Protein–Nucleic Acid Recognition: Statistical Analysis of Atomic Interactions and Influence of DNA Structure

Diane Lejeune,[†] Nicolas Delsaux,[†] Benoît Charloteaux,[§] Annick Thomas,[¶] and Robert Brasseur[‖]*

Centre de Biophysique Moléculaire Numérique, Faculté Universitaire des Sciences Agronomiques, Gembloux, Belgium

ABSTRACT **We analyzed structural features of 11,038 direct atomic contacts (either electrostatic, H-bonds, hydrophobic, or other van der Waals interactions) extracted from 139 protein–DNA and 49 protein–RNA nonhomologous complexes from the Protein Data Bank (PDB). Globally, H-bonds are the most frequent interactions (~50%), followed by van der Waals, hydrophobic, and electrostatic interactions. From the protein viewpoint, hydrophilic amino acids are over-represented in the interaction databases: Positively charged amino acids mainly contact nucleic acid phosphate groups but can also interact with base edges. From the nucleotide point of view, DNA and RNA behave differently: Most protein–DNA interactions involve phosphate atoms, while protein–RNA interactions involve more frequently base edge and ribose atoms. The increased participation of DNA phosphate involves H-bonds rather than salt bridges. A statistical analysis was performed to find the occurrence of amino acid–nucleotide pairs most different from chance. These pairs were analyzed individually. Finally, we studied the conformation of DNA in the interaction sites. Despite the prevalence of B-DNA in the database, our results suggest that A-DNA is favored in the interaction sites. Proteins 2005;61:258–271.**

Key words: **protein–DNA interactions; protein–RNA interactions; DNA conformation; atomic contacts; structural data**

INTRODUCTION

Interactions between proteins and other biological molecules, especially nucleic acids, are the basis of cell life. DNA-binding proteins are coded by 2–3% of the genome in prokaryotes, and 6–7% in eukaryotes.[1] These nucleic acid–binding proteins are involved in fundamental roles as replication, transcription, restriction, and even viral infection.

The recognition of a specific nucleotide sequence by a DNA- or RNA-binding protein is determined by atomic interactions between amino acids and nucleotides. Attempts to determine the rules of this recognition began in the 1970s.[2] Of all possible types of interactions, hydrogen bonds between amino acids and base edges were shown to be the most relevant.[2,3] Most H-bonds involve protein side-chains and contribute to the protein specificity, while interactions between nucleotides and protein backbone seem to be important for the stabilization and orientation of the complex.[4] The specificity due to side-chain H-bonds cannot be explained in terms of a one-to-one mechanism, but rather in terms of a one-to-many interaction, such as bidentate or bifurcated H-bonds, and complex H-bonds contacting several base steps.[2,5]

In addition to the well-recognized H-bonds (NH—O, OH—O), at least two other types of H-bonds have been shown to be involved in protein–nucleic acid interaction. First, the CH—O interactions have been shown to frequently link the C—H groups in the DNA major groove and protein side-chain oxygens.[6] The same CH—O bonds make up 33% of the H-bonds between protein and RNA.[7] Second, water-mediated H-bonds account for 15% of all protein–DNA interactions and are involved in minimizing the electrostatic repulsion of aspartate and glutamate from DNA phosphates. Such interactions are surprisingly frequent in protein–RNA contacts.[5,7,8] See Jayaram and Jain[9] for a recent review of the role of water in protein–DNA interactions.

Ionic interactions between positively charged amino acids and phosphate oxygen are important for stabilizing complexes[7,10] and could be involved in the long-distance preorientation of peptidic chains.[11] In addition, some DNA-binding proteins bend nucleotide chains, modifying the mechanism of DNA recognition by amino acids.[12,13]

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: Ministère de la Région Wallonne; Grant number: 215140 (αBUSTEC) (to D. Lejeune). Grant sponsor: Grant FNRS–Télévie (to N. Delsaux). Grant sponsor: Interuniversity Poles of Attraction Program—Belgian State, Prime Minister's Office, Federal Office for Scientific, Technical, and Cultural Affairs; Grant number: P5/33. Grant sponsor: Ministère de la Région Wallonne; Grant number: 3.4505.02 (FRSM1). Grant number: P5/33. Grant sponsor: National Funds for Scientific Research of Belgium (FNRS); Grant number: 3.4505.02 (FRSM).

[†] These authors contributed equally to this work.

[‖] R. Brasseur is Research Director at the National Funds for Scientific Research of Belgium (FNRS).

[¶] A. Thomas is Research Director at the French Institute of Health and Medical Research (INSERM).

[§] B. Charloteaux is Research Fellow at the FNRS.

*Correspondence to: Robert Brasseur, Centre de Biophysique Moléculaire Numérique, Faculté Universitaire des Sciences Agronomiques, Passage des Déportés 2, B-5030 Gembloux, Belgium. E-mail: brasseur.r@fsagx.ac.be

Received 22 December 2004; Revised 14 March 2005; Accepted 1 April 2005

Published online 24 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20607

The ability of DNA to bend appears to be influenced by its base sequence.[14] The pyrimidine–purine dinucleotide steps allow the DNA chain to adopt more flexible conformations.[15,16] Finally, even though many proteins are homodimeric when they bind DNA, they often bind the nucleic acid chains in an asymmetric manner. Subtle recognition differences between the protein chains have been suggested to play a role in determining specificity.[17] Analyzing the conformation of the binding site on the protein has enabled the classification of protein–DNA complexes into families.[1]

Numerous studies have focused on the atomic contacts in protein–DNA and protein–RNA complexes. Previous articles classified complexes according to structural families (zinc finger, helix–turn–helix, etc.[1]), or focused on a particular nucleotide atom type (especially base edge[5,18]), a particular interaction type (H-bond mainly[3]) or a given complex (i.e., Zif268 zinc finger[19–22]). For a long time, the low number of protein–RNA three-dimensional (3D) structures limited extensive studies for this type of complexes.[7,23–25]

With the aim of discovering universal recognition rules, we analyzed all types of interactions (electrostatic, H-bonds, hydrophobic, and other van der Waals) in both protein–DNA and protein–RNA complexes. For this purpose, we constructed a database of nonhomologous protein–DNA and protein–RNA complexes from the structures available in the Protein Data Bank[26] (PDB) in October 2003. To our knowledge, this database is the largest available of its type. It can be downloaded from our website (http://www.fsagx.ac.be/bp/). Using it, we were able to compare properties of protein–RNA and protein–DNA complexes. One of the main differences is that whereas phosphate is the most frequently involved group in protein–DNA interaction (47%), sugar has this status in RNA complexes (43%).

A statistical analysis revealed the most favored amino acid–nucleotide pairs. Our results confirm the importance of positively charged and polar residues in both types of complexes. In addition, aspartic acid appears as one of the most frequent partners of guanine in protein–RNA complexes. The results also highlight the role of hydrophobic interactions between sugar atoms and aliphatic or aromatic side-chain atoms.

Finally, previous studies analyzed the conformation of the proteic binding site (reviewed by Luscombe et al.[1]). Here, we have analyzed the nucleotide side. Two main types of double-helical DNA structures have been described: A-DNA and classic B-DNA. These structures can be differentiated[27] by using the backbone torsional angle associated with sugar ring (δ), together with the glycosyl torsion angle (χ). Despite the prevalence of B-DNA in the database, our results suggest that A-DNA is favored in the interaction sites.

MATERIALS AND METHODS

Database of Nonhomologous Complexes

To assemble our database of crystal structures, we searched the PDB (October 2003 version)[26] for all protein–nucleic acid complexes with a resolution of 3 Å or less. Next, we discarded homologous complexes using the PISCES server.[28] This software enabled us to select complexes with less than 30% protein sequence identity using local alignments. For each complex, we checked for the existence of a protein–nucleic acid interface and discarded all repeated biological units and any complex composed of cocrystallized molecules. Our final data set contains 49 protein–RNA complexes and 139 protein–DNA complexes (Table I). The PDB files were handled as follows. Hydrogen atoms were added using HyperChem 5.0.[29] Ions, heteroatoms, and water molecules were discarded. Heterobases and modified amino acids were kept.

Tests were made to estimate the influence of database homology and have shown that if homologous protein sequences are added (676 protein–DNA complexes and 143 protein–RNA complexes), similarly qualitative results are obtained (data not shown).

Pex Files and Interacting Pairs

Analysis of the 188 3D structures was performed using the Pex software (Biosiris; Parc Crealys, Belgium). We generated Pex files[30,31] and looked for the shortest contact distance for each couple of amino acid and nucleotide. Two atoms were considered to be in interaction if their centers were within a distance ranging from 1 Å to 5 Å. Each atom pair was added into a new database called the "interaction database." Pex files are accessible on the Centre de Biophysique Moléculaire Numérique (CBMN) website (http://www.fsagx.ac.be/bp/) or can be obtained from author R. Brasseur.

Classification of Amino Acids and Differentiation of Atom Types

For the sake of clarity, amino acids were classified as hydrophilic or hydrophobic, based on the Eisenberg hydrophobicity scale.[32] For hydrophilic amino acids, we distinguished between positively charged (Arg and Lys), negatively charged (Asp and Glu), and polar (Asn, Gln, His, Ser, and Thr) residues, while hydrophobic amino acids were divided into three groups: aliphatic (Ala, Ile, Leu, Met, and Val), aromatic (Phe, Trp, and Tyr), and particular (Cys, Gly, and Pro). We also distinguished atoms belonging to amino acid backbone (*BK*) (N, CA, C, O, H, HA) from those belonging to side-chain (*SC*) (all other amino acid atoms).

For the nucleic acids, we differentiated atoms from phosphate (P, O1P, and O2P), sugar (O2*, O3*, O4*, O5*, C1*, C2*, C3*, C4*, C5*, H1*, H2*, H3*, H4*, H5*, and HO2) or base edge (all other nucleotide atoms).

Interaction Types

All pairs of atoms, one from the protein and the other from the nucleic acid, were classified into one of these four categories: electrostatic interactions for atoms with opposite charges, H-bonds (classic XH—Y H-bonds, where both X and Y are electronegative atoms, but also CH—O H-bonds), and van der Waals interactions, among which we classified a fourth category: hydrophobic interactions involving atom pairs with low electronegativity difference.

TABLE I. PDB Codes of the Complexes Used in the Analysis

A						
1A0A	1A1V	1A3Q	1A73	1AIS	1AM9	1AWC
1B01	1B3T	1BC8	1BDT	1BG1	1BL0	1BPY
1BRN	1C8C	1CEZ	1CF7	1CKT	1CL8	1CW0
1D02	1DC1	1DDN	1DEW	1DFM	1DH3	1DIZ
1DMU	1DP7	1DSZ	1E3O	1ECR	1EFA	1EGW
1ESG	1EWN	1EWQ	1EXJ	1EYG	1F0V	1F44
1F4K	1FIU	1FOK	1FZP	1G38	1G9Z	1GDT
1GT0	1GU4	1GXP	1H6F	1H9D	1HAO	1HCR
1HI0	1HLV	1HWT	1I3J	1I6J	1I7D	1I8M
1IAW	1IC8	1IGN	1J1V	1J75	1JB7	1JE8
1JEY	1JFI	1JJ4	1JMC	1JT0	1JX4	1K3X
1K4T	1K78	1KC6	1KDH	1KU7	1KX5	1L3L
1L3S	1LLM	1LMB	1LQ1	1LRR	1LWY	1M07
1M5R	1MHD	1MJO	1MNN	1MUS	1MW8	1MWI
1N6Q	1NH2	1NKP	1NLW	1NOY	1ODH	1OE4
1ORN	1OUP	1P4E	1P71	1P7H	1PUF	1PV4
1QNA	1QPI	1QPZ	1QRV	1QUM	1R2Z	1REP
1SKN	1T7P	1TC3	1TRO	1TUP	1UBD	1VAS
1ZME	2BOP	2BPA	2CGP	2DRP	2HDD	2IRF
2PJR	2UP1	3HTS	3PVI	6CRO	6MHT	
B						
1A9N	1APG	1ASY	1AV6	1B23	1B7F	1C0A
1CXO	1DDL	1DFU	1DI2	1E6T	1E7K	1EC6
1F7U	1F8V	1FEU	1FFY	1G1X	1GAX	1GTF
1H2C	1H3E	1H4S	1HQ1	1I6U	1IL2	1J1U
1JBR	1JID	1K8W	1KNZ	1KQ2	1LNG	1M8V
1M8X	1MMS	1MZP	1N35	1N78	1NB7	1OOA
1QF6	1QTQ	1RMV	1SER	2A8V	2BBV	2FMT

(**A**) The protein-DNA data set includes 139 nonhomologous complexes with a resolution better than or equal to 3 Å. Underlined PDB codes indicate complexes used in the high resolution (≤2 Å) double helical data set. (**B**) 49 protein-RNA complexes.

Frequencies and Propensities

We compared the frequency of amino acids or DNA structures in the whole databases and in the interaction databases. We defined the propensities of amino acids to interact with nucleic acids and the propensities of DNA structures to interact with proteins (P_i) as follows:

$$P_i = \left(I_i / \sum_i I_i \right) / \left(T_i / \sum_i T_i \right),$$

where I_i is the amount of amino acid i (or DNA structure i) in the interacting data set, and T_i is its amount in the entire structure data set. $\sum_i T_i = 60{,}316$ in protein–DNA and 28,152 in protein–RNA complexes. $\sum_i I_i = 7671$ and 3367 in protein–DNA and protein–RNA complexes, respectively.

A P_i propensity greater than 1.2 indicates a residue (or DNA structure) occurring more frequently in the protein–nucleic acid interfaces than in the whole complexes. A residue (or DNA structure) with a P_i value between 0.8 and 1.2 was considered to be indifferent to the presence of DNA or RNA (or protein), while a residue (or DNA structure) with a P_i lower than 0.8 was considered as disfavored in the interaction sites.

Interaction Matrix and Statistical Analysis

To analyze the interaction pairs, we constructed "interaction matrices" in which rows were amino acids and columns were nucleotides (DNA or RNA).

For all amino acid–nucleotide pairs, we calculated a table of contingency on the basis of the frequency of amino acids and bases in the whole database. Expected values, coming from these contingency tables, are those expected for random occurrence of interactions. The differences between expected and observed frequencies of pairs provided us with the information we needed to perform an independency test of Pearson. We calculated the χ^2 value for the entire table and the associated p value for a χ^2 distribution of Pearson with 57 degrees of freedom [$df = (20 - 1) * (4 - 1)$ for 20 rows and 4 columns]. This p value indicates whether rows and columns are independent or not. Next, we determined which amino acid–nucleotide pairs were statistically favored or disfavored, calculating their individual χ^2 values. Note that the expected frequency was always greater than or equal to 5, so the Pearson χ^2 test was applicable without limitation.

In addition to the interaction matrices for the whole residue, we also analyzed interactions involving either the backbone or the side-chain of amino acids, and the phosphate, sugar, or base edge of nucleotides (see above).

Pair Interactions at an Atomic Level

Statistically significant pairs were classified according to their individual χ^2 values. Then, we selected the most favored pairs for further study (associated p value $< 1 * 10^{-6}$ with 1 degree of freedom).

For each selected pair, we constructed an "atomic interaction matrix" including all possible atomic pairs. We classified the interactions according to the atom type involved, and to the type of interaction: electrostatic, H-bond, hydrophobic, or other van der Waals (as described above).

DNA Structures

From the protein–DNA complex database, we extracted all double-stranded DNA with a resolution equal to or less than 2 Å (underlined PDB codes in Table I). We characterized the two usual types of double-helix structures: A-DNA and B-DNA. Our selection was based on torsional angles: δ (between C5*-C4*-C3*-O3* atoms) and χ (between O4*-C1*-N1-C2 atoms for pyrimidines, and between O4*-C1*-N9-C4 atoms for purines). The definition of A- and B-DNA is based on the work by Lu et al.[27] For the A-DNA, δ ranges from 60° to 110° and χ from 150° to $-140°$; B-DNA has δ values ranging from 70° to 180°, and χ values between $-140°$ and $-60°$.

RESULTS

Database Composition

We prepared two databases: one of 139 protein–DNA complexes and one of 49 protein–RNA complexes. All complexes had a protein sequence identity lower than or equal to 30%, and a resolution of 3 Å or better. The PDB codes of the complexes are listed in Table I. To our knowledge, these databases are the largest sets of nonhomologous protein–nucleic acid complexes at this date, especially for the protein–RNA database.

The protein–DNA complexes database include 60,316 residues (55,811 amino acids and 4505 nucleotides) in 139 complexes, with a mean of 434 residues per complex. The largest complex is made of 2524 residues and the smallest is 74 residues. The 49 protein–RNA complexes are composed of 28,152 residues in total (24,311 amino acids and 3841 nucleotides). This corresponds to a mean of 557 residues per complex, with limit values of 124 and 1874.

Amino Acid Composition of the Databases

Analysis of whole databases

We analyzed the global amino acid content of the databases (60,316 residues for DNA complexes and 28,152 residues for RNA complexes) (see Supplementary Material Figure). In the whole SWISS-PROT, taken as a reference, 54% of the amino acids are hydrophobic and 46% are hydrophilic. Hydrophilic residues are more frequent in the DNA complexes: 51%. This finding correlates with a high proportion of positively charged residues: 7.5% and 8.0% of the residues are Arg and Lys, respectively. In comparison, sequences from the whole SWISS-PROT contain only 5.5% of Arg and 6% of Lys. The hydrophobic portion is reduced, especially in aliphatic and particular residues (Cys, Gly, and Pro): 41% in DNA complexes and 46% in the SWISS-PROT. In the RNA complexes, the hydrophobic-hydrophilic ratio is similar to the whole SWISS-PROT distribution. However, of the hydrophilic residues, the balance between charged and polar residues favors the occurrence of charged residues (26% as compared to 23% in the SWISS-PROT) at the expense of polar residues (19.5% compared to 23%). Finally, aromatic amino acids are equally present (8%) in the three data sets.

Composition of the interaction databases

We then extract all amino acids in direct contact with a nucleotide (as described in the Materials and Methods section). In protein–DNA complexes, there are 7671 direct contacts, and 3367 in protein–RNA complexes. Although the DNA- and RNA-complex databases are already rich in hydrophilic residues (especially positively charged amino acids), these amino acids are even more frequent in interactions with nucleotides [Fig. 1(A and B)]. More than 60% of the interacting amino acids are hydrophilic. Positively charged residues represent 28% of the interactions with DNA and 27% with RNA [Fig. 1(A and B)].

To analyze the specificity of the amino acids in the interaction sites, we calculated the propensities (P_i; see Materials and Methods section) of amino acids to interact. In protein–DNA complexes (Fig. 2, black bars), Arg and Lys have the highest propensity to interact with nucleotides: 1.9 and 1.7, respectively. Asn and His (1.4), Ser, Thr, and Tyr (1.3), and Trp (1.2) are also favored, but to a lesser extent. On the other hand, Leu (0.4), Glu (0.5), Ile, Cys and Asp (0.6), and Ala, Val, and Pro (0.7) are disfavored nucleotide partners. The amino acids interacting with RNA show a similar pattern (Fig. 2, gray bars): Arg (2.1), Lys (1.9), Asn (1.6), His (1.5), Gln and Asp (1.3), and Tyr (1.2) are favored. Ala and Val (0.5), Ile, Leu, and Cys (0.6), Met, Trp, and Glu (0.7), and Phe (0.8) are disfavored. The main differences between the DNA and the RNA complexes are observed for Asp and Trp. Asp has a propensity of 1.3 and a frequency of 7% in the RNA interaction sites as compared to a propensity of 0.6 and a frequency of 3% in the DNA interaction sites [Fig. 1(A and B), white bars, and Fig. 2]. Trp has a propensity of 0.7 in RNA as compared to 1.2 with DNA (Fig. 2).

In brief, these results confirm the importance of the two basic residues (Arg and Lys) and highlight the role of polar amino acids.

Interactions with nucleotide base edges

We considered interactions between protein and DNA–RNA base edges [Fig. 1(A and B), gray bars]. The total number of amino acid–base edge contacts is 1819 for DNA and 1163 for RNA. Positively charged residues of proteins are frequently involved in interactions with DNA base edges. Arg has a propensity of 2.6 for these interactions. Interestingly, Lys is less favored (1.3) but is still well represented (10.5%). Beside positively charged amino acids, the polar residues also interact with base edges and account for 30% of the interacting amino acids. Individual residue propensities are 1.6 (Asn), 1.5 (His), and 1.3 (Gln

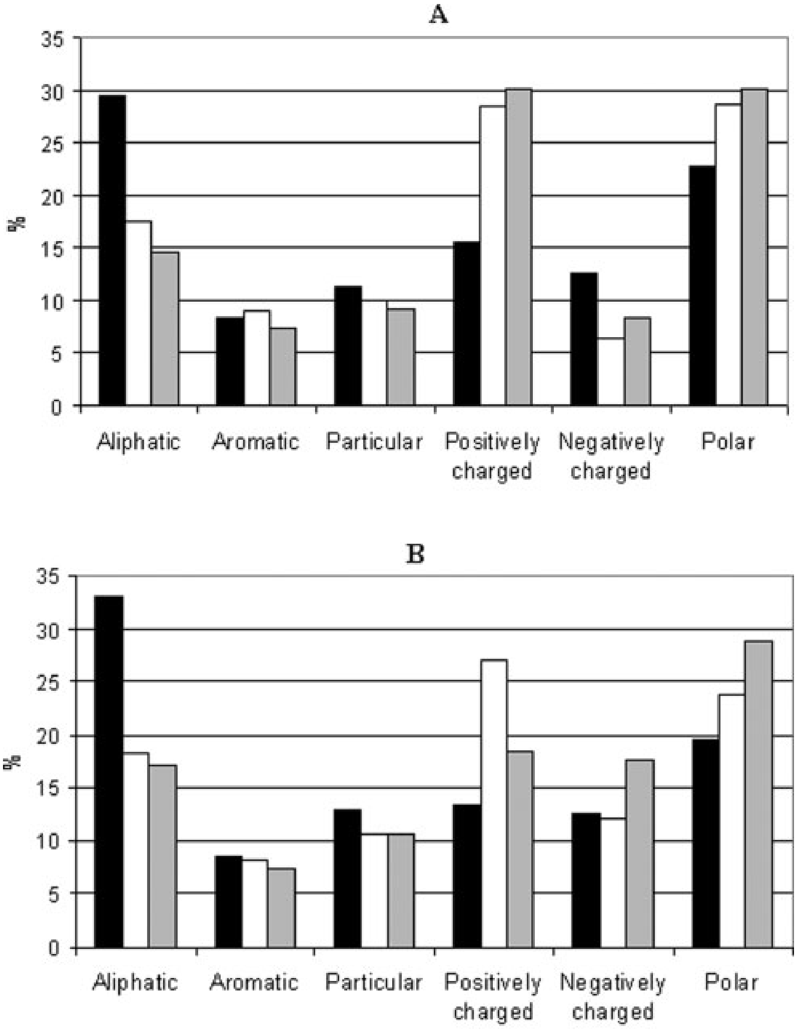

Fig. 1. Frequency distribution of amino acids in the interaction sites of protein–DNA (**A**) and protein–RNA (**B**) complexes. Distribution of amino acids found in the complexes are given as a reference (black bars). White bars correspond to the distribution in the interaction sites, and gray bars correspond to the distribution of amino acids interacting specifically with nucleotide base edges.

and Ser). Hydrophobic residues, except Tyr (1.2), were all disfavored.

For the RNA base edges, interacting propensities of amino acids are different [Fig. 1(B)]. Arg (1.5) and Lys (1.3) are favored, but Asn (2.2), His (2.1), and surprisingly, Asp (2.2) have the highest propensity values. Moreover, Asp is the most frequent amino acid in interaction with RNA base edge (11.5%). As for DNA, hydrophobic residues, except Tyr, are generally disfavored.

Nucleotidic Atom Types Involved in Interactions

Although H-bonds with base edge are important for determining the specificity of protein–nucleic acid interactions, a protein also has contacts with other parts of nucleotides. We clustered nucleotide atoms into three distinct families: those belonging to phosphate, sugar, or base edge (Fig. 3). Unlike the results pertaining to amino acids, these results show clear differences between DNA and RNA complexes.

In protein–DNA complexes, an average of 47% of the interactions involve phosphate atoms, while interactions with base edge account for only 24% (with extreme values of 18% for adenine and 27% for cytosine) (Fig. 3). For the protein–RNA complexes, only 22% of the phosphate atoms come in direct contact with amino acids. Interactions involving base edge and ribose atoms are more numerous than for DNA, with 35% and 43%, respectively (Fig. 3).

This discrepancy probably results from the different conformations of DNA and RNA. Indeed, 118 of the 139 protein–DNA complexes contained double-helical DNA, while 34 out of the 49 RNA complexes contain a single chain of nucleotides. Base edges in double-helical DNA are much less accessible than in single-stranded RNA. The RNA base edge accessibility could explain why bases are involved in more than a third of the RNA contacts with proteins. The greater accessibility of the base edge in RNA complexes results in a decreased number of phosphate atom contacts, while, interestingly, ribose atom contacts

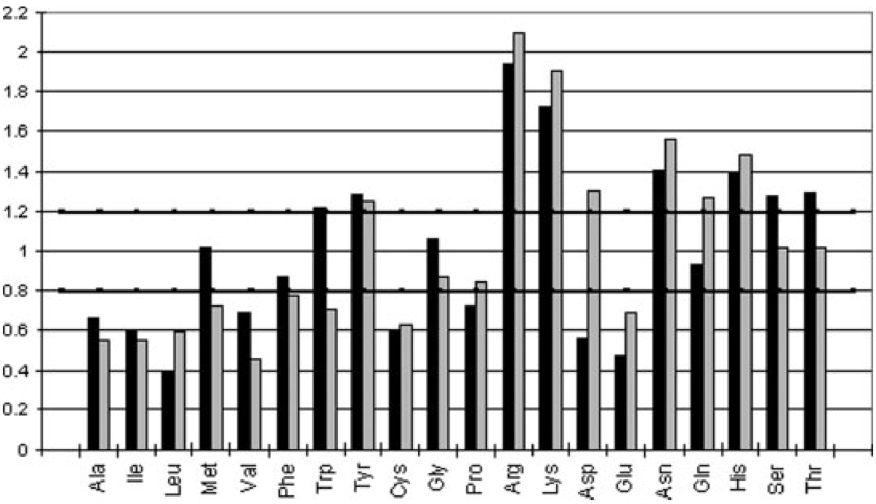

Fig. 2. Amino acid propensities to interact with nucleotides. The propensity values correspond to the frequency of an amino acid in the interaction sites divided by the frequency of the same amino acid in the whole database. Values for DNA are represented with black bars and values for RNA with gray bars. Propensity values higher than 1.2 correspond to favored amino acids, while disfavored amino acids have values lower than 0.8.

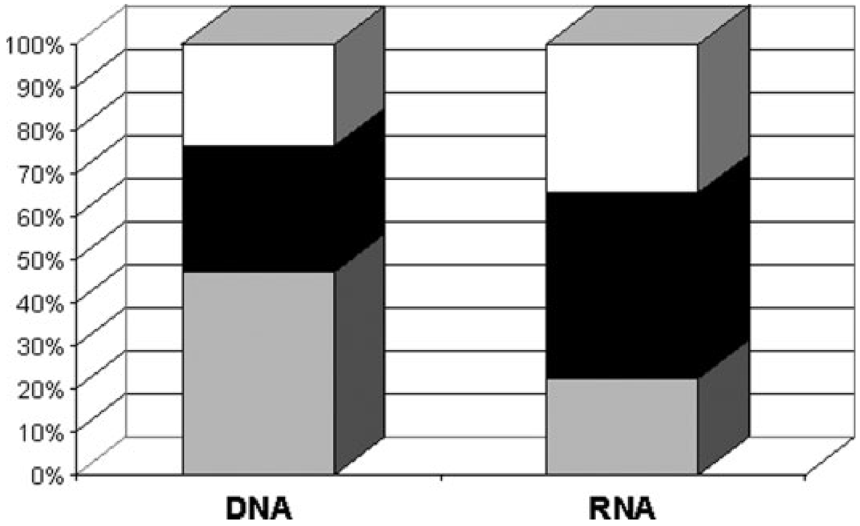

Fig. 3. Frequency distribution of interactions with different nucleic acid parts. Atom types are colored as follows: white, base edge; black, sugar; and gray, phosphate atoms.

are the most represented in the RNA interaction database. This suggests that ribose, the role of which is rarely mentioned in publications, could be important in protein–RNA interactions.

TABLE II. Distributions of Interaction Types in the Interaction Sites of Protein-DNA and Protein-RNA Complexes

	DNA % (Number)	RNA % (Number)
H-bonds	51 (3912)	47 (1582)
Electrostatic	8 (614)	7 (236)
Hydrophobic	19 (1457)	19 (640)
Other van der Waals	22 (1688)	27 (909)

Distribution of Interaction Types

The two distinct behaviors of DNA and RNA nucleotides, in terms of atoms involved in interactions, prompted us to analyze the type of interactions that are concerned. Table II shows the frequency of the four different types of interactions in DNA–protein and RNA–protein complexes. The two distributions are similar, with a predominance of H-bonds (51% and 47%, respectively), and 8% and 7%, respectively, of salt bridges. The similar percentages of salt bridges in DNA and RNA complexes are unexpected given that the interactions with phosphate are twice less frequent in RNA complexes (22%) as in DNA complexes (47%) (Fig. 3). Of the 47% DNA phosphate interactions, only 16% are salt bridges, while of the 22% RNA phosphate interactions, salt bridges are 31%. This difference results

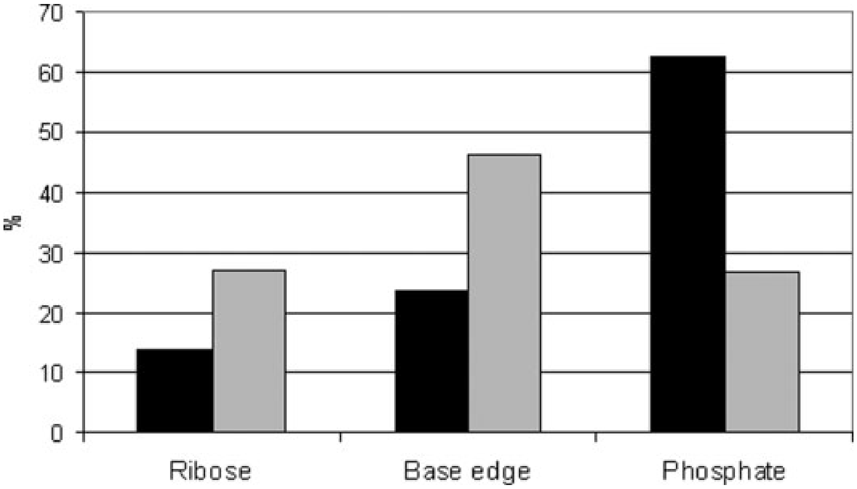

Fig. 4. Distribution of H-bonds according to the nucleotide part. DNA H-bonds correspond to black bars, and RNA H-bonds correspond to gray bars.

TABLE III. Significant Direct Contacts Between Amino Acids and Nucleotides

DNA	A	C	G	T	RNA	A	C	G	U
Ala	↓	↓	↓	↓	Ala	↓	↓	↓	↓
Ile	↓	↓	↓	↓	Ile	↓	↓	—	—
Leu	↓	↓	↓	↓	Leu	↓	↓	—	—
Met	—	—	—	—	Met	—	—	—	—
Val	↓	↓	↓	—	Val	↓	↓	—	↓
Phe	↓	—	—	—	Phe	—	—	—	↓
Trp	—	53 ↑	—	—	Trp	—	—	—	—
Tyr	—	—	102 ↑	95 ↑	Tyr	—	—	—	61 ↑
Cys	—	—	—	↓	Cys	—	—	↓	—
Gly	—	—	—	—	Gly	—	—	79 ↑	↓
Pro	—	—	—	—	Pro	—	—	—	—
Arg	201 ↑	300 ↑	359 ↑	262 ↑	Arg	90 ↑	153 ↑	141 ↑	116 ↑
Asp	↓	—	↓	↓	Asp	—	—	89 ↑	—
Glu	↓	↓	↓	↓	Glu	↓	—	—	↓
Lys	249 ↑	204 ↑	297 ↑	314 ↑	Lys	107 ↑	132 ↑	104 ↑	—
Asn	—	134 ↑	109 ↑	137 ↑	Asn	—	—	—	75 ↑
Gln	—	—	—	—	Gln	—	—	—	—
His	—	—	87 ↑	76 ↑	His	—	—	40 ↑	40 ↑
Ser	—	—	143 ↑	174 ↑	Ser	—	—	—	—
Thr	—	146 ↑	132 ↑	154 ↑	Thr	—	—	—	—

The direction of the arrows indicates whether the number of interactions was higher (up arrow) or lower (down arrow) than expected from chance. Number for favored pairs is highlighted in gray. Pairs that do not contribute significantly to the interaction are represented by dashes, while disfavored pairs are represented by a down arrow.

in a similar occurrence of salt bridges in both kinds of complexes.

Hydrogen bonding analyzed from the nucleotide atoms point of view show a clear difference between DNA and RNA complexes (Fig. 4). In DNA complexes, 62% of the H-bonds involve nucleic phosphate atoms, while in the RNA complexes, H-bonds with phosphates are only 27% of all H-bonds. Although proteins interact twice more frequently with DNA than RNA phosphates, our results show that H-bonds, not salt bridges, are responsible for this increase.

Hydrophobic interactions account for 19% of all protein–nucleic acid interactions (Table II). Pairs of hydrophobic atoms occur between sugar C—H and aliphatic or aromatic side-chain C—H. Hydrophobic interactions are the main type of contacts of sugar. Approximately 63% of the DNA-deoxyribose and 42% of RNA-ribose interactions are hydrophobic contacts. This type of interaction has rarely been taken into account, but could have a stabilizing role in the protein–nucleic acid complexes.

Finally, 22% of protein–DNA interactions, and 27% of protein–RNA interactions are other van der Waals interactions (Table II). On the nucleotide side, these van der Waals interactions mainly involve the sugar–phosphate backbone of nucleic acids [62% in protein–DNA and 60% in protein–RNA interactions (data not shown)]. The presence of an additional hydroxyl group on the ribose cycle of RNA could explain the higher incidence of this type of interaction type in protein–RNA complexes. Concretely, this hydroxyl group is involved in 38% of van der Waals contacts (11% of all protein–RNA interactions). On the amino acid side, although side-chains are the most frequently implicated, the involvement of the protein backbone atoms makes 43% of the protein–DNA interactions and only 27% in the protein–RNA complexes (data not shown).

Analysis of the Amino Acid–Nucleotide Pairs

Matrices of the 80 possible pairs are shown in Table III. The observed frequency of pairs was compared to their expected frequency if interactions were by chance. A statistical analysis shows that pairs have a frequency different from random with an associated p value $\ll 0.001$ (χ^2 value for the entire table = 2101 for protein–DNA pairs and 837 for protein–RNA pairs). In addition, this test enabled us to detect which pairs were important for recognition specificity (boldface numbers in Table III) (see Materials and Methods section).

In DNA complexes, pairs with Ala, Ile, Leu, Val, Glu, and Asp are disfavored (see down arrows in Table III), while pairs with positively charged residues are favored (up arrows in Table III). Asn and Thr are favored except in interaction with A, while His and Ser are favored in interactions with G and T. The results for aromatic residues show that Phe behaves like an aliphatic residue (Phe-A pair is disfavored), while Trp and Tyr adopt behavior similar to polar residues (Trp-C, Tyr-G, and Tyr-T pairs are favored).

In RNA complexes (Table III), pairs with positively charged amino acids are favored. Among polar amino acids, only Asn-U, His-G, and His-U are favored. Globally, pairs with Ala, Leu, Ile, and Val are disfavored. The Phe-U pair is disfavored, while Tyr-U is favored as in DNA complexes. The negatively charged amino acids yielded more surprising results: Only Glu-A and Glu-U are disfavored, while Asp is favored in interactions with G. The fact that the negative charges of Asp and Glu are less damaging in the protein–RNA complexes could be due to the structure of the RNA nucleotides. Indeed, because of the single-stranded structure of RNA, a protein might interact with RNA with no need to overcome the potential energy barrier produced by the negatively charged phosphates.

Analysis of Statistically Significant Pairs

Significant pairs highlighted by the Pearson χ^2 test were classified according to their individual χ^2 values and, the most favored pairs were selected (see Material and Methods section). To analyze these pairs, we made atomic interaction matrices depicting all interactions of a favored pair. In these matrices, atoms were classified according to whether they were from amino acid side-chain or backbone, and from nucleotide phosphate, sugar, or base edge. We also highlighted the type of interaction. Atomic matrices of the most significant pairs are summarized in the Supplementary Material Table.

Most significant pairs in protein–DNA complexes

For the DNA complexes, favored pairs are as follows: Arg-G $\gg$ Arg-C $>$ Lys-G $>$ Lys-T $\gg$ Arg-T $\gg$ Lys-A $\gg$ His-G $>$ Trp-C $>$ Asn-C $>$ Tyr-G $>$ Asn-T $>$ Arg-A $>$ Ser-T $>$ Thr-T $>$ Thr-C $>$ Lys-C $>$ His-T $>$ Thr-G $>$ Tyr-T $>$ Asn-G $>$ Ser-G. The characteristics of the most significant pairs (associated p value $< 1 * 10^{-6}$ with $df = 1$) are described below.

Arg-G. With 359 pairs out of all 7671 pairs, Arg-G is the most frequent, and by far the most favored. Of these 359, 47% involve G base edge (see Supplementary Material Table). This is twice as many as in all nucleotide interactions (24% in Fig. 4 and in Supplementary Material Table, first row). Of the 47%, 87% (145 pairs) are H-bonds between amine hydrogen atoms of Arg and acceptor atoms of G (N7 and O6). The hydrogen atoms on the amine group of Arg are involved in 80% of all Arg-G interactions. Electrostatic interactions make up 28% of all Arg-G pairs (see Supplementary Material Table). Backbone atoms of Arg are rarely involved in direct contact with DNA. Those interactions involve almost exclusively the phosphate oxygens (4.5%).

Arg-C. The second most significant pair is Arg-C. While electrostatic contacts are still numerous (29%), H-bonds between Arg side-chain and C base edge are much less numerous than within the Arg-G pairs (only 22 contacts out of 300; 7.5%) (see Supplementary Material Table). This low incidence of H-bonds could be explained by the fact that cytosine has a single acceptor site on the minor groove and a single donor atom on the major groove. On the other hand, van der Waals interactions between Arg and deoxyribose or base edge atoms correspond to a third of the 300 contacts, and are almost exclusively contacts between positively charged hydrogens of Arg and cytidine hydrogens (see Supplementary Material Table). Finally, 28% of the interactions with sugar atoms are hydrophobic contacts involving the aliphatic hydrogens of the Arg side-chain.

Arg-T. Arg-T pairs are the third most prevalent (262 pairs). With T, Arg behaves in the same way as with C (see above). More than 50% of these contacts involve phosphate atoms, while H-bonds with base edge are less frequent than in Arg-G contacts (see Supplementary Material Table). The low occurrence of H-bonds could be explained by the fact that thymine has only one acceptor site on both the minor and the major grooves.

Lys-G. Like Arg-G pairs, contacts between Lys and guanidine (297 pairs) are mostly H-bonds between the amine hydrogens and the two acceptor atoms of G base edge (N7 and O6; 18%), and electrostatic interactions (30%) (see Supplementary Material Table). The Lys backbone is involved in only 15% of the pairs, 60% of which are H-bonds between Lys N—H and phosphate oxygens.

Lys-T. Of the 314 Lys-T pairs, 62% involve the phosphate of thymidine (mean = 47% in the DNA interacting data set): 57% are electrostatic contacts and 40% are H-bonds. Of the hydrophobic contacts (17%), 67.5% involve sugar C—H, and 32.5% the methyl group of thymine (see Supplementary Material Table).

Lys-A. The main interactions for the 249 Lys-A pairs are electrostatic contacts and H-bonds with 34% and 30.5%, respectively (see Supplementary Material Table).

Hydrogen bonds mainly involve phosphate oxygen atoms (23% of all interactions). Of the Lys-A interactions, 14% are hydrophobic contacts between peptidic side-chain C—H and deoxyribose C—H. Unexpectedly, specific H-bonds with acceptor atoms of adenine (N3 and N7) represent only four out of the 249 contacts.

His-G. The seventh favored pair, His-G, involves mainly H-bonds (75 out of 87), principally linking C—H of the His cycle and oxygen of phosphate (34%) or acceptor atoms of guanine (40%) (see Supplementary Material Table).

Trp-C. Trp represent only 1.5% of the amino acids in interaction with nucleotides, and about half of these interactions (53 out of 120 interactions) occur with cytidine. Of the Trp-C contacts, 51% are H-bonds with phosphate oxygen, and 26% are hydrophobic contacts between side-chain C—H and sugar C—H (see Supplementary Material Table).

Asn-C. Asparagine interacts predominantly with cytidine and mainly via H-bonds (82 contacts out of 134) (see Supplementary Material Table). Backbone atoms are involved in 34% of these contacts. The backbone C—O interact with the N—H group at position 4 of the cytosine and, the backbone N—H with the oxygen of the nucleotide phosphate. Phosphate oxygens are also involved in H-bonds with atoms of the amino acid side-chain (34% of Asn-C H-bonds). Of the 37 van der Waals contacts, 38% involve hydrogen atoms of the Asn amine function and C—H of deoxyribose. Asn-T pairs, also favored, show a similar pattern of interactions.

Tyr-G. Of the 4 possible Tyr-nucleotide pairs, the only pair fitting our selection criteria is with G (102 contacts). The presence of the hydroxyl group on the phenyl cycle enables tyrosine to make H-bonds with acceptor atoms of DNA. Such H-bonds are 30% of all Tyr contacts and correspond to 55% of all H-bonds. Hydrophobic contacts between aromatic C—H and sugar C—H are only 7% of the interactions (see Supplementary Material Table). The polar nature of tyrosine is important in its interactions with DNA. The acceptor atoms of guanine are rarely involved; they occur in only five contacts.

Most significant pairs in protein–RNA complexes

For the RNA complexes, the classification of the favored pairs is as follows: Arg-C > Arg-G ≫ Lys-C > Asp-G ≫ Arg-U > Tyr-U > Asn-U > Lys-G > Gly-G > His-G > Lys-A > His-U > Arg-A. The characteristics of the most significant pairs (associated p value $< 1 * 10^{-6}$ with $df = 1$) are described below.

Arg-C. With 153 out of 3367 pairs in protein–RNA complexes, Arg-C is the most prevalent. Of these 153 contacts, 64% involve amine hydrogens of Arg, and 21% are electrostatic interactions (see Supplementary Material Table). Another 15% are H-bonds with acceptor base edge atoms (O2 and N3), and 9% are hydrophobic contacts between peptidic side-chain C—H and ribose C—H. The hydroxyl group of RNA sugar is involved in 28 contacts out of the 153 cases (18%). Hydroxyl interacts mainly (61%) via H-bonds with amine hydrogens.

Arg-G. Again, amine hydrogens are the most frequently implicated (62%). Of the 141 Arg-G contacts, 24% are salt bridges, and 16% are H-bonds with acceptor atoms of guanine (O6 and N7). Hydrophobic contacts between peptidic backbone C_α-H and ribose C—H account for 8% of the pairs. Backbone atoms form 16% of the interactions (see Supplementary Material Table). This is more than in other Arg-nucleotide pairs, but still about half the mean of RNA interaction in the whole data set (29%) (see Supplementary Material Table, first row). The hydroxyl group of ribose is often involved (14%), but with no clear preference for any Arg atom.

Lys-C. Of the 132 Lys-C pairs, 48% involve phosphate oxygens (67% of these are electrostatic contacts). This is twice the average in all protein–RNA complexes (22%). Hydrophobic contacts make up 15% of the pairs, and 42% of interactions are with side-chain C—H of Lys (see Supplementary Material Table).

Asp-G. The fourth significant pairing involves a commonly favored nucleotide (G) and an unexpectedly favored amino acid: aspartic acid. Of the 89 pairs detected, 64 involve G base edge (72%). This is twice the average in the whole RNA complexes (34%) (see Supplementary Material Table). Of the contacts with base edges, 67% are H-bonds between the carboxylic oxygens of Asp and two hydrogens of guanine: H1 and H2. In double-helical nucleic acids, these two hydrogens are involved in specific H-bonds between G and C. It seems that Asp interacts mainly with unpaired G, by making H-bonds similar to those that stabilize the Watson and Crick helix.[33]

One example of specific Asp-guanine H-bonds takes place in the recognition of tRNA by tyrosyl-tRNA synthetase of *Methanococcus jannaschii* (1J1U; Fig. 5). Tyrosyl-tRNA synthetase recognizes specifically a small number of nucleotides: the anticodon, the characteristic C502-G573 base pair and the discriminator base A574.[35] Asp-G interaction is one of the three most important contacts with anticodon. This point highlights the effectiveness of the method and will be discussed later.

Arg-U. Arg-U pairs (116) behave similarly to Arg-C. Electrostatic interactions are frequent (27%) (see Supplementary Material Table), and contacts between amine hydrogens and the ribose hydroxyl group occur in 16% of all pairs. Amine hydrogens are involved in 30 contacts (26%) with ribose oxygens, but in only 10 specific H-bonds (9%) with base edge acceptor atoms (O4 and O2).

Tyr-U. Results for the 61 Tyr-U pairs are quite unexpected. Backbone atoms, most of which were oxygens, account for 54%, whereas 46% are H-bonds between backbone oxygens and amine hydrogens of uracil (H3). In double-helical nucleic acids, H3 forms a specific H-bond to N1 of adenine. The remaining interactions are mostly pairs involving the hydroxyl group of Tyr (23%), and hydrophobic contacts between ribose C—H and aromatic C—H (15%) (see Supplementary Material Table).

Asn-U. Asn interacts with RNA principally through its amide function (79% of the 75 pairs). Uracil involves

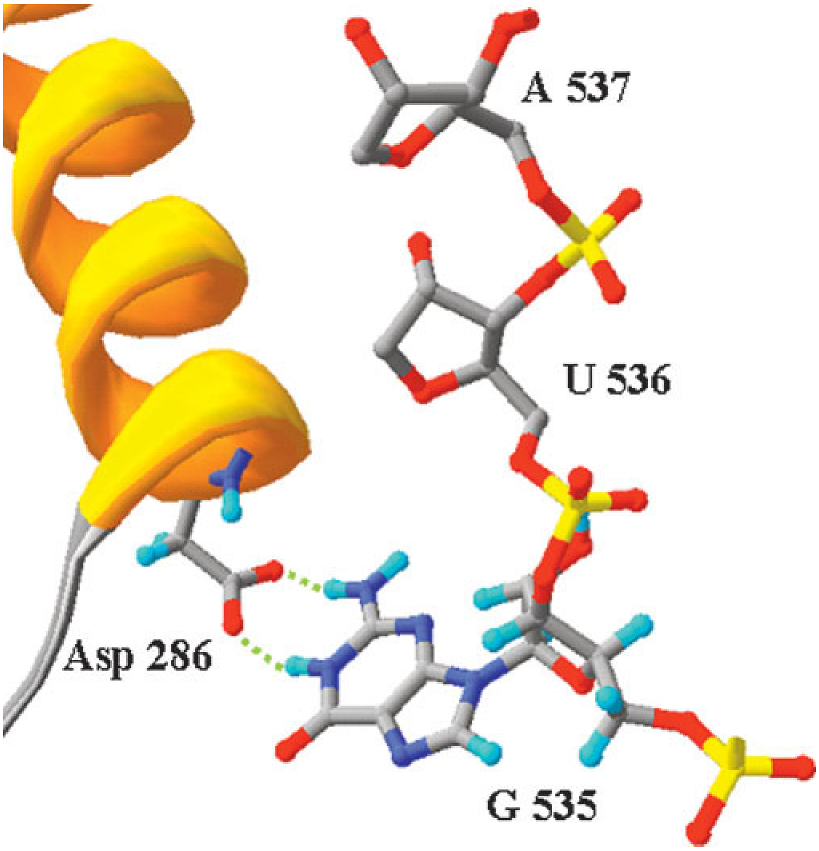

Fig. 5. Anticodon specific recognition by the tyrosyl-tRNA synthetase of *Methanococcus jannaschii* (1J1U). H1 and H2 of G535 are recognized through two H-bonds, with the two oxygen atoms of Asp286. The 5′-GUA-3′ anticodon backbone is shown as a stick model. Amino acids from Leu282 to Ile296 are represented as a ribbon model, and the α-helix is in yellow. All atoms of Asp 286 and G535 are shown in stick model. Image produced using Swiss-PdbViewer.[34]

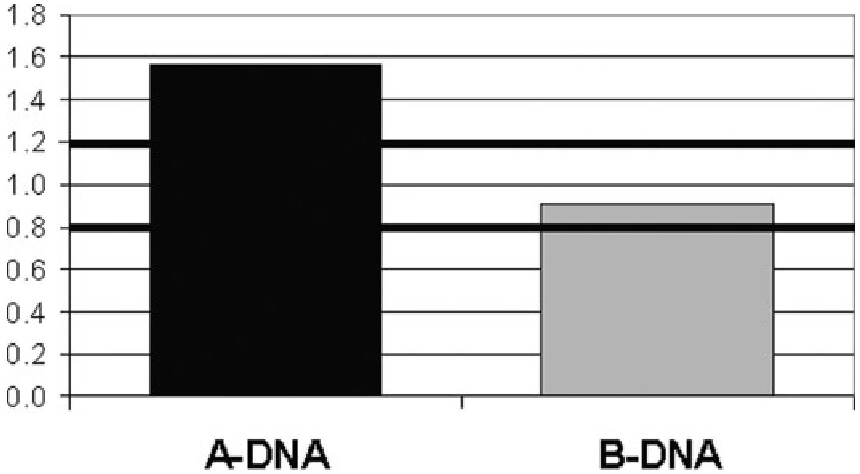

Fig. 6. Protein propensity to interact with the different DNA structures. The propensity values correspond to the frequency of a DNA structure type in the interaction sites divided by the frequency of the same DNA structure type in the whole database. Propensity of A-DNA to interact is represented by a black bar, and that of B-DNA by a gray bar. Propensity values higher than 1.2 correspond to favored DNA structure, whereas disfavored DNA structure should have values lower than 0.8.

mainly base edge atoms (56%) (see Supplementary Material Table). Hydrogen bonds occur between acceptor atoms of U (O2 and O4) and hydrogens of the amine group (33%), and between hydrogens usually involved in specific H-bonds of the nucleotide helix (H3) and the oxygen from the amide group (8%).

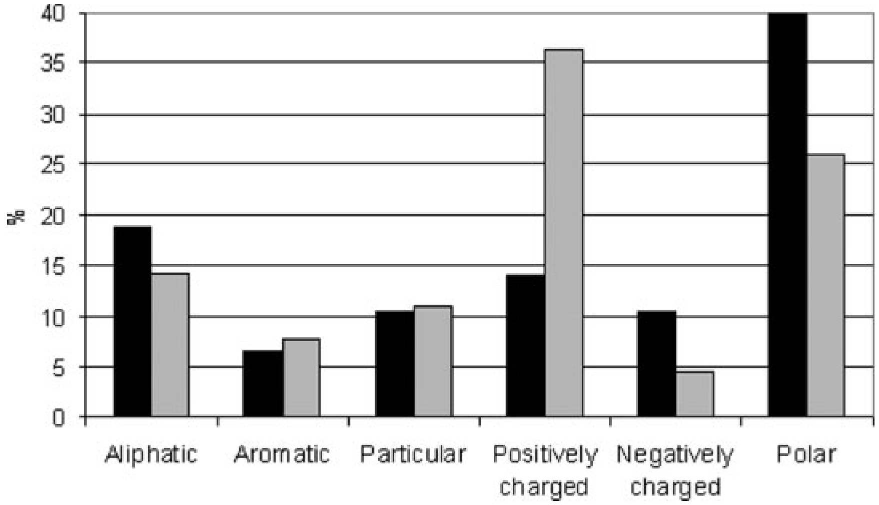

Fig. 7. Frequency distribution of the amino acid families in direct contact with double-helical DNA. The black bars and the gray bars correspond to amino acids interacting with A-DNA and B-DNA, respectively. Results are from the double-helical high-resolution (< 2 Å) data set.

Implication of the DNA Structure in Interactions

As described in the Materials and Methods section, each nucleotide in a double-helical DNA structure was classified as A-DNA or B-DNA, depending on its δ–χ angles. The B-DNA form is the main structure in the whole database of high-resolution double-helical DNA complexes, representing 93% of the 1195 nucleotides. For nucleotides in direct contact with proteins, B-DNA is less frequent but still predominant, whereas the propensity values (Fig. 6) show that A-DNA is largely favored in the interaction sites (1.6). Although A-DNA is favored in the interaction sites, the high proportion of B-DNA in the database prevented us from a definitive conclusion.

The amino acids interacting with these two forms of DNA are distributed quite differently (Fig. 7). Since B-DNA nucleotides are the most frequent, interacting amino acids are not different from what we described up to now [compare gray bars of Fig. 7 to white bars of Fig. 1(A)]. When amino acids interacting with A-DNA are analyzed, Ala, Asn, and Asp are twice as frequent (11%, 15%, and 7.5% instead of 4.5%, 6.5%, and 3%, respectively), while Arg and Lys are less frequent (4% instead of 14.5%, and 9.5% instead of 14%, respectively) [compare black bars of Fig. 7 to white bars of Fig. 1(B)]. Hence, amino acids in interaction with the A-DNA are more similar to those in interaction with RNA.

DISCUSSION

In this work, we analyzed the structural features of a large number of nonhomologous protein–nucleic acid complexes. We extracted 7671 pairs from 139 protein–DNA complexes and 3367 from 49 protein–RNA complexes. We distinguished atoms belonging to protein side-chain or backbone, and atoms belonging to nucleotide phosphate, sugar, or base edge. We also detailed which atomic contacts were involved in the most significant amino acid–nucleotide pairs. Figure 8 shows a summary of the main interactions observed in this study.

Amino Acid Composition

We compared the amino acid composition of our databases to that of SWISS-PROT sequence database. The major difference is for protein–DNA complexes, where hydrophilic residues are more frequent, especially positively charged amino acids. Within interaction sites, hydrophilic residues account for more than 60%. Arg and Lys have the highest propensities to interact with nucleotides, followed by polar residues (His, Ser, and Thr for DNA complexes, and Asn, His, and Gln for RNA complexes). Tyr is also slightly favored. The preference for Arg, Lys, Asn, and Ser has previously been shown in a database of 25 RNA-complexes and 20 ribosomal chains.[7] The high propensity of the two positively charged residues is attributed to an interaction with the negative charge of the nucleotide phosphate.[7,8] Our results show that in DNA, Arg has a higher propensity to interact with base edge than with the rest of the nucleotide, whereas in protein–RNA complexes, Arg interacts preferentially with phosphate groups. Electrostatic interactions are considered to be the main way by which the positively charged residues interact with nucleotides. However, our results demonstrate frequent interactions between basic amino acids and base edge.

Nucleotide Atom Types and Interaction Types

In protein–DNA complexes, 47% of nucleotide atoms in interaction are phosphate atoms, whereas only 24% are base edge atoms. These results are similar to those obtained by others based on 65 double-stranded DNA complexes (43% with phosphate atoms and 27% with base edge atoms[8]). In protein–RNA complexes, there are 1.5 times more contacts with base edges in RNA than in DNA complexes. The dynamic nature of RNA and the diverse structures it can take,[36] together with the composition of our databases (in which 34 of the 49 RNA structures are single-stranded, while 118 DNA complexes out of 139

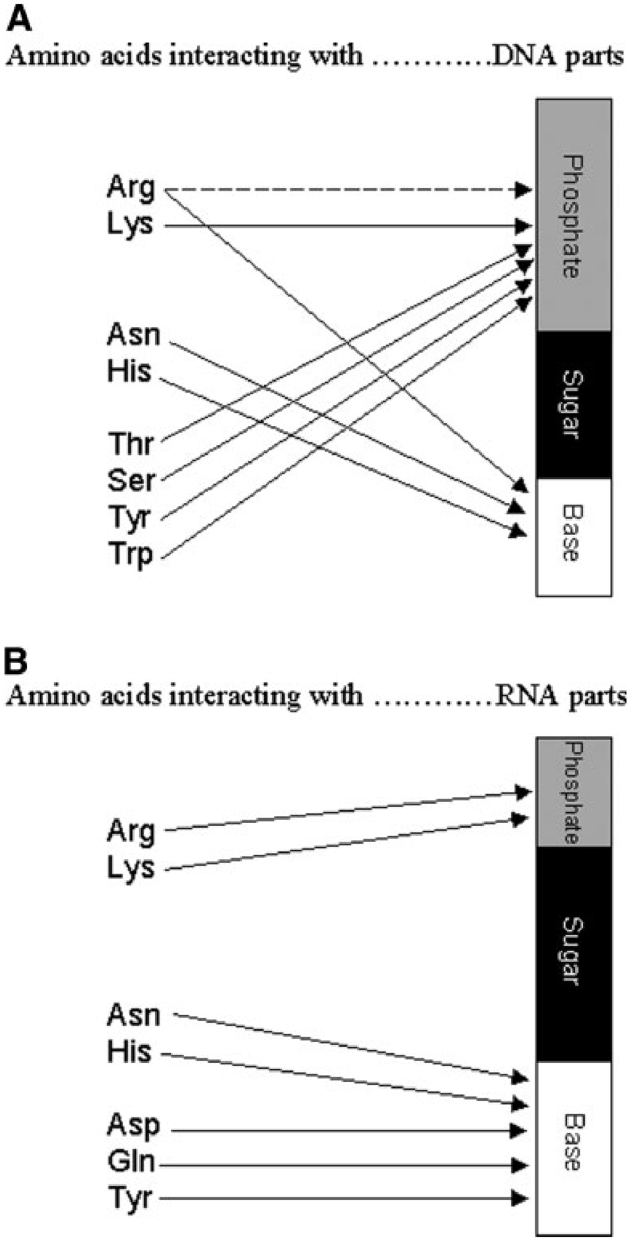
A
Amino acids interacting with ....DNA parts
Arg
Lys
Asn
His
Thr
Ser
Tyr
Trp
Phosphate
Sugar
Base
B
Amino acids interacting with ....RNA parts
Arg
Lys
Asn
His
Asp
Gln
Tyr
Phosphate
Sugar
Base

Fig. 8. Schematic representation of the main interaction observed between favored amino acids and the different parts of DNA (**A**) and RNA (**B**). Amino acids are sorted depending on their propensities to interact with nucleotides.

contained double-helical nucleotides) could explain this difference.

Most studies of protein–nucleic acid interactions have focused on H-bonds, particularly H-bonds with base edge. Here, we considered all types of interactions (H-bonds, electrostatic, hydrophobic, and other van der Waals interactions). The four interaction classes have a similar distribution pattern in both protein–DNA and protein–RNA complexes. Hydrogen bonds are the most frequent, accounting for more than 50% of the contacts. The number of H-bonds concurs with the results of Jeong et al.[37] for RNA complexes, and of Mandel-Gutfreund et al.[3] for DNA complexes. In protein–RNA complexes, 46% of the H-bonds involve base edge atoms. This is similar to the 51/49 ratio (RNA backbone/RNA base edge) found by Jeong et al.[37] In contrast, H-bonds with DNA base edges are only 24% [31% in the study by Mandel-Gutfreund et al.[3]], and most H-bonds are with phosphate (62%). Water-mediated H-bonds could also account in a similar proportion to classic H-bonds.[37] Because the position of water molecules is reasonably accurate for a resolution better than 2Å, and because this limitation would have drastically limited our databases, we have not examined these interactions.

Electrostatic interactions have been shown to be important for long distance preorientation of interacting proteins[11] and for the stabilization of complexes. In our results, they represent 8% of the contacts in DNA complexes and 7% in RNA complexes. This correlates with the large number of positively charged residues involved. Hydrophobic contacts stabilize protein–protein interactions.[38–40] They were rarely taken into account in nucleotide binding.[18,41] In our databases, almost a fifth of the contacts are hydrophobic and involve mainly sugar atoms. These results suggest that, as in protein–protein interaction, hydrophobic contacts may contribute to the stability of protein–nucleic acid complexes, whereas H-bonds with base edge and electrostatic interactions are important for specificity and stability, respectively.

Finally, 22% of all interactions in protein–DNA complexes and 27% in protein–RNA complexes are other van der Waals interactions. These involve mainly nucleic acid backbone (62% in DNA and 60% in RNA complexes) and amino acid side-chains. In a recent article, Jones et al.[24] calculated larger ratios of van der Waals interactions: 76% for double-stranded DNA–protein complexes, 93% for single-stranded DNA–protein complexes, and 92% for RNA–protein complexes. Besides the fact that their RNA and single-stranded DNA data sets contained few complexes (20 and 3, respectively), their more restrictive definition of H-bonds (angle, distance, and atom types) accounts for these broad differences.

Statistically Significant Pairs

Of the 80 possible amino acid–nucleotide pairs, we determined which occurred the most differently from random. In DNA and RNA complexes, pairs with Arg and Lys are always more numerous than by chance, and involve mainly salt bridges or H-bonds with base edges. As reported in other studies,[5,42] Arg-G pairs are widely favored in DNA complexes. The presence of two hydrogen acceptors on the major groove face of guanine could explain this. In RNA complexes, Arg-G pairs are also favored, but H-bonds involving base edge are half as numerous.

Pairs involving polar amino acids are also favored and ~60% of these interactions are H-bonds. Out of the 10 pairs we highlighted in DNA complexes, five were shown to be favored in another study.[43] In RNA complexes, Asp-G pairs are favored, particularly in direct contacts with RNA base edges. This surprising interaction between a negatively charged residue and negatively charged RNA could be due to the ability of the two side-chain oxygen atoms of Asp to make H-bonds with H donor atoms at positions 1 and 2 on guanine.[37] The specific interactions between Asp and guanine are involved in recognition of tRNA anticodon by tRNA synthetase [Fig. 5 for 1J1U complex,[35] and in 1H3E bacterial tyrosil-tRNA synthetase–tRNA complex[44]].

DNA Double-Helical Structure

It has been shown that some DNA-binding proteins interact with bent DNA.[12,13] This DNA deformability clearly plays a role in recognition for many complexes.[45,46] We decided to partially assess this deformability by studying the influence of the DNA structure on interactions. Indeed, Vargason et al.[47] and Dickerson and Ng[48] discussed the transition from B-DNA to A-DNA, and suggested that helical form might influence interactions. To analyze such effects, we classified each nucleotide according to its A-DNA or B-DNA conformation on the basis of χ and δ torsional angles.[27] Our results suggest that the A-DNA is more frequently implicated in protein–DNA interactions than the classical B-DNA conformation. This finding is consistent with a structure-based study performed by Tolstorukov et al.[41] and with experimental results obtained by Elrod-Erickson et al.[21] The nucleotides of the Zif268 zinc finger complex adopt an intermediate conformation between the A- and B-DNA at the binding site, and binding with canonical B-DNA was shown to be less efficient.[21] We find that amino acids interacting with A-DNA are more like those in interaction with RNA. In particular, few basic amino acids are implicated, and Asp is favored in the interacting sites. This correlates with results showing an increased accessibility of DNA bases when amino acids interact with A-DNA nucleotides.[41]

CONCLUSION

This study has presented a structural analysis of protein–nucleic acid recognition mechanisms. The growth of the database of available structures since prior studies allows us to carry out a more global analysis, especially for RNA–protein complexes. We analyzed the interactions at an atomic level and observed the influence of the structure of double-helical DNA in a database of high-resolution complexes.

Positively charged and, to a lesser extent, polar amino acids are clearly favored in the interaction sites. Nevertheless, differences appear between DNA and RNA complexes: Single-stranded RNA chains allow direct contacts with RNA base edges more frequently and easily than with double-helical DNA chains. Moreover, when compared with protein–RNA phosphate contacts, the greater number of interactions with DNA phosphate is due to an increased number of H-bonds, not salt bridges.

In addition to electrostatic interactions and H-bonds with base edges, we also show that hydrophobic contacts are numerous and could have a stabilizing role in protein–nucleic acid interactions.

Statistical analysis has shown that pairs between Arg or Lys and any nucleotide are significantly favored in both protein–DNA and protein–RNA complexes. Polar residues and Tyr are also favored. In protein–RNA complexes, Asp-G pairs are highlighted. We show that these unexpected pairs are involved in the highly specific recognition of the tRNA anticodon, thereby validating our statistical methodology. Finally, we have studied the effects of DNA structure and show that the A-DNA conformation has an advantage for interaction.

REFERENCES

1. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein–DNA complexes. Genome Biol 2000;1:1–10.
2. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci USA 1976;73:804–808.
3. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol 1995;253:370–382.
4. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem 1992;61:1053–1095.
5. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res 2001;29:2860–2874.
6. Mandel-Gutfreund Y, Margalit H, Jernigan RL, Zhurkin VB. A role for CH…O interactions in protein–DNA recognition. J Mol Biol 1998;277:1129–1140.
7. Treger M, Westhof E. Statistical analysis of atomic contacts at RNA–protein interfaces. J Mol Recognit 2001;14:199–214.
8. Nadassy K, Wodak SJ, Janin J. Structural features of protein–nucleic acid recognition sites. Biochemistry 1999;38:1999–2017.
9. Jayaram B, Jain T. The role of water in protein–DNA recognition. Annu Rev Biophys Biomol Struct 2004;33:343–361.
10. Pabo CO, Sauer RT. Protein–DNA recognition. Annu Rev Biochem 1984;53:293–321.
11. Drozdov-Tikhomirov LN, Linde DM, Poroikov VV, Alexandrov AA, Skurida GI. Molecular mechanisms of protein–protein recognition: whether the surface placed charged residues determine the recognition process? J Biomol Struct Dyn 2001;19:279–284.
12. Dickerson RE. DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic Acids Res 1998;26:1906–1926.
13. Dickerson RE, Chiu TK. Helix bending as a factor in protein/DNA recognition. Biopolymers 1998;44:361–403.
14. Zhang Y, Xi Z, Hegde RS, Shakked Z, Crothers DM. Predicting indirect readout effects in protein–DNA interactions. Proc Natl Acad Sci USA 2004;101:8337–8341.
15. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. Proc Natl Acad Sci USA 1998;95:11163–11168.
16. ElHassan MA, Calladine CR. Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. Philos Trans R Soc London A Math Phys Eng Sci 1997;355:43–100.
17. Selvaraj S, Kono H, Sarai A. Specificity of protein–DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. J Mol Biol 2002;322:907–915.
18. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. Nucleic Acids Res 1998;26:2306–2312.
19. Choo Y, Klug A. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. Proc Natl Acad Sci USA 1994;91:11168–11172.
20. Choo Y, Klug A. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc Natl Acad Sci USA 1994;91:11163–11167.
21. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO. Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. Structure 1996;4:1171–1180.
22. Elrod-Erickson M, Benson TE, Pabo CO. High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. Structure 1998;6:451–464.

23. Draper DE. Themes in RNA–protein recognition. J Mol Biol 1999;293:255–270.
24. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein–RNA interactions: a structural analysis. Nucleic Acids Res 2001;29:943–954.
25. Perez-Canadillas JM, Varani G. Recent advances in RNA–protein recognition. Curr Opin Struct Biol 2001;11:53–58.
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
27. Lu XJ, Shakked Z, Olson WK. A-form conformational motifs in ligand-bound DNA structures. J Mol Biol 2000;300:819–840.
28. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
29. Dearing A. Computer-aided molecular modelling: research study or research tool? J Comput Aided Mol Des 1988;2:179–189.
30. Thomas A, Benhabiles N, Meurisse R, Ngwabije R, Brasseur R. Pex, analytical tools for PDB files: II. H-Pex: noncanonical H-bonds in alpha-helices. Proteins 2001;43:37–44.
31. Thomas A, Bouffioux O, Geeurickx D, Brasseur R. Pex, analytical tools for PDB files: I. GF-Pex: basic file to describe a protein. Proteins 2001;43:28–36.
32. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature 1982;299:371–374.
33. Watson JD, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 1953;171:737–738.
34. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
35. Kobayashi T, Nureki O, Ishitani R, Yaremchuk A, Tukalo M, Cusack S, Sakamoto K, Yokoyama S. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. Nat Struct Biol 2003;10:425–432.
36. Szymanski M, Barciszewska MZ, Zywicki M, Barciszewski J. Noncoding RNA transcripts. J Appl Genet 2003;44:1–19.
37. Jeong E, Kim H, Lee SW, Han K. Discovering the interaction propensities of amino acids and nucleotides from protein–RNA complexes. Mol Cells 2003;16:161–167.
38. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 1997;28:333–343.
39. Lo CL, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
40. Chothia C, Janin J. Principles of protein–protein recognition. Nature 1975;256:705–708.
41. Tolstorukov MY, Jernigan RL, Zhurkin VB. Protein–DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. J Mol Biol 2004;337:65–76.
42. Raman B, Guarnaccia C, Nadassy K, Zakhariev S, Pintar A, Zanuttin F, Frigyes D, Acatrinei C, Vindigni A, Pongor G, Pongor S. N(omega)-arginine dimethylation modulates the interaction between a Gly/Arg-rich peptide from human nucleolin and nucleic acids. Nucleic Acids Res 2001;29:3377–3384.
43. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins 1999;35:114–131.
44. Yaremchuk A, Kriklivyi I, Tukalo M, Cusack S. Class I tyrosyl-tRNA synthetase has a class II mode of cognate tRNA recognition. EMBO J 2002;21:3829–3840.
45. Bao G. Mechanics of biomolecules. J Mechanics and Phys Solids 2002;50:2237–2274.
46. Changela A, Perry K, Taneja B, Mondragon A. DNA manipulators: caught in the act. Curr Opin Struct Biol 2003;13:15–22.
47. Vargason JM, Henderson K, Ho PS. A crystallographic map of the transition from B-DNA to A-DNA. Proc Natl Acad Sci USA 2001;98:7265–7270.
48. Dickerson RE, Ng HL. DNA structure from A to B. Proc Natl Acad Sci USA 2001;98:6986–6988.