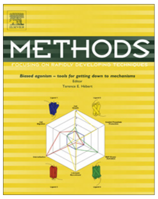
Methods 103 (2016) 138–156

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



Coarse-grained modeling of RNA 3D structure

Wayne K. Dawson [a,*], Maciej Maciejczyk [a,b], Elzbieta J. Jankowska [a], Janusz M. Bujnicki [a,c,*]

[a] Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland
[b] Department of Physics and Biophysics, University of Warmia and Mazury in Olsztyn, ul. Oczapowskiego 4, 10-719 Olsztyn, Poland
[c] Laboratory of Structural Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

ARTICLE INFO

Article history:
Received 13 December 2015
Received in revised form 21 April 2016
Accepted 22 April 2016
Available online 25 April 2016

Keywords:
RNA
Structure
Modeling
Bioinformatics
Simulation
Molecular dynamics
Monte Carlo dynamics

ABSTRACT

Functional RNA molecules depend on three-dimensional (3D) structures to carry out their tasks within the cell. Understanding how these molecules interact to carry out their biological roles requires a detailed knowledge of RNA 3D structure and dynamics as well as thermodynamics, which strongly governs the folding of RNA and RNA-RNA interactions as well as a host of other interactions within the cellular environment. Experimental determination of these properties is difficult, and various computational methods have been developed to model the folding of RNA 3D structures and their interactions with other molecules. However, computational methods also have their limitations, especially when the biological effects demand computation of the dynamics beyond a few hundred nanoseconds. For the researcher confronted with such challenges, a more amenable approach is to resort to coarse-grained modeling to reduce the number of data points and computational demand to a more tractable size, while sacrificing as little critical information as possible. This review presents an introduction to the topic of coarse-grained modeling of RNA 3D structures and dynamics, covering both high- and low-resolution strategies. We discuss how physics-based approaches compare with knowledge based methods that rely on databases of information. In the course of this review, we discuss important aspects in the reasoning process behind building different models and the goals and pitfalls that can result.

Contents

[*] Corresponding authors at: Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland (W.K. Dawson and J.M. Bujnicki).

E-mail addresses: wdawson@genesilico.pl (W.K. Dawson), iamb@genesilico.pl (J.M. Bujnicki).

http://dx.doi.org/10.1016/j.ymeth.2016.04.026
1046-2023/© 2016 The Authors. Published by Elsevier Inc.

1. Introduction

Ribonucleic acid (RNA) is a biological polymer, capable of performing a wide range of functions in the cell. In addition to messenger RNAs, whose primary function is transmission of genetic information from DNA to proteins, numerous other classes of RNA molecules have been found that are involved in a variety of functions such as catalyzing biochemical reactions or performing regulatory roles (reviewed comprehensively in [1]). Most of these functions depend on the three-dimensional (3D) structure, dynamics and thermodynamic properties of the RNA chain. For example, regulatory elements located within mRNA that switch protein production on and off, known as ‘riboswitches’, function owing to their ability to form alternative structures or to undergo transformations between a structured and unstructured state, depending on binding of specific ligands or on sensing environmental changes (reviews: [2,3]). Thus, understanding RNA function beyond protein coding requires a detailed knowledge of RNA 3D structure and dynamics as well as thermodynamics, which strongly governs the folding of RNA and RNA-RNA interactions as well as other interactions within the environment.

Experimental determination of 3D structures of RNA molecules at high resolution is largely owing to the work of X-ray crystallography, and NMR spectroscopy, but it is very laborious and requires substantial expertise and resources (review: [4]). Consequently, the majority of known RNAs (e.g., represented by sequences in the RNAcentral database [5]) remain structurally uncharacterized. To help cull the number of expensive and time consuming experiments that are needed to determine RNA 3D structures, computational methods have been developed to predict these structures based solely on the information about RNA sequences, or by taking into account additional biochemical information [6–8]. These computational methods for RNA structure prediction are advancing rapidly and a variety of new approaches exist. However, to obtain meaningful results from computational methods, it is important to understand the strengths and weaknesses of a given approach.

Over the course of the last 40 years, computational strategies for modeling RNA 3D structure and dynamics have been developed, to some extent based on earlier developments in methodology for modeling protein structure and dynamics [9]. They are centered around two major pillars: all-atom (AA) simulations and coarse-grained approaches [10].

AA methods aim at capturing the essential physics of molecules by modeling the detailed interactions of atoms. One of the most frequently used AA approaches employs Molecular Dynamics (MD), which simulates the time dependent motion of a molecule (or molecules). AA-MD simulations with explicit solvent molecules and ions can be highly informative when the phenomena of interest happen over a very short time window (or time range) extending from a few tens of nanoseconds (ns) to a maximum of a few microseconds (μs). Unfortunately, many important processes that involve biological molecules happen over time scales on the order of milliseconds (ms) [11,12], seconds [13], or even minutes [14]. Using specially designed microprocessors, the longest AA-MD simulation to date extends over 1 ms [15] and it shows sufficient statistical accuracy on a time range up to μs. However, in general, such state of the art tools (both hardware and software) are not easily accessible to most researchers, who must be content with 10–100 ns simulations available with standard tools.

When the biological effects demand computation of the dynamics beyond a few hundred nanoseconds, not only does computational power become an increasingly formidable issue, the vast quantities of data that are generated in such simulations become difficult to process in the usual way of opening a file and evaluating the data in linear fashion [16]. If this were not enough, there is no assurance that such expensive simulations will yield sufficiently accurate results due to the plethora of approximations introduced even at the level of AA-MD. For the researcher confronted with such challenges, a more amenable approach is to resort to coarse-grained (CG) modeling to reduce the number of data points and computational demand to a more tractable size, while sacrificing as little critical information as possible. CG methods aim to model the rough features of (macro)molecular systems by decreasing the resolution from individual atoms to some larger group of atoms while maintaining the essential features of macromolecules responsible for their physico-chemical properties that underlie the biological functions. CG methods can help reduce the heavy (or practically impossible) computation and memory cost of AA simulations, in particular for large molecules [17–20]. Coarse-grained methods were developed either with knowledge-based principles (KBP) or theory-based principles (TBP). Both approaches simplify the complexity of the residue as will be explained in the sections that follow.

CG methods are currently the first line of attack in *de novo* simulations of macromolecular (protein or nucleic acid) folding that use information about sequence alone, because they significantly reduce the number of interaction centers and consequently the number of very time-consuming energy-force evaluations. Rather than generating mountainous quantities of data from months of computations on supercomputers with an uncertain outcome, the simplified 3D conformational space can be sampled on a typical personal computer within a few hours or days and produce a modest file size (GBs). However, washing out the details of the data can come at a very high price of losing critical information. A balance must be struck between a rough picture of the interactions and the details necessary to model critical events. This is also a well-known issue with CG modeling of proteins [20].

The aim of this review is to introduce the concepts and reasoning behind coarse-grained modeling in RNA 3D structure and to point out some of the strengths and potential pitfalls. A variety of excellent reviews on CG methods for proteins are also available [17,21–23]. However, there are few reviews devoted to the application of CG methods to RNA, hence this work is primarily directed to that topic. For select cases where the concepts are generic to all nucleic acids or to all polymers, we will occasionally refer to work associated with DNA and proteins in the interest of *showing the universality or transferability of certain important common principles*. Nevertheless, all references to DNA and proteins are in no way intended to be interpreted as complete or thorough in scope.

We have organized this review as follows. In Section 2, we introduce the terminology and most of the abbreviations used in the remainder of the work. Section 3 discusses all atom methods and Section 4 discusses coarse-grained methods. We have organized Sections 3 and 4 to first explain the core structure unit (the bead), then how the energy function (force field) is defined, how the energy function is explored (sampling), and finally some comments about the strengths and weaknesses (challenges). Because the force field, sampling, and challenges for KBP and TBP are fundamentally different, Section 4 discusses TBP and KBP separately. Section 5 discusses the much broader issues of sampling and Section 6 wraps up the discussion with quality assessment measures of RNA and what methods should be used.

2. Terminology in simulation methods

We begin by introducing terminology common to all-atom (AA) methods and coarse-grained (CG) methods. The details and examples will be discussed in subsequent sections. The general picture of how coarse-grained methods fit within the general scheme of molecular modeling is shown schematically in Fig. 1.

What does coarse-grained mean? At the scale of atoms, we can usually ignore the effects of atomic particles like quarks, gluons, muons, neutrinos, etc. We are looking at the wave function for electrons and atomic nuclei. The atomic nucleus is approximated by a charged point mass and its other properties (e.g., the spin of the nucleus) are neglected. Still, there is coupling between the spin of a proton and the spin of an electron; e.g., nuclear magnetic resonance (NMR) or electron paramagnetic resonance (EPR). Hence, even nuclear physics is not necessarily ignorable. If we look at residues, we lose focus on the atoms. Still, atoms or chemical functional groups may turn out to be of critical issue. If we look at RNA, we largely ignore the details of the individual residues and, if we look at the cell, then we largely ignore the details of proteins, RNA and other molecules. If we look at an organism, then even the details of the cell don't matter. At every level, we ignore certain details to gain other important information in a manageable way.

The art of coarse graining can be described by the following scheme:

- Pick the chemical/biological object and phenomenon of interest (RNA folding, DNA supercoiling, proton transfer).
- Adapt the level of coarse graining (establish the "grain" or "bead" size), which should cover the phenomena of interest (e.g., interactions between ribonucleotide residues [24], behavior of DNA or RNA as an elastic rod [25–29], quantum dynamics [30–32]).
- Design the potential energy function, which describes interactions between the coarse-grained objects. This is the critical and most difficult step, as the potential energy function should be as simple as possible to reduce computational cost, but not too simple because, in such a case, it is likely to produce erroneous results.

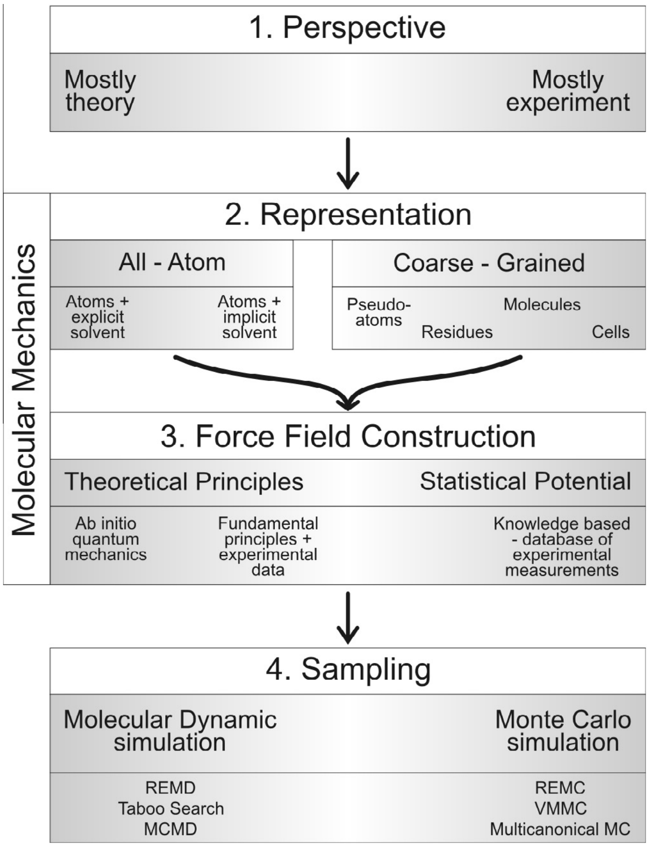


Fig. 1. Graphical description of the relationship between coarse-grained and all-atom simulation methods. Each level/stage of building these models offers a range of approaches which can be combined in many ways. Still, the methods from the left and right sides tend to be used together.

What is a bead? In RNA coarse-graining, a bead can be an atom, a chemical functional group (e.g., purines, pyrimidines, sugars, etc.), or a whole residue (e.g., a nucleoside, an amino acid, etc.) (Fig. 1, box 1). For larger molecules, coarse-graining can extend beyond individual molecules and a bead can represent a local motif or an element of RNA secondary structure (e.g., a helix), an independently folded domain, or a whole molecule. Thus, the grain size ranges typically between roughly one Ångstrom (Å) and one or few nanometers (nm).

Both all-atom (AA) methods and coarse-grained (CG) methods fall under the class of methods known as molecular mechanics (Fig. 1, boxes 2 and 3). Molecular mechanics (MM) emphasizes the actual art of modeling the potential energy surface of a molecule through various interactions within the molecule and between the molecule and its environment: on the atomic level this involves steric interactions, vibrational and rotational forms of energy transfer, the stretching, bending or torsion of chemical bonds, interactions with solvent and ions, etc. The means of modeling of atomic interactions include semi-empirical, pseudo-potential and quantum chemical approaches and combinations thereof. The result is a potential energy function – often referred to as a force field (ff), though a potential energy function is actually more like what it is. MM is probably most often encountered in the parlance of all atom simulations. Some common examples of AA force fields (AAff) are AMBER [33,34], CHARMM [35,36], GROMOS [37], MMFF [38], and OPLS [39,40]. Historically, some of these all atom molecular mechanics (AAMM) software packages have also been known by the name of their molecular mechanics force field; e.g., AMBER, CHARMM and GROMACS. The potential energy function for CG method is also called a force field (CGff), and such ffs must be derived from some description of the mechanical properties of the respective coarse-grained beads. Therefore, in these discussions, we will specify all atom and coarse-grained molecular mechanics as AAMM and CGMM, respectively, and likewise for their respective force fields; AAff and CGff, respectively.

Even at the level of AAMM, the interactions with the solvent or the buffer in the environment can be modeled explicitly or implicitly. When modeled explicitly, all the atoms of the solvent molecules and any buffering charged ions are modeled in the same way as the molecule of interest. When modeled implicitly, the interactions from solvent and the ions are introduced as some averaged interaction. As will be shown in the Sections that follow, this incorporation can be both a blessing and a curse.

A force field can be derived by analyzing known structures using KBP or TBP. The TBP involve employing a variety of equations to describe all the different types of physical interactions. The knowledge-based potentials tend to work the other way around, starting with the answer and relying on the accumulation of statistics to obtain the frequency of a particular interaction or configuration. Most AAMM methods tend to be TBP, though importing some experimental information is basically mixing in some KBP. For CGMM, the TBP involve simplifying the complexity of the residue (e.g., by grouping related atoms together to form a bead and employing implicit solvent) while still retaining enough of the essential physics to generate similar results as the AAMM. On the other hand, KBP in CGMM do not require complex equations describing the interaction of different physical entities, but require gathering massive volumes of information from databases to accumulate sufficient statistics to estimate the probability of a particular configuration. From these probabilities, the approach permits prediction of the energy of such a configuration, where this energy is often called the potential of mean force (PMF) in CGMM strategies [17,41].

All-atom methods permit the most detail and the force fields can often be constructed by obtaining the interaction parameters from QM and/or experimental data. However, even though these methods are much closer to QM in character, all-atom molecular mechanics (AAMM) still involves broad approximations that can be problematical when generalized to all types of chemical bonds and chemical functional groups. Therefore, MM is still an art even at the AAMM level. Force fields are even more difficult to derive for CGMM, because they require more convoluted connections to the original AAMM and still more so the QM [18].

At the bottom of Fig. 1 (box 4) are two major categories of sampling: molecular dynamics (MD) and Monte Carlo (MC) simulation techniques. Sampling describes the way to explore the shape of the multidimensional potential energy surface (defined by the MM force-field), which is complex even for relatively short polymers. This is also clear from experimental evidence [42]. To understand the complex shape of the potential energy surface (force-field) and the corresponding probability distributions in the conformational space of RNA molecules, some form of MD or MC is used by both AA and CG representations.

An MD simulation yields a tangible impression of the time-dependent motion of a molecule in real time; sampling the time dependent conformational characteristics of a given molecule in the environment of surrounding molecules – all of which are represented by the MM force field. The movement of individual atoms is described by their velocities and it is driven by forces which are derivatives of potential energy function. In principle MD trajectories are deterministic; i.e., they depend only on the initial configuration and velocities of atoms and random numbers are not involved in the process of generation of the trajectory. The primary advantage behind MD simulation is that one can measure the real-time dynamics of a molecule: its vibrational modes, diffusion and thermodynamic properties. To simulate how the conformations of a molecule change with respect to time, MD provides the clearest picture. However, since the process is stochastic (it involves successive random variables that form a probability distribution); numerous simulations are required to gain a clear picture of the true time-related dynamics of the molecule under study. MD simulations using CG methods have been attempted on a number of simulated pulling experiments of RNA and DNA [43,44] where an external force is applied to the folded structure to pull it apart. It is also the general approach used in the program Kinefold [45–47], where the results largely agree with the pulling experiments. CG techniques have also been used in DNA and RNA folding [32,48–53].

MC is often a valuable tool for probing the potential energy surface rather than for tracking individual folding trajectories. MC is a stochastic method in which random numbers are used to generate of a series of configurations of an investigated system. MC is similar to energy minimization methods in the sense that both approaches do not use the velocities of particles to describe their movement. The big difference between them is that MC can overcome energy barriers and search for the global minimum whilst energy minimization cannot and, therefore, the latter is strongly dependent on initial configuration. In many cases, one may be more interested in what configurations the entire system will eventually adopt and their respective energies. MC has *mainly* been employed in CG approaches where the focus of the interest is often in exploring the conformational landscape and finding the lowest-energy rather than the dynamics or the folding pathway. It should be stressed here that both MC and MD methods sample the conformational space according to the Boltzmann distribution; therefore, in principle, both methods should converge to the same observables in a very long (infinite) simulation. MC methods are particularly valuable in the development stage of a force field because they can be used to examine and refine the optimum parameters for a given force field. The practical advantage of MC is also that it avoids evaluating forces, which can be a time-consuming process especially for complicated theory-based force-fields (TBff). Showing that an energy function recapitulates the experimental data is taken to indicate that the basic concept of the interactions is understood. Hence, both MD and MC approaches are essential in coarse-grained study of RNA structure, folding and interactions. From MD and MC sampling methods a variety of derivative strategies have been developed to enhance sampling. Some of these are explained in Section 6.1.

The entire methodological framework therefore strongly depends upon the choices made for the representation of the system under consideration. Whereas the individual choices for representation are often quite flexible (Fig. 1); it is often difficult to recognize good choices that don't eventually come with issues. In the sections that follow, we will look more closely at examples of specific choices in the coarse-grained modeling process.

3. All atom molecular mechanics (AAMM)

There are several all-atom simulation methods that use potential energy functions (force fields), which are derived from approximations of the quantum mechanics and sometimes spectroscopic data. Some examples of MD simulation engines include AMBER [54,55], CHARMM [36,56], GROMACS [57,58], TINKER [59,60] and NAMD [61,62]. As mentioned in the introduction, AA-MD simulations with explicit solvent and ions are an essential part of the repertoire of techniques to explore the ns to μs time windows of a molecule (or group of molecules) in some rough but explicit depiction of the true local environment. However, such calculations are often very expensive in terms of calculation time, computer memory, and especially data storage, which can truly explode, even if all the other issues could be endured. Examples of AAMM methods are listed in Table 1 and Supplementary Table 1.

Table 1
Summary of CG and AA applications and their general purpose and method of sampling. For methods using the CG approach, the number of beads per residue is listed, where applicable. Access to these applications (when available) are listed in Supplementary Table 1.

Name	Representation	Force field construction	Sampling
AMBER	AA	TBP	MD
CHARMM	AA	TBP	MD
TINKER	AA	TBP	MD
NAMD	AA	TBP	MD
QRNAS	AA	TBP	MD
FARNA/FARFAR	AA	TBP/KBP	MC
MC-Fold/MC-Sym	AA	KBP	MC
GROMACS	AA and CG	TBP/KBP	MD
openMD	AA and CG	TBP/KBP	MD
YUP	CG, 1	TBP/KBP	MC
NAST	CG, 1	KBP	MD
Kinefold	CG, 1	TBP/KBP	MD
oxRNA	CG, 2	KBP	VMMC
NARES-2P	CG, 2	TBP/KBP	MD
Vfold	CG, 2/3 lattice model	KBP	Dynamic programming
Discrete Molecular Dynamics	CG, 3	TBP	MD
TIP	CG, 3	TBP	MD (Langevin dynamics)
TOPRNA	CG, 3	TBP	MD
SimRNA	CG, 5	KBP	MC
CG (Ren and coworkers)	CG, 5	TBP	MD
RNAkb	CG, 6	KBP	MD
HiRE-RNA	CG, 6–7	TBP	MD
MARTINI	CG, depends on the base	TBP	MD
RAGTOP	CG, helices, loops and junctions	KBP	3D tree graph
ERNWIN	CG, helices, loops and junctions	KBP	MC
Jost and Everaers	CG, lattice model	TBP/KBP	MC

3.1. Definition of beads for AAMM

In AAMM, the definition of the beads is simply the atoms. The interactions of the system of nuclei immersed in electronic cloud (described by QM) is replaced by interactions of spherical beads with electronic charge (partial charge) located at their centers. The beads have a well-defined mass, charge and radius (often according to the Lennard-Jones potential). The intermolecular and intramolecular interactions, and the time evolution of the system is described by Newtonian mechanics and therefore all quantum effects are neglected.

3.2. Definition of the force field for AAMM

The application of theory-based force-fields (TBffs) for AAMM is relatively intuitive. Molecules consist of atoms that interact with other atoms, both within the molecule itself and with the solvent, ions and other molecules in the environment. For the intramolecular interactions, some of these involve covalent bonds: bonds that naturally vibrate (stretch), bend, and, when connected sequentially, can rotate around a torsion axis. Other interactions may be non-bonding but still attractive (e.g., London forces or Coulomb interactions of oppositely charged atoms) or repulsive (Coulomb interactions of similarly charged atoms). These non-bonding interactions can be both intramolecular and intermolecular depending on the particular molecule.

Because the size of a ribonucleotide is small enough to permit computation using QM, it is used to compute the intra-residue interactions of isolated residues and these data are eventually used to assemble an RNA molecule [55]. QM yields the partial charges of the residue (charges located on each atomic nuclei, which approximately reproduce QM-generated electrostatic potential around the residue) and approximates the potential energy for bond stretching, angle bending and dihedral torsion effects. The non-bonding interactions between atoms can be also estimated using QM. These parameters are often supplemented with information from spectroscopic data. Armed with these intra- and interatomic parameters for each residue, one writes the potential energy of a molecule. The sum of these bonding and non-bonding potential energy terms constitutes the potential energy of the system, which is a function of positions of atomic nuclei. The movement of light particles (electrons) is implicitly included in the potential energy function and therefore quantum mechanics can be replaced by classical (Newtonian) mechanics. The common form of the potential energy function used by MM force fields is given by [54,63]:

$$E_{total} = \sum_{bonds} k_r(r-r_o)^2 + \sum_{angles} k_\theta(\theta-\theta_o)^2 + \sum_{dihedrals} \tfrac{V_n}{2}(1+\cos(n\phi-\gamma)) + \sum_{i<j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}\right) + \sum_{H-bonds(i<j)}\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) \quad (1)$$

where $r - r_o$ represents the magnitude of deformation of a covalent bond from its equilibrium length r_o, k_r is a spring constant, $\theta - \theta_o$ is the magnitude of deformation of a covalent bond angle (formed by three atoms) from equilibrium θ_o with a spring constant k_θ, V_n and $1 + \cos(n\phi - \gamma)$ represent the dihedral angle corrections (formed by four connected atoms), the $A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}$ terms approximate the van der Waals interactions and steric repulsion between atom i and j, $q_i q_j/\varepsilon r_{ij}$ is the Coulomb interaction between atoms i and j with dielectric constant ε (where applicable) and $C_{ij}/r_{ij}^{12} - D_{ij}/r_{ij}^{10}$ is the hydrogen bonding interactions between atoms i and j (where applicable). The main reason for writing this explicitly is to show the types of interactions that are evaluated in an AAMM theory-based model (TBM). There are variations to this form; for example, CHARMM applies the Urey-Bradley potential [64,65] and the evaluation of hydrogen bonding is treated using a combination of van der Waals interactions and Coulomb interactions; nevertheless, this equation has largely remained as is. Although involving many interactions, the mathematical physics is handled as though the problem were one of Newtonian mechanics. This is also largely the origin of the notion of molecular dynamics simulation for many people first exposed to these ideas.

For AAMM, QM has played an important role in developing the force fields. QM is also the method of choice used for computation of partial charges located on atomic centers, which cannot be directly measured by any available experimental method. Evaluation of partial charges is very important part of force-field development, because it is responsible for a correct description of weak inter- and intramolecular interactions. In recent years, progress has been made in applying QM approaches in refining RNA and DNA structures [16,30,66,67]. Moreover, QM provides the primary justification for employing MM forces for solving structures for modeling the stacking interactions [16].

3.3. Sampling and implementation in AAMM

AAMM is heavily dominated by MD simulation approaches such as AMBER, CHARMM, GROMOS and TINKER. AAMD in explicit water (complete with ions) is often viewed as the gold standard of MD. However, some degree of coarse-grained approximation and MC methods are also used, particularly with in CG approaches associated with structure prediction [unpublished work].

Coarse-grained methods in the context of AA-MD generally reduce the computational cost by turning to models with implicit solvent and often with restraints that encourage a particular configuration.

For example, Seetin et al. [68] have constructed a simulated annealing protocol that combines RNA secondary structure information as restraints, implicit solvent and a modified version of the standard AMBER force field, in which the interacting residues use a soft core potential instead of the Leonard-Jones potential. Eq. (1) lists the van der Waals interaction terms that also contain a $1/(r - r_o)^{12}$ repulsive interaction, which strongly repels the atoms when $r < r_o$ (smaller than the ideal bond distance). A soft core potential is less repulsive; e.g., increasing by a linear repulsive term for $r < r_o$. This helps avoid a violent response when there is a clash in the structure; e.g., when the temperature is very high. Calculations are carried out using a generalized Born implicit solvent model [69], and the application of restraints is controlled during the simulation. This method has examples of folding simulations with secondary structure and sometimes tertiary structure restraints that reasonably closely reproduce known examples of RNA 3D crystal structures, without capturing local details. For example the model of the group I intron P4–P6 domain structure (158 residues) folded with restraints on secondary structure and tertiary contacts obtained by MOHCA experiments was obtained at 13.3 Å RMSD to the reference crystal structure. Our lab has also developed a RNA structure prediction refinement program QRNAS (Juliusz Stasiewicz and J.M.B., unpublished) that uses a variant of the AMBER force field to fix various clashes in coarse-grained structures resulting from inaccurate bond lengths, angles, base planarity, hydrogen bonds, etc.

Another way to reduce the computational cost of RNA folding simulations is to focus on sampling of structures using some type of empirical or knowledge-based function. FARNA/FARFAR [70–73] is a MC method that generates structural fragments over a large span of conformational space, scores them using the Rosetta potential that includes both knowledge-based and physics-based terms, and then relies on clustering to identify representative structures from the set. The approach works well when the effective length of the sequence (or combined effective length in double-stranded RNA) is less than 40 nt [74]. Much longer sequences require the generation of a formidable number of decoys. Another all-atom representation is the MC-Fold/MC-Sym pipeline [75,76] that combines base pairing energetics with a Bayesian type of scoring function to generate predicted RNA structures. Sequence lengths of approximately 120 nt can often be handled by such an approach.

One of the main ways to speed up computation in AAMM is to use continuum solvent models such as the Generalized Born (GB) model [77] or the Poisson-Boltzmann (PB) model [78]. Applications of MM-PB/SA to RNA hairpin loops showed that some of the important features of the loops were recapitulated [79], and likewise for helices [80]. The correct recapitulation of the sugar pucker and other effects usually could be obtained. In general, continuum solvent models tend to do best when the residues are apolar. Nevertheless, considerable time is spent in all atom simulation computing water, so using continuum solvent models permit far longer simulations.

3.4. Challenges to building good AAMM models

As noted in the introduction, there are some specially designed microprocessors that can increase the simulation to a time window in the order of 1 ms [15]. These processors are highly specialized and have certain limitations of their own due to being first generation [35].

The main drawback for AAMM methods is the extreme computational demand, voluminous data proliferation and weaknesses in the force field that compound these costs. Consequently, AA-MD methods are generally restricted to refining structures pre-folded with other methods or examining the “near global minimum” dynamics of experimentally measured structures obtained from the Protein Data Bank (PDB).

Nevertheless, it is noted that significant improvements in the stability of the all atom force fields have permitted AA simulations of structures like an 18 base pair nucleic acid duplex lasting up to 46 μs [35]. Likewise, simulations on three way junctions (around 80 nt) are now regularly performed up to 100 ns [26,81]. Through constant developments in force fields, some newer versions of force fields for RNA show far more stability than their earlier counterparts and are beginning to succeed at describing with sufficient statistical accuracy biological processes that occur over a time range up to μs.

AAMD has been rather lucky with π-π stacking. Numerous studies [16,30,66,82,83] have shown that the intrastrand stacking is the primary source of stabilization in nucleic acids and emerges from London forces, which can be modeled using the van der Waals equation. However, the electrostatic and van der Waals terms have remained largely unchanged since the mid-90 s and are not sufficiently robust for structure prediction [30]. The radius of gyration (a measure of the compactness of a protein or RNA molecule) still tends to be too small for large molecules (more than 30 residues) compared to the experimentally observed value, indicating that the thermodynamics is not completely recapitulated and may require further work on the force field [84].

Standard continuum solvent models have serious drawbacks when modeling nucleic acids. This appears to be a common problem for highly charged molecules such as aldehydes, carboxylic acid esters, thioethers, fluorine and bromine containing compounds [85] that largely reflects the multitude of assumptions and approximations that have to go into such models [86] and the fact that the charges on the surface of proteins and RNA contain many charged residues that vary extensively over length scales comparable to a water molecule [87]. Part of this problem appears to be the result of dielectric saturation, an effect where charged ions and molecules in close proximity are not screened uniformly by the clouds of counter ions that surround them; in effect, the charge polarization becomes directed toward the local neighboring opposite charge [88]. A recently developed model to overcome this effect introduced a variable directionally dependent dielectric and uses explicit Mg^{2+} ions in the implicit solvent simulation [88,89]. This mixed continuum model in which a small number of explicit molecules such as water or ions are included in QM/MM simulations has proven instructive; however, such approaches must be down with considerable care [23,90]. At any rate, the mixed use of explicit Mg^{2+} within an implicit environment of Na^{+} ions in a continuum solvent strategy appears not to be the only reason for its success.

In simulating pH, some models have been developed for constant pH in implicit solvent [91–93]. However, implicit solvent approaches are prone to yield structures that are too compact and to underestimate the desolvation of buried charges, which tends to over-stabilize buried charge-charge interactions [94]. As a result, efforts have also aimed at developing an explicit solvent method with constant pH, wherein a hybrid model that explicitly contains both the protonated and unprotonated forms of the residue in question is exchanged via MC sampling of protonation states. This approach has also been developed with a focus on RNA [95,96].

When modeling using explicit solvent, one typically chooses a water box with a shell of 10 Å surrounding the solute, which is far too insufficient to model a buffer environment, let alone considering the pH [31]. AAMM has yet to find a way to model ionization effects. Progress has also been made with modeling monovalent and divalent ions [97]. However, divalent ions like Mg^{2+} form a chemical complex $[Mg(H_2O)_6]^{2+}$; hence, the interaction should differ more than simply the radius when comparing Zn^{2+} (for example). There are a host of additional issues: tautomerization, protonation and its dependence of local environment, electronic polarization [98], indeed even water (because water models are far from perfect), partial charges and their dependence on conformation etc. There has been recent significant progress on modeling tautomerization [94], though it remains to be tested extensively.

Nevertheless, for all the deficiencies of AAMM, it is the only method we have in modeling large systems that even considers hydrogen bonding and the most pliant to modeling new ligands and modified residues. Future progress is expected to address some of these points.

4. Coarse-grained molecular mechanics (CGMM)

To develop a coarse grained interaction potential energy function, the modeling context and the degree of simplification are important issues. One must ultimately decide on a set of beads (and even how to represent them; e.g., spherical, ellipsoidal, etc.) and these beads should somehow approximate the original structure. There is no obvious way to move from all-atom models to a simplified model of beads. There are two central issues. First, how many beads are sufficient to define the problem of interest. Second, whether interactions of beads should be derived using TBP or KBP or a combination of these two approaches? Examples of CG applications with respect to the number of beads are listed in Table 1 and more details can be found in Supplementary Table 1. Examples of different bead strategies used in RNA CG modeling are shown in Fig. 2. These will be discussed in the sections that follow.

A critical step in the CG approach is deciding on the representation of the beads: whether the beads themselves represent physical objects, where the interactions use TBP, or whether the beads simply represent positions or orientations that are assigned statistical weights from a database of information; i.e., KBP. The former we call TBP-like and the latter KBP-like. TBP-like methods attempt to reduce the cost of the simulation while maintaining the essential physics by grouping related atoms together into a single bead, whose behavior is governed by physical forces usually obtained from some particular averaging scheme. KBP-like methods take advantage of already available knowledge to reconstruct (from the known answer) the process by which a folded structures forms. KBP-like do not require the construction of beads with physical properties that resemble the original group of atoms that were coarse-grained, rather KBP attempt to catalog the configurations of the molecules in the hopes this compilation of configurations can be used to reconstruct the observed molecule: although there are beads, they tend to serve more the role of *references*. Therefore, the goals of TBP and KBP are somewhat different.

It is worth emphasizing that many CG approaches mix aspects of both KBP and TBP to varying degrees. For example, NARES-2P – a minimal TBM capable of folding double helices – uses KBP for the bonding part of the potential energy function [53] and dipolar-bead TBM of the nucleic acids [52] partially employ KBP with knowledge-based equilibrium distances and angles, but theory-based force constants. A KBP-like approach can generally use statistical weights, but SimRNA [99] adds a generic spherically symmetric function to compute the C4′–C4′ distance to describe the mutual attraction of two chains of an RNA duplex. Likewise, TBP-like approaches often import considerable KBP information to help tune melting temperatures [100] and indeed, the TBP-like approaches often tune their models in terms of known temperature and buffer concentration.

It is important to remember that QM effects do not simply disappear just because we shift to coarse-grained methods. In some cases, QM effects are negligible; however, even interactions such as hydrogen bonds can become far more obscure and even vexing compared to standard all-atom MD methods. Moreover, in crystal structures, some residues are found that form direct bonds with $[Mg(H_2O)_6]^{2+}$, where one of the water molecules is exchanged for an O or N atom on the residue [101]. Prediction program using CG may be completely agnostic about such phenomena. Therefore, one should always be mindful that QM *lurks in the shadows* and the simplifications introduced in all atom MD/MC methods may become even more problematical when extended further with the CG process. In such cases, it may be better to circumvent this by applying multiscale approaches where one finds a CG solution first and gradually moves to more fine-grained solutions [23,102].

Each approach has strong points and weak points, but all such methods are typically time-consuming regardless of which approach is used. At some level, all of them use some form of KBP [50]. In this section, we discuss various methods and show examples.

4.1. Determining the number of beads

The beads can represent parts of a residue, one residue or even several residues. The number of beads, their arrangement and, to some extent, even the force field description is largely at the discretion of the analyst and there have been a number of strategies that have been employed. For example, one way is to blur out the representation of chemical functional groups, such as reducing a methyl group ($-CH_3$) to a single bead. This is commonly applied in models like MARTINI [103,104] and, for example, protein-protein docking programs like Zdock [105–107].

The decision about how many beads to use and which ones will depend in part on what one is trying to ascertain or show. If one only wants to see a rough progression of the folding process, it may be sufficient to use a one bead per residue model. If the interest is to model the interactions of residues with each other, then increasingly more beads may be needed. The interaction between the beads is not easy to model, because the beads can no longer be broken down into force field parameters that are easily measurable by quantum mechanics (QM) methods or experimental techniques in general. Nevertheless, a well-chosen set of beads can be quite powerful when some judicious chemical intuition and well-designed optimization procedure is applied; e.g., Ref. [108] used a dipole modeling approach to reproduce the QM generated electrostatic potential of the bases. Usually interactions between beads must be inferred by integrating the interactions between beads together. Thus, the beads must account for the atoms that have been neglected in the CG process.

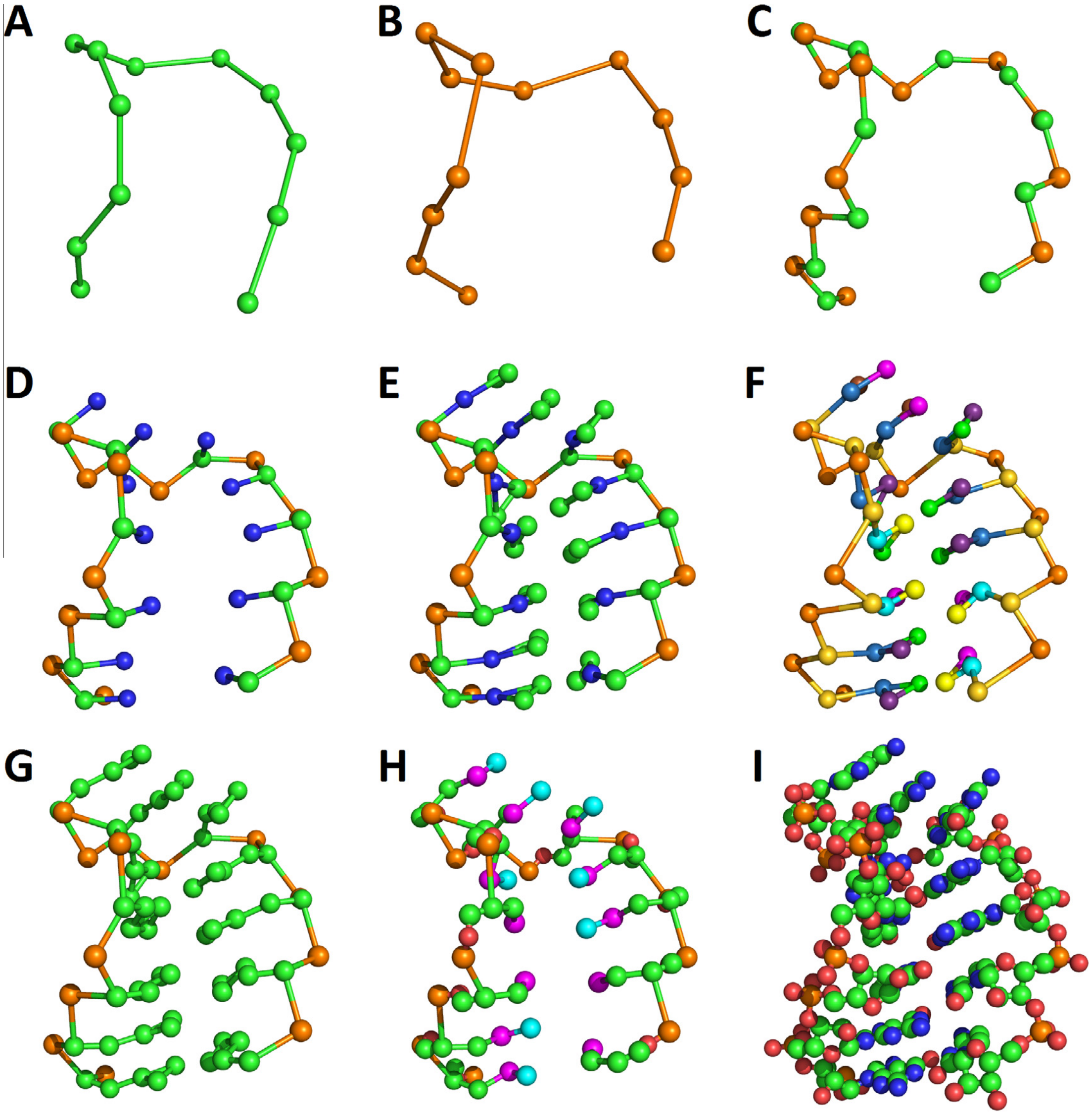


Fig. 2. Examples of real coarse grained models using different numbers of beads and different definitions for the types of beads (PDB id: 2f87). (A) A one bead model using C3′ as used in NAST [24]. (B) A one bead model using P atoms as used in YAMMP [130]. (C) A two bead model with P and C4′ as used in Vfold [121]. (D) A three bead model with P and C4′ and N1 (pyrimidine) or N9 (purine) as used in Vfold [144]. (E) A five bead model as used in SimRNA: P, C4′, and N1, C2, C4 (pyrimidine) or N9, C2, C6 (purine) [99]. (F) A five bead model as used in CG: P, sugar, and different beads for the A, C, G and U bases [137]. (G) A six bead model as used in RNAkb: P, C4′, C1′, C2, C4, C6 [153]. (H) A six or seven bead model used in HiRE-RNA [194]. (I) All the heavy atoms in the structure 2f87.

4.1.1. Representation of the backbone

Since nucleic acid molecules are built from a phosphate, a sugar, and a base, one logical way to proceed is to break down the structure into pieces corresponding to these building blocks. An effective way to build a coarse grained model for an RNA chain is to treat it similarly as it has been done with proteins, namely to consider the sugar and phosphate as the backbone and the bases as a side chains [17,21]. For example, the backbone can be largely approximated by two beads associated with pseudo-bonds between P and C4′ in each residue. Olson and Flory [109–112] showed that the structure of the ribofuranosyl ring and the phosphate (P) bond render a convenient division of the backbone into a virtual bond from P to the sugar (at C4′) and a second virtual bond from C4′ to the next P, Fig. 3. The backbone of RNA and DNA contains six dihedral angles [113,114]: from the 5′ phosphate (P), α (P–O5′, axis), β (O5′–C5′), γ (C5′–C4′), δ (C4′–C3′), ε (C3′–O3′), and ζ (O3′–P). Further, there are the dihedral angles on the ring of the ribofuranosyl sugar ν_3 (C4′–C3′: same as δ) ν_2 (C3′–C2′) and ν_1 (C2′–C1′) that uniquely determine the sugar pucker: 3′-endo or 2′-endo [115], Fig. 3. This suggests an eight-dimensional problem [116]. However, the early work by Olson combined with the study from Pyle and coworkers identified that the vast majority of these conformations could actually be encompassed within the two virtual bonds C4′–P and P–C4′, Fig. 3 (bottom). Although the 3′-*endo* and 2′-*endo* pucker are not *directly* accounted for in this coarse-graining, the virtual bonds formed by C4′–P and P–C4′ are not affected by these differences, and the 3′-*endo* and 2′-*endo* pucker (and even the very rare *exo* cases) can be inferred from the geometry of the interacting bases [116,117]. Furthermore, they introduced the two torsion angles in the notation of the virtual bonds: η (C4′–$\underline{\text{P–C4′}}$–P) and θ (P–$\underline{\text{C4′–P}}$–C4′) [118], where the underlined region indications the bond axis where rotation occurs, Fig. 3 (bottom). Chen and coworkers observed that these dihedral angles largely fall into three general conformation clusters that could be approximated with a diamond (i.e., tetrahedral) lattice [119], at least for short loops where a lattice model is a valid approximation of the degeneracy [120]. Therefore, although in principle an eight-dimensional problem, it can be largely reduced

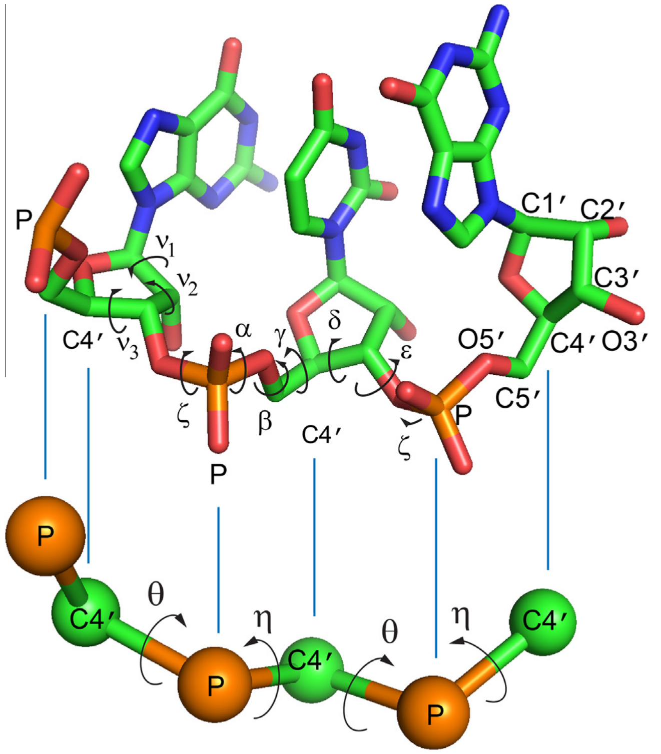


Fig. 3. Definition of angles and bonds along the backbone of RNA. Top: All atom representation (minus the H atoms) showing a short ribonucleotide sequence GUG with the names of the relevant atoms and torsion angles. Bottom: The same chain showing only the two bead per residue coarse-grained representation (composed of P–C4′ and C4′–P virtual bonds) of the same sequence and the pseudo-torsion angles η and θ.

to these two virtual bonds [121]. This approach has been implemented in some approaches that will be discussed later.

Reduction of eight bonds to two pseudobonds that connect two beads is not the only way to cover the RNA backbone. Olson had also shown the possibility of using a single virtual bond on the backbone that connects the phosphate atoms [122]; though this approach required different virtual bonds for 3′-endo and 2′-endo sugar pucker (the two most common conformations). It is also possible to increase the number of beads. Another minimalist way is to use very few beads, but employ very complex potentials [21] to describe the composite interactions in a rigid body nucleotide for both the backbone and the base, as implemented in oxRNA [50] and in models from Liwo, Scheraga and coworkers; e.g., Ref. [53].

It is also possible to consider beads or blocks that represent much larger structures than individual ribonucleotide residues. Examples of this approach occur early on with programs like YAMMP [123,124] that could reduce the whole dsRNA helix to a single bead, and more recent methods like RAGTOP [125] and ERNWIN [126] that reduce the RNA structure to helices, loops and junctions and try to arrange them based on the statistics of loops and junctions. The model of Jost and Everaers [127] uses a lattice model where the stems are treated as a unit and the conformation space of the loop regions is explored. This allows for efficient sampling of all the possible conformations; however, the lattice model also limits any precise transformation between the CG representation and the actual 3D structure.

4.1.2. Representation of the base

If the decision is to divide the structure into monomers and work with beads within these monomers, then the next issue is how to show the base that is attached to the backbone. In general, although nucleic acids have far more regularity in the side chains compared to proteins, the backbone configurations alone are not sufficient to describe important physical properties of the bases. However, the P–C4′ model is uniquely suited to this purpose [119]. The details of these strategies will be discussed in the philosophy sections where many of the strengths and weakness of the approaches can be compared.

4.2. Philosophy behind TBP-like CG approaches

Here, we discuss the methodology, some examples, and some challenges to the TBP-like approach.

4.2.1. The force fields of TBP-like CG methods

Since TBP-like CG methods derive their concepts mostly from the all-atom MD approaches, the concept of a force field and its application to beads instead of individual atoms has considerable appeal. For one thing, biomolecules typically contain a plethora of hydrogens. Why not just ignore them if we can make life easier? The difference is that the beads now represent the collective motion and the interaction of a cluster of atoms is built from the chemical intuition of the person developing the CG representation. There is no obvious way to select or construct the beads and no recipe to derive or estimate the interaction potential as for AAMM [17,18].

For TBP-like CG methods, one seeks to re-parameterize the force field parameters in MD simulations to approximate the AA-MD simulations. In such a model, the binding interactions are approximated from all atom simulation and possibly other auxiliary experimental information, when available [17,18]. These resulting force fields typically have the same form as Eq. (1); however, the interaction potentials represent an integration of the forces over the groups of atoms being approximated by individual beads. For example, when this involves treating a methyl group as a bead, it is relatively easy to understand that the H atoms are "smeared out". However, when a whole residue is treated as a bead, it is not so obvious what to neglect in the pseudo-atoms that one creates.

The core advantage of TBP-CG approaches over TBP-AA is the large reduction of conformational space available for the molecule. There are several consequences of such an operation. First, the number of interactions is reduced, which may speed up the computations. Second, the potential energy function is usually smoothed and, therefore, with the absence of many local minima, the conformation of the molecule can evolve faster. Third, the high frequency vibrations (especially these of protons) are removed and a longer time-step can be used in MD simulation; i.e., a longer real time evolution of the system can be achieved with the same computational effort. However, switching from TBP-AA to TBP-CG usually leads to a more complicated analytical potential. For example, in NARES-2P [53], the spherically symmetric Lennard-Jones potential is replaced by the Gay-Berne ellipsoid of revolution potential and charges are replaced with electric dipoles, which are also used in the dipolar-bead model [52]. Also replacing many covalent bonds with one virtual bond usually leads to larger deviations from harmonicity of the bonding terms of potential energy function.

Nevertheless, in principle, low energy vibrational modes of RNA could be used to model the vibrational modes of the beads and like AA approaches and these parameters could be obtained from IR spectroscopy. However, such information is usually quite difficult to extract from spectroscopic data and rarely all that helpful [18]. Using TBP-CG methods, one is able to simulate the time-dependent dynamics of RNA (only with less resolution) and the simplifications (such as the number of beads and bonds) can be adjusted to the scale of the interaction of interest. Thermodynamic potentials result from the collective sampling of conformations over a sufficiently large time window (amenable to biological time scales), as discussed in Supplement S2.

4.2.2. Sampling and implementation of TBP-like CG methods

In general, TBP with CG potential energy functions tend to be predominantly aimed at MD. However, several structure prediction schemes employ MC techniques. The simplest version of TBP-CG methods uses only one bead per residue. There is a long history of single bead models for DNA and RNA. The earliest 3D modeling programs required secondary structure and tertiary contacts to produce a folded structure [123,128,129], where both these early approaches included building a structure of stems, each of which was defined by 5 beads. One of the first methods was developed by Hubbard et al. [128,129]; where the aim was to study tRNA and part of ribosomal RNA [124,128–130]. Harvey's group introduced harmonic restraints from the secondary structure in the distance geometry approach and made an early attempt at applying electron microscopy restraints with these harmonic potentials [124,130]. Even under harmonic restraints, many misfolded structures were observed [124,130]. In the model from Harvey's group, the stems were first built using one bead, followed by expanding the whole stem to 5 beads, followed by one bead per residue, where the phosphate (P) was used as the reference point [123,124,130] (Fig. 4).

The one bead per residue model YAMMP [124] and the interface package YUP [131] are one of the earliest CG methods for RNA 3D structure modeling. The YAMMP method [124] featured a machine readable script file that was quite complex. The upgraded version of the program called YUP was developed to reduce the input/output complexity [131]. Because of the pre-established harmonic restraints, these models also contain a fair degree of KBP-like modeling concepts; nevertheless, they sit more in the spirit of TBP-like CG perspectives in attempting to import experimental information.

Discrete Molecular Dynamics (DMD) [132,133] is a three bead approach consisting of a phosphate (P), a sugar (S) position and a third bead for the base (B) that is tied to the sugar (Fig. 4). This selection of beads resembles the TIP model first proposed by Thirumalai [134], Section 4.3.2. To evaluate the base–base interactions in DMD, the Turner energy rules [135] are refitted by dividing the base-base interaction into base-edge interactions due to hydrogen bonding and base-face interactions which are the primary source of what is known as stacking. Coupling between bases is worked out from solving a set of equations that are correlated with the Turner base stacking interactions [132,136]. Decoupling the base-base interactions significantly helps in a *de novo* folding approach. The stacking effects are derived from quantum mechanics [16,30,66] and are a consequence of the van der Waals interactions (in particular, the attractive term in the van der Waals equation that represents the London forces). The base pairing weight is established based upon the orientation of two bases being such that they satisfy certain distance relationships and these interactions are penalized when the relationships are not satisfied. Thus, stacking, as observed in real molecules, is achieved and the base-base interactions are identified by the configurations of different bases.

On a slightly different course, Ren and coworkers [137,138] have developed a five bead MD simulation approach where the force field can be run on packages like GROMACS [57,58] and NAMD [61,62]. In this model, instead of forming a virtual bond to the center of the base, a virtual bond points from C4′ to N, where N is N9 for purines and N1 for pyrimidines. The base consists of the N plus the two additional beads: in all, three pseudo-atoms that are bound together to form a triangle shaped structure of beads, Fig. 5A. This helps decouple the orientation of the base from the orientation of the backbone. The model resembles the MARTINI coarse-grained pseudo atom methods that has been developed for studying protein interactions in the lipid bilayer [103,104] and other complex systems. This is a promising approach, but as a hybrid MD simulation approach with knowledge based potentials, it depends on obtaining adequate force field parameters that satisfy the demands of difficult long time-range MD simulations when the chemical properties of the individual atoms are smeared out between several atoms. This is an issue that has been noted by the developers of the MARTINI coarse-grained approach [103]. It is also a familiar problem in modeling solvent interactions with salt bridges in proteins [98]. Nevertheless, progress only comes with trying.

The TBP-like CG model for RNA with the largest number of beads is HiRE-RNA, which contains a backbone of four beads (P–C4′–C5′–O5′) one bead for a sugar (C1′) and one bead for a pyrimidine base (called B_1, where the pyrimidine bead B_1 is bound

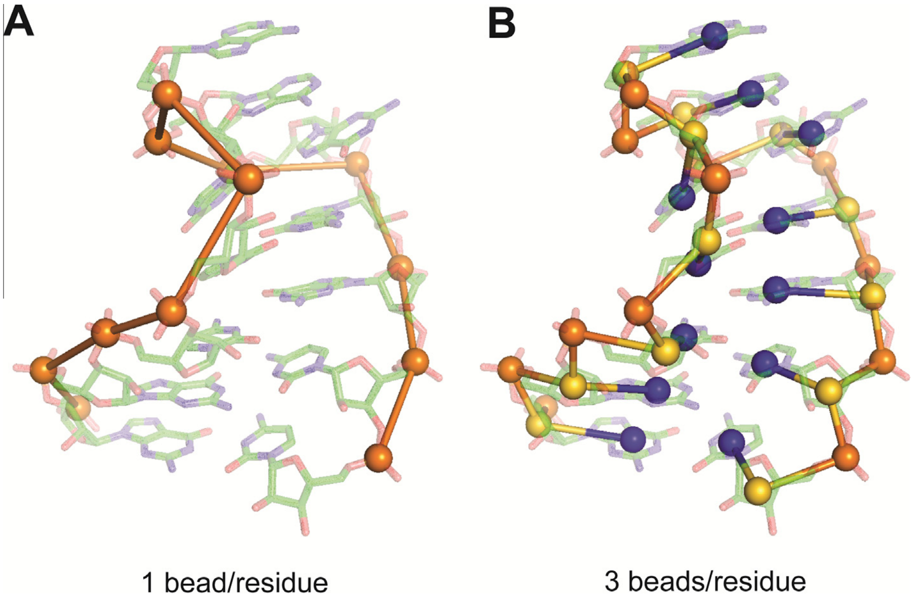


Fig. 4. Examples of theory-based models that use one and three beads. (A) A one bead per residue structure in which the phosphates (P) form the backbone, used in YAMMP [123,124,131]. (B) A three bead per residue PBS model in which the chain is represent as a phosphate (P: same as in panel A) a base (B: the blue colored bead), and a sugar (S: yellow colored bead). Models like this are used by Thirumalai et al. (TIP) [134,148] and by Dokholyan et al. (DMD) [132,190,195].

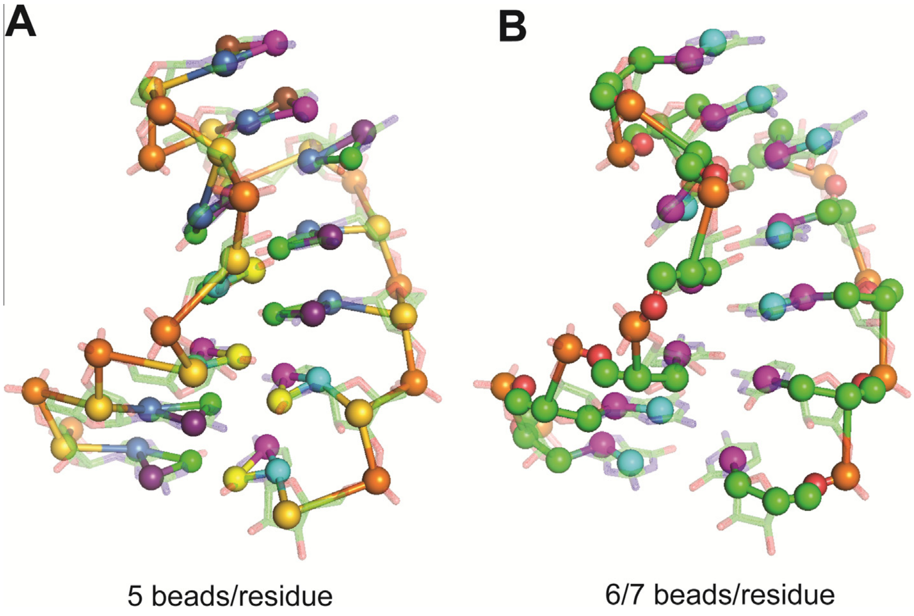


Fig. 5. Examples of theory-based models that use 5–7 bead structures. (A) A 5 bead per residue model with P-S (sugar) for the backbone and three beads for A (CG,CA,N6), C (O2,CU,N6), G (CG,N2,O6) or U (O2,CU,O6) by Ren and coworkers (sometimes called CG) [137,138]. (B) Example of a 6 (purine) or 7 (pyrmidine) bead model with P–C4′–(–C1′–B1)C5′–O5′ (for pyrimidines) or P–C4′–(–C1′–B1–B2)C5′–O5′ (for purines), used in HiRE-RNA [51,194].

to C1′) and two beads for a purine base (B_1-B_2 where the purine bead B_1 is bound to C1′), Fig. 5B. The extra bead on the purine gains a better description of the excluded volume and allows for accounting for some of the differences of purines and pyrimidines.

4.2.3. Challenges to building good TBP-like CG models

The main problems associated with TBP-CG methods are also found with AA methods; only they are even more amplified when CG approaches are used. First, AA-MD methods require very long simulation times to adequately sample the full conformation space [132,139] and this sampling can be seriously hampered by various weaknesses of the force field used [98]. This problem is exacerbated when approximations like CG are applied [103,104]. Unless one introduces complicated potentials, the specificity of the beads is often reduced, and typically issues with polarization are only compounded. As mentioned in Section 3.4, even the rotation of a functional group around a bond changes the partial charges. When carboxylic acid groups or amines interact with each other or with water, the partial charge distributions are also changed - sometimes drastically. In general AAMM, the partial charges are often written in stone before the simulation begins. Though polarizable potentials have been introduced [55,98], it is difficult to derive a universally applicable polarizable potential; hence, they tend to be rather specialized — perhaps this is why they have not been adopted in general. Finally, as pointed out in implicit solvent AAMM, the strategy of these methods to treat ions and water lightly may not always be to their advantage.

Second, since developing an accurate description of the atoms comprising each of the beads is crucial to accurate representation of the essential forces of the system, the selection of a proper set of beads is surely a vexing problem. The assignment of the beads and their interactions with other beads are difficult (if sometimes not impossible) to convey in such a way that they directly correspond to some experimentally measurable quantity [16]. When one tries to integrate a collection of atoms together into a single collective entity, either one must build a complicated potential energy function to describe the new bead (or pseudo-atom), or make very good choices on what sort of beads represent the integrated structure.

Third, it is difficult to carry out a highly systematic approach on the role of beads and what type of forces are essential to retain in a model to recapitulate the characteristics of a real system. Nevertheless, some attempts as determining the number of beads and their positioning have been carried out for DNA [108]. Somewhat revealing in this study, which was directed to the number of beads to represent purine and pyrimidine bases, was that the thymine base required four beads (one extra bead for the methyl structure compared to the uridine base that can be modeled with three beads) and that the guanine required four beads whereas adenine required five beads, one extra to cover the interaction at the thymine base in canonical Watson-Crick base pairing problems.

Therefore, it is wise to be cautious of overinterpretation because the dynamics are likely to be oversimplified and the predictions of thermodynamic parameters may not be sufficiently reliable. Nevertheless, a major aim of MD approaches is to understand the physics, where reproducing the experimental data is taken to be the evidence of having captured this essence of the forces involved.

4.3. Philosophy behind KBP-like CG approaches

Statistical potentials have played a very strong role in attempting to model the process of protein folding [21,140]. To understand KBP, we need to reflect on the concept of potential mean force and the inverse Boltzmann function. The deeper technical matters and derivations are left to the Supplementary information S1 and S2; however, to help motivate the concept and show its basic assumptions, we briefly explain the concept of potential mean force (PMF) and show how this is implemented at a practical level to generate statistical potential models. We then discuss the application of these models.

4.3.1. The force field of KBP-like CG methods

The basic concept behind the potential energy function essentially works backwards from the answer (the observed physical data). Typically, we would use TBP models to generate the thermodynamics (see Supplementary information S2), where the successful reproduction of the observed phenomena is taken to mean that we understand the physics of the system. However, we could suppose things in the opposite way. In principle, a successful TBM could assign an observed configuration between two beads (i and j) to some reference vectors $\{\mathbf{r}\}_{ij}$ with some probability $p_{ij}(\{\mathbf{r}\}_{ij})$. Such vectors can contain a variety of information (e.g., angles, distances) about the mutual orientation of the beads with respect to some reference orientation. The potential mean force ΔE_{ij} is described by the following equation

$$\Delta E_{ij} = -k_B T \ln(p_{ij}(\{\mathbf{r}\}_{ij})/p_{\text{ref}}) \qquad (2)$$

where k_B is the Boltzmann constant, T is the temperature, p_{ref} is the reference probability (essentially some reference state that normalizes the probabilities) and ΔE_{ij} would reflect the observed change in the energy due to the interaction of beads i and j at some reference orientation $\{\mathbf{r}\}_{ij}$. The earliest models simply considered the distance between beads i and j, but far more elaborate strategies can be used. For our purposes of getting the concept, exactly what this $\{\mathbf{r}\}_{ij}$ or p_{ref} etc. are, is not really all that important. Simply imagine that there is some way to get them for the problem at hand.

Assuming we can obtain these data, $p_{ij}(\{\mathbf{r}\}_{ij})$ would be proportional to the number of such observed instances divided by the total number of measured instances p_{ref}. Why then not simply grab this answer and use it to calculate the energy (ΔE_{ij}) from the start? This is the basic idea behind the inverse Boltzmann function. Instead of going through all the labor of obtaining $p_{ij}(\{\mathbf{r}\}_{ij})$ from first principles when we effectively have it already, why not simply *use* $p_{ij}(\{\mathbf{r}\}_{ij})$ to obtain ΔE_{ij}? This ΔE_{ij} is known as the potential mean force (PMF) in the literature [141,142]. In general, the PME describes the relative preference for interactions — usually the distance (and sometimes the orientation) between two pairs. Such a model actually could be extended to three body problems, permitting likely greater accuracy; however, current statistics in the PDB render such considerations impractical. What is of primary importance is sufficient statistics. If only distance is considered, the PMF can be represented as a series of interpolated points. When angle dependent terms are needed, this becomes a 2D array, and when 3D orientations are involved, this becomes a 3D grid. Such information can be difficult to visualize, but to understand that the distance dependent model would involve fitting a curve to a spline, this is the basic picture of the PMF for r_{ij}, and this has been commonly used for proteins with success [21].

In principle, building a statistical potential only requires having sufficient statistics for specified configurations. If a certain configuration is regularly seen when a particular experimental environment is encountered (and particularly if it occurs in many circumstances), then there is a high probability that the same will occur when a new situation of similar context is encountered. If, on the other hand, a certain configuration is rarely seen, it is unlikely to be observed in a new context as well. Hence, if one can just obtain enough statistical data, one need only to define the conditions satisfied by the statistics and the outcome is largely determined by the statistical probabilities.

It is important to mention that, whereas the principle presented above is certainly appealing, there are many issues and pitfalls even for the experienced. Some of these will be mentioned in the sections that follow. For a more complete exposition of these concepts, the reader is referred to the Supplement S1 and the cited references therein. A significant advantage of knowledge-based potentials is that the developer can tune or adjust the potentials to the particular representation of the molecule that is needed for a problem of interest [7]. CG information is generally derived from structural data; typically data obtained from high resolution X-ray structure experiments. Such data can also be obtained from molecular simulations as well.

4.3.2. Sampling and implementation of KBP-like CG methods

Sampling for statistical potentials is more likely to be an MC strategy. This is partly because the statistics for the potential energy functions are still rather rough; hence, it would be difficult to model dynamics with the functions available. Moreover, it is difficult to obtain temperature data so information about the thermodynamics is incomplete. Finally, most of the current use of KBP-like CG methods has been directed to 3D structure prediction, which is better done using MC methods (as opposed to studies on folding pathways and dynamics, which is typically done using TBP-like methods that employ MD).

KBP-like methods already presume the thermodynamic outcome in terms of statistical likelihoods of certain interactions (such as base pairing) from observed examples and have only to decide how to score structures with different base pairing and non-Watson-Crick (nWC) base-base interactions according to the available information. The beads are used to represent unique orientations of a pair of residues for the PMF [141,142] and are therefore basically reference points, where the probabilities are based upon whatever statistics have been obtained and the model of the interactions used. For example, in addition to binding interactions on the beads, the beads may contain some additional excluded volume parameters; however, the potentials are often built on purely statistical weights that are defined by the orientation of the residue and the beads mainly serve as alignment pins in positioning the resulting RNA polymer chain.

A standard one bead per residue model for 3D RNA folding is NAST [24,70], which assumes the C3′ atom as the reference point for the structure and uses knowledge-based potentials, Fig. 6A. NAST requires RNA secondary structure information at the beginning and is helped further if tertiary structure is available. Therefore, as with previously mentioned one bead per residue CG TBP, this single bead approach cannot be used for *de novo* folding. A recent model developed by Carbajal-Tinoco and coworkers [143] uses a single bead and can fold short hairpins successfully without restraints.

The simplest successful *de novo* folding approach involves two beads per nucleotide [144,145]. Both the two and three bead model (Fig. 6B) used in Vfold have proven useful in studying the loop entropy of various hairpins and pseudoknots [146] because all the conformations of a short free strand sequence (up to 9 nucleotides) can be explored on a tetrahedral lattice. Sampling methods such as TOPRNA also work with a three bead model [147]. In the model, stacking is evaluated based on the Turner energy rules [135] and the configuration of the RNA chain in the lattice. Including base stacking interactions in the evaluation of conformations, and comparing the conformations calculated with stacking against those calculated without stacking, it is possible to evaluate the loop entropy of a specified sequence from a short RNA hairpin loop [119].

An off-lattice rigid body model oxRNA uses a structure composed of two beads and multiple interaction centers embodying a rather complicated potential energy function. There are two primary beads: one that approximates the rotation about the backbone and one that locates the nucleic acid base orientation [50]. The method is designed to model the properties of nucleic acids including salt effects and obtains quantitative agreement in the thermodynamics of simple hairpin structures.

Thirumalai and coworkers developed a three interaction point (TIP) model that employs the phosphate, the center of the sugar and the base at a fixed distance from the sugar position [134]. Preceding DMD (Section 4.2.2), this model also divides stacking and hydrogen bonding between bases and these interactions are derived from the Turner rules. Recent developments with TIP include factoring in monovalent salt effects [148]. Similarly, Chen and coworkers have developed the Tightly Bound Ion model to

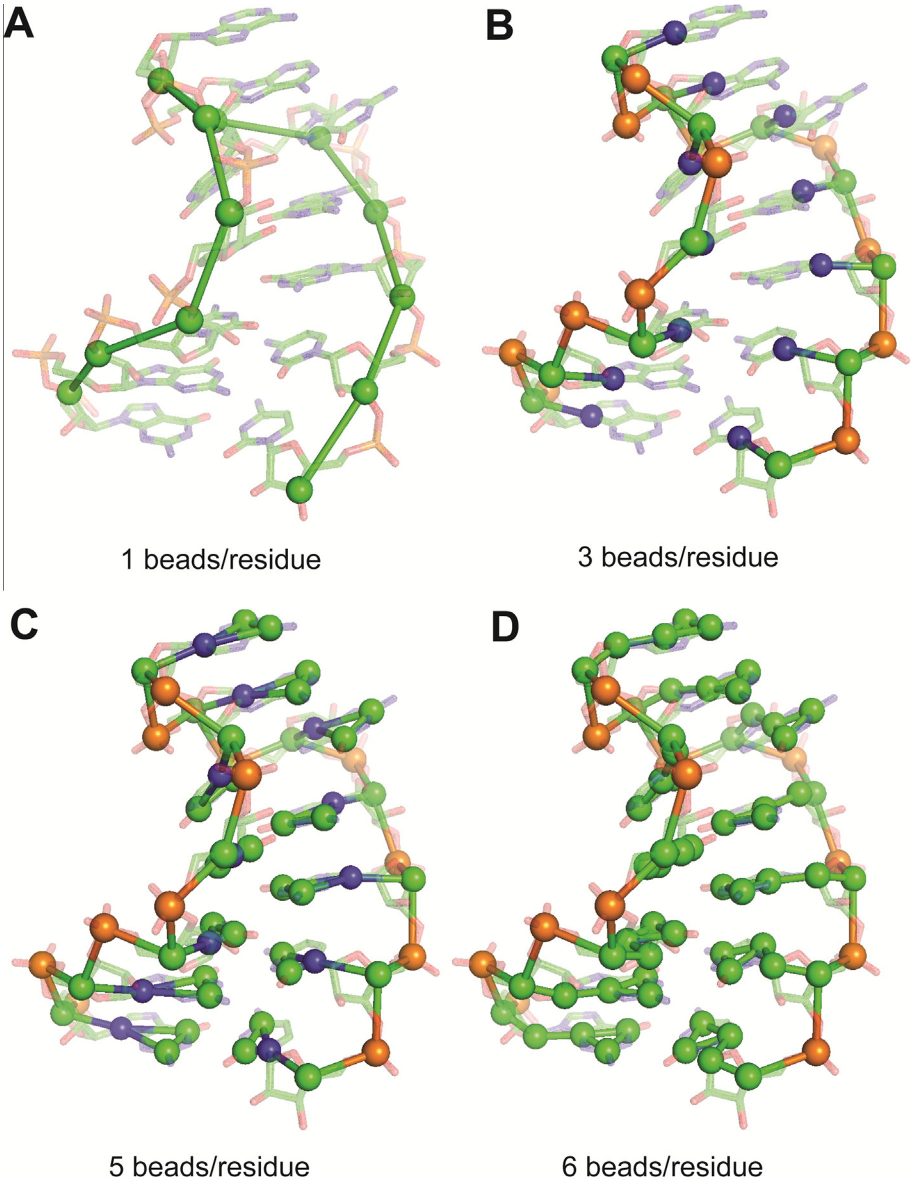


Fig. 6. Examples of knowledge based models with 1, 3, 5 and 6 beads per residue. (A) A one bead per base model of RNA based on the C3′ position, used in NAST [24,196]. (B) A three bead model with P–C4′–N1/N9, used in Vfold [144,146]. (C) A five bead model with P–C4′–(N1, C2, C4: pyrimidine)/(N9, C2, C6: purine), used in SimRNA [99]. (D) A six bead model with P–C4′–C1′(C2, C4, C6), used in RNAkb [153,197].

address Mg^{2+} binding in RNA structures [149–152]; where Mg^{2+} is modeled explicitly in the thermodynamics.

SimRNA [99] utilizes a five bead model where the backbone orientations are described by P and C4′ and the orientation of the bases is defined by 3 beads: N1, C2 and C4 (for the pyrimidine bases) and N9, C2 and C6 (for the purine bases), Fig. 6C. The latter three beads define the plane of the base and therefore clearly define the intersection and orientation of the planes between any pair of bases. Placing the rotational axis of the base at N1/9 is crucial to capturing the orientation of the planes of the base. The relative orientation of two interacting bases is projected onto a grid where the statistics of these configurations are stored. The relative position and orientation of the bases determines the value of the PMF ΔE_{ij} (Eq. (2)). These grids also include such things as interacting base-phosphate and base-sugar orientations. The grids depend on having sufficient statistics to build these potentials. As in the case of Thirumalai's model [134] and DMD [132], stacking and base-base binding are treated as separate interactions and can be adjusted.

Levitt and coworkers introduced a six bead per residue model used with the RNAkb potential: P–C4′ for the backbone and C1′, C2, C4 and C6 for pyrimidines and purines, where the base beads are different for these two base types [153], Fig. 6D. Similar to SimRNA, the model locates a joint at C1′ to help map out the orientation of the plane of the base. However, the focus of the joint emanating from C1′ rather than from the N1/9 position of the base (as done in SimRNA) makes it more difficult to render a one-to-one transformation of the orientation of the base in stacking and base-base interactions, but explicitly identifies the sugar pucker.

4.3.3. Challenges to building good KBP-like CG models

The main issues with statistical potentials are as follows: (1) completeness of the statistical information and (2) the choice of beads demarcating the structure and the nature of their moving parts. Obtaining complete statistical information is problematical because one has difficulty claiming that any set of statistics is sufficiently complete to guarantee that the number of configurations observed is a representative set. The Protein Data Bank (PDB) contains mostly the structures of molecules that have been crystallized or which have been determined by nuclear magnetic resonance, and only a handful of structures determined with other techniques. These molecules with determined structures often have been truncated or otherwise altered to help in the structure determination process. For most of the RNA molecules with experimentally determined sequences, especially those that cannot be crystallized or obtained in high quantity and concentration, structures are not available and are unlikely to be obtained in the near future. Hence, some information is clearly missing. Therefore, a formidable problem in justifying statistical potentials and their result is the fact that the very dataset they depend on is biased toward structures that can be determined using existing high-resolution methods.

Obtaining a good representation of the beads demarcating parts of the structure is also a problem. However, the PMF for models like SimRNA are built from grids of energies that only use the beads to determine the orientation of the ribonucleotide. Inasmuch as the (obtained) statistics satisfy all the conformations and conditions of the RNA molecule, the method promises to be easier to manage than designing complex potentials from TBP. However, because water and ions are almost always implicit in these models, these may conspire to cause overly compact structures and strong interactions between parts that should normally remain in solvent surrounded by a cloud of ions.

Another issue with structure prediction using statistical potentials is finding long range interactions and tertiary structure. Statistical potentials are typically simple pair potentials – whereas they are able to bind residues locally, they typically lack any intrinsic long range tertiary structure information. Recent approaches [154] have taken advantage of the fact that many RNA structures appear to fold hierarchically [155,156] and form recurring local structured motifs, which are predictable from sequence [157]. Hence, with sufficient statistics, it may be possible for the statistical potential to describe three way junctions, four way junctions, various types of loops, etc. of common RNA motifs [154]. This would permit the prediction of the global tertiary structure of RNA to be made based on the assembly of local motifs, junctions, and elements of secondary structure. Within the genre of statistical potentials, there are also approaches that try to determine the arrangement of secondary structure and thus work at a much larger scale of stems and loops. RAGTOP is a graph theory approach that finds the 3D arrangement of the stem-loop structures [125]. ERNWIN uses KBP to carry out a similar procedure of arranging stems [126]. Fragment assembly, at some level, is coarse-grained in the sense that each fragment is treated as a unit and the pieces fitted together.

4.4. An example of using two CG methods

In Fig. 7, we show an example of a flow diagram for using two CG applications: oxRNA (TBP, top) and SimRNA (KBP, bottom) and the approximate flow of carrying out a simulation. On the left, a sequence is entered, the programs translate that sequence into an input structure, the structure goes through a MC simulation, and a final structure is obtained after clustering. The input sequence is obtained from the native structure shown in Fig. 2 (PDB id 2f87). It can be seen that the results of these two methods are qualitatively quite similar for this short sequence and the structures closely resemble those shown in Fig. 2.

For the TBP example, the oxRNA package [50] was designed to directly adhere to physical parameters. Therefore, temperature and salt concentration can be clearly specified; quantitative physical parameters can be used and obtained from this approach. Indeed, the model is even built to evaluate the Kuhn length of RNA and persistence length of dsDNA [158]. The set up specifications for the input file (“input”) are more involved than typical programs. The program is designed to recapitulate the physical properties of RNA with less emphasis on obtaining the “correct” structure (the PDB file reference). As a result, the transformation between the CG and AA representation is not as trivial or one-to-one in character. There are other TBP like DMD that are more oriented to obtaining the native structure.

For the KBP example, the SimRNA program [99] was designed with the intention of ease of input and transformation between the CG reduced representation and the AA structure. The native structure in AA representation is shown on the far right of Fig. 7 and was transformed to reduced representation and back to AA

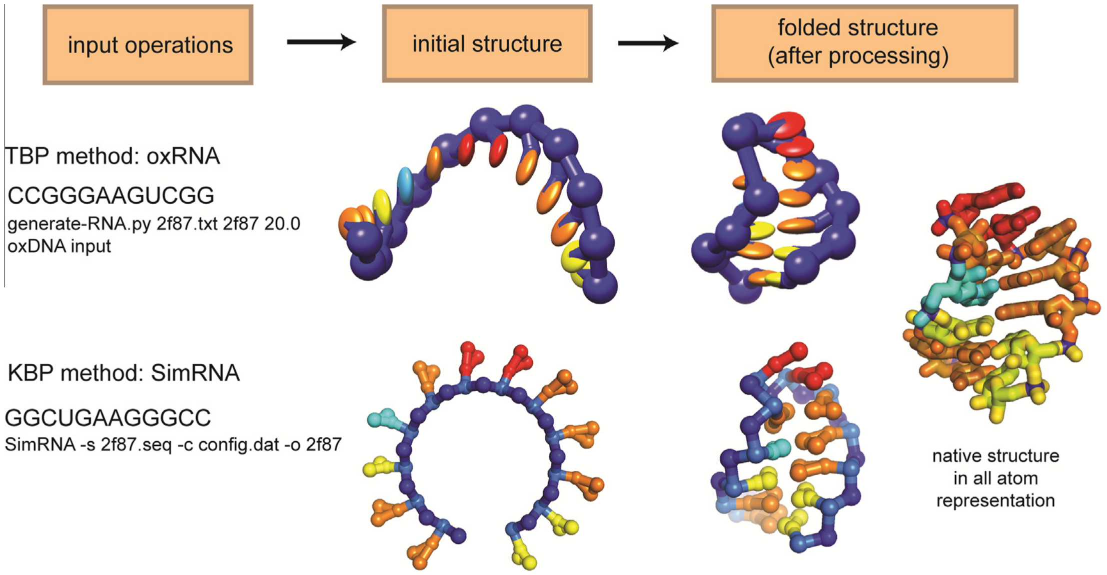


Fig. 7. An example of the approximate flow of coarse-grained simulations for two coarse-grained methods: a theory based method (TBP, top) oxRNA and a knowledge-based method (KBP, bottom). Flowing from left to right, the sequence and command-line files are created and called to initiate the program, an initial structure is created (middle) and a final structure is created at the end of the simulation (right). On the far right is the all-atom native structure. Note that the sequence is entered 5′ to 3′ for SimRNA (more or less standard notation) but 3′ to 5′ for oxRNA. Most of these programs require reading the manual.

with less than 0.5 A RMSD difference, reflecting the one-to-one transformation of coordinates for this short sequence. However, the input value for temperature or the temperature range for simulated annealing is not obvious because the interpretation of T = 1 depends strongly on the particular force field; similarly for other weights. There are other KBP like Vfold that have introduced some quantitative measures for some parameters such as salt interactions and free energy.

These methods and others have been compiled in Table 1 and Supplementary Table 1 for the reader's perusal and consideration.

5. Sampling

Up to this point, the discussion has been focused on comparing and contrasting the aims and features of knowledge based principles and theory based principles in RNA structure and folding. In this Section, we turn briefly to the critical issue of sampling.

The free energy landscape (FEL) generated by 3D RNA structure conformations is typically rife with many local minima. Many of these minima are deep enough that standard optimization algorithms are likely to become trapped. Hence, coverage (or sufficient sampling) of the FEL is an essential part of modeling RNA structure. Both MD and MC sampling methods can easily fall prey to insufficient sampling because they become trapped in a local minimum for the entire simulation without ever escaping – particularly when the sampling is done at only one temperature. Therefore, various strategies have been employed to attempt to expand the coverage of the FEL.

One method to increase sampling is to employ MC with simulated annealing (SA). It has been shown [159–161] that recovering digital images by simulated annealing achieves sufficient coverage if a temperature schedule of the order $T(k) = C/\log(k + 1)$ is followed, where k is the index and C is a constant. However, digital images can be processed independently, whereas a folded polymer is a coupled system which is not as easily separable, though the traveling salesman problem is also NP hard yet can be solved in this way [161].

Some of the earliest methods to speed up and enhance the sampling of the FEL were umbrella sampling and weighted histogram analysis method [162–164], replica exchange Monte Carlo (REMC) and replica exchange Molecular Dynamics (REMD) [165] and metadynamics [166]. The most popular means of enhancing the coverage in 3D RNA structure prediction are REMC and REMD. In such an approach, an initial structure is simulated in a fixed number of temperature shelves and transferred between shelves to help expose it to conformations not normally encountered in a single simulation carried out at ambient temperatures. A somewhat related approach is multiple MC, where different temperatures are used to construct a global 3D polymer lattice distribution and these structures are explored locally using umbrella sampling [167]. Traditionally, REMC and REMD methods have used a linear temperature schedule between the highest and the lowest temperature. A logarithmic temperature schedule appears to yield well overlapping coverage in REMD simulations of proteins [165] and may enhance the quality of sampling coverage closer to ambient temperatures in REMC as well. Further multiplexed REMD (MREMD) has been employed in solving protein folding problems using distributed computing [168], which is not so amenable to standard REMD/REMC.

The program oxRNA [50] employs the virtual move MC (VMMC) algorithm: a cluster-move MC algorithm [169] and an approach where the landscape is explored in a hierarchical fashion [154]. The Rosetta-based fragment assembly MC for RNA (FARNA) approach starts from an extended structure. From there, a random position is chosen in the chain and the torsion angles for three residues are replaced with values taken from a fragment randomly chosen from a database of experimentally determined RNA structures, whereupon the moves are accepted or rejected based upon the Metropolis algorithm [161,170]. Bayesian methods have also been employed in sampling [171].

Recently there have been a variety of promising enhanced sampling methods that have been employed in protein structure prediction that are yet to be used in 3D RNA studies – particularly in MD simulation, though derivatives for MC methods may also be suitable.

One well-established approach is multicanonical MD (McMD) [172–175] and multicanonical MC (McMC) [176,177]. In this approach, the FEL is essentially flattened out so that all configurations have equal probability of forming – rather than being weighted by $\exp(-E/k_BT)$, where E is the energy of a given conformation, k_B is the Boltzmann constant and T is the temperature. The probabilities of the true FEL are later obtained by reweighting the multicanonical distribution to obtain the canonical ensemble.

Another promising approach in MD simulation is Taboo Search [178–182]. Although quaint sounding, it is an easily implemented method wherein independent short simulations are initially carried out to obtain a very rough energy landscape search (at the target temperature). Then, a weighted histogram analysis method is used to find the minimum and maximum energy range. Then, this energy range is then inverted so that the structures close to the best local energy minimum are excluded and structures far from the minimum are selected as "seeds". Finally, the simulation is restarted, mixed with the momenta from other trajectories. If two minima become clustered, they become "taboo" and a new seed is selected. This appears especially promising for rare transitions that are often difficult or perhaps impossible to find with conventional MD simulation approaches [179]; however, it should be remembered that the number of necessary seeds grows exponentially with the number of residues in the protein sequence. Other strategies for finding rare transitions are reviewed in [182].

There are also graph theoretic approaches that in many ways resemble integer programming or dynamic programming. Most of these employ meta-structure approaches of obtaining configurations of RNA secondary structure and converting them to tertiary structure by transformation methods [76,125,126]. However, there have been some approaches in protein structure that work directly from the 3D structure using graph techniques [183–185].

There are also hybrid methods that employ both CGMD to sample the majority of the FEL and carry out refinements with AAMD to obtain more accurate structures known as multiscale methods [186,187]. This procedure can be carried out all the way to quantum mechanics for some problems where there is no possible diffusion of reactants and the boundaries between the QM can be sufficiently isolated from the MM [102]. Such methods are well outside of the scope of this review.

Finally, it is worth commenting on the use of restraints. Restraints tend to reduce the coverage of the conformational space in favor of a particular configuration, dependent strongly on the magnitude of the restraints. For example, weakly applied restraints still permit some level of sampling of alternative states, whereas very strongly applied restraints tend to prevent any movement. In applying methods like REMC or REMD, it is recommended to not apply heavy restraints and even to increase the maximum temperature because even weak restraints tend to inflict a rather heavy bias on the folding and freedom of the structure to change conformation. As a result, folding structures can have difficulty exploring alternative conformations around the desired conformation. Since chains can easily become twisted, this is not a favorable situation. Likewise, in carrying out induced fit simulations, restraints that render the structure too stiff may seriously hamper an induced fit approach.

6. A general assessment of the state of the art

It is well beyond the scope of this work to discuss this topic in detail. However, we will briefly discuss what the strengths and weaknesses of the methods in this section.

6.1. Coarse-grained or all-atom: which is better?

"Which is better" depends largely on what the objectives are. For broadly sampling the FEL, it is more practical for the average researcher to use CG approaches because for AA approaches this is (computationally) a very expensive task. It is also true that uncertainties in the force fields are additive and may contribute even more error when all atom beads are used instead of a few reduced representation beads.

For CG statistical potentials using KBP, the force fields are derived from known structures and therefore carry most of the local thermodynamics with them. However, unless the statistical potentials can morph with the local environment, they are likely to carry too much target specific information. Moreover, the statistics are often derived from the limited set of data available in the PDB which are dominantly for structures that actually crystalize or can be obtained in high concentration necessary for NMR experiments. The question is "what becomes of the statistics if we could combine them with structural data for molecules that don't crystallize?" Therefore, KBP can become a double-edged sword.

MD simulations typically suffer when reduced representation is used, as can be quickly observed when addressing real problems with the MARTINI method (outside of the excellent examples where the method works quite well [103]). On the other hand, if all that is desired is to achieve adequate sampling of the FEL, the reduced representation and less pronounced trapping may be an advantage if one then refines the set of trajectories with conventional AAMD methods.

Finally, as pointed out in Section 3.4, the AAMM methods are hardly a panacea, even if we were granted all the time and computer power in the world. Numerous problems emerge because of the vast simplifications needed to generate the AAMM force fields. For π-π stacking interactions, we are only lucky because the primary effect actually emerges from London forces rather than the often expected quantum mechanics [30]. At any rate, the only methods that can take hydrogen bonding explicitly into consideration (however poorly it may be), are indeed the AAMM-based ones.

It is notable that, for RNA-puzzles, genuine AAMD methods (using explicit water and ions) have not been used so far; though, notably, FARNA/FARFAR (an AAMC method with KBP corrections for ions and water [70,73]) has often come out quite successful [188,189]. First, for AAMD, it is not clear that the base pairing energies in AAff are sufficient to recapitulate the necessary selectivity to be used in these simulations [16]. Second, whatever the method, to find the minimum energy structure and the clusters of structures (right or wrong as the force field decides), we must explore a vast number of conformations. Though CG methods are far from a perfect solution, they are more likely to offer greater coverage at less expense. In this respect, they are of considerable advantage. However, it should be remembered that there are severe time constraints in RNA-puzzles and usually workers are fighting with their own CG methods and refinements until the final minutes. It is possible that a long AAMD simulation in explicit water and ions might help in a final refinement, but we must be assured that CG prediction is reliable first. Therefore, this also reflects some practical realities unrelated to AAMD. It is clear that none of these methods should be blindly used, regardless of how much power one has. At the same time, all of these approaches have merit and should be used for their strengths.

6.2. How do we assess the model accuracy?

RNA molecules are significantly larger than protein molecules composed of the same number of residues, simply because a ribonucleotide residue is much larger than an amino acid residue. The spacing between adjacent ribonucleotides is about 6 Å, while it is about 3.5 Å for proteins, and the diameter of an RNA helix is as much as 23 Å. In general, the accuracy corresponding to RMSD to the 'true' structure of about 10 Å (1 nm) enables visualization of how helices and functional motifs are positioned in three dimensions and is regarded as practically useful. In RNA Puzzles, it has been found that for medium-sized RNAs of 100–300 nt, it is extremely difficult to generate models with RMSD < 10 Å. Since the RMSD depends on the molecule length, a method has been proposed by Dokholyan and his group that estimates what RMSD corresponds to a statistically significant RNA tertiary structure prediction [190].

6.3. When does a model start to become useful?

This is a very tricky question, because it very much depends on the intended use. It is important for the resolution of the model to match (or exceed) the resolution of the question asked. A model such as the one mentioned above (RMSD to the 'true' structure about 10 Å) shows how helices and functional motifs are positioned in three dimensions. It also indicates the mutual position of functionally important residues. However, it would be most likely useless for analyses that require extremely high precision, for instance for docking of small molecule ligands or for quantum-mechanical calculations of chemical reaction mechanisms.

6.4. Is there some way to make a quality assessment?

Unfortunately, it is currently not possible to estimate what is the probability of a given approach (CG, AA, TBP or KBP) to generate a model with a particular accuracy (e.g., with the RMSD less than 10 Å compared to a reference structure). For proteins, results of CASP experiments have shown us that the results of structure prediction vary greatly (for all methods) and it is not possible for any method to provide a simple estimator of model quality a priori, before the model for a given sequence is actually generated [191,192]. Similar results have emerged from the RNA-Puzzles challenges [188,189]. This is because the accuracy of models produced by each method varies significantly from one RNA prediction to another. The accuracy of the models depends on the sequence length and on many issues that are not known before the structure is determined; e.g., the number and type of non-canonical base-pairs, interactions with ions and small-molecule ligands, and the general complexity of the architecture. An assessment of the accuracy of the model can only be attempted using tools such as RASP *after* the structure is generated [193]. However, currently it is not possible to tell (with a high degree of confidence) what is the RMSD of a given model compared to the 'true' structure, without knowledge of that 'true' structure".

7. Summary

We have presented a brief introduction to the topic of coarse-grained modeling covering both all atom approach and lower resolution strategies. We have also discussed how physics based approaches (building from fundamental science) compare with more recent knowledge based methods (using databases of information to solve problems). We have discussed in the course of this review both the strengths and weaknesses of these approaches. It is hoped that the reader will find this experience and advice useful in deciding on what approach is best used for the task of interest.

Acknowledgements

We thank members of the Bujnicki group, in particular Michał Boniecki, Marcin Magnus, and Grzegorz Łach, for stimulating discussions. We also thank the reviewers for their very helpful comments. J.M.B. and M.M. were supported by the European Research Council (ERC, StG Grant RNA+P=123D project number 261351, to J.M.B.). W.K.D. was supported by the European Commission (E.C., REGPOT Grant FishMed contract number 316125, to Jacek Kuźnicki at IIMCB). E.J.J. was supported by the Polish National Science Centre (NCN, Grant 2012/04/A/NZ2/00455 to J.M.B.). The development of bioinformatics servers in the Bujnicki laboratory was funded mainly by the Polish Ministry of Science and Higher Education (MNiSW, Grant POIG.02.03.00-00-003/09 to J.M.B.), and by the statutory funds of the IIMCB. J.M.B. was additionally supported by the "Ideas for Poland" fellowship from the Foundation for Polish Science.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2016.04.026.

References

[1] R.F. Gesteland, The RNA World, third ed., Cold Spring Harbor, 2005.
[2] A. Serganov, E. Nudler, Cell 152 (2013) 17–24.
[3] A. Serganov, D.J. Patel, Curr. Opin. Struct. Biol. 22 (2012) 279–286.
[4] J.A. Doudna, Nat. Struct. Biol. 7 (Suppl) (2000) 954–956.
[5] T.R. Consortium, Nucleic Acids Res. 43 (2015) D123–D129.
[6] M. Magnus, D. Matelska, G. Lach, G. Chojnowski, M.J. Boniecki, E. Purta, W. Dawson, S. Dunin-Horkawicz, J.M. Bujnicki, RNA Biol. 11 (2014) 522–536.
[7] A.Y. Sim, P. Minary, M. Levitt, Curr. Opin. Struct. Biol. 22 (2012) 273–278.
[8] C. Laing, T. Schlick, Curr. Opin. Struct. Biol. 21 (2011) 306–318.
[9] K. Rother, M. Rother, M. Boniecki, T. Puton, J.M. Bujnicki, J. Mol. Model. 17 (2011) 2325–2336.
[10] W.K. Dawson, J.M. Bujnicki, Curr. Opin. Struct. Biol. 37 (2015) 22–28.
[11] L. Green, C.H. Kim, C. Bustamante, I. Tinoco Jr., J. Mol. Biol. 375 (2008) 511–528.
[12] R. Narayanan, Y. Velmurugu, S.V. Kuznetsov, A. Ansari, J. Am. Chem. Soc. 133 (2011) 18767–18774.
[13] G. Chen, J.D. Wen, I. Tinoco Jr., RNA 13 (2007) 2175–2188.
[14] J.C. Schlatterer, J.S. Martin, A. Laederach, M. Brenowitz, PLoS One 9 (2014) e85041.
[15] D.E. Shaw, R. O'Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M.P. Eastwood, D.J. Lerardi, J.L. Klepeis, J.S. Kuskin, R.H. Larson, K. Lindorff-Larsen, P. Maragakis, M.A. Moraes, S. Piana, Y. Shan, B. Towles, in: Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09), ACM, 2009.
[16] J. Sponer, J.E. Sponer, A. Mladek, P. Banas, P. Jurecka, M. Otyepka, Methods 64 (2013) 3–11.
[17] W.G. Noid, J. Chem. Phys. 139 (2013) 090901.
[18] W.G. Noid, Methods Mol. Biol. 924 (2013) 487–531.
[19] W.G. Noid, in: L. Monticelli, E. Salonen (Eds.), Biomol Sim: Meth Prot, Meth Mol Biol, Springer Science + Business Media, New York, 2012.
[20] S. Matysiak, C. Clementi, Arch. Biochem. Biophys. 469 (2008) 29–33.
[21] V. Tozzini, Q. Rev. Biophys. 43 (2010) 333–371.
[22] M.G. Saunders, G.A. Voth, Annu. Rev. Biophys. 42 (2013) 73–93.
[23] S.C. Kamerlin, S. Vicatos, A. Dryga, A. Warshel, Annu. Rev. Phys. Chem. 62 (2011) 41–64.
[24] M.A. Jonikas, R.J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, R.B. Altman, RNA 15 (2009) 189–199.
[25] W.K. Olson, V.B. Zhurkin, Curr. Opin. Struct. Biol. 10 (2000) 286–297.
[26] K. Reblova, J. Sponer, F. Lankas, Nucleic Acids Res. 40 (2012) 6290–6303.
[27] M.G. Munteanu, K. Vlahovicek, S. Parthasarathy, I. Simon, S. Pongor, Trends Biochem. Sci. 23 (1998) 341–347.
[28] A. Balaeff, L. Mahadevan, K. Schulten, Phys. Rev. E 73 (2006) 031919.
[29] D. Swigon, B.D. Coleman, I. Tobias, Biophys. J. 74 (1998) 2515–2530.
[30] J. Sponer, J.E. Sponer, A. Mladek, P. Jurecka, P. Banas, M. Otyepka, Biopolymers 99 (2013) 978–988.
[31] P. Banas, P. Jurecka, N.G. Walter, J. Sponer, M. Otyepka, Methods 49 (2009) 202–216.
[32] P. Bala, P. Grochowski, K. Nowinski, B. Lesyng, J.A. McCammon, Biophys. J. 79 (2000) 1253–1262.
[33] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, J. Comput. Chem. 25 (2004) 1157–1174.
[34] W.D. Cornell, P. Cieplak, C.L. Bayly, I.R. Gould, K.M. Merz Jr., D.M. Ferguson, D. C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, J. Am. Chem. Soc. 117 (1995) 5179–5197.
[35] R. Galindo-Murillo, D.R. Roe, T.E. Cheatham 3rd, Biochim. Biophys. Acta 1850 (2015) 1041–1058.
[36] A.D. MacKerell Jr., N. Banavali, N. Foloppe, Biopolymers 56 (2000) 257–265.
[37] C. Oostenbrink, A. Villa, A.E. Mark, W.F. van Gunsteren, J. Comput. Chem. 25 (2004) 1656–1676.
[38] T.A. Halgren, J. Comput. Chem. 17 (1996) 490–519.
[39] W.L. Jorgensen, D.S. Maxwell, J. Tirada-Rives, J. Am. Chem. Soc. 118 (1996) 11225–11236.
[40] W.L. Jorgensen, J. Tirada-Rives, J. Am. Chem. Soc. 110 (1988) 1657–1666.
[41] M.J. Sippl, J. Comput. Aided Mol. Des. 7 (1993) 473–501.
[42] I. Shcherbakova, S. Mitra, A. Laederach, M. Brenowitz, Curr. Opin. Chem. Biol. 12 (2008) 655–666.
[43] C. Hyeon, D. Thirumalai, Biophys. J. 90 (2006) 3410–3427.
[44] H. Isambert, Methods 49 (2009) 189–196.
[45] A. Xayaphoummine, V. Viasnoff, S. Harlepp, H. Isambert, Nucleic Acids Res. 35 (2007) 614–622.
[46] A. Xayaphoummine, T. Bucher, H. Isambert, Nucleic Acids Res. 33 (2005) W605–W610.
[47] A. Xayaphoummine, T. Bucher, F. Thalmann, H. Isambert, Proc. Natl. Acad. Sci. 100 (2003) 15310–15314.
[48] E.J. Sambriski, V. Ortiz, J.J. de Pablo, J. Phys. Condens. Matter 21 (2009) 034105.
[49] T.E. Ouldridge, P. Sulc, F. Romano, J.P. Doye, A.A. Louis, Nucleic Acids Res. 41 (2013) 8886–8895.
[50] P. Sulc, F. Romano, T.E. Ouldridge, J.P. Doye, A.A. Louis, J. Chem. Phys. 140 (2014) 235102.
[51] T. Cragnolini, P. Derreumaux, S. Pasquali, J. Phys. Chem. B 117 (2013) 8047–8060.
[52] M. Maciejczyk, A. Spasic, A. Liwo, H.A. Scheraga, J. Chem. Theory Comput. 10 (2014) 5020–5035.
[53] Y. He, M. Maciejczyk, S. Oldziej, H.A. Scheraga, A. Liwo, Phys. Rev. Lett. 110 (2013) 098101.
[54] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, P. Kollman, Comput. Phys. Commun. 91 (1995) 1–41.
[55] D.A. Case, T.E. Cheatham 3rd, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, R.J. Woods, J. Comput. Chem. 26 (2005) 1668–1688.
[56] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, J. Comput. Chem. 4 (1983) 187–217.
[57] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, Bioinformatics 29 (2013) 845–854.
[58] H.J.C. Berendsen, D. Vanderspoel, R. Vandrunen, Comput. Phys. Commun. 91 (1995) 43–56.
[59] J.W. Ponder, F.M. Richards, J. Comput. Chem. 8 (1987) 1016–1024.
[60] P. Ren, C. Wu, J.W. Ponder, J. Chem. Theory Comput. 7 (2011) 3143–3161.
[61] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, K. Schulten, J. Comput. Phys. 151 (1999) 283–312.
[62] M.T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L.V. Kale, R.D. Skeel, K. Schulten, Int. J. Supercomput. Appl. High Perform. Comput. 10 (1996) 251–268.
[63] S.J. Weiner, P.A. Kollman, D.T. Nguyen, D.A. Case, J. Comput. Chem. 7 (1986) 230–252.
[64] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, J. Phys. Chem. B 102 (1998) 3586–3616.
[65] J.P. Devlin, I.C. Hisatsune, Spectrochim. Acta 17 (1961) 218–225.
[66] C.A. Morgado, D. Svozil, D.H. Turner, J. Sponer, Phys. Chem. Chem. Phys. 14 (2012) 12580–12591.
[67] A.R. Srinivasan, R.R. Sauers, M.O. Fenley, A.H. Boschitsch, A. Matsumoto, A.V. Colasanti, W.K. Olson, Biophys. Rev. 1 (2009) 13–20.
[68] M.G. Seetin, D.H. Mathews, J. Comput. Chem. (2011).
[69] V. Tsui, D.A. Case, Biopolymers 56 (2001) 275–291.
[70] R. Das, M. Kudaravalli, M. Jonikas, A. Laederach, R. Fong, J.P. Schwans, D. Baker, J.A. Piccirilli, R.B. Altman, D. Herschlag, Proc. Natl. Acad. Sci. 105 (2008) 4144–4149.
[71] R. Das, J. Karanicolas, D. Baker, Nat. Methods 7 (2010) 291–294.
[72] E. Capriotti, T. Norambuena, M.A. Marti-Renom, F. Melo, Bioinformatics 27 (2011) 1086–1093.
[73] C.Y. Cheng, F.C. Chou, R. Das, Methods Enzymol. 553 (2015) 35–64.
[74] R. Das, D. Baker, Proc. Natl. Acad. Sci. 104 (2007) 14664–14669.

[75] M. Parisien, F. Major, Nature 452 (2008) 51–55.
[76] V. Reinharz, F. Major, J. Waldispuhl, Bioinformatics 28 (2012) i207–i214.
[77] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, J. Am. Chem. Soc. 112 (1990) 6127–6129.
[78] B. Honig, K. Sharp, A.S. Yang, J. Phys. Chem. 97 (1993) 1101–1109.
[79] N.J. Deng, P. Cieplak, J. Chem. Theory Comput. 3 (2007) 1435–1450.
[80] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, D.A. Case, J. Am. Chem. Soc. 120 (1998) 9401–9409.
[81] I. Besseova, K. Reblova, N.B. Leontis, J. Sponer, Nucleic Acids Res. 38 (2010) 6247–6264.
[82] J. Florian, J. Sponer, A. Warshel, J. Phys. Chem. B 103 (1999) 884–892.
[83] D. Svozil, P. Hobza, J. Sponer, J. Phys. Chem. B 114 (2010) 1191–1203.
[84] S. Piana, J.L. Klepeis, D.E. Shaw, Curr. Opin. Struct. Biol. 24 (2014) 98–105.
[85] J.L. Knight, C.L. Brooks 3rd, J. Comput. Chem. 32 (2011) 2909–2923.
[86] S.E. McDowell, N. Spackova, J. Sponer, N.G. Walter, Biopolymers 85 (2007) 169–184.
[87] R.C. Harris, B.M. Pettitt, J. Chem. Theory Comput. 11 (2015) 4593–4600.
[88] Y. Liu, E. Haddadian, T.R. Sosnick, K.F. Freed, H. Gong, Biophys. J. 105 (2013) 1248–1257.
[89] J. Kleinjung, F. Fraternali, Curr. Opin. Struct. Biol. 25 (2014) 126–134.
[90] S.C. Kamerlin, M. Haranczyk, A. Warshel, ChemPhysChem 10 (2009) 1125–1134.
[91] M. Dlugosz, J.M. Antosiewicz, Chem. Phys. 302 (2004) 161–170.
[92] M. Dlugosz, J.M. Antosiewicz, A.D. Robertson, Phys. Rev. E 69 (2004).
[93] M. Dlugosz, J.M. Antosiewicz, Z. Naturforsch. A 59 (2004) 873–874.
[94] G.B. Goh, B.S. Hulbert, H. Zhou, C.L. Brooks 3rd, Proteins 82 (2014) 1319–1331.
[95] G.B. Goh, J.L. Knight, C.L. Brooks 3rd, J. Chem. Theory Comput. 8 (2012) 36–46.
[96] G.B. Goh, J.L. Knight, C.L. Brooks 3rd, J. Chem. Theory Comput. 9 (2013) 935–943.
[97] H. Yu, T.W. Whitfield, E. Harder, G. Lamoureux, I. Vorobyov, V.M. Anisimov, A. D. Mackerell Jr., B. Roux, J. Chem. Theory Comput. 6 (2010) 774–786.
[98] I. Leontyev, A. Stuchebrukhov, Phys. Chem. Chem. Phys. 13 (2011) 2613–2626.
[99] M.J. Boniecki, G. Lach, W.K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K.M. Rother, J.M. Bujnicki, Nucleic Acids Res. (2015), http://dx.doi.org/10.1093/nar/gkv1479.
[100] T.E. Ouldridge, A.A. Louis, J.P. Doye, J. Chem. Phys. 134 (2011) 085101.
[101] J.C. Bowman, T.K. Lenz, N.V. Hud, L.D. Williams, Curr. Opin. Struct. Biol. 22 (2012) 262–272.
[102] V. Tozzini, Acc. Chem. Res. 43 (2010) 220–230.
[103] D.H. de Jong, G. Singh, W.F.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schafer, X. Periole, D.P. Tieleman, S.J. Marrink, J. Chem. Theory Comput. 9 (2013) 687–697.
[104] S.J. Marrink, D.P. Tieleman, Chem. Soc. Rev. 42 (2013) 6801–6822.
[105] R. Chen, Z. Weng, Proteins 51 (2003) 397–408.
[106] B.G. Pierce, K. Wiehe, H. Hwang, B.H. Kim, T. Vreven, Z. Weng, Bioinformatics 30 (2014) 1771–1773.
[107] R. Chen, L. Li, Z. Weng, Proteins 52 (2003) 80–87.
[108] M. Maciejczyk, A. Spasic, A. Liwo, H.A. Scheraga, J. Comput. Chem. 31 (2010) 1644–1655.
[109] W.K. Olson, P.J. Flory, Biopolymers 11 (1972) 1–23.
[110] W.K. Olson, P.J. Flory, Biopolymers 11 (1972) 25–56.
[111] W.K. Olson, P.J. Flory, Biopolymers 11 (1972) 57–66.
[112] W.K. Olson, Macromolecules 13 (1980) 721–728.
[113] L.J. Murray, W.B. Arendall 3rd, D.C. Richardson, J.S. Richardson, Proc. Natl. Acad. Sci. 100 (2003) 13904–13909.
[114] A. Takasu, K. Watanabe, G. Kawai, Nucleic Acids 21 (2002) 449–462.
[115] W. Saenger, Principles of Nucleic Acid Structure, Springer-Verlag, New York, 1984.
[116] C.M. Duarte, L.M. Wadley, A.M. Pyle, Nucleic Acids Res. 31 (2003) 4755–4761.
[117] L.M. Wadley, K.S. Keating, C.M. Duarte, A.M. Pyle, J. Mol. Biol. 372 (2007) 942–957.
[118] C.M. Duarte, A.M. Pyle, J. Mol. Biol. 284 (1998) 1465–1478.
[119] L. Liu, S.J. Chen, PLoS One 7 (2012) e48460.
[120] W. Dawson, G. Kawai, J. Comput. Sci. Syst. Biol. 2 (2009) 001–023.
[121] S. Cao, S.J. Chen, RNA 11 (2005) 1884–1897.
[122] W.K. Olson, Macromolecules 8 (1975) 272–275.
[123] A. Malhotra, H.A. Gabb, S.C. Harvey, Curr. Opin. Struct. Biol. 3 (1993) 241–246.
[124] A. Malhotra, R.K.Z. Tan, S.C. Harvey, Biophys. J. 66 (1994) 1777–1795.
[125] N. Kim, M. Zahran, T. Schlick, Methods Enzymol. 553 (2015) 115–135.
[126] P. Kerpedjiev, C. Honer Zu Siederdissen, I.L. Hofacker, RNA 21 (2015) 1110–1121.
[127] D. Jost, R. Everaers, J. Chem. Phys. 132 (2010) 095101.
[128] J.M. Hubbard, J.E. Hearst, Biochemistry 30 (1991) 5458–5465.
[129] J.M. Hubbard, J.E. Hearst, J. Mol. Biol. 221 (1991) 889–907.
[130] A. Malhotra, R.K.Z. Tan, S.C. Harvey, J. Comput. Chem. 15 (1994) 190–199.
[131] R.K. Tan, A.S. Petrov, S.C. Harvey, J. Chem. Theory Comput. 2 (2006) 529–540.
[132] F. Ding, D. Tsao, H. Nie, N.V. Dokholyan, Structure 16 (2008) 1010–1018.
[133] C.M. Gherghe, C.W. Leonard, F. Ding, N.V. Dokholyan, K.M. Weeks, J. Am. Chem. Soc. 131 (2009) 2541–2546.
[134] C. Hyeon, D. Thirumalai, Proc. Natl. Acad. Sci. 102 (2005) 6789–6794.
[135] D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner, J. Mol. Biol. 288 (1999) 911–940.
[136] T. Xia, J. SantaLucia Jr., M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, D.H. Turner, Biochemistry 37 (1998) 14719–14735.
[137] Z. Xia, D.P. Gardner, R.R. Gutell, P. Ren, J. Phys. Chem. B 114 (2010) 13497–13506.
[138] Z. Xia, D.R. Bell, Y. Shi, P. Ren, J. Phys. Chem. B 117 (2013) 3135–3144.
[139] S. Sharma, F. Ding, N.V. Dokholyan, Bioinformatics 24 (2008) 1951–1952.
[140] D.L. Pincus, S.S. Cho, C.B. Hyeon, D. Thirumalai, Prog. Mol. Biol. Transl. 84 (2008) 203.
[141] M.J. Sippl, J. Mol. Biol. 213 (1990) 859–883.
[142] T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro, J. Ferkinghoff-Borg, PLoS One 5 (2010) e13714.
[143] O. Taxilaga-Zetina, P. Pliego-Pastrana, M.D. Carbajal-Tinoco, J. Chem. Phys. 140 (2014) 115106.
[144] S. Cao, S.J. Chen, J. Phys. Chem. B 115 (2011) 4216–4226.
[145] S. Cao, D.P. Giedroc, S.J. Chen, RNA 16 (2010) 538–552.
[146] X.J. Xu, S.J. Chen, Biophys. Rep. (2015).
[147] A.M. Mustoe, H.M. Al-Hashimi, C.L. Brooks 3rd, J. Phys. Chem. B 118 (2014) 2615–2627.
[148] N.A. Denesyuk, D. Thirumalai, J. Phys. Chem. B 117 (2013) 4901–4911.
[149] Z.J. Tan, S.J. Chen, Biophys. J. 92 (2007) 3615–3632.
[150] Y. Zhu, Z. He, S.J. Chen, PLoS One 10 (2015) e0119705.
[151] Z. He, Y. Zhu, S.J. Chen, Phys. Chem. Chem. Phys. 16 (2014) 6367–6375.
[152] Z. He, S.J. Chen, J. Chem. Theory Comput. 8 (2012) 2095–2101.
[153] J. Bernauer, X. Huang, A.Y. Sim, M. Levitt, RNA 17 (2011) 1066–1075.
[154] A.Y.L. Sim, M. Levitt, P. Minary, Proc. Natl. Acad. Sci. 109 (2012) 2890–2895.
[155] P. Brion, E. Westhof, Annu. Rev. Biophys. Biomol. Struct. 26 (1997) 113–137.
[156] I. Tinoco, C. Bustamante, J. Mol. Biol. 293 (1999) 271–281.
[157] L. Jaeger, E. Westhof, N.B. Leontis, Nucleic Acids Res. 29 (2001) 455–463.
[158] W. Dawson, K. Yamamoto, K. Shimizu, G. Kawai, J. Nucleic Acids Invest. 4 (2013) e2.
[159] S. Geman, D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741.
[160] U.H. Hansmann, Y. Okamoto, Curr. Opin. Struct. Biol. 9 (1999) 177–183.
[161] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671–680.
[162] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman, J. Comput. Chem. 16 (1995) 1339–1350.
[163] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman, J.M. Rosenberg, J. Comput. Chem. 13 (1992) 1011–1021.
[164] B. Roux, Comput. Phys. Commun. 91 (1995) 275–278.
[165] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. 314 (1999) 141–151.
[166] A. Laio, M. Parrinello, Proc. Natl. Acad. Sci. 99 (2002) 12562–12566.
[167] M.C. Tesi, E.J. Jans van Rensberg, E. Orlandini, S.G. Whittington, J. Stat. Phys. 82 (1996) 155–181.
[168] Y.M. Rhee, V.S. Pande, Biophys. J. 84 (2003) 775–786.
[169] S. Whitelam, E.H. Feng, M.F. Hagan, P.L. Geissler, Soft Matter 5 (2009) 1251–1262.
[170] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21 (1953) 1087–1092.
[171] J. Frellsen, I. Moltke, M. Thiim, K.V. Mardia, J. Ferkinghoff-Borg, T. Hamelryck, PLoS Comput. Biol. 5 (2009) e1000406.
[172] A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. 61 (1988) 2635–2638.
[173] A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. 63 (1989) 1195–1198.
[174] N. Nakajima, H. Nakamura, A. Kidera, J. Phys. Chem. B 101 (1997) 817–824.
[175] U.H.E. Hansmann, Y. Okamoto, F. Eisenmenger, Chem. Phys. Lett. 259 (1996) 321–330.
[176] B.A. Berg, T. Neuhaus, Phys. Rev. Lett. 68 (1992) 9–12.
[177] W. Janke, Phys. A 254 (1998) 164–178.
[178] F. Glover, Comput. Oper. Res. 13 (1986) 533–549.
[179] R. Harada, Y. Takano, Y. Shigeta, J. Comput. Chem. 36 (2015) 763–772.
[180] S.F. Chekmarev, Phys. Rev. E 64 (2001) 036703.
[181] D. Cvijovic, J. Klinowski, Science 267 (1995) 664–666.
[182] R. Harada, Y. Takano, T. Baba, Y. Shigeta, Phys. Chem. Chem. Phys. 17 (2015) 6155–6173.
[183] T.R. Sosnick, D. Barrick, Curr. Opin. Struct. Biol. 21 (2011) 12–24.
[184] R. Samudrala, J. Moult, J. Mol. Biol. 279 (1998) 287–302.
[185] M. Vassura, L. Margara, P. Fariselli, R. Casadio, Artif. Intell. Med. 45 (2009) 229–237.
[186] R. Harada, A. Kitao, J. Phys. Chem. B 115 (2011) 8806–8812.
[187] A.W. Serohijos, D. Tsygankov, S. Liu, T.C. Elston, N.V. Dokholyan, Phys. Chem. Chem. Phys. 11 (2009) 4840–4850.
[188] Z. Miao, R.W. Adamiak, M.F. Blanchet, M. Boniecki, J.M. Bujnicki, S.J. Chen, C. Cheng, G. Chojnowski, F.C. Chou, P. Cordero, J.A. Cruz, A.R. Ferre-D'Amare, R. Das, F. Ding, N.V. Dokholyan, S. Dunin-Horkawicz, W. Kladwang, A. Krokhotin, G. Lach, M. Magnus, F. Major, T.H. Mann, B. Masquida, D. Matelska, M. Meyer, A. Peselis, M. Popenda, K.J. Purzycka, A. Serganov, J. Stasiewicz, M. Szachniuk, A. Tandon, S. Tian, J. Wang, Y. Xiao, X. Xu, J. Zhang, P. Zhao, T. Zok, E. Westhof, RNA 21 (2015) 1066–1084.
[189] J.A. Cruz, M.F. Blanchet, M. Boniecki, J.M. Bujnicki, S.J. Chen, S. Cao, R. Das, F. Ding, N.V. Dokholyan, S.C. Flores, L. Huang, C.A. Lavender, V. Lisi, F. Major, K. Mikolajczak, D.J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K.M. Weeks, C. Waldsich, M. Wildauer, N.B. Leontis, E. Westhof, RNA 18 (2012) 610–625.
[190] C.E. Hajdin, F. Ding, N.V. Dokholyan, K.M. Weeks, RNA 16 (2010) 1340–1349.
[191] A. Kryshtafovych, J. Moult, P. Bales, J.F. Bazan, M. Biasini, A. Burgin, C. Chen, F. V. Cochran, T.K. Craig, R. Das, D. Fass, C. Garcia-Doval, O. Herzberg, D. Lorimer,

H. Luecke, X. Ma, D.C. Nelson, M.J. van Raaij, F. Rohwer, A. Segall, V. Seguritan, K. Zeth, T. Schwede, Proteins 82 (Suppl. 2) (2014) 26–42.

[192] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Proteins 82 (Suppl. 2) (2014) 1–6.

[193] T. Norambuena, J.F. Cares, E. Capriotti, F. Melo, Bioinformatics 29 (2013) 2649–2650.

[194] S. Pasquali, P. Derreumaux, J. Phys. Chem. B 114 (2010) (1966) 11957–11966.

[195] A. Krokhotin, N.V. Dokholyan, Methods Enzymol. 553 (2015) 65–89.

[196] M.A. Jonikas, R.J. Radmer, R.B. Altman, Bioinformatics 25 (2009) 3259–3266.

[197] M.T. Sykes, M. Levitt, J. Mol. Biol. 351 (2005) 26–38.