Topical Review

Ab initio RNA folding

Tristan Cragnolini1,3, Philippe Derreumaux1,2 and Samuela Pasquali1

1 Laboratoire de Biochimie Théorique UPR 9080 CNRS, Université Paris Diderot, Sorbonne, Paris Cité, IBPC 13 rue Pierre et Marie Curie, 75005 Paris, France
2 Institut Universitaire de France, Boulevard Saint-Michel, 75005 Paris, France
E-mail: samuela.pasquali@ibpc.fr

Received 30 December 2014, revised 18 March 2015
Accepted for publication 24 March 2015
Published 20 May 2015

Abstract
RNA molecules are essential cellular machines performing a wide variety of functions for which a specific three-dimensional structure is required. Over the last several years, the experimental determination of RNA structures through x-ray crystallography and NMR seems to have reached a plateau in the number of structures resolved each year, but as more and more RNA sequences are being discovered, the need for structure prediction tools to complement experimental data is strong. Theoretical approaches to RNA folding have been developed since the late nineties, when the first algorithms for secondary structure prediction appeared. Over the last 10 years a number of prediction methods for 3D structures have been developed, first based on bioinformatics and data-mining, and more recently based on a coarse-grained physical representation of the systems. In this review we are going to present the challenges of RNA structure prediction and the main ideas behind bioinformatic approaches and physics-based approaches. We will focus on the description of the more recent physics-based phenomenological models and on how they are built to include the specificity of the interactions of RNA bases, whose role is critical in folding. Through examples from different models, we will point out the strengths of physics-based approaches, which are able not only to predict equilibrium structures, but also to investigate dynamical and thermodynamical behavior, and the open challenges to include more key interactions ruling RNA folding.

Keywords: RNA, coarse-grained models, force field

(Some figures may appear in colour only in the online journal)

1. Introduction
Over the last 15 years it has been recognized that RNAs play a wide range of functions besides their well-known roles of genetic information carrier (mRNA) and amino acid recruiter (tRNA): microRNA (miRNA) are short sequences regulating genes in the post-transcriptional process, the small interference RNA (RNAi) acts on the gene silencing mechanism, ribozymes are mid-sized (often less than 100 nucleotides) molecules with catalytic properties, and ribosomal RNA constitutes the ribosome together with proteins and can be as large as several thousand nucleotides. The publication of high-resolution x-ray structures revealed that the catalytic activity in the ribosome was carried by RNA, and not the associated proteins [1, 2]. Many more ribozymes have been identified. The RNase P is necessary for the maturation of tRNA, while intron splicing is catalyzed by a protein-RNA complex, the spliceosome [3]. Other ribozymes play a role in metabolic pathways, such as the glucosamine-6-phosphate (glmS) riboyme, regulating the translation of the protein catalyzing the production of glmS [4]. There is also growing interest in the use of RNA for nanotechnology, with the creation of self-assembling systems such as artificial nanorings [5], nanocages [6], nanoscale scaffolds [7, 8], and other nanostructures [9]. More recently, riboswitches have been identified. Those sequences, usually present in the 5’ untranslated region of genes, adopt a specific
fold in the presence or absence of a ligand. The folding of the riboswitch sequence will then regulate the expression of the associated gene. Riboswitches have been identified for purine bases, adenine and guanine, for amino acids, notably tryptophan, as well as organic compounds, such as fluoride. RNA is also a prime candidate for being a key molecule in the emergence of life on earth [10]. Like proteins, the functionality of these molecules depends crucially on their equilibrium structures and their dynamical behavior [11, 12], with distinct active conformations biologically active under different conditions [13]. This poses the problem of understanding RNA folding, which is why and how a specific sequence adopts a specific tertiary structure.

The ENCODE project showed that a large number of non-protein-coding RNA transcripts were produced, most of them with no previously recognized roles [14]. With the explosion of sequencing data, with nearly 200 millions entries contributed to GenBank over the last 30 years, and most DNA being detected as ‘non-coding’, therefore possibly containing the information to synthesize RNAs, structure prediction from sequence is an urgent matter. High resolution experimental techniques for determining three-dimensional structures, such as x-ray crystallography and NMR, are challenging as is shown by the small number of resolved structures in the Nucleic Acids Data Bank (NDB) and by the scarcity of structures with substantially different architectures. Low-resolution techniques, such as SAXS and Cryo-EM, allow for easier access to the raw data, but require extensive modeling to propose a well-resolved structure.

1.1. RNA structural organization

Before entering the details of RNA folding predictions it is useful to outline the different levels of complexity that are involved. RNA, just like DNA, benefits from sequence complementarity, with A pairing with U and G pairing with C. If we have strands with perfect complementary sequences, the structure of the molecule is a perfect helix (for RNA an A-form). In this case predicting the fold of the molecule is rather trivial as the characteristics of the helix (rise, pitch, ...) are well known. But RNA sequences almost never allow for base complementarity along the whole sequence. RNAs are most often single stranded molecules that have sequences incompatible with the formation of long double helices. Nonetheless they can have short portions of complementary sequences giving rise to short helices separated by single stranded regions. Portions of the sequence close by tend to form helices and give rise to hairpins, with a helical stem and a terminating loop of variable size. Helices and single stranded regions arrange in space with the possible formation of base pairs external to helices. Often these contacts exhibit non-canonical pairings, that is, base pairs other than AU or CG, and involve all sides of the base [15].

If for proteins the definition of secondary and tertiary structure comes unambiguously from backbone hydrogen bonds, for RNA the definition is more delicate because base pairing occurs both at intermediate length scales with hairpins, and at large length scales with bonds holding together already formed structures. We adopt the following definition: two base pairs I \circ J, H \circ K, are called nested if I<H<K<J, unrelated if I<J<H<K, linked if I<H<J<K. RNA secondary structure is a set of base pairs in which no two base pairs are linked, that is, every base pair in a secondary structure is either nested or unrelated [16]. If we represent secondary structures as graphs with nucleotides identified with nodes and lines representing base pairs, this is equivalent to saying that secondary structures can be represented by planar graphs in which no lines intersect (figure 1). We can classify basic secondary structures into single stranded regions, hairpins, bulge loops, mismatches, internal loops, and junctions [17].

Nucleotides which are linked form tertiary interactions. Most tertiary interactions involve non-canonical base pairing
or backbone-backbone interaction. Some tertiary structures are adenosine platforms, triplets, helices docking, metal-core motifs, and ribose zippers. Adenosine platforms and triplets involve multiple base pairing, helix docking involves non-canonical pairing as well as backbone-backbone or backbone-base interactions, ribose zippers involve backbone-backbone interactions. Tertiary structures involving canonical pairing are pseudoknots and kissing loops. Pseudoknots and loop-loop interactions tie together single stranded regions.

The definition of the tertiary structure adopted for RNA is clearly different from the definition commonly used for proteins, where the term refers to the global organization of secondary structure elements in space. For RNA we will refer to the three-dimensional global organization as to the architecture.

Early experiments on RNA melting showed that RNA unfolds in a series of discrete steps corresponding to the breaking down of the folding process into localized regions of the structure [18]. More recent analyses on large ribosomal molecules show that RNA has a rich modular structure [11, 19], findings also supported by numerous single molecule pulling experiments [20, 21]. These findings suggest that an RNA molecule possesses a hierarchical structure in which the primary sequence determines the secondary structure, which in turn determines its tertiary folding: the three-dimensional architecture results from the compaction of separate pre-existing and stable elements that form autonomous entities.

Exceptions to this general scheme exist, as is the case for some complex architectures and pseudoknots, where the melting of tertiary structures is not well separated from the melting of secondary structures [22]. Because of the stability of base-pairing at room temperature and of the frustration in the possible secondary structures, RNA can adopt dramatically different conformations all of similar energetic stabilities. Some of these structures have indeed alternative conformations that the molecule can adopt in response to environmental conditions, others are kinetic traps that can lead to the molecule degradation by regulating factors [23–25]. Typically small RNA molecules reach their native state without being trapped in misfolded structures, while long molecules are trapped more easily with increasing chain length [26].

In conclusion, despite what one could naively think based on base complementarity, the folding of RNA can not be considered fully as a hierarchical problem. Whether secondary and tertiary structures can be treated separately depends on the molecule’s size and structural complexity.

1.2. Computational challenges

With the recent access to massive computational resources and with the establishment of reliable atomistic force fields, one would think that numerical studies of RNA folding should not pose a problem. In the late nineties the common belief was that if only a fraction of the resources put into solving protein folding was put into RNA folding, the problem would have been solved quickly [18]. Yet, fifteen years later, despite the increase in human and computational resources, the question is still open. The obvious approach using an atomistic model and simulations proves particularly challenging, and in practice limits the studies to very small molecules and short times [27, 28]. The main difficulty comes from the size of the molecule and the time of structure formation, as several different length scales are involved for both. When RNA renatures, stem-loops form in microseconds, while global architectures can take seconds to minutes to develop. Even if we were able to simulate efficiently at the lowest time scale, which is far from being the case with atomistic simulations, there would be several orders of magnitude in time to look at just one folding event, not to mention any statistical analysis. Concerning size, even just at the level of a secondary structure, molecules with more than a dozen nucleotides have a multitude of possible states and base-pairing space quickly becomes extremely large [29], with different structures separated by large energy barriers. Two additional problems come from the high charge of the RNA backbone, giving rise to crucial interactions with both solvent and ions in solution, and the intrinsic nature of hydrogen bonding and stacking that would require quantum mechanical calculations for accurate results [30]. Because of these challenges, nucleic acid force fields are still far from being as reliable as one would like them to be. AMBER, which among classical atomistic force fields is the one that has been developed more carefully for nucleic acids, works well for small helical structures, but by the admission of its own developers, fails in the study of RNA single stranded molecules, as the configurational changes involved go beyond the testing ground of its parameters [31].

Given the limitations of atomistic simulations, different strategies have been applied to RNA structure prediction and can be loosely organized into three categories [32]: knowledge-based homology models, hybrid bioinformatic methods, and coarse-grained ab initio models.

As is the case for proteins, when the question is that of determining a 3D configuration with the best possible accuracy, homology models based on sequence similarity perform well, provided one can find an already resolved structure that serves as a template [33, 34]. This is rarely the case for single stranded RNA.

The hybrid category comprises a large variety of methods, going from fragment reconstructions [35, 36] to models strongly relying on secondary structure prediction algorithms and 3D scaffolds extracted from the NDB [37]. In general these methods are good in predicting local structures, but have their weakness in the prediction of overall complex architectures, unless experimental additional constraints on the tertiary structure are known.

Coarse-grained ab initio models try to capture the physics of the system, and aim at predicting equilibrium structures as well as folding intermediates and energy landscapes. Atoms are grouped in particles constituting the elementary objects of the model, and a set of forces are defined to generate a dynamic.

In this review we will present the basic principles of the different strategies of physics-based models for RNA folding. To put these developments into context, in section 2 we will discuss the state of the art of single stranded nucleic acid atomistic simulations and in section 3 we will discuss bioinformatic approaches, with the explicit examples...
Section 4 constitutes the main body of the review and focuses on ab initio coarse-grained models. Here we will address the various issues going into building a CG model for RNA: choice of particles to represent the system, choice of the functions describing the interactions, and parametrization. These aspects will be developed in detail for the models that currently have the best prediction capabilities, namely the model by Xia [38], OxAshRNA [39], and HiRE-RNA [40]. We will also discuss the simulation methods employed by coarse-grained models for structure prediction (section 5) and discuss the performance of both bioinformatics and ab initio approaches on some benchmark systems (section 6). Section 7 presents how coarse-grained models can be coupled to existing experimental data to obtain structures fulfilling constraints coming from experiments. We will conclude with a discussion on the open challenges coming from the interplay of ions and RNA and the interactions of bases with other groups, such as phosphates (section 8) and with perspectives on how current ab initio models could evolve to account for the environment in which folding takes place and that ultimately influences the structures active in nature section 9.

RNA and DNA molecules being chemically very similar at a coarse-grained level by neglecting explicitly the OH group, we include in this review models and simulation results for DNA also when the DNA model presents insights into how to consider nucleic acids in general.

2. Atomistic models

Atomistic simulations comprise most of the necessary ingredients to describe the interactions ruling the behavior of RNA molecules. Through well-established empirical potentials, one can in principle access the question of folding and of the dynamics and thermodynamics of the molecule. Typical atomistic force-fields adopt a harmonic description for bond lengths and angles, sinusoidal potentials for dihedral angles, Lennard–Jones potentials to describe long-range Van der Waals interactions, and a description of electrostatics which is dependent upon the treatment of solvent and ions. Coulomb potentials in periodic boundary conditions are used for explicit solvent and ions representation, and Born or generalized Born approximations are used in implicit solvent to avoid the computational cost of having to solve the Poisson–Boltzmann equation. For nucleic acids, historically, the reference atomistic force field is AMBER, which has been developed and tested thoroughly over the past fifteen years, especially on DNA duplexes [41, 42]. More recently, the force field parameters have been adjusted to account for A-form helices, the typical helix formed by RNA, and to correctly represent short loops, even though the results remain in general sensitive to the ionic [43] and water representation chosen [44].

Even with the most recent force field, ff12, simulating single stranded RNA remains an open challenge for at least two reasons.

The first, obvious, problem is in the size of the systems that can be considered, and that limits the applicability of atomistic simulations to small fragments of fewer than a dozen nucleotides. In explicit solvent, the effects of size for nucleic acids can have an even higher impact than for proteins given that RNA and DNA tend to have more elongated configurations, and therefore require larger water boxes, slowing down the simulation even further. In practice, atomistic simulations on large systems are limited to an investigation around the experimentally available structure. To this day, only a few unbiased simulations have been able to describe folding events for molecules no longer than a dozen nucleotides [27, 45–49].

One recent interesting atomistic simulation highlighting the challenges of single stranded nucleic acid folding has been presented by J. Sponer and collaborators [45]. They study the late stages of folding of DNA G-quadruplexes, a motif that can be formed by both RNA and DNA, found on telomeres, and thought to be related to the development of certain cancers [50]. These motifs clearly show the peculiarity of the structures that single stranded DNA or RNA can form, greatly departing from the double helix, in which several bases can interconnect forming hydrogen bonds on all of their three sides (Watson–Crick, Hoogsteen, and Sugar) giving rise to ‘platforms’ of three or four bases. In the case of G-quadruplexes, four G bases come together on a plane forming squares through Watson–Crick and Hoogsteen pairings; several such G-quartet configurations stack on top of each other with intercalating K+ or Na+ ions. In the simulation by Sponer, the native G-quadruplex configurations are destabilized by an initial simulation in the absence of the ions. Under these conditions the quadruplex opens to partially unfolded states that are then used as initial configurations to attempt refolding once natural ionic conditions have been restored. Refolding is studied through Molecular Dynamic (MD) simulations at ambient temperature and physiological ionic conditions (figure 2). It is found that not all the partially unfolded configurations are able to fold back into the experimental quadruplet, but some remain trapped in states with alternative base-pairing organizations. Once alternative base pairs are formed in MD they are too stable...
to break under physiological conditions, and for all practical purposes the molecule can remain trapped indefinitely, making it impossible to better explore its configurational landscape.

The second, more subtle, problem is related to parametrization. Most of the structures currently known for nucleic acids come from double helices, implying that if we can characterize well all degrees of freedom of the systems for configurations close to those adopted by bases forming Watson–Crick pairs, we are left with little information on all other possible structures. Biases in parametrization procedures are common in all fields, but that is especially problematic in the context of RNA folding where we want to investigate configurations that are heavily under-represented in the databases used to train the force-field. Indeed, in single stranded RNA many bases are involved in interactions other than those of a regular double helix and even the sugar-phosphate backbone can be found highly deformed from the typical helix. When structures depart significantly from the double helix, one is no longer guaranteed that the parameter set, or even the functional forms used, are still adequate to describe the system.

The recent work of Garcia and collaborators on RNA tetraloop folding clearly illustrate these difficulties. Through REMD atomistic simulations, Garcia was able to obtain full folding trajectories and thermodynamical information for the folding of three 8 nucleotides hairpins forming hyperstable tetraloops [27]. Garcia was able to reach such a significant result through an important reparametrization of the AMBER-99 force field to obtain better agreement with the thermodynamic and kinetic measurements of RNA monomers and dimers. To derive accurate RNA parameters, the Lennard–Jones interaction had to be fully revisited to correct for the AMBER overestimate of the base-stacking propensities, the imbalance between the syn and anti glycosidic rotamers, and the violation of the contact distances as calculated by quantum chemical means. The new parametrization was obtained through an optimization which included many different experimental sources other than structures of the NDB, including in particular thermodynamic data. Prior attempts to fold hyperstable tetraloops with atomistic details using various folding techniques (fragment reconstruction [51], interactive simulations [52], REMD [46]) failed to predict the non-canonical interactions forming in the loops and responsible for its high stability. The work by Garcia shows that commonly used force-fields for nucleic acids are still far from being optimal, and as a consequence, even if one was able to simulate larger systems using massive computing resources, the results might not be reliable, at least for the time being.

3. Bioinformatic models

In order to predict structures for systems of the size of whole RNA molecules one needs to leave the atomistic description and resort to techniques adopting a simplified vision of the molecule or use bioinformatic methods exploiting existing structural experimental data. In this section we will focus on approaches based on bioinformatics, while we will discuss ab initio models in the next.

The bioinformatics, or ‘hybrid’, category comprises a large variety of methods, going from fragment reconstructions, to models strongly relying on secondary structure prediction algorithms and 3D scaffolds extracted from the NDB. These methods are useful to obtain three-dimensional structures, but provide no information on the dynamics and thermodynamics of the molecule.

Most hybrid models employ a simplified, coarse-grained representation of the molecule, the choice of which depends on the information that is being exploited to make the prediction and on the calculations that follow, experimental structural information extracted from the NDB as a structural library, and, often, secondary structure prediction algorithms.

Given that bioinformatic methods have been the subject of numerous publications and reviews [32, 53–57], we will not give here an exhaustive account of the field, but only an illustration of what can be achieved by integrating various sources of information from bioinformatics and modeling, later to be compared with the performance of ab initio models.

3.1. Secondary structure prediction

Before entering the details of bioinformatic approaches, it is useful to spend a few words on secondary structure predictions, as most methods, to a varying extent, assume hierarchical folding and base their three-dimensional prediction on the prediction of the secondary structure first. The strategy to determine secondary structures consists of looking for the most stable set of base pairing under a simplified free energy scheme. 2D free energies are based on sequence complementarity and are constructed under the assumption that the energy is additive. While early algorithms considered only the energies of single base pairs, current models also include the effect of neighboring base pairs on the free energy, as well as loop lengths and composition. The more recent 2D energy models therefore require a large number of parameters, mostly obtained from calorimetry experiments.

As a first approximation it is assumed that all base pairs are nested. In this case the most stable secondary structure can be found efficiently using dynamic programming algorithms, with a scaling $O(N^2)$, with N the sequence length. A free energy minimization approach also including the evaluation of all possible 2D structures, and not just the most stable ones, and taking into account their contribution to the partition function, has a scaling of $O(N^3)$. The RNAstructure server [58] and ViennaRNA package [59] both implement the latter approach. Pseudoknots, however, are missing from these prediction methods as they are non-nested structures. While approaches similar to dynamic programming can be applied to pseudoknots, they tend to scale to a higher power of N, from $O(N^3)$ to $O(N^6)$, depending on the class of pseudoknots considered [60], severely limiting their application to more complex RNA.

3.2. Hierarchical predictions

Several methods use predictions of the secondary structure as a starting point to determine the 3D configuration [36, 61]. Under the assumption that local contacts form first and
that helices are more stable than tertiary interactions, the knowledge of the secondary structure allows us to greatly reduce the conformational space to be explored in 3D and provides useful information to build a starting configuration then to be refined. The predictive power of these approaches depends directly on the accuracy of the secondary prediction. Most secondary structure prediction methods capable of considering sequences of the size of whole RNAs are based on nested algorithms [36, 62] or account for Watson–Crick pairs only (or both) [61, 63–66]. These methods do not provide reliable information for 3D predictions as they miss representing both pseudoknots and non-canononical tertiary base pairs that are essential for large RNA architectures.

One exception is given by Vfold [37], which allows for the computation of free energies for secondary structures including pseudoknots. This method is based on a coarse-grained representation of the system that allows a direct evaluation of entropy parameters for different RNA motifs. Energies of base-stacking are taken from the Turner energy set [67] including non-canonical pairs in loops, and entropies are estimated by 3-body virtual bond model [68]. In a second stage a 3D coarse-grained scaffold is constructed based on the secondary structure prediction. Helices are modeled by A-forms, and loops and junctions are constructed from fragments from the NDB. An optimization procedure selects the most stable scaffold, after which a fully atomistic model is reconstructed and refined using AMBER energy minimization.

Based on secondary structure predictions, Vfold also makes predictions on melting temperatures and folding intermediates. Currently among the best prediction techniques, it was recently made available to the public through an online server [61].

3.3. Fragment assembly

Fragment-based approaches have been developed for RNA [35, 36]. They use experimentally determined structures to construct a repertoire of smaller elements, or fragments, associated with short sequences. For a given longer sequence, fragments are combined, and the total energy of the system is calculated according to an underlying physical model, which can be atomistic or coarse-grained. The structure obtained can then be refined using a local optimization or a short Monte Carlo run. Given one short sequence is usually associated with a variety of possible fragments, many structures are generated and ranked in energy.

These methods have shown a good ability to reconstruct local structures with high precision, but are less accurate on bigger structures [51]. In particular, the prediction capability of these methods is limited for structures displaying unusual or unknown folds, as they rely on structural repertoires in databases.

FARNA/FARFAR by Das and Baker [35] is the RNA fragment reconstruction method taken from the ROSETTA method [69], so successful for proteins. At the heart of the procedure is a stepwise assembly of fragments composed of 3 nucleotides, and treated in atomistic detail, to generate several million possible conformations for each given sequence [70]. Though more expensive than usual fragment reconstructions, this method allows the sampling of new combinations of nucleotide conformations. The physical potentials implemented in ROSETTA are used to compute the energies and to rank these configurations. Energies can be computed atomistically (FARFAR) or using a simplified coarse-grained potential (FARNA). Both energy models can form non-canonical pairs listed in the Leontis annotation, but are limited in the size of the molecule they can study, making it unfeasible to predict long-range tertiary contacts that stabilize large RNA molecules. FARNA/FARFAR was able to successfully predict the structure of molecules of fewer than 40 nucleotides, correctly reproducing the local backbone deformations induced by non-canonical pairings. The full-atom energy function can be supplemented with harmonic restraints to impose base pairs, typically obtained from secondary structure predictions.

Another example of a fragment-based prediction method is the MC-Fold and MC-Sym pipeline, constructing three-dimensional models from a library of nucleotide cyclic motifs that incorporate all base pairs [36]. Adjacent cyclic motifs share common base pairs and allow the proposal of a secondary structure inclusive of all possible pairings from a fragment reconstruction method. From a sequence, the MC-Fold method generates an ensemble of 2D structures ranked by their probability of occurrence, which is estimated based on the observed probability of the various 2D motifs in the proposed structure, given the sequence. MC-Sym follows a similar approach, but using 3D motifs, which are then combined via a Monte Carlo method to generate 3D structures. This pipeline has shown good accuracy for large structures, and is totally automated. It was able to build the 3D structure of a precursor microRNA and of a frame-shifting segment of HIV.

4. Ab initio coarse-grained models

In order to address the broader question of how a molecule attains its fold, of the different folding pathways, of the response of the system to environmental conditions, and of thermodynamic properties, a physical description of the systems is necessary. In section 2 we discussed how atomistic simulations would in principle provide access to all this information, but in practice they are severely limited in the size of the structures that can be studied in attainable times, making it impossible to address the question of the folding of full RNA molecules. An alternative strategy is to give a simplified representation of the system, focusing only on the degrees of freedom that are thought to be relevant to the folding problem. This can be done either by keeping the atomistic description but freezing some degrees of freedom into rigid bodies, a strategy which has proven useful to address the question of DNA interconversion between B-form and A-form helices [71] but that to our knowledge has not been extensively investigated for RNA single stranded folding, or by adopting a coarse-grained approach, where groups of atoms are replaced by beads with averaged interactions. Ab initio coarse-grained methods try to capture the physics of the system in an effective theory suited for the spatial and
temporal scales involved in folding. Through simulations responding to physical laws, they aim at predicting equilibrium structures as well as folding intermediates, and at investigating the molecule’s energy landscape and thermodynamics.

Over the last few years several coarse-grained models have been proposed to address RNA folding, with different levels of resolution and different complexities of the force-field. We can make a first classification of these models based on the level of resolution adopted to represent a nucleotide.

Maciejczyk and Sheraga developed NARES-2P [72], a 2-particle minimal nucleic acid representation for both DNA and RNA, and showed that dipole interactions between bases are sufficient to drive the formation of double helices from unpaired single strands, with a potential that is entirely physics based, and not specifically designed to reproduce neither nucleic acids structures nor thermodynamical properties.

Hyeon and Thirumalai developed the Three Interaction Site (TIS) model [73], a Go-like model with a specific stabilization term for tetraloops. The model has been used to study the mechanical unfolding of hairpins and the stability of some pseudoknots, observing in particular the dependence of folding pathways and stability upon minor sequence variations in molecules with the same topology [74].

Dokholyan’s group developed iFoldRNA, a 3-bead representation coupled to Discrete Molecular Dynamics (see section 5), an enhanced sampling technique giving access to the vast RNA conformational space. The model has been extensively tested on over 150 molecules with sizes ranging from a dozen to one hundred nucleotides and was used in the investigation of folding pathways to address the question of folding hierarchy.

Plotkin’s group developed a 3-particle DNA model where the beads representing the sugar and the phosphate groups are considered as spherical particles, and bases are treated as ellipsoids [75]. The model was shown to correctly predict the persistence lengths of both single stranded and double stranded DNA, and it has been used to study the temperature dependence of twisting and stacking double helices.

Doye’s group recently developed a rigid model with 5 interaction sites for both DNA and RNA, optimized on thermodynamic properties [39]. The model is shown to be suited for the study of the folding of a small pseudoknot, of the melting of a kissing complex, of the dynamics of a double-helical nanoring, and of hairpin unzipping under the pulling of the extremities. The twin DNA model, OxDNA [76], has been successful in the study of large DNA nanostructures and in reproducing the results of single molecule pulling experiments [77].

A 5-particle RNA model was introduced by Xia and coworkers. In unbiased simulations the model correctly folds several structures of less than 30 nucleotides, including hairpins, duplexes, and pseudoknots [38], and, when coupled to a limited number of base-pairs restraints and experimental data such as those coming from Small Angle x-ray Scattering (SAXS) experiments [78], is able to fold structures of up to about 120 nucleotides.

Lastly, at the resolution of 6 or 7 beads per nucleotide, depending on the base species, we have developed the model HiRE-RNA [40, 79, 80] in 3 successive versions. While the earlier versions were able to correctly predict folds of simple hairpins and duplexes, the most recent development allows us to consider molecules of complex architectures and larger sizes, and to fold a 49 nt triple helix pseudoknot from the knowledge of the sequence only [40], as well as a 80 nt riboswitch when three base-pairing constraints are imposed.

In what follows we are going to discuss some key ingredients for building a sensible coarse-grained model, focusing on designing the force-field, on parametrization, and on how to include the interactions specific to nucleic acids and to RNA in particular.

### 4.1. Coarse-grained representation

The first element going into building a coarse-grained model is the choice of the representation, which, in physics terms, starts with determining what the relevant degrees of freedom of the system are for the process under investigation. One needs to define the number and type of elements that are going to constitute the new particles, grains or beads, which are then going to interact via an appropriate force-field. The choice of the beads reflects the degree of resolution adopted and is directly linked to the ability to reconstruct back an atomistic model from the coarse-grained representation. A detailed model, with many beads, is clearly computationally more costly than a simple model with a few beads, therefore the choice of the representation is a trade off between speed and accuracy and depends sensibly on the questions that are to be addressed by the model. Models with rigid bodies or with two or three beads per base have proven useful to study duplex assembly and melting processes, but lack many details necessary for structure prediction. Models with 5 or more particles, thanks to a more accurate representation of the bases, allowing the definition of more properly stacking and base-pairing interactions, are better suited for structure prediction, but require more computational time and can lack enough sampling for the study of thermodynamic properties.

The models mentioned in the previous section show well the wide range of possible choices. NARES-2P defines 2 beads, one spherical for the phosphate and one elliptical for the base, plus a virtual sugar used in the definition of the relative geometries of the phosphate and the base, but that does not participate in the interactions. TIS and iFoldRNA define 3 spherical beads: a phosphate, a sugar, and a base, positioned at the center of mass of the respective groups. Plotkin’s model defines 2 spherical beads, representing the phosphate and sugar groups, and one elliptical bead representing the base. OxRNA defines the nucleotide as a rigid body composed of 5 interaction centers in different locations depending on the kind of interaction considered. Xia’s model defines 5 particles: a phosphate, a sugar, and 3 particles for the base positioned differently according to the base type. HiRE-RNA defines 6 or 7 particles: four in the positions of the backbone’s heavy atoms P, O5’, C5’ and C4’, one on the sugar C1’, and one or two beads in the center of mass of the aromatic rings of the bases.

In figure 3 we give a summary of the bead representations of the different models.
Figure 3. Coarse-grained representation of various models: NARES-2P (A), iFoldRNA and TIS (B), Plotkin’s (C), OxRNA (D), Xia’s (E), HiRE-RNA (F). Beads centered at the phosphates are shown in green, beads representing sugars are in gray, and beads representing bases are in blue. Lines connecting the beads are in brown when the beads are connected through flexible potentials and in green for rigid bodies.

The choice of the representation is strictly coupled to the choice of force-field. For example, it is clear that if we want to include stacking interactions the model needs to have the possibility of defining a plane for the bases through a sufficient number of beads, through ellipsoids, or through an internal reference frame.

4.2. Up or down?

Once the particles of the model have been chosen, one needs to give them properties on how to interact, namely to define a force-field through potentials. The introduction of a force-field is what makes this approach physical, or ab initio, as opposed to a bioinformatic, data-mining, approach. Once the potentials have been defined, the system obeys classical mechanics, with forces computed as spatial derivatives of potentials, accelerations computed through the inertia law, and particle trajectories obtained through integrations over time.

Among ab initio models we can make the distinction between those for which the force-field is built systematically from the integration of the underlying degrees of freedom, called ‘bottom–up’ models, and ‘top–down’ models that, to a varying extent, make use of experimental data to assign parameters of a potential assigned a priori or to derive statistical potentials all together.

4.2.1. Bottom–up. Bottom–up potentials follow the natural definition of a coarse-grained model, where fast degrees of freedom are integrated over and included in the interactions of the slower variables. Features of these potentials are usually extracted from long atomistic simulations. Two examples of bottom–up potentials are NARES-2P and the model by Plotkin.

In NARES-2P the interaction between bases and phosphates is described through four local interaction energies—bond-stretching harmonic potential, an angle bending sinusoidal potential, a torsion sinusoidal potential, and a sugar-base rotameric potential—and non-local interactions between bases and phosphates. The local energy terms were fitted to the Boltzmann inversion of the respective distributions obtained from the PDB structures of several dozen DNA and RNA molecules through a common structure-based optimization procedure. Non-local terms include a base–base Gay–Berne potential accounting for close contact repulsion and long-range attraction of nonspherical beads [72], a dipolar base–base electrostatic interaction, a base-phosphate and phosphate–phosphate Lennard–Jones potential, and a Debye–Huckel electrostatic potential between phosphates. To describe the anisotropy of the beads representing the bases, the analytical expression of base–base interactions are rather articulate and involve 11 parameters for each one of the allowed 15 base pairs. Both non-local base–base interactions were parametrized fitting potentials of mean force (PMF) computed by the numerical integration of AMBER energy surfaces, through a systematic averaging over the degrees of freedom not represented in the coarse-grained model. To make the calculation feasible, integration was performed on a grid. Potentials derived with this procedure were shown to still represent well the directionality of the potentials computed with AMBER for atomistic structures, i.e. of the potential prior to integration. Some parameters were left free from the PMF fitting procedure and were then adjusted based on the nearest-neighbor parameters of Santa Lucia’s HyTher model [81] in order to reproduce the thermodynamics of DNA folding.

The interaction between phosphates depends significantly on the environment surrounding the charges as it is screened by water molecules and counter ions in solution. Phosphate–phosphate interaction energies were derived from an umbrella sampling atomistic AMBER simulation of two phosphate ions in a TIP3P water box and counter ions. A potential of mean force was then extracted averaging over the degrees of freedom of water molecules and counter-ions.

Another model built from a bottom–up procedure is Plotkin’s DNA model. Its potential is composed of local bond, angle, and dihedral terms, electrostatic interaction between
phosphates, non-local base–base, base-residue, and residue–residue interactions, (where a residue can be either a sugar or a phosphate) accounting for the elliptical shape given to the bases, and base–base hydrogen bonds for base-pairing. The functional form adopted for Van der Waals base–base interactions is a modification of the Gay–Berne potential called RE$^2$ potential [82, 83] with 14 parameters for the ten possible base–base interactions. These parameters are determined by fitting RE$^2$ and the all-atom molecular mechanics force field MM3 [84] with a Buckingham exponential-6 potentials for long-range interactions. The same procedure is adopted for the optimization of base-sugar or base-phosphate parameters for which the interaction potential is the limiting case of one ellipsoid interacting with a sphere, while the sugar–sugar and sugar–phosphate Van der Waals interactions are described by a Lennard–Jones potential with pre-fixed equilibrium distance and depth, to prevent steric overlap of the backbone particles. Base-pairing is not included in the RE$^2$ potential and it is described by a phenomenological 12–10 LJ and sinusoidal angular dependence potential, where the angles account for the relative orientations of the two interacting ellipsoids and are defined through their normal vectors. Geometric parameters parameters are pre-set from the specific configurations of base-pairs and the maximal energy of the pair (bottom of the potential well) is also given a priori based on energy calculations in vacuo and on the experimental evidence of the role of hydration on hydrogen bonds. Electrostatic interactions between phosphates are described by a Debye–Huckel potential with parameters given to represent a system immersed in water and at a fixed ion concentration (200 nM), i.e. fixed screening length. For all local potentials no functional form is assumed a priori, but the forms and parameters of the potential are extracted from equilibrium all atom simulations. In order to obtain results that do not include the effects of other interactions for which the functional form of the potentials are imposed from the start, a modified system with no base–base interactions and minimal Coulombic interactions was designed. The modified system is simulated for 250 ns using CHARMM27 parameter set, in explicit water and with counter ions. The long simulations are required to ensure a good convergence of the extracted potentials. It is interesting to note that potentials derived according to this procedure can adopt significantly different forms from commonly assumed phenomenological potentials between the same set of atoms. For example, of the 11 different bond angles in the coarse-grained model, 5 were fitted by harmonic potentials, while the remaining 6 were better fitted by a double well potential.

As these two examples clearly show, bottom–up potentials still require the assumption of the functional form of most interactions, but the coarse-grained parameters are derived through fitting the underlying atomistic potentials and are obtained either directly through the integration of the atomistic force-field, or indirectly, through PMF computed from atomistic simulations. Bottom–up models find their limits on the validity of the underlying atomistic force fields. As discussed in section 2, this could turn out to be a severe limitation, as for the time being atomistic potentials, even in their most refined version, are known to misrepresent some interactions, such as those occurring between phosphates and bases [85] and the account of ions in solution.

4.2.2. Top–down. Top–down models are effective theories where the phenomenological interactions between the particles of the system are given a priori based on physical intuition and the parameters are obtained through a systematic procedure exploiting different kinds of experimental evidence. Three examples of top–down models are OxRNA, Xia’s model, and HiRE-RNA, all defining different force fields and optimization strategies. OxRNA parameters are fitted to thermodynamic data, while both HiRE-RNA and Xia’s model are optimized based on the structural information of the NDB.

An OxRNA force-field is composed of a backbone interaction term, modeled by a finitely-extendible nonlinear elastic potential, an excluded volume term modeled by a Lennard–Jones potential, and a stacking term and a base-pairing term both modeled by a Morse potential, with the stacking term also including an explicit linear dependence with the temperature, and the coaxial and cross stacking terms both modeled by harmonic potentials. In the absence of sufficient experimental thermodynamic data for small molecules, predictions made with Turner’s nearest-neighbor model (NN-model) were used to derive the melting temperatures for a large set of small RNAs containing different motifs. The fitting of the interaction strengths was done by simulated annealing to find a parameter set minimizing the differences between the melting temperatures calculated via the NN-model and those extracted from OxRNA simulations on the same systems. The OxRNA was first optimized to reproduce the melting temperatures of the structures including only canonical pairs to obtain an averaged parameter set independent of the sequence specificity. In a second stage sequence specificity was introduced for the Watson–Crick and wobble base pairs and optimized to fit the melting temperatures of a large set of short sequences forming hairpins or duplexes.

The Hamiltonian of Xia’s model is composed of a set of bonded terms, including bond stretching, angle bending, and dihedral energy, and a non-bonded effective potential inclusive of both Van der Waals and electrostatic contributions, modeled through a Buckingham potential [78]. Local interaction potentials were derived directly from the Boltzmann inversion of variables distributions obtained from 668 3D structures containing more than five base pairs, resulting in the usual harmonic functions for bond lengths and angle bending, and sinusoidal form for dihedrals. Non-bonded parameters were fitted to reproduce global energy minima and later refined to minimize the difference between energy-minimized coarse-grained structures and their corresponding experimental structures. The parameter set was then validated through the comparison of coarse-grained simulations and an atomistic simulation on a set of 15 molecules.

HiRE-RNA’s potential is composed of local harmonic terms for bond angle stretching, sinusoidal energy for dihedrals, excluded volume, Debye–Huckel electrostatic energy, and specifically designed stacking and base-pairing terms taking into account base orientations [40]. The model
has geometric parameters whose values have been determined from distributions extracted from 200 NDB structures including molecules of varying sizes and topologies; overall energetic parameters, representing the relative weights of the different interaction terms, which are subject to an optimization procedure; and base-pairing energetic parameters, which for the time being are assigned from the start based on the number of hydrogen bonds of the contact, and no longer modified. The optimization procedure is done through a genetic algorithm to find the parameters that better distinguish energetically native structures from decoys [86]. For each structure of a training set we generated 20 decoys including low-energy and high-energy structures. Low-energy decoys were chosen to evenly cover four possible scenarios of high or low rmsd and high or low base-pairing similarity with respect to the native structure, with the goal of covering extensively the different possible conformations adopted by a given sequence. The algorithm mimics an evolutionary process in which the vectors containing the energetic parameters undergo mutations and swapping to obtain a combination of parameters that maximizes the energy difference between the native structures and all the decoys. To optimize with a genetic algorithm, the choice of training set is also important. Since our goal is to have a model that is able to follow a molecule’s large conformational changes, we want to have parameters that allow all possibilities, and that are not biased toward some specific conformations. In particular, for RNA, the risk is to have parameter sets highly favoring helices, given that they are by far the most common structural element in the NDB. We therefore used the concepts of RNA graphs to build a structure database rich in different topologies [29, 87] since this descriptor captures well the different overall organization of the molecule’s structure. From the RAG database [88], we have chosen an equal number of representative structures for each populated topology to be part of our training set. The parameters obtained with this procedure were then tested through long MD simulations on systems of various sizes and showed a significant improvement over the previous parameters calibrated by hand.

Top–down models require the optimization of many parameters, a task that, depending on the detail of the force-field, can quickly become as challenging as parameterizing an atomistic force field. These models rely intrinsically on the availability of experimental data with a direct correspondence to quantities that can be extracted from the model, such as melting temperatures and spatial variable distributions. As we saw for OXRNA, even though in principle melting temperatures are experimentally accessible, in practice for a well-grounded optimization one needs information on so many molecules that the only viable route is to randomly generate structures to simulate and use thermodynamic models to compute their melting temperatures.

Geometric distributions are readily accessible from the NDB, but the choice of the structures used to compute them is critical. As is the case for bioinformatic prediction models, the risk is to bias configurations toward the double helix, given the large majority of nucleic acid structures in the NDB are of this form.

Much harder is assigning relative weights to sequence dependent base-pairing and stacking. Energetic information on these two terms can only be inferred indirectly from thermodynamic data and single molecule pulling experiments, where the contribution of different energy terms cannot be easily disentangled. Base-pairing and base-stacking energies can in principle be computed by quantum mechanics calculations, but for the time being these data are available only for bases in vacuum and in gas phase [89, 90], and it is unclear how they transfer to the context of a molecule under physiological conditions. Assigning base-pairs relative weights is the main difficulty in the parametrization of HRE-RNA, where bases can interact on their three sides in different positions, for a total of 28 possible different pairings. Each pair contains one, two, or three hydrogen bonds. The choice we have made for now has been to give a pair a weight proportional to the number of hydrogen bonds formed, but this is clearly in contrast to the observed overwhelming abundance of GC and AU canonical pairs in the NDB and to the results of the QM calculations. Indeed, we found that to better account for structures we needed to artificially modify these parameters, giving a slightly higher weight to canonical pairs over all the others. From a physical stand point, such a difference could derive from a cooperativity effect of hydrogen bonds. It is to be noted that this problem affects atomistic models as well, and therefore bottom–up models are subject to the same uncertainties.

As it emerges from this brief discussion, parametrization is possibly the main challenge in the development of a force field. Recently, methods have been proposed to consistently integrate both experimental and theoretical data in an automatic parameter optimization procedure [91], and they have been applied as a test case to the parametrization of different water models. So far, these methods are highly expensive, and their efficiency has yet to be demonstrated on more complex models, such as the ones used for protein and RNA.

4.3. Flat or round?

The specificity of nucleic acid interactions in folding is given by base-stacking and hydrogen bonding to give base pairs. Both interactions are dependent on the flat shape of the aromatic rings of the base. Typical coarse-grained models of biomolecules developed in the past for proteins and for lipids represent the newly defined particles (grains) as isotropic spheres. It is the case of the popular Martini force field [92], and of the protein model OPEP [93], that we also develop, for which both the backbone heavy atoms and side chains are represented by spheres of an appropriate size. Such an approach find its reasons in a model that is at low resolution, for which interactions are taken as isotropic and long-ranged, often modeled by Lennard–Jones potentials. The first coarse-grained models developed for RNA also adopted this description. For example, both iFoldRNA and TIS describe a nucleotide as composed of three spherical beads. The specificity of base-pairing and of helix formation are integrated into iFoldRNA by giving a set of distance constraints to the base beads, including both same strand and on the cross strand terms. In TIS it is the Go-like potential that drives the molecule to the native base-pairing and stacking. In neither of
these models are the bases really free in their interactions, in one model because they are constrained by geometry, in the other because they are biased toward the native structure.

If we want to model base behavior realistically, we need to look closer at the base, adopting a relatively high resolution that takes into account the anisotropy of stacking and hydrogen bonding. The more recent nucleic acid coarse-grained models adopt different strategies to take into account base planarity and orientation, going from an ellipsoid base representation (Plotkin and NARES-2P), to introducing an internal reference frame and several interaction sites (OxRNA), or to explicitly defining planes thanks to the high resolution description (HiRE-RNA).

In NARES-2P the bases are subject to an electrostatic dipole–dipole interaction and a non-bonded interaction, including excluded volume and long-range Van der Waals. Both are modeled through anisotropic potentials, with the dipole interaction dependent on the angle of the bases with respect to the distance vector (figure 4(a)) and the long-range potential modeled through a Gay–Berne potential, under which the bases perceive each other as ellipsoids. The interaction between the bases and phosphates is on the other hand modeled through a standard LJ potential, as if the bases were spherical. In the model by Plotkin, the bases are always subject to a through a standard LJ potential, as if the bases were spherical. In the model by Plotkin, the bases are always subject to a

In HiRE-RNA we take advantage of the high resolution of the base representation to define base planes and we construct potentials using the norm vectors perpendicular to the plane. Base-pairing potential is composed of the product of a hydrogen bonding potential, depending on the distance and relative angles of the interacting particles (the base extremities), and a planarity term, where the co-planarity of the bases is implemented through a short-range inverted Gaussian potential dependent on the distance of the particles of one base with respect to the plane defined by the other base (figure 4(d)).

Stacking is also dependent on norm vectors both for base orientation, with a preferred parallel orientation for stacked bases, and to ensure that the stacked bases are coaxial. Our method has the advantage of keeping a spherical particle description and to allow us to easily make the distinctions of the different sides of the base and to define sideways base-pairs, that, as we will discuss in the next section, allows us to define non-canonical and multiple base pairs. The drawbacks are in need of a finer description, and therefore less computational speed, since we need 3 beads for each base to properly define a plane, and in the functional forms of the potentials that become multi-body, with the planarity term requiring the contribution of 6 different beads. Moreover, there are no standard potentials to describe the interactions we want to model, and the functions we have introduced are very much empirical.

4.4. Watson–Crick or non-canonical?

A detailed analysis of RNA structures has shown that there exists of the order of one hundred possible base-pairings between RNA bases since bases have in principle the ability to form hydrogen bonds on all their different sides [94]. Following the classification of base sides introduced by Westhof and Leontis as Watson–Crick (WC), Hoogsteen (H), and Sugar (S), interactions between bases are found to involve all combinations of sides. WC–WC base pairs are the most
common, respecting the canonical DNA pairing scheme A·U, G·C, but all the other pairings of all the sides with each other and of all the bases with each other are also found. Even adopting a simplified view where only one possible pair is formed on each side of the base, considering all 9 possible side-side pairs (WC—WC, WC—H, WC—S, H—H, H—WC, H—S, S—S, S—WC, S—H) in the cys and trans conformation, for all the 12 possible combinations of base kind (GG, GA, GC, GU, AA, AG, AC, AU, CC, CG, CA, CU, UU, UG, UA, UC), we can count over 200 different possible base-pairs. These non-canonical interactions are especially relevant for single stranded molecules that do not have a WC complementary strand immediately accessible, and are therefore specific to RNA (and to ssDNA).

Only a few prediction models take into account the possibility of forming non-canonical pairs. As we have seen in section 3, MC-Fold and FARNA include the possibility of forming non-canonical pairs, but of the ab initio coarse-grained models, most lack the level of detail necessary to describe the base sides.

Only the model by Xia and HiRE-RNA can make the distinction between interactions occurring on different sides of the base. In the model by Xia 3 particles are used to define a base, forming a triangle, with each bead corresponding to a side of the base. The model includes 14 possible nominal pairings, corresponding to all possible base-pairs for the 4 bases and 3 sides, without considering trans and cys conformations, with average distance parameters. Base-pairing is described as part of a generic long-range interaction between beads and depends only on the relative distance between the particles. Constructing pairs based solely on the information of the side on which they occur, implies averaging over many possible different interactions occurring between the two sides, which can vary significantly in equilibrium distances, angles, and strengths, depending on how many hydrogen bonds are formed simultaneously. With a more detailed approach, HiRE-RNA, in its current version, includes 28 different possible interactions occurring on all sides, each associated with a specific set of distances, angles, torsions and number of hydrogen bonds formed. The choice of 28 interactions is quite arbitrary and can be extended to any number of interactions as long as they are sufficiently distinct in interaction centers. For now pairs have been chosen based on their abundance in the NDB, making sure to have at least two or three representatives for each letter pair. For some letter pairs we can account for two distinct interaction sites occurring between the same sides at different geometric centers (figure 5(a)).

The potential energy is given by a narrow inverted Gaussian around the geometric center, and for any given letter pair, we simply add over all possible centers (figure 5(b)). Because of the excluded volumes of the beads, effectively there can only be three interaction centers simultaneously present around a base, one on each side (figure 5(c)). Despite the fact that HiRE-RNA considers at the moment fewer possible interactions than Xia’s model, coupling hydrogen-bonding with planarity allows the capture of fine structural details of base-pairing which can then have a large repercussion on the overall conformation of the molecule, as it is for the formation of triplets and quadruplets.

5. Simulation methods

Ab initio models constitute a physical description of the system. Their natural application is for Molecular Dynamic simulations (MD), where the system is subject to Newton’s equations. If the goal is to observe phenomena occurring on long time scales such as folding, for molecules of the size of most RNAs, simple MD at a fixed temperature is often not sufficient even with the sensible reduction in degrees of freedom of the CG models. In addition to MD, enhanced sampling techniques are commonly used. In this section we are going to review the most prominent enhanced sampling techniques that have been used with the various CG RNA models, with particular attention to Parallel Tempering (Replica Exchange MD—REMD) and Simulated Tempering (ST) [95], both employed by HiRE-RNA, the Discrete Molecular Dynamic technique employed by iFoldRNA, and interactive simulations, an innovative technique, that we are currently testing in combination with HiRE-RNA. Other simulation techniques, such as Monte Carlo and simulated annealing, are also commonly used for structure predictions, but we will limit the discussion here to MD enhanced sampling variants, more naturally linked to the folding process.

5.1. Parallel tempering / replica exchange

In parallel tempering molecular dynamics (PTMD), a number of MD simulations are run concurrently, with different values of control parameter. During the simulation, exchanges of configurations between the replicas are attempted, according to a chosen protocol.

The most common protocol, Temperature Replica Exchange Molecular Dynamics (T-REMD), uses replicas simulated at different temperatures. The attempted exchange between neighboring replicas (figure 6) is done at fixed time intervals and must obey the balance condition in order for each replica to sample the correct ensemble. The

![Figure 5](image-url). (a): Two distinct AC base pairs included in HiRE-RNA, both occurring as cys WC—WC, according to Leontis’ classification, and occurring in the NDB with similar probability among the most frequent AC base pairs. (b): Different possible interaction centers for the interaction between two bases in HiRE-RNA. (c): schematic formation of a base quadruplet in HiRE-RNA. All interactions occurring in natural RNA G-quadruplexes are included in our model.
5.2. Simulated tempering

Simulated Tempering (ST) is a simulation technique that enhances sampling by raising and lowering the temperature sequentially in time. The temperature becomes a dynamical variable, taking values in a discrete range $T_1 < T_2 < \cdots < T_N$. The exchange between temperatures is governed by weights that need to be assigned at the beginning of the simulation to ensure a uniform random walk in temperature space. With the correct weights ST has a higher acceptance ratio than PT [100, 101]; however, given the weights depend on the Helmholtz free energies at each temperature [102], determining them \textit{a priori} is problematic. Recently introduced on-the-fly weights determinations, however, allow us to obtain the weights automatically, greatly simplifying the use of ST [95] (an example is provided in figure 7).

Each ST simulation being independent, it can easily be parallelized at little additional cost, making it ideal to run on a large number of CPUs, allowing for faster data gathering. ST can also be readily generalized to a random walk with several parameters [103], without requiring more than a single simulation, while a similar PT simulation would require an exponentially larger number of replicas.

5.3. Discrete molecular dynamics

Discrete molecular dynamics uses a simplified representation of the energy function, replacing it by discrete step functions. Instead of using the derivative of the energy function to integrate Newton’s equation of motion, a collision detection algorithm based on the ballistic motions of the particles is adopted. Atoms move freely until they collide, and since collisions are purely local, only the nearest neighbors need to be considered for possible collisions. Since only the particles involved in a collision need to be considered, fewer updates are required [104].

This method largely decreases the computational cost compared with MD: the potential functions are much simpler, thus less costly to calculate, and derivatives are not needed. In traditional MD, the evaluation of the energy and force derivatives is the main computational bottleneck. Although better approximations of the original function can be constructed by reducing the step size used when discretizing the potential, this results in an increase in computational cost, until it reaches that of classical MD.

5.4. Interactive simulations

Interactive simulations applied to macromolecular manipulation is now an active field of research [105]. One recent application has been for fitting models into experimentally determined envelopes [106, 107].

Interactive simulations are built on the idea of rendering accurate molecular models real and tangible for scientists. The modest technical requirements allow the setting up of an interactive simulation session on a small laptop computer, simply controlled by a touchpad or a mouse. By including the possibility of interacting directly with the simulation, it is possible to model structural changes in a very intuitive way, and to probe the stability of structures by directly perturbing the structure [108]. This type of approach has seen great success with the emergence of game software challenging players to fold structures by hand. With FoldIt [109], players put in a great performance in predicting protein folds, and were able to
solve a novel protein structure [110]. The strategies used by the top players were shown to outperform the best prediction algorithms published so far [111], and the success of FoldIt inspired other similar projects, notably the EteRNA game for 2D RNA structure prediction [112].

The coarse-grained representation is the natural partner for virtual interactive experiments as it represents an excellent compromise between simulation speed and biological fidelity. Moreover, CG models are in general more robust with respect to user interactions than computations carried out at an atomistic level.

In the context of structure predictions and folding, CG interactive simulations exploit human creativity and constitute a technique complementary to massive computing. While a regular MD of the system runs in the background, guided by our knowledge of the system, we can generate by manipulation very different new conformations that a computer’s calculations may never reach in a finite time. These conformations can then be explored thoroughly and extensively by the enhanced simulations techniques presented earlier. We have recently started to investigate this approach with HiRE-RNA, coupling the model and its simulation engine with the MDDriver software [113] that allows us to guide simulations interactively. By connecting to a network socket, any device with a driver implementing the IMD protocol can connect to the running HiRE-RNA simulation, and the inject user forces the simulations to be altered. We used both the VMD [114] and UnityMol [105] programs to drive our simulations, using either a mouse or a haptic device. Interactive simulations allow us to easily fold and unfold hairpins [93] (see figure 8) and to probe the stability of more complex structures. The software we developed is currently also being used as a teaching tool on university courses. While the students benefit from the virtual reality experience of manipulating a molecular structure, we test the ability of interactive HiRE-RNA to solve folding problems by creativity.

6. Benchmarking results

In this section we are going to discuss some of the results achieved by the various prediction methods, trying to draw comparisons where possible. We will start by discussing the 3D prediction competition set up by E. Westhof in 2012, to which bioinformatic methods mentioned in section 3 participated, together with the \textit{ab initio} method iFoldRNA. The most recent \textit{ab initio} methods did not take part in the
Figure 9. Structures of the three challenges of the RNA puzzle 2012. The native structure is in blue and the best prediction is in yellow.

competition and are harder to compare on specific systems given the novelty of their codes. We will illustrate the results from their most recent publications, gathering similar systems.

6.1. RNA-puzzles

In 2012 the first RNA puzzle competition was launched to benchmark prediction tools [53], in the same spirit as what is being done for proteins in the CASP competition [115]. Different research groups attempt to predict the RNA structures of molecules for which the experimental structure has been determined, but not yet published.

The sequences of three RNA structures were provided as a challenge. The first structure was a dimer (PDB ID 3MEI [116]) with symmetric sequences for the two strands, but for which the crystallized structure displayed two asymmetrical internal loops. The second structure was a square composed of four duplexes, each with the same inner and outer strands (PDB ID 3P59 [117]). The 3D coordinates of the inner strands were provided, leaving the outer strands to be predicted. The third structure was a riboswitch with a three-way junction, PDB ID 3OWI [118]. The sequence of the crystallized structure had been modified at one loop, compared with the sequence given to the contestants.

Other than the methods already discussed (Vfold, FARNA, MCFold, iFoldRNA), three other groups took part in the competition, for a total of seven participants. The group by Bujnicki used the ModeRNA programs, based on sequence homology, to obtain a tentative structure from the known structures of a similar sequence content, and SimRNA to refine the structure through a 3-particle coarse-grained model and inverse Boltzmann potentials [119]. The group by Flores used the program RNABuilder performing Molecular Dynamics simulations in internal coordinates and a rigidification of the parts of the molecule [34]. Their force field consists of torques that act to fold the molecule according to the restraints specified by the users and by stacking, which is the only interaction always present. Information for the restraints comes from experimental evidence, including sequence homology. The group by Santa Lucia used the de novo modeling module RNA123. In this approach the secondary structure is predicted first and is decomposed into constituent motifs such as internal loops, helices, and hairpins, and the 3D structure is assembled by putting together fragments from a motif library.

A complete discussion of the results can be found in the original RNA puzzle paper. The prediction for the duplexes are very close to the crystal structures for all methods. The base pairs found in the experimental structure are recovered with high efficiency, which is not surprising since all the base pairs in the dimer are canonical (see also figure 9). Of all the three challenges, the best prediction was obtained for the square, where the coordinates of one strand were given as the input data. As expected, the helical regions were better predicted than the loops. The limits of all the methods appeared clearly in the prediction of the riboswitch, for which all the proposed structures deviated significantly from the experimental conformation. From a detailed analysis of the different kinds of interactions (canonical base-pairing, non-canonical, stacking), what appears to be most challenging is to predict the formation of non-canonical pairs. Only FARNA/FARFAR was able to recover a percentage (up to 60% depending on the proposed structure) of the non-canonical pairs, but the overall shape of the molecule was missed by the local reconstruction. The method that was able to better capture the overall shape of the molecule was VFold, which on the other hand missed completely the prediction of the non-canonical pairs.

The results of the RNA puzzle competition highlight how for small, simple RNA molecules we now have several well-performing methods to predict the three-dimensional structure. However, these methods only give access to a static picture of the structure and do not give insight into the dynamic of folding, or into the thermodynamics, and do not account for the influence of the external conditions such as ionic conditions or the possible presence of ligands. When challenged with RNA molecules of a larger size and more complex architectures, such as the riboswitch, most methods gave predictions far from the experimental structure and for the most part did not recover the specific base-pairing network necessary for the molecule to hold its shape.

More recently, Xia tested his model with the same challenges posed by the RNA puzzle competition. Coupling their model with secondary structure prediction algorithms providing 14 base-pairs, they were able to fold the riboswitch into the correct topology with an RMSD of 7 Å [78]. This is an important indication that the new generation ab initio
model, representing more realistically base-pairing and with the possibility of forming multiple pairs, is beginning to have prediction capabilities comparable to bioinformatic methods, with the added value of also providing information on the dynamics and thermodynamics of the molecule.

6.2. Pseudoknots

Predicting small pseudoknots is particularly challenging because of the tight configuration adopted by the molecule. Small pseudoknots are stabilized extensively by stacking interactions, and base-pairing alone is not sufficient to hold them together. Both OxRNA and HiRE-RNA, in its latest version, have been able to fold simple pseudoknot topology varying in length from 22 to 34 nucleotides. It is interesting to notice that while the previous versions of our model were not able to predict such structures, the more detailed description of hydrogen bonding and stacking introduced in HiRE-RNA v3 makes it possible to fold tight pseudoknots as well [40]. OxRNA studied the thermodynamics of the experimentally well-documented MMTV pseudoknot, composed of two WC stems of 6 base pairs and 5 base pairs, respectively. Their simulations recovered the double peaked melting temperature also found experimentally [39], which may possibly correspond to a higher temperature transition from an unstructured strand to a hairpin, and a lower temperature transition from a hairpin to a pseudoknot. Our investigation of the smaller 2G1W also supports this behavior. Starting from completely unfolded structures, REMD simulations predict the experimental fold as the most stable structure at room temperature, and the corresponding specific heat curves also exhibit two separate peaks (figure 10). Even though it was not possible at this stage to characterize the intermediate state between unfolded and pseudoknot, ST simulations showed the presence of non-negligible populations of distorted hairpins.

To highlight the importance of non-canonical pairings, we repeated REMD simulations on 2G1W allowing only interactions between Watson–Crick sides. The molecule is still able to reach the native state, but a large number of partially folded or misfolded states are missing, pointing to the importance of considering non-canonical pairs even for molecules in which the native state consists purely of canonical base pairs. Preliminary investigations seem to suggest that the possibility of forming transient non-canonical pairs opens up new folding pathways, more easily interconnecting largely different conformations involving rearrangements of the secondary structure. This could be an important general
Figure 11. Conformational states and transition for 2K96 during a folding simulation. The molecule starts in a completely unfolded state (U) and rapidly transitions to a state where the WC helix is formed (H1). At this stage the molecule can become trapped in misfolded states with alternative base pairings such as T1. On a longer time scale the molecule transitions to a state where the pseudoknot is formed with two WC helices (H2) and can become trapped in configurations where the dangling end folds back on itself, such as T2. Finally, the remaining unpaired strand enters the groove of the first helix making triple contacts and giving rise to the full triple helix.

feature of RNA, since many RNA molecules are known to be able to adopt structures with different architectures in response to the cellular environment and signals.

6.2.1. Triple helix folding. HiRE-RNA was able to fold pseudoknots beyond the simple topology. Indeed it was able to fold the triple helix of the telomerase RNP complex (2K96) [120], a compact structure exhibiting several multiple A triplets, from extended configuration, and without any additional input other than the sequence [40]. Folding was achieved through Replica Exchange Molecular Dynamic simulations using 64 replicas in temperature, ranging from 250 K to 504 K, using a Langevin thermostat and an integration step of 4fs. 2K96 is a 49 nucleotide-long structure composed of a WC hairpin with a dangling end that folds back on itself, inserting into the groove of the hairpin forming triple pairs, and locking into position by base-pairings with the hairpin loop. In REMD, the folding of the triple helix occurred in two steps. In a short first phase the hairpin is formed, while the second phase is characterized by the insertion of the dangling end into the groove. The correct topology was formed at the lower temperatures, after 600 ns REMD time with the molecule reaching a native RMSD of 7–8 Å. After 1.2 μs, the full NMR base pair network was reached and the RMSD lowered to 4.1 Å with respect to the NMR structure (figure 11).

To our knowledge HiRE-RNA is the only model that was able to fold a structure heavily stabilized by multiple base pairs, a result that is even more significant considering no base-pairing biases or other constraints have been used. This is an encouraging result for all coarse-grained ab initio models, as it shows that when the correct ingredients are put into the effective theory, on one hand the predicative capabilities are high, and on the other the model provides some key elements for our understanding of the folding mechanism.

7. Coupling predictions and experimental data

A good ab initio model should in principle obtain 3D predictions starting from the sequence knowledge only. This is the case for HiRE-RNA and for Xia’s model for relatively small RNA molecules. However, the cost of exploring all possible pairings becomes quickly unmanageable as the system size increases. On the other hand, experimental information is often available on the systems of biological interest, and even if minimal, it can be of great use to help models converge to a sound prediction. As we have seen for bioinformatic models, secondary structure predictions can also be integrated into the prediction pipeline with good results for molecules with simple architectures, for which 2D models are reliable.

Experimental information falls into two categories: local, high-resolution data, which include all the information on the spatial proximity of the parts of the molecules, typically base-pairing, which can be obtained by NOESY resonances from NMR, exposed surface data with SHAPE [121], or pairs obtained by secondary structure predictions [122]; and low-resolution, global data giving information on the shape of the molecule, which is obtained by Small Angle x-ray Scattering (SAXS) [123, 124] and Cryo-EM [125–127]. Different information can be integrated by prediction models in various ways. In bioinformatic models, Vfold for example, local base-pairing data can be used to give a complete secondary structure of the molecule, which is then turned into a 3D scaffold constituting the starting point of a refined prediction. In ab initio models, where the folding process is simulated, base-pairing can be introduced as a set of constraints that are implemented at the same time as the molecule folds. This is the strategy we have adopted with HiRE-RNA to fold a riboswitch, for which we are able to recover the overall topology and secondary structure using only 3 base-pair constraints. Global, low-resolution data are less direct to include into models, but they can easily be used as a filter for the proposed structures.

7.1. Local constraints

In ab initio simulations, where a force field is defined, local constraints can be easily implemented in the form of additional potentials, exerting a force on some particles of the system. Adding this information to a simulation can lead to dramatic improvements in its ability to reach the native state by eliminating or hindering the exploration of a large number of possible conformations now incompatible with the imposed data.

Depending on the functional form of this additional potential, different strategies can be adopted. A soft pairing potential, dying off at long distances, provides additional stability to base-pairing once the pair is formed or nearly formed, but it is not able to drive the molecule to the specific conformation starting from a completely different state. A hard
potential, extending to long range, such as a harmonic potential or a potential having a linear behavior at large distances, is able to drive the molecule to form the requested pairs from unfolded configurations. However, care must be taken in the way this potential is added to the simulation. While the observed base-pairs may well be present in the folded structure, if applied too quickly or too strongly, the constraints can dominate the molecule’s behavior, often resulting in entangled states, where the natural behavior of the unconstrained molecule is violated: forcing the constraints at all times in the simulation may stabilize partially folded states, thereby significantly slowing down the folding process.

With HiRE-RNA we decided to adopt a long-range potential, harmonic over short range and linear beyond a 4 Å cutoff, to impose a set of preassigned base-pairs. The switch to a linear potential at a larger distance allows us to modulate the maximum force present in the system, so that even for large deviations from the target constraint values, the system’s forces can be numerically integrated without any problem. To prevent the system from being locked by the constraints in misfolded intermediates, the potential is modulated in time so that the molecule goes through phases when the constraints are active and phases when the molecule can relax unconstrained to re-establish its natural topology. With HiRE-RNA the use of just a few local constraints dramatically improves the success in folding.

7.1.1. Riboswitch. As we have already seen in the discussion of the RNA puzzle results, folding large structures is challenging because of the complex architectures they can adopt, with many possible alternative secondary structures. We have used constrained REMD simulations to obtain the structures of the 79 nts adenine riboswitch (1Y26) [128]. In its NMR state with an adenine ligand the riboswitch adopts a Y-shape with the two upper stems binding through kissing loops. To test whether we could fold this large molecule, we imposed three secondary structure constraints taken from the experimental structure, one on each of the helices. Imposing the time-dependent potential, both simple MD at 300 K and 64 Temperature REMD with a temperature ranging from 250 K to 540 K, with a Langevin thermostat and 4 fs integration time step, we were able to recover the overall organization of the kissing loops with a RMSD of 7–8 Å (figure 12) [40]. The structure is not yet determined with high accuracy, but our results are comparable with the best results obtained by other techniques for ribozymes of similar size.

How to determine the optimal number of constraints to make the simulation converge quickly to the experimental configuration remains an open question. Our example using HiRE-RNA shows that a very limited number of constraints can suffice to obtain the overall shape of the molecule, but clearly a larger number of constraints could improve the precision if occurring in different parts of the molecule. On the other hand, constraints occurring on the same helix do not provide much additional information in the context of folding since when a base-pair is imposed in a simulation, nearby pairs form spontaneously. Many constraints could, however, trap the molecule in entangled states if not applied carefully. Future work will address the question of what is a good set of local constraints, both in terms of how many are needed, and in terms of their distribution in the global architecture.

7.2. Global constraints

SAXS experiments are now becoming common to investigate biomolecular structures since they have the advantage of leaving the molecule unperturbed in its own environment. An x-ray beam is shone for several minutes through a solution containing the molecule, water, and ions at room temperature, and forward scattering is recorded. The collected data are an intensity curve as a function of exchanged momentum q. Because the molecule is free to move in the solution and it is observed over an extended time, the intensity curve is the result of a spatial and temporal average, which could be deconvoluted to give complete structural information only in the limit of q going to infinity. In practice, only the forward scattering has enough intensity to give a clear signal, limiting useful data only to low q values (forward scattering).

Once the atomic structure of a molecule is known, from high-resolution experiments or as a model from simulations, the SAXS intensity curves can be calculated explicitly through Debye formula [129], involving a double summation of atomistic scattering factors of all the atoms of the molecule and of the solvent surrounding it. For large molecules, as is the case for proteins, RNA, and for protein-RNA complexes, these calculations become rather demanding, especially if one needs to repeat them often, as would be the case for a comparison of a simulation trajectory with experimental data where one would need to compute SAXS curves at regular intervals for different simulated configurations. A way of circumventing this problem comes from the consideration that SAXS data are intrinsically at low-resolution, and it is therefore natural to couple them with coarse-grained models. Debye’s theory
can be generalized to grains, instead of atoms, by averaging the atomistic form factors [130, 131]. This technique has been applied successfully for proteins and it is starting to be investigated for RNA molecules as well [132, 133]. A coarse-grained molecular representation was used to perform the SAXS comparisons of structures predicted through MC-Fold and MC-Sym [124, 131]. Theoretical SAXS curves are computed using a 2-bead model (backbone-base) and compared to experimental intensity curves to filter structures that satisfy the experimental data. An important aspect of this process is the treatment of the solvent, since SAXS intensity curves depend on the contrast of the molecule’s scattering with the solvent’s scattering. In the MC-Fold pipeline, the molecule is hydrated by a layer of dummy water molecules and ions. Applying this method to a tRNA, to the P4–P6 intron domain, and to an RNA dimer, shows that filtering with SAXS data is able to select native-like folds. With a similar approach, Xia and coworkers also set up a pipeline where models proposed by simulations with their coarse-grained 5 particles model were filtered by comparison with SAXS amplitudes after having been converted into atomistic structures.

With HiRE-RNA we want to go one step further than simply filtering the simulation results, and we are testing the possibility of integrating the SAXS experimental data ‘on the fly’ to make the simulation converge toward structures compatibles with the low-resolution data (work in progress). Thanks to the important reduction in the number of particles, the SAXS calculations can be performed often during the course of a simulation without significantly slowing it down. The fit between the SAXS theoretical curves and experimental data is integrated as an additional term in the potential computed only with a preset frequency, with times large enough so that the molecule has adopted a significantly different configuration, which can indeed give rise to a different scattering intensity curve, but short enough so as to lead the simulation to converge to the experimental data. As is the case for MC-Fold, hydration is a critical step. For this purpose we introduce a hydration layer of coarse-grained water molecules all around the RNA. The main steps of integrating the SAXS data into simulations is highlighted in figure 13.

8. Open challenges

The physical description of the RNA folding process and the prediction of equilibrium structures is a field still in its early stages. As was shown in the previous discussion, properly accounting for base-pairing and stacking is challenging when one leaves the atomistic description, but it is a necessary step to study RNA molecules of sensible sizes. Other interactions are undoubtedly relevant for RNA folding and have yet to be integrated into ab initio models. Two main additional challenges come from the interaction of the molecule with the surrounding ions and the interaction of the bases with the phosphate groups.

8.1. Ions

Accounting for ions is a challenge for all simulations, atomistic or coarse-grained. In atomistic simulations, with fixed partial charges, the effect of ions is poorly represented because of the lack of polarization [134]. In coarse-grained models, ions are typically not represented explicitly and their effect is taken into account only indirectly. It is clear that neither approach is satisfactory at this stage [135].

In RNA one can distinguish three different roles of ions. The first long-range role is that of counter-ions, present to reduce the electrostatic repulsion of the phosphate groups. The second role is ion condensation around the molecule, constituting a dynamical layer of additional charge on the molecule’s surface. The third role is that of structural ions, where ions are observed in specific positions...
in crystal structures. Most ab initio models account for an electrostatic interaction between the phosphate groups that can take implicitly into account the role of counter-ions, usually modeled though a Debye–Huckel potential, but lack a description of the other two aspects.

Xia and co-workers have shown the impact of introducing structural ions into a small pseudoknot [78]. For 1L2X the effect of the magnesium ions was accounted for by imposing distance restraints between phosphates known experimentally to be in contact with the ion, imposing and stabilizing the shape of the ion pocket throughout the simulation. Their results showed how the structure can indeed be sensibly ameliorated when the structural ions are kept into account.

With the example of HiRE-RNA we want to investigate the effect of ion condensation, and possibly structural ions, and to show how a simulation in the presence of a few explicit ions differs qualitatively from a simulation where an exclusively implicit ion description is used, underlying the importance of a correct ion representation to understand the structure of RNA and the dynamics of ion interactions [136]. These very preliminary results give us some insights into how ion condensation and structural ions could be considered in ab initio models.

8.1.1. Short-range RNA-ions interactions. Two challenges are associated with short-range RNA-ions interactions. First, the charge density in these areas is high enough that most implicit solvent methods, such as Generalized-Born and Poisson–Boltzmann solvers, exhibit divergent results [137]. Also, these methods are ill-suited to represent accurately the strongly constrained geometry that are caused by these interactions. Mg\(^+\) ions, for instance, are known to have an octahedral first coordination shell, constraining the surrounding RNA backbone when tightly bound in an RNA structure. Second, the precise geometries seen around those ions also suggest that the traditional terms used to represent the non-bonded interactions between particles in a force-field, such as the Lennard–Jones or Coulomb terms, may not be adequate either despite progress in their parametrization [134], since they have uniform value for a given radius distance, and would not encompass the important angular aspect of the interaction.

To study the effects of magnesium ions in our simulations, we started supplementing the HiRE-RNA model with an explicit description of magnesium ion particles.

We started with a classical point-charge description of the charged ion in order to observe which phenomenon would be captured by such a model. The most striking result of these simulations, as illustrated in figure 14, is the strong bending of the backbone induced by the ions. Even at very low concentrations, the addition of ions increases the rate of formation of the local hairpins by at least an order of magnitude, favoring compaction. With a concentration of approximately 0.1 mM, a large increase in the speed of compaction can be seen compared with a simulation without ions (figure 15). This is in agreement with the experimental data, where the presence of magnesium ions greatly improves folding rates, and can shift the equilibrium of RNA molecules towards folded structures [20].

8.2. Base-phosphate interactions

It has recently been recognized that hydrogen bonding between bases and the oxygens of the phosphate groups is a common interaction in RNA molecules of complex architectures [85], for example in the S-turn motif present in the sarcin/ricin loop, a part of the ribosomal RNA that forms an essential binding site for elongation factors [138] (figure 16). Currently no structure prediction method, bioinformatic nor physics-based, explicitly accounts for these interactions, which are also misrepresented by atomistic non-polarizable force fields. A statistical investigation of the NDB, followed by quantum calculations, has shown that each base can form stable interactions with the phosphate group. A classification has been proposed for the most common base-phosphate motifs based on the base’s hydrogen fixed by the phosphate’s oxygen.
For each base several possible fixation point are available and quantum calculations are able to investigate the relative stabilities of these bonds in implicit solvent.

In a model such as HiRE-RNA, where the phosphate group corresponds to a bead, and where the bases are described with sufficient accuracy to distinguish their sides, and therefore to differentiate between possible fixation points, a base-phosphate potential term could be easily introduced. As is the case for all other types of interaction, one would need to find the appropriate form for this potential as well as the relative weight of its contribution with respect to all the other interactions. A statistical analysis of the existing structures could be translated into a statistical potential and quantum mechanics calculations could be used as weights. However, consistent with the development of the other potentials of HiRE-RNA, we are considering designing a phenomenological potential derived directly from quantum calculations, and then adjusting its weight through a comparison with existing base-phosphate motifs.

9. Perspectives

RNA structure prediction is a young field where much is yet to be done. Bioinformatic tools are in continuous evolution to better keep account of the experimental data in their predictions, as the example of MC-Fold and SAXS data clearly shows, but it is physics-based ab initio models that have the largest margins of improvement and where much has yet to be accomplished. High-resolution models such as Xia’s and HiRE-RNA have shown their potential as proof of principle and have now to be put to the test in a variety of real systems.

As a start, existing models need to be used for predictions of systems beyond benchmark molecules. It is only with ‘real life’ tests that we will learn their strengths and their weaknesses. This is an essential step in the development of any model. The fact that physical models make predictions also on the dynamics and thermodynamics of the molecule gives access to the comparison with experimental data other than crystal or NMR structures. Non-structural data contribute to our understanding of the relevant terms and folding helps ameliorate the force-fields.

A second line of development is directed toward ameliorating the description of the system to include critical elements left out at first. A successful model should take into account the critical role of ions, as well as base-phosphate hydrogen bonds, implying that new force-fields will have to be designed and parametrized.

A third and more challenging development is to account for the environment and for the conditions under which folding occurs. Unless we limit ourselves to folding experiments in vitro where the molecule has been completely isolated from its natural context, predicting a functional RNA structure implies understanding, and possibly controlling, all the other processes surrounding it. RNA molecules can interact with ligands, which, for example, for riboswitches can drive the formation of one or the other of the possible structures [139], form complexes and interplay with proteins, fold concurrently with their transcription from DNA (co-transcriptional folding) [140], and are subject to the complex cellular machinery of degradation which filters out misfolded structures. It would be presumptuous at this stage to think that we can tackle all of these questions with the tools currently at our disposal, but there are some points worth considering before becoming overspecialized in making predictions for isolated molecules.

9.1. Co-transcriptional folding

One aspect that we can easily access with simulations is co-transcriptional folding versus free folding. In MD simulations where we have access to the dynamics of folding, we can release the degrees of freedom of the molecule a bit at a time and simultaneously let the unfrozen portion of the molecule fold. Such a procedure can be implemented immediately for any model implying MD simulations, and it is currently under investigation with HiRE-RNA. It is then straightforward to mimic some basic biological processes as, for example, the speed at which the polymerase produces the RNA [141] and which would change the speed of release of the frozen degrees of freedom in a simulation (figure 17). Without adding too much difficulty, one could also go one step further and explicitly include a DNA/RNA double helix with the dangling RNA released over time and simultaneously folding. Such studies will help us understand the context in which folding take place and the relative importance of the different phenomena.

9.2. Post-processing

As is shown by riboswitches, RNA molecules can adopt alternative structures of similar energies separated by high barriers, involving the reorganization of the secondary structure and therefore the formation and breaking of many base-pairs. Such barriers are unlikely to be overcome simply by thermal energies at room temperature. In biological systems, it is more probable that during folding the molecule gets trapped in stable states far from the native structure from which it cannot escape on its own, and what we observe as
Figure 17. Cartoon of co-transcriptional folding. As the RNA molecule (black string) is transcribed from the DNA (red tube) thanks to the action of the polymerase (blue donut), it starts folding. The folded structure is in this case the result of a dynamical process that depends also on the speed (arrow) and pauses of the polymerase along the DNA.

A functional structure is the product of a selection process carried out by RNA degradation factors (figure 18). The degradation machinery is itself a complex system, out of our modeling reach, but coarse-grained models coupled to enhanced sampling techniques can generate a large variety of alternative structures and can shed light on the plurality of possible configurations that can co-exist at different stages of the cell’s life and that can be a pool of new functional structures if the environmental conditions change and the need presents. Models such as HiRE-RNA, Xia’s or OX-RNA, can all give access to large sets of very different structures and investigate their structural properties and relative equilibrium.

9.3. Protein partners

RNA molecules often work in conjunction with proteins in an intricate functional network. If the structures of the RNA and of the protein are known, protein-RNA docking programs can give insight into the possible interactions between the two partners under the assumption that neither the protein nor the RNA substantially modify their conformations from unbound to bound [142]. This is the case for some complexes, but not for all. It is well known that the formation of many complexes involves a substantial rearrangement of the protein, or of the RNA, or of both. Examples of structural reorganizations upon binding in complexes are ubiquitous and go under the name of ‘induced fit’ [143, 144]. Because of the large size of a complex, once again atomistic simulations are not suited for the study of systems where large-scale structural changes are involved, but coarse-grained models could stand up to the task. Currently a fully flexible coarse-grained model including a description for both proteins and nucleic acids and for their interactions does not exist. The widely known Martini force-field [92] can describe simultaneously proteins, lipids, and nucleic acids, but it has to impose secondary structures. It can therefore accommodate for molecular flexibility, but not for induced large-scale rearrangements.

HiRE-RNA has a protein coarse-grained model partner named OPEP [93] that has been successfully used over the last 10 years to address protein folding, in particular in the context of amyloid fibers formation in relation to Alzheimer’s disease. At its present stage OPEP can take into account many features specific to amino acid interactions such as hydrogen bond formation to give rise to secondary structures, salt bridges, and hydrodynamic effects. As a simplified model with only 6 particles per amino acid, it can study large-scale rearrangements of single molecules as well as the interplay of several thousands molecules at once. HiRE-RNA and OPEP have been developed in parallel on the same simulation engine with the goal of one day bringing the two together into one model able to describe the behavior of co-existing proteins and RNA, allowing full flexibility for both. Now that HiRE-RNA is converging into a stable version and has given proof of being able to address RNA folding correctly, we are starting to work on developing an interface force field that will take into account occupied volumes, electrostatic interactions, but also specific interactions between bases and amino-acids, making use of the experience we have developed to treat the details of base interactions.

Acknowledgments

The authors wish to thank the French ANR LABEX DYNAMO ANR-11-LABX-0011 for the equipment support, P Stadlbauer and J Sponer for kindly providing the G-quadruplex image, P Sulc, T Ouldridge and J Doye for providing the images of the MMTV pseudoknot, M Baaden and S Doutreligne for the interactive simulation image, and B Caron and Y Laurin for their contribution in the development of HiRE-RNA during their Master internships.

References

[1] Ban N, Nissen P, Hansen J, Moore P B and Steitz T A 2000 The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution Science 289 905–20
[92] Marrink S J and Tieleman D P 2013 Perspective on the Martini model Chem. Soc. Rev. 42 6801–22
[93] Sterpone F et al 2014 The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems Chem. Soc. Rev. 43 4871–93
[94] Lemieux S and Major F 2002 RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire Nucl. Acids Res. 30 4250–63
[95] Nguyen P H, Okamoto Y, Derreumaux P and Nguyen P H 2013 Communication: Simulated tempering with fast on-the-fly weight determination J. Chem. Phys. 138 061102
[96] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method J. Comput. Chem. 13 1011–21
[97] Chodera J D and Shirts M R 2011 Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced mixing J. Chem. Phys. 135 194110
[98] Earl D J and Deem M W 2005 Parallel tempering: theory, applications, and new perspectives Phys. Chem. Chem. Phys. 7 3910
[99] Buchete N V and Hummer G 2008 Peptide folding kinetics from replica exchange molecular dynamics Phys. Rev. E 77 030902
[100] Mitsutake A and Okamoto Y 2000 Replica-exchange simulated tempering method for simulations of frustrated systems Chem. Phys. Lett. 332 131–8
[101] Park S 2008 Comparison of the serial and parallel algorithms of generalized ensemble simulations: an analytical approach Phys. Rev. E 77 016709
[102] Chelli R 2010 Optimal weights in serial generalized-ensemble simulations J. Chem. Theory Comput. 6 1935–50
[103] Mitsutake A and Okamoto Y 2009 From multidimensional replica-exchange method to multidimensional multicanonical algorithm and simulated tempering Phys. Rev. E 79 1–4
[104] Proctor E A, Ding F and Dokholya n N V 2011 Discrete molecular dynamics Wiley Interdisciplinary Rev.: Comput. Mol. Sci. 1 80–92
[105] Lv Z, Tek A, Da Silva F, Empereur-mot C, Chavent M, Baaden M and Silva F D 2013 Game on, science: How video game technology may help biologists tackle visualization challenges PloS One 8 e57990
[106] Birmans S, Rusu M and Wrighgers W 2011 Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes J. Struct. Biol. 173 428–35
[107] Molza A-E, Ferény N, Czijek M, Le Rumeur E, Hubert J-F, Tek a, Laurent B, Baaden M and Delalande O 2014 Innovative interactive flexible docking method for multi-scale reconstruction elucidates dystrophin molecular assembly Faraday Discuss. 169 45–62
[108] Serles M C, Richardson J S, Richardson D C and Brooks F P 1994 Sculpting proteins interactively: continual energy minimization embedded in a graphical modeling system Protein Sci. 3 198–210
[109] Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z and Players F 2010 Predicting protein structures with a multiplayer online game Nature 466 756–60
[110] Khatib F et al 2011 Crystal structure of a monomeric retroviral protease solved by protein folding game players Nat. Struct. Mol. Biol. 18 1175–7
[111] Khatib F, Cooper S, Tyka M D, Xu K, Makedon I, Popovic Z, Baker D and Players F 2011 Algorithm discovery by protein folding game players Proc. Natl Acad. Sci. USA 108 18949–53
[112] Lee J et al 2014 RNA design rules from a massive open laboratory Proc. Natl Acad. Sci. USA 111 2122–7
[113] Delalande O, Ferey N, Grasseau G and Baaden M 2009 Complex molecular assemblies at hand via interactive simulations J. Comput. Chem. 30 2375–87
[114] Humphrey W, Dalke A and Schulten K 1996 VMD: visual molecular dynamics J. Mol. Graph. 14 33–8, 27–8
[115] Moult J 2005 A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction Curr. Opin. Struct. Biol. 15 285–9
[116] Dibrov S M, McLean J, Parsons J and Hermann T 2011 Self-assembling RNA square Proc. Natl Acad. Sci. USA 108 6405–8
[117] Dibrov S, McLean J and Hermann T 2011 Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA Acta Crystallograph. D 67 97–104
[118] Huang L, Sermanov A and Patel D J 2010 Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch Mol. Cell 40 774–86
[119] Rother K, Rother M, Boniecki M, Puton T, Tomala K, Lukasz P and Bujnickji J M 2012 Template-Based, Template-Free Modeling of RNA 3D Structure: Inspirations from Protein Structure Modeling RNA 3D Structure Analysis, Prediction (Nucleic Acids, Molecular Biology vol 27) ed N Leontis and E Westhof (Berlin: Springer) chapter 5
[120] Kim N K, Zhang Q, Zhou J, Theimer C A, Peterson R D and Feigon J 2008 Solution structure and dynamics of the wild-type pseudoknot of human telomerase RNA J. Mol. Biol. 384 1249–61
[121] Weeks K M 2010 Advances in RNA structure analysis by chemical probing Curr. Opin. Struct. Biol. 20 295–304
[122] Mathews D H, Turner D H and Zuker M 2007 RNA secondary structure prediction Current Protocols in Nucleic Acid Chemistry ed S L Beaucage et al chapter 11 unit 11.2
[123] Putnam C D, Hammel M, Hura G L and Tainer J A 2007 X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution Q. Rev. Biophys. 40 191–285
[124] Parisien M and Major F 2012 Determining RNA three-dimensional structures using low-resolution data J. Struct. Biol. 179 252–60
[125] Mitra K and Frank J 2006 Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps Ann. Rev. Biophys. Biomol. Struct. 35 299–317
[126] Whitford P C, Ahmed A, Yu Y, Hennelly S P, Tama F, Spahn C M T, Onuchic J N and Sanbonmatsu K Y 2011 Excited states of ribosome translocation revealed through integrative molecular modeling Proc. Natl Acad. Sci. USA 108 18943–8
[127] Lindert S, Meiler J and McCammon J A 2013 Iterative molecular dynamics-Rosetta protein structure refinement protocol to improve model quality J. Chem. Theory Comput. 9 3843–7
[128] Sermanov A, Yuan Y R, Pikovskaya O, Polonskaia A, Malinina L, Phan A T, Hobartner C, Micura R, Breaker R R and Patel D J 2004 Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs Chem. Biol. 11 1729–41
[129] Debey P 1915 Zerstreuung von Röntgenstrahlen Ann. Phys. 531 809–23
[130] Harker D 1953 The meaning of the average of [F] over large values of the interplanar spacing Acta Crystallograph. 6 731–6
[131] Yang S, Parisien M, Major F and Roux B 2010 RNA structure determination using SAXS data J. Phys. Chem. B 114 10039–48

[132] Yang S, Park S, Makowski L and Roux B 2009 A rapid coarse residue-based computational method for x-ray solution scattering characterization of protein folds and multiple conformational states of large protein complexes Biophys. J. 96 4449–63

[133] Putnam D K, Lowe E W and Meiler J 2013 Reconstruction of SAXS profiles from protein structures Comput. Struct. Biotechnol. J. 8 e201308006

[134] Allnér O, Nilsson L and Villa A 2012 Magnesium ion–water coordination and exchange in biomolecular simulations J. Chem. Theory Comput. 8 1493–502

[135] Bowman J C, Lenz T K, Hud N V and Williams L D 2012 Cations in charge: magnesium ions in RNA folding and catalysis Curr. Opin. Struct. Biol. 22 262–72

[136] Draper D E 2008 RNA folding: thermodynamic and molecular descriptions of the roles of ions Biophys. J. 95 5489–95

[137] Kirmizialtin S, Silalahi A R J, Elber R and Fenley M O 2012 The ionic atmosphere around A-RNA: Poisson–Boltzmann and molecular dynamics simulations Biophys. J. 102 829–38

[138] Correll C C, Wool I G and Munishkin a 1999 The two faces of the Escherichia coli 23 S rRNA saricin/ricin domain: the structure at 1.11 A resolution J. Mol. Biol. 292 275–87

[139] Reining A, Nozinovic S, Schlepckow K, Buhr F, Fürtig B and Schwalbe H 2013 Three-state mechanism couples ligand and temperature sensing in riboswitches Nature 499 355–9

[140] Zhang G and Ignatova Z 2011 Folding at the birth of the nascent chain: coordinating translation with co-translational folding Curr. Opin. Struct. Biol. 21 25–31

[141] Schroeder R S, Grossberger R, Picher A and Waldsich C 2002 RNA folding in vivo Curr. Opin. Struct. Biol. 12 296–300

[142] Lang P T et al 2009 DOCK 6: combining techniques to model RNA-small molecule complexes RNA 15 1219–30

[143] Williamson J R 2000 Induced fit in RNA-protein recognition Nat. Struct. Biol. 7 834–7

[144] Schwalbe H, Buck J, Fürtig B, Noeske J and Wöhner J 2007 Structures of RNA switches: insight into molecular recognition and tertiary structure Angew. Chem. 46 1212–9