12-2010

Quantum Chemical Studies Of Nucleic Acids Can We Construct A Bridge To The Rna Structural Biology And Bioinformatics Communities?

Jiri Sponer,

Judit Sponer

Anton I. Petrov

Neocles B. Leontis
Bowling Green State University, leontis@bgsu.edu

Quantum Chemical Studies of Nucleic Acids: Can We Construct a Bridge to the RNA Structural Biology and Bioinformatics Communities?

Jiří Šponer,*,† Judit E. Šponer,† Anton I. Petrov,‡ and Neocles B. Leontis*,§

Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 61265 Brno, Czech Republic, Department of Biological Sciences, Bowling Green State University, Bowling Green, Ohio 43403, United States, Department of Chemistry, Bowling Green State University, Bowling Green, Ohio 43403, United States

Received: May 13, 2010; Revised Manuscript Received: September 16, 2010

In this feature article, we provide a side-by-side introduction for two research fields: quantum chemical calculations of molecular interaction in nucleic acids and RNA structural bioinformatics. Our main aim is to demonstrate that these research areas, while largely separated in contemporary literature, have substantial potential to complement each other that could significantly contribute to our understanding of the exciting world of nucleic acids. We identify research questions amenable to the combined application of modern ab initio methods and bioinformatics analysis of experimental structures while also assessing the limitations of these approaches. The ultimate aim is to attain valuable physicochemical insights regarding the nature of the fundamental molecular interactions and how they shape RNA structures, dynamics, function, and evolution.

Introduction

Nucleic acids (DNA and RNA) are perhaps the most important biomolecules. In modern cellular organisms, genetic information is encoded in the sequences of exceedingly long molecules of DNA. RNA molecules are produced by copying selected segments of genomic DNA ("genes") in a process called "transcription." RNA, long considered to possess few specialized functions beyond transmission of genetic information from the nucleus to the cytoplasm, has emerged as a central player in all aspects of gene expression and its regulation. The discovery of RNA enzymes (ribozymes) less than 30 years ago set the stage for one of the most exciting revolutions in modern biology.[1] Since the discovery of ribozymes, one major RNA breakthrough has followed another, revealing the biochemical versatility of RNA that enables it to play major roles in so many biological functions. In addition to their familiar roles as messenger-RNAs (mRNA) and transfer-RNAs (tRNA), RNA molecules are integral components of the cellular machineries for mRNA splicing (the splicesome),[2] for mRNA-directed protein synthesis (the ribosome),[3] and for sequence-directed protein targeting and transport to different membranes or compartments of the cell (the signal-recognition particle).[4] The ribosome, one of the most complex biomolecular machines ever evolved, is essentially a ribozyme, although its function has been refined during evolution by recruitment of several dozen protein partners. RNA forms the catalytic core and the primary functional centers of the ribosome, which successively bind the correct amino-acid-bearing tRNAs to synthesize proteins by linking together the amino acids in the order specified by an mRNA template. The catalytic core of the splicesome is also composed of RNA and likely evolved from self-splicing, autocatalytic RNA introns. Splicing is a fundamental modification of RNA after transcription, in which large RNA segments, called introns,[5] are removed and the remaining parts, called exons, are covalently joined ("spliced") to produce the mature mRNA. Early in the evolution of life, splicing was probably largely autocatalytic.

The discovery of catalytic RNA provided a new paradigm for theories of the origin of life by resolving the "chicken-or-the-egg" conundrum of which came first, DNA or protein. In modern cells, DNA codes for proteins, but proteins are needed to copy DNA in a process called "replication" that occurs when cells divide. Because RNA appears to be chemically capable of serving simultaneously as an information carrier as well as a self-replicator, it is likely that primitive cellular life was RNA-based.[6] In modern cells, RNA has ceded information storage to DNA and most catalytic functions to proteins. However, RNA has retained some of the decisive roles it probably had in primitive, hypothetical, RNA-based life forms, while acquiring new ones in the course of evolution. Although <2% of the human genomic DNA directly encodes protein sequences, over 80% of the genome is actually transcribed into RNA at some point in the life cycle. In other words, the vast majority of the genome is transcribed into nonprotein coding RNAs (ncRNA), including well-known functional RNAs, such as ribosomal (rRNA) and tRNAs. However, the functions of most ncRNAs are unknown.[7] For example, the recent discovery of small RNA molecules (micro-RNA) that regulate gene expression at multiple levels was a complete surprise.[8] Thus, the structural complexity and functional versatility of RNA molecules is much greater than that of DNA.

The last 30 years have witnessed parallel breakthrough developments in the application of physics to chemistry and biochemistry. It has long been anticipated that quantum mechanics (QM) would provide the ultimate description and understanding of molecular systems by describing electronic structures in terms of fundamental physical principles. However, meaningful practical applications of quantum mechanics (quantum chemistry) had to await the development of sufficiently powerful computer hardware and software, something achieved only in the last two decades. Subsequently, QM methods have

* Corresponding authors. (J.Š.) Phone: +420 541517133. Fax: +420 541212179. E-mail: sponer@ncbr.chemi.muni.cz. (N.B.L.) Phone: +1 419 372 8663. Fax: +1 419 372 9809. E-mail: leontis@bgsu.edu.

† Academy of Sciences of the Czech Republic.

‡ Department of Biological Sciences, Bowling Green State University.

§ Department of Chemistry, Bowling Green State University.

10.1021/jp104361m
Published on Web 11/04/2010

Jiří Šponer (1964) is presently Head of the Department of Structure and Dynamics of Nucleic Acids, Institute of Biophysics, Academy of Sciences of the Czech Republic (ASCR), Brno, Czech Republic; Professor of Biomolecular Chemistry at Palacký University, Olomouc, Czech Republic, and Masaryk University, Brno; and senior researcher at the Institute of Organic Chemistry and Biochemistry, ASCR, Prague, Czech Republic. He earned his M.Sc. and Ph.D. at Masaryk University, Brno. Since 1992, he has been primarily associated with ASCR, initially working mainly with Pavel Hobza. His primary research interests are applications of modern quantum-chemical and molecular simulation methods in studies of structure, dynamics, function, and evolution of nucleic acids.

Judit E. Šponer is a senior researcher at the Department of Structure and Dynamics of Nucleic Acids, Institute of Biophysics, Academy of Sciences of the Czech Republic (ASCR), Brno. She obtained her M.Sc. at the Eötvös University and her Ph.D. at the Technical University in Budapest. Her current research interest is the quantum chemical modeling of nucleic acids and their components.

Anton I. Petrov took his B.Sc. degree in Biology from St. Petersburg State University, Russia, in 2007. He is currently pursuing doctoral studies in Bioinformatics at Bowling Green State University, Bowling Green, Ohio.

Neocles B. Leontis is Professor of Chemistry at Bowling Green State University (Ohio), where he has taught since 1987. He earned his B.S. in Chemistry at The Ohio State University and his Ph.D. in Biophysical Chemistry from Yale University, working with Peter Moore. He carried out postdoctoral research with David Engelke at the University of Michigan. His research interests are in RNA structural bioinformatics, RNA modeling, and RNA nanoscale self-assembly. He served as convener of the RNA Ontology Consortium 2005−2009. Currently, he is serving as a Program Director in Molecular and Cellular Biology at the National Science Foundation. (Photo courtesy of Bowling Green State University, Bowling Green, Ohio.)

emerged as powerful tools in many areas of modern chemical research. It is now, in principle, possible to carry out on modest laptop computers QM calculations that were unthinkable 20 years ago or that required the most powerful supercomputers 15 years ago. These developments have made it possible to address challenging chemical problems by close collaboration between theoretical, computational, and experimental approaches, as exemplified, for example, by the studies of Hobza and Schlag on the benzene dimer,[9] the basic results of which remain unchallenged to this day.

The QM field continues on a track of rapid and sustained methodological development, illustrated by recent innovations in density functional theory.[10] By contrast, the biomolecular modeling field, based on the description of molecular systems using classic potential functions, has not enjoyed this level of sustained progress. Despite intense efforts, for example, to develop polarization force fields,[11] variants of second-generation pair-additive force fields, first developed in the 1990s,[12] remain dominant in modeling studies of nucleic acids. Despite the enormous complexity of macromolecular biological systems, which challenges the application of theoretical approaches, QM has provided interesting and relevant results that could not be obtained by any other techniques. From early on, nucleic acids, with their well-defined molecular interactions, especially base stacking and base pairing, have been favorable targets for QM computations.[13] For example, QM calculations have clarified the physicochemical origins of base stacking, as will be detailed below.[14] QM studies revealed new phenomena, such as the intrinsic propensity of amino groups of nucleic acid bases to undergo partial sp^3 pyramidalization.[15] Modern QM calculations, carried out with expansion to complete basis sets of atomic orbitals and inclusion of higher-order electron correlation effects, provide accurate energies of molecular interactions.[16] QM calculations have addressed additional aspects of nucleic acids that are beyond the scope of this review, including metal−nucleic acid interactions, proton transfer processes, electronically excited states, effects of radiation and reactive free radicals, and chemical aspects of theories of the origin of life. QM calculations play decisive roles in parametrization of biomolecular force fields.[17]

Nevertheless, one must concede that the direct impact of QM studies on structural biology, biochemistry, and bioinformatics has remained limited. For example, the basic research of the effect of base stacking on the local conformational variability of B-DNA and the classification of RNA base pairing were accomplished without considering QM data.[18,19] The literature of nucleic acid quantum chemistry and nucleic acid structural biology and bioinformatics remain largely segregated, reflecting the lack of significant interaction between the respective communities. What are the reasons for this state of affairs? We question the facile suggestion that QM research is less relevant to structural biology than it is to other fields of molecular sciences. Because QM approaches are based on fundamental physical principles, they are the most sophisticated tools available to directly study the specific local interactions that occur widely in macromolecular structures. In the nucleic acid context, the local interactions occur between bases (base-stacking and base-pairing) or between bases and backbone moieties (e.g., base-phosphate) or involve interactions with solvent molecules, including ions. The usefulness of QM to study similar noncovalent interactions has been widely accepted in many other areas of chemistry.[20] Thus, the striking lack of interest in high-quality QM results relevant to structural biology and bioinformatics is puzzling. Obscure, outdated, and even incorrect models too often continue to circulate. A common Biochemistry textbook[21] continues to publish outdated stacking energies obtained in the 1970s by at the time affordable semiempirical approaches that are wildly in error by modern calculational standards.

What can be done to increase communication between practitioners of quantum chemistry and biochemistry, structural biology, and bioinformatics? In this feature article, we compare the methodological approaches of QM and structural bioinformatics as they pertain to molecular interactions in RNA and provide suggestions as to how these two fields could profit from greater interaction and cooperation.

The lack of communication between the QM and structural biology communities may have its origin in one salient feature of QM calculations. To be tractable, QM calculations must be carried out on sufficiently small model systems, of the order of dozens to no more than about 100+ atoms. These model systems are studied in complete isolation, that is, largely in the absence of solvent. Although QM calculations provide very accurate and physically complete descriptions of the molecular interactions in these model systems and exactly the same interactions are indeed present in nucleic acids, their influence is always realized within a context of a multitude of other effects that produces a delicate balance among all molecular interactions. This balance is so exceptionally complex that it is very difficult to correlate the calculated data derived for isolated model systems with relevant experimentally measurable quantities of interacting systems. For example, with proper attention to the choice of geometry, QM calculations can provide accurate descriptions of base stacking energies.[22] These calculations, however, do not correlate well with experimentally derived thermodynamic parameters for nucleic acids, obtained in water solution at moderate ionic strength (e.g., 1.0 M NaCl) and physiological temperatures.[23] This does not imply that the QM stacking data are irrelevant: QM calculations do provide valid and correct descriptions of one of the dominant forces in nucleic acids, information that cannot be collected by any other technique. At the same time, the experimental thermodynamic data, because of the complexity of the interactions, do not, in fact, provide unambiguous measures of the strength of the direct base−base interactions. The experimental measurements reflect the overall free energies associated with a given nucleic acid structure and sequence and are not dissectible into the contributions of individual interactions that would be equivalent to the QM data. Thus, QM calculations and thermodynamics experiments reflect different aspects of the base stacking phenomenon, and although both descriptions are valid, neither is complete by itself. To achieve the best possible insight into base stacking, we need to integrate both sources of information.

In summary, we need to keep in mind that biomolecular systems are exceptionally complex, and only a fraction of problems in biology can be addressed in a comprehensive manner by computations. It is crucial not to overrate the capabilities of QM tools in biology. Nevertheless, considering the importance and richness of biology, it pays to apply computations to accessible problems, where they can provide valuable, and often unexpected, insights. Overinterpretation of computational results sometimes occurs in another area of computational chemistry, nucleic acid molecular modeling based on molecular mechanics (MM) force fields. Some studies push molecular dynamics (MD) simulations entirely beyond the limits of the force fields, a practice that reflects a lack of attention to the limitations imposed by the underlying approximations of current techniques.[24] This is usually not so great a problem in the QM literature because most QM practitioners are attentive to the limitations of accuracy. The greater difficulty of extrapolating results obtained from QM studies of small model systems to intact biomolecular systems, however, can easily lead to oversimplification, even when the model studies per se are correctly executed.

The parallel application of the QM and MM methodologies to the same nucleic acid systems has great potential, but is rarely attempted. The interpretation of many QM studies would profit from complementary, explicit solvent, classical MD simulations. Likewise, simulation studies can gain credibility when the limitations of force field approximations are assessed by insights from QM calculations. Wisely combining the two computational techniques gives us more space for maneuvering when facing the daunting challenges posed by biochemical and biological systems. The direct integration of QM and MM descriptions has produced genuine hybrid QM/MM methods, suitable for studies of enzymatic reactions.[25]

Another source of misunderstanding between the QM and structural biology communities concerns problems of scientific communication, which in principle can be solved. QM studies are written in a style and terminology that limits accessibility to nonexpert readers. Consequently, potentially valuable results and insights do not reach the audience that would most benefit. At the same time, most structural biologists and bioinformaticists largely ignore the QM literature, assuming that it does not contain relevant information. This is partially understandable, because only a few nonspecialists have time to follow the computational literature in sufficient depth, given the huge volume of new biological literature they need to follow in their own fields. In addition, it is difficult for nonexperts to sift out the most relevant studies from the large number of papers in the computational literature. Nonetheless, peremptory dismissals of theory tip too far in the other direction, and the result is that a significant number of carefully done theoretical and computational studies have been ignored, which could aid in the interpretation and understanding of much experimental data.

Purposes of QM Calculations. QM calculations quantify crucial properties of molecular systems. For some of these properties, QM calculations represent the only available tool

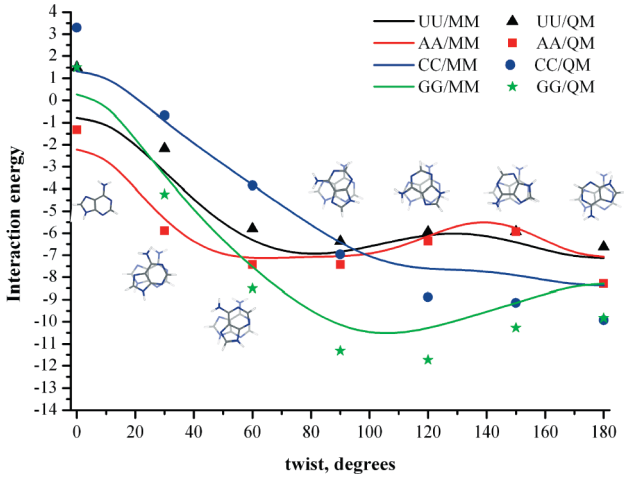


Figure 1. Modern QM theory to base stacking.[16f] The dependence of base stacking energy on twist angle between the nucleobases in A/A, U/U, C/C, and G/G base−base stacks, calculated for undisplaced face-to-back nucleobase dimers. The geometries of A/A stacks are shown to illustrate the twist angle. The solid lines represent force field calculations using the Cornell et al. (AMBER)[12] MM force field with point charges derived to fit the electrostatic potential of the monomers at the MP2 level of theory. MP2 charges are more appropriate for direct comparison with the QM data than the more polar HF charge distributions recommended for condensed phase simulations.[14b] The QM data (the specific symbols) are calculated with the MP2 method expanded to the complete basis set (CBS) of atomic orbitals and corrected for higher-level electron correlation effects. Although the agreement between the force field and QM is not perfect, there are no major deviations. The Figure illustrates one of the key applications of QM methods: point by point comparison of the rigorous data with approximate descriptions.[14b] In this particular case, the lack of substantial deviations between QM and force field over the whole potential energy surface indirectly clarifies the nature of base stacking.[16]

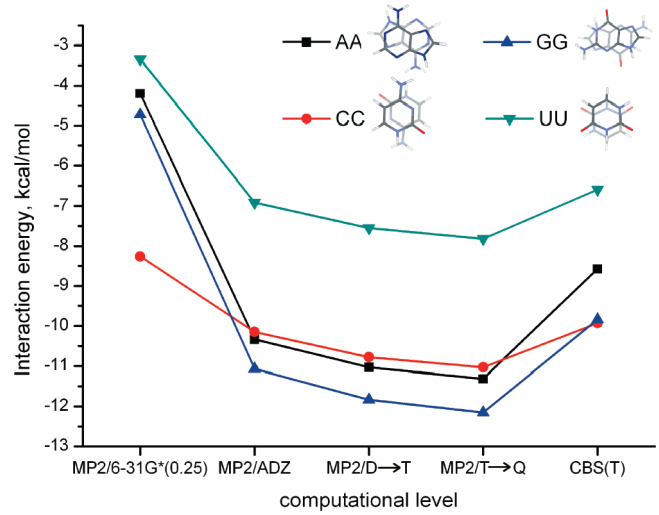


Figure 2. Convergence of stacking energies for antiparallel undisplaced face to back arrangements of A/A, U/U, C/C, and G/G stacks. The data show MP2/6-31G*(0.25) method (reference method in 1990s), MP2/aug-cc-pVDZ (ADZ) calculations, MP2/CBS calculations using aug-cc-pVDZ → aug-cc-pVTZ (D → T) and aug-cc-pVTZ → aug-cc-pVQZ (T → Q) extrapolations and the final MP2/CBS T → Q calculations corrected for the CCSD(T) contribution with a small basis set (CBS(T)).[16]

of contemporary science. QM methods can calculate some properties with high accuracy while providing qualitative insights for others. The most important of these is the intrinsic electronic energy, which is a function of the molecular geometry, defined as the exact Cartesian coordinates of all atoms (i.e., nuclei) of the molecule. By calculating the energy on a grid of varying geometric coordinates, potential energy surfaces can be constructed point by point (Figure 1[12,14b,16]). The calculated intrinsic energy for a given, fixed geometry corresponds to a hypothetical measurement of the energy at zero Kelvin temperature. This differs from the averaged energies obtainable by any real experiment, which is necessarily performed at nonzero temperature on an ensemble of populated structures. Thus, QM calculations have one considerable advantage over experiment: they can investigate the properties of any geometry of choice, including geometries that would not be populated in experiments of model complexes, but which occur when the model system is embedded in real nucleic acid structures. In principle, even the nature of fleeting transition states occurring during conformational changes or chemical reactions can be investigated.

Typically, QM is applied to calculate interaction energies of model molecular complexes representing interactions that occur in intact macromolecular nucleic acids. Examples include base pairs, base stacks, base−backbone, and base−solvent interactions. The interaction energy is defined as the difference between the energy of the interaction complex in a given geometry, specified in Cartesian coordinates, and the energies of the corresponding monomers when they are separated to infinity so that they do not interact. QM energies calculated on completely isolated systems (in vacuo; i.e., in the gas phase), are called "intrinsic energies". For gas phase calculations, modern QM methods can achieve high chemical accuracy; that is, deviations of ∼0.5−1.0 kcal/mol from true values for interactions between two bases (Figure 2).[16,26] This is a qualified estimate with respect to hypothetical (i.e., unknown) values that would be obtained by fully converged calculations. For comparison, the gas phase interaction energies of AU and GC Watson Crick base pairs are ∼ −15 and −30 kcal/mol, respectively, while optimal configurations of base-on-base stacks possess interaction energies of ∼−10 kcal/mol.[16]

The capability of QM calculations to include solvent effects on conformational preferences and interaction energies is limited by lack of accurate methods to model the solution phase.[27] Nonetheless, recently continuum solvent techniques are becoming increasingly popular in the quantum chemistry of nucleic acids. These methods are common in that they treat the solvent as a dielectric continuum, which creates an interaction potential around the solute molecule. There are plenty of variants of this technique, which differ in the formalism used to express this interaction potential. Among them, the COSMO[28] and MST[29] models are rather suitable to characterize the strength of intermolecular interactions in nucleic acids, albeit the results should not be overinterpreted (see also below).[16e,30] In addition, recently, several new continuum solvent techniques have become available, such as the IEF-MST[31] and SMx[32] methods, which, on the basis of results of blind tests on nucleobase derivatives, seem to provide very promising computational platforms for future studies. A similar performance can be expected also from the COSMO-RS method,[33] which combines the original COSMO formalism[28] with a statistical treatment of surface interactions.

Although QM can in principle provide very accurate values of the intrinsic energies of interacting systems, not all results obtained by calculation are biologically relevant. The quality of calculations can suffer because of either an inappropriate choice of QM method or an unsuitable geometry. With presently available QM methods, the latter problem is more significant. Use of X-ray structures (which are inevitably averaged and influenced by data/refinement errors) can lead to major and

uncontrollable defects in accurate interaction energy calculations for a variety of reasons.[16f,22,26]

Although QM calculations of base stacking, base pairing, and related interactions can be carried out routinely, it is more difficult to perform realistic calculations to obtain the energies of models of the flexible sugar−phosphate backbone, which can assume a large range of conformations.[34] An important source of error in calculations of the backbone is the artifact known as "basis set superposition error" (BSSE), which is a spurious unphysical stabilization of molecular contacts in variational QM computations using finite basis sets of atomic orbitals.[14c,26] BSSE can be eliminated in a straightforward fashion from calculations of intermolecular complexes but not from those of intramolecular interactions.

Large problems also arise when trying to carry out computations relevant to macromolecular nucleic acids, due to the uncompensated charges of the phosphates. Thus, it is best to avoid including more than one phosphate group in model systems. Optimizations of flexible backbone tend to produce geometries that do not occur in polymeric nucleic acids and instead form intramolecular contacts that prevent biochemically relevant energy analysis. In recent studies investigating sugar−phosphate−sugar DNA model systems, we found it necessary to freeze all dihedral angles, so as to keep the system under control and match experimental or target values.[35] Our attempts to adequately include stacked bases into the backbone calculations have not succeeded (unpublished data). We find that in the absence of constraints, the backbone tends to deviate from biochemically relevant geometries. QM calculations of the nucleic acid backbone are still rather rare.[36] Further information about various aspects of QM calculations can be found in the literature.[26]

Structure/energy QM calculations can be supplemented by electron density and energy decomposition analyses. There are methods to analyze the electronic density, which can be directly derived from the wave function. These methods divide the electronic density into components, which can be assigned to classical chemical bonds. Among them, perhaps the natural bonding orbital (NBO) analysis[37] or Bader's atoms in molecules (AIM) model[38] are the most accurate and widespread. For example, a combination of these two techniques can be successfully applied to evaluate contribution of individual H-bonds in complex interaction networks. Such calculations rule out the presence of weak C−H···O H-bond in canonical AU and AT base pairs.[39] Common problem of these methods is that it is not easily possible to translate the knowledge of the fine aspects of the electronic structure into direct information about energetics. This limits practical applicability of such analyses to structural biology problems or force field derivation.

Another extension of basic QM description is energy decomposition. Various energy decomposition schemes have been elaborated to disclose the physicochemical nature of the stabilizing forces acting in intermolecular complexes. Some of them (e.g. analysis of the frontier molecular orbitals) have been applied to analyze base pairing in DNA.[40] The SAPT (symmetry adapted perturbation theory) method has been employed to evaluate the balance of stabilizing forces in RNA base pairs, tertiary interactions,[41] and base stacking.[42] Decomposition was utilized in parametrization of the specialized polarizable SIBFA force field,[43] which is useful for model calculations of interactions between metal cations and nucleic acid components.[44] The usual limitation of the decompositions is that they are not fully unambiguous. They also decompose the interaction energy into a set of large, exponentially growing (in absolute values) terms. This makes the decompositions very sensitive to small variations of interatomic distances and impractical for force field derivation and biochemical/bioinformatics analyses. Note that although energy calculations correspond to observable (i.e., "real", physically existing) quantities, electron topology analyses and energy decompositions require arbitrary decisions. The nature of a H-bond as derived by decomposition in the gas phase 0 K minimum geometry is not fully representative for a biomolecular H-bond that extensively fluctuates and competes with other interactions. Among decomposition schemes, SAPT and its faster DFT-SAPT variant are considered as most physically based.[16f,45]

In summary, the goal of QM computations is to provide quantitative understanding of the energetics and physicochemical nature of the interactions that structure RNA molecules. This information, when wisely used, can improve our ability to predict RNA structure from sequence and gain insight into its function. Thus, computations represent a potentially useful complement to RNA structural bioinformatics, where such physicochemical insights are lacking.

What Is the Scope of RNA Structural Bioinformatics? A major goal of RNA bioinformatics is to identify all genes of ncRNA (see above) in genomes. Entire sequenced genomes are accumulating rapidly in sequence databases. Transcriptome projects have demonstrated that much of the genome is transcribed (i.e., copied into RNA) and that most of the RNA produced is ncRNA.[7a,b] Evidence is rapidly accumulating that much of this RNA production plays critical roles in gene regulation, development, adaptation to environmental changes, and evolutionary plasticity.[7c,46] Still, the structures and functions of most ncRNAs remain unknown. Thus, another task is to predict the secondary (2D) and tertiary (3D) structures of ncRNAs identified in the genomic sequences or discovered by transcriptome projects. In addition, one seeks to identify possible protein and RNA interaction partners for ncRNAs. Finally, one would like to predict the possible functions of ncRNAs or to better understand their mechanisms of action, including their dynamics.

Hierarchical Structure of RNA. Like DNA, RNA is an unbranched, linear polymer composed of four nucleotide units: A, C, G, and U. In RNA, the base uracil (U) replaces thymine (T), found in DNA. Each nucleotide consists of a planar aromatic base attached to a five-member sugar moiety (ribose in RNA or 2′-deoxyribose in DNA) and a phosphate group. Nucleic acid chains (RNA or DNA) result from phospho−diester linkages between successive sugar residues, with phosphate groups linking the 3′-carbon of each sugar to the 5′-carbon of the next sugar, leaving a free (unlinked) 5′- position at the beginning of each chain (the "5′- end") and a free 3′- position at the other end (the "3′- end"). Thus, nucleic acid chains are asymmetric; that is, 5′-ACGU-3′ is a different molecule from 3′-ACGU-5′. The 5′-end is considered the beginning because that is where chain synthesis begins in living organisms.

The ribose 2′-OH group (absent in DNA) induces profound differences between DNA and RNA. It makes RNA chemically less stable than DNA (by assisting in autochain cleavage), so DNA is better suited for stably coding large genomes. Because the 2′-OH group is a versatile H-bond donor and acceptor, its presence enhances the ability of RNA to create complex architectures not available to DNA.[19a,47] As detailed below, the 2′−OH makes possible a whole range of non-Watson−Crick base pairs not found in DNA and facilitates compact packing of RNA helices. Evolution has exploited the versatile self-interaction properties of RNA to generate an incredible diversity

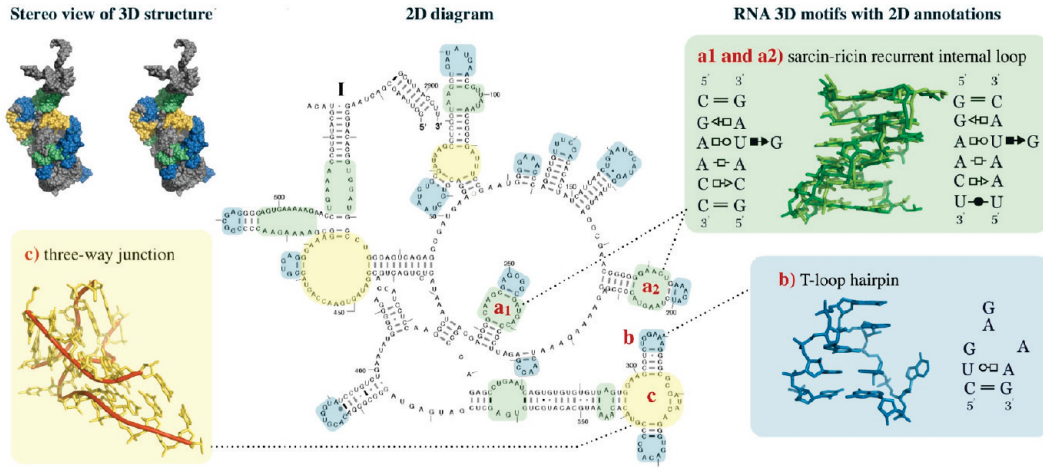


Figure 3. Domain I of 23S rRNA from *Escherichia coli* with internal loops (green), hairpin loops (blue), and multihelix junction loops (yellow) highlighted both on the stereo view of the 3D structure (upper left) and on the 2D diagram (center). The 2D diagram shows canonical base pairs (marked by short lines) and GU wobbles (dots). The other nucleotides are formally unpaired and form 2D loops (see the text for explanation). In reality, the 2D loops are precisely structured RNA elements. Locations of several RNA 3D motifs are marked on the 2D diagram as a1, a2, b, and c. Two instances of a recurrent sarcin−ricin motif (a1 and a2) found in this region are superimposed, and their interactions are annotated according to the Leontis−Westhof[19a] classification in the green inset. The blue inset (b) shows a T-loop hairpin with an annotation, and the yellow inset (c) depicts a multihelix junction loop (the backbone of each of the three strands is traced by a red ribbon).

of RNA structures capable of a large range of specific RNA−RNA, RNA−protein, RNA−DNA and RNA−small molecule or −ion interactions of great biological importance. The chemical difference between U and T is rather subtle.[48] U is essentially a T lacking the methyl group at position C5 (carbon-5) of the base. The methyl group contributes to more efficient base stacking and, thus, subtly improves helix stability.[22,23] For recent gas phase analysis of the effect of the methyl group on pairing and stacking, see also ref 49. The most important role of the 5-methyl group of T is likely related to DNA repair. Cytosines occasionally deaminate to uracils. The presence of a U (T lacking the 5-methyl group) in DNA marks sites at which a C → U conversion has occurred, requiring repair.[21,50]

DNA occurs largely as a dimeric, double-helical complex comprising two long complementary strands, which associate antiparallel to each other by forming exclusively AT and GC Watson−Crick (WC) base pairs, that is, canonical base pairs. The base pairs stack to form the regular B-form right-handed double helix. In contrast, RNA molecules are single-stranded. Nevertheless, they can also form usually short (see below) antiparallel double helices by folding back upon themselves to align WC complementary stretches of sequence. In addition to the canonical AU and GC WC base pairs, A-RNA double helices contain a significant fraction of GU "wobble" base pairs (see Figure S1 in the Supporting Information). Canonical RNA double helices alternate with regions of nucleotides that do not form canonical base pairs, that is, that are nominally unpaired. The *secondary* (*2D*) *structure* of an RNA is a summary of the adjacent canonical base pairs formed when an RNA molecule folds. Drawings representing the 2D structures of RNA molecules often show only the nested canonical base pairs. All the remaining nucleotides are shown as unpaired "loops" (see below) in the 2D plots (Figure 3[19a]). Many of the nominally unpaired nucleotides, however, form noncanonical (non-Watson−Crick, non-WC) base pairs. All these terms will be explained in detail below.

The *tertiary* (*3D*) *structure* refers to the non-WC and long-range interactions that stabilize the exact RNA three-dimensional structure. Predicting RNA 2D and 3D structures starting from sequence is a challenging and multistep process.[51] The WC base pairs determine the basic folding and contribute most of the thermodynamic stability to the folded 3D structure. Thus, structure prediction usually starts with prediction of 2D structure.[52] In other words, the 2D structure is "separable" so that to a good first approximation, most RNAs fold so as to minimize the free energy of the 2D structure. Approximately (only) 60% of bases in structured RNAs form canonical base pairs. However, the tertiary interactions can also contribute decisively to the overall free energy of RNA molecules, especially in those cases that part or all of a molecule can form two or more distinct 3D structures having comparable free energies. In fact, the ability to form more than one structure is essential to the function of some RNAs. Environmental factors, interactions with other molecules, or subtle effects of the kinetics of folding may affect which structure is finally realized under specific conditions. For RNAs with length up to ∼700 nucleotides, contemporary methods for predicting the 2D structure by computational folding of a single sequence achieve ∼70% accuracy.[52] This is calculated as the percentage of correctly predicted WC base pairs minus predicted base pairs that do not occur. This accuracy is considerably improved when additional experimental data are available; for example, chemical or enzymatic probing data of the folded RNA molecule.[53] Probing data can identify nucleotides, which are more likely to belong to 2D structure loops vs WC paired helices. Folding programs allow one to include these data as constraints.[53,54]

Predictions of 2D structure can also be improved by knowledge of additional homologous sequences that are sufficiently, but not excessively, diverged.[55] The success of comparative sequence analysis (CSA) methods is based on the idea that random mutations that occur during evolutionary processes are not equally likely to be passed on to progeny. Natural selection rapidly eliminates mutations that disrupt the 3D structure in ways that block the proper function of RNA molecules. Moreover, natural selection favors compensating mutations that restore function to molecules whose function is compromised by the initial mutation. Thus, the 2D and 3D structures of homologous RNA molecules tend to diverge much more slowly than their sequences. By identifying compensating mutations that preserve WC complementarity at equivalent sequence positions in homologous RNAs, one obtains reliable

evidence for conserved base pairs that belong to the common 2D structure. If nucleotides "*i*" and "*j*" in the RNA sequence form canonical base pair, then conservation of the 2D structure requires that a mutation at position "*i*" be accompanied by a compensatory mutation at position "*j*" to maintain the WC base pair and the functional structure.

An accurate 2D structure provides the necessary basis for prediction of the 3D structure, but it is not sufficient by itself. Despite considerable progress in structure prediction methods, the only reliable way to obtain atomic-resolution 3D structures of new RNA molecules remains X-ray crystallography. Sequence alignment and CSA can also play a role in modeling RNA 3D structure[56] and are especially efficient if one or more exemplar X-ray structures are available. When an X-ray structure of a given RNA of one organism is available, it is possible to deduce molecular interactions of equivalent RNAs of other species by aligning their sequences to the known structure. Sequence alignment means arranging the sequences of two or multiple RNAs to identify regions that mutually correspond because of structural or evolutionary relationships between them.[57] Unless the sequences to be aligned are nearly identical, structural alignment of RNA molecules requires simultaneously determining the 2D structure. When correctly constructed and properly annotated, sequence alignments allow one to infer for each RNA sequence the base pairs and other interactions that form at positions equivalent to the "parent" X-ray structure. Accurate alignments allow one to identify evolutionary conserved motifs, sequence patterns that form characteristic RNA 3D "building blocks". The quality of alignments can thus be improved using sequence signatures known to form specific 3D molecular building blocks and interactions. We suggest that advanced QM and MM computations can substantially enrich the RNA structural bioinformatics by providing additional insights into the physical chemistry of molecular interactions determining the sequence signatures. Guided by phylogenetic analysis and 3D bioinformatics, computations can be used to explore and analyze the effects of base substitutions not yet observed in the available experimental structures.[47a,58]

Watson−Crick Base Pairs. The most frequent base pairs in RNA molecules are those that compose canonical A-form double helices, the Watson−Crick AU and GC (canonical) base pairs. They have the special property of being exactly superposable on each other, so we say that GC and AU canonical base pairs are "isosteric". In fact, GC and AU pairs are self-isosteric, in the sense that AU superposes on UA and GC superposes on CG. Thus, all four WC pairs, GC, CG, AU, and UA are mutually isosteric. The structural consequence of this isostericity is that the canonical A-RNA double helix has a regular, periodic, and largely sequence-independent 3D shape. The biological consequence is that mutations that substitute, for example, a UA base pair with a GC, CG, or AU do not change the 3D structure of the helix to which the mutated bases belong. If nucleotides "*i*" and "*j*" in the RNA chain form a conserved X_iY_j canonical base pair in the X-ray structure, sequence alignments generally reveal *covariation* (alternation) of CG, GC, AU, and UA at corresponding positions of homologous RNA molecules, even those from distantly related organisms.

The free energy of an RNA helix is an important biochemical parameter. It depends on the length and sequence of the helix and can be quantified by the free energy released upon RNA chain folding. Because of the asymmetry of RNA chains, a CG pair stacked on a GC pair (i.e., 5′-GC-3′ paired with 3′-CG-5′) is different from a GC pair stacked on a CG pair (i.e., 5′-CG-3′ paired with 3′-GC-5′). As in DNA, there are 10 unique dinucleotide sequences (base pair steps) formed by canonical base pairs in RNA (and 21 including GU "wobble" pairs). Measured thermodynamic (TD) parameters for these 10 canonical steps are called nearest-neighbor parameters and constitute the core for predicting secondary structure from base sequence.[23a] The relative contributions of base pairing and base stacking to the thermodynamics of RNA are not known, but it is assumed that the two interactions are roughly of equal importance.[59] Although TD parameters for canonical base pair steps are well established,[52] extension of the TD predictions to non-WC duplexes and motifs, which would dramatically improve 2D predictions, is limited by lack of experimental data.[52] Carefully designed computations of molecular interactions could contribute to finding or at least rationalizing the TD rules for these RNA elements, which are difficult to access by experiment. This is one of the main areas where computational chemists should direct their efforts.[58c,60,61]

"Wobble" Base Pairs. The next most frequent base pair is the GU "wobble" base pair.[47a] For optimal H-bonding between the WC edges of G and U, a lateral shift of the U toward the major (deep) groove is required (see Figure S1 in the Supporting Information). This perturbation is relatively minor and does not greatly distort the A-form double helix, that is, the GU wobble pair is nearly isosteric with the AU and GC WC base pairs. Thus, GU wobble pairs occur frequently within or at the ends of WC helices and are thermodynamically quite stable, on a par with WC AU pairs. The lateral shift nevertheless creates a pocket in the minor groove that can be occupied by a water molecule, the O2′ hydroxyl of another nucleotide, or a phosphate group and is often used for RNA tertiary interactions. Importantly, the GU wobble is not self-isosteric; that is, GU is not superposable with UG. Consequently, GU is rarely observed to covary with UG in RNA 3D structures or correctly constructed sequence alignments.[47a]

Ribosome Decoding: Shape vs Energy. One significant problem that evolution has had to solve is how to discriminate between GU wobble and canonical base pairs formed between the first and second positions of the codons of mRNAs and the anticodons of tRNAs.[3e] Because of its genuine thermodynamic stability, the GU wobble pairs can participate in stable codon−anticodon interactions between mRNA and tRNA. It is not necessary to prevent a GU pair from forming at the third codon−anticodon position because for most amino acids, the genetic code is degenerate at this position, in the sense that more than one codon base (up to four) will be decoded as the same amino acid. However, acceptance of GU wobble in the first two positions would mean acceptance of "near-cognate" tRNAs and subsequent insertion of incorrect amino acids into the growing protein chain. Thus, the ribosome decoding center located on the small ribosomal subunit utilizes a sophisticated network of dynamical molecular interactions to discriminate between the shape of GU wobble pairs (formed by near-cognate tRNA binding) and the near-isosteric shapes of the canonical base pairs (formed by cognate tRNA).[3e] An important lesson for anyone who makes calculations is that the most critical stage of ribosomal decoding relies not on differences in the energies of wobble versus canonical base pairing, but on precise monitoring of the exact shapes of the base pairs formed between mRNA and tRNA.

The basic principle of decoding cannot be deduced from any studies of stability of codon−anticodon base pairing. Such supremacy of shape over energy is common in biology. Thus, during DNA replication, DNA polymerase also monitors the

shape of bases and base pairs to ensure that the correct DNA base is inserted to form a canonical base pair with the base in the template strand. This fact has been demonstrated by efficient replication of isosteric nonpolar nucleobase analogs that cannot form H-bonded base pairs but which mimic the shape of the natural base pairs.[62] However, there is evidence that some other classes of DNA polymerases involved in DNA repair directly recognize DNA base pairing and its stability.[63] This is therefore another lesson illustrating the enormous complexity and variability of biomolecular recognition processes. We cannot expect to find a simple set of universally valid rules. For every rule one tries to formulate, evolution finds other ways to achieve optimal function. Thus, the energy of molecular interactions, while often "silent" in biochemical processes, remains an integral part of the overall picture. In specific cases, the intrinsic energetics can play decisive roles. Exactly how evolution uses energy to achieve biomolecular recognition and functional dynamics must be determined on a case-by-case basis. Clearly, the role of energetics in biomolecular interactions cannot be ignored without serious misunderstandings. This underlines the potential usefulness of appropriate computational efforts in getting in-depth insights into the balance of forces in the individual systems and recognition patterns.

RNA 3D Motifs. Canonical RNA helices tend to be short, generally less than about 12 consecutive WC base pairs. Longer stretches of canonical RNA base pairs are probably too monotonous and too stable to be useful for evolution of complex and often dynamical RNA molecules and RNA-based biomolecular machines. Computations on canonical base pairs and A-RNA helices give only limited information about functional RNAs. As noted above, RNA secondary structure consists of short canonical helices punctuated by nominally unpaired segments forming what appear as "loops" in planar 2D representations. At the level of the secondary structure, loops consist of one or more strand segments and can accordingly be classified in three basic types (Figure 3): (1) *hairpin loops* consist of a single continuous strand segment folded on itself and terminate a helix; (2) *internal loops* comprise two strand segments and occur between two helices; and (3) *multihelix junction loops* consist of three or more strand segments and occur where three or more helices meet.

Figure 3 shows a part of the 2D structure of 23S (large subunit)[3] rRNA and is annotated to illustrate examples of each kind of loop. The term "loop" causes confusion for those unfamiliar with RNA 3D structure as it evokes the idea of unstructured, floppy chain segments. However, most "loops" in RNA molecules that function by virtue of their 3D structure are, in fact, precisely structured, including the most common apical hairpin loops.[64] The nucleotides of structured loops form multiple interactions with each other and frequently with other parts of the same RNA or with other molecules. Such structures are called "RNA 3D motifs".[65] Thus, RNA "loops" are generally the most interesting and functionally important parts of RNA molecules and are frequently recurrent, highly specific molecular building blocks.

What Kinds of Interactions Structure RNA 3D Motifs? RNA motifs largely lack WC base pairs. They are usually rich in non-WC base pairs, as well as base stacking and a variety of base−backbone interactions. Some internal loops are fully paired duplexes, but because they consist of non-WC base pairs, their backbone structures deviate substantially from A-form helices. Many hairpin (or terminal) loops are highly structured and have few nucleotides that are not paired or stacked. Examples include the two most common hairpin loops, the "UNCG" and "GNRA" tetraloops. They usually have four nucleotides and conform to the indicated consensus sequences, where "N" indicates any nucleotide and "R" purine. These hairpin loops are 3D motifs in the sense that they have strictly defined molecular shapes stabilized by characteristic invariant signature molecular interactions. In each case, the first and fourth nucleotides of the loop form non-WC base pairs. Junction loops can have enormous structural complexity.[66] The characteristic molecular interactions in all three classes of RNA loops represent genuine targets for systematic computational studies to clarify the role of molecular energetics in relating the observed sequence preferences to the conserved 3D structures.

Properties of RNA 3D Motifs. RNA 3D motifs are ordered arrays of non-WC base pairs under sequence constraints. This means that it is usually not possible to change just one base without having to change others to keep the functional RNA motif (i.e. 3D structure). The general properties of RNA 3D motifs include the following: (1) They are modular, in the sense that they can occur as discrete units. This makes it hard to experimentally dissect the effect of individual interactions because disrupting one non-WC base pair can cause the entire motif to collapse. (2) They are autonomous; that is, they can occur in different molecular contexts, folding into their characteristic geometry dictated by their specific sequence independently of the context. For example, sarcin/ricin motifs ("SR loops", Figure 3) occur in internal loops or in multihelix junctions in many different molecules. (3) They are recurrent, in the sense that they occur in different molecules or different places in the same molecule. The same motif can evolve convergently in different molecules; that is, evolution is finding multiple times independently the same 3D arrangement. (4) They are multipurpose. For example, the same motif can participate in proteins binding in one context and various kinds of RNA−RNA interactions in others. Some RNA motifs, such as UNCG hairpin tetraloops, appear to largely function by nucleating RNA folding because of their local stabilities. Other RNA motifs, such as GNRA hairpin loops, appear to function primarily by mediating RNA tertiary interactions. Thus, almost every GNRA loop in the ribosome forms a tertiary interaction, whereas almost no UNCG loop does so. V-shaped Kink-turn internal loops play primary roles in protein assisted RNA folding[67] and also can act as anisotropic flexible elbow.[68]

RNA 3D motifs can be in principle predicted from sequence. By detecting their characteristic signature sequences in ncRNA sequences, their occurrence in the folded RNA may be inferred. In favorable cases, we can infer their likely role in the functional structure. However, the occurrence of a sequence potentially forming an established RNA building block does not always guarantee its actual formation because the surrounding context can also play a role. Some well-studied 3D motifs exhibit most or all of the above listed properties. They include internal loops such as loop E from 5S rRNA and the Sarcin-Ricin loop;[69] various kink-turns;[67] C-loops; the prominent hairpin loops UNCG and GNRA; and the anticodon loop and T-loop, both from tRNA. Our unpublished data suggest that the current 3D database has ∼100 distinct internal loops, some of which occur so far in only one instance. Thus, we have currently ∼100 distinct modular RNA building blocks that are used to construct RNAs. Thermodynamic parameters have been determined for only a small number of 3D RNA motifs for use in energy-based RNA 2D prediction programs.[52] Sequence- and knowledge-based approaches for predicting the 3D structures of small RNA 3D motifs are promising.[51b,70]

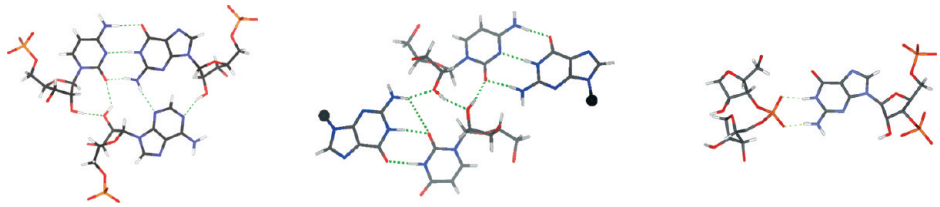
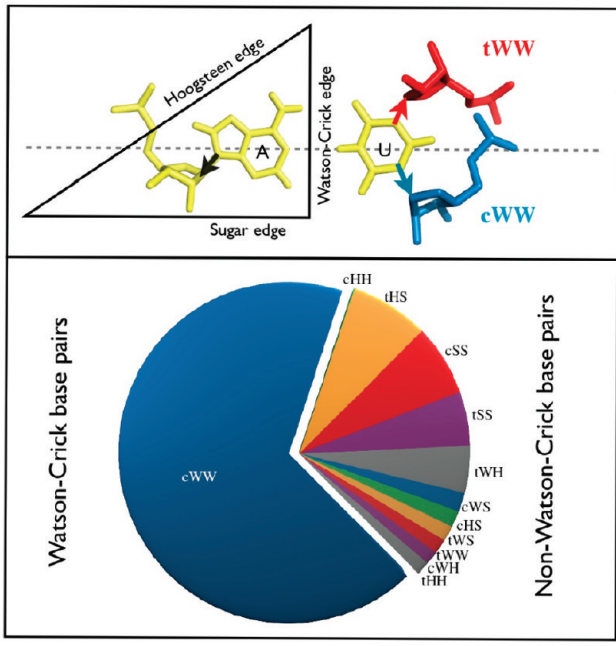


Figure 4. Examples of RNA base pairing involving backbone atoms. Left: A-minor type I GCA triple interaction is the most frequent tertiary interaction in structured RNAs. Middle: Packing interaction GCUG quartet is another powerful and recurrent tertiary interaction. Right: Example of base−phospate "base pair" interaction. Note the dominant role of the backbone functional groups in the interactions. All interactions are highly sequence-specific.



No.	Glycosidic Bond Orientation	Interacting Edges NT1	Interacting Edges NT2	Abbreviation	Symbol	Triangle Abstraction
1	*Cis*	Watson-Crick	Watson-Crick	cWW		
2	*Trans*	Watson-Crick	Watson-Crick	tWW		
3	*Cis*	Watson-Crick	Hoogsteen	cWH		
		Hoogsteen	Watson-Crick	cHW		
4	*Trans*	Watson-Crick	Hoogsteen	tWH		
		Hoogsteen	Watson-Crick	tHW		
5	*Cis*	Watson-Crick	Sugar Edge	cWS		
		Sugar Edge	Watson-Crick	cSW		
6	*Trans*	Watson-Crick	Sugar Edge	tWS		
		Sugar Edge	Watson-Crick	tSW		
7	*Cis*	Hoogsteen	Hoogsteen	cHH		
8	*Trans*	Hoogsteen	Hoogsteen	tHH		
9	*Cis*	Hoogsteen	Sugar Edge	cHS		
		Sugar Edge	Hoogsteen	cSH		
10	*Trans*	Hoogsteen	Sugar Edge	tHS		
		Sugar Edge	Hoogsteen	tSH		
11	*Cis*	Sugar Edge (priority)	Sugar Edge	cSs		
		Sugar Edge	Sugar Edge (priority)	csS		
12	*Trans*	Sugar Edge (priority)	Sugar Edge	tSs		
		Sugar Edge	Sugar Edge (priority)	tsS		

Figure 5. Base pair classification and occurrence. Upper left: AU cis and trans Watson−Crick Watson−Crick base pairs superimposed. The arrows indicate the direction of the glycosidic bond. The adenine belongs to both base pairs. The triangle abstraction of a nucleobase is overlaid onto the adenine to demonstrate the three nucleotide edges available for base pairing. Right: Twelve geometric families. Each base pair family is defined by the interacting edges of the bases and the relative orientation of the glycosidic bonds (columns 2−4). Abbreviations and symbols for representing base pair families in text and secondary structures are shown in columns 5 and 6. Column 7 shows an abstract representation of each family using triangles to represent the bases, where the hypotenuse represents the Hoogsteen edge. The shaded cells denote base pairs in the cis orientation. Lower left: Base pair frequencies (how frequent are the 12 base pair families in percent) based on their occurrence in rRNA structures (adapted from ref 19b).

Not all recurrent RNA building blocks are autonomous. For example, in isolation, the 5′-UAA/5′-GAN internal loop forms a fully base-paired noncanonical double helix, basically consistent with standard thermodynamics predictions. This topology is not used by evolution. Nevertheless, in specific tertiary contexts, this loop is completely remodeled into an RNA module serving in tertiary interactions.[58b,71]

The amazingly variable 3D RNA motifs represent some of the most attractive targets for all kinds of advanced QM and MM computations. They contain many interesting molecular interactions; they are small enough to be tractable and sufficiently structured to define the computational task. RNA motifs are enormously important, and there is a desperate need for quantitative insights in light of the lack of data for use in thermodynamics-based prediction algorithms.

Non-Watson−Crick Base Pairs and RNA 3D Motifs. RNA nucleobases form a bewildering variety of base pairs. Only in the early 2000s, after a critical mass of RNA 3D structures become available, did the general principles for cataloguing RNA base pairs emerge. The decisive step forward was to extend the RNA base pair definition to include base−sugar and sugar−sugar hydrogen bonding (Figure 4).[19] In fact, some RNA base pairs contain no direct base−base H-bonds and still are biochemically highly relevant. The generalized principle of RNA base pairing (the "Leontis−Westhof" classification) states that each RNA nucleobase can pair with another base using one of three base edges: the Watson−Crick edge (W), the Hoogsteen edge (H), or the sugar edge (S). The 2′-OH of the ribose is considered part of the sugar edge, and generally contributes to base-pairing interactions involving this edge, a feature that makes RNA distinct from DNA. Thus, base-pairing can occur by bases interacting edge-to-edge in six ways: W edge to W edge (WW), W to H (WH), W to S (WS), H to H (HH), H to S (HS), or S to S (SS).[19a,64a] Further, the edges can come together in cis or trans, depending on whether the glycosidic bonds attaching the sugars are on the same or the opposite side of the axis joining the base centers. This leads to 12 basic geometric families of RNA base pairs (Figure 5[19b]). The families are marked by using "c" or "t" to refer to cis or trans, and the capital letters W, C, and H to refer to the edges. Thus "tWH" stands for the trans Watson−Crick/Hoogsteen family.

Note that within the individual families, only certain base combinations can form. For example, there is no cWW GG base pair. In addition, it is not sufficiently specific to refer to a

base pair only by its base combination. For example "AG base pair" could refer to cWW, tHS, tWS, cWH, tWH, cWS, cHH, tHH, cHS, cSS, and tSS AG arrangements. This illustrates the difficulty of predicting non-WC base pairs from sequence. The individual families contain up to 12 or 16 distinct base combinations, depending on whether they are self-symmetric. Some families typically occur as part of larger contexts, forming base triples or quadruples (Figure 4). In addition to the standard classification, there are additional planar interactions involving bifurcated hydrogen bonds, inserted solvent molecules, or single H-bonds that are not included in the classification.[19a] The canonical "WC" base pairs belong to the cWW family and, in addition, require GC and AU nucleotide combinations. In addition, GU wobble belongs to the cWW family. The remaining cWW base pairs are already referred to as non-WC base pairs; that is, noncanonical and non-WC are synonyms.

In structured RNAs, 30% or more of base pairs are non-WC base pairs, that is, all pairs other than cWW AU or GC pairs.[19b] Furthermore, in contrast to canonical base pairs, some non-WC base pairs possess alternative *substates*. Let us consider the A-minor I triplet in Figure 4, left, which consists of cWW GC, tSS AG, and cSS CA base pairs. The A-minor I tSS AG base pair is evidently not optimally intrinsically paired because the adenine nucleoside is also involved in the cSS base pair. Still, this specific observed tSS AG geometry is dominant in experimental structures and has been identified by structural bioinformatics. It is due to the enormous frequency of A-minor I interactions. Existence of substates means that a given base pair may adopt several competing microarrangements. Substates are even more important for larger contexts such as triples and quadruples.

Consider again the A-minor I triple. Figure 4 shows its fully paired (direct) variant. However, in some observed instances, its cSS CA interaction is water-mediated, with water molecule inserted between the two 2′O groups (Supporting Information Figure S2).[68a] In MD simulations, the triple often fluctuates between direct and water-mediated geometries, which creates energetically flat (anharmonic) triplet system.[68a] Such flexibility of RNA base pairing is functionally important because, for example, dynamical water insertion in the A-minor I interaction contributes to elbow-like flexibility of kink-turn motifs.[68a] Starting from the A-minor I interaction, the adenine nucleotide can slide along the CG base pair in both directions and create a number of additional alternative substates known as A-minor II, A-minor III, and A-minor 0 interactions (Supporting Information Figure S2).

Similar flexibility is known also for the phosphate-in-pocket interactions, where phosphate groups are inserted into minor groove of a double helix, as shown by the quadruple in Figure 4, middle.[47a] In such cases, different parts of the RNA molecule can slide relative to each other over several angstroms. Sometimes the ribosomal structures even show nucleosides that are properly arranged to make some interaction but are too far from each other. This is known as potential interactions.[47a,f] Potential interactions can be converted into real interactions upon conformational changes. This indicates that substates provided by RNA base pairing are likely of outmost importance for functional dynamics of large RNAs and ribonucleoprotein systems.

Some base pairs occur frequently in RNAs, and others are infrequent (Figure 5). The frequencies of occurrence (how often has evolution been using a given pair) of each base combination for each base pair family has been determined using a nonredundant set of atomic-resolution X-ray structures from PDB and also by using rRNA sequence alignments.[19b] The frequencies of occurrence of each base combination within each geometric base pair family result from their relative stabilities and shapes. In other words, each base pair possesses at least three characteristic features: (i) an intrinsic capability for base pairing, (ii) a shape that determines its structural compatibility with the overall RNA structure, and (iii) specific capabilities to contribute to functionally interesting RNA architectures. Each of these three factors contributes to the frequency with which the given base pair is selected by evolution. Generally, the geometric family is very conserved by evolution if the motif is conserved. Changing the geometric family of a single base pair can completely change the 3D motif structure. The factors that determine which geometric base pair family forms in a given context, or which base combinations occur most frequently for a given geometric family, are not fully known and constitute another area where computations are needed.

Robustness of Base Pairing Families. The base pairing classification was proposed almost 10 years ago, before sufficient numbers of atomic resolution structures had been determined to provide examples of all base pair combinations in each family. The classification suggested additional (at that time unobserved) base combinations for certain families.[19a] Recently, the compilation was updated using RNA 3D structures available in 2009.[19b] This analysis provided experimental confirmation for almost all base pairs predicted in 2002. Even more significantly, no additional base pairs (absent in the 2002 compilation) have been found.[19b]

Isostericity Principle. Above, we illustrated the importance of molecular shape in structural biology. For RNA base pairing, this can be formulated as the RNA base pair isostericity principle. During evolution, natural selection typically eliminates those base mutations that disrupt the 3D structure of the RNA molecule, preventing it from achieving its function. For bases involved in edge-to-edge pairing, substitutions resulting in base combinations that cannot form the right base pair family or that produce a nonisosteric base pair are likely to disrupt structure and function. Thus, only isosteric or near-isosteric base substitutions are found at corresponding positions of homologous RNA molecules or recurrent RNA 3D motifs.[19b]

Why is the shape of the base pairs so important? It is because it determines the position and direction of the attached RNA backbone and, thus, the RNA topology. The 12 isosteric families are mutually nonisosteric, and even within a given family, not all base pairs are mutually isosteric. Thus, a given family can be split into several isosteric or near isosteric subfamilies. For example, in the cWW family, GC and AU are isosteric with each other, whereas GC and GU are near isosteric. However GA and GC or even GU and UG are nonisosteric (Figure 6[72]). Analysis of the available RNA 3D structures and sequences shows that the RNA isostericity principle is one of the most powerful constraints of RNA sequence evolution. Normally, evolution very strictly conserves base pair shapes, and only such substitutions are realized, which can be isosteric or near isosteric. This demonstrates the dominance of the 3D structure over the primary sequence and dominance of the shape over the intrinsic energetics of molecular interactions.

Of course, energetics also plays a role when there is a large difference in stability. Thus, although "wobble" cWW GU and AC base pairs are isosteric to each other, AC is considerably less stable, owing in part to the need to protonate the A(N1) position to allow H-bonding with C(O2). (The AC base pair is protonated.) Thus, AC is much less frequent than GU. Further, GU and AC cWW base pairs have entirely different electrostatic

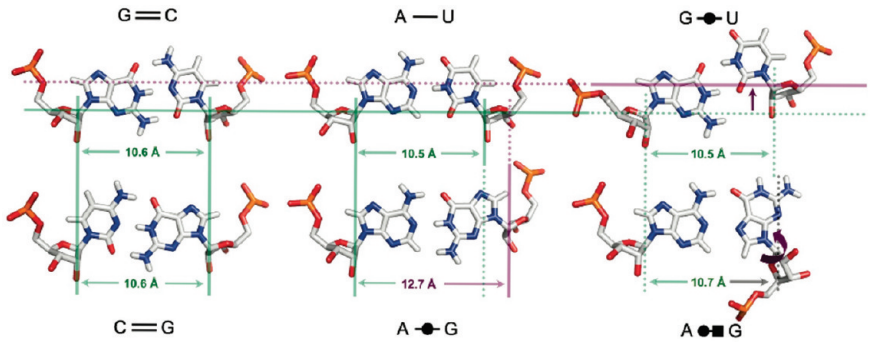


Figure 6. Isosteric relationships between base pairs. Two base pairs are isosteric when they meet three criteria: (1) The C1′−C1′ distances are the same, (2) the paired bases are related by same rotations in 3D space, and (3) H-bonds are formed between equivalent base positions. The cWW GC, CG, and AU base pairs (upper and lower left and upper center) meet all three criteria and are isosteric to each other, as shown. The cWW AG pair (lower center) and GU pair (upper right) belong to the same geometric family, so the paired bases are related by the same 3D rotation. However, the cWW AG pair has a significantly longer C1′−C1′ distance and thus is not isosteric to the other pairs, even though it meets the other two criteria. The C1′−C1′ distance in the cWW GU (wobble) pair is about the same as in canonical pairs, but the U is shifted toward the major groove, so H-bonding does not occur between the same positions as in the other cWW pairs. This change is more subtle, so GU is considered near isosteric to the canonical cWW pairs AU, UA, GC, and CG, consistent with its ability to substitute in Watson−Crick helices for these pairs. The last example, cWH AG (lower right), has about the same C1′−C1′ distance as the canonical cWW pairs, but belongs to a different geometric family. The bases are related by a very different 3D rotation so the base pair it is nonisosteric to all cWW base pairs (from ref 72).

potentials, which may affect stability of their involvement in triples and quadruples. Nevertheless, the AC quite often to a certain extent covary with GU in sequences. However, there are tertiary interactions that specifically require the wobble shape of cWW GU while covariation with AC is forbidden. A textbook example is the GU/CG tertiary quartet known as packing interaction (Figure 4), which is destabilized by the electrostatic potential of AC cWW.[47a]

The isostericity concept extends traditional views of sequence conservation. Many RNA 3D motifs that are not conserved at the level of sequence are entirely conserved when considering base pair isostericity. It should be added, however, that occasionally, we observe more complex scenarios; for example, replacement of one RNA 3D motif by another that uses entirely different base interactions to achieve the same function (motif swap).[47a] Still, important physicochemical properties such as overall topology or flexibility can be conserved. These cases are not predictable, even with the aid of structural bioinformatics, and show the almost endless complexity of RNA molecules.[47a,68b,73]

The fundamental importance of the base pairing became highlighted by recent comparative analysis of the available atomic-resolution 3D structures of the rRNAs of *Escherichia coli* (*E.c.*) and *Thermus thermophilus* (*T.t.*), two distantly related bacteria.[19b] (Figure 7[19b]). All base pairs were identified using the FR3D program,[64a] and the corresponding structures were aligned base pair by base pair. The analysis shows that over 90% of the base pairs (non-WC as well as canonical base pairs) belong to the conserved (core) structures common to these two highly diverged bacterial species, and are thus likely also common to most other bacterial rRNA molecules. Moreover, the aligned base pairs were almost 100% conserved as to geometric base pair family (type) between the two structures. For example, if two bases of the conserved core of the *E.c.* structure formed a tWH base pair, then the corresponding bases in the *T.t.* structure also form a tWH base pair. Amazingly, in 98% of cases, the corresponding base pairs in the two structures were found to be isosteric or near isosteric. This reveals an enormous degree of conservation, which is not visible when considering only the sequence information.

DNA vs RNA Difference from Perspective of Computational Chemists. Figure 6 illustrates the well-known fact that the hydrogen-bond donor and acceptor properties on the exposed

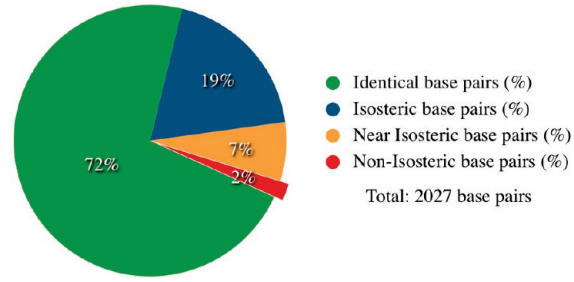


Figure 7. Comparison of corresponding base pairs in the 3D structural alignment of *E.c.* and *T.t.* 5S, 16S, and 23S rRNAs based on the IsoDiscrepancy Index (IDI), a qualitative measure of isostericity.[19b] There are total 2027 base pairs that belong to the conserved core of the bacterial ribosome and, therefore, are seen in both ribosomes at equivalent positions. The majority of the base pairs in the corresponding positions of the 3D alignment are identical (shown in green, IDI close to 0). The next largest group, isosteric base pairs, have IDI ≤ 2.0 (shown in blue). Near-isosteric base pairs (yellow) are characterized by 2.0 < IDI ≤ 3.3. Only 2% of all base pairs are nonisosteric (IDI > 3.3). The Figure illustrates how strictly evolution preserves the isosteric base pairs, even in such distantly related bacteria as *E.c.* and *T.t.*

faces of the canonical base pairs (GC, CG, UA, and AU) are sequence-specific. This in turn influences the interaction with other molecules. Therefore, isosteric pairs can have different hydrogen-bond patterns with the interacting molecules. In B-DNA molecules, many proteins directly read the functional groups in both major and minor groove and, thus, distinguish the B-DNA sequence. Further, the B-DNA double helix possesses fine sequence-dependent conformational variability (irregularity) that is of outmost importance for the majority of DNA-related molecular recognition processes. Numerous computational studies have been devoted to sequence dependence of B-DNA and molecular recognition of base functional groups. The RNA molecular recognition processes are strikingly different. The core of RNA interactions involves all kinds of the noncanonical structural features, starting from single bulges and non-WC base pairs up to complex motifs and architectures (Figure 3). They provide incomparably more variability than canonical duplexes. This is also the reason why the word "mismatch" base pair (noncanonical base pair in B-DNA, usually perturbing the molecule) is rather inappropriate when discussing RNA structures. These are functional base pairs, not mismatches. Thus, evolution does not need to extensively

experiment with fine local conformational variations of canonical A-RNA helix and utilize specifically the base functional groups in the canonical helix grooves. Purely canonical A-RNA helices are anyway only short, 2−10 base pairs in the ribosome (cf. Figure 3) and even shorter in mRNAs. This is in contrast with the extremely long B-DNA canonical duplexes. (For the sake of completeness, longer 20+ base pair canonical A-RNA duplexes are important in RNA interference.)

The functional groups in the deep A-form major groove are quite inaccessible. We in no case suggest that the base functional groups are entirely unimportant, but definitely the common picture is that RNAs, in a given position, typically freely alternate all four possible canonical base pairs (CG, GC, AU, and UA). GC base pairs are more common in RNA, probably because they are more stable, whereas thermophilic organisms show an increased content of GC base pairs. The 2′OH groups that often interact with base exocyclic groups can act as both donor and acceptor. Local variation of a canonical helix that would be considered excitingly large by a DNA researcher would not be noticed by an RNA researcher. Therefore, the two most favorable problems of DNA computational studies are less relevant for RNA. We do not suggest that A-RNA does not possess sequence-dependent local variations, and recent simulations have revealed a large effect of sequence on inclination, roll, and major groove width of A-RNA.[74] However, RNA is usually not about canonical helices. Contemporary nucleic acids computational literature remains visibly concentrated on DNA, strikingly contrasting recent trends in biology and biochemistry. On the other side, RNA molecules offer a much wider (basically unlimited) range of problems than can be amenable to computations. And there is yet another advantage: although the RNA molecules are definitely more complex and at first sight more difficult to describe than DNA, we usually do not need to study fine details, such as a few degrees variation of the helical twist. This increases chances that the studied effects can be properly reflected despite limitations of the computational methods.

QM Calculations on RNA Base Pairs. We have carried out basic QM computations on all six "sugar edge" base pair families.[30b,75] The calculations demonstrate that classifications excluding the sugar moieties would not respect the basic physical chemistry of the interactions. Upon inclusion of the riboses, the QM calculations are quite consistent with the classification; that is, the RNA base pairing principles emerge directly from the intrinsic interactions. For many of the sugar-edge base pairs, the calculated gas phase geometries closely resemble exemplar (centroid) base pairs extracted from the RNA structure database.[19b] The exemplar base pair structure is calculated as the most representative instance in the database of experimental structures. The sugar edge base pairs generally profit from relatively large dispersion attraction. This indicates that they are more hydrophobic compared with canonical base pairs, which may support formation of tertiary interactions. To carry out structural optimizations of some isolated base pairs, it was necessary to impose specific geometrical constraints to keep them close to the experimental geometries. In some cases, the computations predict additional H-bonds not seen in the X-ray structures. Some of these are artifacts due to the absence of the RNA context, in which such H-bonds violate optimal RNA topologies. However, the calculations may also detect the capability of the base pairs to occasionally form interactions that are missed by bioinformatics. We are currently scrutinizing this issue in more detail for those cases where the QM predicted minimal energy structures differ visibly from the exemplar structures. In most cases, we indeed found one or more instances

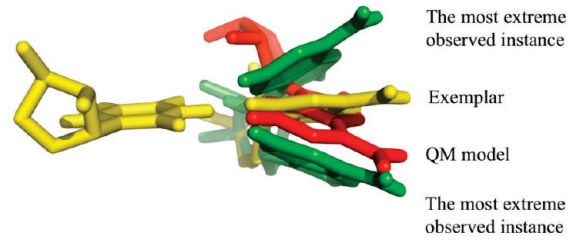


Figure 8. The tWS UA base pair is shown as represented by an exemplar structure (centroid, in yellow), two most diverse observed instances of tWS UA base pair (in green), and a calculated QM model[75b] (red). Although not as planar as the exemplar, the QM model is within the range of observed variation.

in the 3D database very similar to the calculated structure, even when it differed from the exemplar (centroid). For most base pairs involving one or more sugar edges, the exemplar tends to be rather planar, but the individual instances in the database exhibit a large range of interbase angles. The calculated optima for these base pairs tend to be nonplanar, but within the observed range (Figure 8). Work is in progress to extend the RNA base pair computations by considering base pairs in specific structural contexts (with additional aid of computer simulations). Such targeted studies can bring the QM and bioinformatics data into a really intimate relation.[76]

For some base pairs, QM predicts the capability to form substates with amino-acceptor interactions.[15a,77] It is presently difficult to assess their relevance, since the resolution of X-ray structures is low, and the possibility of such interactions is considered neither in X-ray crystallography nor in 3D bioinformatics. In the first crystallographic study reporting interaction with nonplanar amino group (1.9 Å resolution B-DNA−DAPI complex), the interaction was identified after unsuccessful refinement attempts to eliminate a presumably repulsive amidinium−amino interaction.[78] The interaction was then explained using QM computations. Had the crystallographer been less patient, the interaction would have been completely misunderstood. Thus, if experimental structures are analyzed with presumption, many rare but functionally important cases deviating from usual expectations are missed.

Extension of Base Pairing Classification to Base - Phosphate Interactions: Combining Structural Bioinformatics Directly with QM Computations. For some geometric base pair families, certain base combinations occur much more frequently than others.[19b] In some cases, it is evident that additional interactions not included in the standard classification play a role. Thus, we have complemented the base pair classification by introducing a classification of base-phosphate (BPh) interactions,[79] which has substantially expanded the number and kind of interactions that can be annotated in RNA 3D structures. This has been the first joint study simultaneously applying the tools of bioinformatics and QM calculation to a major class of interactions in RNA. Approximately 12% of the nucleotides in rRNA form direct internucleotide H-bonds between nucleobase donor atoms and phosphate oxygen acceptor atoms. The bioinformatics analysis provided the initial classification of the BPh interactions, and this was subsequently refined by QM calculations, which uncovered the physicochemical differences between the various binding types. The BPh interactions obey the isostericity principle and impose significant evolutionary constraints on the RNA sequences. Further, we found correlation between the calculated intrinsic stabilities of the BPh interactions and their occurrence frequencies (Figure 9[79]). Those BPh interactions that are calculated as intrinsically

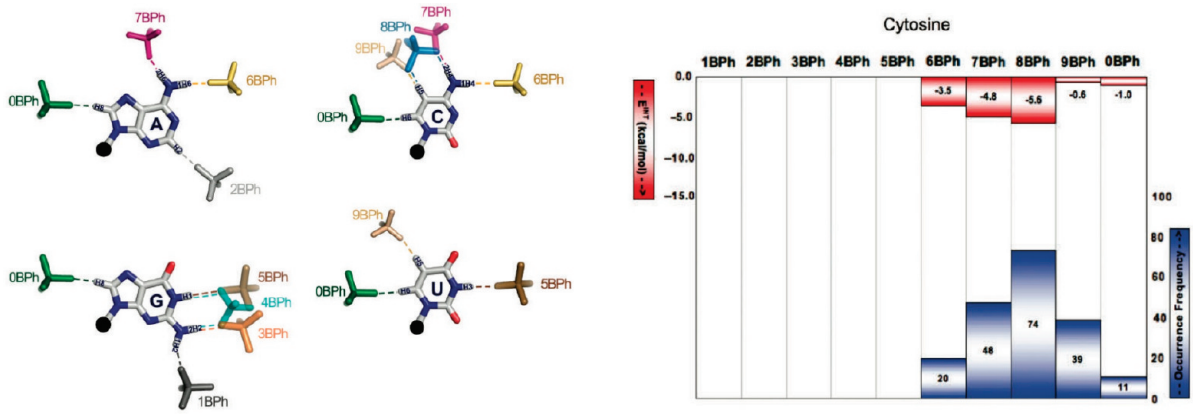


Figure 9. Left: Proposed nomenclature for BPh interactions and superpositions of idealized BPh interactions observed in RNA 3D crystal structures for each base. H-bonds are indicated with dashed lines. Compare to Figure 4 for a representative example. BPh categories are numbered 0−9, starting at the H6 (pyrimidine) or H8 (purine) base positions. BPh interactions that involve equivalent functional groups on different bases are grouped together: 0BPh (A, C, G, U), 5BPh (G, U), 6BPh (A, C), 7BPh (A, C), and 9BPh (C, U). Right: Comparison of calculated BPh interaction energies (red) and BPh occurrence frequencies (blue) from a reduced-redundancy set of crystal structures for cytosine. Adapted from ref 79.

more stable tend to occur more frequently in biological RNA structures. This is a clear indication that natural selection at the level of RNA sequence is, to a certain extent, sensitive also to molecular interaction energies.

We have also carried out preliminary analysis correlating QM-calculated interaction energies of RNA base pairs and their occurrence frequencies (unpublished data). In contrast to BPh interactions, we do not see clear correlations. This, however, is not so surprising because the analysis averages over all base pairs in all their contexts, which basically means comparing apples and oranges. The role of energy may only become apparent upon considering specific interaction contexts. We now analyze RNA interactions in their different specific contexts, which is the ultimate way how computations can aid the bioinformatics. Structural bioinformatics relies heavily on known 3D structures to construct databases of possible RNA motifs and interactions. However, 3D structures do not provide direct information about energies. Thus, structural biology and bioinformatics are biased toward purely structural data. By applying modern computational approaches, the energy dimension can be added to the 3D structures.[79]

Base Stacking in RNA. RNA nucleobases are planar, one-atom-thick entities and can interact by stacking on each other like two plates. Base stacking interactions are just as important as base pairs for stabilizing RNA 3D structures. Base stacking can occur between individual bases, as occurs in tertiary interactions involving looped out bases, but more often occurs between two base pairs, as occurs in helices and many local motifs comprising non-WC base pairs. Common also are stacking arrangements involving base triples and quartets. Coaxial stacking of helices is one of the fundamental driving forces of RNA folding. Each nucleobase has two distinct planar faces considering its 5′−3′ position in RNA. Thus, two distinct bases can stack on each other in four unique ways with regard to which base faces are in contact. However, stacking is actually a continuum of possible geometries because the bases can slide in two dimensions and rotate relative to one another while remaining stacked. Thus, it has proven difficult so far to distinguish clear-cut subclasses of the stacking relations, although it is apparent that some base combinations prefer some stacking arrangements over others. Classification of base stacking interactions will require a combination of structural and energy data with substantial input from computations.

Two contributions dominate the direct stacking interactions (Figure 1): van der Waals interaction, which is a combination of short-range repulsion effects and dispersion attraction, which is roughly the Lennard-Jones term of MM force fields; and the electrostatic interaction, which is roughly the Coulombic term of MM force fields.[14b] The former term maximizes the overlap of the bases while minimizing steric clashes, vertical compressions and gaps between bases.[80] The latter term is responsible for the orientational component of stacking. There are no substantial specific "$\pi-\pi$" or "aromatic" effects associated with base stacking attributable to the delocalized π-electron cloud, and the currently used functions employed in MM force fields account well for base stacking.[14b] However, the electrostatic contribution to stacking free energies is dramatically counterbalanced by solvent screening effects,[27b] which can even, in appropriate structural context, stabilize stacking of consecutive positively charged base pairs.[81] This illustrates the fundamental problem in the interpretation of stacking calculations for biologically relevant contexts: the degree of solvent screening modulation of the stability of stacking interactions is highly context-dependent.[16f]

How To Compare Computed and Experimental Data? Base Stacking As the Case Example. As noted above, clarification of the nature of intrinsic base stacking and its subsequent quantification (cf. Figures 1 and 2) is widely accepted as a substantial success of QM methodology. Therefore, let us provide some additional data which illustrate what can be clarified by QM computations and what cannot. The best currently available QM methods (MP2 calculations extrapolated to the complete basis set of atomic orbitals supplemented by higher-order electron correlation calculations with smaller basis set) can derive highly accurate interaction energy for any single $x-y-z$ geometry of a pair of two nucleic acid bases (see above and Figures 1 and 2).[16] Selection of relevant $x-y-z$ geometry is nowadays a larger source of uncertainty than the interaction energy evaluations per se.

Let us consider calculation of interaction energy between two bases. The monomer geometries need to be sufficiently relaxed using a good-quality method. Substantially unrelaxed geometries may compromise the electronic structure (incorrect dipoles, for example), which may bias the intermolecular terms. There are two limit cases of such stacking calculations, both valid: (i) Calculations carried out with rigid monomers (relaxed in isolation) neglect the intramolecular relaxation processes due to intermolecular interactions. (ii) Computations carried out with complete relaxation of the whole system fully include mutual structural adaptation of the interacting monomers, as would

TABLE 1: Energies of Base Stacking in B-DNA Base Pair Steps (i.e., between two consecutive base pairs in B-DNA geometry)[a]

5′-XY-3′	2006 QM pair	2006 QM + mb	1997 QM pair	AMBER/MP2	exp ΔG	exp ΔH	2006 QM water	Ornstein
GG═CC	−13.7	−11.2	−11.5	−13.8	−1.8	−8.0	−9.4	−8.3
GC	−16.6	−15.8	−14.1	−15.6	−2.3	−9.8	−10.3	−9.7
CG	−18.4	−17.3	−13.8	−16.3	−2.1	−10.6	−9.2	−14.6
AA═TT[b]	−13.1	−13.1	−12.0	−14.7			−9.9	−5.4
AT	−13.3	−13.3	−11.6	−15.6	−0.7	−7.2	−11.9	−3.8
TA	−13.0	−12.8	−11.2	−14.2	−0.6	−7.2	−9.2	−6.7
AG═CT	−14.3	−12.5	−12.2	−14.9	−1.2	−8.4	−9.8	−9.8
GA═TC	−13.6	−12.9	−12.1	−13.7	−1.5	−7.8	−10.2	−6.8
AC═GT	−14.2	−13.4	−12.3	−14.6	−1.4	−8.2	−10.2	−6.6
CA═TG	−16.0	−15.1	−12.5	−15.7	−1.4	−8.4	−9.2	−10.5
AA(prop)[c]	−14.7	−14.7		−15.8	−1.4	−8.5	−11.5	

[a] 2006 QM pair: 2006 QM reference data calculated as sum of four base−base terms.[16e] Idealized geometries with helical twist of 36°. Propeller twist and vertical separation between base pairs are optimized. 2006 QM + mb: the preceding data with added many-body term. 1997 QM: 1997 reference QM calculations.[14e] AMBER/MP2: Cornell et al. force field.[12] The atomic charges derived with inclusion of electron correlation effects (MP2 method): see the text. Exp ΔG: reference experimental values of B-DNA base pair step free energies.[23b] Exp ΔH: the corresponding enthalpies.[23b] 2006 QM water: the 2006 QM reference values corrected for water solvent screening effects using continuum solvent model, obtained by combining the "2006 QM + mb" data with Table 7 B3LYP data from ref 16e. Ornstein: forty-year-old semiempirical QM calculations.[13a] [b] Propeller twist 0. [c] Propeller twist was optimized and has a value of −20°.

occur in the gas phase. The two stacked bases are visibly deformed toward each other. Neither approach is perfect. Deformations of monomers seen upon full gas phase optimization are exaggerated compared with stacking in biomolecular systems, condensed phase, or solid state experiments, where the bases are surrounded by other interacting partners from all sides. The approximation of entirely rigid monomers is also not fully realistic. The later calculations, however, allow sampling of the conformational space. These two approaches can be combined by freezing the intermolecular geometry while at least partially relaxing the monomers.[58c] H-bonded base pairs need to be always relaxed to allow full optimization of the H-bonds that includes stretching of the X−H covalent bonds of the donors.

Table 1[14e,16e,23b] compiles computed and experimental data for stacking between two consecutive base pairs for all 10 independent steps in B-DNA. (The available theoretical B-DNA data are more complete than A-RNA data, although the main conclusions would be the same.) The "2006 QM pair" column presents the current reference values obtained using idealized geometries of base pair steps with helical twist 36°, optimized propeller twist, and optimal vertical separation.[16e] Optimized base pair geometries were used as the starting structures to construct the base pair step geometries with rigid monomer structures. For the 5′-AA-3′ step, due to its prominent propeller twisting tendency, we show geometries with 0 and optimized (−20°) propeller twist. The energies were calculated as a sum of four pair base−base stacking contributions. Energies of H-bonded base pairs are not included to get pure stacking energies. The second "2006 QM + mb" column shows stacking energies upon addition of the many-body term, that is, nonadditivity of base stacking. The B-DNA stacking energies are −13.1 to −18.4 for the pair additive calculation and −11.2 to −17.3 after adding the many-body term. The most interesting step is the 5′-GG-3′ one, which has rather unfavorable intrastrand electrostatics due to the two intrastrand GG and CC homostacks of highly polar G (dipole moment of ∼6.6 D) and C (dipole moment of ∼6.4 D). T and A have dipoles of ∼4.3 and ∼2.6 D.[14d] The GG step is the only one with a significant mb term of +2.2 kcal/mol, meaning that the stacking is anticooperative. For more details see the literature 16e.

The third column (1997 QM pair) shows the 1997 benchmark calculations, within the pair approximation, for similar but not identical geometries.[14e] There is a meaningful agreement between the first and third column data, showing that the 1997 calculations were already qualitatively correct. The fourth column shows AMBER Cornell et al. force field[12] calculations that are in very reasonable agreement with the reference data. It illustrates the success of this particular force field in description of base stacking that stems from using atomic charges derived to reproduce the electrostatic potentials around the monomers. The calculations in Table 1 were done with charge derivation using QM method with inclusion of electron correlation, which allows consistent comparison between the QM and force field data. The actual simulation force field is derived using the same basic scheme but with the uncorrelated HF approximation, which overpolarizes the monomer charge distributions. HF-derived charges may be more suitable for condensed phase simulations with nonpolarizable force fields, since real molecules are polarized by water. The utilization of QM methods to verify and parametrize force fields is straightforward, and via improving force fields, the QM methods indirectly influence our knowledge of nucleic acids. Nevertheless, the force fields remain inevitably approximate. Force fields neglect polarization and charge transfer effects. Force fields do not allow to describe nonclassical H-bonds and amino-acceptor interactions utilizing the intrinsic flexibility of amino groups.[14c,d,15,77,78] Force fields are inherently less accurate in description of backbone topologies, which must be carefully tuned by nonphysical dihedral "cosine" force field terms. The Lennard-Jones empirical potential with excessively steep r^{-12} repulsion term (r, interatomic distance) is inexact in description of close interatomic contacts, since correct description of the short-range repulsion would require an exponential term.[80b]

Columns 5 and 6 bring the reference experimental nearest-neighbor ΔG and ΔH parameters for B-DNA base pair steps. Thermodynamics (TD) measurements and the derived TD stability parameters have enormous impact on 2D predictions and many other applications. Nevertheless, the forces determining the TD data are not fully understood. The Table shows that it is not easy to directly compare the TD reference data with the QM reference data. As already pointed out, QM and TD data represent a valid description of the interactions upon different conditions. QM calculations show the net stacking energy for a given $x-y-z$ geometry. The experiment shows the overall thermodynamics stability associated with a given stacked base pair step in the context of B-DNA, which includes not only stacking, but also base pairing, all populated geometries,

the presence of backbone, all solvent effects, ion binding, etc. Still, the experimental data reflect some of the gas phase trends. The stability order increases with the number of GC base pairs (0, 1, or 2), which reflects the intrinsic stability of GC and AT base pairs. For the GC, CG, and GG base pair steps, the enthalpies reflect the gas phase stability order, mainly the relatively low stability of the GG step. This agreement can nevertheless be incidental. The measured TD parameters may often be radically affected by incidental contributions that differ from case to case, rather than being determined by the most fundamental forces, such as base stacking.[60b,61]

Let us assume that we deal with two configurations. One of them is optimally hydrated by an integer number of waters and the other is not. The second configuration may be penalized due to unoptimal distribution of hydration sites. As another example, let us consider the guanine to inosine (I) substitution in a GC WC base pair. The GC and IC base pairs are isosteric, and except for the missing NH_2 group, the electronic structures of I and G (such as the dipole moment magnitude and orientation) are rather similar.[14d] So we do not expect a radical effect of such substitution on base stacking and pairing, except for preventing minor groove clash in the B-DNA for the respective 5′-PyrPur-3′ step. Yet, there is a striking ∼1.6 kcal/mol free energy difference between equivalent G → I substitutions in canonical B-DNA and A-RNA.[82] TD studies sometimes attempt to rationalize the measured trends by intrinsic stacking and base-pairing interactions, but not always considering the corresponding modern physical-chemistry data. Some of the assumptions are then not in agreement with modern physical chemistry of molecular interactions. We suggest that if TD experiments are discussed using the intrinsic molecular interactions such as stacking and base pairing, it should be done using modern physical chemistry computations.[83] If a meaningful correlation between TD data and the intrinsic forces does not exist, then it should be understood as result of the overall complex balance of molecular forces.

Inclusion of solvent effects could bring the calculations closer to experiment. Relatively straightforward approach is to use continuum solvent approximations (see above). Thus, the seventh column of Table 1 gives the 2006 QM reference calculations extended by continuum solvent (water) calculations by combing (using B3LYP data) results of Tables 4 and 7 of ref 16e. Such calculations include the effect of solvent screening of the electrostatic interactions; however, they still do not allow direct comparison with the experiments because many other contributions remain excluded. The calculations are still using just a single static geometry, do not include solute entropy terms, do not include the rest of the NA molecule (i.e., the two stacked base pairs are fully immersed to water), and do not include specific water binding, etc. In fact, comparing the QM data for GC, CG, and GG steps with the TD data, we see that the above-noted correlation for ΔH is lost. So linking such computations (for systems as complex as nucleic acids) to existing experiments is not trivial. One important feature revealed by such calculations is nevertheless clear. The solvent screening is effectively suppressing (or counterbalancing) the electrostatic energy contribution to stacking, which dominates the orientation dependence of base stacking in the gas phase.[27b] It agrees with the empirical experience.

For decades, structural biologists have rationalized stability of stacking as dispersion-controlled and hydrophobic interaction based solely on the degree of mutual overlap of the stacked bases, not considering the mutual orientations of nucleobases which vary widely. This simple approach, which is equivalent to switching off the electrostatic term in computations, is quite insightful in linking structural data with biochemical data. Apparently, orientation of stacking geometries in nucleic acids is not determined by the electrostatic part of stacking, definitely not to the extent seen in gas phase computations. From this point of view, telling that modern QM calculations revealed the role of dispersion forces in nucleic acids is not a fully accurate statement. The role of dispersion has been well-known in experimental science and has been quite accurately included, even in the oldest empirical force fields. The right statement is that the correct evaluation of the (roughly known) dispersion forces in QM computations has been achieved upon inclusion of a large portion of the intermolecular electron correlation effects as the last step to reach chemical accuracy and to provide the ultimate and unambiguous picture of the interaction. The fundamental issue of the degree of expression/attenuation of electrostatic effects in nucleic acids awaits an in-depth physical chemistry analysis because it likely varies from context to context while evolution is utilizing this variability.[14d]

To complete the comparison, the last column of the Table shows B-DNA stacking energy data derived in 1978 by Ornstein et al.[13a] The stacking energies range from −4 to −15 kcal/mol, whereas the stability order has no correspondence to modern QM calculations. Similarly, the 1962 data by deVoe and Tinoco (from −2 kcal/mol, AT and CG steps to −16 kcal/mol, GC step)[84] and 1976 data by Kudriatskaya and Danilov (from −7 kcal/mol for the AT step to −24 kcal/mol for the GC step)[85] do not resemble modern calculations. This reflects the insurmountable limitations of these pioneering calculations in prehistoric quantum chemistry before the advance of modern computers. It is, however, hardly justified to use these older calculations in any discussions of molecular interactions, as still sometimes happens. Actually, the first calculations capable of giving a meaningful stacking estimate are the 1988 first ab initio data by Aida,[86] being in the range of ∼−7 to −12 kcal/mol and basically reflecting the correct order of stacking stabilities (See Figure 3 in ref 14e). Thus, meaningful ab initio QM data is available for more than 20 years in the literature, albeit the first 1988 attempt could get only a fraction of the dispersion energy. Within 20 years, the ab initio calculations matured to chemical accuracy and completeness in stacking calculations.[16f]

Table 2 compares base stacking in 10 unique B-DNA and A-RNA steps, compiled from the recent study by Svozil et al.[22] In contrast to Table 1, the geometries are now derived from explicit solvent MD simulation trajectories. The approximate nature of the force field means that the populated structures are not perfect (helical twist is underestimated, etc.) and introduce errors into the calculations. Nevertheless, the simulation allows one to monitor the genuine thermal fluctuations of stacking instead of using just a single static geometry. Thus, for each step sequence, 10−50 individual geometries of A-RNA and 50 for B-DNA are evaluated.

The first column gives the average value of stacking energy; the next column, the standard deviation; and the subsequent two columns give maximum and minimum values of the calculated stacking energies. The geometries are based on 400-ps averaged portions of trajectories, which is a substantial smoothing of the thermal fluctuations. Individual snapshots would be even more diverse. The force field geometries are replaced by QM-optimized geometries of bases, and the stacking energy is calculated using the fast DFT-D method (see ref 16f for an explanation). It would not be tractable to use the best calculations for almost 1000-base pair step stacking evaluations. However, to make Table 2 comparable to Table 1, the DFT-D

TABLE 2: Comparison of Base Stacking in B-DNA and A-RNA Base Pair Steps Based on Evaluation of Series of 400-ps Averaged Geometries along Explicit Solvent Simulation Trajectories[22] [a]

	B-DNA				A-RNA			
5′-XY-3′	AVG	SD	MAX	MIN	AVG	SD	MAX	MIN
GG	−9.72	0.42	−11.68	−8.89	−8.57	0.40	−9.26	−7.31
GC	−15.88	0.55	−16.94	−14.63	−15.72	0.50	−16.63	−14.72
CG	−15.75	0.40	−16.50	−14.53	−15.21	0.56	−16.06	−14.45
AA	−12.87	0.50	−14.09	−11.73	−10.12	0.23	−10.53	−9.75
AT(U)	−13.63	0.23	−14.08	−12.97	−10.31	0.19	−10.77	−9.77
T(U)A	−13.44	0.77	−14.41	−11.58	−13.42	0.33	−13.89	−12.61
AG═CT(U)	−12.55	0.26	−13.27	−12.08	−11.38	0.15	−11.84	−11.08
GA═T(U)C	−12.28	0.30	−12.99	−11.62	−11.65	0.41	−12.70	−10.95
AC═GT(U)	−13.70	0.53	−14.65	−12.26	−11.43	0.22	−11.77	−10.88
CA═T(U)G	−13.00	0.38	−14.14	−11.68	−12.26	0.28	−12.76	−11.55

[a] AVG, SD, MAX, and MIN stand for the averaged value of base stacking, standard deviation, the maximum value, and the minimum value. The energies are derived using DFT-D approach and are further corrected using the highest-quality calculations carried out for one single geometry of each step; see the text.

energies are adjusted by the highest-accuracy calculations done for single A-RNA and B-DNA geometry of each sequence. This correction is in the range of −1.14 to −2.06 kcal/mol. Thus, the data in the present Table 2 are compiled from Tables 1, 2, 3, and 4 of the original work.[22]

The stacking energy varies significantly along the trajectories. Note that all the single geometries are meaningful and representative. Thus, the inevitable conclusion is that base stacking cannot be completely represented on the basis of single geometries, irrespective of how carefully these geometries are designed and selected. The A-RNA and B-DNA intrinsic stacking energies are similar, and most of the DNA/RNA energy differences in the Table 2 can be rationalized by presence of T in DNA and U in RNA.[22] For the sake of completeness, let us reiterate that utilization of experimental geometries is also not problem-free.[16f] The experiments provide static and averaged structures while even modest data and refinement coordinate errors of X-ray structures may substantially bias energy calculations. In summary, despite that quantum chemistry nowadays provides very accurate structure-energy relation (energy as a function of geometry) for base stacking, finding fully transparent links to various experimental data is not straightforward.

RNA Sugar−Phosphate Backbone. The sugar−phosphate backbone is chemically monotonous (sequence-independent) and contains consecutive single bonds with a substantial freedom for correlated torsional rotations. It has often been assumed that the backbone plays a rather passive role in structuring (as opposed to stabilizing) nucleic acids. According to this "base-centered" view of nucleic acid structure, interactions directly involving nucleobases are decisive in organizing the 3D structure.[18] However, others have suggested that backbone conformational preferences are also crucial. We take the view that both noncovalent molecular interactions and backbone internal conformational preferences are important.

Backbone torsional angles are highly correlated, reflecting the intrinsic conformational preferences and topological requirements of nucleic acids.[34] However, the role of backbone conformational preferences in determining 3D structures is less understood than the role of nucleobases, both theoretically and experimentally. Characterization of the sugar−phosphate backbone is a formidable task for QM investigations (see above). In addition, the MM force fields have limited accuracy, in part because the use of conformation-independent atomic charges is inadequate to properly describe the energetics of the flexible phosphodiester chains. Although it is possible to determine the positions of the nucleobases and the centers of phosphate groups quite reliably by X-ray diffraction, even at moderate resolution (∼2.5 Å), it is much more difficult at the same resolution to determine the precise backbone conformation, especially for the sugar atoms. A classification of RNA backbone conformations has been proposed,[34a] but some of the less populated backbone families may be artifacts of the limits of the resolution. Moreover, some individual geometries do not fit any of the suggested families. Future QM calculations will bring new data to bear on the problem of classifying sugar−phosphate conformational preferences and their energetics. The calculations will be complicated by the ribose 2′-OH group.

RNA as a Big Jigsaw Toy, Marvelous LEGO or a Russian Doll. To choose model systems for computations, it is instructive to think about large RNAs such as the rRNAs as toys composed of intricate, interlocking parts, like puzzles, LEGOs, or Russian dolls. The isostericity principle with precisely shaped non-WC interactions resembles a complex jigsaw puzzle. Natural selection favors "pieces" that preserve the local RNA shape, but also requires adequate interaction energy, although the most stable interaction is not necessarily the best. Many of the interactions possess substates, which are important for functional dynamics. QM calculations can help to elucidate the principles governing the individual interactions that put each jigsaw puzzle in the right place.

The ribosome also works like a sophisticated LEGO toy. It utilizes recurrent modular building blocks and is highly dynamical. Dynamics is critical for the function, and is not evident from individual structures alone. Much current work using a variety of experimental and computational tools aims to characterize the functional dynamics of the ribosome and other RNA-based molecular machines.[3f,87] Large RNA-based nanomachines, such as the ribosome, work in the regime of high viscosity and very low inertia so that the essential principles of their function are quite different from those of macroscopic machines. Molecular machines are subject to persistent large thermal fluctuations. They use chemical energy to rectify random thermal fluctuations into directional motions. These functional motions are largely driven by stochastic processes, in which fluctuations are of utmost importance. The structural data represent static pictures of the molecular machine, averaged over a large number of particles, over the time scale of the experiments, and with limited resolution. Therefore, theory could bring important insights to bear on the relation between thermal fluctuations and flexibility. Obviously, this is a task for MD technique utilizing classical force fields. Nevertheless, force fields are very approximate and never perfect. QM calculations

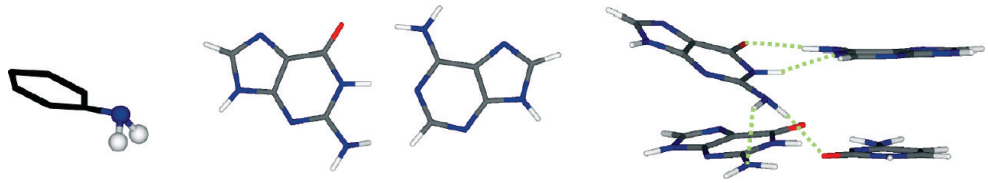


Figure 10. Amino groups of isolated bases are intrinsically nonplanar due to partial sp^3 hybridization of the amino group nitrogen atoms[15a] which is not included in MM force fields. Left: scheme of the pyramidalization, which means that the sum of amino group valence angles is less than 360°. The amino groups are planarized by primary in-plane H-bonds (canonical base pairs) but the amino group flexibility can stabilize specific interactions with out-of-plane (with respect to the nucleobase) distribution of donors and acceptors. Such local environments are rather common in folded RNAs. Middle: cWW GA base pair. Right: typical stacking in a 5′-GG-3′/5′-AC-3′ base pair step seen in RNA X-ray structures, with the cWW GA base pair stacked on top of the canonical GC one.[58c] The base positions are taken from the experiment; the positions of the hydrogens are predicted via QM. The profound nonplanarity of the GA base pair is its intrinsic gas phase feature that remains fully expressed in the experimental structures. The nonplanar guanine amino group is involved in an out-of-plane H-bond, which is a characteristic interaction of the 5′-GG-3′/5′-AC-3′ internal loop and provides a constraint on the RNA sequence.

will play a large role in future refinements of the MD force fields and in assessing their limitations for specific types of interactions and molecular architectures. A combination of QM calculations and RNA structural bioinformatics could provide important feedback to modify the force fields in a targeted manner to improve the description of specific types of RNA interactions, submotifs, and motifs, even when full scale force field reparameterization is not achievable.

Last but not the least, RNA architectures are hierarchical, resembling Russian dolls. Typically a given RNA structural interaction pattern or motif (with its associated sequence signature) includes subpatterns or submotifs while it also participates in larger motifs and contexts.[73] This complicates the definition of model systems for computations. However, systematic computations could bring important insights into the basic physical chemistry principles of the RNA structural hierarchy.

From Intrinsic Interactions to Covariation of RNA Sequences. Above, we have discussed the relationship between physicochemical insights provided by modern theoretical computations and RNA structural bioinformatics, pointing out with examples the many reasons these two research areas can benefit from close cooperation. We conclude with one final instructive example, which shows that when we know what to track down, we can find surprising relations ranging from subtle gas phase effects through structural and thermodynamics data up to evolutionary covariation patterns. The GA cWW base pair is stabilized by two primary H-bonds while the guanine N2 amino group and adenine C2 are juxtaposed (Figure 10[15a,58c]). The latter interaction is repulsive in the planar conformation. Thus, the base pair undergoes large propeller twisting (counter-rotation of the bases) around its major groove edge. This positions the minor groove guanine amino group away from the adenine plane (Figure 10). The unpaired amino group also utilizes its genuine capability to assume a pyramidal geometry. It exposes the G(N2) lone pair to interact with the A(H2) while the amino group hydrogens can form out-of-plane H-bonds with adjacent base pairs.

The characteristic gas phase geometry can be found in many RNA and DNA X-ray structures.[88] However, these data were not properly interpreted until recently. The excessive propeller twisting was ad hoc attributed to the base stacking, whereas in reality, base stacking oppose it because it prefers parallel bases. The structural studies overlooked the stabilizing cross-strand out-of-plane H-bonds between the guanine amino group and the O2 of pyrimidine of the adjacent canonical base pair. Complementary insights emerged from NMR and thermodynamics studies.[60a,c,89] The cWW GA base pair typically occurs in the following sequence context in duplexes: 5′-RG-3′/5′-AY-3′, (R is A or G and Y is C or U/T). To allow the out-of-plane H-bond to form, Y must be located 3′ to the adenine with which the G is paired. Reversal of the adjacent canonical base pair (5′-YG-3′/5′-AR-3′) abolishes this interaction. Indeed, in the latter sequence context, the GA base pair adopts the tSH ("sheared") geometry, with an H-bond between G(N2) and A(N7). Thus, the GA base pair has context-dependent geometry. Finally, bioinformatics guided by QM calculations revealed that observation of cWW GA base pair with the out-of-plane interaction in parent RNA X-ray structure implies lack of GA to AG covariation in homologous sequences despite isostericity of cWW GA and AG base pairs.[58c]

Conclusions

This article jointly presents, for the first time, QM physical chemistry and structural bioinformatics perspectives of forces and rules that shape RNA structures. We suggest that convergence is possible between these heretofore quite separated research areas. Their synergy presents substantial potential to deepen our understanding of RNA structure, dynamics, and evolution. We have tried to highlight the unique contributions that each field provides and the value of respecting each field's unique perspective and limitations to obtain reliable and useful results. In conclusion, we propose that carrying out high-quality computations of recurrent structural motifs identified in experimental structures is especially useful in analyzing specific structural motifs and interactions. In this manner, the computations provide in-depth insights compensating some of the intrinsic weaknesses of structural bioinformatics, which include bias toward structural data and basically an understandable trend to derive conclusions on the basis of the most representative data. Important but specific (less frequent) strategies of molecular adaptation that evolution has discovered can be easily overlooked. In general, the research is initiated by bioinformatic analysis of structural data, because these data represent the primary source of our information and also because 3D structures are decisive for RNA evolution and function (as exemplified by the isostericity principle). The bioinformatics will provide the initial set of targets for the computations, literally guiding the computations through the overwhelming complexity of RNAs. Then, computations can provide, in turn, much-needed insights that can lead to further hypothesis that can be tested by bioinformatics or experiments. Nonetheless, the reverse scenario, in which insight from physical chemistry is used to uncover novel structural and sequence patterns, can also be fruitful. This interdisciplinary research should be preferably done with close interaction between RNA bioinformatics and computational researchers, with substantial involve-

ment from both sides. This on one side stems from the enormous complexity of RNA structural biology and evolution and on the other side also from the fact that it is not as easy to carry out competent computations as many researchers, who are not computational specialists, assume. When such cooperation is achieved, we will be rewarded by unique physicochemical insights, which will advance our understanding of RNA structural biology.

Acknowledgment. This work was supported by the Grant Agency of the Academy of Sciences of the Czech Republic (CR), grant IAA400040802; Grant Agency of the CR, grants 203/09/1476 and P208/10/2302; Ministry of Education of the CR, LC06030; Academy of Sciences of the CR, AV0Z50040507 and AV0Z50040702; grant from the National Institutes of Health (2 R15GM055898-05); and by a grant from the National Science Foundation (Research Coordination Network Grant no. 0443508).

Supporting Information Available: Figures illustrating the canonical GC and wobble GU base pairs as well as the possible substates of A-minor interactions. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) (a) Kruger, K.; Grabowski, P. J.; Zaug, A. J.; Sands, J.; Gottschling, D. E.; Cech, T. R. *Cell* **1982**, *31*, 147–157. (b) Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. *Cell* **1983**, *35*, 849–857.

(2) Wahl, M. C.; Will, C. L.; Luhrmann, R. *Cell* **2009**, *136*, 701–718.

(3) (a) Frank, J.; Zhu, J.; Penczek, P.; Li, Y. H.; Srivastava, S.; Verschoor, A.; Radermacher, M.; Grassucci, R.; Lata, R. K.; Agrawal, R. K. *Nature* **1995**, *376*, 441–444. (b) Yusupov, M. M.; Yusupova, G. Z.; Baucom, A.; Lieberman, K.; Earnest, T. N.; Cate, J. H. D.; Noller, H. F. *Science* **2001**, *292*, 883–896. (c) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. *Nature* **2000**, *407*, 327–339. (d) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 905–920. (e) Ramakrishnan, V.; Moore, P. B. *Curr. Opin. Struct. Biol.* **2001**, *11*, 144–154. (f) Mitra, K.; Frank, J. *Annu. Rev. Biophys. Biomol.* **2006**, *35*, 299–317. (g) Korostelev, A.; Ermolenko, D. N.; Noller, H. F. *Curr. Opin. Chem. Biol.* **2008**, *12*, 674–683.

(4) (a) Halic, M.; Becker, T.; Pool, M. R.; Spahn, C. M. T.; Grassucci, R. A.; Frank, J.; Beckmann, R. *Nature* **2004**, *427*, 808–814. (b) Egea, P. F.; Stroud, R. M.; Walter, P. *Curr. Opin. Struct. Biol.* **2005**, *15*, 213–220.

(5) Catania, F.; Gao, X.; Scofield, D. G. *J. Hered.* **2009**, *100*, 591–596.

(6) Gesteland, R. F.; Cech, T. R.; Atkins, J. F. *The RNA World*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2006.

(7) (a) Birney, E.; Stamatoyannopoulos, J. A.; Dutta, A.; et al. *Nature* **2007**, *447*, 799–816. (b) Carninci, P. *Trends Genet.* **2006**, *22*, 501–510. (c) Pheasant, M.; Mattick, J. S. *Genome Res.* **2007**, *17*, 1245–1253.

(8) Fire, A.; Xu, S. Q.; Montgomery, M. K.; Kostas, S. A.; Driver, S. E.; Mello, C. C. *Nature* **1998**, *391*, 806–811.

(9) (a) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790–18794. (b) Hobza, P.; Selzle, H. L.; Schlag, E. W. *Chem. Rev.* **1994**, *94*, 1767–1785.

(10) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.

(11) (a) Cieplak, P.; Dupradeau, F. Y.; Duan, Y.; Wang, J. M. *J. Phys.: Condens. Matter* **2009**, *21*, 333102. (b) Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(12) 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. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(13) (a) Ornstein, R. L.; Rein, R.; Breen, D. L.; Macelroy, R. D. *Biopolymers* **1978**, *17*, 2341–2360. (b) Langlet, J.; Claverie, P.; Caron, F.; Boeuve, J. C. *Int. J. Quantum Chem.* **1981**, *20*, 299–338. (c) Forner, W.; Otto, P.; Ladik, J. *Chem. Phys.* **1984**, *86*, 49–56. (d) Gresh, N.; Pullman, B. *Int. J. Quantum Chem.* **1985**, *12*, 49–56. (e) Aida, M.; Nagata, C. *Int. J. Quantum Chem.* **1986**, *29*, 1253–1261. (f) Hobza, P.; Sandorfy, C. *J. Am. Chem. Soc.* **1987**, *109*, 1302–1307. (g) Anwander, E. H. S.; Probst, M. M.; Rode, B. M. *Biopolymers* **1990**, *29*, 757–769. (h) Colson, A. O.; Besler, B.; Close, D. M.; Sevilla, M. D. *J. Phys. Chem.* **1992**, *96*, 661–668.

(14) (a) Hobza, P.; Sponer, J.; Polasek, M. *J. Am. Chem. Soc.* **1995**, *117*, 792–798. (b) Sponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem.* **1996**, *100*, 5590–5596. (c) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247–3276. (d) Sponer, J.; Leszczynski, J.; Hobza, P. *Biopolymers* **2001**, *61*, 3–31. (e) Sponer, J.; Gabb, H. A.; Leszczynski, J.; Hobza, P. *Biophys. J.* **1997**, *73*, 76–87.

(15) (a) Sponer, J.; Hobza, P. *J. Phys. Chem.* **1994**, *98*, 3161–3164. (b) Sponer, J.; Hobza, P. *J. Am. Chem. Soc.* **1994**, *116*, 709–714.

(16) (a) Hobza, P.; Sponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808. (b) Leininger, M. L.; Nielsen, I. M. B.; Colvin, M. E.; Janssen, C. L. *J. Phys. Chem. A* **2002**, *106*, 3850–3854. (c) Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613. (d) Sponer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151. (e) Sponer, J.; Jurecka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. *Chem.—Eur. J.* **2006**, *12*, 2854–2865. (f) Sponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595–2610.

(17) (a) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829. (b) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862. (c) Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(18) (a) Prive, G. G.; Yanagi, K.; Dickerson, R. E. *J. Mol. Biol.* **1991**, *217*, 177–199. (b) Suzuki, M.; Amano, N.; Kakinuma, J.; Tateno, M. *J. Mol. Biol.* **1997**, *274*, 421–435.

(19) (a) Leontis, N. B.; Stombaugh, J.; Westhof, E. *Nucleic Acids Res.* **2002**, *30*, 3497–3531. (b) Stombaugh, J.; Zirbel, C. L.; Westhof, E.; Leontis, N. B. *Nucleic Acids Res.* **2009**, *37*, 2294–2312.

(20) de Vries, M. S.; Hobza, P. *Annu. Rev. Phys. Chem.* **2007**, *58*, 585–612.

(21) Voet, D.; Voet, J. G.; Pratt, C. W. *Fundamentals of Biochemistry*, 3rd ed.; Wiley & Sons, Inc.: New York, 2008.

(22) Svozil, D.; Hobza, P.; Sponer, J. *J. Phys. Chem. B* **2010**, *114*, 1191–1203.

(23) (a) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. *J. Mol. Biol.* **1999**, *288*, 911–940. (b) SantaLucia, J. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1460–1465.

(24) Ditzler, M. A.; Otyepka, M.; Sponer, J.; Walter, N. G. *Acc. Chem. Res.* **2010**, *43*, 40–47.

(25) Banas, P.; Jurecka, P.; Walter, N. G.; Sponer, J.; Otyepka, M. *Methods* **2009**, *49*, 202–216.

(26) Sponer, J.; Lankas, F. *Computational Studies of RNA and DNA*; Springer: Dordrecht, 2006.

(27) (a) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093. (b) Florian, J.; Sponer, J.; Warshel, A. *J. Phys. Chem. B* **1999**, *103*, 884–892.

(28) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(29) (a) Bachs, M.; Luque, F. J.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 446–454. (b) Luque, F. J.; Curutchet, C.; Munoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpi, J. L.; Orozco, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3827–3836.

(30) (a) Sponer, J. E.; Reblova, K.; Mokdad, A.; Sychrovsky, V.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2007**, *111*, 9153–9164. (b) Mladek, A.; Sharma, P.; Mitra, A.; Bhattacharyya, D.; Sponer, J.; Sponer, J. E. *J. Phys. Chem. B* **2009**, *113*, 1743–1755.

(31) (a) Soteras, I.; Forti, F.; Orozco, M.; Luque, F. J. *J. Phys. Chem. B* **2009**, *113*, 9330–9334. (b) Soteras, I.; Orozco, M.; Luque, F. J. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 281–291. (c) Curutchet, C.; Orozco, M.; Luque, F. J. *J. Comput. Chem.* **2001**, *22*, 1180–1193. (d) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. *J. Mol. Struct.: THEOCHEM* **2005**, *727*, 29–40.

(32) (a) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 317–333. (b) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011–2033. (c) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396. (d) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 2447–2464.

(33) (a) Klamt, A.; Diedenhofen, M. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 357–360. (b) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(34) (a) Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M. *RNA* **2008**, *14*, 465–481. (b) Svozil, D.; Kalina, J.; Omelka, M.; Schneider, B. *Nucleic Acids Res.* **2008**, *36*, 3690–3706.

(35) Svozil, D.; Sponer, J. E.; Marchan, I.; Perez, A.; Cheatham, T. E.; Forti, F.; Luque, F. J.; Orozco, M.; Sponer, J. *J. Phys. Chem. B* **2008**, *112*, 8188–8197.

(36) (a) Millen, A. L.; Manderville, R. A.; Wetmore, S. D. *J. Phys. Chem. B* **2010**, *114*, 4373–4382. (b) Palamarchuk, G. V.; Shishkin, O. V.; Gorb, L.; Leszczynski, J. *J. Biomol. Struct. Dyn.* **2009**, *26*, 653–661. (c) MacKerell, A. D. *J. Phys. Chem. B* **2009**, *113*, 3235–3244. (d) Foloppe, N.; Hartmann, B.; Nilsson, L.; MacKerell, A. D. *Biophys. J.* **2002**, *82*, 1554–1569. (e) Foloppe, N.; Nilsson, L.; MacKerell, A. D. *Biopolymers* **2001**, *61*, 61–76. (f) Hocquet, A.; Leulliot, N.; Ghomi, M. *J. Phys. Chem. B* **2000**, *104*, 4560–4568. (g) Florian, J.; Strajbl, M.; Warshel, A. *J. Am. Chem.*

Soc. **1998**, *120*, 7959–7966. (h) Sychrovsky, V.; Vokacova, Z.; Sponer, J.; Spackova, N.; Schneider, B. *J. Phys. Chem. B* **2006**, *110*, 22894–22902.
(37) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.
(38) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon Press: Oxford, 1990.
(39) Hobza, P.; Sponer, J.; Cubero, E.; Orozco, M.; Luque, F. J. *J. Phys. Chem. B* **2000**, *104*, 6286–6292.
(40) Guerra, C. F.; Bickelhaupt, F. M. *Angew. Chem., Int. Ed.* **2002**, *41*, 2092–2095.
(41) Sponer, J.; Zgarbova, M.; Jurecka, P.; Riley, K. E.; Sponer, J. E.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1166–1179.
(42) Hesselmann, A.; Jansen, G.; Schutz, M. *J. Am. Chem. Soc.* **2006**, *128*, 11730–11731.
(43) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856–882.
(44) (a) Gresh, N.; Sponer, J. E.; Spackova, N.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2003**, *107*, 8669–8681. (b) Gresh, N.; Sponer, J. *J. Phys. Chem. B* **1999**, *103*, 11415–11427.
(45) (a) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930. (b) Hesselmann, A.; Jansen, G.; Schutz, M. *J. Chem. Phys.* **2005**, *122*, 014103.
(46) (a) Mattick, J. S. *J. Exp. Biol.* **2007**, *210*, 1526–1547. (b) Beniaminov, A.; Westhof, E.; Krol, A. *RNA* **2008**, *14*, 1270–1275.
(47) (a) Mokdad, A.; Krasovska, M. V.; Sponer, J.; Leontis, N. B. *Nucleic Acids Res.* **2006**, *34*, 1326–1341. (b) Leontis, N. B.; Westhof, E. *RNA* **2001**, *7*, 499–512. (c) Nissen, P.; Ippolito, J. A.; Ban, N.; Moore, P. B.; Steitz, T. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4899–4903. (d) Tamura, M.; Holbrook, S. R. *J. Mol. Biol.* **2002**, *320*, 455–474. (e) Gagnon, M. G.; Steinberg, S. V. *RNA* **2002**, *8*, 873–877. (f) Noller, H. F. *Science* **2005**, *309*, 1508–1514.
(48) (a) Swart, M.; Guerra, C. F.; Bickelhaupt, F. M. *J. Am. Chem. Soc.* **2004**, *126*, 16718–16719. (b) Perez, A.; Sponer, J.; Jurecka, P.; Hobza, P.; Luque, F. J.; Orozco, M. *Chem.—Eur. J.* **2005**, *11*, 5062–5066.
(49) Acosta-Silva, C.; Branchadell, V.; Bertran, J.; Oliva, A. *J. Phys. Chem. B* **2010**, *114*, 10217–10227.
(50) (a) Vertessy, B. G.; Toth, J. *Acc. Chem. Res.* **2009**, *42*, 97–106. (b) Krokan, H. E.; Drablos, F.; Slupphaug, G. *Oncogene* **2002**, *21*, 8935–8948.
(51) (a) Parisien, M.; Major, F. *Nature* **2008**, *452*, 51–55. (b) Das, R.; Karanicolas, J.; Baker, D. *Nat. Methods* **2010**, *7*, 291–294.
(52) Mathews, D. H.; Turner, D. H. *Curr. Opin. Struct. Biol.* **2006**, *16*, 270–278.
(53) Deigan, K. E.; Li, T. W.; Mathews, D. H.; Weeks, K. M. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 97–102.
(54) Mathews, D.; Disney, M.; Childs, J.; Schroeder, S.; Zuker, M.; Turner, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7287–7292.
(55) (a) Eddy, S. R.; Durbin, R. *Nucleic Acids Res.* **1994**, *22*, 2079–2088. (b) Gardner, P. P.; Giegerich, R. *BMC Bioinformatics* **2004**, *5*, 140.
(56) (a) Michel, F.; Westhof, E. *J. Mol. Biol.* **1990**, *216*, 585–610. (b) Massire, C.; Jaeger, L.; Westhof, E. *J. Mol. Biol.* **1998**, *279*, 773–793. (c) Michel, F.; Costa, M.; Massire, C.; Westhof, E. *Methods Enzymol.* **2000**, *317*, 491–510.
(57) Brown, J. W.; Birmingham, A.; Griffiths, P. E.; Jossinet, F.; Kachouri-Lafond, R.; Knight, R.; Lang, B. F.; Leontis, N.; Steger, G.; Stombaugh, J.; Westhof, E. *RNA* **2009**, *15*, 1623–1631.
(58) (a) Reblova, K.; Spackova, N.; Stefl, R.; Csaszar, K.; Koca, J.; Leontis, N. B.; Sponer, J. *Biophys. J.* **2003**, *84*, 3564–3582. (b) Reblova, K.; Strelcova, Z.; Kulhanek, P.; Besseova, I.; Mathews, D. H.; Van Nostrand, K.; Yildirim, I.; Turner, D. H.; Sponer, J. *J. Chem. Theory Comput.* **2010**, *6*, 910–929. (c) Sponer, J.; Mokdad, A.; Sponer, J. E.; Spackova, N.; Leszczynski, J.; Leontis, N. B. *J. Mol. Biol.* **2003**, *330*, 967–978.
(59) Freier, S. M.; Sugimoto, N.; Sinclair, A.; Alkema, D.; Neilson, T.; Kierzek, R.; Caruthers, M. H.; Turner, D. H. *Biochemistry* **1986**, *25*, 3214–3219.
(60) (a) Yildirim, I.; Turner, D. H. *Biochemistry* **2005**, *44*, 13225–13234. (b) Kopitz, H.; Zivkovic, A.; Engels, J. W.; Gohlke, H. *ChemBioChem* **2008**, *9*, 2619–2622. (c) Yildirim, I.; Stern, H. A.; Sponer, J.; Spackova, N.; Turner, D. H. *J. Chem. Theory Comput.* **2009**, *5*, 2088–2100.
(61) Koller, A. N.; Bozilovic, J.; Engels, J. W.; Gohlke, H. *Nucleic Acids Res.* **2010**, *38*, 3133–3146.
(62) Kool, E. T. *Annu. Rev. Biochem.* **2002**, *71*, 191–219.
(63) Krueger, A. T.; Kool, E. T. *Curr. Opin. Chem. Biol.* **2007**, *11*, 588–594.
(64) (a) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. *J. Math. Biol.* **2008**, *56*, 215–252. (b) Nozinovic, S.; Furtig, B.; Jonker, H. R. A.; Richter, C.; Schwalbe, H. *Nucleic Acids Res.* **2010**, *38*, 683–694.
(65) (a) Leontis, N. B.; Lescoute, A.; Westhof, E. *Curr. Opin. Struct. Biol.* **2006**, *16*, 279–287. (b) Lescoute, A.; Leontis, N. B.; Massire, C.; Westhof, E. *Nucleic Acids Res.* **2005**, *33*, 2395–2409.
(66) (a) Lescoute, A.; Westhof, E. *RNA* **2006**, *12*, 83–93. (b) Laing, C.; Schlick, T. *J. Mol. Biol.* **2009**, *390*, 547–559.
(67) Klein, D. J.; Schmeing, T. M.; Moore, P. B.; Steitz, T. A. *EMBO J.* **2001**, *20*, 4214–4221.
(68) (a) Razga, F.; Koca, J.; Sponer, J.; Leontis, N. B. *Biophys. J.* **2005**, *88*, 3466–3485. (b) Reblova, K.; Razga, F.; Li, W.; Gao, H. X.; Frank, J.; Sponer, J. *Nucleic Acids Res.* **2010**, *38*, 1325–1340.
(69) (a) Correll, C. C.; Freeborn, B.; Moore, P. B.; Steitz, T. A. *Cell* **1997**, *91*, 705–712. (b) Correll, C. C.; Munishkin, A.; Chan, Y. L.; Ren, Z.; Wool, I. G.; Steitz, T. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13436–13441. (c) Correll, C. C.; Wool, I. G.; Munishkin, A. *J. Mol. Biol.* **1999**, *292*, 275–287.
(70) Parisien, M.; Cruz, J. A.; Westhof, E.; Major, F. *RNA* **2009**, *15*, 1875–1885.
(71) (a) Shankar, N.; Kennedy, S. D.; Chen, G.; Krugh, T. R.; Turner, D. H. *Biochemistry* **2006**, *45*, 11776–11789. (b) Lee, J. C.; Gutell, R. R.; Russell, R. *J. Mol. Biol.* **2006**, *360*, 978–988.
(72) Nasalean, L.; Stombaugh, J.; Zirbel, C.; Leontis, N.; Walter, N. G.; Woodson, S.; Batey, R. T. *Non-Protein Coding RNAs* **2009**, *13*, 1–26.
(73) Jaeger, L.; Verzemnieks, E. J.; Geary, C. *Nucleic Acids Res.* **2009**, *37*, 215–230.
(74) Besseova, I.; Otyepka, M.; Reblova, K.; Sponer, J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10701–10711.
(75) (a) Sponer, J. E.; Spackova, N.; Kulhanek, P.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. A* **2005**, *109*, 2292–2301. (b) Sponer, J. E.; Spackova, N.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2005**, *109*, 11399–11410. (c) Sponer, J. E.; Leszczynski, J.; Sychrovsky, V.; Sponer, J. *J. Phys. Chem. B* **2005**, *109*, 18680–18689. (d) Sharma, P.; Sponer, J. E.; Sponer, J.; Sharma, S.; Bhattacharyya, D.; Mitra, A. *J. Phys. Chem. B* **2010**, *114*, 3307–3320.
(76) (a) Sharma, P.; Mitra, A.; Sharma, S.; Singh, H.; Bhattacharyya, D. *J. Biomol. Struct. Dyn.* **2008**, *25*, 709–732. (b) Oliva, R.; Tramontano, A.; Cavallo, L. *RNA* **2007**, *13*, 1427–1436. (c) Oliva, R.; Cavallo, L.; Tramontano, A. *Nucleic Acids Res.* **2006**, *34*, 865–879. (d) Sharma, P.; Chawla, M.; Sharma, S.; Mitra, A. *RNA* **2010**, *16*, 942–957.
(77) Luisi, B.; Orozco, M.; Sponer, J.; Luque, F. J.; Shakked, Z. *J. Mol. Biol.* **1998**, *279*, 1123–1136.
(78) Vlieghe, D.; Sponer, J.; Van Meervelt, L. *Biochemistry* **1999**, *38*, 16443–16451.
(79) Zirbel, C. L.; Sponer, J. E.; Sponer, J.; Stombaugh, J.; Leontis, N. B. *Nucleic Acids Res.* **2009**, *37*, 4898–4918.
(80) (a) Sponer, J.; Kypr, J. *J. Biomol. Struct. Dyn.* **1993**, *11*, 277–292. (b) Morgado, C. A.; Jurecka, P.; Svozil, D.; Hobza, P.; Sponer, J. *J. Chem. Theory Comput.* **2009**, *5*, 1524–1544.
(81) Spackova, N.; Berger, I.; Egli, M.; Sponer, J. *J. Am. Chem. Soc.* **1998**, *120*, 6147–6151.
(82) (a) Siegfried, N. A.; Metzger, S. L.; Bevilacqua, P. C. *Biochemistry* **2007**, *46*, 172–181. (b) Siegfried, N. A.; Kierzek, R.; Bevilacqua, P. C. *J. Am. Chem. Soc.* **2010**, *132*, 5342–5344.
(83) Meneni, S. R.; Shell, S. M.; Gao, L.; Jurecka, P.; Lee, W.; Sponer, J.; Zou, Y.; Chiarelli, M. P.; Cho, B. P. *Biochemistry* **2007**, *46*, 11263–11278.
(84) Devoe, H.; Tinoco, I. *J. Mol. Biol.* **1962**, *4*, 500–517.
(85) Kudritskaya, Z. G.; Danilov, V. I. *J. Theor. Biol.* **1976**, *59*, 303–318.
(86) Aida, M. *J. Theor. Biol.* **1988**, *130*, 327–335.
(87) (a) Ninio, J. *Biochimie* **2006**, *88*, 963–992. (b) Blanchard, S. C.; Kim, H. D.; Gonzalez, R. L.; Puglisi, J. D.; Chu, S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12893–12898. (c) Frank, J.; Spahn, C. M. T. *Rep. Prog. Phys.* **2006**, *69*, 1383–1417. (d) Spirin, A. S. *FEBS Lett.* **2002**, *514*, 2–10.
(88) (a) Prive, G. G.; Heinemann, U.; Chandrasegaran, S.; Kan, L. S.; Kopka, M. L.; Dickerson, R. E. *Science* **1987**, *238*, 498–504. (b) Ennifar, E.; Yusupov, M.; Walter, P.; Marquet, R.; Ehresmann, B.; Ehresmann, C.; Dumas, P. *Structure* **1999**, *7*, 1439–1449.
(89) (a) Wu, M.; Turner, D. H. *Biochemistry* **1996**, *35*, 9677–9689. (b) Santalucia, J.; Turner, D. H. *Biochemistry* **1993**, *32*, 12612–12623.

JP104361M