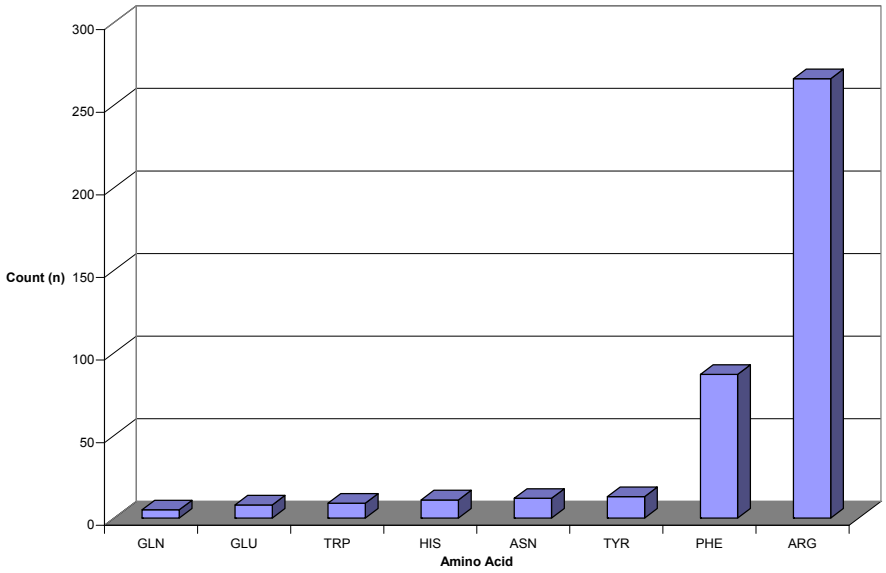
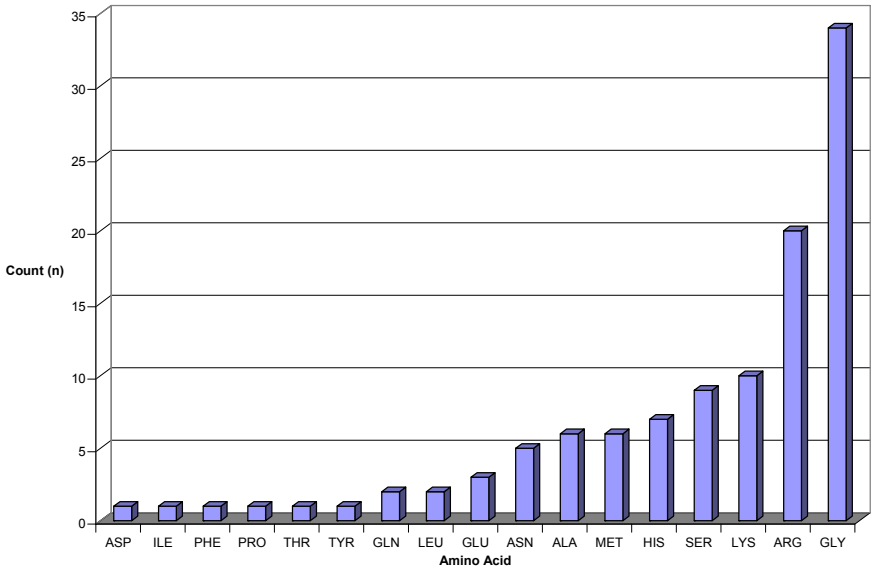


Figure 4. 10: Sp^2 hybridized backbone close contacts with thymine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^2 hybridized amino-acid side-chain and backbone atoms with thymine, classified by amino-acid type in the minor groove.
Overall thymine base numbers within research dataset: thymine – 5433

Source Data Appendix 17

4.3.3.4 Sp^3 Hybridized Atom Contacts with Thymine in the Minor Groove

Sp^3 Hybridized Side-chain Atom Contacts

Figure 4.11 describes sp^3 hybridized amino-acid side-chain atom interactions with thymine by amino acid type in the minor groove for the 499 non-redundant structures chosen. For sp^3 hybridized atoms closely interacting (≤ 3.4 Å) with thymine, the side-chain counts are far greater than the backbone atom counts. Within this dataset, there are 430 close contacts in the minor groove between thymine and sp^3 hybridized side-chain atoms in proteins.

As shown in Figure 4.9, lysine has the largest number close contacts between sp^3 hybridized side-chain atoms and thymine in the minor groove (166 close contacts). For lysine, 104 of a total 166 sp^3 hybridized side-chain atom close contacts are between the N_Z atom of the NH_3^+ group on its side chain and thymine. More specifically, 99 of the 104 are between the N_Z atom on lysine and the O2 atom on thymine.

Sp^3 Hybridized Backbone Atom Contacts

As shown in Figure 4.12 thymine interacts to a very limited extent with sp^3 hybridized amino-acid backbone atoms in the minor groove. Among the 499 structures studied, there are only 41 close contacts (≤ 3.4 Å) involving sp^3 hybridized backbone atoms from amino acids. Similar to the results obtained from major groove close interactions, glycine shows the highest number of close contacts with thymine involving sp^3 hybridized backbone atoms (30 close contacts) in the minor groove. Although this is the highest number comparatively, this is still a very small absolute number of contacts. The small size of the glycine probably aids its ability to allow backbone atoms to interact closely in the minor groove. For the other amino acids that show close contact between

backbone atoms and thymine in the minor groove, there is not another dominant interaction. Only glycine shows a much larger number close contacts with sp^3 hybridized backbone atoms and thymine.

Figure 4. 11: Sp^3 hybridized side-chain close contacts with thymine by amino acid in the minor groove

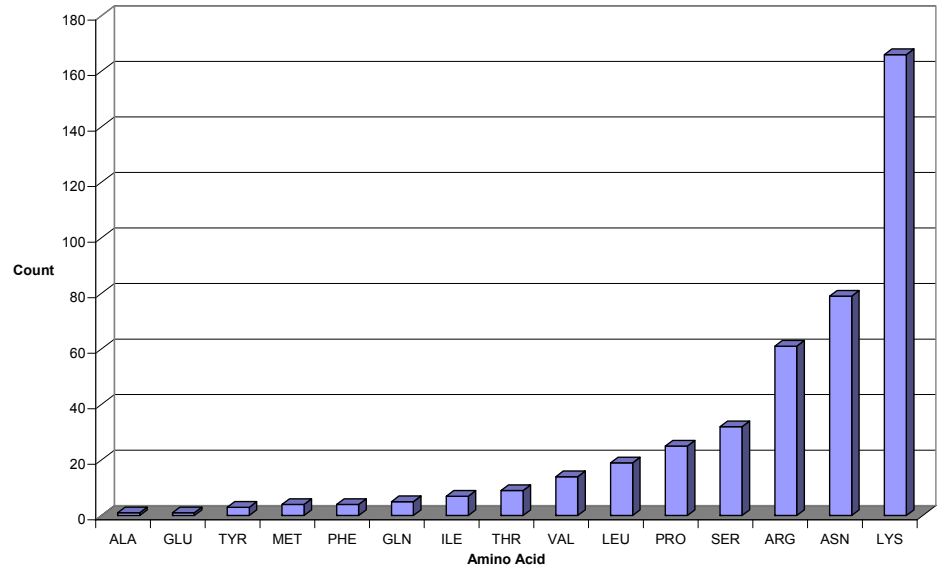
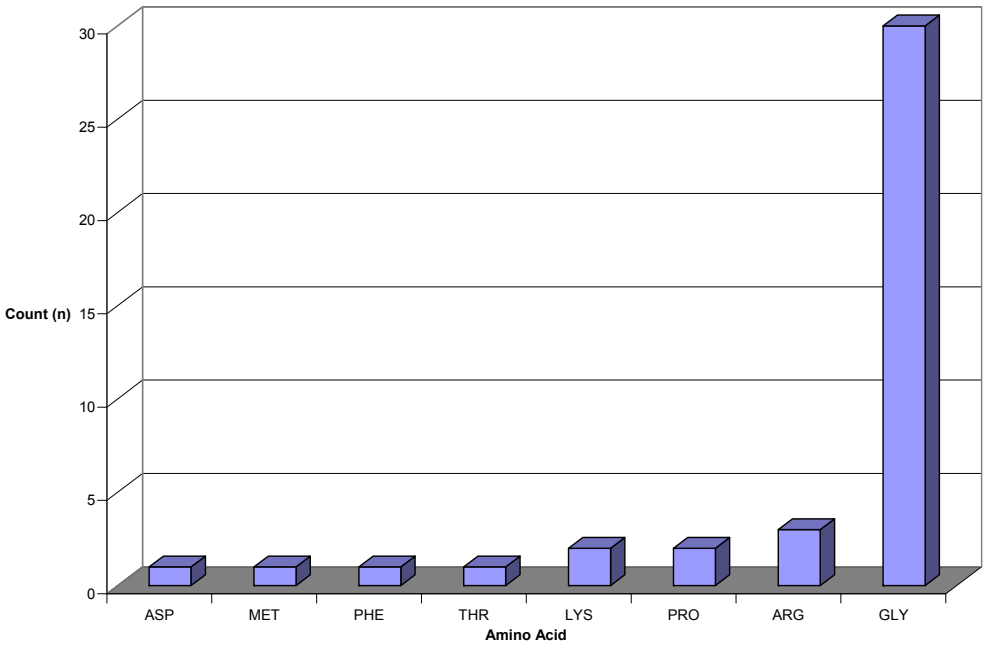


Figure 4. 12: Sp^3 hybridized backbone close contacts with thymine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^3 hybridized amino-acid side-chain and backbone atoms with thymine, classified by amino-acid type in the minor groove.
Overall thymine base numbers within research dataset: thymine – 5433

Source Data Appendix 17

4.3.3.5 Sp^2 Hybridized Atom Contacts with Guanine in the Minor Groove

Sp^2 Hybridized Side-chain Atom Contacts

Figure 4.13 shows specific amino-acid sp^2 side-chain atom interactions (≤ 3.4 Å) with guanine in the minor groove among the 499 non-redundant structures studied. For guanine, the largest comparative number of sp^2 side-chain atom close interactions in the minor groove occurs with arginine. Although guanine shows the largest comparative number of sp^2 side-chain atom close contacts with arginine in both the major and minor groove, there are far fewer absolute contacts in the minor groove than in the major groove (98 versus 1323, respectively). As discussed in Section 3.3.3.5, the very high number of sp^2 hybridized atom close-contacts (≤ 3.4 Å) between guanine and arginine side-chain atoms in the major groove are due largely to interactions with the nitrogen atoms (N_{H1} and N_{H2}) on the arginine side chain. Furthermore, a significant number of the close contacts specifically involve the O6 atom of guanine with the N_{H1} and N_{H2} atoms on the side chain of arginine in the major groove. Within the minor groove, the close interactions with arginine are not specifically dominated by a particular base atom. The close interactions involving the N_{H1} and N_{H2} atoms occur with N2, C2 as well as N3 on guanine. Of the 98 close interactions between guanine and arginine in the minor groove, 45 are between the N_{H1} atom on arginine and guanine, 41 are between the N_{H2} atom on arginine and guanine and 12 are between the C_Z atom on arginine and guanine. Table 4.1 describes the specific close interactions between the N_{H1} and N_{H2} atoms and guanine in the minor groove.

Table 4. 1: Close contacts between guanine and the N_{H1} and N_{H2} atoms of arginine in the minor groove

Guanine Atom	Close Contacts with N_{H1}	Close Contacts with N_{H2}
C2	2	0
C4	2	2
N2	18	25
N3	22	13
N6	0	1

Glutamine and asparagine also show a comparatively high number of sp^2 hybridized side-chain atom close contacts with guanine (42 and 38, respectively) in the minor groove. This is not surprising, since the carbonyl oxygen atom on the side chain of both of these amino acids show high potential for hydrogen bonding with the NH_2 in the S2 position of guanine in the minor groove. Within this dataset there are 37 interactions between the O_{E1} atom of glutamine and guanine and 34 close contacts between the O_{D1} of asparagine and guanine. Among the other amino-acid interactions noted in Figure 4.13, all others show a small absolute number of close sp^2 hybridized side-chain atom interactions with guanine.

Sp^2 Hybridized Backbone Atom Contacts

Comparison of Figure 4.14 and 3.14 illustrates that there much less disparity in the number of close contacts ($\leq$ 3.4 Å) between sp^2 hybridized backbone atoms and guanine in the major and minor grooves (179 and 119, respectively). Moreover, in both the major and minor grooves glycine shows the highest comparative number of close contacts involving sp^2 hybridized backbone atoms (53 and 28, respectively). These close contacts involve the N atom in the peptide bond of proteins containing glycine, as well as

the carbonyl O. In the minor groove, among the 28 total close contacts between guanine and glycine sp^2 hybridized backbone atoms, 7 involve the N atom in glycine (within the protein peptide bond) and 21 involve the carbonyl O and glycine. Once again the small size of glycine may make it better able to develop backbone interactions in the minor groove.

As can be seen in Figure 4.14, tryptophan and alanine show a comparatively high number of sp^2 hybridized backbone atom close contacts with guanine (16 and 9 close contacts, respectively). Interestingly both of these amino acids have heterocyclic rings in their side chains. The planar shape of these heterocyclic rings allows them to fit into the minor groove.

Figure 4. 13: Sp^2 hybridized side-chain close contacts with guanine by amino acid in the minor groove

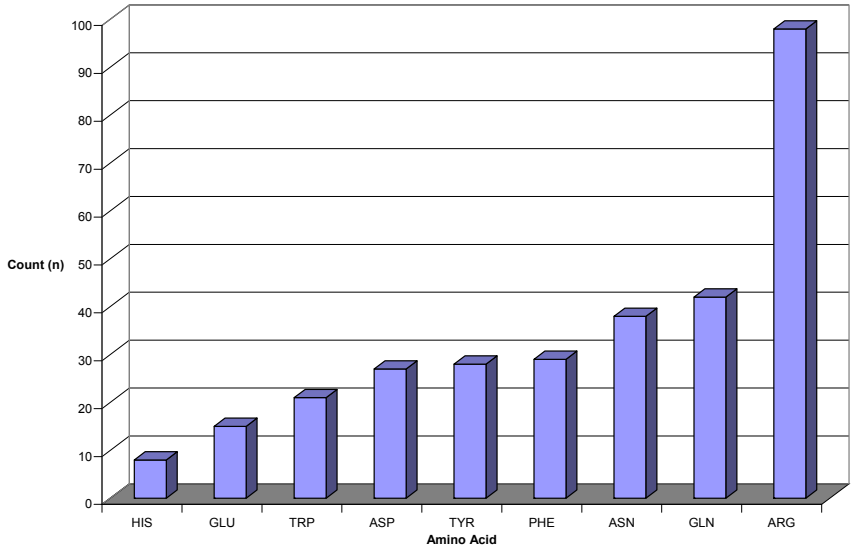
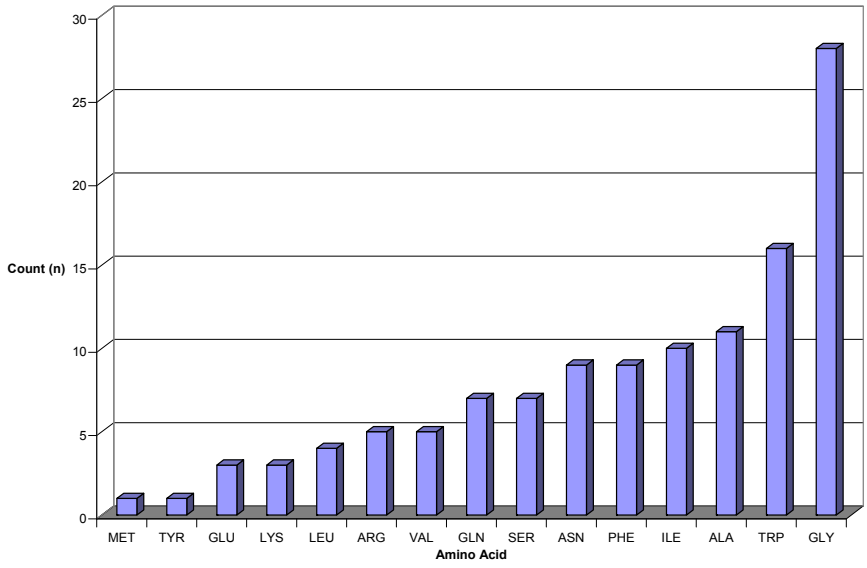


Figure 4. 14: Sp^2 hybridized backbone close contacts with guanine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^2 hybridized amino-acid side-chain and backbone atoms with guanine, classified by amino acid type in the minor groove.
Overall guanine base numbers within research dataset: guanine – 5399

Source Data Appendix 18

4.3.3.6 Sp^3 Hybridized Atom Contacts with Guanine in the Minor Groove

Sp^3 Hybridized Side-chain Atom Contacts

Among the 499 protein-DNA structures studied, guanine shows a similar number of sp^3 hybridized side-chain atom close contacts (≤ 3.4 Å) as the other three bases (adenine, thymine and cytosine) in the minor groove. As described in Figure 4.4, for the minor groove, guanine shows a comparable number of close contacts with sp^3 hybridized side-chain atoms as adenine and cytosine, and slightly less than thymine. There are a total of 229 close contacts between sp^3 hybridized side-chain atoms and the guanine base in the minor groove, noted in Figure 4.4. A large number of amino acids interact with guanine. Moreover, there does not seem to be dominance for a particular amino acid or type of amino acid in the minor groove. Lysine has the highest absolute number of close contacts between sp^3 side-chain atoms with guanine. Within this dataset there are 41 close contacts between guanine and lysine sp^3 hybridized side-chain atoms in the minor groove. Out of the total 41 sp^3 hybridized atom close contacts with lysine, 27 involve the N_Z atom of the NH_3^+ group, 12 involve the C_E atom and 2 involve the C_B atom of lysine. Once again close contacts with the N_Z atom of lysine are dominated by interactions with the electron rich N2 and the N3 atoms on guanine. There are 10 close contacts between the N_Z atom on lysine and the N2 atom on guanine and 17 close contacts between the N_Z atom on lysine and the N3 atom on guanine.

Asparagine, glutamine and serine also show a comparatively high number of close sp^3 side-chain atom interactions with guanine (39, 36, and 33 close contacts, respectively). The OH group on serine can act as a hydrogen bond donor. In addition, the relatively small size of serine makes it both electronically and sterically favorable for

interactions in the minor groove. For asparagine and glutamine, the N_{D2} and N_{E2} atoms, respectively, provide the largest number of close contacts since each can act as a good hydrogen bond donor for interactions with guanine.

Sp^3 Hybridized Backbone Atom Contacts

Figure 4.4 shows that guanine experiences a total of 24 close contacts (≤ 3.4 Å) with sp^3 hybridized backbone atoms. As described in Figure 4.16, glycine shows the highest number of sp^3 hybridized backbone atom close contacts within the dataset studied (14 close contacts). These sp^3 hybridized backbone atom close contacts are solely the result of the interaction of the C_A atom in the amino-acid backbone with guanine. Furthermore, 10 of these 14 close contacts are with the N3 atom on guanine. Because of the very small size of glycine, it more easily fits into the minor groove so that its backbone can come into close contact with the base.

Figure 4. 15: Sp^3 hybridized side-chain close contacts with guanine by amino acid in the minor groove

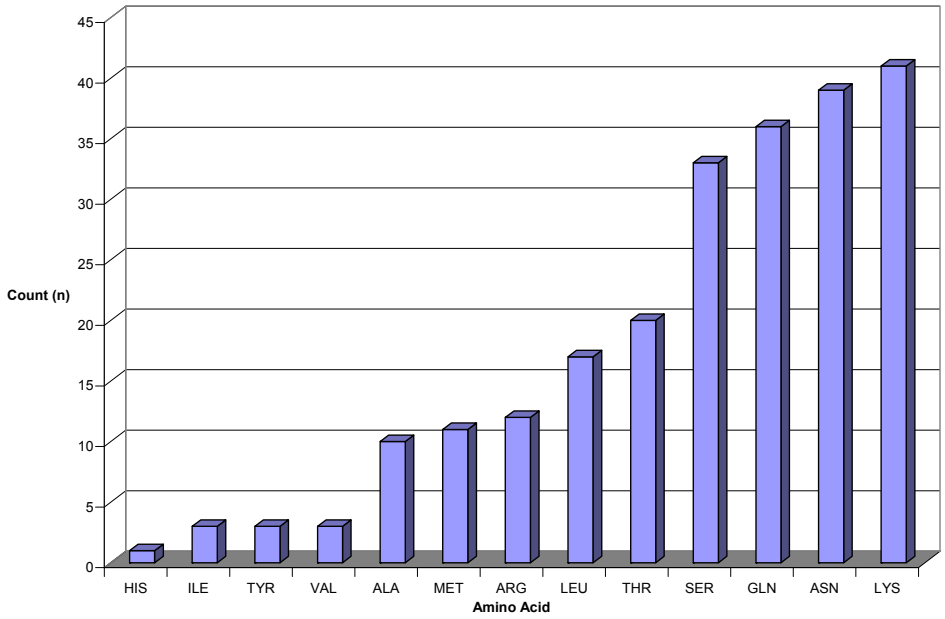
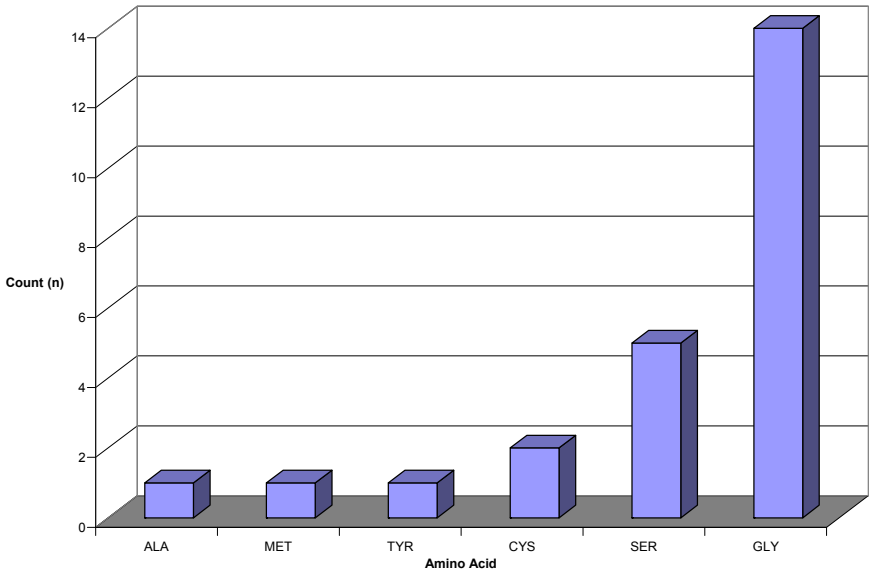


Figure 4. 16: Sp^3 hybridized backbone close contacts with guanine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^3 hybridized amino-acid side-chain and backbone atoms with guanine, classified by amino acid type in the minor groove.
Overall guanine base numbers within research dataset: guanine – 5399

Source Data Appendix 18

4.3.3.7 Sp^2 Hybridized Atom Contacts with Cytosine in the Minor Groove

Sp^2 Hybridized Side-chain Atom Contacts

As shown in Figure 4.4, cytosine experiences the lowest comparative number of close contacts (≤ 3.4Å) with sp^2 hybridized side-chain atoms in the minor groove of the four nucleic acids studied, and much fewer than its Watson-Crick partner guanine. Furthermore, unlike its partner guanine, cytosine exhibits fewer close contacts with sp^2 hybridized side-chain atoms than sp^3 hybridized side-chain atoms. It is also noteworthy that the pyrimidine bases, thymine and cytosine, both show a higher number of close contacts in the minor groove with sp^3 hybridized side-chain atoms than with sp^2 hybridized side-chain atoms.

As described in Figure 4.17, cytosine shows by far the greatest comparative number of sp^2 hybridized side-chain atom close contacts (≤ 3.4Å) with arginine in the dataset studied in the minor groove. As shown, there are 132 close contacts between sp^2 hybridized atoms on the arginine side chain and cytosine in the minor groove. These interactions are dominated by the close contacts with the sp^2 hybridized nitrogen atoms on the side chain of arginine. There are 50 close contacts involving the N_{H1} atom and 73 close contacts involving the N_{H2} atom on arginine with cytosine. Furthermore, these data once again clearly demonstrate the dual nature of the nitrogen atoms on the side chain of arginine in proteins. Both nitrogen atoms interact with the electron rich atom O2 atom, as well the non-polar N1 on cytosine. There are 44 close contacts involving the N_{H1} atom and 50 close contacts involving the N_{H2} atom of arginine with the O2 atom on cytosine in the minor groove.

Sp^2 Hybridized Backbone Atom Contacts

As described in Figure 4.4, among the 499 structures studied, cytosine shows the smallest comparative number of close contacts (≤ 3.4Å) in the minor groove with amino acid sp^2 hybridized backbone atoms of the four Watson-Crick bases studied. This is highly contrary to interactions in the major groove where cytosine exhibits the highest number of close contacts with sp^2 hybridized backbone atoms. As described in Section 3.3.3.7, the dominant interaction in the major groove is the interaction of the sp^2 hybridized carbonyl in the backbone of all amino acids with the NH_2 on cytosine at the W2′ position. However, in the minor groove, the presence of carbonyl at the S1′ position dominates close interaction and these interactions are primarily with sp^3 hybridized atoms.

Figure 4.18 shows that in the minor groove, within the dataset studied, there does not appear to be a dominant close interaction. A number of amino acids show close interaction between sp^2 hybridized backbone atoms and cytosine, but the numbers are all low, <15 close interactions for all encountered.

Figure 4. 17: Sp^2 hybridized side-chain close contacts with cytosine by amino acid in the minor groove

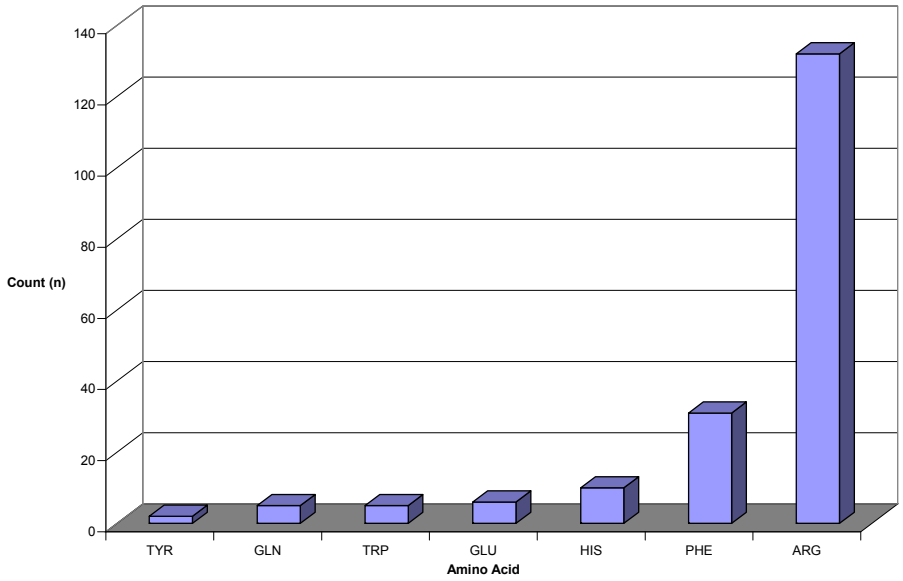
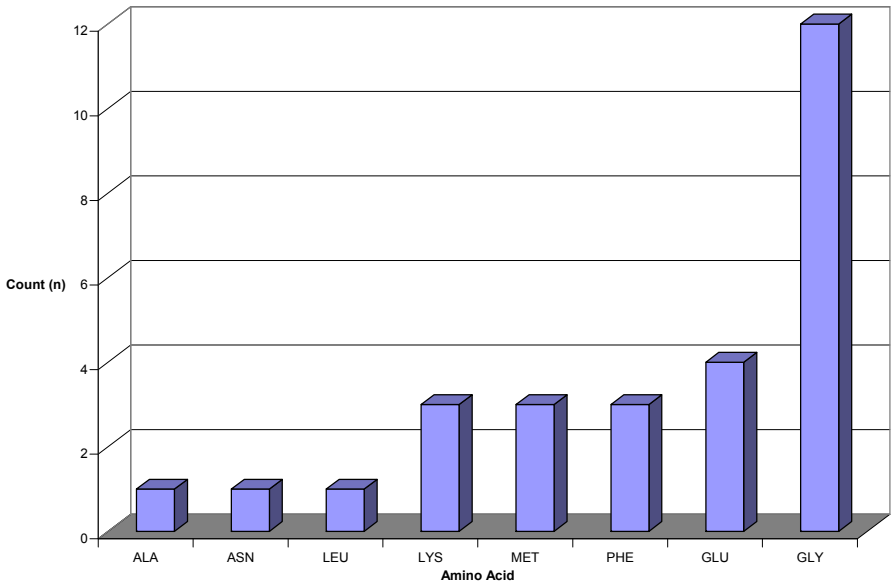


Figure 4. 18: Sp^2 hybridized backbone interactions with cytosine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^2 hybridized amino-acid side-chain and backbone atoms with cytosine, classified by amino acid type in the minor groove.
Overall cytosine base numbers within research dataset: cytosine – 5375

Source Data Appendix 19

4.3.3.8 Sp^3 Hybridized Atom Contacts with Cytosine in the Minor Groove

Sp^3 Hybridized Side-chain Atom Contacts

As shown in Figure 4.4, cytosine displays a slightly lower number of close contacts (≤ 3.4Å) than guanine and adenine in the minor groove with sp^3 hybridized amino-acid side-chain atoms within the dataset examined (241 close contacts). However, cytosine shows a much lower comparative number of close contacts than thymine in the minor groove with sp^3 hybridized amino-acid side-chain atoms within the dataset examined. In comparison, thymine shows 411 close contacts with sp^3 hybridized side-chain atoms in the minor groove for this dataset.

As described in Figure 4.19, cytosine shows the highest number of close contacts with sp^3 hybridized atoms on the lysine side chain. There are 98 close contacts with sp^3 hybridized atoms on the lysine side chain. For lysine, these close contacts are mainly with the N_Z and C_E atoms of the side chain; however, there are close contacts with the C_D and C_G atoms. Of the 98 close contacts involving sp^3 hybridized side-chain contacts with cytosine, 56 of these involve the N_Z atom and 31 involve the C_E atom on lysine. For the N_Z atom, 55 of 56 close contacts are with the O2 atom on cytosine.

Arginine also has a high comparative number of close contacts between sp^3 hybridized side-chain atoms and cytosine. For arginine, these sp^3 hybridized side-chain close contacts are mainly with the sp^3 hybridized N_E atom on its side chain. For arginine, of the 52 sp^3 hybridized side-chain close contacts with cytosine, 27 of these involve the N_E atom. Furthermore, for arginine 25 of these 27 close contacts are with the O2 carbonyl atom in the S1′ position on cytosine.

Sp^3 Hybridized Backbone Atom Contacts

As shown in Figure 4.4, cytosine exhibits an extremely small number of close interactions ($\leq$ 3.4Å) in the minor groove with sp^3 hybridized backbone atoms on proteins in the dataset studied. For the 499 structures investigated, there were only 16 close contacts with sp^3 hybridized backbone atoms in the minor groove and they all involve glycine. Furthermore, all of these close contacts are between the C_A atom on the glycine backbone and cytosine.

Figure 4. 19: Sp^3 hybridized side-chain close contacts with cytosine by amino acid in the minor groove

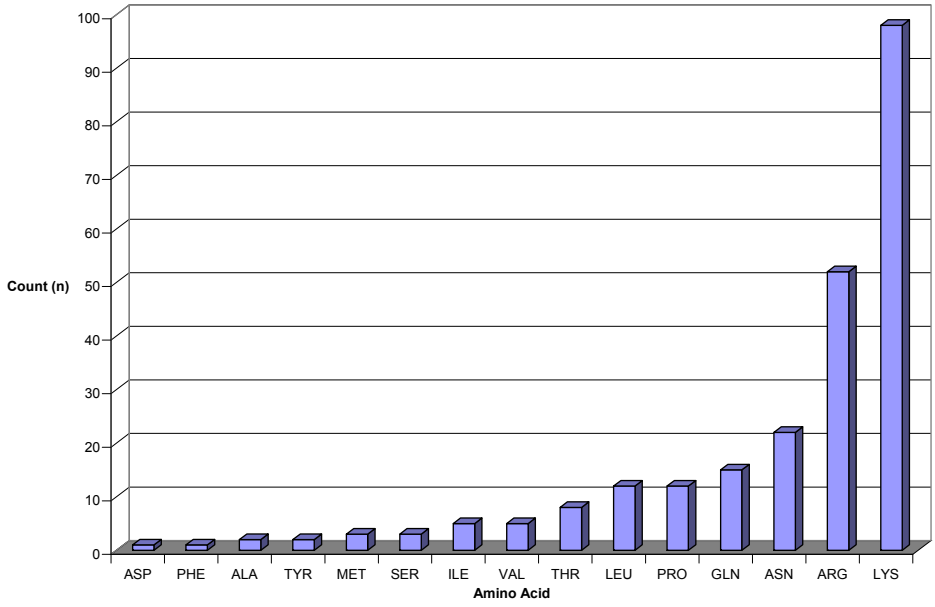
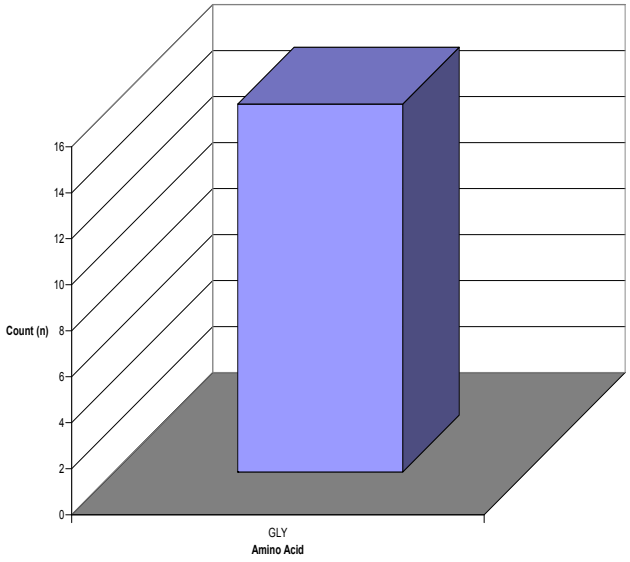


Figure 4. 20: Sp^3 hybridized backbone close contacts with cytosine by amino acid in the minor groove



These tables display close interactions (≤ 3.4 Å) between sp^3 hybridized amino-acid side-chain and backbone atoms with cytosine, classified by amino acid type in the minor groove.
Overall cytosine base numbers within research dataset: cytosine - 5375
Source Data Appendix 19

4.4 Minor Groove Close Contacts Summary Discussion

The following summary represents information about minor groove close contacts obtained from the dataset (499 representative structures) studied in this research**.**

4.4.1 General Comparison of Close Contacts

In the minor groove, as shown in Figure 4.1, along the edges of the A-T base pair, there are a higher absolute number of close contacts (≤ 3.4Å) than along the edges of the G-C base pair. Moreover, there is a greater imbalance between sp^2 and sp^3 hybridized atom close-contacts for the A-T base pair than the G-C base pair. This finding is opposite to the results obtained from this dataset for the major groove. As previously discussed, in the major groove, the G-C base pair exhibits a higher number of close contacts and a greater imbalance between sp^2 and sp^3 hybridized atom close contacts along the edges than the A-T base pair. As noted in Section 3.4, an important distinction between the A-T and G-C base pairs in the minor groove is that the G-C base pair has the potential for three hydrogen bonds between the guanine and cytosine bases, but the A-T base pair has only two potential hydrogen-bonding sites between the adenine and thymine bases. From these data, it appears this allows for a greater number of close contacts by proteins for the A-T base pair than the G-C base pair in the minor groove.

As observed in Figures 4.2 and 4.3, among the 499 non-redundant structures, the carbonyl (O2, thymine and cytosine) and amine (N2, guanine) functional groups contribute significantly to the number of close contacts experienced in the minor groove. For thymine, the dominant site for close interactions is the O2 atom in the S1′ position. The O2 atom of thymine exhibits a much greater number of close contacts, overall and with sp^2 hybridized atoms, than the O2 atom of cytosine, in the minor groove. In the

minor groove, the O2 atom of cytosine has the potential of hydrogen bonding with the N2 atom of guanine, however for the O2 atom on thymine there is no potential hydrogen bonding site on its Watson-Crick base-pair partner, adenine. This may allow the O2 atom of thymine to be more "accessible" to hydrogen bonding and electrostatic interactions with proteins in the minor groove than the O2 atom of cytosine.

In addition, in the minor groove, the electronegative N3 atoms of guanine and adenine account for a large number of close contacts found in this research. This is not unanticipated since these atoms offer the ability to form a hydrogen bond, as well as offer the potential for strong electrostatic attractions. Although the N3 atom in the S1 position of both purine bases offers a large comparative contribution to close contact numbers, there are more close contacts for adenine than guanine. As can be seen in Figure 4.2, for adenine, the N3 atom in the S1 position dominates close interactions ($\leq$ 3.4Å) in the minor groove. This is logical since the N3 of guanine is next to the bulky amine group in the S2 position. This group may hinder close contacts with N3 for guanine in the minor groove.

On the other hand, in Figure 4.3 it is observed that for guanine, the N2 atom in the S2 position dominates close contacts for this base in the minor groove. There are more close contacts with the N2 atom in the S2 position of guanine than with the N3 atom in the S1 position. From these data it is apparent that in the minor groove the N3 atom within the amine on guanine offers a very favorable hydrogen-bonding site for protein atoms and is more accessible than the N3 atom in the S1 position. For cytosine, the O2 atom in the S1′ position shows the highest degree of close interactions for this base. This is the only potentially good hydrogen-bonding site for cytosine in the minor groove.

4.4.2 Hybridization Specific Close Contacts

Thymine shows the highest overall number of close contacts (≤ 3.4Å) in the minor groove, of the four bases studied, among the 499 non-redundant structures selected. Thymine also exhibits the highest comparative number of contacts with sp^3 hybridized amino-acid atoms in the minor groove. The pyrimidine bases, thymine and cytosine, show a dominance of close contacts with sp^3 hybridized atoms over sp^2 hybridized atoms in the minor groove. For the purine bases, adenine and guanine, contacts with sp^2 hybridized atoms dominate over those with sp^3 hybridized atoms. The dominance is greater for guanine than adenine.

For adenine, the sp^2 hybridized atoms of arginine show the greatest number of close contacts among the 499 non-redundant structures in this data set, within the minor groove. Details of the specific interactions are described in Section 4.3.3.1. The extremely large and nearly equal numbers of close interactions between the electronegative N3 atom of adenine and the N_{H1} and N_{H2} atoms of arginine once again give good evidence of the high sp^2 character and potential delocalization of charge between these atoms. As previously stated, the extent of close interaction varies by protein-DNA structure and is a potential area of further investigation.

For thymine, the higher relative number of close interacts with sp^3 hybridized atoms over sp^2 hybridized atoms is logical since the main hydrogen-bonding site on thymine in the minor groove is the electronegative O2 atom. Lysine shows the highest comparative number of sp^3 hybridized atom close contacts with thymine in the major groove within this dataset. As shown in Figure 1.13, the amine of lysine retains a positive charge. Therefore, this group interacts strongly with the electronegative O2 atom in a

hydrogen bond donor-acceptor relationship. For thymine, although the number of close contacts with sp^3 hybridized atoms is greater than the number of close contacts with sp^2 hybridized atoms, there are still significant numbers close contacts with sp^2 hybridized atoms present in this dataset. Arginine shows the highest number of close interactions between sp^2 hybridized atoms and thymine and these interactions are primarily between the N_{H1} and N_{H2} atoms of arginine and the O2 atom of thymine.

Guanine shows far fewer close interactions in the minor groove than in the major groove. Guanine shows more close interactions with sp^2 hybridized atoms than sp^3 hybridized atoms in the minor groove. For guanine, the highest numbers of close interactions with sp^2 hybridized atoms involve the N_{H1} and N_{H2} atoms of arginine. For guanine, there is low specificity in terms of close contacts with sp^3 hybridized atoms. Lysine, asparagine, glutamine and serine all show a similar number of close contacts involving sp^3 hybridized side-chain atoms.

There are a higher number of close contacts between cytosine and sp^3 hybridized atoms than between cytosine and sp^2 hybridized atoms in the minor groove among the 499 non-redundant structures. Most of the close contacts with cytosine involve the O2 atom of this base. Although lysine shows the highest number of sp^3 hybridized atom close contacts with cytosine, arginine shows a significant number of close contacts involving both sp^3 and sp^2 hybridized atoms and cytosine.

Among the amino acids containing cyclic species with delocalized electrons present in this research database, it is noteworthy that phenylalanine shows an appreciable number of close contacts in the minor groove. For phenylalanine, the greatest numbers of close contacts involve the cyclic sp^2 hybridized atoms and these

contacts are highest with the adenine and thymine bases. As described in Sections 4.3.3.1 and 4.3.3.3, although all of the heterocyclic carbons in phenylalanine interact with adenine and thymine, it is not to the same extent. The C_Z atom shows the highest number of close contacts with these bases in the minor groove. The position of the C_Z atom in the ring at the end of the side chain appears to make it more accessible to close contact. Phenylalanine also has a small number of sp^2 hybridized atom close contacts with guanine in the minor groove. The N2 atom of guanine offers good accessibility for the cyclic atoms in the minor groove. Interestingly, phenylalanine also shows a small number of sp^3 backbone atoms that come into close contacts with guanine in the minor groove. The planar structure of the cyclic ring may allow the backbone atoms of phenylalanine to approach guanine closely and also intercalate into the DNA structure in the minor groove.

4.4.3 Backbone and Side-chain Atom Close Contacts

As shown in Figure 4.4, close contacts with side-chain atoms far outnumber close contacts with backbone atoms in the minor groove for all bases studied. Although the absolute number of close backbone interactions overall is greater in the major groove, comparison of Figure 4.4 to Figure 3.4 shows that backbone atoms make a higher relative contribution to close contacts in the minor groove than in the major groove. Guanine and thymine have the highest comparative number of close contacts involving backbone atoms in the minor groove. For both guanine and thymine, these close contacts occur with a wide range of amino acids. Seventeen different amino acids show close contacts involving sp^2 hybridized backbone atoms with thymine in the minor groove. Additionally, 8 different amino acids show close contacts involving sp^3 hybridized

backbone atoms. For guanine, 15 different amino acids show close contacts involving sp^2 hybridized backbone atoms and 6 show close contacts involving sp^3 hybridized atoms.

By far, glycine shows the highest number of close contacts of any of the amino acids in the dataset studied. Because of its small size, and lack of a side chain (only H is attached to the backbone for glycine), the glycine backbone is extremely accessible. Glycine dominates backbone close contacts in the minor groove with all bases except adenine. For adenine, there are a greater number of close interactions involving backbone atoms with arginine than glycine.

CHAPTER 5
CONCLUSION

DNA-protein contacts show distinct microenvironments. Hybridization is a notable atomic feature that contributes to the distribution of charge within a protein. Hybridization assignments were made for all atoms within the 20 common amino acids that comprise proteins. Each atom has been assigned a hybridization of sp^2 or sp^3. Atoms with delocalized electrons have been assigned as sp^2 hybridized.

The Protein Data Bank (PDB) was utilized as the source for obtaining DNA-protein structures for investigation. Because the PDB contains such a large amount of data, there is significant redundancy in terms of structure and sequence. In order to study a more manageable amount of discreet data, a dataset of 499 non-redundant structures was generated. These structures were derived from information in the Protein Data Bank up to June 2011. Structures were chosen from a set of structures of high resolution, > 2.5Å. Results were clustered based on CLATH, SCOP and Pfam descriptions of proteins. This was an arduous process due to overlaps and inconsistencies between classifications generated by the three methods. Although each of the 499 non-redundant structures has been categorized based on the available methods, further development of protein description would be beneficial in order to bring the three methods into better alignment.

This investigation demonstrates that for the four DNA bases studied (adenine, thymine, guanine and cytosine), the potential hydrogen bonding sites, designated as W (major groove) and S (minor groove) are primary locations for close interaction with protein atoms in the major and minor grooves. The electrostatic as well as the steric environment of the base atoms contribute to the number and type of close contacts it

experiences with proteins. Although generalizations can be made in terms of hybridization-dependent DNA-protein interactions, the data in this research show hydrogen bond donor-acceptor relationships, as well as electrostatic (positive-negative) attractions, are the primary features regulating close contacts in the major grooves. Hydrogen bonds between amine and carbonyl functional groups contribute to the largest number of close contacts revealed in this research.

In general, atoms with a more electropositive environment show a greater number of close contacts with sp^2 hybridized atoms, and conversely atoms with a more electronegative environment show a greater number of close contacts with sp^3 hybridized atoms. However, for proteins comprised of amino acids with delocalized electrons, it is the charge (positive or negative) on the functional group, as well as the structure (cyclic or aliphatic) that have the greatest impact on the number and type of close contacts.

Arginine is the most basic amino acid, and exists in a cationic form containing a positively charged guanidinium group in neutral, acidic and basic environments. An important feature of arginine is the delocalization of electrons present in the guanidinium group. Since the nitrogen atoms on the side chain (N_{H1} and N_{H2}) can not be distinguished from each other, they are both treated as sp^2 hybridized with delocalization of charge. As a result of its structure and the presence of the positively charged guanidinium group, arginine acts a very good hydrogen bond donor. Arginine can actually donate several hydrogen bonds. Therefore, although most sp^2 hybridized side-chain atoms act as hydrogen bond acceptors (e.g. carbonyl) and interact to the greatest extent with electropositive species, the sp^2 hybridized atoms of arginine do not fit this generalization. The N_{H1} and N_{H2} atoms of arginine show a very large number of close interactions with

the carbonyl O6 atom as well as the electronegative N7 atom of guanine in the major groove. This is consistent with the postulate by Seeman *et al.* that arginine recognize guanine in the major groove[8]. Arginine also shows a comparatively high number of close contacts with the carbonyl O2 atom of thymine in the minor groove. A review of all close contacts involving the N_{H1} and N_{H2} atoms in the major and minor groove shows approximate equality in the number of close contacts of each type, providing support for the guanidinium structure and delocalization of charge. Further study of specific DNA-protein structures containing arginine to determine how arginine functions within these structures has significant merit and is a topic for further investigation.

Asparagine and glutamine, with their amide side chains, show a relatively high number of close contacts in the major groove. Since the amide side chain contains sp^2 hybridized carbonyl atoms and sp^3 hybridized amine atoms, these amino acids show a considerable number of close contacts with the four bases studied. The numbers of sp^3 hybridized atom close contacts are highest with adenine. This is consistent with Seeman's postulate that asparagine and glutamine recognize adenine in the major groove. In the minor groove, asparagine, and to a lesser extent glutamine, again show modestly high number of close contacts between the sp^3 hybridized side chain atoms and the four bases studied. Although asparagine and glutamine have the potential for delocalized electrons as a result of the amide side chain, the results of this research indicate that the N_{D2} atom of asparagine and the N_{E2} atom of glutamine are most appropriately given a hybridization assignment of sp^3. Close contacts involving the N_{D2} and N_{E2} atoms occur primarily with base atoms that are in a negative environment.

In the minor groove, there are a modest number of close contacts between the sp^2 hybridized side chain atoms in the heterocyclic ring of phenylalanine and the four bases under consideration. The atoms in the heterocyclic ring do not show equal numbers of close contacts with base atoms. The sp^2 hybridized C_Z atom of phenylalanine has the highest comparative number of close contacts with base atoms. The results of this research suggest that the planar structure of phenylalanine makes it well suited for close contact in the minor groove, indicating the planar ring is partially intercalated in the DNA structure and that these interactions potentially occur at sites of deformity.

The results of this research also show a difference in backbone versus side chain atom close contacts. For all bases, in both the major and minor grooves, close contacts with side chain atoms far outnumber close contacts with backbone atoms. Furthermore, in both the major and minor grooves, glycine shows the highest number of close contacts between its backbone atoms and the four bases. These backbone atom close contacts involve both sp^2 and sp^3 hybridized atoms of glycine. The small size of glycine makes it well suited for close contacts its backbone atoms in both the major and minor groove.

References

1. Smith E., and R. Hill, and I.R. Lehman et al. *Principles of Biochemistry*. 7th edn.. New York: McGraw-Hill Book Company, 1983.
2. Lodish, H., and A. Berk, and S. L. Zipursky. et al. *Molecular Cell Biology*. 4th edn.. New York: WH Freeman, 2000.
3. Watson, J.D., and F.H.C. Crick. "Molecular Structure of Nucleic Acids A Structure for Deoxyribose Nucleic Acids." *Nature* 171:737-738, 1953.
4. Branden, C, and J. Tooze. *Introduction to Protein Structure*. 2nd edn.. New York: Garland Publishing, 1999.
5. Wing, R., and H. Drew, and T. Takano, and C. Broka, and S. Tanaka, and K. Itakura, and R. Dickerson. "Crystal Structure Analysis of a Complete Turn of β-DNA." *Nature* 287: 755-758, 1980.
6. Pabo, C., and R. Sauer. "Protein-DNA Recognition." *Annu Rev Biochem* 53:293-321, 1984.
7. Lilley, D. *DNA-Protein Structural Interactions*, Oxford: Oxford University Press, 1995.
8. Seeman, N, and J. Rosenberg, and A. Rich. "Sequence-specific recognition of double helical nucleic acids by proteins." *Proc. Natl. Acad. Sci. USA* 73 (3): 804-808, 1976.
9. Li, Yun. "Understanding DNA-protein interactions from the nucleic-acid perspective." Ph.D. Thesis, Rutgers University, New Brunswick, NJ, 2001.
10. Kielkopf, Clara L., and Sarah White, and Jason W. Szewczyk, and James M. Turner, and Eldon E. Baird and Peter B. Dervan and Douglas Rees. "A structural basis for recognition of A-T and T-A base pairs in the minor groove of β-DNA." *Science* 282:111-115, 1998.
11. White, Sarah, and Jason W. Szewczyk, and James M. Turner, and Eldon E. Baird, and Peter B.Dervan. "Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands." *Nature*, 391: 468-471, 1998.
12. Travers, A.. *DNA-Protein Interactions*, London: Chapman Hill, 1993.
13. Mandel-Gutfreund, Yael, and Ora Schueler, and Hanah Margalit. "Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles." *J. Mol. Biol.*, 253(2):370-382, 1995.
14. Calladine, C.R. and H. R. Drew . *Understanding DNA; the Molecule and How it Works.* 2nd ed., London: Academic Press, 1997.
15. Xiang-Jun Lu and Wilma K. Olson. "3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures." *Nucleic Acid Research*, 31(17):5108-5121, 2003.

16. W.F. Cooper, G.A. Clark, and C.R. Hare. "A simple quantitative molecular orbital theory." *J. Chem. Educ.,* 48:247, 1971.

17. Pauling, Linus. "The nature of the chemical bond. III The transition from one extreme bond type to another." *J. Am. Chem. Society*, 54: 988-1003, 1932.

18. Milner-White, E. James. " The partial charge of the nitrogen atom in peptide bonds." *Protein Science,* 6:2477-2482, 1997.

19. Pauling, Linus and Robert B. Corey "Atomic coordinates and structure factors for two helical configurations of the polypeptide chain." *Proc. Natl. Acad. Sci. USA, 37: 235-40, 1951.*

20. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

21. "The Biotechnology Project: Proteins." The Biotechnology Project: Homepage. 2005. Web. 14 June 2011. http://biotech.matcmadison.edu/resources/proteins/labManual/chapter_2.htm

22. Pauling, Linus and Carl Neimann. "The structure of proteins." *J. Am. Chem. Soc.* 61, 1860-1867 1939.

23. Bernstein, F.C. and T.F. Koetzle, and G.J.Williams, and E.E. Meyer Jr., and M.D. Brice, and J.R. Rodgers, and O. Kennard, and T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A computer-based archival file for macromolecular structures." *J. of. Mol. Biol.,* 112 (1977): 535.

24. Orengo,C.A., and A.D. Michie, and S. Jones, and D. T. Jones, and M. B. Swindells, and J. M. Thornton. "CATH--a hierarchic classification of protein domain structures". *Structure* **5** (8): 1093–1108. 1997.

25. Murzin, A.G., and S. E. Brenner, and T. Hubbard, and C. Chothia. "SCOP: a structural classification of proteins database for the investigation of sequences and structures". *J. Mol. Biol.* **247** (4): 536–40, 1995.

26.Sonnhammer, E.L and S. R. Eddy, and R. Durbin. "Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins: Structure, Function, and Bioinformatics." *Protein* 28: 405–420, 1997.

27.Prlić, Andreas, and Spencer Bliven and Peter W. Rose and Wolfgang F. Bluhm and Chris Bizon and Adam Godzik and Philip E. Bourne "Pre-calculated protein structure alignments at the RCSB PDB website" *Bioinformatics* 26: 2983-2985, 2010.

28. Moreland, J. L., and A. Granada, and O. V. Buzko, and Q. Zhang, and P.E. Bourne "The Molecular Biology Toolkit (MBT): A modular platform for developing molecular visualization applications". *BMC Bioinformatics* 6:21. 2005.

29. The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.

30. Luger, K., and A. Mäder, and R. Richmond., and D. Sargent, and T. Richmond. "Crystal structure of the nucleosome core particle at 2.8 A resolution." *Nature* 389: 251-60, 1997.

31. Spiegelman B., and R. Henirich. "Biological control through transcriptional coactivators." *Cell* 119 (2): 157-167, 2004.

32. Carl R. Kemnitz and Mark J. Loewen J. "Amide resonance" correlates with a breadth of C-N rotation barriers." *J. Am. Chem. Soc*. 129(9) 2521 – 2528, 2007.

Appendix 1A
Listing of NDB/PDB File Numbers of DNA-Protein Complexes Studied

PDB ID	NDB ID	PDB ID	NDB ID	PDB ID	NDB ID	PDB ID	NDB ID
2BNW	2BNW	3AAF	NA0357	1BDT	PD0035	1QN3	PD0164
2BQ3	2BQ3	3KXT	NA0385	1B95	PD0038	1F44	PD0166
2C28	2C28	3L2C	NA0394	1B8I	PD0042	1F4K	PD0167
2C6Y	2C6Y	3KZ8	NA0395	4CRX	PD0047	1F4R	PD0168
2C7A	2C7A	3KTU	NA0405	6PAX	PD0050	1F6O	PD0172
2C7O	2C7O	3L4J	NA0416	1CKT	PD0051	1MJ2	PD0173
2C7P	2C7P	3LNQ	NA0441	1SSP	PD0052	1FIU	PD0177
2C7Q	2C7Q	3L8B	NA0449	2SSP	PD0053	1GD2	PD0180
2C9L	2C9L	3LWH	NA0459	1QPZ	PD0056	1FYL	PD0183
2J6S	2J6S	3LWI	NA0460	1QRH	PD0062	1FYM	PD0184
2J6U	2J6U	3AFA	NA0467	1QSS	PD0066	1G2F	PD0187
2JEI	2JEI	3LWL	NA0469	1QUM	PD0068	1G38	PD0188
2JEJ	2JEJ	3LWM	NA0470	1D3U	PD0070	1G9Y	PD0189
2V4R	2V4R	2WTF	NA0483	1BY4	PD0071	1HU0	PD0195
2VBJ	2VBJ	3M9E	NA0485	3HTS	PD0073	1I3J	PD0200
2VBL	2VBL	2WBS	NA0486	1B72	PD0075	1IAW	PD0207
2VBN	2VBN	2WBU	NA0487	2IRF	PD0076	1IG7	PD0211
2VBO	2VBO	3MFH	NA0489	1DC1	PD0085	1IJW	PD0213
2VE9	2VE9	3MDA	NA0493	1CEZ	PD0086	1JE8	PD0219
2VJU	2VJU	3MDC	NA0494	1BF4	PD0088	1JEY	PD0220
2VLA	2VLA	3M9N	NA0499	1HWT	PD0089	1JFT	PD0224
2VOA	2VOA	3M9O	NA0500	2HAP	PD0090	1E3O	PD0225
2VS7	2VS7	3MGV	NA0506	1DIZ	PD0099	1JJ4	PD0227
2VS8	2VS8	3M4A	NA0514	1CW0	PD0100	1JK2	PD0231
2VY1	2VY1	3MR2	NA0532	1DFM	PD0101	1JKO	PD0234
2VY2	2VY2	3MR3	NA0534	2CRX	PD0103	1JNM	PD0241
2W35	2W35	3MX4	NA0542	1DMU	PD0108	1MJM	PD0245
2W36	2W36	3MR4	NA0543	1QRV	PD0110	1MJO	PD0246
2W7N	2W7N	3MVA	NA0548	1DP7	PD0111	1MJQ	PD0248
2WIW	2WIW	3N7Q	NA0570	1DSZ	PD0115	1JX4	PD0251
2X6V	2X6V	3NAE	NA0583	1DUX	PD0116	1JXL	PD0252
3A46	NA0056	3NDH	NA0589	1EBM	PD0117	1K3W	PD0253
3HXO	NA0063	3MR5	NA0598	1EGW	PD0121	1K4T	PD0256
3A4K	NA0070	3MR6	NA0602	1QPI	PD0122	1K61	PD0257
3IAY	NA0085	3NGO	NA0603	1QAI	PD0126	1K78	PD0259
3IAG	NA0092	2XCS	NA0645	1EMH	PD0127	1K79	PD0260
3IGC	NA0104	2XC9	NA0717	1EOO	PD0132	1K82	PD0264
3IKT	NA0118	2XCA	NA0718	1ESG	PD0139	1KU7	PD0284
3IMB	NA0121	1CRX	PD0003	1EWQ	PD0142	1KX5	PD0287
3JR5	NA0141	3PVI	PD0006	1EYU	PD0147	1H89	PD0289
3JSO	NA0150	9ANT	PD0007	1D02	PD0151	1H8A	PD0290
3JTG	NA0167	1A6Y	PD0008	1QN4	PD0154	1L1T	PD0292
3JXZ	NA0172	1BGB	PD0010	1QN5	PD0155	1L3L	PD0298
3JY1	NA0173	1RV5	PD0013	1QN6	PD0156	1J59	PD0299
3JXB	NA0181	3HDD	PD0016	1QN7	PD0157	1L3S	PD0300
3K57	NA0201	1BC8	PD0020	1QN8	PD0158	1L3U	PD0302
3K58	NA0202	1B3T	PD0024	1QN9	PD0159	1L3V	PD0303
3K5M	NA0206	1BC7	PD0027	1QNA	PD0160	1H6F	PD0311
3KDE	NA0235	2KTQ	PD0030	1QNB	PD0161	1LLM	PD0312
3KHR	NA0322	3KTQ	PD0032	1QNC	PD0162	1LQ1	PD0314
3KMD	NA0356	4KTQ	PD0033	1QNE	PD0163	1LV5	PD0317

Appendix 1A cont.
Listing of NDB/PDB File Numbers of DNA-Protein Complexes Studied

PDB ID	NDB ID	PDB ID	NDB ID	PDB ID	NDB ID	PDB ID	NDB ID
2BNW	2BNW	3AAF	NA0357	1BDT	PD0035	1QN3	PD0164
2BQ3	2BQ3	3KXT	NA0385	1B95	PD0038	1F44	PD0166
2C28	2C28	3L2C	NA0394	1B8I	PD0042	1F4K	PD0167
2C6Y	2C6Y	3KZ8	NA0395	4CRX	PD0047	1F4R	PD0168
2C7A	2C7A	3KTU	NA0405	6PAX	PD0050	1F6O	PD0172
2C7O	2C7O	3L4J	NA0416	1CKT	PD0051	1MJ2	PD0173
2C7P	2C7P	3LNQ	NA0441	1SSP	PD0052	1FIU	PD0177
2C7Q	2C7Q	3L8B	NA0449	2SSP	PD0053	1GD2	PD0180
2C9L	2C9L	3LWH	NA0459	1QPZ	PD0056	1FYL	PD0183
2J6S	2J6S	3LWI	NA0460	1QRH	PD0062	1FYM	PD0184
2J6U	2J6U	3AFA	NA0467	1QSS	PD0066	1G2F	PD0187
2JEI	2JEI	3LWL	NA0469	1QUM	PD0068	1G38	PD0188
2JEJ	2JEJ	3LWM	NA0470	1D3U	PD0070	1G9Y	PD0189
2V4R	2V4R	2WTF	NA0483	1BY4	PD0071	1HU0	PD0195
2VBJ	2VBJ	3M9E	NA0485	3HTS	PD0073	1I3J	PD0200
2VBL	2VBL	2WBS	NA0486	1B72	PD0075	1IAW	PD0207
2VBN	2VBN	2WBU	NA0487	2IRF	PD0076	1IG7	PD0211
2VBO	2VBO	3MFH	NA0489	1DC1	PD0085	1IJW	PD0213
2VE9	2VE9	3MDA	NA0493	1CEZ	PD0086	1JE8	PD0219
2VJU	2VJU	3MDC	NA0494	1BF4	PD0088	1JEY	PD0220
2VLA	2VLA	3M9N	NA0499	1HWT	PD0089	1JFT	PD0224
2VOA	2VOA	3M9O	NA0500	2HAP	PD0090	1E3O	PD0225
2VS7	2VS7	3MGV	NA0506	1DIZ	PD0099	1JJ4	PD0227
2VS8	2VS8	3M4A	NA0514	1CW0	PD0100	1JK2	PD0231
2VY1	2VY1	3MR2	NA0532	1DFM	PD0101	1JKO	PD0234
2VY2	2VY2	3MR3	NA0534	2CRX	PD0103	1JNM	PD0241
2W35	2W35	3MX4	NA0542	1DMU	PD0108	1MJM	PD0245
2W36	2W36	3MR4	NA0543	1QRV	PD0110	1MJO	PD0246
2W7N	2W7N	3MVA	NA0548	1DP7	PD0111	1MJQ	PD0248
2WIW	2WIW	3N7Q	NA0570	1DSZ	PD0115	1JX4	PD0251
2X6V	2X6V	3NAE	NA0583	1DUX	PD0116	1JXL	PD0252
3A46	NA0056	3NDH	NA0589	1EBM	PD0117	1K3W	PD0253
3HXO	NA0063	3MR5	NA0598	1EGW	PD0121	1K4T	PD0256
3A4K	NA0070	3MR6	NA0602	1QPI	PD0122	1K61	PD0257
3IAY	NA0085	3NGO	NA0603	1QAI	PD0126	1K78	PD0259
3IAG	NA0092	2XCS	NA0645	1EMH	PD0127	1K79	PD0260
3IGC	NA0104	2XC9	NA0717	1EOO	PD0132	1K82	PD0264
3IKT	NA0118	2XCA	NA0718	1ESG	PD0139	1KU7	PD0284
3IMB	NA0121	1CRX	PD0003	1EWQ	PD0142	1KX5	PD0287
3JR5	NA0141	3PVI	PD0006	1EYU	PD0147	1H89	PD0289
3JSO	NA0150	9ANT	PD0007	1D02	PD0151	1H8A	PD0290
3JTG	NA0167	1A6Y	PD0008	1QN4	PD0154	1L1T	PD0292
3JXZ	NA0172	1BGB	PD0010	1QN5	PD0155	1L3L	PD0298
3JY1	NA0173	1RV5	PD0013	1QN6	PD0156	1J59	PD0299
3JXB	NA0181	3HDD	PD0016	1QN7	PD0157	1L3S	PD0300
3K57	NA0201	1BC8	PD0020	1QN8	PD0158	1L3U	PD0302
3K58	NA0202	1B3T	PD0024	1QN9	PD0159	1L3V	PD0303
3K5M	NA0206	1BC7	PD0027	1QNA	PD0160	1H6F	PD0311
3KDE	NA0235	2KTQ	PD0030	1QNB	PD0161	1LLM	PD0312
3KHR	NA0322	3KTQ	PD0032	1QNC	PD0162	1LQ1	PD0314
3KMD	NA0356	4KTQ	PD0033	1QNE	PD0163	1LV5	PD0317

Appendix 1A cont.
Listing of NDB/PDB File Numbers of DNA-Protein Complexes Studied

PDB ID	NDB ID	PDB ID	NDB ID	PDB ID	NDB ID	PDB ID	NDB ID
1IXY	PD0333	1TX3	PD0577	2DTU	PD0837	2PYL	PD1012
1M5X	PD0335	1U8B	PD0584	2HOF	PD0838	2Z3X	PD1014
1MNN	PD0341	1XO0	PD0601	2HR1	PD0842	2Q2T	PD1017
1MOW	PD0342	1XSL	PD0604	2HT0	PD0843	2QHB	PD1020
1MUS	PD0350	1XYI	PD0607	2HVH	PD0844	2QOJ	PD1029
1N48	PD0358	1WTE	PD0608	2HVI	PD0845	2R1J	PD1037
1N4L	PD0359	1WTO	PD0609	2HW3	PD0846	2RDJ	PD1044
1N6J	PD0363	1WTV	PD0613	2I06	PD0849	2RBF	PD1052
1NH2	PD0369	1YF3	PD0622	2I13	PD0851	3BEP	PD1054
1J1V	PD0371	1YO5	PD0627	2I9G	PD0865	3BIE	PD1055
1NLW	PD0387	1YQK	PD0629	2IBS	PD0867	3BS1	PD1057
1J3E	PD0390	1YQM	PD0631	2IBT	PD0868	3BI3	PD1059
1NVP	PD0393	1YQR	PD0632	2IA6	PD0869	3BKZ	PD1060
1OMH	PD0394	1VRL	PD0634	2IBK	PD0870	3BM3	PD1062
1ORN	PD0396	1ZG1	PD0645	2IH4	PD0876	3BRD	PD1067
1ORP	PD0397	1ZG5	PD0646	2IH5	PD0877	3BRF	PD1068
1OUP	PD0403	1ZET	PD0647	2IIE	PD0878	3BRG	PD1069
1OWF	PD0406	1ZNS	PD0657	2IHM	PD0880	3C0W	PD1075
1OZJ	PD0409	1ZRF	PD0673	2IMW	PD0881	3C0X	PD1076
1P47	PD0424	2CV5	PD0676	2ITL	PD0882	3C1B	PD1077
1P71	PD0430	1ZS4	PD0687	2NL8	PD0883	3C2I	PD1080
1P78	PD0431	2A07	PD0689	2ISZ	PD0884	3C29	PD1082
1PJJ	PD0447	2A66	PD0691	2E1C	PD0889	3C58	PD1092
1PP7	PD0452	2ADY	PD0694	2NOB	PD0891	3CBB	PD1094
1PT3	PD0454	2AGQ	PD0699	2NOE	PD0892	2ZKD	PD1096
1PUF	PD0455	2ALZ	PD0700	2NOF	PD0893	3CLZ	PD1099
1PVP	PD0457	2ATA	PD0705	2NOH	PD0894	3CMY	PD1100
1R0O	PD0471	2ASD	PD0708	2NOI	PD0895	3COQ	PD1101
1R2Z	PD0474	2ASJ	PD0709	2NOZ	PD0897	3COA	PD1104
1R4O	PD0475	2EUV	PD0728	2IS6	PD0901	3CFP	PD1105
1R71	PD0480	2EX5	PD0738	2NP6	PD0904	3CFR	PD1106
1R7M	PD0481	2D5V	PD0739	2NP7	PD0905	3CQ8	PD1111
1R8E	PD0483	2FCC	PD0747	2NTC	PD0908	3C25	PD1131
1RIO	PD0492	2FLD	PD0765	2NQB	PD0914	2ZO0	PD1133
1RH6	PD0494	2GEQ	PD0791	2O49	PD0946	2ZO1	PD1134
1RZT	PD0503	2GIG	PD0793	2O4A	PD0947	3D0A	PD1137
1S0O	PD0510	2GIH	PD0794	2O6M	PD0950	3D0P	PD1139
1SA3	PD0515	2GII	PD0795	2E52	PD0956	3C5G	PD1141
1SKR	PD0522	2GIJ	PD0796	2OAA	PD0958	2QL2	PD1152
1SXQ	PD0534	2DPJ	PD0805	2OFI	PD0960	3DPG	PD1157
1S9F	PD0535	2H1K	PD0810	2OG0	PD0961	3DVO	PD1161
1T2T	PD0538	2H27	PD0813	2EA0	PD0968	3DW9	PD1162
1T3N	PD0543	2H7H	PD0818	2OPF	PD0969	3E54	PD1179
1WD0	PD0544	2HAN	PD0822	2OXV	PD0974	3E6C	PD1180
1WD1	PD0545	2HHQ	PD0824	2EFW	PD0975	3ECP	PD1182
1T9I	PD0547	2HHS	PD0825	2OWO	PD0978	3EEO	PD1183
1TDZ	PD0551	2HHT	PD0826	2P66	PD0983	3EPG	PD1189
1TEZ	PD0552	2HHU	PD0827	2YVH	PD0993	3EXJ	PD1191
1TK0	PD0554	2HHV	PD0828	2PYO	PD1003	3EYZ	PD1198
1RM1	PD0559	2HHW	PD0829	2PI0	PD1009	3EZ5	PD1199
1U0C	PD0561	2HHX	PD0830	2PYJ	PD1011	3F8J	PD1204

Appendix 1A cont.
Listing of NDB/PDB File Numbers of DNA-Protein Complexes Studied

PDB ID	NDB ID	PDB ID	NDB ID
3FDQ	PD1212	1PDN	PDR018
3EY1	PD1213	2NLL	PDR021
3G00	PD1221	1TUP	PDR027
3G0R	PD1222	1AIS	PDR031
3G2C	PD1223	1A3Q	PDR032
3G2D	PD1224	1AU7	PDR034
3FYL	PD1225	1MNM	PDR036
3G3Y	PD1228	1AZP	PDR047
3G73	PD1230	1AZQ	PDR048
3G6P	PD1231	1AKH	PDR049
3G6Q	PD1232	2RAM	PDR051
3G6R	PD1233	1BL0	PDR056
3G6T	PD1234	1HCQ	PDRC03
3G6U	PD1235	1YTB	PDT012
3G97	PD1238	PDT013	PDT013
3G99	PD1239	1NFK	PDT015
3GA6	PD1240	1YRN	PDT028
3G9I	PD1241	2DGC	PDT029
3G9J	PD1242	1LAT	PDT030
3G9M	PD1243	1PUE	PDT033
3G0Q	PD1246	1CDW	PDT034
3G6V	PD1247	1IGN	PDT035
3GIJ	PD1253	1YTF	PDT036
3GDX	PD1257	1UBD	PDT038
3GQC	PD1271	1AAY	PDT039
3GXQ	PD1276	1IHF	PDT040
3H0D	PD1278	2HDD	PDT043
3GV5	PD1280	1ZME	PDT044
3H40	PD1285	1XBR	PDT045
3H4D	PD1287	1AWC	PDT048
3H8X	PD1290	1A1F	PDT055
1DNK	PDE005	1A1H	PDT057
2DNJ	PDE006	1A1K	PDT058
1HCR	PDE009	1AM9	PDT062
1BPX	PDE0126	1SKN	PDT064
1TC3	PDE0128	1MEY	PDTB41
2BDP	PDE0131	2BOP	PDV001
4BDP	PDE0133	2O5I	PH0036
1RVA	PDE014	2QKB	PH0041
1A31	PDE0142	3BSU	PH0045
1IPP	PDE0144		
1BHM	PDE020		
1FJL	PDE025		
6MHT	PDE141		
3CRO	PDR001		
2OR1	PDR004		
1TRO	PDR009		
1LMB	PDR010		
1RPE	PDR011		
1TRR	PDR013		

Appendix 1B
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0883	2NL8	DNA binding protein	3 layer sandwich	none	replication binding protein
PD0908	2NTC	DNA binding protein	3 layer sandwich	origin of replication binding protein	origin of replication binding protein
PD0882	2ITL	DNA binding protein	3 layer sandwich	origin of replication binding protein	origin of replication binding protein
PD0142	1EWQ	Replication DNA	3 layer sandwich/2 layer sandwich/orthogonal bundle	DNA repair protein	none
PD0192	1HLV	DNA binding protein	orthogonal bundle	DNA/RNA-binding 3-helical bundle	none
PD0070	1D3U	DNA Gene regulation	2 layer sandwich/orthogonal bundle	TATA-box binding protein-like/Cyclin-like	Transcription factor
PD0694	2ADY	Aptosis/DNA	sandwich	P53 like transciption factor	P53 binding domain
PD0705	2ATA	Aptosis/DNA	sandwich	P53 like transciption factor	P53 binding domain
NA0467	3AFA	Structural Protein/DNA	orthogonal bundle	none	histone
PD1014	2Z3X	DNA binding protein	none	none	spore protein
PD1020	2QHB	DNA binding protein	none	DNA/RNA-binding 3-helical bundle	none
PD1212	3FDQ	Transport protein	up down bundle	none	Acyl CoA binding protein
PD1276	3GXQ	DNA binding protein	none	none	none
NA0235	3KDE	DNA binding protein	none	none	THAP domain
NA0356	3KMD	DNA binding protein	none	none	P53 binding domain
NA0357	3AAF	DNA binding protein	none	none	RQC domain
NA0385	3KXT	DNA binding protein	none	none	Chromatin protein cren7

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
NA0459	3LWH	DNA binding protein	none	none	Chromatin protein cren7
NA0460	3LWI	DNA binding protein	none	none	Chromatin protein cren7
PH0036	2O5I	DNA directed RNA polymerase	2 layer sandwich/beta complex	none	DNA directed RNA polymerase
PD0006	3PVI	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction endonuclease
PD0062	1QRH	Hydrolase/DNA	4 layer sandwich	restriction endonuclease	restriction endonuclease
PD0068	1QUM	Hydrolase/DNA	alpha beta barrel	TIM beta/alpha-barrel	TIM barrel
2VOA	2VOA	Lyase	4 layer sandwich	none	endonuclease/ exonuclease
2VERSUS7	2VERSUS7	DNA binding protein	roll endonuclease	none	none
2VERSUS8	2VERSUS8	DNA binding protein	roll endonuclease	none	none
PD0010	1BGB	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction endonuclease
PD0013	1RV5	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction endonuclease
PD0038	1B95	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction endonuclease
PD0085	1DC1	Hydrolase/DNA	3 layer sandwich/orthogonal bundle	restriction endonuclease	restriction endonuclease
PD0100	1CW0	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	endonuclease
PD0101	1DFM	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	DNA restriction modification
PD0108	1DMU	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	none

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0132	1EOO	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	DNA restriction modification
PD0139	1ESG	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	DNA restriction modification
PD0147	1EYU	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	DNA restriction modification
PD0151	1D02	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	DNA restriction modification
PD0177	1FIU	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction enzyme
PD0189	1G9Y	Hydrolase/DNA	roll endonuclease	Homig endonuclease	Homig endonuclease
PD0200	1I3J	Hydrolase/DNA	none	DNA-binding domain of intron-encoded endonucleases	none
PD0207	1IAW	Hydrolase/DNA	3 layer sandwich/orthogonal bundle	none	restriction nuclease
PD0335	1M5X	Hydrolase/DNA	roll endonuclease	Homig endonuclease	LAGLIDADG endonuclease
PD0342	1MOW	Hydrolase/DNA	roll endonuclease	Homig endonuclease	LAGLIDADG endonuclease
PD0403	1OUP	Hydrolase/DNA	none	His-Me-finger endonuclease	endonuclease I
PD0481	1R7M	Hydrolase/DNA	roll endonuclease	Homig endonuclease	LAGLIDADG endonuclease
PD0502	1RXW	Hydrolase/DNA	none	SAM/PIN domain like	XPG terminal domain
PD0515	1SA3	Hydrolase/DNA	none	restriction endonuclease	restriction endonuclease
PD0538	1T2T	Hydrolase/DNA	none	DNA binding domain of intron-encoded endonculeases	none

Appendix 1B cont.

Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0547	1T9I	Hydrolase/DNA	roll endonuclease	Homig endonuclease	LAGLIDADG endonuclease
PD0561	1U0C	Hydrolase/DNA	roll endonuclease	Homig endonuclease	LAGLIDADG endonuclease
PD0577	1TX3	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction nuclease
PD0608	1WTE	Hydrolase/DNA	3 layer sandwich/orthogonal bundle	restriction endonuclease	none
PDE014	1RVA	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction nuclease
PDE0144	1IPP	Transcription DNA	alpha beta complex	His-Me-finger endonuclease	none
PDE020	1BHM	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	restriction nuclease
2VBJ	2VBJ	Hydrolase	roll endonuclease	none	LAGLIDADG endonuclease
2VBL	2VBL	Hydrolase	roll endonuclease	none	LAGLIDADG endonuclease
2VBN	2VBN	Hydrolase	roll endonuclease	none	LAGLIDADG endonuclease
2VBO	2VBO	Hydrolase	roll endonuclease	none	LAGLIDADG endonuclease
2VLA	2VLA	Hydrolase	none	none	restriction endonuclease
2W35	2W35	Hydrolase	none	none	Endonuclease V
2W36	2W36	Hydrolase	none	none	Endonuclease V
2WIW	2WIW	Hydrolase/DNA	3 layer sandwich	none	Archael junction resolve
NA0056	3A46	Hydrolase	none	none	Formamidopyrimidine-DNA glycosylase N-terminal domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
NA0070	3A4K	Hydrolase/DNA	none	none	restriction endonuclease
NA0121	3IMB	Hydrolase/DNA	none	none	none
NA0141	3JR5	Lyase/DNA	none	none	DNA glycoslase
NA0150	3JSO	Hydrolase/DNA	none	none	Lexa DNA binding domain
NA0172	3JXZ	Hydrolase/DNA	none	none	DNA alkylation repair enzyme
NA0173	3JY1	Hydrolase/DNA	none	none	DNA alkylation repair enzyme
NA0405	3KTU	Hydrolase Lyase/DNA	none	none	DNA glycoslase
NA0542	3MX4	Hydrolase/DNA	none	none	restriction endonuclease
NA0589	3NDH	Hydrolase/DNA	none	none	none
NA0603	3NGO	Hydrolase/DNA	none	none	endonuclease/exonuclease
PD0052	1SSP	Hydrolase/DNA	3 layer sandwich	uracil-DNA glycosylase	uracil-DNA glycosylase
PD0053	2SSP	Protein/DNA	3 layer sandwich	uracil-DNA glycosylase	uracil-DNA glycosylase
PD0099	1DIZ	Hydrolase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA repair protein
PD0117	1EBM	Lyase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycoslase
PD0127	1EMH	Hydrolase/DNA	3 layer sandwich	uracil-DNA glycosylase	uracil-DNA glycosylase
PD0168	1F4R	Hydrolase/DNA	roll	DNA glycosylase	DNA glycosylase
PD0172	1F6O	Hydrolase/DNA	roll	DNA glycosylase	DNA glycosylase

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0195	1HU0	Hydrolase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase
PD0253	1K3W	Hydrolase/DNA	none	DNA repair protein	DNA glycosylase
PD0264	1K82	Hydrolase/DNA	none	DNA repair protein/Zn Finger	DNA glycosylase/Zn Finger
PD0292	1L1T	Hydrolase/DNA	none	DNA repair protein/Zn Finger	none
D0396	1ORN	Hydrolase/DNA	orthogonal bundle	DNA glycosylase	none
PD0397	1ORP	Hydrolase/DNA	orthogonal bundle	DNA glycosylase	none
PD0447	1PJJ	Hydrolase/DNA	none	DNA repair protein	DNA glycoslase
PD0452	1TEZ	Lyase/DNA	3 layer sandwich/orthogonal bundle/alpha horseshoe	Cryptochrome/photolyase FAD-binding domain	DNA_photolyase
PD0474	1R2Z	Hydrolase/DNA	none	Glucocorticoid receptor-like (DNA-binding domain)	DNA glycosylase
PD0496	1RRS	Hydrolase/DNA	alpha beta complex/orthogonal bundle	DNA glycosylase	DNA repair protein/endonuclease
PD0551	1TDZ	Hydrolase/DNA	none	Glucocorticoid receptor-like (DNA-binding domain)	DNA glycosylase
PD0577	1TX3	Hydrolase/DNA	3 layer sandwich	restriction endonuclease	Restrictin endonculease
PD0629	1YQK	Hydrolase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase
PD0631	1YQM	Hydrolase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase
PD0631	1YQM	Hydrolase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0632	1YQR	Hydrolase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase
PD0634	1VRL	Hydrolase/DNA	alpha beta complex/orthogonal bundle	DNA glycosylase/Nudix	DNA repair protein/endonuclease
PD0657	1ZNS	Hydrolase/DNA	alpha beta complex	none	none
PD0738	2EX5	Hydrolase	none	none	LAGLIDADG endonuclease
PD0747	2FCC	Hydrolase/DNA	orthogonal bundle	T4 endonuclease V	endonculease V
PD0765	2FLD	Hydrolase/DNA	Roll	none	LAGLIDADG endonuclease
PD0793	2GIG	Hydrolase/DNA	2 layer sandwich	none	Restrictin endonclease
PD0794	2GIH	Hydrolase/DNA	3 layer sandwich	none	Restrictin endonclease
PD0795	2GII	Hydrolase/DNA	3 layer sandwich	none	Restrictin endonclease
PD0796	2GIJ	Hydrolase/lyase/DNA	3 layer sandwich	none	Restrictin endonclease
PD0891	2NOB	Hydrolase/lyase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase/DNA repair protein
PD0892	2NOE	Hydrolase/lyase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase/DNA repair protein
PD0893	2NOF	Hydrolase/lyase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase/DNA repair protein
PD0894	2NOH	Hydrolase/lyase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase/DNA repair protein
PD0895	2NOI	Hydrolase/lyase/DNA	2 layer sandwich/orthogonal bundle	none	DNA glycosylase/DNA repair protein

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0897	2NOZ	Hydrolase/lyase/DNA	2 layer sandwich/orthogonal bundle	DNA glycosylase/TATA box	DNA glycosylase/DNA repair protein
PD0901	2IS6	Hydrolase/DNA	3 layer sandwich/orthogonal bundle	none	UvrD/REP helicase
PD0950	2O6M	Hydrolase/DNA	alpha beta complex	none	none
PD0956	2E52	Hydrolase/DNA	none	none	none
PD0958	2OAA	3 Methyladenine DNA Glycosylase I/dna	orthogonal bundle	none	methyadenine glycosylase
PD0960	2OFI	3 Methyladenine DNA Glycosylase I/dna	orthogonal bundle	none	methyadenine glycosylase
PD0968	2EA0	Hydrolase/DNA	none	Glucocorticoid receptor-like (DNA-binding domain)	DNA glycosylase
PD0969	2OPF	Hydrolase/DNA	none	Glucocorticoid receptor-like (DNA-binding domain)	DNA glycosylase
PD0974	2OXV	Hydrolase/DNA	3 layer sandwich	none	Restrictin endonclease
PD1029	2QOJ	Hydrolase/DNA	Roll	none	LAGLIDADG endonuclease
PD1062	3BM3	Hydrolase/DNA	none	none	EcoRII C terminal
PD1075	3C0W	Hydrolase/DNA	Roll	none	LAGLIDADG endonuclease
PD1076	3C0X	Oxidoreductase(oxygen Receptor)	2 layer sandwich/3 layer sandwich	FAD/NAD(P)-binding domain	GMC oxidoreductase
PD1092	3C58	Hydrolase/DNA	none	none	DNA glycosylase
PD1131	3C25	Hydrolase/DNA	none	none	restriction endonuclease
PD1139	3D0P	Hydrolase/DNA	none	Ribonuclease H-like	RNase H

Appendix 1B cont.

Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD1157	3DPG	Hydrolase/DNA	none	none	restriction endonuclease
PD1161	3DVO	Hydrolase/DNA	none	none	restriction endonuclease
PD1162	3DW9	Hydrolase/DNA	none	none	restriction endonuclease
PD1179	3E54	Hydrolase/DNA	none	none	LAGLIDADG endonuclease
PD1213	3EY1	Hydrolase/DNA	none	none	RNase H
PD1221	3G00	Hydrolase/DNA	none	none	Endonuclease/Exonuclease/phosphatase family
PD1222	3G0R	Hydrolase/DNA	none	none	Endonuclease/Exonuclease/phosphatase family
PD1223	3G2C	Hydrolase/DNA	none	none	Endonuclease/Exonuclease/phosphatase family
PD1224	3G2D	Hydrolase/DNA	none	none	Endonuclease/Exonuclease/phosphatase family
PD1228	3G3Y	Hydrolase/DNA	none	none	Endonuclease/Exonuclease/phosphatase family
PD1240	3GA6	Hydrolase/DNA	none	none	Endonuclease/Exonuclease/phosphatase family
PH0041	2QKB	Uncharacterized	none	none	none
PH0045	3BSU	Hydrolase/DNA	3 layer sandwich	none	Caulimovirus viroplasmin
PD0978	2OWO	Ligase/DNA	none	none	DNA ligase

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD1017	2Q2T	Ligase/DNA	none	DNA ligase/nucleic acid binding protein	ATP dependent DNA ligase domain
PD1096	2ZKD	Ligase	none	PUA domain-like	YDG/SRA domain
PD1099	3CLZ	Ligase	none	PUA domain-like	YDG/SRA domain
PD1133	2ZO0	Ligase/DNA	none	PUA domain-like	YDG/SRA domain
PD1134	2ZO1	Ligase/DNA	none	PUA domain-like	YDG/SRA domain
PD1204	3F8J	Ligase/DNA	none	none	YDG/SRA domain
PD0454	1PT3	Hydrolase/DNA	alpha beta complex	His-Me finger endonucleases	none
PDE005	1DNK	Hydrolase/DNA	4 layer sandwich	DNase I-like	Endonuclease/Exonuclease/phosphatase family
PDE006	2DNJ	Hydrolase/DNA	4 layer sandwich	DNase I-like	Endonuclease/Exonuclease/phosphatase family
2C7O	2C7O	Transferase	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases	C-5 cytosine-specific DNA methylase
2C7P	2C7P	Transferase	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases	C-5 cytosine-specific DNA methylase
2C7Q	2C7Q	Transferase	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases	C-5 cytosine-specific DNA methylase
PD0188	1G38	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases	Eco57I restriction-modification methylase
PD0622	1YF3	Transferase/DNA	2 layer sandwich/orthogonal bundle	S-adenosyl-L-methionine-dependent methyltransferases	D12 class N6 adenine-specific DNA methyltransferase

Appendix 1B cont.

Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0842	2HR1	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases	C-5 cytosine-specific DNA methylase
PD0867	2IBS	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases /DNA methylase specificity domain	Eco57I restriction-modification methylase
PD0868	2IBT	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases /DNA methylase specificity domain	Eco57I restriction-modification methylase
PD0876	2IH4	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases /DNA methylase specificity domain	Eco57I restriction-modification methylase
PD0877	2IH5	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases /DNA methylase specificity domain	Eco57I restriction-modification methylase
PD0904	2NP6	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases /DNA methylase specificity domain	Eco57I restriction-modification methylase
PD0905	2NP7	Transferase/DNA	3 layer sandwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases /DNA methylase specificity domain	Eco57I restriction-modification methylase

Appendix 1B cont.

Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD1183	3EEO	Transferase/DNA	3 layer sandwich/alpha beta complex	none	C-5 cytosine-specific DNA methylase
PD0552	1TEZ	Lyase/DNA	3 layer sandwich/orthogonal bundle/alpha horseshoe	Cryptochrome/photolyase FAD-binding domain/Cryptochrome/photolyase, N-terminal domain	DNA photolyase/FAD binding domain of DNA photolyase
2BQ3	2BQ3	Transferase	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
2C28	2C28	Polymerase	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
2J6S	2J6S	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
2J6U	2J6U	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
2JEI	2JEI	Transferase	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
2JEJ	2JEJ	Transferase	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
2V4R	2V4R	Transferase	2 layer sandwich/orthogonal bundle	DNA RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0085	3IAY	Transferase/DNA	none	none	DNA polymerase family B
NA0201	3K57	Transferase/DNA	none	none	DNA polymerase family B
NA0202	3K58	Transferase/DNA	none	none	DNA polymerase family B
NA0206	3K5M	Transferase/DNA	none	none	DNA polymerase family B
NA0322	3KHR	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0449	3L8B	Transferase/DNA	none	none	DNA polymerase family A
NA0469	3LWL	Transferase/DNA	none	none	DNA polymerase family B/Taq polymerase, exonuclease
NA0470	3LWM	Transferase/DNA	none	none	DNA polymerase family B/Taq polymerase, exonuclease

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
NA0483	2WTF	Transferase/DNA	2 layer sandwich	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0489	3MFH	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0493	3MDA	Lyase , Transferase/dna	none	none	Fingers domain of DNA polymerase lambda
NA0494	3MDC	Lyase , Transferase/dna	none	none	Fingers domain of DNA polymerase lambda
NA0499	3M9N	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0500	3M9O	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0532	3MR2	Transferase/DNA	none	none	impB/mucB/samB family
NA0534	3MR3	Transferase/DNA	none	none	impB/mucB/samB family
NA0543	3MR4	Transferase/DNA	none	none	impB/mucB/samB family

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
NA0583	3NAE	Transferase/DNA	none	none	impB/mucB/samB family
NA0598	3MR5	Transferase/DNA	none	none	impB/mucB/samB family
NA0602	3MR6	Transferase/DNA	none	none	impB/mucB/samB family
NA0717	2XC9	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
NA0718	2XCA	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0030	2KTQ	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	Taq polymerase, exonuclease/DNA polymerase family A
PD0032	3KTQ	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	Taq polymerase, exonuclease/DNA polymerase family A
PD0033	4KTQ	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	Taq polymerase, exonuclease/DNA polymerase family A
PD0066	1QSS	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	Taq polymerase, exonuclease/DNA polymerase family A

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0251	1JX4	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0252	1JXL	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0300	1L3S	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	DNA polymerase family A
PD0302	1L3U	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	DNA polymerase family A
PD0303	1L3V	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	DNA polymerase family A
PD0317	1LV5	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polymerase	DNA polymerase family A
PD0358	1N48	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polymerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0503	1RZT	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA polymerase beta, N-terminal domain-like/PsbU/PolX domain-like/Nucleotidyltransferase	Fingers domain of DNA polymerase lambda
PD0510	1S0O	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0522	1SKR	Transferase/electron Transport/dna	2 layer sandwich/3 layer sandwich/orthogonal bundle/Up-down bundle	DNA/RNA Polyerase/Ribonculease H-like/Thioredoxin-like	DNA polymerase family A/Thioredoxin
PD0535	1S9F	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0543	1T3N	Replication/dna	2 layer sandwich/orthogonal bundle	DNA/RNA PolyeraseLesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0554	1TK0	Transferase/electron Transport/dna	2 layer sandwich/3 layer sandwich/orthogonal bundle/Up-down bundle	DNA/RNA Polyerase/Ribonculease H-like/Thioredoxin-like	DNA polymerase family A/Thioredoxin

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0604	1XSL	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA polymerase beta, N-terminal domain-like/PsbU/PolX domain-like/Nucleotidyltransferase	Fingers domain of DNA polymerase lambda
PD0647	1ZET	Replication/dna	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0699	2AGQ	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0700	2ALZ	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0708	2ASD	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0709	2ASJ	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0805	2DPJ	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0824	2HHQ	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0825	2HHS	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0826	2HHT	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0827	2HHU	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0828	2HHV	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0829	2HHW	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0830	2HHX	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family A
PD0837	2DTU	Transferase/DNA	2 layer sandwich	none	DNA polymerase family B, exonuclease domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0838	2HOF	Recombination DNA	orthogonal bundle	lambda integrase-like, N-terminal domain/DNA breaking-rejoining enzymes	Phage integrase family
PD0844	2HVH	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonuclease H-like	DNA polymerase family A
PD0845	2HVI	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonuclease H-like	DNA polymerase family A
PD0846	2HW3	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	DNA/RNA Polyerase/Ribonuclease H-like	DNA polymerase family A
PD0865	2I9G	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA polymerase beta, N-terminal domain-like/PsbU/PolX domain-like/Nucleotidyltransferase/PsbU/PolX domain-like	Fingers domain of DNA polymerase lambda
PD0869	2IA6	Transferase/DNA	2 layer sandwich/orthogonal bundle	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0870	2IBK	Transferase/DNA	2 layer sandwich/orthogonal bundle	polymerase (Y-family), little finger domain Lesion bypass DNA polymerase (Y-family), little finger domain /DNA/RNA polymerases	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0880	2IHM	Transferase/DNA	none	none	Fingers domain of DNA polymerase lambda
PD0881	2IMW	Transferase/DNA	2 layer sandwich/orthogonal bundle	polymerase (Y-family), little finger domain Lesion bypass DNA polymerase (Y-family), little finger domain /DNA/RNA polymerases	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD0983	2P66	Transferase , Lyase/dna	2 layer sandwich/orthogonal bundle	DNA polymerase beta, N-terminal domain-like/PsbU/PolX domain-like/Nucleotidyltransferase	Fingers domain of DNA polymerase lambda
PD1011	2PYJ	Replication , Transferase/dna	2 layer sandwich/orthogonal bundle/Alpha-Beta Complex/irregular	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase type B, organellar and viral
PD1012	2PYL	Replication , Transferase/dna	2 layer sandwich/orthogonal bundle/Alpha-Beta Complex/irregular	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase type B, organellar and viral
PD1044	2RDJ	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA Polyerase/Lesion bypass DNA polymerase (Y-family), little finger domain	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD1054	3BEP	Transferase , Transcription/dna	RolI	DNA clamp	DNA polymerase III beta subunit, N-terminal domain
PD1105	3CFP	Transferase/DNA	none	none	DNA polymerase family B, exonuclease domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD1106	3CFR	Transferase/DNA	none	none	DNA polymerase family B, exonuclease domain
PD1111	3CQ8	Transferase/DNA	none	DNA/RNA Polyerase/Ribonculease H-like	DNA polymerase family B, exonuclease domain
PD1141	3C5G	Transferase , Lyase/dna	2 layer sandwich/orthogonal bundle	DNA polymerase beta, N-terminal domain-like/PsbU/PolX domain-like/Nucleotidyltransferase/PsbU/PolX domain-like	Fingers domain of DNA polymerase lambda
PD1189	3EPG	Transferase/DNA	2 layer sandwich/orthogonal bundle	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD1198	3EYZ	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	none	DNA polymerase family A
PD1199	3EZ5	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	none	DNA polymerase family A
PD1247	3G6V	Replication DNA	2 layer sandwich/orthogonal bundle	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD1253	3GIJ	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD1257	3GDX	Transferase/DNA	2 layer sandwich/orthogonal bundle	none	Fingers domain of DNA polymerase lambda
PD1271	3GQC	Transferase/DNA	none	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD1280	3GV5	Transferase/DNA	2 layer sandwich/orthogonal bundle	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD1285	3H40	Replication DNA	2 layer sandwich/orthogonal bundle	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PD1287	3H4D	Replication DNA	2 layer sandwich/orthogonal bundle	none	impB/mucB/samB family/IMS family HHH motif /impB/mucB/samB family C-terminal
PDE0126	1BPX	Transferase/DNA	2 layer sandwich/orthogonal bundle	DNA polymerase beta, N-terminal domain-like/PsbU/PolX domain-like/Nucleotidyltransferase/PsbU/PolX domain-like	Fingers domain of DNA polymerase lambda
PDE0131	2BDP	Transferase/DNA	2 layer sandwich/orthogonal bundle	Ribonuclease H-like/DNA/RNA polymerases	DNA polymerase family A

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PDE0133	4BDP	Transferase/DNA	2 layer sandwich/orthogonal bundle/up-down bundle	Ribonuclease H-like/DNA/RNA polymerases	DNA polymerase family A
PD0878	2IIE	Recombination DNA	none	none	Bacterial DNA-binding protein
PD1082	3C29	Recombination DNA	orthogonal bundle	none	Phage integrase family
PD1182	3ECP	Recombination DNA	alpha beta complex/orthogonal bundle	Ribonuclease H-like	Transposase DDE domain /Transposase Tn5 dimerisation domain
PD0849	2I06	Replication DNA	2 layer sandwich/3 layer sandwich	Replication terminator protein (Tus)	DNA replication terminus site-binding protein
2VJU	2VJU	DNA binding protein	none	Transposase IS200-like	Transposase IS200 like
PD1060	3BKZ	DNA binding protein	none	Transposase IS200-like	Transposase IS200 like
PD0371	1J1V	Replication/dna	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Domain
PD0494	1RH6	Protein/dna	orthogonal bundle	Putative DNA-binding domain	Excisionase-like protein
PD0024	1B3T	Protein/dna	2 layer sandwich	Viral DNA-binding domain	Epstein Barr virus nuclear antigen-1, DNA-binding domain
PD0227	1JJ4	Transcription/dna	2 layer sandwich	Viral DNA-binding domain	E2 (early) protein, C terminal
PD1055	3BIE	Oxidoreductase/dna	sandwich	none	none
PD1059	3BI3	Oxidoreductase/dna	sandwich	none	none
PD1290	3H8X	Oxidoreductase/dna	none	none	2OG-Fe(II) oxygenase superfamily
PDV001	2BOP	Transcription/dna	2 layer sandwich	Viral DNA-binding domain	E2 (early) protein, C terminal

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0219	1JE8	Transcription/dna	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Bacterial regulatory proteins, luxR family
PD0008	1A6Y	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PD0071	1BY4	Gene Regulation/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PD0115	1DSZ	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PD0471	1R0O	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PDR021	2NLL	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PDRC03	1HCQ	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)/Oestrogen receptor
PDT030	1LAT	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PD0475	1R4O	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PD0333	1IXY	Transferase/dna	3 layer sandwich	UDP-Glycosyltransferase/glycogen phosphorylase	Bacteriophage T4 beta-glucosyltransferase
PD0534	1SXQ	Transferase/dna	3 layer sandwich	UDP-Glycosyltransferase/glycogen phosphorylase	Bacteriophage T4 beta-glucosyltransferase

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0386	1NKP	Transcription/dna	Irregular	HLH, helix-loop-helix DNA-binding domain	Helix-loop-helix DNA-binding domain /Myc leucine zipper domain
PD0387	1NJW	Transferase/dna	3 layer sandwich/orthogonal bundle/up-down bundle	Ribonuclease H-like/DNA/RNA polymerases	DNA polymerase family A
PD0584	1U8B	Metal Binding Protein/dna	3 layer sandwich/orthogonal bundle	none	Metal binding domain of Ada /Bacterial regulatory helix-turn-helix proteins, AraC family
PDT062	1AM9	Helix-loop-helix domains	2 layer sandwich/up-down bundle	DNA/RNA polymerases/Ribonuclease H-like	DNA polymerase family A
PD0287	1KX5	Structural Protein/dna	Orthogonal bundle	Histone-fold	Core histone H2A/H2B/H3/H4
PD0051	1CKT	Gene Regulation/dna	Orthogonal bundle	HMG-box	HMG (high mobility group) box
PD0110	1QRV	Gene Regulation/dna	Orthogonal bundle	HMG-box	HMG (high mobility group) box
NA0441	3LNQ	Gene regulation/dna	none	none	Homeobox domain
PD0406	1OWF	Transcription/dna	Irregular	IHF-like DNA-binding proteins	Bacterial DNA-binding protein
PD0430	1P71	DNA Binding Protein/dna	Irregular	IHF-like DNA-binding proteins	Bacterial DNA-binding protein
PD0431	1R0O	Transcription/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PDT040	1IHF	Transcription/dna	Irregular	IHF-like DNA-binding proteins	Bacterial DNA-binding protein
PD0311	1H6F	Transcription Factor	sandwich	p53-like transcription factors	T-box
PD0341	1MNN	Transcription/dna	sandwich	p53-like transcription factors	NDT80 / PhoG like DNA-binding family

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0728	2EUV	Cell Cycle/dna	sandwich	p53-like transcription factors	NDT80 / PhoG like DNA-binding family
PD1067	3BRD	DNA Binding Protein/dna	sandwich/trefoil	p53-like transcription factors/E set domains/DNA-binding protein LAG-1	LAG1, DNA binding /Beta-trefoil
PD1068	3BRF	DNA Binding Protein/dna	sandwich/trefoil	p53-like transcription factors/E set domains/DNA-binding protein LAG-1	LAG1, DNA binding /Beta-trefoil/IPT/TIG domain
PD1069	3BRG	DNA Binding Protein/dna	sandwich/trefoil	none	LAG1, DNA binding /Beta-trefoil
PDR027	1TUP	Antitumor Protein/dna	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type (two domains)
PDR032	1A3Q	Transcription/dna	sandwich	E set domains/p53-like transcription factors	Rel homology domain (RHD)
PDR051	2RAM	Transcription/dna	sandwich	E set domains/p53-like transcription factors	Rel homology domain (RHD)
PDT015	1NFK	Transcription/dna	sandwich	E set domains/p53-like transcription factors	Rel homology domain (RHD)
PDT045	1XBR	Transcription/dna	sandwich	p53-like transcription factors	T-box
PD0480	1R71	Transcription/dna	none	DNA/RNA-binding 3-helical bundle	ParB-like nuclease domain/KorB domain
PD0056	1QPZ	Transcription/dna	3 layer sandwich/orthogonal bundle	lambda repressor-like DNA-binding domains/Periplasmic binding protein-like	Bacterial regulatory proteins, lacI family/Periplasmic binding proteins and sugar binding domain of LacI family

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0224	1JFT	Transcription/dna	3 layer sandwich/orthogonal bundle	lambda repressor-like DNA-binding domains/Periplasmic binding protein-like	Bacterial regulatory proteins, lacI family/Periplasmic binding proteins and sugar binding domain of LacI family
PDR001	3CRO	Transcription/dna	Orthogonal bundle	lambda repressor-like DNA-binding domains	Helix-turn-helix
PDR004	2OR1	Gene Regulation/dna	Orthogonal bundle	lambda repressor-like DNA-binding domains	Helix-turn-helix
PDR010	1LMB	Transcription/dna	Orthogonal bundle	lambda repressor-like DNA-binding domains	Helix-turn-helix
PDR011	1RPE	Gene Regulation/dna	Orthogonal bundle	lambda repressor-like DNA-binding domains	Helix-turn-helix
PDR015	1PER	Transcription/dna	Orthogonal bundle	lambda repressor-like DNA-binding domains	Helix-turn-helix
2C9L	2C9L	Viral Protein	none	Leucine zipper domain	bZIP transcription factor
PD0180	1GD2	Transcription/dna	up-down bundle	Leucine zipper domain	bZIP transcription factor
PD0241	1JMN	Transcription/dna	2 layer sanwich	Leucine zipper domain	bZIP transcription factor
PD0290	1H8A	Transcription/dna	Orthogonal bundle/up-down bundle	Leucine zipper domain/homeodomain-like	Basic region leucine zipper/Myb-like DNA-binding domain
PD0818	2H7H	Viral Protein/dna	up-down bundle	Leucine zipper domain	bZIP transcription factor
PDT029	2DGC	Transcription/dna	2 layer sandwich	Leucine zipper domain	bZIP transcription factor
2VE9	2VE9	Transport Protein	none	DNA/RNA-binding 3-helical bundle	Ftsk gamma domain
PD0007	9ANT	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0016	3HDD	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain/Engrailed homeobox C-terminal signature domain
PD0020	1BC8	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain
PD0027	1BC7	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain
PD0042	1B8I	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PD0050	6PAX	Gene Regulation/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Paired box' domain
PD0073	3HTS	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	HSF-type DNA-binding
PD0075	1B72	Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PD0076	2IRF	Gene Regulation/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Interferon regulatory factor transcription factor
PD0111	1DP7	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	RFX DNA-binding domain
PD0116	1DUX	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain
PD0167	1F4K	Replication/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Replication terminator protein
PD0183	1FYL	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	HSF-type DNA-binding
PD0184	1FYM	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	HSF-type DNA-binding
PD0211	1IG7	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0225	1E3O	Transcription Factor	Orthogonal bundle	DNA/RNA-binding 3-helical bundle/lambda repressor-like DNA-binding domains	Homeobox domain/Pou domain - N-terminal to homeobox domain
PD0257	1K61	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PD0259	1K78	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain/Paired box' domain
PD0260	1K79	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain
PD0299	1J59	Gene Regulation/dna	Orthogonal bundle/sandwich	DNA/RNA-binding 3-helical bundle/Double-stranded beta-helix	Cyclic nucleotide-binding domain/Bacterial regulatory proteins, crp family
PD0455	1PUF	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PD0673	1ZRF	Gene Regulation/dna	sandwich/orthogonal bundle	DNA/RNA-binding 3-helical bundle	Cyclic nucleotide-binding domain/Bacterial regulatory proteins, crp family
PD0813	2H27	Tranferase/DNA	none	DNA/RNA-binding 3-helical bundle	Sigma-70, region 4
PD0975	2EFW	Replication/dna	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Replication terminator protein
PDE025	1FJL	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PDR018	1PDN	Gene Regulation/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Paired box' domain
PDR034	1AU7	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle/lambda repressor-like DNA-binding domains	Homeobox domain/Pou domain - N-terminal to homeobox domain
PDR049	1AKH	DNA Binding Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PDR056	1BL0	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Bacterial regulatory helix-turn-helix proteins, AraC family
PDT013		Transcription/dna			
PDT028	1YRN	DNA Binding Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PDT031	1JGG	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PDT033	1PUE	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain
PDT035	1IGN	DNA Binding Protein/dna	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Myb-like DNA-binding domain/Rap1, DNA-binding
PDT043	2HDD	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Homeobox domain
PDT048	1AWC	Transcription/DNA	orthogonal bundle/alpha horseshoe	DNA/RNA-binding 3-helical bundle/beta-hairpin-alpha-hairpin repeat	Ets-domain/Ankyrin repeat
PD0483	1R8E	Transcription/DNA	orthogonal bundle/alpha-beta barrel/up-down bundle	Putative DNA-binding domain	MerR family regulatory protein /Bacterial transcription activator, effector binding domain
PD0961	2OG0	DNA Binding Protein/dna	orthogonal bundle	Putative DNA-binding domain	Excisionase-like protein
PD0003	1CRX	Replication/dna	orthogonal bundle	lambda integrase-like, N-terminal domain/DNA breaking-rejoining enzymes	Phage integrase family
NA0506	3MGV	Isomerase/dna	none	none	Phage integrase family

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0047	4CRX	Protein/dna	orthogonal bundle	lambda integrase-like, N-terminal domain/DNA breaking-rejoining enzymes	Phage integrase family
PD0103	2CRX	Hydrolase , Ligase/dna	orthogonal bundle	DNA breaking-rejoining enzymes/lambda integrase-like, N-terminal domain	Phage integrase family
PD0166	1F44	Hydrolase , Ligase/dna	orthogonal bundle	DNA breaking-rejoining enzymes/lambda integrase-like, N-terminal domain	Phage integrase family
PD0213	1IJW	DNA Binding Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Helix-turn-helix domain of resolvase
PD0234	1JKO	DNA Binding Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Helix-turn-helix domain of resolvase
PD0457	1PVP	Recombination/dna	orthogonal bundle	DNA breaking-rejoining enzymes/lambda integrase-like, N-terminal domain	Phage integrase family
PD0601	1XO0	Hydrolase , Ligase/dna	orthogonal bundle	DNA breaking-rejoining enzymes/lambda integrase-like, N-terminal domain	Phage integrase family
PDE009	1HCR	DNA Binding Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Helix-turn-helix domain of resolvase
PD0126	1QAI	Transferase/dna	2 layer sandwich/roll	DNA/RNA polymerases	RNase H/Integrase core domain
PD0359	1N4L	Transferase/dna	2 layer sandwich/roll	DNA/RNA polymerases	RNase H/Integrase core domain
2BNW	2BNW	DNA Binding/regulatory Protein	orthogonal bundle	Ribbon-helix-helix	Omega Transcriptional Repressor

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0035	1BDT	Gene Regulation/dna	Othogonal bundle	Ribbon-helix-helix	Arc-like DNA binding domain
PD0173	1MJ2	Transcription/dna	Othogonal bundle	Ribbon-helix-helix	Met Apo-repressor, MetJ
PD0245	1MJM	Transcription/dna	Othogonal bundle	Ribbon-helix-helix	Met Apo-repressor, MetJ
PD0246	1MJO	Transcription/dna	Othogonal bundle	Ribbon-helix-helix	Met Apo-repressor, MetJ
PD0248	1MJQ	Transcription/dna	Othogonal bundle	Ribbon-helix-helix	Met Apo-repressor, MetJ
PD1052	2RBF	Oxidoreductase/dna	none	none	none
PD0086	1CEZ	Transferase/dna	2 layer sandwich/orthogonal bundle	DNA/RNA polymerases	DNA-dependent RNA polymerase
PD0390	1J3E	Replication/dna	up-down bundle	Replication modulator SeqA, C-terminal DNA-binding domain	SeqA protein
PD0492	1RIO	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle/lambda repressor-like DNA-binding domains	Helix-turn-helix/Sigma-70, region 4
PD0284	1KU7	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle bundle	Sigma-70, region 4
PDT064	1SKN	Transcription/dna	Orthogonal bundle	A DNA-binding domain in eukaryotic transcription factors	none
PD0409	1OZJ	Transcription/dna	Alpha beta complex	SMAD MH1 domain	MH1 domain
PD0314	1LQ1	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle bundle	Sporulation initiation factor Spo0A C terminal
PD0220	1JEY	DNA Binding Protein/dna	3 layer sandwich/orthogonal bundle/beta barrel	SPOC domain-like/vWA-like	Ku70/Ku80 N-terminal alpha/beta domain/SAP domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0121	1EGW	Transcription/dna	none	SRF-like	SRF-type transcription factor (DNA-binding and dimerisation domain)
PD0363	1N6J	Transcription/dna	none	SRF-like	SRF-type transcription factor (DNA-binding and dimerisation domain)
PD0028	1BG1	Transcription/dna	2 layer sandwich/orthogonal bundle/sanwich/up down bundle	STAT/p53-like transcription factors/SH2 domain	STAT protein, all-alpha domain/SH2 domain
PD0676	2CV5	Structural Protein/DNA	orthogonal bundle	Histone-fold	Core histone H2A/H2B/H3/H4
PD0914	2NQB	Structural Protein/DNA	orthogonal bundle	Histone-fold	Core histone H2A/H2B/H3/H5
PD1003	2PYO	Structural Protein/DNA	orthogonal bundle	Histone-fold	Core histone H2A/H2B/H3/H6
PD1077	3C1B	Structural Protein/DNA	orthogonal bundle	Histone-fold	Core histone H2A/H2B/H3/H7
PD0154	1QN4	Tata Box Binding Protein (tbp	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0155	1QN5	Tata Box Binding Protein (tbp	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0156	1QN6	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0157	1QN7	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0158	1QN8	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0159	1QN9	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0160	1QNA	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0161	1QNB	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0162	1QNC	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0163	1QNE	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PD0164	1QN3	Tata Box Binding Protein (tbp)	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PDT012	1YTB	Transcription/dna	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)
PDT034	1CDW	Transcription/dna	2 layer sandwich	TATA-box binding protein-like	Transcription factor TFIID (or TATA-binding protein, TBP)

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0122	1QPI	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle/Tetracyclin repressor-like, C-terminal domain	Bacterial regulatory proteins, tetR family/Tetracyclin repressor, C-terminal all-alpha domain
PD0369	1NH2	Transcription/dna	2 layer sandwich/orthogonal bundle/roll	TATA-box binding protein-like/Transcription factor IIA (TFIIA), alpha-helical domain	Transcription factor TFIID (or TATA-binding protein, TBP)/Transcription factor IIA, alpha/beta subunit
PD0393	1NVP	Transcription/dna	2 layer sandwich/orthogonal bundle/roll	TATA-box binding protein-like/Transcription factor IIA (TFIIA), alpha-helical domain	Transcription factor TFIID (or TATA-binding protein, TBP)/Transcription factor IIA, alpha/beta subunit
PDT036	1YTF	Transcription/dna	2 layer sandwich/orthogonal bundle/roll	TATA-box binding protein-like/Transcription factor IIA (TFIIA), alpha-helical domain	Transcription factor TFIID (or TATA-binding protein, TBP)/Transcription factor IIA, alpha/beta subunit
PDR031	1AIS	Transcription/dna	2 layer sandwich/orthogonal bundle	TATA-box binding protein-like/Cyclin-like	Transcription factor TFIID (or TATA-binding protein, TBP)/Transcription factor TFIIB repeat
PD0088	1BF4	DNA Binding Protein/dna	Beta Barrel	Chromo domain-like	7kD DNA-binding domain
PD0544	1WD0	Structural Protein/dna	Beta Barrel	Chromo domain-like	7kD DNA-binding domain
PD0545	1WD1	Structural Protein/dna	Beta Barrel	Chromo domain-like	7kD DNA-binding domain
PD0607	1XYI	Structural Protein/dna	Beta Barrel	Chromo domain-like	7kD DNA-binding domain
PD0609	1WTO	Structural Protein/dna	Beta Barrel	none	7kD DNA-binding domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0613	1WTV	DNA Binding Protein/dna	Beta Barrel	none	7kD DNA-binding domain
PDR047	1AZP	DNA Binding Protein/dna	Beta Barrel	Chromo domain-like	7kD DNA-binding domain
PDR048	1AZQ	DNA Binding Protein/dna	Beta Barrel	Chromo domain-like	7kD DNA-binding domain
PDR036	1MNM	Transcription/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle/SRF-like	Homeobox domain/SRF-type transcription factor (DNA-binding and dimerisation domain)
PD0289	1H89	Transcription/dna	Orthogonal bundle/up-down bundle	DNA/RNA-binding 3-helical bundle	Basic region leucine zipper/Myb-like DNA-binding domain
NA0104	3IGC	Isomerase/dna	none	none	Viral DNA topoisomerase I, N-terminal/Eukaryotic DNA topoisomerase I, catalytic core
NA0416	3L4J	Isomerase/dna	2 layer sandwich/orthogonal bundle/Alpha-Beta Complex	none	DNA gyrase/topoisomerase IV, subunit A
NA0514	3M4A	Isomerase	none	none	Eukaryotic DNA topoisomerase I, catalytic core
NA0645	2XCS	Isomerase/dna	none	none	Toprim domain /DNA gyrase B subunit, carboxyl terminus/DNA gyrase/topoisomerase IV, subunit A

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0256	1K4T	Isomerase/dna	Orthogonal bundle/alpha-beta complex/alpha complex	Long alpha-hairpin/DNA breaking-rejoining enzymes/Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment	Eukaryotic DNA topoisomerase I, DNA binding fragment
PDE0142	1A31	Isomerase/dna	Orthogonal bundle/alpha-beta complex/alpha complex	Long alpha-hairpin/DNA breaking-rejoining enzymes/Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment	Eukaryotic DNA topoisomerase I, DNA binding fragment
2C6Y	2C6Y	Transcription regulation	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Fork head domain
2VY1	2VY1	Transcription	none	none	Floricaula / Leafy protein
2VY2	2VY2	Transcription	none	none	Floricaula / Leafy protein
2W7N	2W7N	Transcription/DNA	none	none	none
2X6V	2X6V	Transcription/DNA	sandwich	none	T-box
NA0092	3IAG	Transcription/DNA	sandwich/trefoil	none	LAG1, DNA binding/Beta-trefoil/IPT/TIG domain
NA0118	3IKT	DNA Binding Protein/dna	3 layer sandwich/orthogonal bundle	none	Putative DNA-binding protein N-terminus/CoA binding domain
NA0167	3JTG	Transcription	orthogonal bundle	none	Ets-domain
NA0181	3JXB	Transcription regulation	orthogonal bundle	none	Helix-turn-helix
NA0394	3L2C	Transcription/DNA	none	none	Fork head domain
NA0395	3KZ8	Transcription/DNA	none	none	P53 DNA-binding domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
NA0485	3M9E	Transcription/DNA	none	none	Zinc finger, C4 type
NA0486	2WBS	Transcription/DNA	none	none	Zinc finger, C2H2 type
NA0487	2WBU	Transcription/DNA	none	none	Zinc finger, C2H2 type
NA0548	3MVA	Transcription/DNA	none	none	mTERF
NA0570	3N7Q	Transcription , Replication/dna	none	none	mTERF
PD0559	1RM1	Transcription/DNA	2 layer sandwich/orthogonal bundle/roll	TATA-box binding protein-like/Transcription factor IIA (TFIIA), beta-barrel domain	Transcription factor TFIID (or TATA-binding protein, TBP/Transcription initiation factor IIA, gamma subunit, helical domain/Transcription factor IIA, alpha/beta subunit
PD0627	1YO5	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Ets-domain
PD0645	1ZG1	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Bacterial regulatory proteins, luxR family
PD0646	1ZG5	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Bacterial regulatory proteins, luxR family
PD0687	1ZS4	Transcription/DNA	none	lambda repressor-like DNA-binding domains	Bacteriophage CII protein
PD0689	2A07	Transcription/DNA	none	DNA/RNA-binding 3-helical bundle	Fork head domain
PD0691	2A66	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD0739	2D5V	Transcription/DNA	none	none	CUT domain/Homeobox domain
PD0791	2GEQ	Transcription/DNA	sandwich	none	P53 DNA-binding domain
PD0810	2H1K	Transcription/DNA	none	none	Homeobox domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0822	2HAN	Transcription/DNA	2 layer sandwich	Glucocorticoid receptor-like (DNA-binding domain)	Zinc finger, C4 type
PD0843	2HT0	Transcription/DNA	Irregular	IHF-like DNA-binding proteins	Bacterial DNA-binding protein
PD0884	2ISZ	Transcription/DNA	orthogonal bundle	DNA/RNA-binding 3-helical bundle/Iron-dependent repressor protein, dimerization domain	Iron dependent repressor, N-terminal DNA binding domain
PD0889	2E1C	Transcription/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA-binding 3-helical bundle /Ferredoxin-like	Bacterial regulatory protein, arsR family/AsnC family
PD0946	2O49	Transcription/DNA	none	lambda repressor-like DNA-binding domains	CUT domeain
PD0947	2O4A	Transcription/DNA	none	lambda repressor-like DNA-binding domains	CUT domeain
PD0993	2YVH	Transcription/DNA	orthogonal bundle	none	Bacterial regulatory proteins, tetR family
PD1009	2PI0	Transcription Activator/dna	orthogonal bundle	DNA/RNA-binding 3-helical bundle	Interferon regulatory factor transcription factor
PD1037	2R1J	Transcription/DNA	orthogonal bundle	lambda repressor-like DNA-binding domains	Helix-turn-helix
PD1057	3BS1	Transcription Regulator	none	none	LytTr DNA-binding domain
PD1080	3C2I	Transcription Regulator	2 layer sandwich	none	Methyl-CpG binding domain
PD1094	3CBB	Transcription Regulator/dna	2 layer sandwich	none	Zinc finger, C4 type
PD1100	3CMY	Transcription/DNA	none	none	Homeobox domain
PD1101	3COQ	Transcription/DNA	none	Zn2/Cys6 DNA-binding domain/Leucine zipper domain	Fungal Zn(2)-Cys(6) binuclear cluster domain/Gal4-like dimerisation domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD1104	3COA	Transcription/DNA	none	none	Fork head domain
PD1152	2QL2	Transcription/DNA	Irregular	none	Helix-loop-helix DNA-binding domain/Neuronal helix-loop-helix transcription factor
PD1180	3E6C	Transcription Regulator/dna	none	DNA/RNA-binding 3-helical bundle/Double-stranded beta-helix	Cyclic nucleotide-binding domain
PD1191	3EXJ	Transcription/DNA	none	none	P53 DNA-binding domain
PD1225	3FYL	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1230	3G73	Transcription/DNA	none	none	Fork head domain
PD1231	3G6P	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1232	3G6Q	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1233	3G6R	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1234	3G6T	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1235	3G6U	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1238	3G97	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1239	3G99	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1241	3G9I	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1242	3G9J	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1243	3G9M	Transcription/DNA	2 layer sandwich	none	Zinc finger, C4 type
PD1278	3H0D	Transcription/DNA	none	none	Firmicute transcriptional repressor of class III stress genes (CtsR)
PD0350	1MUS	Transcription/DNA	Orthogonal bundle/alpha-beta complex	Ribonuclease H-like	Transposase DDE domain/Transposase Tn5 dimerisation domain

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0394	1OMH	Tranferase/DNA	none	Origin of replication-binding domain, RBD- like	TrwC relaxase
PDE0128	1TC3	DNA Binding Protein/dna	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Tc3 transposase
PDE141	6MHT	Tranferase/DNA	3 layer sanwich/alpha beta complex	S-adenosyl-L-methionine-dependent methyltransferases	C-5 cytosine-specific DNA methylase
PD0298	1L3L	Transcription/DNA	2 layer sandwich/orthogonal bundle	DNA/RNA-binding 3-helical bundle/Profilin-like	Autoinducer binding domain/Bacterial regulatory proteins, luxR family
PDR009	1TRO	Transcription/DNA	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Trp repressor protein
PDR013	1TRR	Transcription/DNA	Orthogonal bundle	DNA/RNA-binding 3-helical bundle	Trp repressor protein
NA0063	3HXO	Blood Clotting/blood Clotting Regulator	3 layer sandwich	none	von Willebrand factor type A domain
PDT055	1A1F	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type
PDT057	1A1H	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type
PDT058	1A1K	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type
PDTB41	1MEY	Transcription/DNA	2 layer sandwich	Zinc finger design	none
2C7A	2C7A	Receptor/dna	2 layer sandwich	none	Progesterone receptor /Zinc finger, C4 type
PD0187	1G2F	Transcription/DNA	2 layer sandwich	Zinc finger design	Zinc finger, C2H2 type
PD0231	1JK2	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type
PD0424	1P47	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type

Appendix 1B cont.
Classification of Non-Redundant Structure in NAPID

NDB ID	PDB ID	PDB Classification	CATH	SCOP	PFAM
PD0851	2I13	DNA Binding Protein/dna	none	Zinc finger design	Zinc finger, C2H2 type
PDT038	1UBD	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type
PDT039	1AAY	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	Zinc finger, C2H2 type
PD0089	1HWT	Gene Regulation/dna	none	Zn2/Cys6 DNA-binding domain/Parallel coiled-coil	none
PD0090	1HAP	Hydrolase/hydrolase Inhibitor/dna	Beta Barrel	Trypsin-like serine proteases	Trypsin/Thrombin light chain
PD0312	1LLM	Transcription/DNA	2 layer sandwich	beta-beta-alpha zinc fingers	bZIP transcription factor

Appendix 2
SQL Query Used to Generate Hybridization Data

```
USE NAPID;

SELECT DISTINCT resName, name, groupType

FROM proteinAtoms

ORDER BY resName, name;
```

Appendix 3
SQL Query Used to Create AAHybridization Table

```
CREATE TABLE 'AAHybridization'
('AAHybridID' varchar(10) NOT NULL,
'resName' varchar(5) NOT NULL
'name' varchar(5) NOT NULL
'grouptype' varchar(20) NOT NULL,
'Hybridization' varchar(3) NOT NULL,
PRIMARY KEY('AAHybridID'));
```

Appendix 4
Hybridization Data Stored in AAHybridization Table

Amino acid	Atom Name	Group Type	Hybridization	Amino acid	Atom Name	Group Type	Hybridization
ALA (A)	C	Backbone	sp2	CYS (C)	OXT	Backbone	sp2
ALA (A)	CA	Backbone	sp3	CYS (C)	SG	Side chain	sp3
ALA (A)	CB	Side chain	sp3	GLN (Q)	C	Backbone	sp2
ALA (A)	N	Backbone	sp2	GLN (Q)	CA	Backbone	sp3
ALA (A)	O	Backbone	sp2	GLN (Q)	CB	Side chain	sp3
ALA (A)	OXT	Backbone	sp2	GLN (Q)	CD	Side chain	sp2
ARG (R)	C	Backbone	sp2	GLN (Q)	CG	Side chain	sp3
ARG (R)	CA	Backbone	sp3	GLN (Q)	N	Backbone	sp2
ARG (R)	CB	Side chain	sp3	GLN (Q)	NE2	Side chain	sp3
ARG (R)	CD	Side chain	sp3	GLN (Q)	O	Backbone	sp2
ARG (R)	CG	Side chain	sp3	GLN (Q)	OE1	Side chain	sp2
ARG (R)	CZ	Side chain	sp2	GLN (Q)	OXT	Backbone	sp2
ARG (R)	N	Backbone	sp2	GLU (E)	OE1	Side chain	sp2
ARG (R)	NE	Side chain	sp3	GLU (E)	O	Backbone	sp2
ARG (R)	NH1	Side chain	sp2	GLU (E)	N	Backbone	sp3
ARG (R)	NH2	Side chain	sp2	GLU (E)	CG	Side chain	sp3
ARG (R)	O	Backbone	sp2	GLU (E)	CD	Side chain	sp2
ARG (R)	OXT	Backbone	sp2	GLU (E)	CB	Side chain	sp3
ASN (N)	C	Backbone	sp2	GLU (E)	CA	Backbone	sp3
ASN (N)	CA	Backbone	sp3	GLU (E)	C	Backbone	sp2
ASN (N)	CB	Side chain	sp3	GLU (E)	OE2	Side chain	sp2
ASN (N)	CG	Side chain	sp2	GLU (E)	OXT	Backbone	sp2
ASN (N)	N	Backbone	sp2	GLY (G)	C	Backbone	sp2
ASN (N)	ND2	Side chain	sp3	GLY (G)	CA	Backbone	sp3
ASN (N)	O	Backbone	sp2	GLY (G)	N	Backbone	sp2
ASN (N)	OD1	Side chain	sp2	GLY (G)	O	Backbone	sp2
ASN (N)	OXT	Backbone	sp2	GLY (G)	OXT	Backbone	sp2
ASP (D)	C	Backbone	sp2	HIS (H)	C	Backbone	sp2
ASP (D)	CA	Backbone	sp3	HIS (H)	CA	Backbone	sp3
ASP (D)	CB	Side chain	sp3	HIS (H)	CB	Side chain	sp3
ASP (D)	CG	Side chain	sp2	HIS (H)	CD2	Side chain	sp2
ASP (D)	N	Backbone	sp2	HIS (H)	CE1	Side chain	sp2
ASP (D)	O	Backbone	sp2	HIS (H)	CG	Side chain	sp2
ASP (D)	OD1	Side chain	sp2	HIS (H)	N	Backbone	sp3
ASP (D)	OD2	Side chain	sp2	HIS (H)	ND1	Side chain	sp2
ASP (D)	OXT	Backbone	sp2	HIS (H)	NE2	Side chain	sp2
CYS (C)	C	Backbone	sp2	HIS (H)	O	Backbone	sp2
CYS (C)	CA	Backbone	sp3	HIS (H)	OXT	Backbone	sp2
CYS (C)	CB	Side chain	sp3				
CYS (C)	N	Backbone	sp2				
CYS (C)	O	Backbone	sp2				

Appendix 4 cont.
Hybridization Data Stored in AAHybridization Table

Amino acid	Atom Name	Group Type	Hybridization	Amino acid	Atom Name	Group Type	Hybridization
ILE (I)	C	Backbone	sp2	PHE (F)	C	Backbone	sp2
ILE (I)	CA	Backbone	sp3	PHE (F)	CA	Backbone	sp3
ILE (I)	CB	Side chain	sp3	PHE (F)	CB	Side chain	sp3
ILE (I)	CD1	Side chain	sp3	PHE (F)	CD1	Side chain	sp2
ILE (I)	CG1	Side chain	sp3	PHE (F)	CD2	Side chain	sp2
ILE (I)	CG2	Side chain	sp3	PHE (F)	CE1	Side chain	sp2
ILE (I)	N	Backbone	sp2	PHE (F)	CE2	Side chain	sp2
ILE (I)	O	Backbone	sp2	PHE (F)	CG1	Side chain	sp2
ILE (I)	OXT	Backbone	sp2	PHE (F)	CZ	Side chain	sp2
LEU (L)	C	Backbone	sp2	PHE (F)	N	Backbone	sp2
LEU (L)	CA	Backbone	sp3	PHE (F)	O	Backbone	sp2
LEU (L)	CB	Side chain	sp3	PHE (F)	OXT	Backbone	sp2
LEU (L)	CD1	Side chain	sp3	PRO (P)	C	Backbone	sp2
LEU (L)	CD2	Side chain	sp3	PRO (P)	CA	Backbone	sp3
LEU (L)	CG	Side chain	sp3	PRO (P)	CB	Side chain	sp3
LEU (L)	N	Backbone	sp2	PRO (P)	CD	Side chain	sp3
LEU (L)	O	Backbone	sp2	PRO (P)	CG	Side chain	sp3
LEU (L)	OXT	Backbone	sp2	PRO (P)	N	Backbone	sp2
LYS (K)	C	Backbone	sp2	PRO (P)	O	Backbone	sp2
LYS (K)	CA	Backbone	sp3	PRO (P)	OXT	Backbone	sp2
LYS (K)	CB	Side chain	sp3	SER (S)	C	Backbone	sp2
LYS (K)	CD	Side chain	sp3	SER (S)	CA	Backbone	sp3
LYS (K)	CE	Side chain	sp3	SER (S)	CB	Side chain	sp3
LYS (K)	CG	Side chain	sp3	SER (S)	O	Backbone	sp2
LYS (K)	N	Backbone	sp2	SER (S)	N	Backbone	sp2
LYS	NZ	Side chain	sp3	SER (S)	OG	Side	sp3

(K)						chain	
LYS (K)	O	Backbone	sp2	SER (S)	OXT	Backbone	sp2
LYS (K)	OXT	Backbone	sp2	THR (T)	C	Backbone	sp3
MET (M)	C	Backbone	sp2	THR (T)	CA	Backbone	sp3
MET (M)	CA	Backbone	sp3	THR (T)	CB	Side chain	sp3
MET (M)	CB	Side chain	sp3	THR (T)	CG2	Side chain	sp3
MET (M)	CE	Side chain	sp3	THR (T)	N	Backbone	sp2
MET (M)	CG	Side chain	sp3	THR (T)	O	Backbone	sp2
MET (M)	N	Backbone	sp2	THR (T)	OG1	Side chain	sp3
MET (M)	O	Backbone	sp2	THR (T)	OXT	Backbone	sp2
MET (M)	OXT	Backbone	sp2				
MET (M)	SD	Side chain	sp3				

Appendix 4 cont.
Hybridization Data Stored in AAHybridization Table

Amino acid	Atom Name	Group Type	Hybridization
TRP (W)	C	Backbone	sp2
TRP (W)	CA	Backbone	sp3
TRP (W)	CB	Side chain	sp3
TRP (W)	CD1	Side chain	sp2
TRP (W)	CD2	Side chain	sp2
TRP (W)	CE2	Side chain	sp2
TRP (W)	CE3	Side chain	sp2
TRP (W)	CG	Side chain	sp2
TRP (W)	CH2	Side chain	sp2
TRP (W)	CZ2	Side chain	sp2
TRP (W)	CZ3	Side chain	sp2
TRP (W)	N	Backbone	sp2
TRP (W)	NE1	Side chain	sp2
TRP (W)	O	Backbone	sp2
TRP (W)	OXT	Backbone	sp2
TYR (Y)	C	Backbone	sp2
TYR (Y)	CA	Backbone	sp3
TYR (Y)	CB	Side chain	sp3
TYR (Y)	CD1	Side chain	sp2
TYR (Y)	CD2	Side chain	sp2
TYR (Y)	CE1	Side chain	sp2
TYR (Y)	CE2	Side chain	sp2
TYR (Y)	CG	Side chain	sp2
TYR (Y)	CZ	Side chain	sp2
TYR (Y)	N	Backbone	sp2
TYR (Y)	O	Backbone	sp2
TYR (Y)	OXT	Backbone	sp2
VAL (V)	C	Backbone	sp2
VAL (V)	CA	Backbone	sp3
VAL (V)	CB	Side chain	sp3
VAL (V)	CG1	Side chain	sp3
VAL (V)	CG2	Side chain	sp3
VAL (V)	N	Backbone	sp3
VAL (V)	O	Backbone	sp2
VAL (V)	OXT	Backbone	sp3

Appendix 5A
SQL for Loading Data into AAHybridization Table in NAPID

```
Mysql> load data local infile 'AAHybridization.tab' into table AAhybridization;
```

Appendix 5B
Sample SQL for Extracting Distance Between Base and Protein Atoms in NAPID

```
SELECT t2.distance, t3.resName, t2name, t4.resName, t4.name, t5hybridization
FROM  structures t1, naProteinContacts t2, naAtoms t3, proteinAtoms t4,
        AAHybridization t5
WHERE  ndbid = (refer to Appendix 1A for list of reference structures)
and t1.pdbid = t2.pdbid
and  t2.naAtomID = t3.id and t2.proteinAtomID = t4.id
and t4resName = t5resName and t4.name = t5.name
and t2.distance ≤ 3.4 and t2.location = 'major groove'
and t3.grouptype = 'Base' and t3.resName = 'G'
and t5.Hybridization = 'sp2';
```

The SQL clause t2.naAtomID = t3.id establishes the linkages between naProteinContacts table and naAtoms table. The clause t2.proteinAtomID = t4.id establishes the linkage between naProteinContacts table and proteinAtoms table. The clause t4.resname = t5.resName and t4.name = t5.name establish the linkage between protinAtoms and AAHybridization table. The clauses t2.distance ≤ 3.4, t2.location = 'major groove, t3.grouptype = 'Base', t3.resName = 'G' and t5.Hybridization = 'sp^2' create the specifications for finding contacts ≤ 3.4Å in the major groove with guanine and amino-acid sp^2 hybridizations. Note, the statement "ndbid = (Refer to Appendix 1A…) limits the data to the 499 non-redundant structures discussed in Section 2.1.3.

Appendix 6
Source Data for Figure 3.1

Base	Base Type	W1 Description	W2 Discription	Distance	Count sp^2	Count sp^3
Adenine	Purine	N - More Polar	NH2 - More Hindered	≤3.0	157	152
Adenine	Purine	N - More Polar	NH2 - More Hindered	>3.0 and ≤3.2	147	149
Adenine	Purine	N - More Polar	NH2 - More Hindered	>3.2 and ≤3.4	247	229
Thymine	Pyrimidine	C-Me - Nonpolar	NH2 - More Hindered	≤3.0	157	167
Thymine	Pyrimidine	C-Me - Nonpolar	NH2 - More Hindered	>3.0 and ≤3.2	100	136
Thymine	Pyrimidine	C-Me - Nonpolar	NH2 - More Hindered	>3.2 and ≤3.4	324	355
Guanine	Purine	N - More Polar	C=O - Less Hindered	≤3.0	895	524
Guanine	Purine	N - More Polar	C=O - Less Hindered	>3.0 and ≤3.2	414	307
Guanine	Purine	N - More Polar	C=O - Less Hindered	>3.2 and ≤3.4	545	515
Cytosine	Pyrimidine	C-H- Nonpolar	C=O - Less Hindered	≤3.0	283	54
Cytosine	Pyrimidine	C-H- Nonpolar	C=O - Less Hindered	>3.0 and ≤3.2	252	67
Cytosine	Pyrimidine	C-H- Nonpolar	C=O - Less Hindered	>3.2 and ≤3.4	522	174
				Total	4043	2829

Appendix 7
Source Data for Figure 3.2 and Figure 3.3

Atom Type	Adenine sp^2	Adenine sp^3
N7	129	298
N6	326	109
N1	5	7
C8	60	81
C6	10	12
C5	10	14
Total	540	521

Atom Type	Guanine sp^2	Guanine sp^3
O6	783	708
N7	809	466
N1	11	1
C8	152	62
C6	61	43
C5	16	33
Total	1832	1313

Atom Type	Thymine sp^2	Thymine sp^3
O4	192	469
N3	5	5
C6	40	19
C5M	195	144
C5	45	10
C4	15	10
Total	492	1149

Atom Type	Cytosine sp^2	Cytosine sp^3
N4	572	140
N3	4	0
C6	94	35
C5	319	103
C4	52	5
Total	1041	283

Appendix 8
Source Data for Figure 3.4

Base	Hybridization	Backbone Count	Sidechain Count
Adenine	sp^2	69	481
Adenine	sp^3	13	516
Thymine	sp^2	63	434
Thymine	sp^3	29	628
Guanine	sp^2	179	1674
Guanine	sp^3	54	1291
Cytosine	sp^2	305	751
Cytosine	sp^3	11	283
	Total	723	6058

Base Description	Count
AT Backbone	174
CG Backbone	549
G Backbone	233
C Backbone	316
A Backbone	82
T Backbone	92
AT Sidechain	2059
GC Sidechain	3999
G Sidechain	2965
C Sidechain	1034
A sidechain	997
T sidechain	1062

Appendix 9
Source Data for Figures 3.5, 3.6, 3.8 and 3.9

Adenine sp^2 Side Chain	
Amino Acid	**Count**
PHE	4
TYR	4
ASP	16
HIS	29
GLU	36
GLN	87
ASN	152
ARG	153
TOTAL	481

Adenine sp^2 Backbone	
Amino Acid	**Count**
HIS	2
GLN	2
LYS	4
TYR	5
PRO	5
GLY	7
ALA	7
SER	8
ASN	14
THR	15
TOTAL	69

Adenine sp^3 Side Chain	
Amino Acid	**Count**
ALA	3
CYS	4
VAL	4
PHE	5
LEU	6
ILE	7
GLU	9
PRO	11
MET	16
ARG	17
SER	53
THR	62
LYS	79
GLN	108
ASN	132
TOTAL	516

Adenine sp^3 Backbone	
Amino Acid	**Count**
GLN	1
HIS	1
GLY	5
SER	6
TOTAL	13

Appendix 10
Source Data for Figures 3.10 through 3.13

Thymine sp^2 Side Chain	
Amino Acid	**Count**
ASP	1
TYR	12
GLU	19
PHE	27
ASN	28
TRP	36
GLN	38
HIS	72
ARG	201
TOTAL	434

Thymine sp^2 Backbone	
Amino Acid	**Count**
CYS	1
PRO	1
TRP	1
TYR	1
HIS	2
LEU	2
PHE	2
GLU	3
THR	3
ASP	4
GLY	4
ARG	5
GLN	5
SER	5
ASN	6
LYS	8
ALA	10
TOTAL	63

Thymine sp^3 Side Chain	
Amino Acid	**Count**
PHE	2
TRP	2
CYS	5
LEU	5
ILE	6
HIS	9
GLU	12
VAL	13
MET	19
PRO	25
ALA	34
ASN	55
SER	59
ARG	86
GLN	95
LYS	96
THR	105
TOTAL	628

Thymine sp^3 Backbone	
Amino Acid	**Count**
ALA	1
GLN	1
GLU	1
HIS	2
LEU	2
ASN	4
GLY	18
TOTAL	29

Appendix 11
Source Data for Figures 3.14 through 3.17

Guanine sp^2 Side Chain	
Amino Acid	**Counts**
ASP	2
TRP	2
TYR	2
PHE	9
GLU	12
ASN	23
GLN	43
HIS	258
ARG	1323
TOTAL	1674

Guanine sp^2 Backbone	
Amino Acid	**Count**
ILE	1
VAL	1
THR	3
ALA	4
ARG	4
MET	4
PHE	4
TYR	4
LEU	6
GLN	8
LYS	10
HIS	16
ASN	20
SER	41
GLY	53
TOTAL	179

Guanine sp^3 Side Chain	
Amino Acid	**Count**
ASP	1
CYS	1
MET	2
HIS	3
VAL	3
PRO	4
TRP	4
LEU	5
ILE	9
THR	18
GLN	62
ASN	85
SER	119
ARG	289
LYS	686
TOTAL	1291

Guanine sp^3 Backbone	
Amino Acid	**Count**
LEU	2
LYS	4
SER	11
GLY	37
TOTAL	54

Appendix 12
Source Data for Figures 3.18 through 3.23

Cytosine sp^2 Side Chain	
Amino Acid	**Counts**
TYR	7
PHE	8
GLN	18
TRP	23
HIS	45
ASN	58
ASP	159
ARG	203
GLU	230
TOTAL	751

Cytosine sp^2 Backbone	
Amino Acid	**Count**
LEU	1
MET	1
TYR	1
GLU	2
GLN	6
HIS	6
TRP	6
PHE	11
PRO	11
ASP	15
SER	26
GLY	27
LYS	28
ARG	31
THR	31
ALA	39
ASN	63
TOTAL	305

Cytosine sp^3 Side Chain	
Amino Acid	**Count**
ASP	1
LEU	1
GLU	2
TRP	2
TYR	2
VAL	2
HIS	3
ALA	4
MET	5
CYS	20
GLN	20
LYS	21
ARG	47
ASN	47
SER	50
THR	56
TOTAL	283

Cytosine sp^3 Backbone	
Amino Acid	**Count**
GLN	2
GLY	3
ARG	6
TOTAL	11

Specific Cytosine Interactions with sp^2 Hybridized Backbone Atoms

Amino Acid	**naAtomname**	**Count**
ALA	C4	2
	C5	9
	N4	27
ASN	C4	4
	C5	18
	C6	8
	N4	33
ARG	C4	1
	C5	4
	C6	2
	N4	17
GLY	C4	2
	C5	2
	C6	3
	N4	19
LYS	C4	4
	C5	9
	N4	15
SER	C5	12
	C6	3
	N4	11
TRP	C4	2
	C5	2
	N4	2

Amino Acid	**C4**	**C5**	**C6**	**N4**
ALA	2	9	0	27
ASN	4	18	8	33
ARG	1	4	2	17
GLY	2	2	3	19
LYS	4	9	0	15
SER	12	0	3	11
TRP	2	2	0	2

Appendix 13
Source Data for Counts Figure 4.1

Base	Base Type	W1 Description	W2 Discription	Distance	Count sp^2	Count sp^3
Adenine	Purine	N - More Polar	NH2 - More Hindered	≤3.0	96	90
Adenine	Purine	N - More Polar	NH2 - More Hindered	>3.0 and ≤3.2	116	83
Adenine	Purine	N - More Polar	NH2 - More Hindered	>3.2 and <3.4	265	192
Thymine	Pyrimidine	C-Me - Nonpolar	NH2 - More Hindered	≤3.0	96	186
Thymine	Pyrimidine	C-Me - Nonpolar	NH2 - More Hindered	>3.0 and ≤3.2	137	114
Thymine	Pyrimidine	C-Me - Nonpolar	NH2 - More Hindered	>3.2 and <3.4	270	212
Guanine	Purine	N - More Polar	C=O - Less Hindered	≤3.0	121	69
Guanine	Purine	N - More Polar	C=O - Less Hindered	>3.0 and ≤3.2	127	123
Guanine	Purine	N - More Polar	C=O - Less Hindered	>3.2 and <3.4	289	174
Cytosine	Pyrimidine	C-H- Nonpolar	C=O - Less Hindered	≤3.0	78	79
Cytosine	Pyrimidine	C-H- Nonpolar	C=O - Less Hindered	>3.0 and ≤3.2	86	62
Cytosine	Pyrimidine	C-H- Nonpolar	C=O - Less Hindered	>3.2 and <3.4	139	167
				Total	1820	1551

Appendix 14
Source Data for Figure 4.2 and Figure 4.3

Atom Type	Adenine sp^2	Adenine sp^3
N9	10	1
N3	246	238
N1	10	10
C4	31	8
C2	65	36
Total	362	293

Atom Type	Guanine sp^2	Guanine sp^3
N9	4	7
N3	84	120
N2	226	81
N1	46	11
C8	1	0
C4	13	12
C2	35	17
Total	409	248

Atom Type	Thymine sp^2	Thymine sp^3
O2	427	451
N3	43	7
N1	6	3
C6	1	0
C2	36	10
Total	513	471

Atom Type	Cytosine sp^2	Cytosine sp^3
O2	154	243
N3	40	4
N1	0	2
C2	23	7
Total	217	256

Appendix 15
Source Data for Counts Figure 4.4

Base	Hybridization	Backbone Count	Sidechain Count
Adenine	sp^2	73	298
Adenine	sp^3	21	276
Thymine	sp^2	110	411
Thymine	sp^3	41	430
Guanine	sp^2	119	306
Guanine	sp^3	24	229
Cytosine	sp^2	28	191
Cytosine	sp^3	16	241
	Total	432	2382

Base Description	Count
AT Backbone	245
CG Backbone	187
G Backbone	143
C Backbone	44
A Backbone	94
T Backbone	151
AT Sidechain	1415
GC Sidechain	967
A Sidechain	574
T Sidechain	841
G Sidechain	535
C Sidchain	432

Appendix 16
Source Data for Figures 4.5 through 4.8

Adenine sp^2 Side Chain	
Amino Acid	Counts
GLN	2
GLU	2
ASN	3
HIS	5
ASP	6
TRP	9
TYR	15
PHE	61
ARG	195
TOTAL	298

Adenine sp^2 Backbone	
Amino Acid	Counts
CYS	1
ILE	1
SER	1
GLN	2
PRO	2
LEU	3
ASN	4
GLY	9
ARG	50
TOTAL	73

Adenine sp^3 Side Chain	
Amino Acid	Count
ALA	1
ASP	1
HIS	1
TRP	1
ILE	2
VAL	7
GLN	11
LEU	13
SER	21
PRO	29
ARG	30
THR	33
LYS	58
ASN	68
TOTAL	276

Adenine sp^3 Backbone	
Amino Acid	Count
ARG	1
LEU	2
SER	2
GLY	16
TOTAL	21

Appendix 17
Source Data for Figures 4.9 through 4.12

Thymine sp^2 Side Chain	
Amino Acid	**Counts**
GLN	5
GLU	8
TRP	9
HIS	11
ASN	12
TYR	13
PHE	87
ARG	266
TOTAL	411

Thymine sp^2 Backbone	
Amino Acid	**Counts**
ASP	1
ILE	1
PHE	1
PRO	1
THR	1
TYR	1
GLN	2
LEU	2
GLU	3
ASN	5
ALA	6
MET	6
HIS	7
SER	9
LYS	10
ARG	20
GLY	34
TOTAL	110

Thymine sp^3 Side Chain	
Amino Acid	**Counts**
ALA	1
GLU	1
TYR	3
MET	4
PHE	4
GLN	5
ILE	7
THR	9
VAL	14
LEU	19
PRO	25
SER	32
ARG	61
ASN	79
LYS	166
TOTAL	430

Thymine sp^3 Backbone	
Amino Acid	**Counts**
ASP	1
MET	1
PHE	1
THR	1
LYS	2
PRO	2
ARG	3
GLY	30
TOTAL	41

Appendix 18
Source Data for Figures 4.13 through 4.16

Guanine sp^2 Side Chain	
Amino Acid	**Counts**
HIS	8
GLU	15
TRP	21
ASP	27
TYR	28
PHE	29
ASN	38
GLN	42
ARG	98
TOTAL	306

Guanine sp^2 Side Backbone	
Amino Acid	**Counts**
MET	1
TYR	1
GLU	3
LYS	3
LEU	4
ARG	5
VAL	5
GLN	7
SER	7
ASN	9
PHE	9
ILE	10
ALA	11
TRP	16
GLY	28
TOTAL	119

Guanine sp^3 Side Chain	
Amino Acid	**Counts**
HIS	1
ILE	3
TYR	3
VAL	3
ALA	10
MET	11
ARG	12
LEU	17
THR	20
SER	33
GLN	36
ASN	39
LYS	41
TOTAL	229

Guanine sp^3 Backbone	
Amino Acid	**Counts**
ALA	1
MET	1
TYR	1
CYS	2
SER	5
GLY	14
TOTAL	24

Appendix 19
Source Data for Figures 4.17 through 4.20

Cytosine sp^2 Side Chain	
Amino Acid	**Counts**
TYR	2
GLN	5
TRP	5
GLU	6
HIS	10
PHE	31
ARG	132
TOTAL	191

Cytosine sp^2 Backbone	
Amino Acid	**Counts**
ALA	1
ASN	1
LEU	1
LYS	3
MET	3
PHE	3
GLU	4
GLY	12
TOTAL	28

Cytosine sp^3 Side Chain	
Amino Acid	**Counts**
ASP	1
PHE	1
ALA	2
TYR	2
MET	3
SER	3
ILE	5
VAL	5
THR	8
LEU	12
PRO	12
GLN	15
ASN	22
ARG	52
LYS	98
TOTAL	241

Cytosine sp^3 Backbone	
Amino Acid	**Counts**
GLY	16
TOTAL	16

VITA

Laura O'Grady
Candidate for the Degree of
Master of Science/Arts

Thesis: **UNDERSTANDING DNA-PROTEIN HYBRIDIZATION-DEPENDENT INTERACTIONS**
Major Field: Biophysical Chemistry

Laura O'Grady
5 Cymbeline Drive
Old Bridge, New Jersey 08857
e-mail: ltrogrady@optonline.net
Home Telephone: (732) 679 5021

Summary

- Thirty years experience in the pharmaceutical industry
- Demonstrated achiever with exceptional knowledge of clinical trial management
- Skilled project manager with expertise in the administration of worldwide clinical operations from early development activity through regulatory filing
- Extensive people management experience
- Ability to master new concepts quickly, working well under pressure, and communicating ideas clearly and effectively
- Subject matter expert in patient recruitment tracking, study initiation, medical data review study close out activities, and clinical study reporting
- Experience in outsourcing
- Research experience in drug design and synthesis
- Expertise in 3-D molecular modeling with experience in Unix and Fortran programming

Education

M.S. Biophysical Chemistry
Rutgers University Graduate School, *New Brunswick, NJ*
Research in Biophysical Chemistry studying Protein-DNA interactions
42 credits completed, 3 credits in progress
3.9 G.P.A. (out of 4.0)

B.A. Chemistry *1981*
Douglass College, *New Brunswick, NJ*
Concentration in biophysical and organic chemistry with Honors Thesis in DNA conformation analysis. Graduated with Highest Honors and High Distinction in Chemistry
3.98 G.P.A. (out of 4.0)

Career History & Accomplishments

Merck & Co. **1981 to Present**

Senior Clinical Program Manager/Senior Clinical Operations Sp *10/ 200 6 to Pre sent*
CROps /Senior Clinical Trial Manager,
Clinical Development Program Management

- **Lead program operations manager for Zetia/Vytorin/EZ Ato**
  - Manage cross-functional team supporting program and trial level activities that includes but is not limited to study start up, country/site selection, investigators' meetings, recruitment, and study close-out
  - Monitor submission and regulatory activities
  - Supervise cross-functional personnel assigned to program
  - Resource and manpower estimates for budget development
  - Oversight of Contract Research Organization operational activities for program
- **Manager of 6-8 direct reports**
  - Includes objective setting, monthly update meetings, periodic development reviews, management of issues and conflict resolution
- **Subject matter expert for recruitment and study close out**
- **Member of the department training team**
- **Key member of team responsible for development of departm website.**

Clinical Associate, Clinical Research Endocrinology and Metabolism *2/2 002 to 10/ 200 6*

- **Manager of a high priority diabetes program**
  - Coordinated all operational activities across Phase II and III trials.
  - Managed protocol and clinical study report development for program as well submission of IND safety reports.
  - Responsible for staffing and manpower estimates.
  - Supervised and mentored 41 medical program coordinators assigned to program.
  - Oversight responsibility for cross-functional team members
  - Managed development of all Phase III protocols and clinical study reports.
  - Oversight of laboratory data review and adverse experience reporting.
  - Authored Long-term Safety Report for the NDA/WMA of key diabetic compound.
  - Co-authored Safety Update Report to the NDA/WMA of key diabetic compound.
- **Manager of 9 direct reports and 3 indirect reports**
  - Included objective setting, monthly update meetings, periodic development reviews.

Senior Medical Program Coordinator, Clinical Research Endocrinology and Metabolism *2/1998 to 2/2002*

- **Medical Program Coordinator for Merck Schering Plough J Venture**
  - Coordinated activities for a key anti-hypercholesterolemia study involving a rare disease, including protocol and consent form development, investigator meeting coordination, recruitment, data collection, laboratory monitoring, adverse experience reporting, budgets and clinical study report writing.
  - Involved creative and diligent management since disease is so rare.
  - This study was submitted to request expedited review for the NDA for an important anti-atherosclerotic medication.
  - Supervised clinical staff assigned to study.
- **Lead medical program coordinator for critical phase II diab program**
  - Managed activities across program trials, including protocol development, trial site selection, consent form development, investigator meeting coordination, recruitment, data collection, laboratory monitoring, adverse experience reporting, budgets and clinical study report writing.
  - Supervised staff assigned to program

Medical Program Coordinator, Clinical Research Endocrinology and Metabolism *7/1996 to 2/1998*

- **Medical Program Coordinator for a high priority hypercholesterolemia program**
  - Lead coordinator for numerous trials.
  - Responsible for site selection, budget development, consent form development, central laboratory data monitoring, data collection oversight, patient data review, adverse experience reporting and monitoring and generation of clinical study reports.
  - Developed and presented posters for major medical conferences.
  - Key contributor to the development of Merck's first remote electronic data system.
- **Key contributor to WMA for program, which included development of safety counts tables**

Medical Data Coordination Specialist/Senior Medical Data Coordinator, Clinical Biostatistics and Research Data Systems *9/1986 to 7/1996*

- **Lead specialist for Phase II/III endocrine programs as well a ophthalmology program**
  - Supported development of data capture and reporting tools.
  - Performed review of patient data and generated data reports.
  - Supervised staff assigned to program.
  - Managed activities for two outsourced studies.

- Contributed to numerous important NDA/WMA submissions
- **Lead specialist for critical bone density study involving oversight of Contract Research Organization activities**
- **Departmental profit plan administrator including administration of resource and space allocation**

Synthetic Organic Chemist, Membrane and Arthritis Research *6/1 981 to 9/1 986*

- **Synthesis of organic compounds for in-vo testing**
  - Expertise in 90 and 200 MHz NMR, IR, UV, gas spectrometers**,** chromatography and polarimetry.

Memberships & Awards

- Member, Project Management Institute, 2009 (PMP Certification in progress)
- Divisional Staff Awards, Merck 2002 and 2006
- Annual Incentive Awards and Stock Option Awards, Merck Yearly
- John B. Zajac Chemistry Award, Rutgers University
- American Institute of Chemists Award, Rutgers University

Publications

Synthesis and Acylation of 2-Nitro-11H-Dibenzo [b,e][1,4] Dioxepin. William K. Hagmann, **Laura A. O'Grady**, Conrad P. Dorn, James P. Springer. J. Heterocyclic Chem. 23, 673 (1986).

Synthesis and Antiinflammatory/Analgesic Activities of 11H-Dibenzo [b,e][1,4] Dioxepinacetic Acids. William K. Hagmann, Conrad P. Dorn, Robert A. Frankshun, **Laura A. O'Grady**, Philip J. Bailey, Anita Rackham and Harry W. Dougherty. Journal of Medicinal Chem., 1986, 29, 1436.

Inhibition of Human Leukocyte Elastase by C-2 substituted Cephalosporin sulfones. Bonnie M. Ashe, M. Ellen Dahlgren, James B. Doherty, William K. Hagmann, William B. Knight, **Laura A. O'Grady**, Alan L. Maycock. M. Hazel Weston. Eur. Journal of Med. Chem., 24. (1989), 599-604.

Inhibition of Human Leukocyte Elastase. 1. Inhibition by C-7-Substituted Cephalosporin tert-Butyl Esters. James B. Doherty, Bonnie M. Ashe, Perer L. Barker, Thomas J. Blacklock, John W. Butcher, Gilbert O. Chandler, M. Ellen Dahlgren, Philip Davies, Conrad P. Dorn, Jr., Paul E. Finke, Raymond A. Firestone, William K. Hagmann, Thomas Halgren, Wilson B. Knight, Alan L. Maycock, Manuel A. Navia, **Laura A. O'Grady**, Judity M. Pisano, Shrenik K. Shah, Kevan R. Thompson, Hazel Weston and Morris Zimmerman. Journal of Medicinal Chem., 1990, 33, 2513.

Glycolipids as Host Resistance Stimulators. Mitree M. Ponpipom, William K. Hagmann, **Laura A. O'Grady**, Jesse J. Jackson, David D. Wood and Hans J. Zweerink. Journal of Medicinal Chem., 1990 33, 861.

Prevention of Human Leukocyte Elastase - Mediated Lung Damage by 3-Alkyl-4-Azetidinones. William K. Hagmann, Shrenik K. Shah, Conrad P. Dorn, **Laura A. O'Grady**, Jeffrey J. Hale, Paul E. Finke, Kevan R. Thompson, Karen A. Brause, Bonnie M. Ashe, Hazel Weston, M. Ellen Dahlgren, Alan L. Maycock, Pam S. Dellea, Karen M. Hand, Donald G. Osinga, Robert J. Bonney, Philip Davies, Daniel S. Fletcher, James B. Doherty. Bioorganic and Med. Chem. Letters, Vol. 1, No. 10, 545-550, 1991.

Inhibition of Human Leukocyte Elastase. 4. Selection of a Substituted Cephalosporin (L-658,758) as a Topical Aerosol. Paul E. Finke, Shrenik K. Shah, Bonnie M. Ashe, Richard G. Ball, Thomas J. Blacklock, Robert J. Bonney, Karen A. Brause, Gilbert O. Chandler, Meredith Cotton, Philip Davies, Pam S. Dellea, Conrad P. Dorn Jr., Daniel S. Fletcher, **Laura A. O'Grady**, William K. Hagmann, Karen M. Hand, Wilson B. Knight, Alan L. Maycock, Richard A. Mumford, Donald G. Osinga, Paul Sohar, Dean R. Thompson, Hazel Weston, James B. Doherty. Journal of Medicinal Chemistry, Vol. 35, No. 21, pp 3731-3744, 1992.

Comparison of Effects of Simvastatin Versus Atorvastatin on High-Density Lipoprotein Cholesterol and Apolipoproteins A-1 Levels John J.P. Kastelein, MS, PhD, Jonathan L. Isaacsohn, MD, Leiv Ose, MS, Donald B. Hunninghake, MD, Jiri Frohlich, MD, Michael H. Davidson, MD, Rafik Habib, MD, Carlos A. Dujovne, MD, John R. Crouse III, MD, Minzhi LIu, PhD, Michael R. Melino, PhD, **Laura O'Grady**, BA,

Michele Mercuri, MD, PhD, and Yale B. Mitchel, MD, for the Simvastatin Atorvastatin HDL Study Group The American Journal of Cardiology, Vol. 86, NO. 2 pp 221-223, 15 July 2000.
A comparison of Simvastatin and Atorvastatin up to Maximal Recommended Doses in a Large Multicenter Randomized Clinical Trial D. Roger Illingworth, John R. Crouse III, Donald B. Hunninghake, Michael H. Davidson, Ivan D. Escobar, Anton F. H. Stalenhoef, Gyorgy Paragh, Patrick, T.S. Ma, Minzhi Liu, Michael R. Melino, **Laura O'Grady**, Michele Mercrui and Yale B. Mitchel for the Simvastatin Atorvastatin HDL Study Group Current Medical Research and Opinion, Vol. 17, No. 1, 2001