

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Strategies for specific recognition of RNA using hydrogen bonding

Permalink

<https://escholarship.org/uc/item/0996z4k4>

Author

Cheng, Alan C.

Publication Date

2002

Peer reviewed|Thesis/dissertation

**Strategies for specific recognition
of RNA using hydrogen bonding**

by

Alan C. Cheng

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

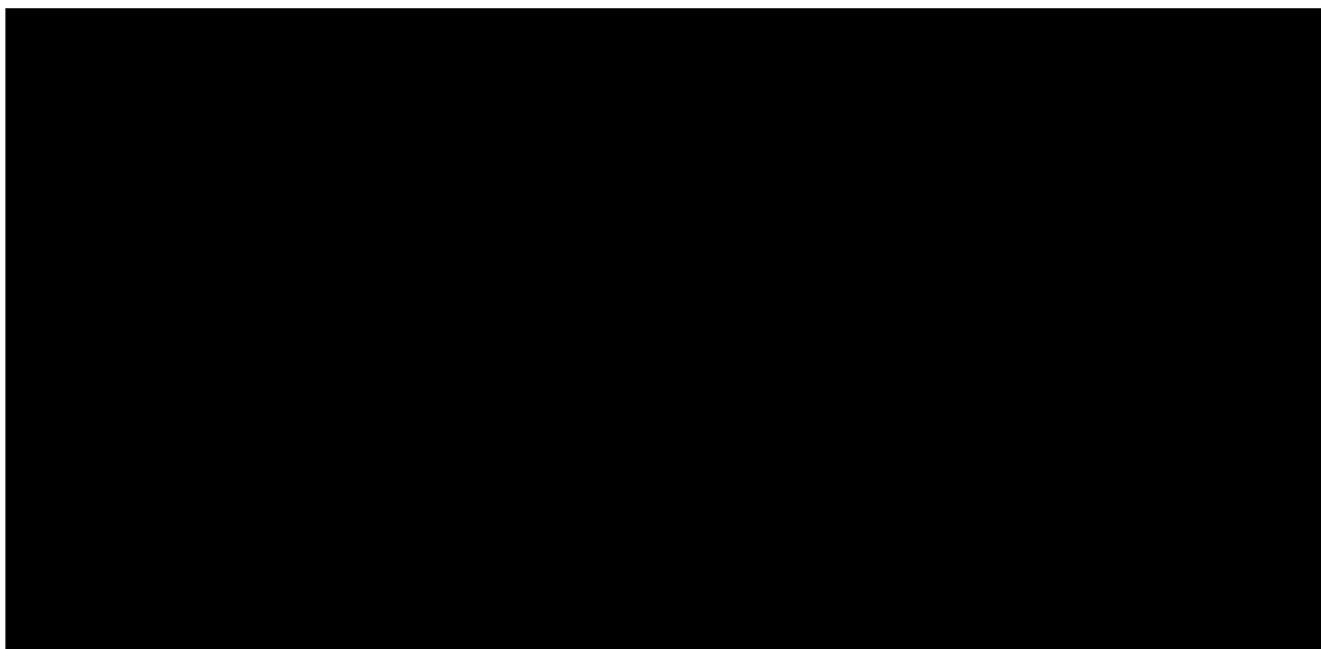
Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Intentionally Blank

Intentionally Blank

Preface

I would like to first acknowledge my parents and sisters for their unconditional caring and support. I would also like to acknowledge the wonderful people and friends I've met during my time at UCSF, not least of which are the people I encountered in the Frankel lab, including: Ami Bhatt, Valerie Calabro, Donna Campisi, Lily Chen, William Chen, Chandreyee Das, Steve Edgcomb, John Fisher, Cynthia Fuhrmann, Kazuo Harada, Cindy Honchell, Ashwini Jambhekar, Steve Landt, Aenoch Lynn, Emily Mai, Damien McColl, Rob Nakamura, Jemmy Nejim, Ralph Peteranderl, Jim Robertson, Colin Smith, Roying Tan, Chris Tang, Bernhard Walberer, Bixun Wang, Hadas Zehavi. I would like to acknowledge the many faculty I was privileged to interact with: Jim Cleaver, Ken Dill, Dennis Deen, Lily Jan and Dan Minor, Dick Shafer, Tack Kuntz, Fred Cohen, and many others. I am grateful for the invaluable help of Julie Ransom, and the members of my thesis and orals committees: David Agard, Liz Blackburn, Tom Ferrin, Peter Kollman, and Wendell Lim. In particular, David Agard was supportive and helpful throughout my graduate career. Peter Kollman was a great teacher, inspiration, and motivator. And last but not least, I am indebted to the tremendous guidance, insight, and teaching of Alan Frankel, without whom this thesis certainly would not exist today.

Abstract

Specificity of protein-RNA recognition is essential to biological processes and viral lifecycles, and is determined in large part by hydrogen bonded interactions between amino acid side chains and nucleic acids. To understand the interaction preferences and to develop strategies for specific recognition of RNAs, we have calculated and analyzed all possible hydrogen bonded patterns between side chains and units of RNA structure, including unpaired bases, the 53 possible RNA base pairs, A-form RNA triplets, and B-form DNA. We find 32 possible bidentate interactions between side chains and the six unpaired bases, and we use quantum chemical methods to explore the relative energetic contributions of these interactions, and the correlation with experimentally-observed frequencies. We find 186 spanning interactions to base pairs, of which a three-hydrogen bonded Arg-Wobble base pair may be particularly biologically relevant. We provide supporting evidence for the interaction, and suggest an experiment to directly address the presence of three hydrogen bonds. We model possible complex multi-step interactions to idealized DNA and RNA helices, and find that 3' cross strand interactions are significantly more favorable in ARNA. Finally, we propose an approach to looking at non-ideal nucleic helices, and we propose a method for generating highly specific ligands to tertiary RNAs.

Table of Contents

Chapter 1	Introduction.....	1
Chapter 2	Modeled databases of amino acid-base and amino acid-base pair hydrogen-bonding interactions.....	12
Chapter 3	<i>Ab-initio</i> interaction energies of amino acid side-chain analogues hydrogen bonding with nucleic acid bases, and correlations with observed frequencies.....	62
Chapter 4	Strategies for discrimination of A-form RNA and B-form DNA by single protein side chains.....	85
Chapter 5	Future Directions.....	143
Appendix A	Protocol for building helix interaction databases.....	158
Appendix B	Schema and sample SQL scripts for helix interaction databases.....	164
Appendix C	NailMine: An user-friendly program for mining the interaction databases	171

List of Tables

2.1	Numbers of computed amino acid-base and amino acid-base pair interactions	49
2.2	Nucleic acid-protein complexes from the PDB utilized in this study	50
2.3	Observed amino acid-base interactions	51
2.4	Calculated amino acid-base pair spanning interactions	52
3.1	Ranking of side chain-base interaction energies	77
3.2	Observed Ser/Thr/Tyr-A intermolecular hydrogen bonds	78
3.3	Secondary structures of Asn/Gln-A interactions found in crystal and NMR structures	79
4.1	Number of interactions for ARNA and BDNA found after each filter step	123
4.2	Number of interactions to phosphate and sugars in ARNA and BDNA, broken out by the four types of base patterns	124
4.3	Characteristics of modeled interactions to ARNA	125
4.4	Characteristics of modeled interactions to BDNA	127
4.5	Base specific amino acid recognition of triplet sequences	129
4.6	Interactions from the PDB involving an O4' and at least one base atom ...	131
4.7	Interactions found in the PDB involving an O2' and at least one base atom	132
5.1	Results of quantum chemical energy calculations	149
B.1	Atom codes used in database for amino acids and nucleic acids	167

List of Figures/Illustrations

2.1	Schematic of the WASABI search method	53
2.2	Generation of conformational diversity by DIVERSIGEN	54
2.3	Interaction of Ser and Tyr with a G:G base pair demonstrating a steric clash with Ser but not Tyr	55
2.4	Amino acid-base interactions with two or more hydrogen bonds	56
2.5	Spanning interactions utilizing three hydrogen bonds	57
2.6	Observed spanning interactions	58
2.7	Possible and observed spanning interactions to the Watson-Crick base pairs	59
2.8	Possible spanning interactions to the G:U wobble pair	60
2.9	Similarity of a G-G:U base triple and a modeled Arg-G:U wobble interaction	61
3.1	All two hydrogen bond interactions between a side-chain and a base, as modeled using WASABI	83
3.2	HF/6-31G** geometry optimization of the models	84
4.1	Overview of A-form and B-form nucleic acids	132
4.2	A-form RNA base patterns	133
4.3	B-form DNA base patterns	134
4.4	Sampling of modeled interactions	135
4.5	Example of an interaction of the type X_Z, where the middle base pair is not directly contacted but is partially sequence specific	136

4.6	Three-hydrogen bond interactions in ARNA and BDNA, divided into X_Z and XYZ type interactions	137
4.7	All non-bifurcated hydrogen bonding interaction models involving the ribose 2'OH donor or O2' acceptor, broken out by base pattern	138
4.8	ARNA difference figure	139
4.9	BDNA difference figure	140
4.10	Summary of observed base-specific interactions involving more than one step	141
5.1	Reference interactions used for Arg-GU energy calculations	
5.2	Modeled complex of the arginine-GU interaction and reduced model used in quantum chemical calculations	
5.3	Optimized geometries of the Arg-Wobble interaction at four successive levels of theory	150
B.1	Database schema for helixdb2	166

Chapter 1

Introduction

Intricate cellular processes in animals, plants, bacteria, and viruses underlie life as we know it. Molecules binding to RNA are essential in the processes of life, and the degree of specificity of these interactions is critical to establishing and maintaining harmony in cells. For instance, in the inherited disease thalassemia, a mutation at the spliceosome-RNA interface [Gorman 2000] leads to aberrant splicing of the hemoglobin gene, resulting in potentially fatal bone, heart, liver and spleen dysfunction. In AIDS, the binding of HIV Tat protein to TAR RNA and the binding of HIV Rev protein to RRE RNA are essential events for viral propagation. Specific ligand-RNA recognition is key to the effectiveness of the antibiotic streptomycin which functions by interfering with bacterial ribosomal RNA specifically over mammalian ribosomal RNA [Brodersen 2000] [Carter 2000] [Schlunzen 2001].

Specificity in protein-RNA interactions is determined by a number of molecular interactions, including charge-charge, hydrogen bonding, van der Waals (vdw), and hydrophobic interactions. A single hydrogen bond interaction has been reported to provide 0.5-1.5 kcal/mol stabilization, whereas a single vdw interaction provides ~0.2 kcal/mol, and hydrophobic interactions provide ~20cal/A² [Creighton 1983]. The energetic contribution of charge-charge interactions varies widely in biological milieu, and is highly dependent on the surrounding environment. Na⁺ and Cl⁻, for instance, have a strong 120 kcal/mol interaction in the gas phase, but have only a 1.5 kcal/mol intermolecular affinity in water. Hydrogen bonds are good targets for computational modeling approaches to understanding specificity of interactions to nucleic acids because (1) they are a dominant player in protein-nucleic interactions and represent two-thirds of all base-specific contacts [Luscombe 2001] [Mandel-Gutfreund 1995], (2) their angle and

distance preferences are strong and well characterized [Taylor 1983] [Taylor 1984] [Taylor 1984], and (3) there is a clear basis for specificity, that is, we can model specificity based on two simple assumptions: hydrogen bonds need to be satisfied, and multiple hydrogen bonds lead to increased specificity.

Intermolecular hydrogen bonds have long been postulated to be important in specificity of protein-nucleic interactions. Seeman, Rosenberg and Rich, in a key 1976 work [Seeman 1976], systematically examined the possible hydrogen bonding interactions between amino acid side chains and groups along the edges of the Watson-Crick DNA base pairs. They concluded that interactions involving two hydrogen bonds would be required to uniquely distinguish each base pair from the others, and inferred that two hydrogen bonds from a single functional group would specify a site with higher precision than from two independent groups. Based on this prediction, they predicted the occurrence of two interactions: an arginine-guanine (Arg-G) interaction, where the guanidinium group of Arg donates two hydrogen bonds to the O6 and N7 hydrogen bond acceptors of G, and an asparagine-adenine (Asn-A) interaction, where the carboxamide makes two hydrogen bonds to the N6 donor and N7 acceptor of A. The two predictions, made before a single protein-nucleic structure had been solved, have been shown to be important in protein-nucleic recognition. A recent study [Luscombe 2001] done with a set of 129 high-resolution crystal structures of protein-DNA complexes resulted in 54 examples of bidentate interactions between single side-chains and single bases, of which 43% is the mentioned Arg-G interaction and another 26% is the Asn/Gln-A interaction. Inclusion of another 25 out of 26 Arg-G interactions that are classified as single hydrogen bonding interactions but are likely to be bidentate interactions [Luscombe 2001] would

shift the percentage of interactions of 61% Arg-G interactions and 18% Asn/Gln-A. Thus 70 to 80% of the observed bidentate interactions were predicted by Seeman et al from simple consideration of possible hydrogen bonding patterns.

Efforts to identify a simple and general “recognition code” for protein-DNA interactions have been largely unsuccessful however. Common interaction patterns have been found within a given structural context such as the zinc-finger or helix-turn-helix [Suzuki 1994] [Wolfe 2000]. A statistical study of 52 protein-DNA complexes derived a set of amino acid-base and amino acid-backbone propensities, based on hydrogen bonding, hydrophobic interactions, and the relative positioning of C atoms, that could reasonably predict the preferred binding sites for several DNA-binding proteins [Kono 1999]. Even within a given structural context, however, variations in nucleic and protein backbone geometries can lead to differences in “recognition codes” [Elrod-Erickson 1998][Wolfe 2000]. In addition, dynamics and kinetics of induced fit and conformational capture can be important factors in specific recognition [Frankel 1998] [Williamson 2000], but are currently not easily considered.

However, the successes of the Seeman et al. and Kono et al. predictions and the clear importance of hydrogen bonding at protein-nucleic interfaces [Luscombe 2001] demonstrate that studies of specificity strategies for discrimination of nucleic acids provide useful foundations for understanding protein-nucleic specificity. The analysis of protein-RNA interactions is at an earlier stage compared to analysis of protein-DNA interactions, with only about 45 available structures that represent only a small subset of possible RNA tertiary elements. However, some characteristics are beginning to emerge [Allers 2001] [Treger 2001] [Jones 2001]. With respect to hydrogen bonding, perhaps

the most obvious difference between RNA and DNA complexes is the use of the ribose 2'OH group in about a quarter of all hydrogen bonds [Treger 2001]. In addition, of all the observed base-specific interactions, hydrogen bonds appear less dominant than in DNA complexes [Jones 2001] [Allers 2001] [Treger 2001], probably because a significant number of bases are not stacked within Watson-Crick duplexes and consequently some bases are sequestered from solvent via van der Waals interactions with the protein. Nevertheless, the importance of hydrogen bonding for RNA-binding specificity is as apparent for RNA as it is for DNA.

In RNA, a great variety of non-canonical features augment the standard base-paired Watson-Crick helices. Not only can RNA involve 51 base pairs in addition to the 2 Watson-Crick base pairs [Walberer 2002], but RNA can also involve a variety of loops, bulges, and flipped out bases that contribute to formation of exquisite tertiary structures [Sanger 1984]. RNA binding proteins often target non-canonical features to gain specific recognition [Draper 1999].

The construction and analysis of databases of all possible hydrogen-bonding interactions between amino acids and units of RNA structure thus can be useful in the understanding, prediction, and design of proteins and peptides that bind specifically to RNAs. The enumerated interactions can be useful by itself as a reference database. For instance, given mutational data identifying the importance of a base and amino acid in protein-RNA recognition for a particular complex, the database of interactions can suggest recognition and binding modes based on hydrogen bonding. I can also pick out interactions that can be especially useful in gaining recognition specificity. From the design perspective, the specific interactions I identify may be useful in the design of

proteins that specifically recognize a given RNA. While the Seeman et al. modeling studies are highly successful in predicting the frequency of interactions, the commonality of the predicted Arg-G and Asn-A make them less useful in designing specificity. However, prediction of interactions to unique noncanonical RNA features does not suffer the same issue, and thus are likely to be significantly more useful in the design and analysis of specific RNA-protein binding.

In this work, I build and analyze all possible hydrogen bonding patterns for amino acids interacting with progressively more complicated units of RNA structure. Chapters 2, 3, and 4 are completed or nearly completed manuscripts that describe the work done for single bases, base-pairs, and, lastly, canonical helices. Chapter 5 surveys the future directions of this work and steps taken toward achieving them, and presents a method for design of RNA-binding small molecule ligands.

Chapter 2 introduces the general method that I use, which is encapsulated in a program called WASABI. The method is applied to predict bidentate hydrogen-bonding interactions to single bases. Single bases are found flipped out into solution in RNA bulges, such as that found in HIV RRE, and loops, such as that found in U1A snRNA, providing novel recognition elements for proteins. I look at spanning interactions to the 53 possible RNA base pairs. Spanning interactions involve a minimum of one hydrogen bond to each base of the base pair, allowing recognition of the base pair as a unit. Analysis of the spanning interactions suggests strategies for specificity, including a three hydrogen-bond Arg spanning interaction to the major groove of the wobble base pair.

In Chapter 3, quantum-chemical methods are used to derive a ranking of our predicted hydrogen-bonded interactions between amino acids and single bases. The

results of the calculation provide a relative measure of the inherent stability of each interaction pattern, and can be used as a basis for understanding other contributors to specificity. I find a good correlation between our interaction energy rankings and observed occurrences of the same interactions in the Protein Data Bank (PDB).

In Chapter 4, the WASABI method is applied to the next level of RNA structure, the idealized ARNA, and I examine how differences in the helical rise and displacement of ARNA and BDNA give rise to significant differences in protein recognition strategies. Because small peptide motifs, such as alpha helices and beta turns, can be used to specifically recognize either DNA or RNA, one question I explore is how single amino acids making multiple hydrogen bonds can be used to differentiate one nucleic helix form over another. In practice such discrimination will be only one component of recognition, but it can be important in small peptides where a few residues are tasked to specify a complicated site.

Chapter 5 describes a proposed experiment for testing an interesting Arg-Wobble interaction from Chapter 2, and future directions for exploring specific recognition to more complex RNA structures. Extending the studies of idealized helices, I propose a method of looking at specificity to non-ideal nucleic helices. Interactions between molecules and tertiary RNA structures are just beginning to be explored [Cheng 2001]. I hypothesize the importance of hydrogen bonds in ligand recognition at ribosomal binding sites, and suggest a possible plan for experimental design of specific RNA-binding drugs that might be used to disrupt microbial ribosomal function and alter disease progression in cases such as AIDS and Fragile X.

References

Allers, J & Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.* **311**, 75-86.

Brodersen, D. E., Clemons, W. M. Jr., Carter, A. P., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. (2000). The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit. *Cell.* **103**, 1143-1154.

Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B.T. & Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature.* **407**, 340-348.

Cheng, A. C., Calabro, V., Frankel, A. D. (2001). Design of RNA-binding proteins and ligands. *Curr. Opin. Struct. Biol.* **11**, 478-484.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. 2nd Edition. W. H. Freeman and Company, New York.

Draper, D. E. (1999). *Themes in RNA-protein recognition.* *J Mol. Biol.* **293**, 255-270.

Elrod-Erickson, M., Benson, T. E. & Pabo, C. O. (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure*. **6**, 451-464.

Frankel, A. D. & Smith, C. A. (1998). Induced folding in RNA-protein recognition: more than a simple molecular handshake. *Cell*. **92**, 149-151.

Gorman, L., Mercatante, D. R. & Kole, R. (2000). Restoration of correct splicing of thalassemic beta-globin pre-mRNA by modified U1 snRNAs. *J Biol. Chem.* **275**, 35914-35919.

Jones, S., Daley D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. (2001). Protein-RNA interactions: a structural analysis. *Nucl. Acids Res.* **29**, 943-54.

Kono, H. & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*. **35**, 114-131.

Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860-2874.

Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370-382.

Miller, J. C. & Pabo, C. O. (2001). Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J. Mol. Biol.* **313**, 309-315.

Saenger, W. (1984). *Principles of nucleic acid structure*. Springer-Verlag. New York.

Schlunzen, F., Zarivach, R., Harms, J., Bashan, A., Tocilj, A., Albrecht, R., Yonath A. & Franceschi, F. (2001). Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature*. **413**, 814-821.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci.* **73**, 804-808.

Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl. Acad. Sci.* **91**, 12357-12361.

Taylor, R., Kennard, O. & Versichel, W. (1983). Geometry of the imino-carbonyl (N-H...O:C) hydrogen bond. 1. Lone-pair directionality. *J. Am. Chem. Soc.* **105**, 5761-5766.

Taylor, R., Kennard, O. & Versichel, W. (1984). Geometry of the N-H...O=C hydrogen bond. 2. Three-center ("Bifurcated") and four-center ("Trifurcated") bonds. *J. Am. Chem. Soc.* **106**, 244-248.

Taylor, R., Kennard, O. & Versichel, W. (1984). The Geometry of the N-H...O=C hydrogen bond. 3. Hydrogen-bond distances and angles. *Acta Cryst. B.* **40**, 280-288.

Treger, M. & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol. Recognition.* **14**, 199-214.

Williamson, J. R. (2000). Induced fit in RNA-protein recognition. *Nature Struct. Biol.* **7**: 834-837.

Wolfe, S. A., Nekludova, L. & Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Ann. Rev. Biophys. Biomol. Struct.* **29**, 183-212.

Chapter 2

1. The first part of the chapter discusses the importance of understanding the context of the data being analyzed. This includes identifying the source of the data, the methods used to collect it, and any potential biases or limitations. The second part of the chapter focuses on the various statistical techniques used to analyze data, including descriptive statistics, inferential statistics, and regression analysis. The third part of the chapter discusses the importance of interpreting the results of the analysis in the context of the research question. The fourth part of the chapter discusses the importance of communicating the results of the analysis to a non-technical audience. The fifth part of the chapter discusses the importance of ethical considerations in data analysis.

**Modeled databases of amino acid-base and amino acid-base pair
hydrogen-bonding interactions**

Alan C. Cheng*, William W. Chen, Cynthia N. Fuhrmann, and Alan D. Frankel

Department of Biochemistry and Biophysics

and

*Graduate Group in Biophysics

University of California, San Francisco

San Francisco, CA 94143-0448

Address correspondence to: Alan Frankel
Department of Biochemistry and Biophysics
UCSF
513 Parnassus Avenue
San Francisco, CA 94143-0448

Telephone: 415-476-9994
FAX: 415-502-4315
e-mail: frankel@cgl.ucsf.edu

ABSTRACT

Sequence-specific protein-RNA recognition is determined in part by hydrogen bonding interactions between amino acid side chains and nucleotide bases. To examine the repertoire of possible interactions, we have calculated all geometrically plausible arrangements in which amino acids hydrogen bond to unpaired bases, such as those found in RNA bulges and loops, or to the 53 possible RNA base pairs. We find 32 possible interactions that involve two or more hydrogen bonds to the six unpaired bases (including protonated adenine and cytosine), 17 of which have been observed. We find 186 “spanning” interactions to base pairs in which the amino acid hydrogen bonds to both bases, in principle allowing particular base pairs to be selectively targeted. Of the spanning interactions, 4 are to the Watson-Crick pairs and 15 are to the G:U wobble pair commonly found in RNA structures, including an interesting arrangement involving three hydrogen bonds to the Arg guanidinium group. A systematic analysis of the computed databases reveals that interactions involving two hydrogen bonds to U (or T) can occur only if the base is unpaired, suggesting a possible role for bulged Us in protein recognition. In general, the distribution of donors and acceptors on the bases allows Asn/Gln, which has both acceptor and donor groups, to make numerous interactions, and Asp/Glu, which has two acceptors, to make relatively few. The databases highlight some general characteristics of amino acid-base hydrogen bonding and may be useful for analyzing experimental data, devising tests of proposed interactions, and designing novel interactions.

INTRODUCTION

The ability of proteins to recognize specific RNA sites is important in many biological systems but the “rules” governing these interactions are not well understood. This is in part because the structural database of protein-RNA complexes is still relatively limited (despite the recent addition of the ribosomal subunit structures) and, perhaps more importantly, because RNA structures are so diverse. In addition to Watson-Crick helices, RNAs often contain non-Watson-Crick base pairs, unpaired bases such as those in bulges and loops, and base triples or other higher order tertiary interactions (Draper, 1999; Hermann & Patel, 1999). Therefore many of the principles of protein-DNA recognition inferred from the large number of solved structures (Luscombe *et al.*, 2001; Pabo & Nekludova, 2000) do not apply to RNA.

Despite the current gaps in knowledge, it is apparent that one important determinant of specificity in both DNA and RNA complexes is the complementary nature of hydrogen bonding interactions between polar groups on the protein side chains and nucleic acid bases. A seminal study by Seeman *et al.* conducted before the structure of even a single protein-nucleic acid complex had been solved (Seeman *et al.*, 1976) systematically examined the possible hydrogen bonding interactions between amino acid side chains and groups along the edges of the Watson-Crick base pairs. They concluded that interactions involving two hydrogen bonds would be required to uniquely distinguish each base pair from the others, and inferred that two hydrogen bonds from a single functional group would specify a site with higher precision than from two independent groups, analogous to the “chelate effect” in which formation of one bond favors formation of additional bonds by an increase in effective concentration (Creighton, 1993). Based on their analysis, Seeman *et al.* predicted two interactions in which precisely positioned side chains in the DNA major groove could discriminate amongst all the base pairs: one in which the guanidinium group of Arg donates two hydrogen bonds to the O6

and N7 acceptor groups of guanine (G) and a second in which the carboxamide group of Asn (or Gln) hydrogen bonds to the N7 acceptor and N6 donor groups of adenine (A). These interactions are indeed the most commonly observed in protein-DNA complexes (Luscombe et al., 2001; Lustig & Jernigan, 1995; Mandel-Gutfreund *et al.*, 1995; Pabo & Sauer, 1992), and the importance of such direct amino acid-base hydrogen bonds in determining sequence specificity has been confirmed by many structure-function studies.

Several detailed studies have analyzed the interactions observed in protein-DNA complexes, partly in efforts to determine whether a “recognition code” exists for DNA double helices (Choo & Klug, 1997; Jones *et al.*, 1999; Kono & Sarai, 1999; Luscombe et al., 2001; Lustig & Jernigan, 1995; Mandel-Gutfreund & Margalit, 1998; Mandel-Gutfreund et al., 1995; Pabo & Nekludova, 2000; Suzuki, 1994). It seems clear that while no simple code exists, some common interaction patterns can be found, particularly within a given structural context such as the zinc finger or helix-turn-helix motif (Choo & Klug, 1997; Pabo & Nekludova, 2000; Suzuki & Yagi, 1994). Hydrogen bonding interactions to the bases comprise about two-thirds of all base-specific contacts (Luscombe et al., 2001), and interactions involving two hydrogen bonds from the side chain are dominated by the Arg-G and Asn(Gln)-A interactions described above. The only other frequent bidentate interaction utilizes the amino group of Lys to hydrogen bond to the O6 and N7 acceptors of guanine, although other two hydrogen-bond interactions are found that utilize bifurcated bonds or donors and acceptors from the peptide backbone (Luscombe et al., 2001). A statistical survey of 28 protein-DNA complexes found that side chains possessing both donor and acceptor atoms more frequently use the donor atom for hydrogen bonding (Mandel-Gutfreund et al., 1995). Despite the importance of direct amino acid-base hydrogen bonds in determining DNA sequence-specificity, it is clear that many other types of interactions are used, including water-mediated hydrogen bonds, van der Waals contacts, and interactions to the sugar-phosphate backbone, and that the structural context in which the interactions are

presented is an inherent part of the recognition process (Jones et al., 1999; Kono & Sarai, 1999; Luscombe et al., 2001; Pabo & Nekludova, 2000; Suzuki, 1994). A statistical study of 52 protein-DNA complexes derived a set of amino acid-base and amino acid-backbone propensities, based on hydrogen bonding, hydrophobic interactions, and the relative positioning of C α atoms, that could reasonably predict the preferred binding sites for several DNA-binding proteins (Kono & Sarai, 1999).

The analysis of protein-RNA interactions is at an earlier stage and the available structures represent only a small subset of possible RNA tertiary elements, but some characteristics are beginning to emerge (Allers & Shamoo, 2001; Draper, 1999; Jones *et al.*, 2001; Steitz, 1999; Treger & Westhof, 2001). With respect to hydrogen bonding, perhaps the most obvious difference between RNA and DNA complexes is the use of the ribose 2'OH group in about a quarter of all hydrogen bonds (Treger & Westhof, 2001). In addition, of all the observed base-specific interactions, hydrogen bonds appear less dominant than in DNA complexes (Allers & Shamoo, 2001; Jones et al., 2001; Treger & Westhof, 2001), probably because a significant number of bases are not stacked within Watson-Crick duplexes and consequently some bases are sequestered from solvent via van der Waals interactions with the protein. Nevertheless, the importance of hydrogen bonding for RNA-binding specificity is as apparent for RNA as it is for DNA. Here we report the construction of databases of all possible hydrogen-bonding interactions between amino acids and bases or base pairs that can occur in RNA structures. The problem is more complex than that faced by Seeman et al. (Seeman et al., 1976) in that many more RNA base configurations are possible beyond those in Watson-Crick helices and thus a systematic computational approach is required. The databases include interactions between unpaired bases, such as those found in bulges or loops, and non-Watson-Crick base pairs, some of which involve multiple hydrogen bonds that may be used to uniquely recognize bases in particular structural contexts. Because noncanonical features are important recognition elements of RNA structures, the database of

interactions may be even more useful for RNAs than it is for DNAs in analysis of recognition strategies. The databases may be useful not only for analyzing existing interactions but perhaps also for designing novel interactions in RNA-binding proteins or peptides.

METHODS

Database Construction

The approach for generating all possible hydrogen-bonding arrangements of amino acid side chains with bases or bases pairs utilizes simple geometric and steric criteria and is illustrated in Fig. 1A. The program WASABI (What Are the Specific Amino acid-Base Interactions) first forms a single linear hydrogen bond between a side chain and base for every possible combination of donor and acceptor groups, and then samples the allowable three-dimensional conformations by rotating the side chain around each of five angles (pivoting around the donor or acceptor atoms) while still maintaining the initial bond. Rotations in only five of the six planes are sufficient to sample the available space because one rotation is symmetric along the axis of the hydrogen bond and would be redundant for the donor and acceptor sites. Each conformation is evaluated for the formation of additional hydrogen bonds (using the parameters shown in Fig. 1B) and the absence of steric clashes. The “best” conformation, as judged by a simple scoring function that favors linear hydrogen bonds, is identified and all unique hydrogen-bonding arrangements are stored, including those with single hydrogen bonds. This three-dimensional search algorithm is an adaptation of a two-dimensional version used to generate all possible base-base combinations (Walberer, 2000; Walberer *et al.*, 2002).

Conformational searches were performed utilizing the nine fixed hydrogen-bonding side chain moieties shown in Fig. 1C, including unprotonated and protonated forms of histidine, and either with the six RNA bases (A, C, G, U, A+, C+) or 53 possible RNA base pairs generated by Walberer *et al.* (Walberer *et al.*, 2002). In addition, we constructed interactions with the additional DNA base, thymine, and the 17 possible base pairs that utilize thymine (Walberer, 2000). These types of amino acid-DNA interactions may occur in the context of single-stranded sites or in helices with extruded bases.

For our steric parameters, van der Waals radii taken from AMBER parm94 (Cornell *et al.*, 1995) were divided by $2^{1/6}$ to approximate hard sphere radii (Israelachvili, 1989); these radii were further reduced by 0.8 to include geometries slightly outside a reasonable steric range. Polar hydrogens were assigned van der Waals radii of 0.2 and also were further reduced by the 0.8 steric parameter. Because the polar hydrogens had such small radii, we implemented a filter that removed conformations in which two polar hydrogens were closer than 2.5Å, thereby eliminating unfavorable arrangements with two nearby positive charges. Amino acid side chains were constructed using the LEAP package with param96 residue definitions (Cornell *et al.*, 1995).

The parameters used to define a hydrogen bond (Fig. 1B) were chosen based on the analysis of small molecule high-resolution crystal structures (Taylor & Kennard, 1984; Taylor *et al.*, 1983) and on an empirical test of the donor angle parameter. A maximum distance of 3.4Å was used for all hydrogen bonds. We estimated acceptor and donor parameters from the small molecule analysis, which reported only acceptor-hydrogen-donor atom angles, by assuming a hydrogen-acceptor length of 2.0Å and a donor-hydrogen bond length of 1.0Å and using donor angle = $\sin^{-1}(D)$, where D is as described previously (Taylor & Kennard, 1984). For the nitrogen acceptor angle, we found that values of $0\pm 22^\circ$ for two-center bonds and $0\pm 45^\circ$ for three-center bonds were 3 standard deviations from the mean and therefore included >99% of observed hydrogen bonds. Thus, we used a nitrogen acceptor angle of $0\pm 50^\circ$ to include slightly unreasonable geometries and to allow us to incrementally sample conformations using a 4° step size. We used an oxygen acceptor angle of $0\pm 90^\circ$ for similar reasons. To determine our donor angle parameter, we performed a set of WASABI calculations using angles of $0\pm 30^\circ$, 32° , 34° , 36° , 38° , and 40° and found that $0\pm 36^\circ$ generated all known interactions (see below) and that at least some of the additional conformations generated using a $0\pm 38^\circ$ angle appeared reasonable by inspection. A similar empirical approach was used to select the $0\pm 18^\circ$ donor angle parameter used to construct the base-base interaction database, which

was substantially more restrictive due to the planar nature of the conformational search (Walberer et al., 2002).

As mentioned above, WASABI generates multiple conformations with the same hydrogen bonding arrangement and thus we devised an empirical scoring function in order to select a representative conformation with as planar an arrangement as possible. Scores (S) were calculated as follows: $S = w_1 * (1 + [A/r]^4 - [B/r]^2) + w_2 * \sin^2 d + w_3 * \sin^2 a + w_4 * \sin^2 p$, where $w_1:w_2:w_3:w_4 = 100:30:1:1$ for oxygen acceptors and $100:30:10:0$ for nitrogen acceptors, a is the acceptor angle, d is the donor angle, p is the angle between the plane of the side chain and the plane of the base or base pair, r is the distance between the non-hydrogen donor and acceptor, and A and B are parameterized to mean hydrogen bond distances of 2.95Å for N-O bonds, 2.73Å for O-O bonds, and 2.90Å for N-N bonds, which are average distances calculated from the database of known protein-nucleic acid complexes and similar to those previously reported (Baker & Hubbard, 1984; Jeffrey & Saenger, 1991; Saenger, 1984; Taylor & Kennard, 1984; Taylor et al., 1983).

Finally, we wished to ensure that each calculated arrangement could accommodate a nucleotide backbone and complete amino acid side chain. We added C2' endo or C3' endo ribose sugars (generated using AMBER parm94 parameters) to each base in a combinatorial manner, rotating the sugars by 360° around the C1'-N1 bond in 2° increments. We similarly added all amino acid rotamers (Dunbrack & Cohen, 1997) (August 10, 1999 release) in a combinatorial manner and identified any model in which no set of sugar and rotamer conformations could be accommodated sterically. These models were analyzed further using DIVERSIGEN as described below. Although we used one hydrogen bond moiety to represent Asn(Gln), Asp(Glu), and Ser(Thr) side chains (Fig. 1C), rotamers of all represented amino acids also were added for these final steric tests. Interestingly, two interactions involving bifurcated hydrogen bonds with Ser and Thr were found to be sterically impossible but could occur with Tyr. Despite the

larger size of the Tyr side chain, the planarity of the aromatic ring makes the interaction more favorable than with the Ser or Thr side chains (see Results and Discussion).

Diversity Generator

Each hydrogen-bonding arrangement is represented in our databases by a single conformation, however many three-dimensional conformations typically are possible for each arrangement. We constructed a diversity generating program, DIVERSIGEN, that begins with one conformation and creates a set of conformations chosen to represent the sterically accessible space for the particular hydrogen-bonding arrangement. The program generates ~10,000 conformations (or 1000-4000 in a few particularly sterically restricted cases) which then are clustered into a specified number (10 to ~10,000). For arrangements that could not accommodate the nucleotide sugars and side chain rotamers (see above), we generated 10 conformations and tested each for its ability to accommodate the sugars and rotamers. Those few arrangements that could not (see Table 1) were eliminated from the databases.

To generate conformational diversity, first the length of each hydrogen bond in a particular arrangement is set to three values, corresponding to short, median, and long distances that cover the experimentally observed range for each type of donor-acceptor pair. Next, the same five angles varied in the WASABI search again are incrementally varied, beginning with a large step size, and hydrogen bond distances and angles are monitored and steric tests performed using the parameters described above to retain plausible conformations with the appropriate hydrogen bonds. Conformations generated using each of the starting hydrogen bond lengths are retained, until a total of ~10,000 are generated. The step size used for each of the angles varied is adjusted iteratively to achieve the desired 10,000 conformations. These conformations then are clustered into the desired number of representatives, chosen to cover conformational space as completely as possible. It is particularly difficult to achieve a good representation when

choosing a small number of conformations to represent a broad space. To assess whether a chosen set of conformations reasonably represents the space sampled, we define similarity of any two conformations as the Euclidean distance between the five parameters of the WASABI search. For conformations **a** and **b**, with parameter coordinates $\{a_1, a_2 \dots a_6\}$, $\{b_1, b_2 \dots b_6\}$, Euclidean distance (d_E) is defined by $d_E = \|\mathbf{a}-\mathbf{b}\|$. For any given d_E , conformations may be grouped and represented by one suitable conformation as long as they fall within the same d_E range. Thus, we are defining conformational similarity based on hydrogen bonding parameters and not on the r.m.s.d. of three-dimensional coordinates. The clustering routine produces conformations within each hydrogen bond class (short, median, long), with their number proportional to the number of conformations found for each in the WASABI search. If one of the hydrogen bond lengths produces few possible conformations, then few, if any, with that hydrogen bond length may be found in the clustered set.

Database Search of Observed Interactions

We identified all hydrogen bonding interactions between amino acids and bases or base pairs in protein-DNA and protein-RNA complexes in the PDB (September 12, 2000 release, and the 30S ribosome (1jff); see Table 2). For this search, we slightly relaxed the hydrogen bonding parameters, using a donor angle of $0\pm 40^\circ$ and a maximum distance of 3.5\AA to ensure that no plausible interactions would be missed. Of the 295 protein-DNA complexes examined, 253 were crystal structures, 17 were averaged, energy-minimized NMR structures, and 25 were NMR ensembles. Of the 76 protein-RNA complexes examined, 54 were crystal structures, 6 were averaged, energy-minimized NMR structures, and 16 were NMR ensembles. Only crystal structures with $<3.5\text{\AA}$ resolution were used, and protons were added using InsightII (Biosym) or AMBER PROTONATE. Polymerases and topoisomerases were not examined, and for crystal structures with multiple complexes in a unit cell, only one representative was included. We made no

other attempts to remove other possible sources of redundancy, including similar structures solved by more than one group, similar structures reported at different levels of resolution or refinement, or mutant structures (one interaction was observed only in a mutant; see Results and Discussion). For the NMR structures, we scored the presence of an interaction if it was observed in the averaged structure or any member of an ensemble. Our goal for this study was to gather all the observed interactions rather than to compile precise statistics.

RESULTS AND DISCUSSION

To better understand the ways in which RNA sites might be recognized by proteins in a base-specific manner, we calculated “complete” databases of all possible hydrogen bonding interactions between amino acid side chains and either the six unpaired bases (A, C, G, U, A+, and C+) or the 53 possible RNA base pairs in planar conformations (Walberer et al., 2002). Although our focus is primarily on RNA, we also constructed databases that include thymine or the 17 possible thymine-containing base pairs that might be found in single-stranded DNA structures. A simple geometric algorithm (WASABI; Fig. 1A) was utilized in which a single hydrogen bond was first formed between a hydrogen-bonding donor or acceptor of a side chain moiety (Fig. 1C) and a complementary group on a base, followed by a systematic conformational search that identified additional possible hydrogen bonds in sterically plausible configurations. Each of five hydrogen bond angles (Fig. 1B) was varied in 4° steps such that no other donor or acceptor on any of the amino acid side chains would move by more than 0.6Å. The hydrogen bonding and steric parameters used were slightly beyond what would be considered energetically favorable to help ensure completeness of the databases. A single conformation was chosen to represent each unique hydrogen-bonding arrangement using a scoring function that attempted to maintain relatively planar geometries when possible (see Methods). One limitation to the WASABI algorithm is that the length of the initial hydrogen bond remains fixed during the conformational search, but its length subsequently may be varied to generate multiple conformations of any hydrogen-bonding arrangement (see below). The databases contain all possible amino acid-base and amino acid-base pair arrangements with one or more hydrogen bond (Table 1), but we focus primarily on interactions containing two or more bonds that have defined orientations and may contribute to high binding specificity.

In addition to the simple steric criteria applied by WASABI to the side chains and bases, we also wished to ensure that each hydrogen-bonding arrangement could accommodate the nucleic acid backbone and at least one reasonable conformation of a full amino acid side chain. We added C2'- and C3'-endo conformations of the ribose sugar and all amino acid rotamers (Dunbrack & Cohen, 1997) in a combinatorial manner to all arrangements and found that every interaction to the unpaired bases was acceptable whereas 170 interactions to the base pairs were sterically restricted. Given that the WASABI search utilized only planar conformations of the base pairs (Walberer et al., 2002) and output only a single conformation for each hydrogen-bonding arrangement, we wished to generate a larger set of plausible conformations for any individual arrangement and to re-examine their abilities to accommodate the backbone moieties. We constructed the DIVERSIGEN algorithm, which generates a specified number of output conformations for a single hydrogen-bonding arrangement (see Methods; Fig. 2), and found that only 12 arrangements to the base pairs remained sterically impossible when multiple conformations were tested (Table 1). DIVERSIGEN may be used to generate multiple conformations of amino acid interactions with bases or base pairs (Fig. 2A) or of interactions between bases (Fig. 2B).

The modeled interactions were constructed using the hydrogen-bonding moieties shown in Fig. 1C, assuming that interactions involving the carboxyl groups of Asp and Glu, the carboxamide groups of Asn and Gln, and the hydroxyl groups of Ser and Thr would be redundant. Indeed, we found that all arrangements could accommodate the extra lengths of the Glu, Gln, and Thr side chains. Initially we considered the hydroxyl-containing moiety of Tyr as separate from Ser(Thr), but subsequently found that all Ser(Thr) interactions could be represented by Tyr interactions despite the extra bulk of the Tyr ring. Interestingly, two Tyr-base pair arrangements cannot occur with Ser, both involving a bifurcated hydrogen bond to one base that is located close to the sugar of the second base (Fig. 3). All interactions with thymine were possible with uracil. Thus, the

databases were appropriately filtered for all types of redundant interactions, including those with His⁺ and pseudo-symmetric arrangements with Asn(Gln) and Arg moieties that produce structural redundancies (Table 1).

Amino Acid-Base Interactions

There are 344 unique ways for the amino acid hydrogen-bonding moieties to interact with the bases, including A⁺ and C⁺ (Table 1). Of these, 225 utilize single hydrogen bonds, representing all possible acceptor-donor combinations. Some of the interactions, even with one hydrogen bond, require that the amino acid be roughly perpendicular to the plane of the base because of sterics, particularly for His interactions to adenine N3. To focus the analysis, we removed bifurcated hydrogen-bonding interactions and those involving the protonated bases (A⁺, C⁺) that do not form hydrogen bonds to the extra proton (Table 1). Thus, there are 32 unique interactions involving two or more hydrogen bonds between amino acid side chains and the unpaired bases (shown in Fig. 4). For two of the Asn(Gln) interactions, we present arrangements that include bifurcated bonds that probably are more stable than the non-bifurcated versions in the database. Of the 32 possible interactions, 12 involve Asn(Gln) and 8 involve Ser(Thr/Tyr). Both types of side chains show potential interactions to all bases except A⁺, and their dominance likely reflects the high probability of pairing with complementary donor and acceptor groups on the bases, as is also true for base-base interactions (Walberer et al., 2002). Asp(Glu), with two acceptors, shows 5 interactions, including one to A⁺ not possible with the unprotonated base, and none to U. Arg, with five hydrogen donors on its guanidinium group, allows only 4 interactions, all to C and G.

To help evaluate the completeness of our database and to determine whether any rules might be inferred from known interactions, we identified amino acid-base hydrogen bonds in protein-nucleic acid complexes in the PDB (Table 2), using slightly relaxed hydrogen bond parameters (see Methods) to help ensure that no plausible interactions

would be missed. All observed interactions are found in our database, including 17 of the 32 possible two-hydrogen bonded arrangements (Table 3; Fig. 4). There are 12 types of interactions in DNA complexes, including 5 in the major groove and 2 in the minor groove of Watson-Crick helices, with the Arg-G and Asn(Gln)-A interactions (#26 and #13, Fig. 4) predicted by Seeman *et al.* (Seeman *et al.*, 1976) dominating, as previously observed (Luscombe *et al.*, 2001; Lustig & Jernigan, 1995; Mandel-Gutfreund *et al.*, 1995; Pabo & Sauer, 1992). Only 5 types of DNA interactions are found in which amino acids form two hydrogen bonds to a Watson-Crick face. Asp (or Glu)-C⁺ interactions (#29, Fig. 4) are observed in HhaI and HaeIII methylase complexes in which cytosines are extruded from the DNA helix (Klimasauskas *et al.*, 1994; Reinisch *et al.*, 1995), Asp-G interactions (#28, Fig. 4) are observed in two telomere-binding protein complexes (Ding *et al.*, 1999; Horvath *et al.*, 1998), a Lys-C interaction (#6, Fig. 4) is observed in a nucleocapsid-single-stranded DNA complex (Morellet *et al.*, 1998), Asn (or Gln)-A interactions (#16, Fig. 4) are observed in an RNaseB-DNA complex (Ko *et al.*, 1996), and a Ser-U interaction is observed in a reverse transcriptase complex (# 3, Fig. 4) (Najmudin *et al.*, 2000).

Despite the small database of protein-RNA complexes, the greater diversity of amino acid-base interactions already seems apparent. There are 13 types of interactions to RNAs, including 6 in which amino acids form two hydrogen bonds to a Watson-Crick face (Table 3). Of these, Gln-U, Arg-C, Ser-C, and Ser-C⁺ interactions (#4, #8, #12, #31; Fig. 4) have been observed only in RNA complexes, the Lys-C and Asp-C⁺ interactions (#6, #29) mentioned above only have been observed in DNA complexes, and the Asp-G interaction (#28) has been observed in both DNA and RNA complexes, including RNA complexes with TRAP and threonyl-tRNA synthetase (Antson *et al.*, 1999; Sankaranarayanan *et al.*, 1999). The Asp-G interaction in the TRAP complex is essential for binding (Elliott *et al.*, 1999), and also is observed in the binding of GTP by G proteins, where binding specificity can be switched to xanthine (XTP) by a compensatory

switch to the donor and acceptor arrangement of Asn (Powers & Walter, 1995). In addition to recognition of the Watson-Crick faces of the bases, some interactions to the major or minor groove faces are found in unpaired or non-Watson-Crick pairing contexts (Table 3), adding further to the diversity of interactions seen with RNAs. For recognition of RNA Watson-Crick pairs, the Arg-G interaction is the most common, as for DNA, but the Asn(Gln)-A interaction is not observed at all, as noted previously (Allers & Shamoo, 2001).

Five of the 32 interactions in our calculated database were found only because of the wide 38° donor angle used (see Fig. 4 legend). We suspect that these arrangements may not be energetically favorable, and none have yet been observed. To preliminarily assess whether our models are energetically reasonable, we calculated in vacuo interaction energies of the 28 arrangements involving the four unprotonated bases using quantum chemical methods and found that the five interactions near the edge of our parameter range were unstable (data not shown). Of the remaining 23 arrangements, all but Lys-G (#25) appear quite reasonable, with good hydrogen bond geometries and favorable interaction energies (A.C.C. and A.D.F., in preparation).

Amino Acid-Base Pair Interactions

One potentially attractive strategy to uniquely recognize portions of an RNA involves simultaneous hydrogen bonding to both partners of a non-Watson-Crick base pair. The Rev-RRE interaction appears to utilize such a strategy to recognize an unusual G:A base pair (Battiste *et al.*, 1996; Ye *et al.*, 1996). To systematically examine the possible amino acid interactions with base pairs, we constructed a database using the 53 possible RNA base pairs that are bridged by two or more hydrogen bonds (and 17 additional pairs that include thymine) (Walberer *et al.*, 2002). After removing bifurcated and redundant interactions, as for the unpaired bases, we identified 186 “spanning” interactions in which two or more hydrogen bonds bridge across each pair (Table 1).

Table 4 lists all interactions by the 53 RNA base pairs defined by Walberer et al. (Walberer et al., 2002). The database is dominated by interactions with Asn(Gln) (77 arrangements), as for the unpaired bases, but in contrast, interactions with Arg are common (64 arrangements) and interactions with Asp(Glu) are very rare (3 arrangements). Interestingly, very few Arg interactions are possible to the purine-purine base pairs (just 3 arrangements) but are common to the purine-pyrimidine and pyrimidine-pyrimidine pairs.

Of the 186 possible spanning interactions, nine potentially form three hydrogen bonds from Asn(Gln) or Arg side chains (Fig. 5). The four interactions involving Arg are to G:U wobble or reverse wobble base pairs (see below) whereas the five interactions involving Asn(Gln) are to four unusual base pairs, two G:G and two G:C⁺ pairs. These six base pairs are among the most commonly used for all spanning interactions (base pairs #8, 10, 18, 20, 31, 32; Table 4), reflecting the diversity of their donor and acceptor groups. Some of these pairs also are observed to form potential base triple interactions in which a third base, rather than an amino acid, is used to span the base pair (Walberer et al., 2002).

Eight types of spanning interactions have been observed (Fig. 6), including four to Watson-Crick pairs and one to a G:U wobble pair (discussed below), and five of the eight are in RNA complexes. In the crystal structure of a spliceosomal U2B''-U2A' protein complex to a U2 snRNA hairpin (Price *et al.*, 1998), Lys20 makes a spanning interaction to a U:U base pair located in the loop (Fig. 6). This U:U pair provides an important hinge that orients the loop relative to the stem and helps explain differences in binding specificity between the U2 complex and a related U1A-hairpin complex (Oubridge *et al.*, 1994). In NMR structures of an HIV Rev peptide bound to an RRE hairpin or to a related RNA aptamer (Battiste et al., 1996; Ye et al., 1996), the carboxamide of Asn40 hydrogen bonds to both bases of an important G:A base pair (Fig. 6). The position of Asn40 in the two Rev peptide-RNA complexes is well-defined by the NMR data, but the Asn-G:A

hydrogen bonding arrangements appear to differ (Fig. 6). It is not yet clear whether the difference in these spanning interactions reflects the slightly different RNA contexts in which the G:A pair is presented or inaccuracies in the structures. A tight RRE-binding peptide identified from a combinatorial library probably utilizes a Gln side chain, instead of Asn, in the context of a polyarginine framework to form a spanning interaction to the G:A pair (Tan & Frankel, 1998).

Our database of 186 spanning interactions contains only six of the eight observed cases, reflecting some limitations of our modeling approach. We did not identify one of the Rev-RRE Asn-G:A interactions (Fig. 6) because the G:A pair is especially nonplanar in the complex (Battiste *et al.*, 1996). Nevertheless, this spanning interaction was readily identified when we first used DIVERSIGEN to generate 10 conformations of the G:A pair (Fig. 2B) and then used WASABI to generate all possible amino acid hydrogen bonding interactions. Thus, subsequent databases may take into account an even wider three-dimensional structural diversity of base pairings. The second observed spanning interaction not present in our modeled database is seen in the crystal structure of EcoRV bound to its cognate GATATC DNA site (Winkler *et al.*, 1993). In this case, a Thr-A:T spanning interaction in the middle of the site places two polar hydrogens at a distance of 1.87Å (Fig. 6). WASABI eliminates such polar hydrogen “clashes” when the distance is less than 2.5Å, and the spanning interaction was identified by changing the parameter to 1.8Å. Subsequent crystal structures of EcoRV bound to the same GATATC site but with different flanking sequences indicate that Thr186 may make only one hydrogen bond to the O4 of T and that Asn185 may hydrogen bond to the N6 of the paired A (Horton & Perona, 1998a; Horton & Perona, 1998b; Kostrewa & Winkler, 1995; Perona & Martin, 1997).

Spanning Interactions to Watson-Crick and Wobble Base Pairs

Because Watson-Crick base pairs dominate in nucleic acid structures, followed by G:U wobble base pairs in RNAs (Masquida & Westhof, 2000; Varani & McClain, 2000), we examined their possible spanning interactions in more detail. We found four possible interactions to the two Watson-Crick base pairs (Fig. 7). Asn(Gln) can span either the major or minor groove of a G:C pair and the major groove of an A:U(T) pair, whereas Arg can span the minor groove of an A:U(T) pair. Two of these interactions have been observed, including an Asn-A:T major groove interaction in a c-Myb-DNA complex and Asn-G:C minor groove interactions in both EndoIV-DNA and Gln tRNA synthetase-tRNA complexes (Fig. 5) (Arnez & Steitz, 1996; Hosfield *et al.*, 1998; Ogata *et al.*, 1994). In the c-Myb complex (Ogata *et al.*, 1994), Asn183 hydrogen bonds to both partners of an A:T pair in one of 25 members of an NMR ensemble. While seemingly not well populated, the interaction is within the constraints of the experimental data and mutation of Asn183 to Ala severely reduces binding activity (Gabrielsen *et al.*, 1991). In the crystal structure of the EndoIV complex (Hosfield *et al.*, 1998), an Asn35 interaction to a G:C pair represents the only direct side chain-base hydrogen bonds in the complex. However, EndoIV is a DNA base excision repair endonuclease that recognizes abasic nucleotides within a protein pocket, so the role of a base-specific spanning interaction is unclear. In the crystal structure of a Gln tRNA synthetase mutant bound to its cognate tRNA (Arnez & Steitz, 1996), Asn235 hydrogen bonds to both bases of the G3:C70 base pair in the minor groove of the acceptor stem. Asn is able to make an additional hydrogen bond to the G:C base pair compared to the wild-type Asp side chain, consistent with a lowered K_M for the mutant enzyme corresponding to a gain in binding free energy of ~ 1.3 kcal/mol.

In addition to the observed spanning interactions with the Watson-Crick pairs, our studies suggest two other possible arrangements (Fig. 7). A spanning interaction of Asn(Gln) with a G:C pair in the major groove seems especially plausible given that the arrangement of donors and acceptors on a G:C pair are relatively symmetric in both the

major and minor grooves (Fig 7), and given the precedent of the minor groove interaction. However, it is unclear how well such an interaction would discriminate between base pairs because Asn(Gln) can similarly span the major groove of an A:U(T) pair (Fig. 7). In contrast, the Asn(Gln) minor groove spanning interaction, observed in the Gln tRNA synthetase and EndoIV structures, can uniquely distinguish the donor/acceptor arrangements among all base pairs, as can a possible spanning interaction of Arg in the A:T minor groove (Fig. 7).

The wobble G:U base pair is very common in RNA structures, and our database contains 16 possible spanning arrangements utilizing Arg, Lys, Asn(Gln), and Ser(Thr/Tyr) side chains (Fig. 8A). The Arg and Lys interactions can occur only in the major groove, the Ser(Thr/Tyr) interaction only in the minor groove, and the Asn(Gln) interactions in both grooves. Seven of the 11 Arg-G:U hydrogen bonding arrangements require a nonplanar orientation of the guanidinium group relative to the base pair (Fig. 8B). One spanning interaction between Lys and a G:U wobble pair has been observed in the NMR structure of L30 bound to a hairpin site in its mRNA (Mao *et al.*, 1999) (Fig. 6). In the ensemble of 21 structures, 18 show the spanning hydrogen bonding arrangement between Lys28 and G10:U60, with 31 intermolecular NOEs defining the position of Lys28. The Lys28 side chain also appears to make two additional hydrogen bonds to the surrounding RNA tertiary structure formed by this terminal G:U pair of a helix and an adjacent internal loop. This network of hydrogen bonding interactions may explain why substituting Lys28 with Ala reduces RNA-binding affinity by 20-30-fold (Mao *et al.*, 1999).

Two spanning arrangements of Arg to the G:U wobble pair involve three hydrogen bonds and appear particularly favorable (Fig. 8). In both cases, the three hydrogen bonds donated by the two different faces of the guanidinium group have reasonable geometries, and preliminary energy minimization and quantum chemical geometry calculations indicate that the proposed interactions are stable (data not shown).

Interestingly, a model of the *Drosophila* homolog of the U2B''-U2A'-snRNA complex described above places Arg52 in the major groove of a G:U wobble pair, where the Lys-U:U spanning interaction is found (Price et al., 1998). It will be interesting if an Arg-G:U spanning interaction is found in the *Drosophila* complex. Indirect experimental evidence for this type of interaction also is provided by the existence of a G-G:U base triple in tRNA^{Asp} (Westhof *et al.*, 1985). In this triple, part of the Watson-Crick face of G forms three hydrogen bonds to a G:U wobble pair, presenting three donors in an arrangement virtually identical to that of the guanidinium group (Fig. 9). Thus, the Arg-G:U and G-G:U interactions can be considered "pseudo-isomorphic". Our calculated databases of base triples contain many types of spanning interactions (Walberer et al., 2002), and two additional A+-G:U and C+-G:U arrangements are found that are pseudo-isomorphic to the proposed Arg-G:U interaction. The nearly equivalent arrangement of donors on the Arg guanidinium group and guanine base led to competition experiments that helped identify the guanosine-binding site in the Tetrahymena group I intron (Michel *et al.*, 1989; Yarus, 1988).

In principle, several side chains might be used to discriminate between a G:U wobble pair and the Watson-Crick pairs. From inspection of Table 4, Lys, Ser, or Arg are able to form spanning interactions to the wobble pair but not to the Watson-Crick pairs, whereas Asn can span both types. Thus, if Lys, Ser, or Arg were positioned between the bases of a pair, accurate discrimination might be possible. We favor Arg for this purpose given its potential to form the three hydrogen-bonded interaction described above.

Characteristics of Unpaired U Bases and Asp(Glu) Side Chains

Given our databases of amino acid interactions with unpaired bases and all possible base pairs, we identified those two hydrogen-bonded interactions possible only in an unpaired context. Such interactions are candidates for recognizing bases in bulges or loops exclusively via hydrogen bonding. Interestingly, every amino acid interaction to

U (or T) can form two hydrogen bonds to a base only in the absence of any type of base pairing (assuming two hydrogen bonds are required in a base pair). We rationalize this finding by noting that U (or T) possesses a total of only three donor and acceptor groups and thus cannot simultaneously form two hydrogen bonds to both another base and to an amino acid, nor can bifurcated bonds be made to the middle N3 donor group. Thus, U bases in RNA bulges and loops in principle could be specified uniquely by two hydrogen bonds, although the loss of hydrogen bonding to water may make such interactions unfavorable.

A previous analysis of protein-DNA complexes revealed that interactions with Asp and Glu are rarely observed, and it was suggested that this probably reflects unfavorable electrostatic interactions between the negatively charged carboxyl group and DNA backbone (Jones et al., 1999). Our databases of doubly hydrogen-bonded amino acid-base and amino acid-base pair interactions also reveals a rare occurrence of Asp(Glu) interactions that, in this case, reflects the limited number of adjacent donor group arrangements on the bases. For unpaired bases, 6 of the 32 possible arrangements involve Asp(Glu) (Fig. 4), but only 3 have favorable hydrogen bond geometries. Of these, two interactions are to the protonated bases (A⁺ and C⁺) and one is to G. Interestingly, Asp-C⁺ and Asp-G interactions already have been observed (see above) despite the involvement of the Watson-Crick face. For base pair spanning interactions, there are only 3 possible Asp(Glu) interactions, two to noncanonical G:C and C:A⁺ purine-pyrimidine pairs, one to a C:C⁺ pyrimidine-pyrimidine pair, and none to any purine-purine pair. It seems clear that, in addition to unfavorable electrostatic interactions, the arrangement of donors on the bases inherently disfavors hydrogen-bonded Asp(Glu) interactions. The rarity of hydrogen bonding possibilities for Asp(Glu) may present a good strategy for base-specific recognition and, indeed, two interactions to unpaired bases already have been observed, despite the small size of the RNA structural database.

Conclusions

We have systematically calculated “complete” databases of hydrogen bonding interactions between amino acids and unpaired bases and between amino acids and all possible base pairs. By examining interactions involving at least two hydrogen bonds, one can begin to see the types of interactions that may contribute to base-specific recognition of some of the unique elements of RNA structure. These doubly hydrogen-bonded interactions may be considered analogous to the types of interactions identified by Seeman et al. for recognition of DNA base pairs in Watson-Crick helices (Seeman et al., 1976). However, given the complexity of RNA structure and the diverse interactions possible, our databases clearly represent just small subsets of possible base-specific interactions, excluding, for example, interactions with water molecules, groups on the backbone, hydrophobic moieties, and others. In addition, some approximations in our calculations, such as the use of planar base pairs, place further limits on the completeness of the databases. Nevertheless, the databases contain virtually all observed interactions of this hydrogen bonding class, and we have identified several interesting new interactions that we expect ultimately will be observed as the number of structures of protein-RNA complexes grows. In addition, the databases provide starting points for designing novel sequence-specific RNA binding proteins or peptides whose interactions are guided largely by hydrogen bonding interactions, and also may be useful in constructing change-of-specificity mutants for known complexes.

We have placed the databases, named NAIL (Nucleic Acid Interaction Libraries), on a graphical web site (see <http://nail.ucsf.edu>) and have devised a set of filters that can be used to sort through the databases by criteria such as: number of hydrogen bonds, type of amino acid, and type of base or base pair (see (Walberer et al., 2002)).

ACKNOWLEDGEMENTS

We thank Peter Kollman, David Agard, Wendell Lim, and Steve Landt and others members of the Frankel lab for helpful discussions, James Robertson for help with quantum chemical calculations, Wei Wang for advice on energetic calculations, David Konerding for computational advice, and Steve Landt, Valerie Calabro, Steve Edgcomb, and Chandreyee Das for comments on the manuscript. We thank the Computer Graphics Laboratory (UCSF) for use of computing resources. This work is supported by NIH grants GM47478 and GM56531 (A.D.F.) and NIH Training grants GM08284 and GM08388 (A.C.C.).

REFERENCES

Allers, J. & Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.* **311**, 75-86.

Antson, A. A., Dodson, E. J., Dodson, G., Greaves, R. B., Chen, X. & Gollnick, P. (1999). Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **401**, 235-242.

Arnez, J. G. & Steitz, T. A. (1996). Crystal structures of three misacylating mutants of *Escherichia coli* glutamyl-tRNA synthetase complexed with tRNA(Gln) and ATP. *Biochemistry* **35**(47), 14725-14733.

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97-179.

Battiste, J. L., Mao, H., Rao, N. S., Tan, R., Muhandiram, D. R., Kay, L. E., Frankel, A. D. & Williamson, J. R. (1996). Alpha helix major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science* **273**, 1547-1551.

Choo, Y. & Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Curr. Op. Struct. Biol.* **7**, 117-125.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**(19), 5179-5197.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. 2nd edit, W. H. Freeman and Co., New York.

Ding, J., Hayashi, M. K., Zhang, Y., Manche, L., Krainer, A. R. & Xu, R. M. (1999). Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev.* **13**, 1102-1115.

Draper, D. E. (1999). Themes in RNA-protein recognition. *J. Mol. Biol.* **293**, 255-270.

Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical analysis of protein sidechain rotamer preferences. *Prot. Sci.* **6**, 1661-1681.

Elliott, M. B., Gottlieb, P. A. & Gollnick, P. (1999). Probing the TRAP-RNA interaction with nucleoside analogs. *RNA* **5**, 1277-1289.

Gabrielsen, O. S., Sentenac, A. & Fromageot, P. (1991). Specific DNA binding by c-Myb: evidence for a double helix-turn-helix-related motif. *Science* **253**(5024), 1140-3.

Hermann, T. & Patel, D. J. (1999). Stitching together RNA tertiary architectures. *J. Mol. Biol.* **294**, 829-849.

Horton, N. C. & Perona, J. J. (1998a). Recognition of flanking DNA sequences by EcoRV endonuclease involves alternative patterns of water-mediated contacts. *J. Biol. Chem.* **273**, 21721-21729.

Horton, N. C. & Perona, J. J. (1998b). Role of protein-induced bending in the specificity of DNA recognition: crystal structure of EcoRV endonuclease complexed with d(AAAGAT) + d(ATCTT). *J. Mol. Biol.* **277**, 779-787.

Horvath, M. P., Schweiker, V. L., Bevilacqua, J. M., Ruggles, J. A. & Schultz, S. C. (1998). Crystal structure of the *Oxytricha nova* telomere and binding protein complexed with single strand DNA. *Cell* **95**, 963-974.

Hosfield, D. J., Mol, C. D., Shen, B. & Tainer, J. A. (1998). Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity. *Cell* **95**, 135-146.

Israelachvili, J. N. (1989). *Intermolecular and Surface Forces*, Academic Press.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in biological molecules*, Springer, Berlin, Heidelberg.

Jones, S., Daley, D. T. A., Luscombe, N. M., Berman, H. & Thornton, J. M. (2001). Protein-RNA interactions: a structural analysis. *Nucl. Acids Res.* **29**, 943-954.

Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein-DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877-896.

Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. *Cell* **76**(2), 357-369.

Ko, T. P., Williams, R. & McPherson, A. (1996). Structure of a ribonuclease B+d(pA)₄ complex. *Acta Cryst.* **D52**, 160-164.

Kono, H. & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **35**(1), 114-131.

Kostrewa, D. & Winkler, F. K. (1995). Mg²⁺ binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution. *Biochemistry* **34**(2), 683-696.

Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860-2874.

Lustig, B. & Jernigan, R. L. (1995). Consistencies of individual DNA base-amino acid interactions in structures and sequences. *Nucl. Acids Res.* **23**, 4707-4711.

Mandel-Gutfreund, Y. & Margalit, H. (1998). Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucl. Acids Res.* **26**, 2306-2312.

Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *Journal of Molecular Biology* **253**(2), 370-382.

Mao, H., White, S. A. & Williamson, J. R. (1999). A novel loop-loop recognition motif in the yeast ribosomal protein L30 autoregulatory RNA complex. *Nat. Struct. Biol.* **6**, 1139-1147.

Masquida, B. & Westhof, E. (2000). On the wobble GoU and related pairs. *RNA* **6**, 9-15.

Michel, F., Hanna, M., Green, R., Bartel, D. P. & Szostak, J. W. (1989). The guanosine binding site of the Tetrahymena ribozyme. *Nature* **342**(6248), 391-395.

Morellet, N., Demene, H., Teilleux, V., Huynh-Dinh, T., de Rocquigny, H., Fournie-Zaluski, M. C. & Rocques, B. P. (1998). Structure of the complex between the HIV-1 nucleocapsid protein NCp7 and the single-stranded pentanucleotide d(ACGCC). *J. Mol. Biol.* **283**, 419-434.

Najmudin, S., Cote, M. L., Sun, D., Yohannan, S., Montano, S. P., Gu, J. & Georgiadis, M. M. (2000). Crystal structures of an N-terminal fragment from Moloney murine leukemia virus reverse transcriptase complexed with nucleic acid: functional implications for template-primer binding to the fingers domain. *J. Mol. Biol.* **296**, 613-632.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. & Nishimura, Y. (1994). Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79**(4), 639-648.

Oubridge, C., Ito, N., Evans, P. R., Teo, C. H. & Nagai, K. (1994). Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**(6505), 432-438.

Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597-624.

Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* **61**(1053), 1053-95.

Perona, J. & Martin, A. (1997). Conformational transitions and structural deformability of EcoRV endonuclease revealed by crystallographic analysis. *Journal of Molecular Biology* **273**(1), 207-225.

Powers, T. & Walter, P. (1995). Reciprocal stimulation of GTP hydrolysis by two directly interacting GTPases. *Science* **269**, 1422-1424.

Price, S. R., Evans, P. R. & Nagai, K. (1998). Crystal structure of the spliceosomal U2B'-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394**(6694), 645-650.

Reinisch, K. M., Chen, L., Verdine, G. L. & Lipscomb, W. N. (1995). The crystal structure of HaeIII methyltransferase covalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. *Cell* **82**, 143-153.

Saenger, W. (1984). *Principles of nucleic acid structure*, Springer-Verlag, New York.

Sankaranarayanan, R., Dock-Bregeon, A.-C., Romby, P., Caillet, J., Springer, M., Rees, B., Ehresmann, C., Ehresmann, B. & Moras, D. (1999). The structure of Threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* **97**, 371-381.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA* **73**, 804-808.

Steitz, T. A. (1999). RNA recognition by proteins. In *The RNA World, 2nd Ed.* (Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds.), pp. 427-450. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Suzuki, M. (1994). A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure* **2**, 317-326.

Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl. Acad. Sci. USA* **91**, 12357-12361.

Tan, R. & Frankel, A. D. (1998). A novel glutamine-RNA interaction identified by screening libraries in mammalian cells. *Proc. Natl. Acad. Sci. USA* **95**(8), 4247-4252.

Taylor, R. & Kennard, O. (1984). Hydrogen-Bond Geometry in Organic Crystals. *Acc. Chem Res.* **17**, 320-326.

Taylor, R., Kennard, O. & Versichel, W. (1983). Geometry of the N-H--O=C Hydrogen Bond. *J. Am. Chem Soc.* **105**, 5761-5766 **105**, 5761-5766.

Treger, M. & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recog.* **14**, 199-214.

Varani, G. & McClain, W. H. (2000). The G:U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse systems. *EMBO Rep.* **1**(18-23).

Walberer, B. J. (2000). Construction and analysis of a complete database of hydrogen-bonded base combinations, University of California, San Francisco.

Walberer, B. J., Cheng, A. C. & Frankel, A. D. (2002). Diversity of hydrogen-bonded base interactions in nucleic acids. , submitted.

Westhof, E., Dumas, P. & Moras, D. (1985). Crystallographic refinement of yeast aspartic acid transfer RNA. *J. Mol. Biol.* **184**(1), 119-145.

Winkler, F., Banner, D., Oefner, C., Tsernoglou, D., Brown, R., Heathman, S., Bryan, R., Martin, P., Petratos, K. & Wilson, K. (1993). The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO Journal* **12**(5), 1781-95.

Yarus, M. (1988). A specific amino acid binding site composed of RNA. *Science* **240**(4860), 1751-1758.

Ye, X., Gorin, A., Ellington, A. D. & Patel, D. J. (1996). Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex. *Nat. Struct. Biol.* **3**, 1026-1033.

TABLES AND FIGURE LEGENDS

Table 1. Numbers of computed amino acid-base and amino acid-base pair interactions. The raw output from WASABI was filtered to remove sterically restricted conformations, calculational and structural redundancies, and bifurcated and single hydrogen-bonded interactions, as described in the text. Bifurcated interactions refer only to those in which a donor or acceptor atom on the amino acid is simultaneously involved in two hydrogen bonds to a base and does not include those in which a bifurcated bond exists between two bases of a base pair. The narrow donor angle parameter used to construct the base pairs limits the number of bifurcated bonds (Walberer et al., 2002).

Table 2. Nucleic acid-protein complexes from the PDB utilized in this study.

Table 3. Observed amino acid-base interactions. Numbers refer to the interactions shown in Fig. 4. Face refers to the interacting surface of the base (Watson-Crick, major groove, minor groove) as located in a Watson-Crick helix. The observed cases in DNA and RNA are indicated, with details (protein name, pdb identifier, residue) provided for bases that are not in a Watson-Crick pair.

Table 4. Calculated amino acid-base pair spanning interactions. Base pairs are listed according to the numbering of Walberer et al. (Walberer et al., 2002). For cases in which the protonated base atom did not form an additional hydrogen bond in the pair, the number of the corresponding unprotonated pair is followed by +. The A+C pair (12+) is the only case with interactions not observed with the unprotonated partner. The arrangements marked with brackets indicate base arrangements in which a U is flipped, presenting essentially an identical donor and acceptor arrangement. The interactions observed with the related pairs are identical in all but one case.

Figure 1. (A) Schematic of the WASABI search method. (B) The three parameters used to define a hydrogen bond: the acceptor angle, the donor angle, and distance between heavy atoms, with parameters listed for the different atom types. (C) The nine hydrogen-bonding moieties of the amino acid side chains, with arrows indicating donor and acceptor positions.

Figure 2. Generation of conformational diversity by DIVERSIGEN. (A) Ten representative conformations of an Asn-C:C+ base pair interaction. The starting model calculated by WASABI, shown in yellow, cannot accommodate the sugar backbone and full amino acid side chain, the sterically allowed conformations are shown in blue, and clashing models are shown in gray. (B) Ten conformations of the G:A base pair found in the RRE (Battiste et al., 1996; Ye et al., 1996), beginning with the initial planar conformation shown in yellow. The blue conformations represent those in which Asn can form spanning interactions, as described in the text, whereas the starting yellow conformation and gray conformations can not.

Figure 3. Interaction of Ser and Tyr with a G:G base pair, demonstrating a steric clash with Ser (yellow) but not Tyr (red). The bifurcated hydrogen bonds are indicated.

Figure 4. Amino acid-base interactions with two or more hydrogen bonds. Interactions are grouped by base type, and are subgrouped by the interacting face. Bifurcated versions of two interactions are shown (#10, #22) that probably are more stable than the related two hydrogen-bonded versions also present in the database. In these cases, the side chain must be placed considerably out of the plane of the base to avoid forming the additional bifurcated interaction. Five interactions (#5, #7, #9, #15, #23) only are observed at the

edge of our parameter range (requiring a donor angle of 38°) and may not be energetically reasonable.

Figure 5. Spanning interactions utilizing three hydrogen bonds. The base pair numbering is that of Walberer et al. (Walberer et al., 2002). The top set show interactions with wobble-type arrangements and the bottom set shows three interactions of Asn(Gln). Two symmetric interactions are found between Asn(Gln) and base pair 18.

Figure 6. Observed spanning interactions. The top set shows interactions to non-Watson-Crick base pairs in RNAs, and the bottom set shows interactions to Watson-Crick pairs both in DNA and RNA. PDB identifiers are shown in parentheses, and references are provided in the text.

Figure 7. Possible and observed spanning interactions to the Watson-Crick base pairs. The Asn-A:T major groove interaction has been observed in a c-Myb-DNA complex (Ogata et al., 1994) and Asn-G:C minor groove interactions have been observed in EndoIV-DNA (Hosfield et al., 1998) and Gln tRNA synthetase-tRNA (Arnez & Steitz, 1996) complexes.

Figure 8. Possible spanning interactions to the G:U wobble pair. (A) Interactions in which the side chains are nearly coplanar with the base pair and (B) interactions that require nonplanar orientations.

Figure 9. Similarity of a G-G:U base triple and a modeled Arg-G:U wobble interaction. The base triple has previously been observed in tRNA^{Asp} (Westhof et al., 1985).

Table 1. Numbers of computed amino acid-base and amino acid-base pair interactions

	Single Bases	Base Pairs
WASABI output	470	7730
remove backbone incompatible	470	7718
remove U/T redundancies	426	5819
remove Asn/Gln, Arg, His+ redundancies	385	5076
remove Tyr redundancies	344	4612
remove bifurcated interactions	261	3519
remove single H-bond interactions	36	457
remove A+, C+ redundancies	32	423
remove non-spanning base pair interactions	N/A	186

Table 2. Nucleic acid-protein complexes from the PDB utilized in this study

DNA-protein complexes						RNA-protein complexes	
NMR	X-ray					NMR	X-ray
185d	1a02	1c7y	1hwt	1ruo	2pud	1a1t	1a34
193d	1a0a	1c9b	1if1	1rv5	2pue	1a4t	1a9n
1a66	1a1f	1c9r	1ign	1rva	2puf	1aju	1aq3
1a6b	1a1g	1ca5	1ihf	1rvb	2pug	1akx	1aq4
1ahd	1a1h	1ca6	1ijs	1rvc	2pvi	1arj	1asy
1b69	1a1i	1cbv	1ipp	1skn	2ram	1aud	1asz
1bbx	1a1j	1cdw	1jmc	1srs	2rve	1biv	1av6
1bj6	1a1k	1cez	1lat	1ssp	2ssp	1ck5	1b23
1c7u	1a1l	1cf6	1lau	1svc	2up1	1ck8	1b7f
1c1g	1a3q	1cgp	1lli	1t7p	3bam	1cn8	1c0a
1dsc	1a6y	1cit	1lmb	1tau	3cro	1cn9	1c9s
1dsd	1a73	1ckq	1mdy	1tc3	3crx	1d6k	1cvj
1fja	1a74	1cl8	1mey	1tf6	3hdd	1dz5	1cwp
1gcc	1aay	1clq	1mhd	1tgh	3hts	1ekz	1cx0
1hry	1ais	1cma	1mht	1tro	3mht	1etf	1d9f
1hrz	1akh	1cqt	1mj2	1trr	3orc	1etg	1dfu
1lcc	1am9	1crx	1mjm	1tsr	3pjr	1exy	1di2
1lcd	1an2	1cw0	1mjo	1tup	3pvi	1koc	1dk1
1mse	1an4	1cyq	1mjp	1uaa	4crx	1mnb	1drz
1msf	1aoi	1cz0	1mjq	1ubd	4dpv	1qfq	1dul
1nk2	1apl	1d02	1mnm	1vas	4mht	1ull	1dzs
1nk3	1au7	1d0e	1mvm	1vix	4rve	484d	1ec6
1rcs	1awc	1d1u	1nfk	1vol	4skn		1efw
1tf3	1az0	1d2i	1noy	1vpw	5crx		1eiy
1tn9	1azp	1d3u	1oct	1wet	5mht		1ekc
1yui	1azq	1d5y	1otc	1xbr	6cro		1euq
1yuj	1b01	1d66	1par	1yrn	6mht		1euy
2da8	1b3t	1dct	1pdn	1ysa	6pax		1exd
2ezd	1b72	1ddn	1per	1ytb	7mht		1ffy
2eze	1b8i	1dfm	1pnr	1ytf	8mht		1fjf
2ezf	1b97	1dgc	1pue	1zaa	9ant		1gtr
2ezg	1bc7	1diz	1pvi	1zay	9mht		1gts
2gat	1bc8	1dnk	1pyi	1zme			1mms
2hdc	1bdh	1dp7	1qai	2bam			1qa6
2lef	1bdi	1dsz	1qaj	2bop			1qf6
2stt	1bdt	1du0	1qbj	2bpf			1qln
2stw	1bdv	1ecr	1qp0	2cgp			1qrs
3gat	1ber	1ej9	1qp4	2crx			1qrt
4gat	1bf4	1eqz	1qp7	2dgc			1qru
5gat	1bf5	1eri	1qpi	2dnj			1qtq
6gat	1bg1	1eww	1qps	2drp			1qu2
7gat	1bgb	1eyu	1qpz	2gli			1qu3
	1bhm	1f3i	1qqa	2hap			1ser
	1bi0	1fjl	1qqb	2hdd			1ttt
	1bnk	1flo	1qrh	2hmi			1urn
	1bnz	1fok	1qri	2irf			1zdj
	1bp7	1fos	1qrv	2kfn			1zdi
	1bpx	1gdt	1qsl	2kfz			1zdz
	1bpy	1glu	1qum	2kzm			1zdk
	1bpz	1hao	1ram	2kzz			2a8v
	1bsu	1hap	1rbj	2nll			2abbv
	1bua	1hcq	1rcn	2pjr			2fnt
	1bvo	1hcr	1rep	2pua			5msf
	1c0w	1hdd	1rtd	2pub			6msf
	1hlo	1hut	1run	2puc			7msf

Table 3. Observed amino acid-base interactions

Interaction	Face	Number (DNA)	Number (RNA)	Observed interactions
Ser/Thr/Tyr-U (#3)	WC	1	1	DNA: reverse transcriptase (1d0e) Ser67 RNA: Sxl (1b7f) Tyr164
Asn/Gln-U (#4)	WC	0	5	RNA: AspRS (1asy) Gln138, (1asz) Gln138, (1c0a) Gln46, (1efw) Gln47; GlnRS (1euq) Gln517
Lys-C (#6)	WC	1	0	DNA: nucleocapsid (1bj6) Lys34
Arg-C (#8)	WC	0	1	RNA: S15/S16/S18 (1ekc) Arg74
Ser/Thr/Lys-C (#12)	WC	0	1	RNA: AspRS (1asz) Ser329
Asn/Gln-A (#13)	Major	67	0	
Ser/Thr/Tyr-A (#14)	Major	7	7	RNA: U1A (1aud) Tyr12, (1dz5) Ser45, Thr88, Tyr12; MS2 coat (5msf) Thr45, (6msf) Thr45, (7msf) Thr45
Asn/Gln-A (#16)	WC	2	0	DNA: RNaseB (1rbj) Gln69, Asn71
Ser/Thr/Tyr-A (#17)	Major	0	1	RNA: U2B'/A' (1a9n) Ser91
Asn/Gln-G (#18)	Minor	4	2	DNA: telomere BP (1otc) Gln135
Ser/Thr/Tyr-G (#19)	Minor	2	1	
Lys-G (#25)	Major	18	3	RNA: L30 (1ck8, 1cn9) Lys28
Arg-G (#26)	Major	142	16	DNA: telomere BP (1otc) Arg274 RNA: AspRS (1c0a) Arg222; Rev (1ull) Arg6, (484d) Arg41
Arg-G (#27)	Major	3	2	
Asp/Glu-G (#28)	WC	3	3	DNA: telomere BP (1otc) Asp225, Glu45; UP1 (2up1) Asp42 RNA: TRAP (1c9s) Asp39, Glu36; ThrRS (1qf6) Glu600
Asp/Glu-C+ (#29)	WC	3	0	DNA: HaeIII (1dct) Glu109; HhaI (1mht) Glu119, (4mht) Glu119
Ser/Thr/Tyr-C+ (#31)	WC	0	2	RNA: U1A (1aud) Tyr12, (1dz5) Tyr12

Table 4. Calculated amino acid-base pair spanning interactions

Pur-Pur		Total	D/E	H	H+	K	N/Q	R	S/T/Y
AA 1									
AA 15		1				1			
AA 16									
AA+ 1+									
AA+ 16+									
A+A+16+									
GA 21									
GA 22		2					2		
GA 23		2					2		
GA 24									
GA+ 21+									
GA+ 23+									
GA+ 38		1				1			
GA+ 39									
GG 17		2				1		1	
GG 18		13		2		8	1	2	
GG 19									
GG 20		10		1	1	5	2	1	
Total	31	0	3	0	1	20	3	4	

Pur-Pyr		Total	D/E	H	H+	K	N/Q	R	S/T/Y
AC 12		2					1		1
AC 13		5			1		1	2	1
AC+ 35		1					1		
AC+ 37		2					2		
A+C 12+		2					2		
A+C 34		3	1	1			1		
A+C 36		2					2		
A+C+35+									
[AU 1		2					1	1	
[AU 3		4					1	3	
[AU 2		1					1		
[AU 4		1					1		
[A+U 2+									
[A+U 4+									
GC 5		2					2		
GC 6		5	1			1	1	2	
GC 7		4					3		1
GC+ 31		9		1			5	1	2
GC+ 32		8		1		1	4	2	
GC+ 33		1					1		
[GU 8		16			1	1	2	11	1
[GU 10		16			1	1	2	11	1
[GU 9		1					1		
[GU 11		1					1		
Total	88	2	3	3	4	36	33	7	

Pyr-Pyr		Total	D/E	H	H+	K	N/Q	R	S/T/Y
CC 25		4					2		2
CC+ 40		5	1			1	1	2	
CC+ 41		2					2		
C+C+42		5					2	1	2
[UC 29		5				1	2	2	
[UC 30		5				1	2	2	
[UC+ 43		7				1	2	3	1
[UC+ 44		7				1	2	3	1
[UU 26		9				2	2	5	
[UU 27		9				2	2	5	
[UU 28		9				2	2	5	
Total	67	1	0	0	11	21	28	6	

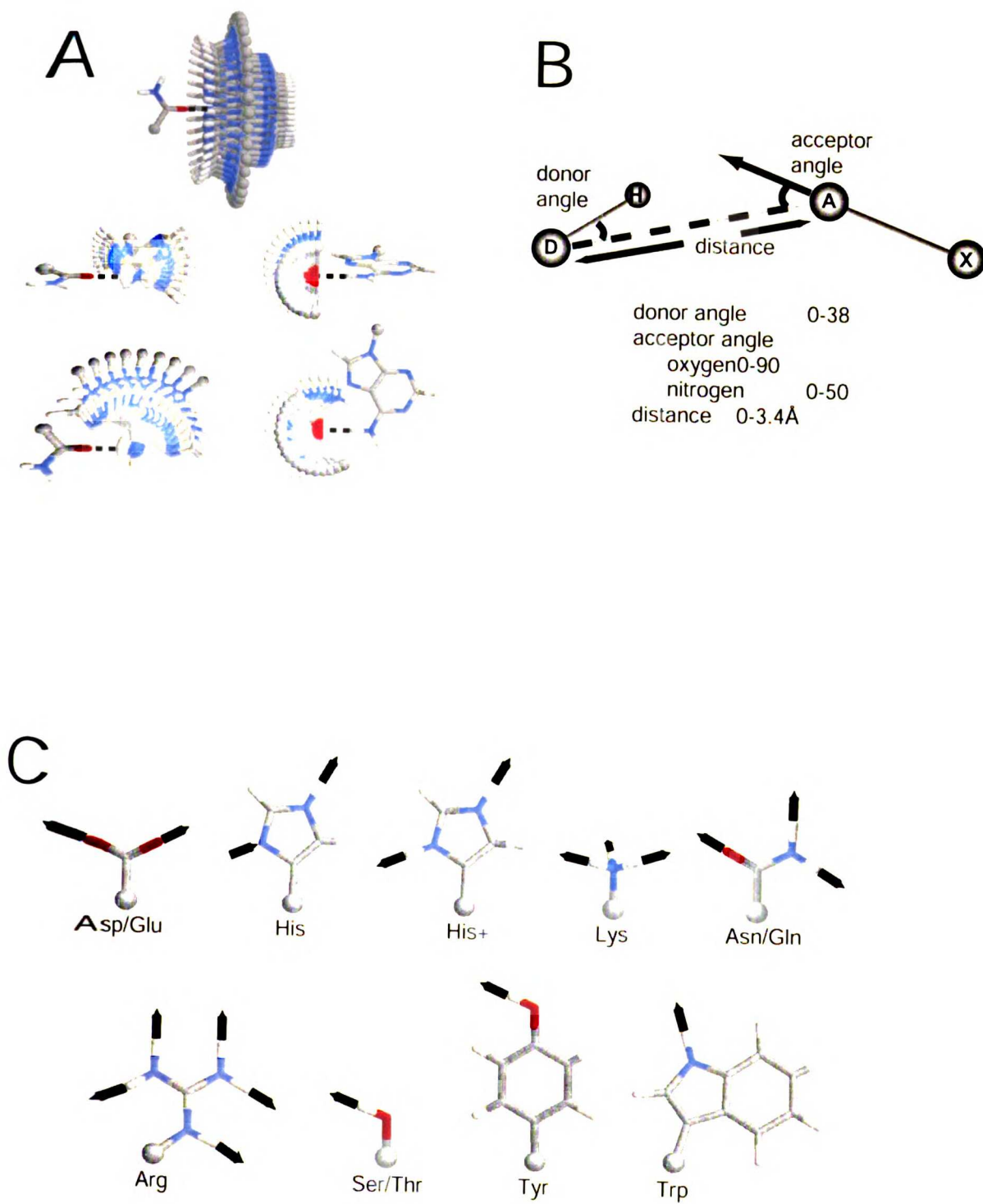


Figure 2.1

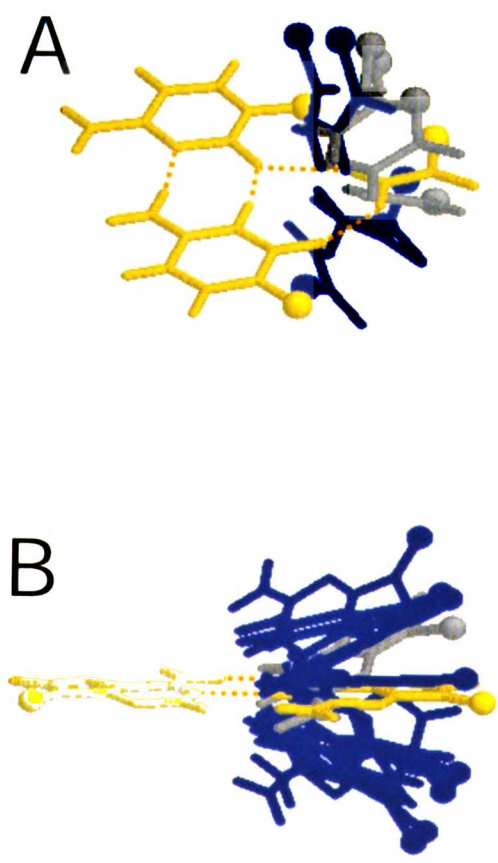


Figure 2.2

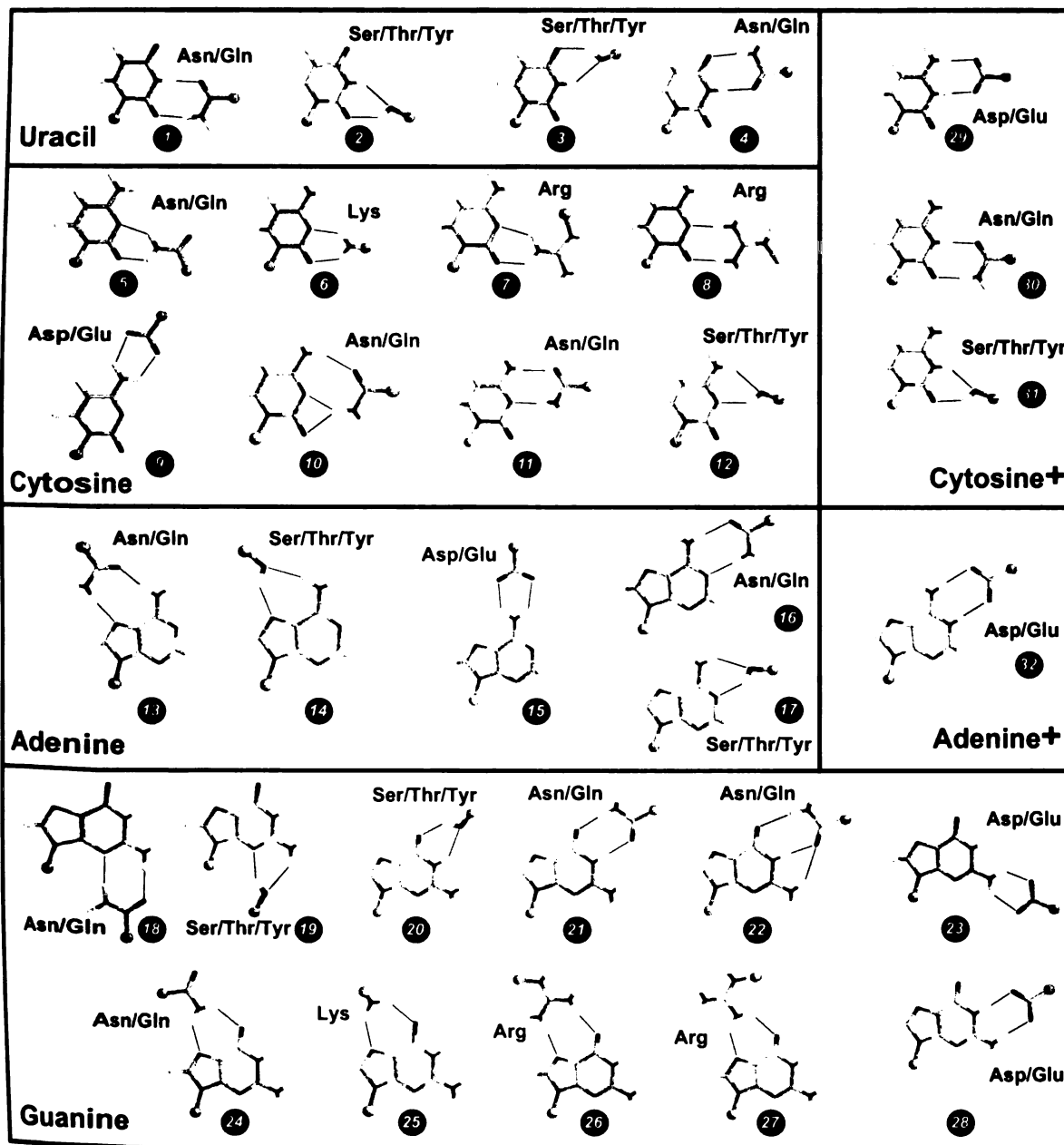


Figure 2.4

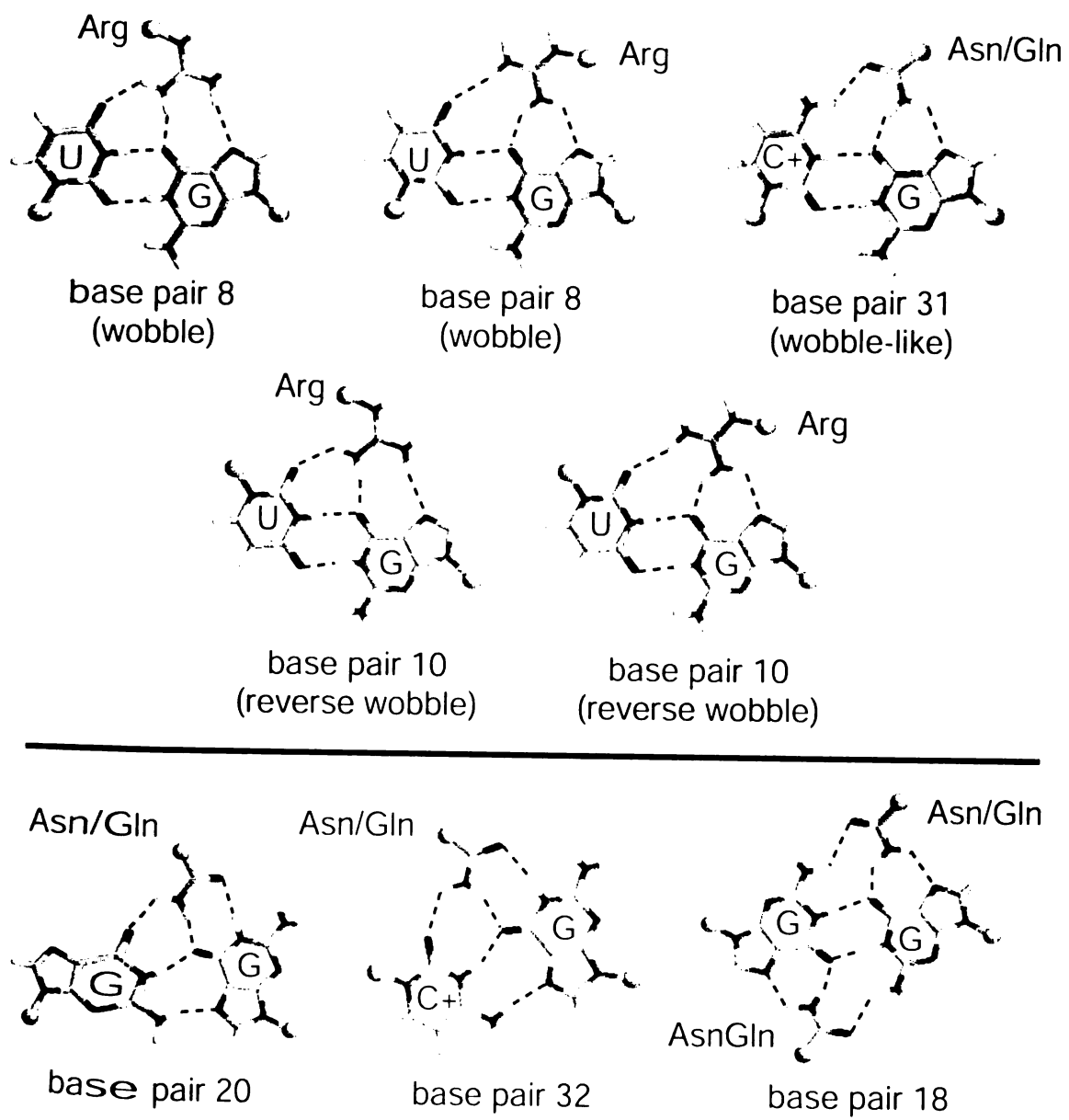


Figure 2.5

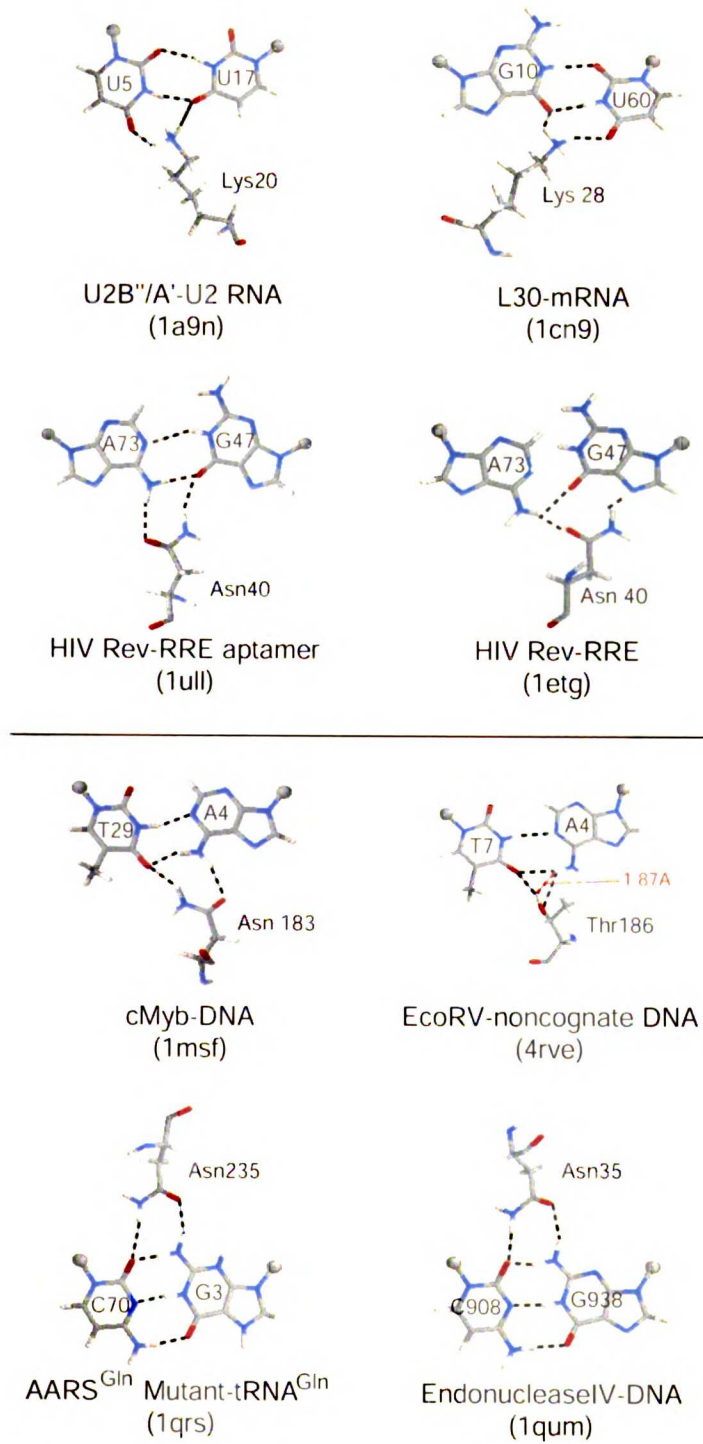


Figure 2.6

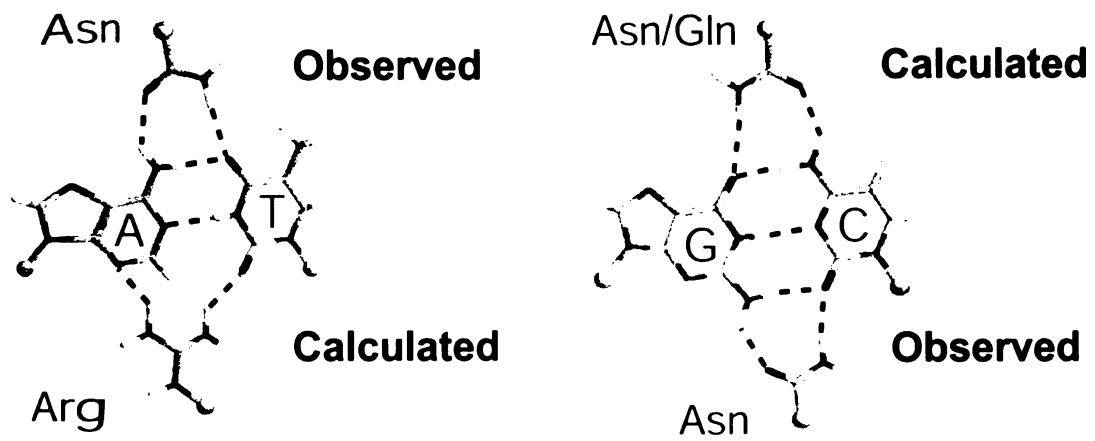


Figure 2.7

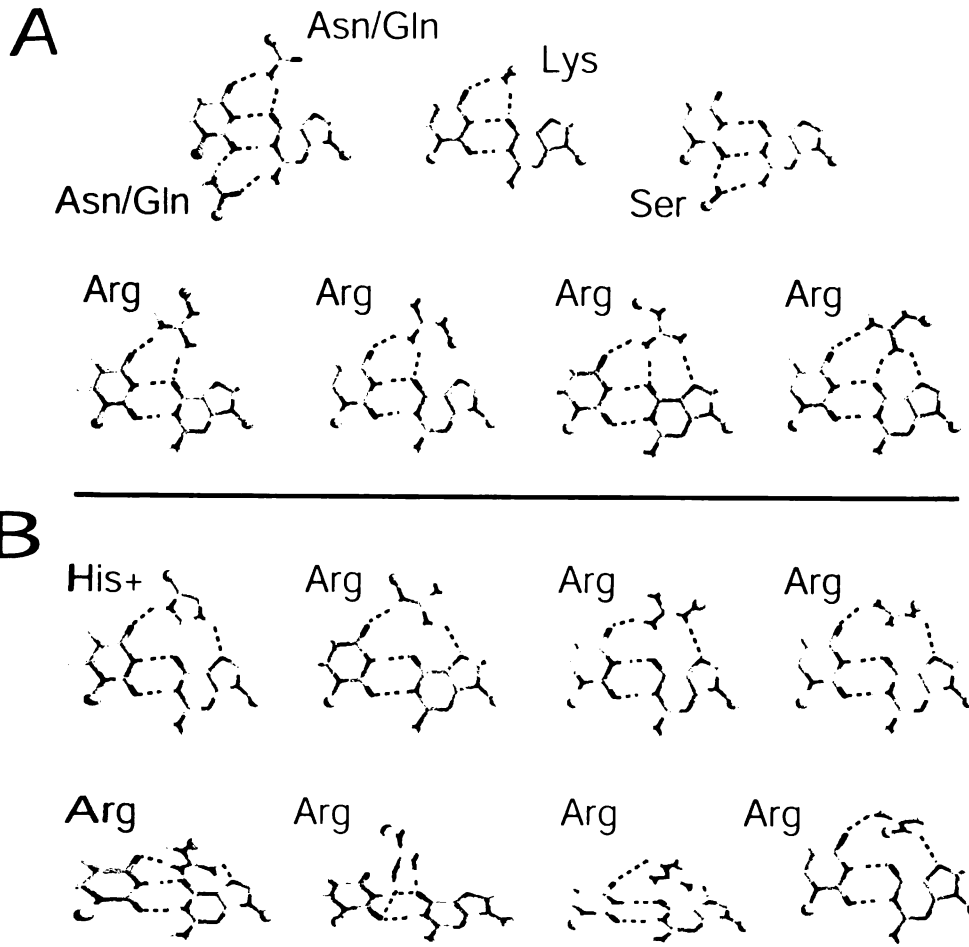


Figure 2.8



Figure 2.9

Chapter 3

**Ab-initio interaction energies of amino acid side-chain
analogues hydrogen bonding with nucleic acid bases,
and correlations with observed frequencies**

Alan C. Cheng and Alan D. Frankel

Graduate Group in Biophysics and Department of Biochemistry and Biophysics,

University of California San Francisco, San Francisco, CA 94143-0448

Abstract

We have performed a first ab-initio study of the interaction of hydrogen-bonding protein sidechain analogues with nucleic acid bases. We have generated models of all bidentate hydrogen bonding interactions with nucleic acid bases, and then used ab-initio quantum chemical calculations to rank-order the models. The rank ordering of the possible interaction models provides insights into the energetics of the possible interactions. We look at the role bifurcated hydrogen bonds play in the conformation of Asn interactions to the Watson-Crick face of Guanine and Cytosine. In investigating the observed occurrences of Asn and Ser/Thr/Tyr interactions to the Hoogsteen face of Adenine, we find that although Asn and Ser/Thr/Tyr have nearly identical interaction energies with Adenine, the Asn-A interactions are much more common in DNA, and the Ser/Thr/Tyr-A interactions appear to be more common in RNA. A look at occurrences of these two interactions in the Protein Databank suggests the Ser/Thr/Tyr interactions are correlated with beta-strand and turn structures. Our calculations of the intrinsic stability of discrete models of specific hydrogen-bonded is an important step in beginning to understand in detail the contributors to specificity in protein-RNA interactions.

Introduction

Hydrogen bonding is an important contributor to specificity in protein-nucleic interactions. Seeman et al. first postulated in 1976 [1] that sequence specific recognition of DNA by proteins could be achieved via intermolecular bidentate hydrogen bonds between side chains and bases in the DNA major groove. Their prediction of two such interactions, Arg-G and Asn/Gln-A (models #21 and #10 in Fig 1), has since been strongly supported by experimental data [2]. In RNA, bases are not limited to Watson-Crick pairing as in DNA, and can be unpaired or form non-canonical base pairs [3]. These non-canonical features are often key specific recognition features of functional RNAs. Keeping in mind that an important contributor to specificity is interaction energy, we try to answer three questions related to specific recognition in RNAs. What are the possible ways to recognize RNA bases with hydrogen bonds? Of the variety of ways to target non-canonical elements, which ones have the most favorable interaction energy? And finally, how do the interaction energies correspond to observed occurrences in the PDB?

We have previously generated 28 possible bidentate hydrogen-bonding patterns between protein side-chains and the four unprotonated RNA bases (A,G,C,U) using an exhaustive geometric search [1c]. In this study we use an ab-initio quantum chemical approach to evaluate the interaction energy of these interactions, and correlate the rankings to the observed interaction statistics in experimentally determined structures found in the PDB [1][4]. Previous ab-initio studies have looked at hydrogen bonding between nucleic acids and between amino acids [5], however, we are not aware of previous studies looking at protein-nucleic interactions.

Ab-initio computational methods based on quantum mechanics have been highly successful in calculating molecular properties such as interaction energies with high accuracy and generality. Hartree-Fock (HF) theory can be used to calculate such properties with increasing accuracy as larger basis sets that more accurately model molecular orbitals are used (e.g., 3-21G, 6-31G, 6-31G**). However Hartree-Fock does not adequately represent electron correlation. A perturbational theory called second-order Moller-Plesset perturbation theory (MP2) can be used to account for such effects, and thus generally provides more accurate intermolecular interaction energies. There are more accurate theories, as mentioned below, but the tradeoff for increasing accuracy with HF and MP2 methods is exponentially increasing computation time.

The development of methods such as localized MP2 (LMP2) [6] and current workstation performance allow us to perform ab-initio calculations on our ~30 atom models. A study of hydrogen-bonded formamides, formamidines and DNA bases [5b] concluded that MP2/6-31G** energies underestimate the stabilization energy by only 0.2-1.3kcal/mol compared to aug-cc-pVDZ, and a very large aug-cc-pVQZ basis is estimated in the study to bring an additional 0.3 kcal/mol of stabilization.. Because we look at interaction energies ranging from 10-50 kcal/mol, a medium LMP2/6-31G**//HF/6-31G** [7] calculation should be usefully accurate for rank-ordering our models.

Method

To generate all possible hydrogen bonded patterns between amino acids and RNA bases, we wrote a program called WASABI (What Are the Specific Amino-acid Base Interactions) [2c] that forms an initial hydrogen bond, and then exhaustively generates between 25 and 80 million conformations within a hydrogen-bonding space defined by small crystal studies [12]. The hydrogen bonding space is defined using a distance range and two angles, the hydrogen-donor-acceptor angle (donor angle) and the donor-acceptor-acceptor direction angle (acceptor angle). The acceptor angle was allowed to vary between 0° and 50° for nitrogens, and between 0° and 90° for oxygens. The donor angle was allowed to vary between 0° and 38°, and the distance was allowed to vary between 0Å and 3.4Å. Sterics were checked for each allowed conformation using AMBER van-der Waals radii scaled by 0.8.

Quantum-chemical calculations were performed using Jaguar [7] on a dual-1Ghz PentiumIII running Linux. Models were geometry-optimized using HF calculations with a 6-31G** basis set, and point energies were evaluated at the LMP2/6-31G**//HF/6-31G** level. Interaction energies were computed as $\Delta E = E_{\text{complex}} - E_{\text{base}} - E_{\text{aa}}$, and included BSSE corrections. Energies of the components were computed using optimized geometries from the complex. BSSE corrections were performed using the counterpoise method [8a-b] only on the HF component of the interaction energy [5c].

The September 11, 2000, version of the Protein Data Bank (PDB) [4] was searched for observed hydrogen bonds using an automated approach [2c]. Crystal structures, average minimized NMR structures, and ensemble NMR structures were

considered. Crystal structures used had better than 3.5Å resolution and were protonated with InsightII (Biosym). Each model in ensemble NMR structures was analyzed separately. Nonspecific polymerases and topoisomerases were filtered out because of the large number of mutant structures and because of the non-specificity of the interactions.

Results and Discussion

We have written a program to generate all possible hydrogen bonding patterns between two molecules by doing an exhaustive 3D search. [2c]. An initial hydrogen bond is made, and then 25-80 million conformations are generated within the preferred space of the hydrogen bond, as defined by small crystal studies [12]. Conformations that confer additional hydrogen bonds are kept. All conformations with identical hydrogen-bonding patterns are compared and a representative is kept. Using this method, we found 28 unique ways protein side chains can make two hydrogen bonds to a neutral base (Figure 1).

Our previously modeled interactions can be divided into four groups based on the donor angle of the hydrogen bond: (1) five “38°” interactions that include donor angles on the edge of our parameter range (not shown), (2) two initially-bifurcated three-hydrogen bond interactions, (3) six interactions involving a charged side chain, and (4) all others. The 38° group interactions were not preserved by geometry optimization, and thus do not constitute an inherently stable interaction. (James Robertson, ACC, ADF, unpublished). The optimized geometries of the interactions we performed calculations for are shown in Figure 2.

The two bifurcated interactions, models #7 (Asn-C) and #18 (Asn-G), have similarity to two-hydrogen bonded non-bifurcated interactions, models #8 and #17. In both cases, the bifurcated and non-bifurcated interactions minimize to identical structures (Figure 2), and in the case of Asn-G (#17 and #18), the heavy-atom hydrogen bond distances are shifted slightly to allow for a weak 3.4Å hydrogen bond to the third group. In Asn-C (#7 and #8), the third hydrogen bond is not present in the optimized

structure, since it has a donor-acceptor distance of 3.8Å. In the rank order of the model interactions, these two initially bifurcated interactions have the greatest stability of the non-charged models.

Looking at table 1, we see that interactions involving charged side chains (Lys, Asp, Arg) are more favorable, which correlates with mutational analysis that has shown the importance of charged hydrogen bonds in specific recognition [9]. Within the charged interactions Lys-G #20 is 10 kcal/mol more favorable than a similar Arg-G #21 interaction, although Arg-G constitutes at least 43% of bidentate interactions while Lys-G constitutes less than 15% [2b]. We speculate that this is due to the greater dynamic flexibility of Lys that makes it difficult to gain specific positioning in competition with the negatively charged phosphates on the nucleic backbone. Lys-G #20 remains the third most common bidentate interaction [2b][2c], presumably due to its intrinsic stability. Asp/Glu-G #23, involving the N1 and N2 of G, is also calculated to be 10 kcal/mol more favorable than Arg-G #21. Because Asp/Glu are shorter side chains, are repelled by the phosphate backbone, and have desolvation penalties that should be similar to Arg-G #21, we suggest that Asp/Glu-G may be an important interaction for specific recognition of RNA. The lower frequency of an exposed G Watson-Crick face may be responsible for the observed frequencies. We note that this interaction is essential to the specificity of the TRAP complex, where it is repeated 17 times (Glu36 and Asp39) [10]. The strong favorable interaction and the involvement of a non-canonical feature makes Asp/Glu-G a potentially influential specificity strategy.

All the interactions that have not been observed in the PDB, with the exception of one, involve the Watson-Crick face. Such interactions are expected to occur mainly

with RNA, and there are currently not sufficient numbers of structures to make a conclusive statement. The one exception, Asn/Gln-G #19 involves the major groove of G, and is incidentally ranked second to last in energetic favorability. Looking at the two most common interactions in DNA, we find that Arg-G #21 and Asn-A #10 have interaction energies of -36.54 and -14.85 kcal/mol, which correlates with experimentally observed occurrence rates. A careful PDB survey [2b] of bidentate protein-DNA interactions found that 43% are Arg-G (model #21) and 26% are Asn/Gln-A (model #10) interactions. Inclusion of Arg-G interactions likely to be bidentate [2b] shifts the percentages to 61% Arg-G interactions and 18% Asn/Gln-A. While the correlation is good, we note that we have not considered differences in desolvation.

Although Asn/Gln #10 is the second most common bidentate interaction with DNA, the interaction is perplexingly rare with RNA [2a][2c]. In RNA, Ser/Thr/Tyr-A #11 interactions take the place of second-most common interaction, making up 15% of interactions [2c], and no Asn/Gln-A #10 interactions are observed. While the dataset of 45 bidentate side chain-RNA hydrogen bonding interactions [2c] is small, the difference is striking. Our study finds Asn/Gln-A and Ser-A to have similar energies of -14.85 and -14.80 kcal/mol, respectively. Other factors must contribute to the difference between DNA and RNA. One possibility is that the deep, narrow major groove of A-form RNA can more easily allow the hydroxyl of Ser/Thr/Tyr than the bulkier carboxamide of Asn/Gln. This steric restriction is not the single determinant of the difference in interaction frequencies, since both A-form and B-form helices can accommodate either interaction [ACC and ADF, in preparation]. Looking at the PDB

data, we found that the Ser/Thr/Tyr-A interactions observed so far with both DNA and RNA exclusively involve amino acids in non-helical protein structures (Table 2). Asn/Gln-A interactions can also be found in such structures, but helical regions predominate at 46% of interactions (Table 3). Others have previously shown that the protein backbone conformation is an important determinant of what side chain-DNA interactions are possible [11]. Ser/Thr/Tyr appears unable to make a bidentate interaction in the context of a helix in a groove, and is thus more suited for the varied binding modes and more diverse RNA tertiary structures found at RNA-protein interfaces.

Conclusion

Clearly effects other than interaction energy impact the selection of a particular hydrogen-bonding scheme for specificity. By having a measure of the intrinsic interaction affinity for discrete models of specificity, we can begin to deconvolute the various contributions to specificity. This first study of protein side chain-nucleic base interactions provides gas-phase energies. By using solvent models such as the Onsager and PCM models, it should be possible to provide a measure of the effect of solvation on the interaction energies. Recent molecular-mechanics/continuum solvation schemes [13] for calculating free energies of interactions may provide an even better measure. Finally, simple experimental systems will be needed to confirm and assess the various contributions.

Acknowledgement

We thank Jan Jensen, Preston Snee, Martin Head-Gordon, David Agard, and Wendell Lim, and James Robertson for advice. This work is supported by NIH grants GM47478 and GM56531 (A.D.F.) and NIH Training grants GM08284 and GM08388 (A.C.C.).

References

- (1) Seeman, N.C.; Rosenberg, J.M.; Rich, A. *Proc. Natl. Acad. Sci.* **1976** 73, 804-8.
- (2) (a) Allers, J, Shamoo, Y. *J. Mol. Biol.* **2001** 311:75-86 (b) Luscombe, N. M., Laskowski, R. A., Thornton J. M. *Nucleic Acids Research* **2001**, 29, 2860-2874. (c) Cheng, A.C.; Chen W.; Fuhrmann C.; Frankel AD. Submitted to *J. Mol. Biol.* **2002**.
- (3) Walberer, B.J.; Cheng, A. C.; Frankel, A.D. Submitted to *J. Mol Biol.* 2002
- (4) (a) Berman, H. M.; Westbrook J.; Feng, Z.; Gilliland, G; Bhat, T. N.; Weissig, H.; Shindyalov I. N.; Bourne P. E. *Nucleic Acids Research* **2000** 28, 235-242. (b) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E. Jr; Brice, M.D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Bio* **1997** 112, 535-542
- (5) Hobza P., Kabelac M, Sponer J, Mejzlik P, Vondrasek J. *J Comp Chem* **1997** 18, 1136-1150 (b) Sponer J.; Hobza P. *J. Phys. Chem. A* **2000** 104, 4592-4597 (c) Kim K.; Friesner, R. A. *J. Am. Chem. Soc.* **1997** 119, 12952-12961
- (6) (a) S. Sæbø and P. Pulay, *Theor. Chim. Acta* **1986** 69, 357 (b) S. Sæbø and P. Pulay, *Ann. Rev. Phys. Chem.* **1993** 44, 213 (c) S. Sæbø, W. Tong, and P. Pulay, *J. Chem. Phys.* **1993** 98, 2170
- (7) Jaguar Software v 4.1 Schrodinger, Inc. Portland, Oregon
- (8) (a) Boys SF, Bernardi F. *Mol. Phys.* **1970** 19:553 (b) Jansen HB.; Ross P. *Chem. Phys. Lett.* **1969** 3, 140-143 (c) Schuetz M, Rauhut G, Werner H-J. *J. Phys. Chem. A* **1998** 102, 5997-6003

- (9) Fersht AR, Shi J-P, Knill-Jones J, Lowe DM, Wilkinson AJ, Blow DM, Brick P, Carter P, Waye MMY, Winter G. *Nature* **1985** 314, 235-238.
- (10) Antson AA, Dodson EJ, Dodson G, Greaves RB, Chen X, Gollnick P. *Nature* **1999** 401, 235-42
- (11) (a) Pabo, C., Nekludova, L. *J. Mol. Biol.* **2000** 301, 597-624 (b) Suzuki, M. *Structure* **1994** 2, 317-326. (c) Kono, H.; Sarai A. *Proteins* **1999** 35, 114-31.
- (12) (a) Taylor, R., Kennard, O., Versichel, W. *J. Am. Chem. Soc.* **1983** 105, 5761-5766 (b) Taylor, R., Kennard, O., Versichel, W. *J. Am. Chem. Soc.* **1984** 106, 244-248 (c) Taylor, R., Kennard, O. & Versichel, W. *Acta Cryst. B* **1984** 40, 280-288.
- (13) Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan Y., Wang W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., Cheatham, T. E. 3rd *Acc Chem Res.* **2000** 33, 889-97.

Legends

Figure 3.1. All two hydrogen bond interactions between an side-chain and a base, as modeled using WASABI. Interactions where the hydrogen bond donor angle is marginal (between 36° and 38°) are labeled with letters A through E. All other interactions are labeled with numbers 1 through 23. Interactions seen in the PDB are indicated by a star, and the two interactions that involve a bifurcated three hydrogen bond interactions are outlined by a box. Models are grouped by the base face used in the interaction.

Figure 3.2. HF/6-31G** geometry optimization of the models. Each geometry-optimized model is represented by a top view and a edge view. The edge view is taken perpendicular to the interaction direction. In the top view, expected hydrogen bonds have been drawn in.

Table 3.1. Ranking of side chain-base interaction energies.

Rank	Model	Base Face	Observed in PDB?	$\Delta E(\text{LMP2})$ kcal/mol	$\Delta E(\text{HF})$ kcal/mol
1	Lys-G #20	major	yes	-47.34	-49.47
2	Asp/Glu-G #23	watson-crick	yes	-43.41	-47.49
3	Lys-C #5	watson-crick	yes	-41.48	-46.97
4	Arg-G #21	major	yes	-36.54	-39.12
5	Arg-C #6	watson-crick	yes	-35.29	-38.31
6	Arg-G #22	major	yes	-33.16	-35.81
7	Asn/Gln-G #18	watson-crick	no	-20.48	-22.22
8	Asn/Gln-G #17	watson-crick	no	-20.48	-22.21
9	Asn/Gln-C #8	watson-crick	no	-18.95	-17.85
10	Asn/Gln-C #7	watson-crick	no	-18.94	-17.84
11	Asn/Gln-A #12	watson-crick	no	-15.61	-13.59
12	Asn/Gln-G #14	minor	yes	-15.43	-13.82
13	Ser-G #16	watson-crick	no	-15.33	-14.29
14	Asn/Gln-A #10	major	yes	-14.85	-13.02
15	Ser-A #11	major	yes	-14.80	-12.26
16	Ser-C #9	watson-crick	yes	-14.72	-13.01
17	Ser-G #15	minor	yes	-13.41	-10.99
18	Ser-A #13	watson-crick	yes	-13.15	-11.16
19	Ser-U #3	watson-crick	yes	-12.46	-11.13
20	Asn/Gln-U #1	watson-crick	no	-12.30	-13.09
21	Asn/Gln-U #4	watson-crick	yes	-12.19	-14.07
22	Asn/Gln-G #19	major	no	-10.21	-9.24
23	Ser-U #2	watson-crick	no	-9.85	-10.41

Table 3.2. Observed Ser/Thr/Tyr-A intermolecular hydrogen bonds

PDB id	type	Complex	Interaction	Structure
1RBJ	crystal	Ribonuclease	Thr45:A201	mid-strand
5MSF	crystal	MS2 coat protein-aptamer	Thr45:A11	mid-strand
1DZ5	nmr	U1A-Pie RNA	Ser45:A25	strand end
1DZ5	nmr	U1A-Pie RNA	Thr88:A44	strand end
1BP7	crystal	CreI endonuclease-DNA	Tyr33:A13	beta turn
1TN9	nmr	TN916 integrase-DNA	Tyr40:A120	mid-strand

Table 3.3. Secondary structures of Asn/Gln-A interactions found in crystal and NMR structures

A. Secondary structures of Asn-A (Model #10) interactions found in crystal structures

XTAL: AsnND2 : Ade N7 AsnOD1 : AdeN6				
<i>Pdbid</i>	<i>aa</i>	<i>base</i>	<i>secondary structure</i>	<i>function</i>
1az0	A 185	D 805	turn	ecorv endonuclease
1b97	A 185	D 5	turn	ecorv endonuclease
1bgb	B 185	D 905	turn	ecorv endonuclease
1rv5	A 185	D 5	turn	ecorv endonuclease
1rvb	A 185	D 5	turn	ecorv endonuclease
4rve	A 185	D 4	turn	ecorv endonuclease
1bsu	A 185	D 805	turn	endonuclease
1bua	A 185	D 805	turn	endonuclease
1fok	A 13	B 905	turn	foki restriction endonuclease
1cvj	B 100	N 2	turn	polydenylate binding protein 1
1a1f	A 121	B 9	helix	zinc finger peptide
1akh	A 120	C 26	helix	mating-type protein
1apl	C 182	A 16	helix	mat alpha2 homeodomain
1b72	A 253	D 13	helix	homeobox protein
1b72	B 286	D 9	helix	homeobox protein
1cqt	A 151	M 210	helix	pou domain
1du0	A 51	C 213	helix	engrailed homeodomai
1du0	B 151	D 301	helix	engrailed homeodomai
1fjl	A 51	D 5	helix	homeodomain
1hdd	D 51	B 22	helix	Engrailed homeodomain complex
1mey	C 19	A 10	helix	consensus zinc finger protein
1mnm	D 182	F 51	helix	mcm1 transcriptional regulator
1ym	A 120	C 26	helix	mating-type protein a-1
2drp	A 125	B 10	helix	tramtrack protein (zinc-finger)
2drp	A 155	B 7	helix	tramtrack protein (zinc-finger)
2hdd	A 51	C 13	helix	engrailed homeodomain
3hdd	A 51	C 213	helix	engrailed homeodomain
9ant	A 51	D 220	helix	antennapedia protein
1d02	A 117	C 4	distorted helix	type ii restriction enzyme muni
2pvi	A 140	C 7	distorted sheet	type ii restriction enzyme pvuii
1pvi	B 140	D 7	distorted sheet	pvuii endonuclease
3pvi	A 140	C 7	distorted sheet	pvuii endonuclease
1eyu	B 140	D 7	distorted sheet	type ii restriction enzyme pvuii

Table 3.3 (continued)

B. Secondary structures of Gln-A (Model #10) interactions found in crystal structures

XTAL: GlnNE2 : AdeN7 GlnOE1 : AdeN6				
<i>pdbid</i>	<i>aa</i>	<i>base</i>	<i>secondary structure</i>	<i>function</i>
1fok	A 12	C 936	turn	foki endonuclease
1mey	C 16	A 11	turn	zinc finger
1mey	C 44	A 8	turn	zinc finger
1ubd	C 396	B 28	turn	zinc finger
1a02	N 571	A 4005	helix	transcription complex
1au7	A 44	C 460	helix	pou-domain
1cqt	A 44	M 204	helix	pou domain
1lli	A 44	D 4	helix	Lambda repressor mutant
1lmb	3 44	1 4	helix	Lambda repressor
3cro	R 28	A 5	helix	434 cro protein
1a1h	A 118	B 10	helix	qgsr zinc finger peptide
1oct	C 44	A 204	helix	Oct-1 (pou domain)
1bdt	D 9	F 7	strand	arc repressor
1bdv	D 9	F 7	strand	arc repressor
1par	A 9	F 18	strand	arc repressor
1a73	A 63	D 17	strand	homing endonuclease
1a74	A 63	D 417	strand	homing endonuclease
1cyq	A 63	D 517	strand	homing endonuclease
1cyq	B 263	C 417	strand	homing endonuclease
1cz0	A 63	D 517	strand	homing endonuclease
1cz0	B 263	C 417	strand	homing endonuclease
1evw	A 63	O 16	strand	homing endonuclease
1ipp	A 63	D 217	strand	homing endonuclease
1bp7	B 38	1 4	strand	i-crei endonuclease
1ecr	A 250	B 312	strand	replication terminator protein
1f3i	A 243	M 115	strand	tn5 transposase
1cez	A 758	T 8	strand	t7 RNA polymerase
1qln	A 758	T 13	strand	t7 RNA polymerase

Table 3.3 (continued)

C. Secondary structures of Asn-A (Model #10) interactions found in NMR structures

NMR: AsnND2 : AdeN7 AsnOD1 : AdeN6				
<i>pdbid</i>	<i>aa</i>	<i>base</i>	<i>secondary structure</i>	<i>function</i>
1mse	C 183	A 4	helix	c-myb
1msf	C 179	A 5	helix	c-myb
1msf	C 183	A 4	helix	c-myb
1tf3	A 89	E 5	helix	transcription factor iiaa
1yuj	A 48	B 105	helix	gaga-factor

Table 3.3 (continued)

D. Secondary structures of Asn-A (Model #10) interactions found in NMR structures

NMR: GlnNE2 : AdeN7 GlnOE1 : AdeN6				
<i>pdbid</i>	<i>aa</i>	<i>base</i>	<i>secondary structure</i>	<i>function</i>
1a66	A 176	B 320	distorted strand	nfatc1

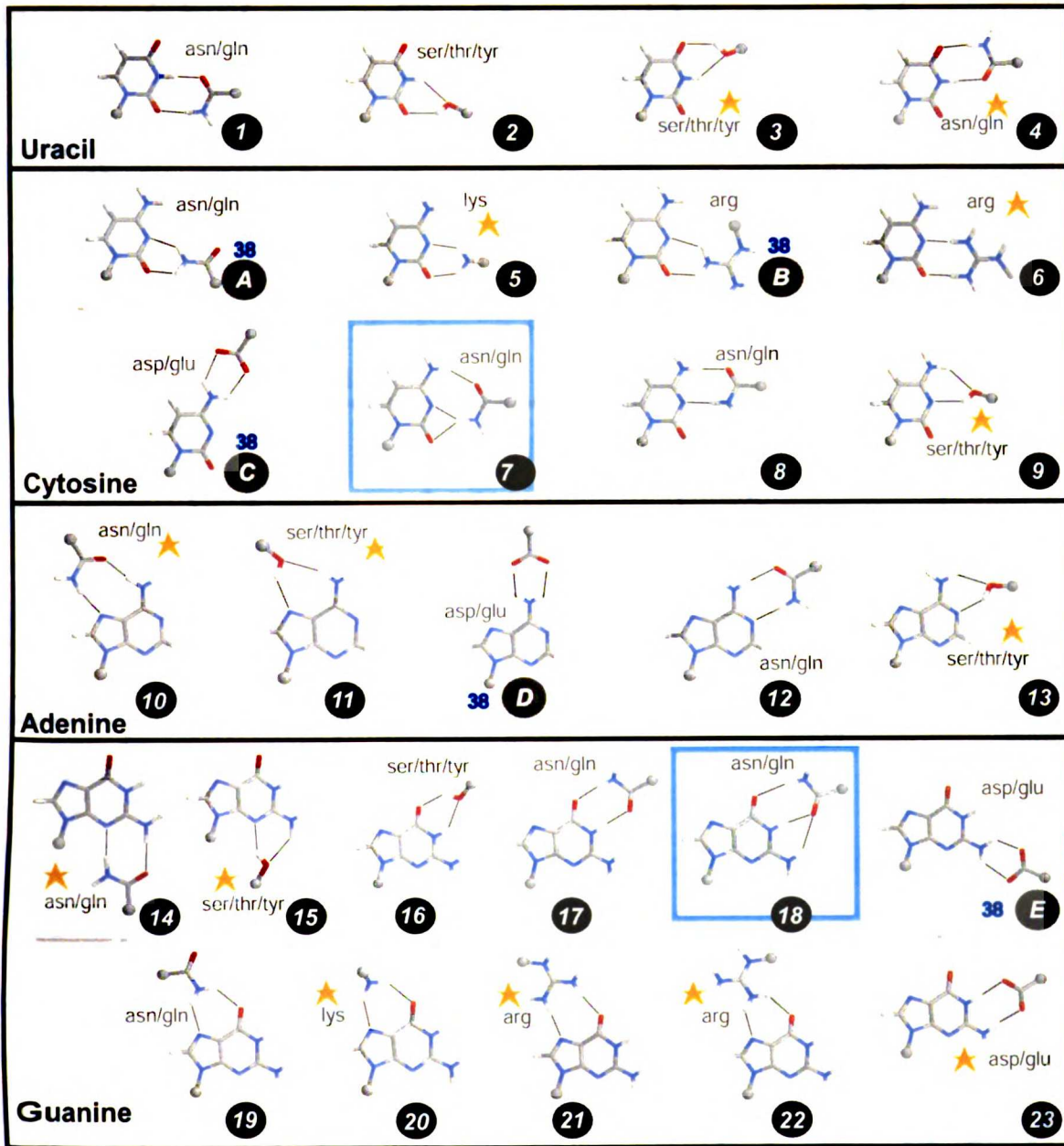


Figure 3.1

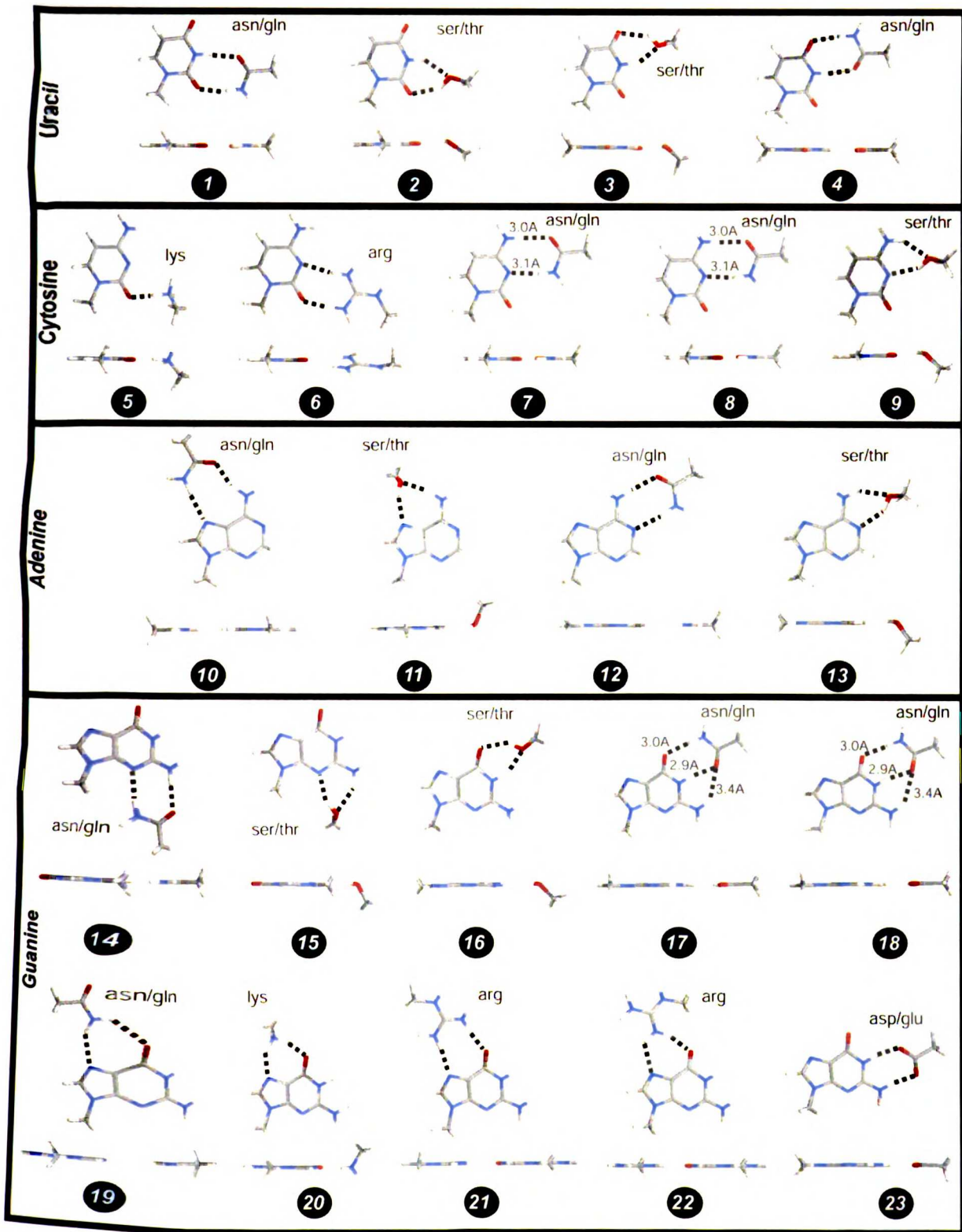


Figure 3.2

Chapter 4

**Strategies for discrimination of A-form RNA
and B-form DNA by single protein sidechains**

Alan C. Cheng and Alan D. Frankel

Graduate Group in Biophysics
and
Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, CA 94143-0448

Abstract

Sequence-specific recognition of nucleic acids by proteins is key to biological processes, and hydrogen bonding has been shown to be a significant interaction in determining specificity. Recent studies have highlighted the importance of hydrogen bonding networks involving multiple steps in specific recognition of duplex DNA. Protein-RNA interactions have more complex tertiary structures, but duplex RNA often remains an important structural feature. We have exhaustively modeled all possible hydrogen bonding networks made by protein side chains to idealized A-form RNAs and B-form DNAs, and we have identified strategies for specific recognition in the context of the canonical helices. Because of twist differences in ARNA and BDNA, it is nearly impossible to make a 3' cross-strand interaction in the major groove of BDNA, whereas such interactions are possible in ARNA. On the other hand, there are few multi-step minor groove interactions in ARNA, whereas the orientation of the sugar O4' in BDNA allows for numerous strategies for multi-step interactions in BDNA. We show how the backbone phosphate and sugars influence interactions in the major and minor grooves of ARNA and BDNA, and how the addition of the backbone leads to new single base interactions that are specific to one canonical helix. We enumerate all possible interaction patterns, and find the most common pattern in our database of models is to adjacent steps on the same strand, which correlates well with the observed frequencies of multi-step patterns in the PDB. Finally, we show several three-step strategies for sequence specific recognition of a triplet sequence by a single amino acid, and highlight several strategies that are unique to one of the canonical helices. The strategies illustrate

Introduction

Specific recognition of RNA and DNA is essential to biological function, and a number of recent studies have explored recognition of RNA (Tregar *et al.*, 2001; Jones *et al.*, 2001; Allers & Shamo, 2001) and DNA (Nadassy *et al.*, 2001; Luscombe *et al.*, 2001) (Pabo *et al.*, 2000; Jones *et al.*, 1999) based on the repertoire of known structures in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Berman *et al.*, 2001).

A variety of atomic interactions are possible at the protein-nucleic interface, including van der Waals contacts, water-mediated hydrogen bonds, hydrophobic contacts, stacking contacts, electrostatic contacts, and hydrogen bonds. While all these occur in interfaces, side chain-base hydrogen bonds have for a long time been hypothesized to be generally the most significant in sequence-specific recognition (Seeman *et al.*, 1976).

Recent studies have shown that DNA has one intermolecular hydrogen bond per 125\AA^2 of interface area (Nadassy *et al.*, 1999), whereas RNA has been shown to have even more hydrogen bonds at the interface, with one bond per 100\AA^2 (Luscombe *et al.*, 2001). These hydrogen bonds can be base-specific or made to the backbone sugar and phosphates. In both nucleic acids, about 60% of hydrogen bonding interactions are made to the phosphate and sugar backbone, with the remaining made to the helical structures. Statistical studies (Luscombe *et al.*, 2001; Jones *et al.*, 1999) suggest that interactions are preferentially made to base atoms rather than backbone atoms in DNA, supporting the belief that backbone interactions are largely nonspecific in nearly canonical nucleic helices. On the protein side, 70-75% of interactions involve the side chain, as opposed to the protein backbone with both DNA (Luscombe *et al.*, 2001) and RNA (Tregar *et al.*, 2001).

While other interactions occur and may be more numerous, hydrogen bonds generally provide the greatest specificity in protein-nucleic acid recognition, and show statistically significant preferences in amino acid-nucleic interactions (Luscombe *et al.*, 2001) (Mandel-Gutfreund *et al.*, JMB). Seeman *et al.* first suggested side chain interactions to single bases were important in DNA-protein recognition. More recently, studies (Luscombe *et al.*, 2001) (Suzuki, 1994) have begun to explore side chain recognition to multiple steps of DNA, since multiple hydrogen bonds across multiple steps are likely to confer greater specificity. In RNA, the data is sparser. However, from the detailed analysis cited above, and from structure-function studies on protein-RNA complexes (Draper, 1999), it is clear that base-specific hydrogen bonds to tertiary structures are important for sequence-specific recognition.

While non-canonical features are often involved in RNA recognition, the duplex helix is common and frequently found in RNA tertiary structures. At binding sites, duplex RNA often is accompanied by non-canonical features that make the major groove more accessible (Weeks & Crothers, 1993). Complex interactions to the duplex RNA structure can be and are used to increase specificity to the binding regions often found at non-canonical sites.

Duplex RNA is almost always A-form, as opposed to the B-form that idealized DNA duplexes form. The differences in the two forms are highlighted in Figure 1. The salient features that affect complex hydrogen bonding patterns to the sites are the difference in rise and displacement from the helical axis. While the BDNA base pairs are centered along the helical axis, ARNA base pairs are displaced from the helical axis by about 4.5Å, resulting in the base pairs tilting into the major groove relative to each other

(Saenger, 1984) (Blackburn & Gait, 1996). We will see this tilt and displacement has a particularly large effect on the array of donors and acceptors in the minor groove compared to BDNA. ARNA has one more base per turn compared to the 10 bases per turn of DNA, due to its enlarged helix. This results in a smaller twist per base pair in ARNA ($360^\circ/11 = 32.7^\circ$) compared to BDNA ($360^\circ/10 = 36^\circ$), as well as less of a rise. (Saenger, 1984) (Blackburn & Gait, 1996) Both the rise and twist have effects on the arrangement of donors and acceptors, and the possible complex interactions that an amino acid can make.

Recent studies support a general model where discrimination between A-form and B-form nucleic acids in a non-specific manner often involves large complementary surface areas as well as backbone interactions (Ryter & Schultz, 1998), and addition of complex hydrogen-bonding interactions involving multiple steps can be used to gain specific recognition. We have previously enumerated all possible hydrogen bonding patterns between side chains and RNA bases and non-canonical base pairs. (Cheng *et al.*, 2002) To address sequence-specific recognition using complex interactions, we have extended our modeling study to generate models of all possible side chain interactions to idealized A-form and B-form helices where at least one hydrogen bond is to a base. Based on the assumptions that hydrogen bonding groups in general must be satisfied, and that multiple hydrogen bonds from single side chains can be used to enhance binding specificity, we highlight the most interesting interactions, as well as the difference in specificity strategies possible for ARNA compared to BDNA. Because small peptide motifs, such as alpha helices and beta turns, can be used to specifically recognize either DNA or RNA (Frankel, 2000), one question we explore is how single amino acids

making multiple hydrogen bonds can be used to differentiate one helix form over another. In practice such discrimination will be only one component of recognition, but the possibility can be important in small peptide recognition motifs such as the arginine rich motif, where a few residues are tasked to specify a complex site with high specificity.

Methods

Molecular representation

The amino acid side-chains are generated using the LEaP package of AMBER, using the param94 residue definitions (Cornell *et al.*, 1995) as previously described (Cheng *et al.*, 2002). Idealized A-form RNA and B-form DNA Watson-Crick duplexes were built using the Nucleic Acid Builder (NAB) package (Macke 1998), based on the AMBER param94 residue definitions (Cornell *et al.*, 1995) and fibre diffraction data (Arnott *et al.*, 1973) included with NAB. To create each set of duplexes, we generated the 32 non-redundant triplet sequences, and used a program written using the NAB language to generate the duplexes. A-form RNA triplets were generated as is, and B-form DNA triplets were flanked by 8 base pairs on both the 5' and 3' side. The intent is to present the triplet sequences in biologically relevant contexts. RNA duplexes may be flanked by sterically non-restrictive non-canonical motifs, while DNA duplexes are commonly found in the context of an extended helix. The flanking sequence used, AGTCAGTC, provided a steric context, and was not searched for hydrogen bonding. We found that 7 flanking base pairs were required to define the steric space of all amino acid moieties in a continuous B-form DNA helix.

Database construction

WASABI (Cheng *et al.*, 2002) was used to perform an exhaustive geometric search for hydrogen bonds between amino acid hydrogen-bonding moieties and idealized ARNA and BDNA nucleic acids as described above. The details of the search are presented in Cheng *et al.*, 2002, and in essence what is done is a search with five degrees

of freedom of the allowed hydrogen bonding space as defined by studies of hydrogen bonds in small crystals (Taylor *et al.*, 1984), with consideration for van der Waals sterics and polar hydrogen clashes. Four degree step sizes are used, and the allowed hydrogen bond space for is defined as +/- 38 degrees for donors, +/-50 degrees for nitrogen acceptors, +/-90 degrees for oxygen acceptors, and non-hydrogen donor-acceptor distances of $\leq 3.4\text{\AA}$. The lower bound of the hydrogen bond distance is set by hard sphere steric radii, as described previously (Cheng *et al.*, 2002). One representative conformation of each hydrogen-bonding pattern is saved. To generate our database of interactions in ARNA and BDNA helices, we performed a search with each side-chain moiety against each of the 32 triplet sequences in each helix. We saved all unique hydrogen bonding patterns with two or more hydrogen bonds that had at least one hydrogen bond to a base atom (as opposed to backbone atoms). The database construction required two weeks of calculation on four 1-Ghz PentiumIII and one 1.4Ghz Pentium4 processors. About one trillion conformations were sampled.

As will be seen in the discussion section, for ARNA we get ten additional interactions to the Watson-Crick face of the bases that were not found in our previous search for side chain-base pair interactions (Cheng *et al.*, 2002). This is mainly due to the base pairs in the previous study being planar, while the ARNA duplex base pairs have a propeller twist of 13.75° . This represents a limitation of our modeling nucleic acids as rigid moieties, and shows that moderate changes in the conformation of the nucleic acid can lead to a different set of interactions. We have attempted to address this issue with loose hydrogen bond parameters. However, one potentially negative consequence of that is illustrated by these interactions, which have a requirement of a $37^\circ - 38^\circ$ donor angle,

may not be realistic. In general, interactions close to the edge of the donor parameter range in particular have to be considered potentially marginal.

In addition to those generated for ARNA and BDNA, we also generated databases of interactions for ADNA triplets and ADNA flanked helices, which involved 1,833 and 1,812 interactions respectively. Thus the flanking structure sterically precludes 21 of the 1,833 interactions that are possible in the triplet form. This number reduces to only four when we restrict the donor angle to $\leq 25^\circ$. The ADNA databases served as controls for our model-generating algorithm: of the 1,812 ADNA flanked interactions, two were not found in the ADNA triplet set. If we eliminate interactions that are at the parameter limits, within 0.001Å of the distance cutoff or within 0.01° of the hydrogen bond angle cutoff, all interactions found in the flanked version are found in the triplet version. The ADNA databases served mainly as controls and will not be further discussed in this work.

Generation of parameter restricted databases

We took all models generated by the search, and created two sets that had models satisfying more restrictive hydrogen bond parameters. Although geometries do not correspond directly to interaction free energies, because other interaction components are also important, hydrogen bonds do have clear geometric preferences. Because the hydrogen bond strength is most sensitive to perturbations in donor angle and distance, these two parameters were used as a sieve to cull out sets with more ideal hydrogen bonds. We will refer to the sets as “good” and “nearly ideal” to reflect the ideality of the hydrogen bonds in the model. The complete set is referred to as “all”. In the “good” set, all hydrogen bonds were required to have donor angles ≤ 25 degrees, and non-

hydrogen donor-acceptor distances of ≤ 3.1 Å. In the “nearly ideal” set, the same hydrogen bond distance cutoff was used, but donor angles were restricted to ≤ 10 degrees.

Analysis of databases

The resulting models from the WASABI search were processed and characterized, then loaded into a MySQL database system. Programs written in Java 1.2 and using the SQL language were used to filter the models to arrive at the final working databases of models. The construction and analysis of the databases is summarized in Table 1, which provides numbers for each of the processing steps we will now describe.

The results of the WASABI search are first filtered for target redundancies, such that multiple interactions are represented by one model if identical atoms on a side-chain make an interaction to identical atoms on the identical base types in identical configurations. Since the triplet sequences are non-redundant, the filter in practice only removes redundant interactions involving only two-steps of the nucleic acid (two sequential base pairs). Next, to generate a representative dataset of all interactions, amino acid “redundancies” are removed such that similar interactions with each side-chain are consolidated and represented by one model. For instance, interactions involving similar “faces” of the guanidinium group of arginine are considered similar, and only one representative is kept. We also remove all Tyr interactions that can be represented by Ser interactions, further simplifying the database. Next, we remove interactions that are not fixed, that is, interactions that involve a single bifurcated hydrogen bond where only one atom on one of the molecules is involved in a hydrogen

bond. The result is a representative set of interactions where at least two pairs of atoms are hydrogen bonded. Removal of all bifurcated hydrogen bonds further simplifies the database to a “processed database” that provides a reasonably non-redundant working database. To focus on complex interactions that involve multiple steps of a nucleic acid helix, we can remove interactions involving a single base or single base pair. We also can remove interactions that are not base specific, that is, interactions that do not make at least one hydrogen bond to the base face of each base position recognized. This eliminates interactions that may recognize one base, but interacts with the backbone at the second position.

By comparing the ARNA and BDNA sets of models, one can cull out interactions that can occur in one helix but not in the other. Performing the same filtering steps done above for the “all” interaction dataset results in a representative set of non-bifurcated unique interactions. We note that when we look at differences, not all interactions that are different in one parameter-restricted set will be different in another parameter-restricted set. Clearly a less restrictive set (e.g., going from “good” to “all”) can have more diverse interactions. However, it is also possible that a more restrictive set will have interactions not found in a less restrictive set (e.g., going from “all” to “good”). This is because an interaction found in the “good” ARNA set, for instance, may not be found in the “good” BDNA set, but can be found in the “all” parameter-restricted set. When we compute the differences in one set, we don’t see interactions of a poorer geometry found in a less restrictive set. We performed our calculations in this way as an approximate way to account for our use of relatively loose hydrogen-bonding parameters. If an interaction is found in the two more stringent model sets but not in the all-inclusive

model set, this interaction may still be relatively unique, because of the poorer hydrogen bonds found in the all-inclusive model set. The difference in hydrogen bond quality can be thought of as providing an interaction quality gap corresponding to a hydrogen-bonding energy “gap”.

Observed interactions

We identified all base-specific protein-RNA hydrogen-bonding interactions in crystal structures of 3.5Å or better resolution found in the November 11, 2001, release of the PDB. The same angle parameters used for the search, and a relaxed 3.5Å distance cutoff was used for identifying hydrogen bonds in the PDB. NMR structures of protein-nucleic interactions were not considered because the backbone is generally not well determined. Luscombe et al (Luscombe *et al.*, 2001), have performed a study of base-specific hydrogen bonding by amino acids to DNA based on a 1998 release of the PDB, and we have utilized that here for our studies.

Results and Discussion

We have generated all possible hydrogen-bonding patterns between amino acid side chains and canonical A-form RNA and B-form DNA molecules by doing an exhaustive 3D search (Cheng *et al.*, 2001). For every donor/acceptor combination an initial hydrogen bond is made, and then 25-80 million conformations are generated within the preferred space of the hydrogen bond, as defined by small crystal studies (Taylor *et al.*, 1984). Conformations that confer additional hydrogen bonds are compared, and a representative of each hydrogen-bonding pattern is kept. Because we wish to look at interactions that depend on sequence, we only consider interactions that make at least one hydrogen bond to a base.

Using a series of filters (see Methods and Table 1), conformations with similar hydrogen-bonding interactions are compared and a representative is kept, thereby reducing the number of interactions from 48,000 total interactions in BDNA and 41,000 interactions in ARNA to a representative 2000 interactions in BDNA and 1000 interactions in ARNA. We have also have constructed difference databases for ARNA and BDNA where interactions common between the two are deleted.

We classify interactions into groups based on the geometric quality of the hydrogen bond donor angle and distance (Cheng *et al.*, 2002). The "all" set includes all interactions found in our search, which allows a donor angle of $\pm 38^\circ$, and non-hydrogen donor-acceptor distance of $\leq 3.4\text{\AA}$ (see methods). Our "good" set is a moderately restricted set that requires interactions to have all donor angles $\leq 25^\circ$, and all hydrogen bond distances $\leq 3.1\text{\AA}$. The "nearly ideal" set includes only interactions that have nearly

ideal hydrogen bonds, where all donor angles are $\leq 10^\circ$, and all hydrogen bond distances are $\leq 3.1 \text{ \AA}$. While the parameters used don't directly correspond to interaction free energies, the parameters do represent the strongest preferences of the hydrogen bond (Verischel *et al.*, 1984; Jeffrey & Saenger, 1991). The "good" set includes 275 interactions to ARNA and 363 interactions to BDNA, and the "nearly ideal" set includes 77 interactions to ARNA and 85 interactions to BDNA.

We also analyze models based on whether they are "base-specific". Base-specific interactions are required to have at least one hydrogen bond to the base of each position recognized, and constitute a set analogous to spanning interactions for base pairs (Cheng *et al.*, 2002), and two-hydrogen bonded interactions for single bases (Cheng *et al.*, 2002; Seeman *et al.*, 1976). By requiring at least one hydrogen bond to each base, we can separate out sequence specific recognition strategies.

Overview of the database

Comparing the total number of interactions in each data set can alone give insights into recognition of the nucleic acids. While the search in both helices results in comparable numbers of interactions ($\sim 48,000$ versus $\sim 41,000$), removal of target redundancy (Table 1) reduces the number of ARNA interactions down to $\sim 3,500$ of the original, whereas it only reduces the number of BDNA interactions down to $\sim 7,300$ of the original. Because the target redundancy filter can only remove two-step interactions, this indicates that ARNA has more two-step interactions than BDNA. In fact, for all non-bifurcated interactions, there are ~ 800 ARNA two-step models vs. ~ 650 BDNA two-step models. Taking out interactions that are not base-specific eliminates the difference,

because, compared to BDNA, ARNA has many more two-step interactions that require backbone contacts (Table 2), and most of these interactions are not base-specific.

More interactions in ARNA than BDNA are possible in the major groove for all interactions as well as for base-specific interactions (Fig 2 and 3). The only exception is for 3' cross-strand interactions (base patterns 221 and 230), discussed below. In BDNA, more interactions are possible in the minor groove. The exceptions here are for related 5' cross-strand interactions (base patterns 220, 231 321, and 334). Same strand step interactions (base pattern 222) in ARNA and BDNA have similar numbers of interactions in the major and minor groove. But these interactions are far outnumbered by interactions in the minor groove.

In both helices, base-patterns 221 and 230 have more possibilities for recognition in the minor groove, and base patterns 220 and 231 have more possibilities for recognition in the major groove. Base patterns 221 and 230 are related 3' cross-strand interactions, and base patterns 220 and 231 are related 5' cross-strand interactions. This is a result of the right-handed twist of the nucleic acids, which results in 5' cross-strand interactions being more easily facilitated in the major groove and 3' cross-strand interactions being easier to facilitate in the minor groove. Because the twist in BDNA is greater than in ARNA, it is nearly impossible to make a 3' cross-strand interaction in the major groove, thus making it a possibly good strategy for differentiating the two helix forms in the major groove. In fact in the BDNA major groove there are only 3 possible 3' cross-strand (base pattern 221) interactions, whereas there are 25 for ARNA. In the “good” set, there are no interactions with BDNA and seven with ARNA. In the “nearly ideal” set there is still one interaction to ARNA, shown in Figure 4a. Looking at the

difference databases, for the 3' cross-strand interaction pattern, 22 of "all" ARNA interactions, as expected, are unique to ARNA. This strength of this strategy is strongly correlated with the twist parameter, and is expected to fall out for non-ideal DNA nucleic acids that have twists substantially closer to the 32.7° found in ARNA and further from the ideal twist of 36° .

Looking at the set of unique interactions found in the difference databases, we see that BDNA has triple the number of unique interactions as ARNA (not shown). The difference databases of models also reflect the general distribution seen in all interactions. For BDNA, the majority of the unique interactions are found in the minor groove, with over half found in three-step helices (Table 4). The three-step interactions in turn largely reflects the strategic placement of backbone atoms in the BDNA minor groove, and the less useful orientation of backbone atoms in the ARNA minor groove, as described below.

We have listed all possible base patterns involving 2 or more steps in Figures 2 and 3, and highlighted the possible interactions in the major and minor groove of ARNA and BDNA helices, respectively. Counts for base-specific models are also listed. It is convenient to divide the interactions into 4 groups of base-patterns based on sequence recognition and the number of base steps involved: the single step (single base and base-pair), the two-step, the three-step where only the end bases are involved in direct hydrogen bond recognition, and the three-step where all three steps are involved in direct hydrogen bonds with the amino acid side-chain. The later three groups are labeled XY, X_Z, and XYZ in Figures 3 and 4, and tables 2-4. We will now discuss each group in turn.

Interactions to a single base step have been discussed previously (Cheng *et al.*, 2002) without consideration of the backbone. Our new study includes the backbone, and shows that the four Watson-Crick spanning interactions are also possible in the context of extended BDNA and triplet ARNA helices. Furthermore, all eight side-chain interactions to single bases in a base-pair are possible in both ARNA and BDNA.

For ARNA, we also get ten additional interactions that span the base pair and involve the Watson-Crick face of the bases due to the use of a triplet instead of an extended helix (also see Methods). These interactions were not found in our previous search for side chain-base pair interactions (Cheng *et al.*, 2002) because the base pairs in the previous study were planar, while the ARNA duplex base pairs have a propeller twist of 13.75° . Seven of the interactions involve bifurcated hydrogen bonds to the Watson-Crick positions of the bases. This represents a limitation of our modeling nucleic acids as rigid moieties, and illustrates how moderate changes in the conformation of the nucleic acid can lead to a different set of interactions. We have attempted to address this issue with loose hydrogen bond parameters. However, one potentially negative consequence of that is illustrated by these interactions, which have a requirement of a $37^\circ - 38^\circ$ donor angle and may not be realistic. Also, the microenvironment of the ARNA in the context of its complete functional structure can place additional, context-specific constraints.

As expected, the presence of the backbone allows for more diversity of base-specific interactions. With ARNA single bases, there are 29 interactions to a base and phosphate and 8 interactions to a base and sugar, usually the 2'OH donor with the O4' also possible. For the ARNA base pairs, there are three interactions involving the minor groove guanine N2 amine of one strand and the sugar on the other strand. Two of the

interactions are a consequence of the 2'OH present in ARNA (see below), and the other involves the O4'. With BDNA there are 16 interactions involving a base and a phosphate, and 18 interactions involving a base and the sugar, usually the O4' with the O3' also possible. There are no backbone interactions to BDNA that span a single step. Interestingly, there are two interactions with two hydrogen bonds to a single base, and one to a backbone atom. Shown in Figure 4b and 4c, one is unique to the ARNA major groove, and the other is unique to the BDNA minor groove.

Turning to the two-step interactions (XY), we see in tables 3 and 4, and in figures 2 and 3, that ARNA and BDNA have similar numbers of interactions. The most common interaction type in both helices is with adjacent steps of the same strand, or base pattern 222. As we will see, this correlates well with the observed frequency of two-step interactions in the PDB. The cross-strand patterns (220 and 221) allow the second most number of interaction patterns. An example of a base-specific interaction with base pattern 220 is shown in Figure 4d. In table 3 and 4, we see that a variety of amino acid contacts can be made to two steps, and as expected, Arg, Asn/Gln make the most interactions due to their geometry and number of donor/acceptor groups. We note that two-step interactions are likely to be more common and can be accommodated in more ways than base-pair spanning interactions. For both BDNA and ARNA, there are over a hundred ways to span across two base steps, while there are only four ways to span across a base pair in a continuous helix. This is the result of the many more arrangements of donors and acceptors possible in two steps than for single base pairs.

In the three step interactions, there are few interactions where only the two base pairs at the end are directly recognized (referred to herein as X_Z). In ARNA, all of

these occur in the major groove and with two hydrogen bonds to bases. This is because of the base-pair tip and helical displacement into the minor groove that results in a bulged surface that makes it impossible to span across the minor groove of ARNA. On the other hand, in BDNA, X_Z interactions occur primarily in the minor groove, and involve the O4'. The O4' is easily accessible in the minor groove, and in fact has an O4' to O4' three-step spanning distance of 7.0Å (see Fig 4e). Hence, most of the interactions are not base-specific. In other words, it is difficult to navigate the array of donor and acceptors along three steps without making at least one contact per step unless the sugar O4' is involved. While no base-specific contacts are made to the middle base pair, the middle base pair can still be recognized either by steric or electrostatic repulsion. One instance of specific recognition of the middle base pair is shown in Figure 5, where the middle base pair can be either AT or TA, but not GC or CG because the N2 amine protrudes into the minor groove and would sterically clash with the arginine guanadinium. We have clustered all three hydrogen bond X_Z-type interactions and schematized the strategy for the data set of moderately restricted hydrogen bonds in Figure 6.

For three-step interactions where all three steps are recognized (XYZ), a minimum of three hydrogen bonds is required for recognition. No XYZ interactions occur in the minor groove of ARNA for the reasons presented above. In the major groove of ARNA, there are 63 possible strategies, all of which are base-specific. One of these is “good”, and it is presented in Figure 6. In BDNA, we find 19 interactions in the major groove, again all base-specific. In the minor groove, 988 interactions are found, of which 98 are base-specific. The “good” interactions are shown in Figure 6. Because of the Watson-Crick nature of duplex helices, the base-specific interactions we found to all

three-steps can be used to specify a triplet sequence using only a single amino acid. The specific sequences are presented in Table 5. Only two base specific interaction are “good”, specifying an ATC or GAT sequence in the minor groove of BDNA, and UGU or ACA in the major groove of ARNA (see figure 6). In general, most interactions in BDNA involving all three steps involve the O4' and are not base-specific, while all ARNA three step interactions are base-specific.

In summary, for all three-step interactions (X_Z or XYZ), we found it is only possible to make interactions to ARNA in the major groove (table 2). In BDNA, 98% of the interaction types are found in the minor groove, largely due to the favorable orientation and positioning of the sugar O4'.

Backbone interactions

With regards to backbone interactions, our modeling indicates that phosphate interactions generally occur in the major groove, while sugar interactions can only occur in the minor groove (see Table 2). Of the backbone hydrogen-bonded interactions in the major groove, two-thirds are to the phosphate oxygen O2P, and the other third is to the O5', which is the phosphate oxygen connected to the 5' carbon of the sugar. The numbers of models are distributed similarly in both ARNA and BDNA.

The major groove of ARNA places the phosphate backbone closer to the bases, because the bases are displaced from the helix axis, and thus allows more diverse interactions that can simultaneously recognize one base and the backbone. This holds true for interactions to base-pairs and two steps, but does not hold true for interactions to three steps in ARNA. It is impossible to make a three-step interaction in ARNA

involving any backbone atom, which is result of the wider helix, resulting in the inability to recognize three steps of ARNA unless interactions are made to the bases, in an orientation perpendicular to the base pairs.

In the minor groove, there are many more interactions possible in BDNA due in large part to the possibility of hydrogen bonding to the sugar O4' while hydrogen bonded to the base. The acceptor O4' is angled towards the minor groove in BDNA, whereas it is directed along the helix backbone in ARNA. This results in the O4' being able to make geometrically moderate hydrogen bonds while recognizing a base in BDNA in many arrangements (see Figure 9 for examples), while in ARNA the O4' can only make geometrically moderate hydrogen bonds with the adjacent base 5' to it, and possibly with its adjoining base. In terms of base-patterns (Figure 3, 4) for the “good” models, six base-patterns allow use of the O4' in BDNA, while only one base-pattern (pattern 222) allows use of the O4' in ARNA (Fig 4f). A three-step interaction involving the O4' is not possible in ARNA, while it is involved in 90% of tri-step interactions in BDNA. In BDNA the O4' distances are shorter across the 3' ends of the helix, and in fact a three step spanning interaction spans the equivalent rise of a bit larger than two steps, as shown in Figure 4e. Interactions to BDNA, but not to ARNA, can also involve the O3' which connects the sugar to the phosphate on the 3' end of the residue. However, the majority of interactions involve the O4' in BDNA. Thus in the BDNA minor groove, the O4' of the sugar provides a convenient and structure-specific recognition feature, and is a dominant strategy in the formation of three-step interactions. Interactions to the O3' are also unique to BDNA.

In ARNA, the extra 2' OH provides a structure-specific recognition feature, although it is generally used as a donor, as opposed to the O4' being used as an acceptor. We show all interaction strategies found involving a 2'OH in Figure 7. As is seen there, the use of the 2'OH is largely dominated by Asn/Gln (see Fig 4g for an example), although Arg can also use the O2' in limited orientations. The only "good" interactions possible with the 2'OH involve the adjacent base (base-pattern 220) and the adjoining base (base pattern 110, not shown in Figure 2). One additional base pattern (base-pattern 221) is possible if we consider all interactions. Ser/Thr/Tyr and Asp/Glu can only make interactions with hydrogen bonds to the adjoining base, and only Asn/Gln and Arg can make interactions that involve more than one step. There are no three step interactions that involve the ribose 2'OH. Many of the Asn/Gln and all of the Arg interactions involve the O2' acceptor instead of the 2'OH donor. Because of the strong directionality of hydrogen donors (Versichel *et al.*, 1984; Jeffrey & Saenger, 1991), and the orientation of the O2', interactions involving the 2'OH are limited for ARNA.

In general the lack of a good hydrogen bonding strategy for multiple steps in ARNA, and the good orientation of the O4' in BDNA allows for unique recognition of multiple steps in BDNA by minor groove hydrogen bonding. However, for a single base, the O4' and 2'OH can provide a structure specific readout in the minor groove of BDNA or the major groove of ARNA.

Bifurcated interactions

While bifurcated interactions are generally not discussed in the literature, we would like to point out that while most bifurcated interactions are variants of non-

bifurcated interactions, some bifurcated interactions are unique. In ARNA, all major groove interactions using base pattern 230, and all interactions using base patterns 240 and 336 (Fig 2) are bifurcated. There are 11, 2, and 15 interactions, respectively, possible with these base patterns. These interactions, however, are not “good” and they all drop out when we place moderate restrictions on the hydrogen bond geometry. These interactions require bifurcated bonds because either the side chain sits parallel to the base plane and in the middle of a base step (for base patterns 230 and 240), or the side-chain is oriented perpendicular to the base plane along the groove (see Fig 4h). Inclusion of bifurcated hydrogen bonds allows for a maximum of 5 hydrogen bonds instead of the 4 with non-bifurcated hydrogen bond patterns. In general the bifurcated interaction has poor hydrogen bond geometry. In fact, out of the 327 interactions found for ARNA that have “good” hydrogen bonds, only 19 interactions are bifurcated. In the set of nearly ideal interactions, there are no bifurcated interactions.

Unique strategies for ARNA and BDNA recognition

As mentioned before, we have calculated a database of differences in interactions to ARNA versus BDNA. We did this by taking the databases of all interactions, and then subtracting from each database the interactions common between the two canonical helices. This serves two purposes. It allows us to find interactions in DNA that are absolutely sensitive to conversion from B-form to A-form. These interactions are presented in Figures 8 and 9, and do not include ribose 2'OH interactions. In another respect, the difference database allows us to look at interactions that can uniquely differentiate ARNA and BDNA in the context of a small peptide. For small peptides to

bind specifically to particular sequences, they may need to maximize the specificity derived from every amino acid member. By making unique interactions, the peptide can not only potentially differentiate between ARNA and the plethora of BDNA in the cell, but also do so in a partially sequence-specific manner, as we shall see.

Many of the readouts we found have some degeneracy, and hence are not absolutely specific to a unique sequence, but most do present some specificity. One example of sequence specificity in the minor groove of BDNA, however, can be seen in Figure 9 in a two-step interaction involving Arg making three hydrogen bonds to AdeN3/ThyO2. The specificity for the two steps of AT over a GC or CG step can be attributed to the presence of a N2 amine in G, which provides steric bulk as well as polar groups that can clash with the guanidinium group of Arg.

We will only discuss the two-step interactions here, because, as discussed above, the majority of three step interactions are unique to their respective canonical helix forms. For the "moderately"-restricted hydrogen bonds, all three-step interactions are unique to their respective canonical helix forms. We note that this result is in line with the idea that multiple hydrogen bonds confer increased specificity.

For two-step interactions, the results are different. The numbers of interactions are comparable in ARNA and BDNA (Tables 2 and 3). We calculated a difference database between the two forms, and we present the best strategies in Figures 8 and 9. We arrived at these strategies by only including interactions that are "good" or "nearly ideal" with respect to hydrogen bond quality, and requiring that an interaction be present in at least two of the three sets. The later requirement provides an approximate hydrogen bonding energy "gap", as described in the methods.

The difference database presented in Figures 8 and 9 allow us to get a sense for the strategies available in ARNA and BDNA, as well as get a sense for how the strategies differ in the two nucleic helices. By using the particular strategies shown in Figures 8 and 9, we can gain some sequence specificity in recognition. All interactions shown are either same strand (base pattern 222), 5' cross strand (base pattern 220), or 3' cross strand (base pattern 221) interactions. As we will see, these are also the only base-specific two-step interactions observed so far in the PDB. Of the interaction types shown, a third to a half are base-specific (also see table 3 and 4).

One interesting interaction that is particularly specific for A-form RNA is the His⁺ interaction (see Figure 8 and Figure 4i and 4j). This interaction recognizes both the absence of the 5'methyl on uracil, as well as the structural arrangement of the backbone using a hydrogen bond to the phosphate oxygen. There are two possible interactions types, one involving the O5' of the phosphate, and one involving the O2P of the phosphate. The first is more general, because it allows any base at the position adjoining the phosphate recognized. In the later case, a purine in the position adjoining the O2P phosphate atom provides a more favorable hydrogen bonding geometry because the O2 bulge does not clash with the histidine imidazole.

Comparison with observed interactions

In 2001, Luscombe *et al.* (Luscombe *et al.*, 2001), published a survey of observed side chain-nucleic acid interactions which included a survey of base-specific interactions to multiple steps of DNA. We found half the interactions between side chain and BDNAs found in the study to be unique to BDNA over ARNA in our study, and all

observed interactions were also found in our databases. This result could be expected purely because roughly a third of all interactions that we found turn out to be unique to a canonical helix form. For bent helices such as the TATA binding complex (Juo *et al.*, 1996), we might not expect this to be the case, however, none of the observed base-specific interactions occur in complexes involving bent DNA. A summary of the interactions found is shown in Figure 10. For ARNAs, we performed a PDB search for base-specific interactions, and visually inspected candidate structures for Watson-Crick base-pairing and duplex structure. The two interactions we found (shown in Figure 10) are both unique to ARNA. Mutagenesis data is required to validate the importance of the interactions we highlight, although in general we would expect multiply hydrogen bonded amino acids to be essential for specific recognition.

We also identified interactions in the PDB involving the O4' or O2', and at least one base atom. We have only included DNA interactions that are B-form, while for RNA we separated interactions to A-form RNA and non-canonical RNA.

Interestingly, all RNA interactions we identified involving the sugar occur in non-canonical, non-A-form regions (Tables 6 and 7). In all cases for RNA, O4' and O2' stepping interactions (222) involve the sugar and the base 5' adjacent to it, as found with our models. With the O2' RNA interactions there is one interaction (indicated by a *) that involves nearly A-form RNA. This interaction involves Watson-Crick base-paired nucleotides, but is in a highly-twisted helix in the 30S ribosome, involving a 5' cross-strand (220) hydrogen bond pattern between Asn73, A737 O2' acceptor, and G670 N2 donor. The interaction is not present in our modeled dataset, and the closest interaction to the one found involves the same bases and atom types, but in a stepping fashion (base

pattern 222, Fig 9) instead of a cross-strand. The high twist allows the G670 N2 to be placed approximately where the N2 is placed in our model. While the number of interactions observed is small, it is striking that so far the O2' is involved in sequence-specific recognition only in non-canonical regions.

For BDNA, all interactions occurred in the minor groove as expected, and 60% of interactions involved single bases, with the remaining 40% involving adjacent bases on base steps (base pattern 222), labeled “steps” in Table 6. This is in line with expectations from our modeling. There are insufficient numbers of interactions to do a significant validation of our predictions. The B-form DNA results support our modeling, although we expect that as more diverse structures are solved we will find the variety of non-ideal B-form will contribute to differences, as seen for ARNA.

CONCLUSION

We have discussed amino acid side chain recognition of nucleic acids in the context of idealized duplex RNA and DNA, and we have presented several strategies that are unique to either ARNA or BDNA. Although many hydrogen-bonding networks involve more than a single amino acid, a single amino acid when used strategically can aid specific or partially specific recognition. The presence of a multiply hydrogen bonded protein side-chain may be an intricate design of nature, but there is a limited number of strategies available. We have attempted to model and discuss them here. Further studies will need to address the diversity of non-idealized BDNA helices, and how they affect the strategies presented here.

ACKNOWLEDGEMENTS

We thank Peter Kollman, David Agard, Wendell Lim, Steve Landt, and other members of the Frankel Lab for helpful discussions. We thank David Case for his help in use of NAB, and Cynthia Fuhrmann for help developing the PDB search used here. This work was supported by NIH grants GM47478 and GM56531 (A.D.F.) and by NIH Training grants GM08284 and GM08388 (A.C.C.).

References

Allers, J & Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.* **311**, 75-86.

Amott, S., Hukins, D. W. L.; Dover, S. D.; Fuller, W. & Hodgson, A. R. (1973). Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations. X-ray diffraction analyses of the molecule conformations of (polyadenylic acid) and (polyinosinic acid).(polycytidylic acid). *J.Mol. Biol.* **81**, 107-22.

Antson, A. A. (2000). Single-stranded-RNA binding proteins. *Curr Opin Struct Biol.* **10**, 87-94.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1997). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Blackburn, G. M. & Gait, M. J. (1996). *Nucleic acids in chemistry and biology*. Oxford University Press, New York.

Cheng, A. C., Chen, W., Fuhrmann, C. & Frankel, A. D. Specificity in hydrogen-bonding of protein side-chains to simple units of RNA structure. Submitted to *J. Mol. Biol.*

Cornell, W. D. & Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.

Draper, D. E. (1999). Themes in RNA-protein recognition. *J Mol Biol.* **293**, 255-270.

Frankel, A. D. (2000). Fitting peptides into the RNA world. *Curr. Opin. Struct. Biol.* 2000 **10**, 332-340.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in biological molecules*. Springer-Verlag.

Jones, S., Daley D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. (2001). Protein-RNA interactions: a structural analysis. *Nucl. Acids Res.* **29**, 943-54.

Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J. & Dickerson, R. E. (1996). How proteins recognize the TATA box. *J. Mol. Biol.* **261**, 239-254.

Lu, X. J. & Olson, W. K. (2001). *3DNA v1.2, a 3-dimensional nucleic acid structural analysis and rebuilding software package*. Rutgers University.

Luscombe NM, Laskowski RA & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860-2874.

Macke T. & Case, D.A. (1998). Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids*, American Chemical Society, Washington, DC., 379-393.

Nadassy, K., Wodak, S. J., & Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999-2017.

Pabo, C. & Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597-624.

Ryter, J. M. & Schultz, S. C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17**, 7505-13.

Saenger, W., *Principles of Nucleic Acid Structure*. Springer Verlag, New York, 1984.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci.* **73**, 804-8.

Suzuki, M. (1994). A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317-26.

Taylor, R., Kennard, O., & Versichel, W. (1983). Geometry of the imino-carbonyl (N-H...O:C) hydrogen bond. 1. Lone-pair directionality. *J. Am. Chem. Soc.* **105**, 5761-5766.

Taylor, R., Kennard, O. & Versichel, W. (1984). Geometry of the N-H...O=C hydrogen bond. 2. Three-center ("Bifurcated") and four-center ("Trifurcated") bonds. *J. Am. Chem. Soc.* **106**, 244-248.

Taylor, R., Kennard, O. & Versichel, W. (1984). The Geometry of the N-H...O=C hydrogen bond. 3. Hydrogen-bond distances and angles. *Acta Cryst. B*, **40**, 280-288.

Treger, M. & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol. Recognition*. **14**, 199-214.

Walberer, B. J., Cheng, A. C. & Frankel A. D. Diversity of hydrogen-bonded base interactions in nucleic acids. Submitted to *J. Mol. Biol.*

Weeks, K. M. & Crothers, D. M. (1993). Major groove accessibility of RNA. *Science*, **261**, 1574-7.

FIGURE LEGENDS

Figure 1. Overview of A-form and B-form nucleic acids. (A) Molecular representation of idealized A-form and B-form helices. Note ARNA's characteristic enlarged helix, greater tilt, and smaller distance between base steps compared to BDNA. (B) Helical parameters that can result in different recognition strategies for ARNA and BDNA. (C) Donor and acceptor arrays in the major and minor grooves. The values and canonical helices presented in (A) and (B) are from the model structures used in this study. The local helical parameters given are the averages calculated using 3DNA (Lu and Olson, 1999; Lu et al, 1999).

Figure 2. A form RNA base patterns. Interaction counts are broken out by major and minor groove interactions, and are also broken out by all interactions, “all”, and by base-specific interactions (see text), “sp”. Base patterns with zero total interaction counts are grayed out.

Figure 3. B form DNA base patterns. Interaction counts are broken out by major and minor groove interactions, and are also broken out by all interactions, “all”, and by base-specific interactions (see text), “sp”. Base patterns with zero total interaction counts are grayed out.

Figure 4. Sampling of modeled interactions. See the text for discussions of these interactions. (A) The geometrically-best example of a 3' cross strand interaction (base pattern 221) to the major groove of ARNA. This interaction involves a base-specific arginine interaction to two guanines. (B) Asn recognizing a single base in the BDNA

minor groove with three hydrogen bonds including one to the O4' (base pattern 110). (C) Asn recognizing a single base in the ARNA major groove with three hydrogen bonds including one to the phosphate O2P (base pattern 110). (D) An example of a base-specific 5' cross-strand interaction (base pattern 220) in the major groove of ARNA. This interaction involves arginine making hydrogen bonds to two guanines. (E) BDNA minor groove interaction involving the sugar O4' and arg. Arginine makes hydrogen bonds utilizing a 3' cross strand interaction (base pattern 336) between sugar O4' atoms across three steps. (F) ARNA minor groove interaction involving the sugar O4' and Arg (base pattern 222). The position of the 3' cross strand sugar O4' is highlighted in blue. (G) Asn interaction to ARNA involving the ribose 2'OH and O4' in the minor groove (base pattern 222). (H) Asp interaction to the ARNA major groove in a 336 base pattern involving bifurcated hydrogen bonds. (I) Top view of a potentially very specific three hydrogen bond bifurcated His⁺ interaction to uracil and the phosphate on the nucleotide 5' to the uracil (base pattern 222). (J) Side view of the interaction in I.

Figure 5. Example of an interaction of the type X_Z, where the middle base pair is not directly contacted but is partially sequence specific. The G-C base pair is not possible in the middle position because of the steric bulk presented by the N2 amine in the minor groove.

Figure 6. Three hydrogen bond interactions in ARNA and BDNA, divided by X_Z and XYZ type interactions. Base-specific interactions are labeled. Hydrogen bonds are

shown in green, and the base pattern for the interaction is shown in parentheses. Any indicates any base is possible at the position.

Figure 7. All non-bifurcated hydrogen bonding interaction models involving the ribose 2'OH donor or O2' acceptor, broken out by base pattern.

Figure 8. ARNA difference figure. Schematic representation of feasible strategies for distinguishing idealized ARNA helices from idealized BDNA helices. The base pattern of each interaction is indicated on the left side, and its appearance in the three levels of hydrogen bonding quality is indicated on the right using "X"s.

Figure 9. BDNA difference figure. Schematic representation of feasible strategies for distinguishing idealized BDNA helices from idealized ARNA helices. The base pattern of each interaction is indicated on the left side, and its appearance in the three levels of hydrogen bonding quality is indicated on the right using "X"s.

Figure 10. Summary of observed base-specific interactions involving more than one step. (A) schematic representations of interactions found in duplex RNA helices. (B) Summary of observed interactions found in duplex DNA helices by Luscombe *et al.* The starred interaction was placed 3' to 5' in the original paper, and we have corrected the positioning here.

Table 4.1. Number of interactions for ARNA and BDNA found after each filter step

	ARNA	BDNA
WASABI output	41,496	48,242
remove target redundancy	3,499	7,259
remove amino acid redundancy	1,859	4,240
remove Tyr/Ser redundants	1,815	4,137
allow only fixed interactions	1,720	3,803
remove all bifurcated hbonds	970	1,990
processed database	970	1,990
remove single step interactions	908	1,944
allow only base specific interactions	409	517
use moderately restrictive hbond parameters	93	147
use stringently restrictive hbond parameters	20	26

Table 4.2. Number of interactions to phosphate and sugars in ARNA and BDNA, broken out by the four types of base patterns.

Backbone interactions		ARNA		BDNA	
		major	minor	major	minor
Single	Phosphate + Sugar	0	0	0	0
	Phosphate involved	19	0	12	4
	Sugar involved	0	21	0	18
	Neither	12	10	8	4
XY	Phosphate + Sugar	0	4	0	0
	Phosphate involved	307	0	137	24
	Sugar involved	0	225	0	232
	Neither	178	79	154	101
X_Z	Phosphate + Sugar	0	0	0	32
	Phosphate involved	0	0	0	0
	Sugar involved	0	0	0	204
	Neither	52	0	11	42
XYZ	Phosphate + Sugar	0	0	0	0
	Phosphate involved	0	0	0	0
	Sugar involved	0	0	0	914
	Neither	63	0	19	74
Total		631	339	341	1649

Table 4.3. Characteristics of modeled interactions to ARNA

a. A-form RNA Major Groove All interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	1	161	323	8	17	29	35	115	269	12	0
	moderate	0	10	148	4	4	14	12	32	86	6	0
	stringent	0	0	47	1	4	0	0	9	29	4	0
X_Z	none	0	0	52	2	6	7	0	16	21	0	0
	moderate	0	0	4	0	0	2	0	2	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	0	63	0	0	0	0	0	9	54	0	0
	moderate	0	1	0	0	0	0	0	0	1	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

b. A-form RNA Minor Groove All interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	0	44	264	10	24	12	19	93	142	8	0
	moderate	0	4	98	0	0	0	19	42	37	4	0
	stringent	0	0	21	0	0	0	11	7	3	0	0
X_Z	none	0	0	0	0	0	0	0	0	0	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	0	0	0	0	0	0	0	0	0	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

Table 4.3 (continued)

c. A-form RNA Major Groove Base specific interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	1	37	161	8	5	5	19	67	83	12	0
	moderate	0	2	67	4	0	0	12	20	27	6	0
	stringent	0	0	11	1	0	0	0	5	1	4	0
X_Z	none	0	0	52	2	6	7	0	16	21	0	0
	moderate	0	0	4	0	0	2	0	2	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	0	63	0	0	0	0	0	9	54	0	0
	moderate	0	1	0	0	0	0	0	0	1	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

d. A-form RNA Minor Groove Base specific interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	0	16	79	2	0	0	3	30	56	4	0
	moderate	0	4	15	0	0	0	3	11	5	0	0
	stringent	0	0	9	0	0	0	3	3	3	0	0
X_Z	none	0	0	0	0	0	0	0	0	0	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	0	0	0	0	0	0	0	0	0	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

Table 4.4. Characteristics of modeled interactions to BDNA

a. B-form DNA Major groove All interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	0	67	224	7	11	16	15	72	161	9	0
	moderate	0	2	89	5	4	9	11	25	37	0	0
	stringent	0	0	11	0	3	0	3	5	0	0	0
X_Z	none	0	1	10	0	0	3	0	1	7	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	1	18	0	0	0	0	0	3	16	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

b. B-form DNA Minor groove All interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	0	106	251	2	8	10	50	86	185	16	0
	moderate	0	1	156	1	0	4	30	41	75	6	0
	stringent	0	0	34	0	0	0	9	9	12	4	0
X_Z	none	12	102	164	0	10	30	4	34	196	4	0
	moderate	0	15	73	0	8	0	0	16	64	0	0
	stringent	0	0	36	0	0	0	0	4	32	0	0
XYZ	none	158	830	0	0	0	0	72	68	848	0	0
	moderate	0	21	0	0	0	0	0	8	13	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

Table 4.4 (continued)

c. B-form DNA Major groove Base specific interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	0	22	141	7	4	2	15	53	73	9	0
	moderate	0	1	64	5	0	0	11	21	28	0	0
	stringent	0	0	8	0	0	0	3	5	0	0	0
X_Z	none	0	1	10	0	0	3	0	1	7	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	1	18	0	0	0	0	0	3	16	0	0
	moderate	0	0	0	0	0	0	0	0	0	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

d. B-form DNA Minor groove Base specific interactions

Base Pattern	Parameter Restriction	# of Hbonds			Side chain							
		4	3	2	D/E	H	H+	K	N/Q	R	S/T/Y	W
XY	none	0	42	101	2	0	0	18	30	85	8	0
	moderate	0	1	72	1	0	0	14	21	35	2	0
	stringent	0	0	18	0	0	0	9	9	0	0	0
X_Z	none	12	29	42	0	2	10	0	10	61	0	0
	moderate	0	0	8	0	0	0	0	0	8	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0
XYZ	none	24	74	0	0	0	0	0	4	94	0	0
	moderate	0	1	0	0	0	0	0	0	1	0	0
	stringent	0	0	0	0	0	0	0	0	0	0	0

Table 4.5.

A. Base specific amino acid recognition of X_Z triplet sequences. Base-specific interaction models in ARNA (A) and BDNA (B) are broken out by amino acid across the top, and base pattern and sequence composition across the bottom. Each sequence also represents its complement. For instance, A_C represents the sequences that has AAC, ACC, AGC, ATC.

Triplet Sequence	Arg		Asn/Gln		Asp/Glu		His		His+		Lys		Ser/Thr/Tyr	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B
A_A/T_T	3	8	1	1	0	0	0	0	1	2	0	0	0	0
A_G/C_T	0	8	0	4	0	0	0	1	0	1	0	0	0	0
A_C/G_T	4	7	2	0	0	0	1	0	1	1	0	0	0	0
X_Z A_T/A_T	2	4	1	0	0	0	0	0	1	1	0	0	0	0
C_A/T_G	3	7	1	1	0	0	0	0	1	1	0	0	0	0
C_C/G_G	2	8	1	4	0	0	0	1	0	1	0	0	0	0
C_G/C_G	0	5	0	1	0	0	0	0	0	1	0	0	0	0
G_A/T_C	3	11	4	0	1	0	2	0	1	2	0	0	0	0
G_C/G_C	2	5	4	0	0	0	2	0	1	1	0	0	0	0
T_A/T_A	2	5	2	0	1	0	1	0	1	2	0	0	0	0



Table 4.5 (continued)

B. Base specific amino acid recognition of XYZ triplets in ARNA (A) and BDNA (B).

Both the sequence and the complementary sequence are listed.

XYZ	Triplet Sequence	Arg		Asn/Gln		Asp/Glu		His		His+		Lys		Ser/Thr/Tyr	
		A	B	A	B	A	B	A	B	A	B	A	B	A	B
	AAA / TTT	0	7	0	1	0	0	0	0	0	0	0	0	0	0
	AAC / GTT	0	8	0	0	0	0	0	0	0	0	0	0	0	0
	AAG / CTT	0	5	0	1	0	0	0	0	0	0	0	0	0	0
	AAT / ATT	1	8	0	0	0	0	0	0	0	0	0	0	0	0
	ACA / TGT	5	0	0	0	0	0	0	0	0	0	0	0	0	0
	ACC / GGT	5	0	0	0	0	0	0	0	0	0	0	0	0	0
	ACG / CGT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ACT / AGT	2	0	1	0	0	0	0	0	0	0	0	0	0	0
	AGA / TCT	0	1	0	1	0	0	0	0	0	0	0	0	0	0
	AGC / GCT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	AGG / CCT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ATA / TAT	2	8	0	1	0	0	0	0	0	0	0	0	0	0
	ATC / GAT	1	8	0	0	0	0	0	0	0	0	0	0	0	0
	ATG / CAT	0	6	0	1	0	0	0	0	0	0	0	0	0	0
	CAA / TTG	2	4	0	0	0	0	0	0	0	0	0	0	0	0
	CAC / GTG	0	6	0	1	0	0	0	0	0	0	0	0	0	0
	CAG / CTG	0	4	0	0	0	0	0	0	0	0	0	0	0	0
	CCA / TGG	7	0	0	0	0	0	0	0	0	0	0	0	0	0
	CCC / GGG	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CCG / CGG	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CGA / TCG	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	CGC / GCG	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	CTA / TAG	3	3	0	0	0	0	0	0	0	0	0	0	0	0
	GAA / TTC	2	10	0	0	0	0	0	0	0	0	0	0	0	0
	GAC / GTC	2	8	0	0	0	0	0	0	0	0	0	0	0	0
	GAG / CTC	0	5	0	1	0	0	0	0	0	0	0	0	0	0
	GCA / TGC	2	1	2	0	0	0	0	0	0	0	0	0	0	0
	GCC / GGC	2	0	1	0	0	0	0	0	0	0	0	0	0	0
	GGA / TCC	2	2	1	0	0	0	0	0	0	0	0	0	0	0
	GTA / TAC	4	8	1	0	0	0	0	0	0	0	0	0	0	0
	TAA / TTA	5	6	1	0	0	0	0	0	0	0	0	0	0	0
	TCA / TGA	4	2	2	0	0	0	0	0	0	0	0	0	0	0

Table 4.6. Interactions from the PDB involving an O4' and at least one base atom

**DNA O4' bidentate interactions
(B-form DNA)**

AA	Base	PDB Id	Intrxn Type	Groove
Lys360	A5, C6	1ign	steps	minor
Lys394	C19, C20	1t7p	steps	minor
Lys800	T108, G109	1ig9	steps	minor
Lys201	G15	4crx	single	minor
Lys116	C706	1g38	single	minor
Arg131	G8, A9	1dc1	steps	minor
Arg131	G7, G8	1dc1	steps	minor
Arg400	A24, A25	1jey	steps	minor
Arg203	T12	1tc3	single	minor
Arg59	A14	1hwt	single	minor
Arg105	T7	1ig7	single	minor
Arg57	T12	2hap	single	minor
Arg5	T211	3hdd	single	minor
Arg5	T518	9ant	single	minor
Arg243	T17	1crx, 1f44	single	minor
Arg305	T201	1jgg	single	minor
Arg243	T33	3crx	single	minor
Asn420	G1006	1qsl	steps	minor
Asn35	C908, G909	1qum	steps	minor
Gln63	C9	1fiu	single	minor

**RNA O4' bidentate interactions
(non A-form)**

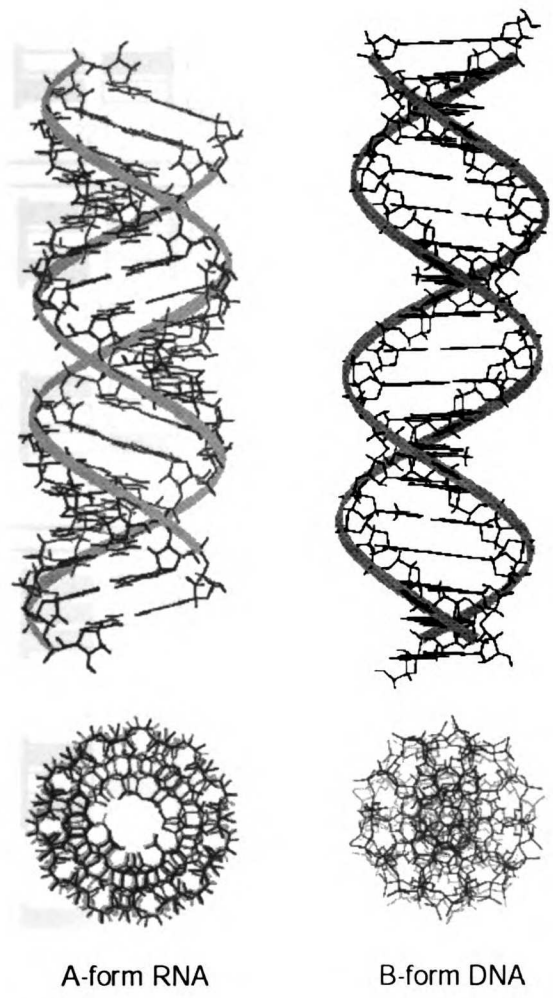
AA	Base	PDB Id	Intrxn Type	Groove
Tyr205	G1, C2	1qf6	steps	minor
Arg570	C911, U912	1gax	steps	minor
Arg435	C534, U535	1g59	steps	minor
Arg570	C924, U912	1gax	spans	minor
Gln121	G634, C638	1asz	spans	minor

Table 4.7. Interactions found in the PDB involving an O2' and at least one base atom.

The interaction indicated by a '*' is discussed in the text.

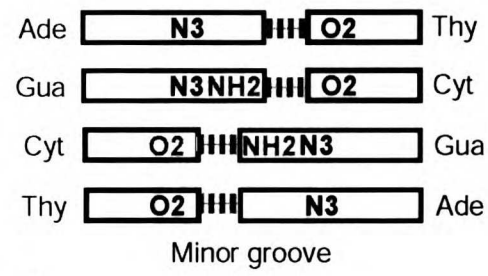
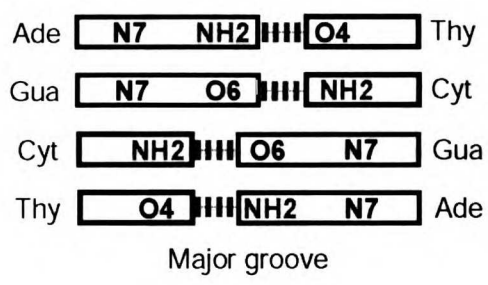
RNA O2' bidentate interactions (not A-form)

AA	Base	PDB Id	Intrxn Type	Groove
Ser151	G30, U39	1dk1	spans	minor
Ser51	G666, U740	1g1x	spans	minor
Ser35	A102	1c9s	single	minor
Thr3	C877	1fjg, 1hnz, 1libk, 1libm	single	minor
Lys110	A9	1jbt	single	minor
Lys61	C13	1zdk	single	minor
Arg412	C934	1qtq	steps	minor
Arg25	C1192	1libm, 1hnw	steps	minor
Arg10	U1376, U1346	1hnw, 1hnx	spans	major
Arg53	A140, C163	1hq1	spans	minor
Arg25	C1192, U1070, C1069	1hnz	spans	minor
Arg16	G1127, C1147	1hnz	spans	minor
Arg25	C1069, C1192	1hnx, 1libk	spans	minor
Arg609	U36	1qf6	single	minor
Asn73 *	A737, G670	1fjg, 1libk	spans	minor
Asn15	C875, C826	1fjg, 1hnw, 1hnx, 1libl, 1libm	spans	minor



	A-form RNA	B-form DNA
basepairs/turn	11	10
twist	32.7	36.0
rise	2.81A	3.38A
tilt	16.0	-5.0
shift	-13.7	0.0
propeller twist	-4.62A	0.13A

B



C

Figure 4.1



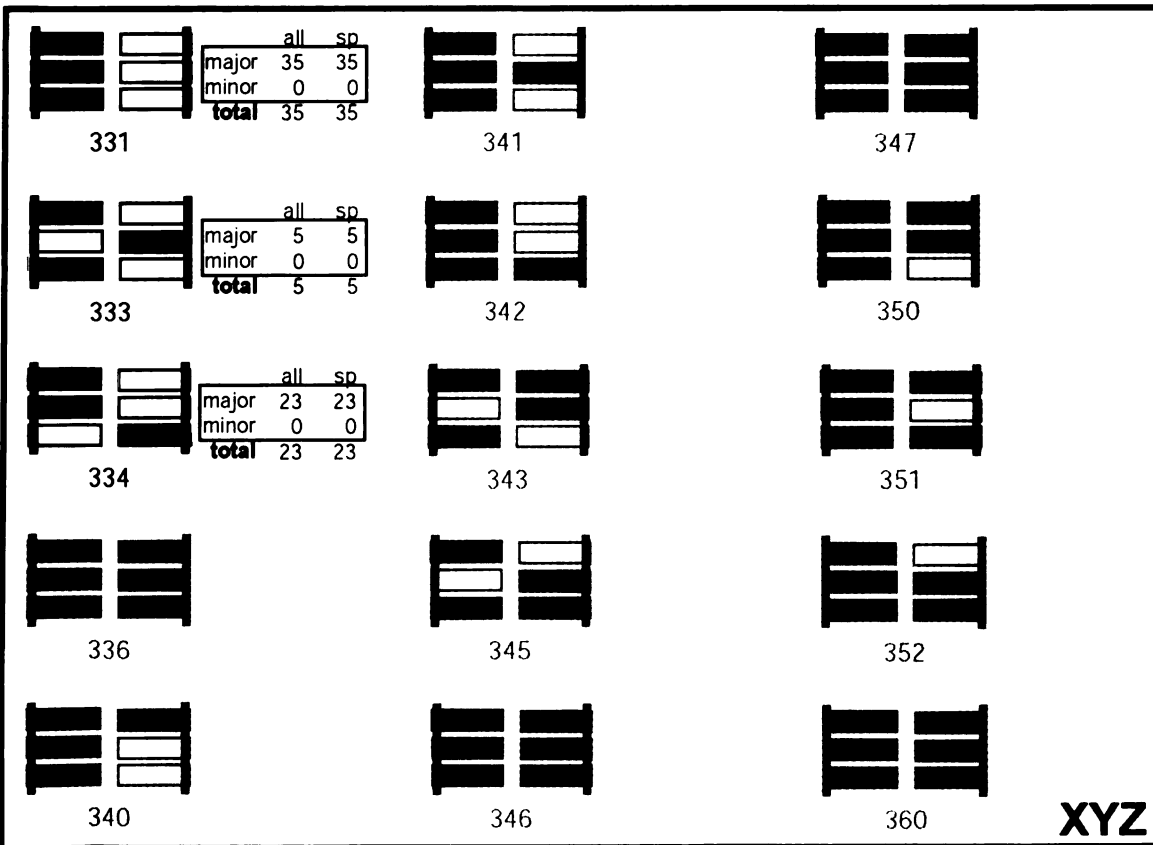
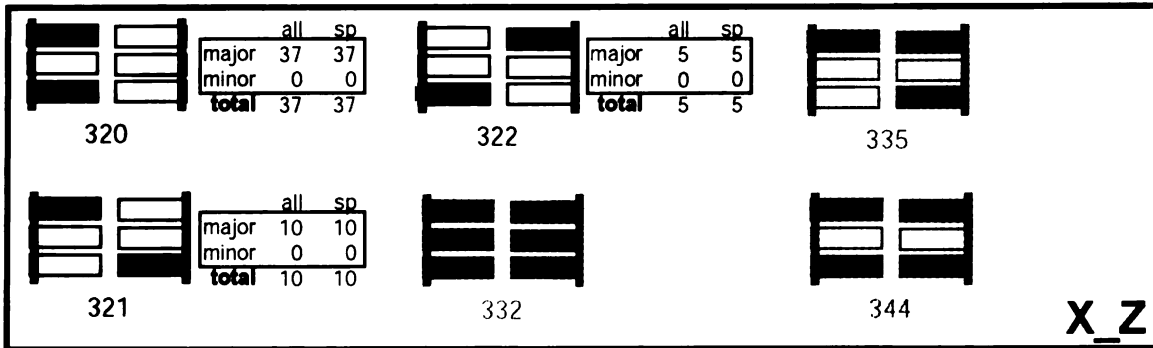
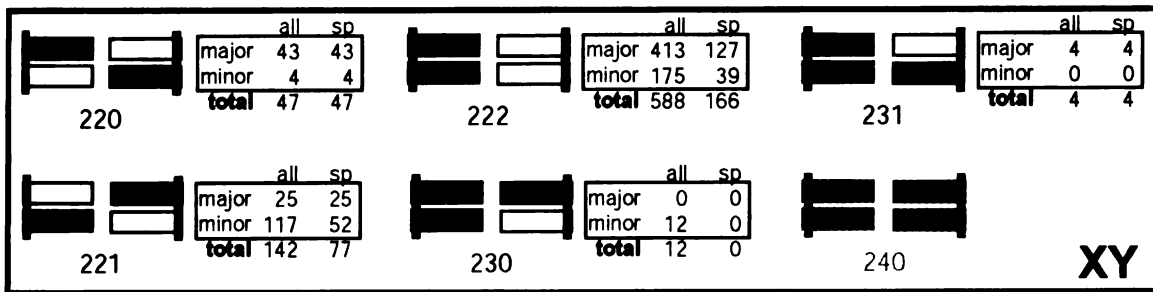


Figure 4.2



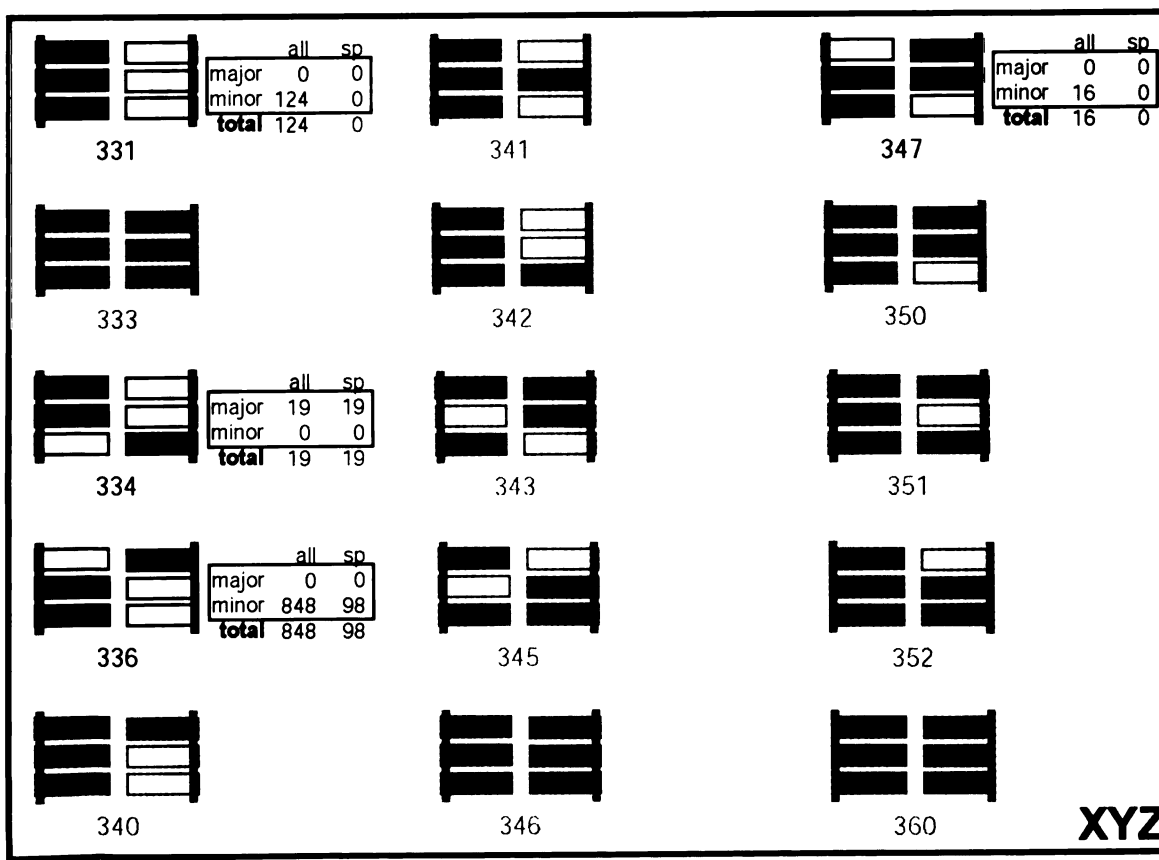
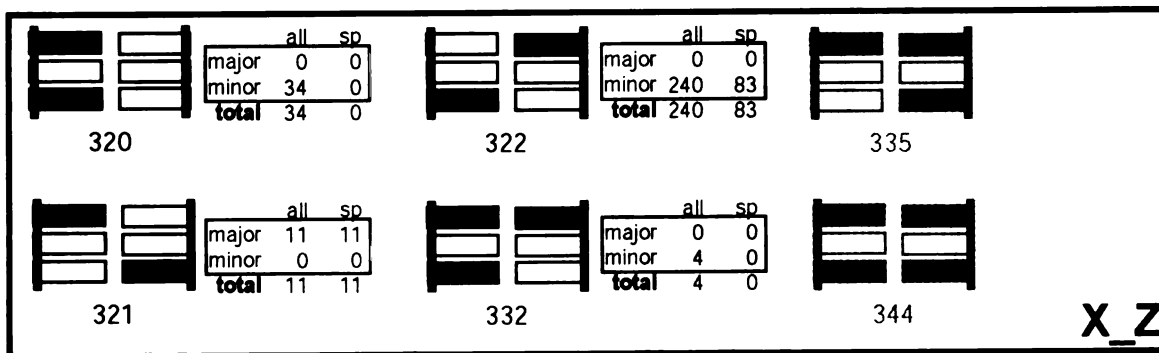
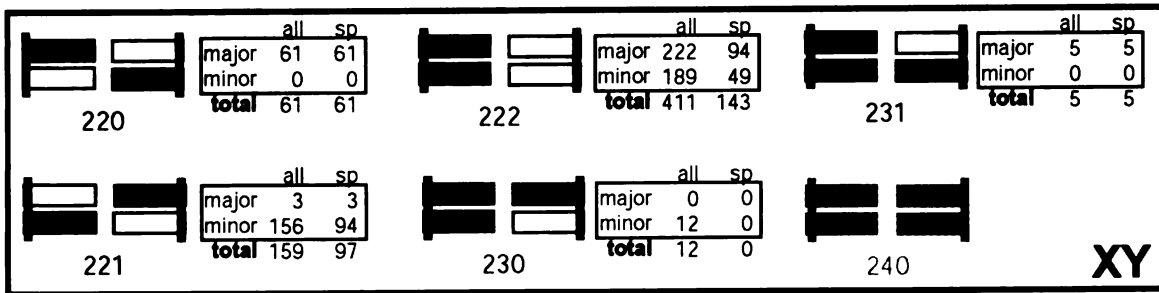


Figure 4.3



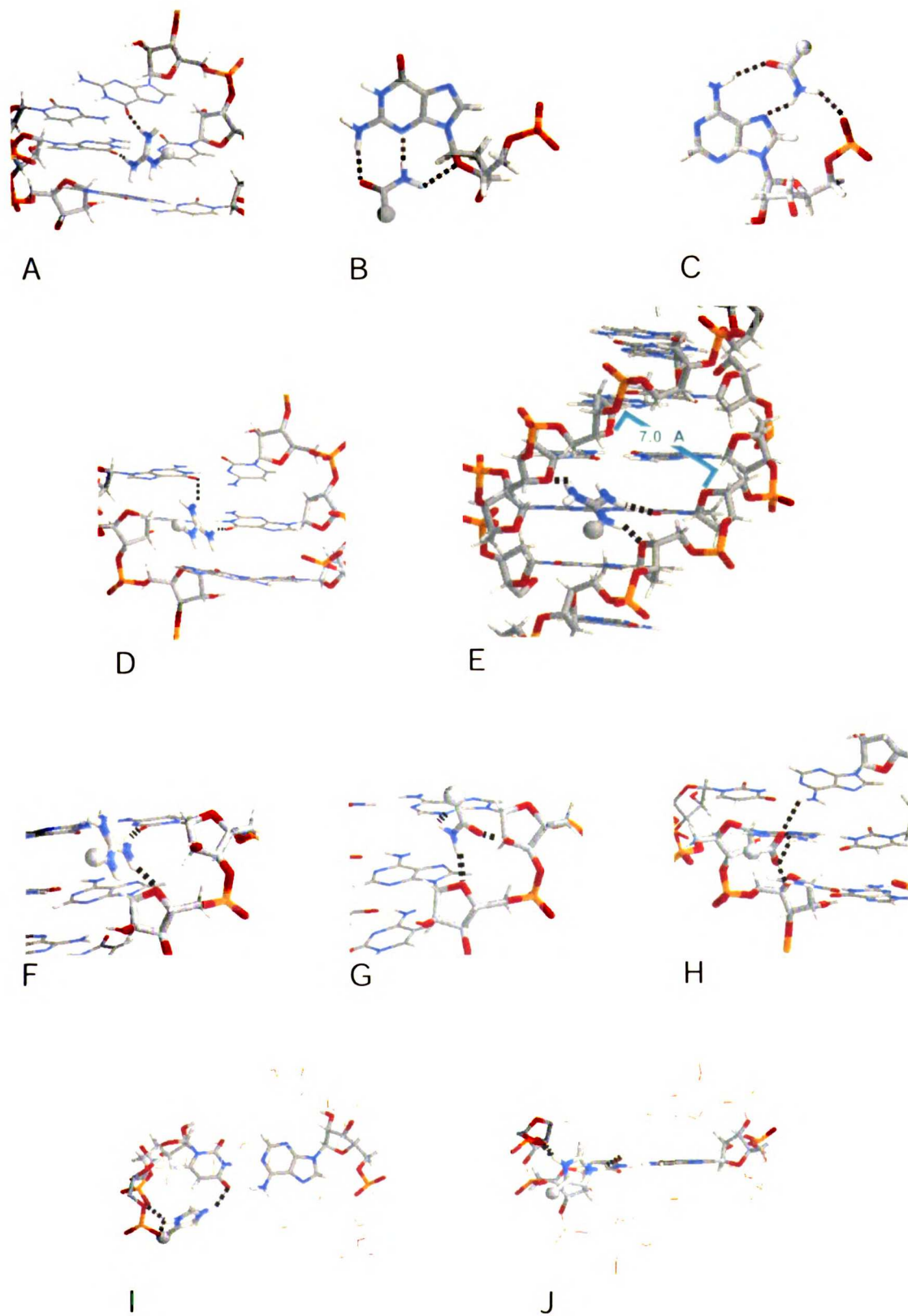


Figure 4.4



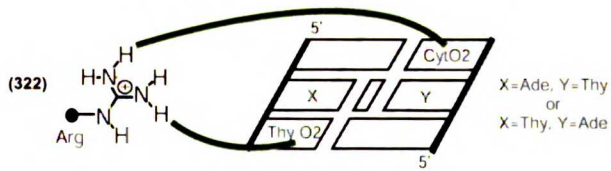
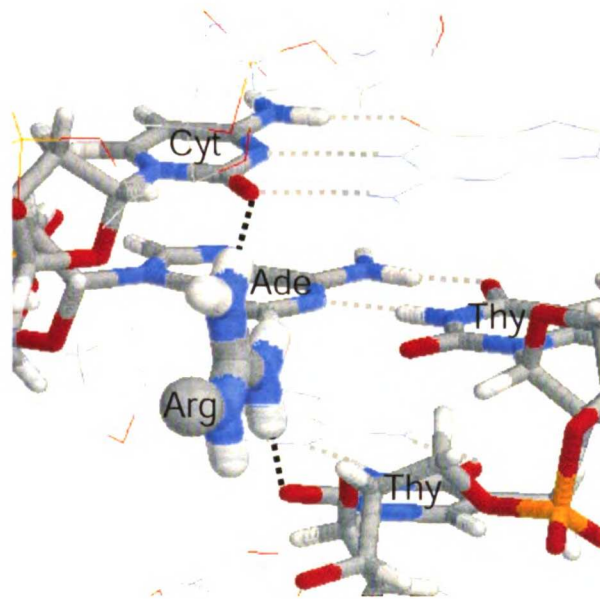


Figure 4.5



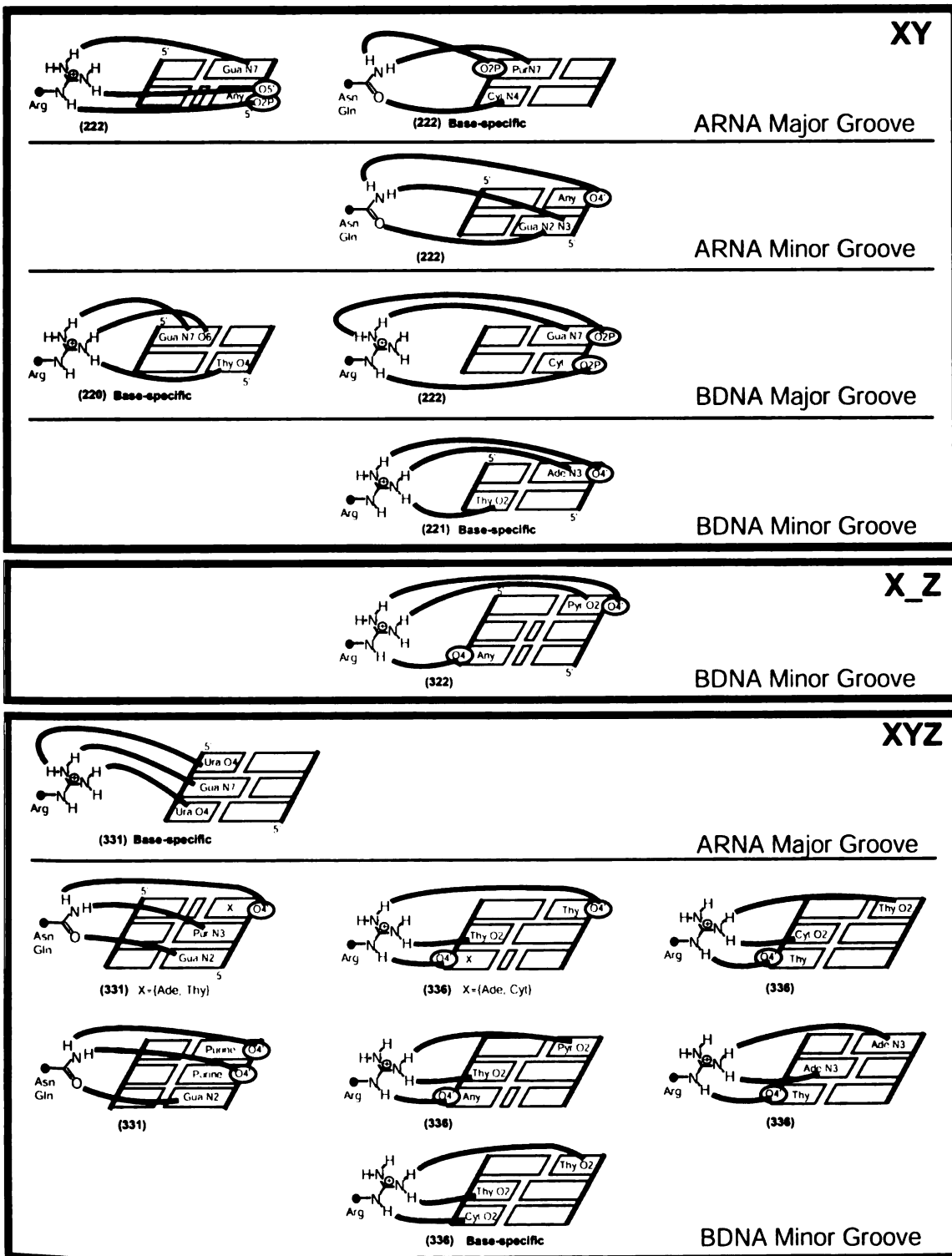


Figure 4.6



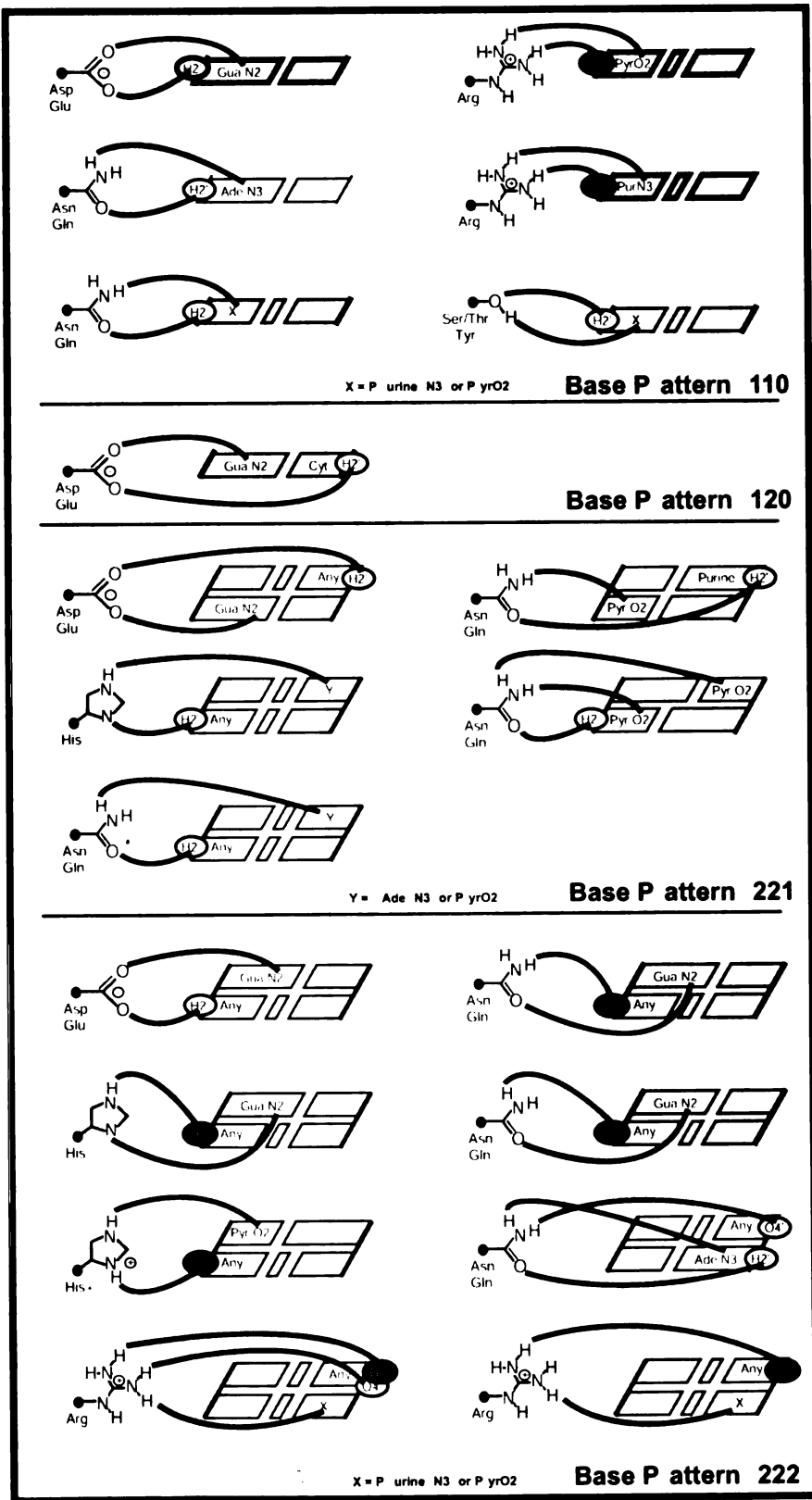


Figure 4.7



ARNa Major Groove

Stringent Moderate None

Stringent Moderate None

(222)		X-(Ade N6, Cyt N4) Y-(Ade N6, Cyt N4) X-(Cyt N4) Y-(Ade N6)	X	X	X	(220)					X	X
(222)			X	X	X	(221)		X-Gua O6 Y-Gua O6 X-Ura O4 Y-(Gua O6, Ura O4)			X	X
(222)			X	X		(222)		X-(Ade, Gua) X-(Cyt, Ura)			X	X
(222)		"None" set allows any base in bottom position	X	X		(222)					X	X
(221)		X-(Gua O6, Ura O4) "None" set also allows Gua O6 in top position	X	X		(222)		X-Any Y-Ade X-(Ade, Gua, Cyt) Y-Gua X-Ura Y-Gua			X	X
(222)		X-(Gua O6, Ura O4) Y-(Gua N7, Ura O4)	X	X		(222)					X	X
(220)			X	X		(222)					X	X
(221)			X	X		(222)					X	X
(222)			X	X		(222)					X	X
(222)			X	X	X	(222)		"None" set also includes Gua N7, Ura O4 in the top position			X	X
(222)		X-(Ade N7, Gua N7, Gua O6) X-(Ura O4)	X	X	X	(222)		both the 2hb and 3hb interactions are possible			X	X
(222)			X	X		(222)		X-(Ura O4) X-(Gua O6)			X	X
(222)		X-Gua N7, Y-Ura O4 X-Ura O4 Y-Gua N7	X	X		(222)		X-(Ade N7, Gua N7, Ura O4) X-(Gua O6)			X	X

ARNa Minor Groove

Stringent Moderate None

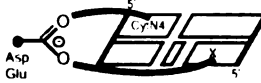

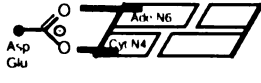
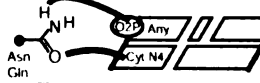
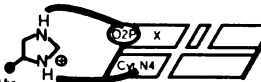



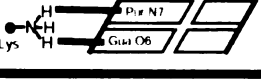

Stringent Moderate None

(221)			X	X		(221)		bifurcated version			X	X
-------	--	--	---	---	--	-------	--	--------------------	--	--	---	---

Figure 4.8



BDNA Major Groove

	Stringent	Moderate	None		Stringent	Moderate	None
(226)  X - [Adk-N6 Cyt-N4]		X	X	(222)  X - [Adk-N7, Gua-N7] X - [Thy-O4]		X	X
(226) 		X	X	(222) 			X
(222)  X - [Adk- Gua-, Cyt-] X - [Thy-]	X	X	X	(220)  X - [Gua- O6, Thy O4]			X
(226) 		X	X	(222) 			X
(222) 		X	X	(222) 			X

BDNA Minor Groove









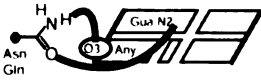
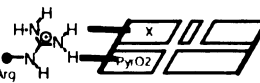
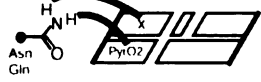
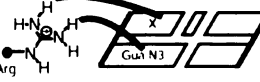
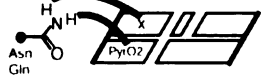
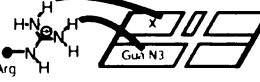
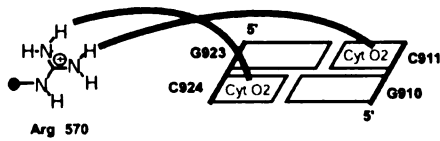
	Stringent	Moderate	None		Stringent	Moderate	None
(221)  X - [Pyr-O2] X - [Pur-N3]	X	X	X	(221)  X - [Adk-N3, ThyO2]		X	X
(222)  X - [Adk-N3, ThyO2]		X	X	(221)  X - [Gua-, Cyt-] X - [Adk-, Thy-]		X	X
(221)  X - [Adk-N3, Cyt-O2, Thy-O2]		X	X	(221)  X - [Adk-N3, Cyt-O2, ThyO2] in "none" set. X includes: GuaN3			X
(221)  X - Adk- Y - [Adk-N3, GuaN2, Cyt-O2, ThyO2] X - Gua- Y - [Cyt-O2, ThyO2]		X	X	(221)  bifurcated			X
(222) 		X	X	(221) 			X
(222) 		X	X	(222)  X - [Adk-N3, ThyO2]			X
(222) 		X	X	(222)  X - [Adk-N3, GuaN3, Cyt-O2, ThyO2]			X

Figure 4.9



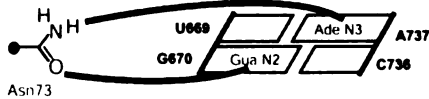
2gax
(Val-AARS with tRNA)

base pattern 221
minor groove



1fg, 1g1x, 1hww
(30S/16S ribosome)

base pattern 221
minor groove



A

DNA Sequence	Same strand (pattern 222)				5' Cross strand (pattern 220)			
	Arg	Lys	Asn/Gln	Asp/Glu	Arg	Lys	Asn/Gln	Ser/Thr
AA/TT	0	0	1	0	0	0	0	0
AC/GT	1	1	2 *	1	0	0	0	0
AG/CT	0	1	0	0	0	0	0	0
AT/AT	0	0	0	0	0	0	0	0
CA/TG	0	0	0	1	0	0	1	2
CC/GG	1	4	0	0	0	0	0	1
CG/CG	0	0	0	0	0	0	0	0
GA/TC	0	0	0	0	5	1	1	0
GC/GC	0	0	0	0	0	0	0	0
TA/TA	0	0	0	0	0	0	0	0
<hr/>								
AA/TT	0	0	1	0	1	0	0	0
AC/GT	1	0	0	0	0	0	0	0
AG/CT	0	0	0	0	0	0	0	0
AT/AT	0	0	2	0	0	1	0	0
CA/TG	0	0	0	0	0	0	0	0
CC/GG	0	0	0	0	0	0	0	0
CG/CG	0	0	0	0	0	0	0	0
GA/TC	1	0	0	0	0	0	1	0
GC/GC	0	0	0	0	0	0	0	0
TA/TA	0	0	0	0	0	0	0	0

B

Figure 4.10

Chapter 5

Future Directions



WASABI has been used to generate models of all possible hydrogen-bonded interactions between amino acid side chains and units of RNA structure. Analyzing the resulting models has led to proposed strategies for specific recognition of RNA at several levels of RNA structure, from single bases that can be involved in recognition at loops and bulges, to non-canonical base pairs, to canonical A-form and B-form nucleic helices. The next steps are threefold: (1) experimentally validate predictions, (2) expand the repertoire of RNA structural units by looking at non-idealized helices, and (3) extend the method for use as a structure-based drug design tool.

Testing a predicted Arg-GU interaction

In our systematically generated set of amino acid- base and amino acid-base pair hydrogen-bonding patterns, we have predicted a number of interesting interactions, including ten interactions with single bases and 180 spanning interactions. Nine of the predicted spanning interactions involve three hydrogen bonds, and one in particular is formed between an arginine and a wobble GU base pair. This interaction is particularly interesting because arginines are quite common in RNA-binding proteins, and wobble base pairs are an important recognition feature in structured RNAs.

We have uncovered three lines of evidence that support such an interaction, with two of them involving analogy to known interactions, and one line of evidence involving high-accuracy calculated gas phase energies.

The first line of evidence derives from homology modeling that suggests the presence of the interaction in a *Drosophila* SNF complex with a U2 snRNA. The crystal structure of a homologous complex, human U2A/U2B'' snRNP bound to U2 snRNA

hairpin, has been solved [Price 1998], and one key interaction in the complex is an Arg52 bidentate hydrogen-bonded interaction to the closing GC base-pair of the loop. Mutation of either the Arg or the GC base-pair abolishes binding. The homologous interaction in the *Drosophila* complex is Arg 52 hydrogen bonding to the major groove of a GU wobble, suggesting the possible use of the predicted three-hydrogen bond interaction between the wobble and an arginine.

A second line of evidence comes from the “pseudo-isomorphic” base-triple found in a crystal structure of a tRNA^{Asp}, and was discussed in Chapter 2 (see Figure 2.9). We will return to the idea of a “pseudo-isomorphic” base triple when we discuss the design of experiments for validation of the predicted wobble-GU.

The final line of evidence comes from our gas-phase energy calculations using quantum chemical methods. To evaluate the energetic favorability of the Arg-wobble interaction, we first performed geometry optimization of the complex, and then compared the optimized interaction energy of the complex to two well-characterized, experimentally observed spanning interactions to base pairs. Shown in Figure 5.1, the two reference interactions are the Asn-GA interaction from the Rev-RRE complex (PDB id: 1ull) [Ye 1996] and the Asn-GC interaction from a mutated Gln aminoacyl-tRNA synthetase bound to tRNA^{Gln} (PDB id: 1qrs) [Arnez 1996].

We reduced the number of atoms in the complexes for computational tractability, since Hartree-Fock calculations scale approximately with N^4 , where N is the number of atoms. The reduced Arg-wobble is shown in Figure 5.2. We used an incremental strategy to obtain the HF/6-31G** optimized geometries of the complexes by energy-minimizing structures using the Amber force-field, then STO-3G, then HF/3-21G, then



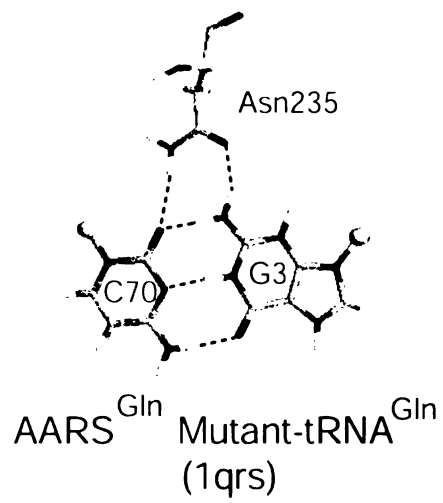
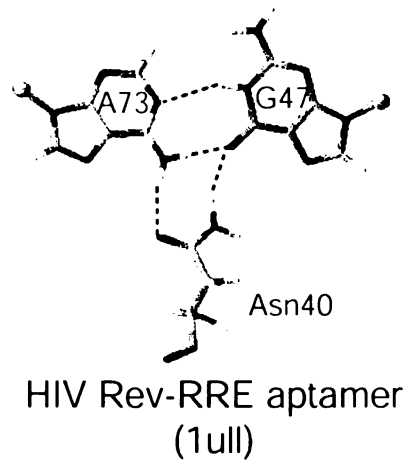


Figure 5.1 Reference interactions used for Arg-GU energy calculations



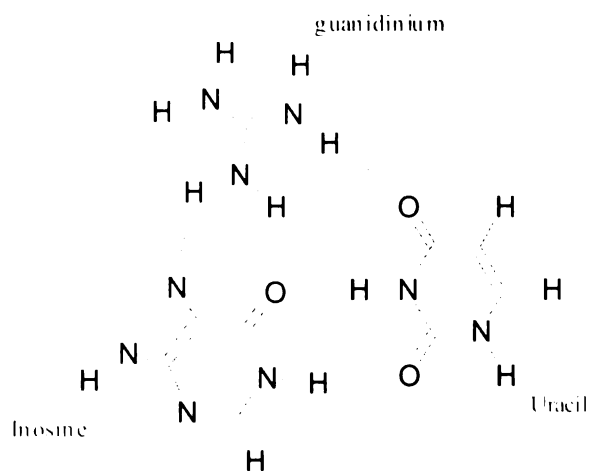
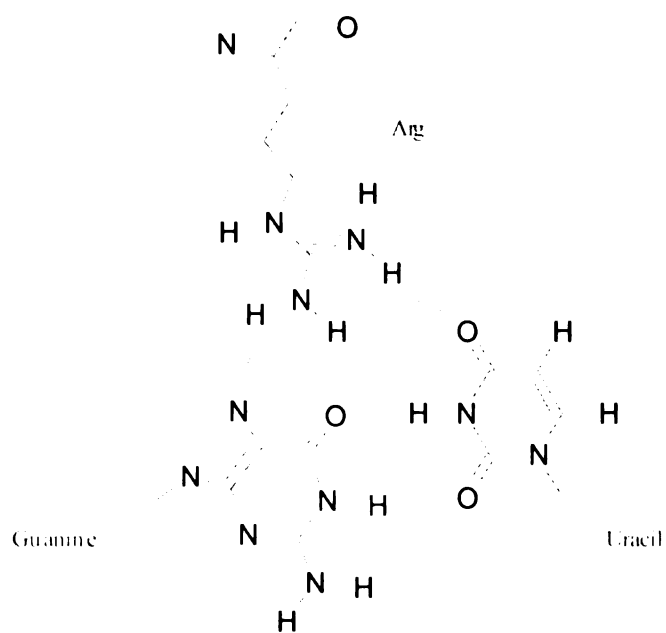


Figure 5.2 Modeled complex of the arginine-GU interaction (top), and reduced model used in quantum chemical calculations (bottom)



HF/6-31G, and finally HF/6-31G** levels of theory. We then calculated LMP2//HF/6-31G** point energies for each of the complexes and their components, with BSSE corrections performed for the HF/6-31G** component of the interaction energy using the counterpoise method.

Our calculated gas phase interaction energies for the three interactions are shown in Table 5.1, and support the favorability of the Arg-Wobble interaction. It has an interaction energy of -40.0 kcal/mol compared to the reference interactions that have interaction energies of -15.0 kcal/mol (Asn-GA) and -17.6 kcal/mol (Asn-GC). The Arg-Wobble is significantly more stable, probably at least in part due to the additional hydrogen bonds, although the presence of a charged hydrogen bond also makes a strong contribution. Further support for the geometry of the predicted interaction comes from the fact that, for every level of theory applied, energy-minimization maintains both the presence of three hydrogen bonds, and the approximately planar conformation of the interaction complex (Fig 5.3).

While the three lines of evidence are suggestive, more conclusive evidence can be obtained by directly measuring the presence of three hydrogen bonds in an experimental system by NMR. Using the idea that an isomorphous base triple can represent the Arg-Wobble interaction, it may be possible to create a triple-helix in solution by forming GGU base triples connected by stable tetraloops. A⁺GU and C⁺GU are also isomorphous to the wobble-GU interaction, and in these cases the A⁺ and C⁺ represent the Arg side-chain. A system utilizing the protonated bases may serve the purpose of more accurately representing the positively-charged Arg side-chain, as well as allow for pH titration of the proton. By altering the pH, we can in theory control the presence and absence of the



	E(HF/6-31G**)	E(LMP2)
Arg-wobble	-1101.479156	-1104.668873
Arg	-204.5457855	-205.163668
Wobble base pair	-896.8612607	-899.4386003
ΔE	-0.072109576	-0.066605016
BSSE correction	0.002921209	0.002921209
ΔE (corrected)	-0.069188367	-0.063683807
ΔE (kcal/mol)	-43.41570039	-39.96158864
1ull	-1156.917333	-1160.366049
Asn	-207.9882461	-208.5999997
GA base pair	-948.9033907	-951.7372471
ΔE	-0.025696032	-0.028802495
BSSE correction	0.004890565	0.004890565
ΔE (corrected)	-0.020805468	-0.023911931
ΔE (kcal/mol)	-13.05543096	-15.00473655
1qrs	-1141.055204	-1144.40974
Asn	-207.9871921	-208.5995458
GC base pair	-933.0276116	-935.7772171
ΔE	-0.040400743	-0.032977273
BSSE correction	0.004911616	0.004911616
ΔE (corrected)	-0.035489126	-0.028065656
ΔE (kcal/mol)	-22.26942662	-17.61119937

Table 5.1 Results of quantum chemical energy calculations. Energies are in Hartrees unless otherwise stated.



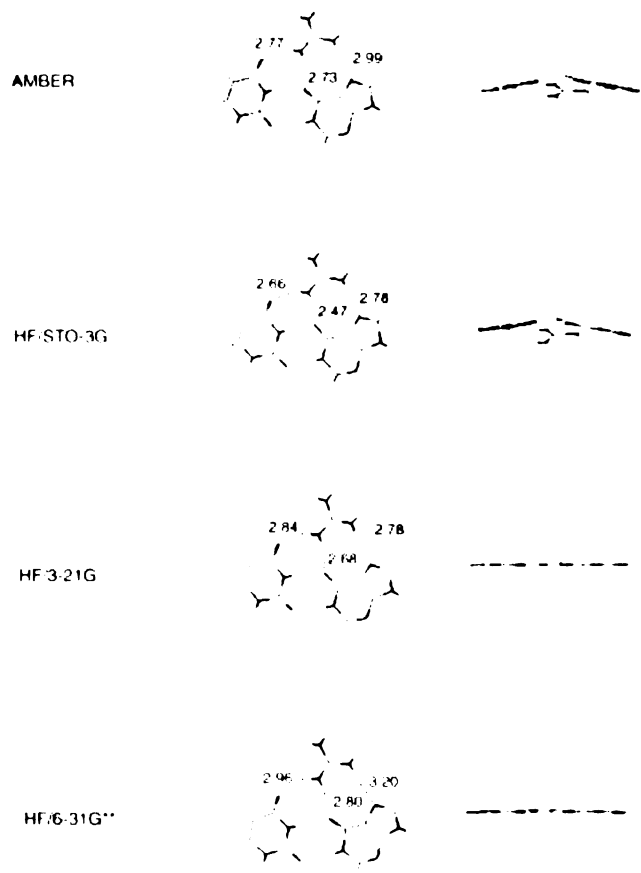


Figure 5.3 Optimized geometries of the Arg-Wobble interaction at four successive levels of theory



predicted third hydrogen bond. Calorimetric measurements may be used to determine the enthalpic contribution of the additional hydrogen bonds. NMR spectroscopy provides a more direct way to detect the presence of hydrogen-bonded protons [Pervushin 1998]. The advantage of using the triple helix system is the precise control over the positioning of the interaction. I attempted some preliminary experiments in which guanidine was titrated into a solution containing double-stranded GU base pairs, but the results were inconclusive.

Extending helix analysis to non-canonical helices

We have used WASABI to search for amino acid preferences for particular base pairs and base steps in the context of idealized A-form and B-form nucleic acids. While permissive parameters can account for slight non-idealities, repeating the studies with nucleic acids that populate the spectrum between A-form and B-form DNA would provide a more complete analysis. Workers in the field have pointed out that intermediate conformations are important in several protein-DNA complexes. A slightly distorted B-form DNA conformation can give rise to an enlarged major groove that is important in formation of complexes involving GLI zinc-finger, trp, glucocorticoid, Zif268 zinc finger, MetJ, engrailed homeodomain, and Tramtrack proteins [Nekludova 1994]. By tracking the presence and absence of each modeled interaction across the spectrum of nucleic conformations, we can categorize specific hydrogen bonded interactions as either broad spectrum or specific to a particular set of conformations. These results of these studies may help us understand why particular interaction patterns are utilized.



Possible approaches to structure-based drug design

With the increasing availability of structural data, RNA targets have become more amenable to structure-based design techniques. We can extend our approach to predict binding modes of small molecule ligands, and ultimately design ligands that specifically bind RNA sites. Using the WASABI approach, specific hydrogen-bonding pockets can be identified in a given RNA structure, and subsequently linked up to predict molecules and binding modes with significantly enhanced affinity and specificity. This fragment-based approach is used in the SAR by NMR [Shuker 1996] and disulfide tethering methods for protein interfaces [Erlanson 2000], as well as in the computational MCSS [Miranker 1991], LUDI [Bohm 1992], and CAVEAT [Lauri 1994] approaches. A recent method called SMOG has been reported to successfully discover picomolar inhibitors. It uses a fragment based approach and a Monte-Carlo search to grow a ligand from a seed site, where additional fragments were added based on a statistical potential scoring function [Grzybowski 2002]. These approaches have been designed for protein targets. The use of WASABI in a fragment-based approach may be particularly appropriate for RNA because of the dominance of hydrogen-bonding in specific interactions. The SMOG approach may be combined with our hydrogen-bonding approach for RNAs. Identification of specifically-binding fragments might also be combined with structure-based combinatorial library techniques for targeted library design.

For each of these approaches, WASABI can be used to list interactions made by small molecule hydrogen bonding moieties to RNAs much as has been done for the reduced amino acid sidechains. A set of functional moieties common to a large number



of small molecules needs to be generated and then used to probe RNA binding sites for pockets where multiple hydrogen bonds can form. Since multiple hydrogen bonds will result in limited group mobility due to the geometric constraints placed by hydrogen bonds, it is practical in this case to list a representative set of possible conformations for the interaction. The Diversigen algorithm (see Chapter 2) can be used to generate a set of conformers that evenly represents the allowed conformational space without breaking the hydrogen-bonding pattern. As an initial test that can be performed with limited chemical synthesis capacity, one can take pairs of specific interaction conformations and ask whether any small molecules in a publically-available database such as the ACD connect the two interactions. Compounds from the ACD can be purchased for testing in an in-vitro or in-vivo assay. In theory greater than pair-wise combinations of multiple-hydrogen bonded moieties can be used to gain multiplicatively increased specificity and binding. Initially, several publically available programs such as CAVEAT, NEWLEAD, BUILDER, and LUDI can be used to search for connectors. CAVEAT [Lauri and Bartlett 1994] is the most suited of these for searching 3D databases based on positioned fragments. Searches using this program are fast, taking on the order of a minute, and uncertainty in the 3D structure of pockets can be taken into account since CAVEAT allows play in its angle and distance search parameters. The major drawback with CAVEAT is that it does not take into account steric clashes, although a separate program can be written to "retest" any hits.

In-silico validation of the method can be done first by attempting to replicate structures of known ligand-RNA interactions. The co-crystal structures of the 30S ribosome with the antibiotics paromomycin, streptomycin, and spectinomycin [Erlanson



2000] provide a good test, and initial steps have been taken towards replicating the binding modes present in the structures. Because performing WASABI searches against the whole 30S ribosome is prohibitively expensive given current available hardware, a feature was built in WASABI that allows docking at predefined sites while using the whole ribosome to define the steric microenvironment for the ligand. Paromomycin was run against the A site of an apo- version of the 30S ribosome structure 1bk.pdb by restricting the target donors and acceptors to residues 1403 to 1410 and 1488 to 1496 on the 16S rRNA, and running WASABI against the full ribosome. The run required about a month of processor time on a dual PIII-1Ghz., partly due to the presence of 30 hydrogen-bonding atoms on paromomycin. An initial run was unable to reproduce the conformation found in the crystal structure. However, an analysis of the structure reveals that a vdw scaling factor of 0.75 instead of the usual 0.80 is required to replicate the crystal structure. It is interesting however that the maximum number of hydrogen bonds calculated between paromomycin and the A-site was seven, the same number observed in the crystal structure. Thus, the preliminary work on validation of ligand binding to the 30S ribosome has prompted the hypothesis that binding of antibiotics to the 16S rRNA may optimize the number of hydrogen bonds. Whether this remains true with the reduced vdw radii remains to be seen, and significant work will be necessary to confirm this hypothesis, as well as compute binding modes in several "negative-control" sites on the 16S. Significant computational resources are needed to increase the feasibility of this study, and we note that large-scale parallelization of WASABI, though not currently implemented, is straightforward.



In conclusion, experimental ligand design coupled with model refinement in a closed-loop fashion will ultimately be essential for a better understanding of the influence of various forces, especially hydrogen-bonding, in the specificity of RNA-ligand and RNA-protein interactions.

References

Arnez JG, Steitz TA. Crystal structures of three misacylating mutants of Escherichia coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP. *Biochemistry*. 1996 Nov 26;35(47):14725-33.

Carter AP, Clemons WM, Brodersen DE, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*. 2000 Sep 21;407(6802):340-8

Erlanson, D.A., Braisted, A.C., Raphael, D.R., Randal, M., Stroud, R.M., Gordon, E.M., Wells, J.A. Site-Directed Ligand Discovery *PNAS* 19: 9367-9372. (2000)

Grzybowski BA, Ishchenko AV, Kim CY, Topalov G, Chapman R, Christianson DW, Whitesides GM, Shakhnovich EI. Combinatorial computational method gives new picomolar ligands for a known enzyme. *PNAS* 2002 Feb 5;99(3):1270-3.

Nekludova, L; Pabo, C.O. Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes *PNAS* 1994 91:6948-6952

Pervushin K, Ono A, Fernandez C, Szyperski T, Kainosho M, Wuthrich K. NMR scalar couplings across Watson-Crick base pair hydrogen bonds in DNA observed by transverse relaxation-optimized spectroscopy. *PNAS* 1998 Nov 24;95(24):14147-51.

Price SR, Evans PR, Nagai K Crystal structure of the spliceosomal U2B''-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature*. 1998 Aug 13; 394(6694):645-50.

Scaria PV, Shafer RH. Calorimetric analysis of triple helices targeted to the d(G3A4G3).d(C3T4C3) duplex. *Biochemistry*. 1996 Aug 20;35(33):10985-94.

Shuker SB, Hajduk PJ, Meadows RP, Fesik SW Discovering high-affinity ligands for proteins: SAR by NMR. *Science*. 1996 Nov 29; 274(5292):1531-4.

Ye X, Gorin A, Ellington AD, Patel DJ Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex. *Nat Struct Biol*. 1996 Dec;3(12):1026-33.



Appendix A

Protocol for building helix

interaction databases



Overview

The following is the protocol used to generate, filter, and mine the amino acid-canonical helix databases discussed in Chapter 4. In part A, canonical helices are created using fibre diffraction data. In part B, the executables required for the WASABI run and pre- and post-run processing are moved into a directory, and a set of scripts are executed to generate an organized directory structure for the run. In part C, the WASABI runs are started, and post-processing is performed after completion of all runs. Note that the Makefile used to control the WASABI run can be modified to take advantage of parallel processors. In part D, the models are loaded into a relational database. The models are further processed using relational database functionality via Java programs. And in part E, the procedures for logging into the database are shown. Once logged in, a user can mine the database using SQL. Examples of this are provided in Appendix B.

A. Create ARNA and ADNA triplet helices, and ADNA and BDNA flanked helices

Generate Combinations (combinations/)

```
make3helixcombos > make3helixcombos.output
```

Create Helices (ahelices_dna/ ahelices_rna/ b_helices_dna/)

Use NAB programs to make the triplet and flanked double helices. Output is

*.ambpdb files in the current directory.

makeadna.nab creates A-form DNAs

makebdna.nab creates B-form DNAs

makearna.nab creates A-form RNAs

Create Brookhaven PDB formatted files

Run BabelScript to change AMBER PDB format to Brookhaven PDB format.

Creates *.pdb files.

```
babel -v -ipdb aaa.ambpdb -opdb aaa.pdb
```

Instead of executing this command one at a time, I create a csh script:

```
babelscript.sh
```

Fix B-DNA residue numbers

This program modifies the original pdb files.

```
bdna_renumber.pl *.pdb
```

Extract triplet double helices from flanked double helices

```
extracthelix.pl <flanked dir> <triplet dir>
```

B. Setup Run Directories for Helices

Create directories

For example create run_triplet_adna/, run_triplet_rna/,

```
run_flanked_adna/, run_flanked_bdna/ directories.
```

Copy files into the directories

Copy into the run directories just created:

- 1) helix models created in part A (*.pdb)
- 2) wasabi3d executables
- 3) preprocess executable
- 4) scriptgen executable
- 5) setupRunDirectories executable and supporting files:


```
setupRunDirectories.pl*
setupRunDirectories.list
setupRunDirectories.makefile
```

6) amino acid models (*sm*.pdb)

7) Makefile (this will control all the runs in the directory)

8) reprocess executable

9) redundancy filters:

```
redundantfilterDKNRH, redundantfilterDKNR,
redundantfilterDKN, redundantfilterDK
```

Preprocess the helices

```
preprocess *.pdb
```

Create run directory structure

```
make setup
```

(use make clean if you screw up)

C. Run the Search

Run the search!

```
make run >& run.out &
```

Eliminate extraneous files from run directories and reprocess output files

```
make neater
```

Create filter lists

```
make redundantfilter
```



Eliminate more extraneous files from run dirs

```
make evenneater
```

Make list of all model files for hbond angle program run in part D

```
make listing
```

D. Load into MySQL Database

All these programs are run in the edu/ucsf/load directory

Load models into MySQL

Make sure to: (1) check LoadModels.java file FILTERREDUNDANTS switch (2)

check LoadModels.java defines for FILTER, HELIXGEOMETRY, NUCLEICTYPE (3)

use the correct targetlist (dna or rna).

```
java load.LoadModels
```

Calculate hbond angles

Copy listing from the run directory into a new directory under load/. Will also

need to copy the hbonds program over. The hbonds run usually runs overnight:

```
mkdir hbondangles
cd hbondangles
cp -/run_triplet_arna/arna4t_listing .
hbonds arna4t_listing >& arna4t_listing.output &
```

Add hbond angle information into the models database

Run this from the load/ directory and make sure to set correct listing file (eg,

arna4t_listing)

```
java load.LoadHbondAngles
```



Remove redundant interactions in the databases

Make sure to set databases you want to run the program on in the

FilterRedundant.java program.

```
java load.FilterRedundant
```

E. Mine MySQL Database

All these programs are run in the edu/ucsf/mine directory

Get differences between two databases

Make sure to set parameters in the GetDifferences.java file

```
java mine.GetDifferences
```

Login to MySQL and use helixdb2 for data mining!

```
Mysql -u acheng -p
```

```
Password: acheng
```

Generating tables with only interactions that have nunique>1

This is a sample sql script to create the “nunique” tables. I like to do this for the

*_filtered_ptf and *_unique_*_ptf tables.

```
CREATE TABLE arna_4t_models_filtered_ptf_nunique
SELECT * FROM arna_4t_models_filtered_ptf WHERE nunique >1;
CREATE TABLE arna_4t_hbonds_filtered_ptf_nunique
SELECT hb.* FROM arna_4t_hbonds_filtered_ptf AS hb
INNER JOIN arna_4t_models_filtered_ptf_nunique AS md
ON md.outputnumber = hb.outputnumber;
```


Appendix B

Schema and sample SQL scripts

for helix interaction databases

Overview

The set of WASABI models for sidechain-helix interactions was loaded into a MySQL relational database to ease data processing and allow leveraging of MySQL built-in functions for robust, accurate data mining and filtering. The final database of interaction models is labeled “helixdb2”, and is provided on the back-up CDs. This database file can be loaded into any MySQL system, and should be easily loadable into an Oracle RDB system as well. 282 tables of data were built in the analysis of the helix interactions, and they are described below. The 282 tables can be divided into 141 pairs of tables that follow the following database schema.

Database Schema

The schema consists of two tables, a “models” table that contains general information for a particular model, and a second “hbonds” table that contains detailed information on each hydrogen bond. Each interaction model is indexed by a unique outputnumber that serves as the primary key in the models table and a foreign key in the hbonds table. The actual schema is shown in figure B.1. Amino acid and base interaction atoms are coded by number as shown in table B.1.



code	aa	atom
1	Asp	OD1
2	Asp	OD2
3	His	ND1
4	His	NE2 (HE2)
5	His+	ND1 (HE1)
6	His+	NE2 (HE2)
7	Lys	NZ (HZ1)
8	Lys	NZ (HZ2)
9	Lys	NZ (HZ3)
10	Asn	ND2 (1HD2)
11	Asn	ND2 (2HD2)
12	Asn	OD1
13	Arg	NH1 (1HH1)
14	Arg	NH1 (2HH1)
15	Arg	NH2 (1HH2)
16	Arg	NH2 (2HH2)
17	Arg	NE (HE)
18	Ser	OG
19	Ser	OG (HG)
20	Trp	NE1 (HE1)
21	Tyr	OH
22	Tyr	OH (HH)

A

code	aa	atom
1	Ade	N1
2	Ade	N3
3	Ade	N7
4	Ade	N6 (1H6)
5	Ade	N6 (2H6)
6	Ade+	N1 (H1)
7	Ade+	N3
8	Ade+	N7
9	Ade+	N6 (1H6)
10	Ade+	N6 (2H6)
11	Gua	N1 (H1)
12	Gua	N2 (1H2)
13	Gua	N2 (2H2)
14	Gua	N3
15	Gua	O6
16	Gua	N7
17	Cyt	O2
18	Cyt	N3
19	Cyt	N4 (1H4)
20	Cyt	N4 (2H4)
21	Cyt+	O2
22	Cyt+	N3 (H3)
23	Cyt+	N4 (1H4)
24	Cyt+	N4 (2H4)
25	Thy	O2
26	Thy	N3 (H3)
27	Thy	O4'
28	Ura	O2
29	Ura	N3 (H3)
30	Ura	O4'
100	backbone	O2'
101	backbone	H2'
102	backbone	O3'
103	backbone	O4'
104	backbone	O5'
105	backbone	O1P
106	backbone	O2P
107	backbone	H3T
108	backbone	H5T

B

Table B.1. Atom codes used in database for (A) amino acids and (B) nucleic acids.


```

TABLE models (
  outputnumber    INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
  filename        VARCHAR(100) NOT NULL,
  targetname     VARCHAR(3),
  helixgeometry  ENUM('A','B','Z'),
  nucleictype    ENUM('rna','dna'),
  aminoacid      CHAR(1),
  score          INTEGER,
  numberhbonds   TINYINT UNSIGNED,
  nunique        TINYINT UNSIGNED,
  numberbases    TINYINT UNSIGNED,
  basesteps      TINYINT UNSIGNED,
  basepattern    INTEGER UNSIGNED,
  bifurcated     ENUM('Y','N'),
  minorgroove    ENUM('Y','N'),
  majorgroove    ENUM('Y','N'),
  involvedbase_A ENUM('Y','N'),
  involvedbase_G ENUM('Y','N'),
  involvedbase_C ENUM('Y','N'),
  involvedbase_T ENUM('Y','N'),
  involvedbase_U ENUM('Y','N'),
  involvedbase_AP ENUM('Y','N'),
  involvedbase_CP ENUM('Y','N'),
  involvedphosphate ENUM('Y','N'),
  involvedsugar  ENUM('Y','N'),
  PRIMARY KEY (outputnumber)
);

TABLE hbonds (
  hbond_id       INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
  outputnumber   INTEGER UNSIGNED NOT NULL,
  aa             CHAR(1),
  aa_atom        SMALLINT UNSIGNED,
  base_chain     CHAR(1),
  base_res_no    INTEGER,
  base_res_type  CHAR(1),
  base_atom      SMALLINT UNSIGNED,
  angle          FLOAT UNSIGNED,
  angle_donor    FLOAT UNSIGNED,
  angle_acceptor FLOAT UNSIGNED,
  angle_planar   FLOAT UNSIGNED,
  distance       FLOAT UNSIGNED,
  primary key(hbond_id)
);

```

Figure B.1. Database schema for helixdb2



Tables Names

There are 141 pairs of tables in helixdb2. The following key words are used in the table names to indicate it's state.

10b_ = "nearly ideal" interactions with 10 degree or lower donor angles only

25b_ = "good" interactions with 25 degree or lower donor angles only

all_ = "all" interactions. Tables not 10b or 25b are also "all" interactions

adna_4t_ = triplet A-form DNA used

arna_4t_ = triplet A-form RNA used

hbonds = hbonds table, as described in the schema description

models = models table, as described in the schema description

filtered = filtered for nucleic acid pattern redundancies (described in Chapter 4)

ptf = filtered for amino acid pattern redundancies (described in Chapter 4)

nunique = excludes single bifurcated hydrogen bonding interactions

unique_adnas = unique interactions between ADNA triplet and ADNA flanked

unique_triplet = unique interactions between ARNA and ADNA triplets

unique_spec = unique interactions between ARNA triplet and BDNA flanked

unique_flank = unique interactions between ADNA and BDNA flanked

ribose = includes only 2'OH ribose interactions

BaseSpec = includes only base-specific interactions (described in Chapter 4)



Example SQL scripts

The term “databases” below will refer to a MySQL table, in order to be consistent with the terminology used in Chapter 4.

Count number of interactions in “nearly ideal” tables (databases)

```
# complete interaction databases
select count(*) from 10b_arna_4t_models_filtered_ptf_nunique ;
select count(*) from 10b_arna_4t_hbonds_filtered_ptf_nunique;
select count(*) from 10b_adna_4t_models_filtered_ptf_nunique;
select count(*) from 10b_adna_4t_hbonds_filtered_ptf_nunique;
select count(*) from 10b_fadna_4f_models_filtered_ptf_nunique;
select count(*) from 10b_fadna_4f_hbonds_filtered_ptf_nunique;
select count(*) from 10b_fbdna_4f_models_filtered_ptf_nunique;
select count(*) from 10b_fbdna_4f_hbonds_filtered_ptf_nunique;

# difference databases
select count(*) from 10b_arna_4t_models_ribose_ptf_nunique;
select count(*) from 10b_arna_4t_models_unique_triplet_ptf_nunique;
select count(*) from 10b_adna_4t_models_unique_triplet_ptf_nunique;
select count(*) from 10b_fadna_4f_models_unique_flank_ptf_nunique;
select count(*) from 10b_fbdna_4f_models_unique_flank_ptf_nunique;
select count(*) from 10b_adna_4t_models_unique_adnas_ptf_nunique;
select count(*) from 10b_fadna_4f_models_unique_adnas_ptf_nunique;
select count(*) from 10b_arna_4t_models_unique_spec_ptf_nunique;
select count(*) from 10b_fbdna_4f_models_unique_spec_ptf_nunique;
```

Break down the number of models by groove and number of hydrogen bonds

```
SELECT majorgroove, minorgroove, numberhbonds, count(*)
FROM 10b_arna_4t_models_filtered_ptf_nunique
GROUP BY majorgroove, minorgroove, numberhbonds;
```

Break down the number of models in the major groove

```
SELECT basepattern, count(*)
FROM 10b_arna_4t_models_filtered_ptf_nunique
WHERE majorgroove='Y' and minorgroove='N'
GROUP BY basepattern;
```

Break down the number of models by amino acid

```
SELECT aminoacid, max(nunique)
FROM arna_4t_models_filtered_ptf_nunique
GROUP BY aminoacid;
```


Get maximum number of hydrogen bonds found for each amino acid

```
SELECT aminoacid, max(numberhbonds),max(nunique)
FROM fbdna_4f_models_filtered_ptf_nunique
GROUP BY aminoacid;
```

Create “nearly ideal” (angle_{donor}≤10°) interaction database from “all” database

```
# get list of output numbers that satisfy the criteria
CREATE TEMPORARY TABLE nearlyideal
SELECT outputnumber, count(*) AS count
FROM arna_4t_hbonds
WHERE angle_donor≤10 and distance≤3.1
GROUP BY outputnumber;

# create new models table
CREATE TABLE 10b_arna_4t_models
SELECT md.* FROM arna_4t_models AS md
INNER JOIN nearlyideal AS tb
ON tb.outputnumber = md.outputnumber
WHERE md.numberhbonds = tb.count and md.nunique>1;

# create new hbonds table
CREATE TABLE 10b_arna_4t_hbonds
SELECT hb.* FROM arna_4t_hbonds AS hb
INNER JOIN 10b_arna_4t_models AS md
ON md.outputnumber=hb.outputnumber;
```

Get number of hbonds broken out by base pattern groups (xy,x_z,xyz)

```
# XY
SELECT numberhbonds, count(*) FROM arna_4t_models_unique_spec_ptf_nunique
WHERE majorgroove='N' and minorgroove='Y'
and bifurcated='N' and aminoacid!='Y'
and (basepattern=220 or basepattern=222 or basepattern=231
or basepattern=221 or basepattern=230 or basepattern=240)
GROUP BY numberhbonds;

# X_Z
SELECT numberhbonds, count(*) FROM arna_4t_models_unique_spec_ptf_nunique
WHERE majorgroove='N' and minorgroove='Y'
and bifurcated='N' and aminoacid!='Y'
and (basepattern=320 or basepattern=321 or basepattern=322
or basepattern=332 or basepattern=335 or basepattern=344)
GROUP BY numberhbonds;

# XYZ
SELECT numberhbonds, count(*) FROM arna_4t_models_unique_spec_ptf_nunique
WHERE majorgroove='N' and minorgroove='Y'
and bifurcated='N' and aminoacid!='Y'
and (basepattern=331 or basepattern=333 or basepattern=334
or basepattern=336 or basepattern=340 or basepattern=341
or basepattern=342 or basepattern=343 or basepattern=345
or basepattern=346 or basepattern=347 or basepattern=350
or basepattern=351 or basepattern=352 or basepattern=360)
GROUP BY numberhbonds;
```



Appendix C

NailMine:

**An user-friendly program for
mining the interaction databases**



Overview

Nailmine is a set of integrated C programs that allows a user to interactively filter the base : base, base : amino-acid, base-pair : amino-acid, and helix : amino acid databases. Users can input constraints, either from experimental evidence or otherwise, to retrieve models of interest. Nailmine has a menu-driven interface for ease-of-use.

Installation

Nailmine is distributed as 2 separate tarballs:

1) `nailmine.tgz` contains

- main program executables
- source code for the programs that mine `WASABI`, and `diversigen`.
- the model databases for the `base-aa` and `basepair-aa` models
- the programs and databases for the `base-base` models, aka `bambi`.

2) `helixaa.tgz` contains model databases for the `aa-helix` models You will need at least

3Gb, and preferably 5Gb of free disk space in the directory you install `nailmine` in.

Directions for installing and compiling `nailmine` is provided in the `readme.txt` file found on installation CD #1.

Running Nailmine

Nailmine can be executed by typing `nailmine` from the shell. The main menu is shown on the next page.

* * * WELCOME TO THE NAIL DATABASE MINING PROGRAM * * *

Please select one of the following choices:

- 1 Mine Base-Base interaction database
- 2 Mine Amino Acid-Base interaction database
- 3 Mine Canonical Helices interaction database
- 4 Use Diversigen conformation generator
- 5 (Internal) Run NAIL on a directory
- 6 I would like to quit

Please enter the number of your choice:

Menu options 1-3 correspond to the aforementioned available databases. option 2 includes both single base and base-pair interactions to amino acids. option 4 is a C-program wrapper for Diversigen. option 5 generates a script and instructions to create a NAIL website for a directory of interaction models. The process is quite awkward, but can be done if there is immense need. option 1 is the BAMBI program written by Bernhard Walberer, and option 4 is self-explanatory. The main menus for options 2 and 3 will be covered here.

Selection of option 2 brings up a menu that allows you to select either single base or base pair interactions with amino acids. The two options presents you with similar menus and filtering options. As an example, we will follow the first option. Selection of option 1 for single base interactions presents you with a self-explanatory menu shown on the next page.

```
* * * WELCOME TO THE AA-BASE DB MINING PROGRAM * * *  
      344 interactions present in database!
```

Which dataset would you like to start with?

- 1 exclude Bifurcated hydrogen bonds
- 2 Include ALL 344 interactions

- 3 Exit to Main MINE menu

Please enter the number of your choice:

Selecting option 1 brings you to the main menu below.

```
* * * WELCOME TO THE AA-BASE DB MINING PROGRAM * * *  
      261 interactions currently selected!
```

Which features would you like to select:

- 1 Base Composition
- 2 Amino Acid Composition
- 3 Number of Hydrogen Bonds
- 4 Exclude Bifurcated Interactions
- 5 Require Interaction Planarity
- 6 Select Interaction Atoms
- 7 Exclude redundant A+, C+ interactions

- 8 RESET all models to "selected"
- 9 QUIT-- print a list of models
- 10 QUIT-- print a list AND save models
- 11 QUIT-- save nothing
- 12 Advanced Black Magic (Alan Cheng only)

Please enter the number of your choice:

Option 1 allows you to select only interactions involving a particular base, or only purines or pyrimidines. Option 2 allows you to select only interactions involving a selected amino acid. Option 3 allows you to chose an exact number or a range of number of hydrogen bonds involved. Option 4 excludes interactions that include bifurcated

hydrogen bonds. Option 5 selects only interactions that can be formed in the plane, and accesses a database of interactions found using a 2D search in the plane. Option 6 allows you to select only interactions that involve a particular base or amino acid atom. This option can be repeated to further narrow down the set of selected interactions. Option 7 excludes interactions to the protonated base pairs that are represented by the non-protonated versions of the base pairs. In essence this means eliminating interactions to A⁺ and C⁺ that don't involve intermolecular hydrogen bonds to the extra proton. Option 8 selects all interactions in the database and undoes all filtering done. Option 9 allows printing of a list of models selected into a user-specified file. Option 10 also prints a list, and furthermore puts pdb files of all models into a user-specified directory. Option 11 allows you to quit to the main nailmine menu. Option 12 allows printing of a database file that can be saved and loaded into a personal copy of nailmine. To load the database into nailmine, a user will need to replace the applicable database file in their personal `~/nailmine/program/bin` directory with the filtered database file. This is admittedly clumsy, but can be done if one is compelled.

Selecting the "Mine Canonical Helices" option from the main nailmine menu brings up the menu on the next page.

* * * WELCOME TO THE AA-CANONICAL HELIX DB MINING PROGRAM * * *

Which interaction dataset would you like to start with?

- 1 All interactions
- 2 Base Specific interactions
- 3 Difference databases of interactions
- 4 Exit to Main MINE menu

Please enter the number of your choice:

The menu items here are self explanatory from a reading of Chapter 4. Selecting option 1 results in the self-explanatory menu below.

* * * WELCOME TO THE AA-CANONICAL HELIX DB MINING PROGRAM * * *

Which all interaction dataset would you like to use?

- 1 arna triplets
- 2 adna triplets
- 3 adna flanked
- 4 bdna flanked
- 5 Exit to Main MINE menu

Please enter the number of your choice:

Selecting option 1 brings up the filtering menu shown on the next page.

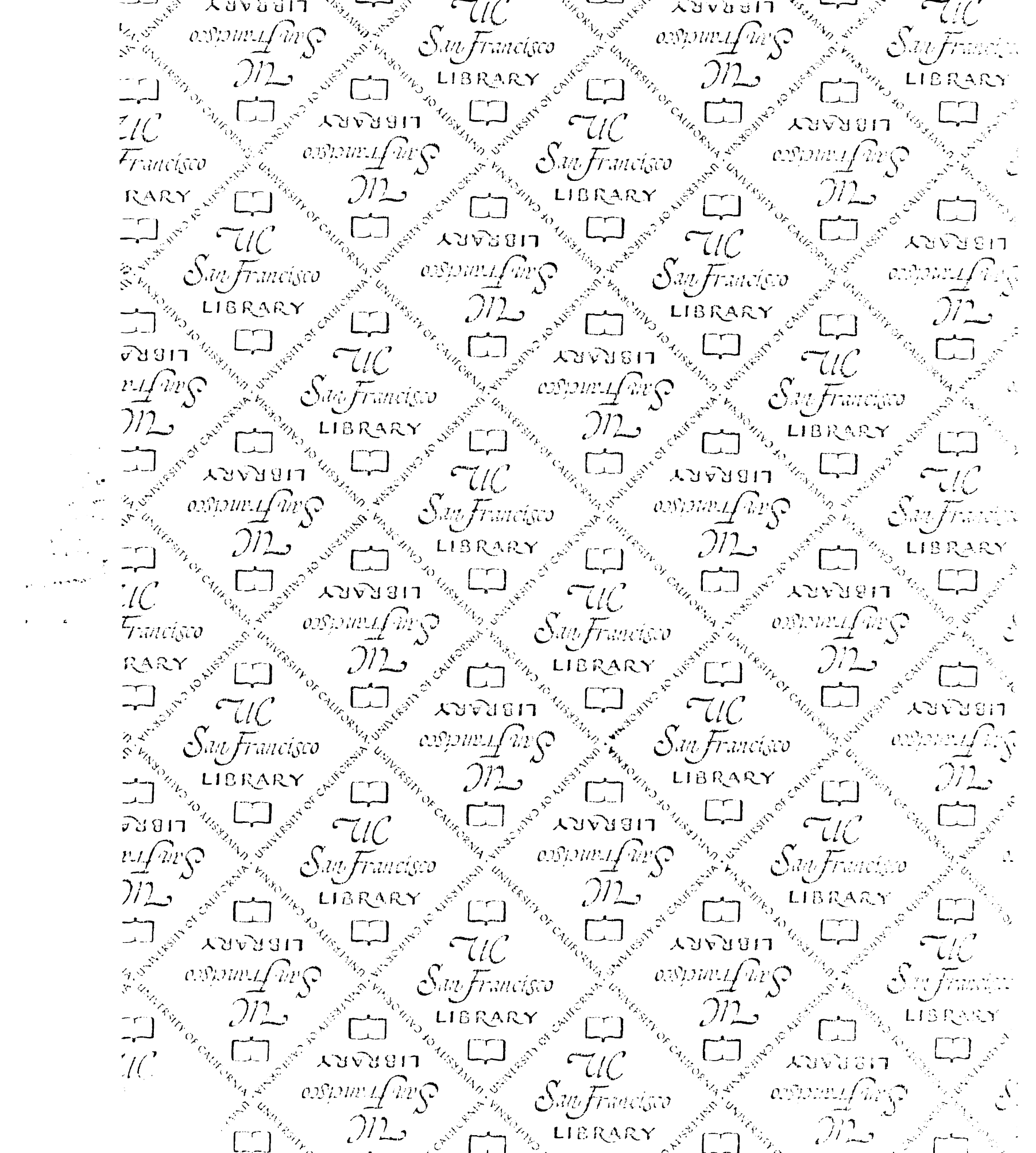
* * * WELCOME TO THE AA-BASEPAIR DB MINING PROGRAM * * *
1859 interactions currently selected!

Which features would you like to select:

- | | |
|-----------------------------------|--------------------------------------|
| 1 Amino Acid involved | 12 Hbond Angle range |
| 2 Bases involved | 13 Hbond Donor Angle range |
| 3 Backbone involvement | 14 Hbond Distance range |
| 4 Number of hydrogen bonds | 15 Select Interaction Atoms |
| 5 Number of unique hydrogen bonds | 16 Exclude tyrosines |
| 6 Exclude Bifurcated Interactions | 17 Select basepattern type |
| 7 Number of bases involved | 18 RESET all models to "selected" |
| 8 Number of basesteps involved | |
| 9 Basepattern involved | 19 Leave- Print list of models |
| 10 Minor Groove Interactions only | 20 Leave- Print list AND save models |
| 11 Major Groove Interactions only | 21 Leave- Save nothing |

Please enter the number of your choice:

Many of the options here are similar to those in the base-amino acid menu or are self-explanatory, and only the novel options will be reviewed. Option 3 allows selection of only interactions that involve a sugar, or alternatively only interactions that involve a phosphate. Selection of a particular backbone interaction atom can be done via option 16. Option 9 allows selection of a particular base pattern via the basepattern numbers presented in Chapter 4. Option 17 can be used for grosser selection of basepattern types (ie, single step, XY, X_Z, or XYZ). Options 12, 13, and 14 allow restriction to only interactions that have all hydrogen bonds with less than a given angle or distance. Option 12 refers to the acceptor-hydrogen-donor angle. Option 16 allows exclusion of Tyr because Ser can represent all Tyr interactions.



For reference

Not to be taken from the room.

