A short guide for molecular dynamics simulations of RNA systems

Yaser Hashem, Pascal Auffinger *

Architecture et Réactivité de l’ARN, Université Louis Pasteur de Strasbourg, IBMC, CNRS, 15 rue René Descartes, 67084 Strasbourg Cedex, France

Abstract

As a result of important methodological advances and of the rapid growth of experimental data, the number of molecular dynamics (MD) simulations related to RNA systems has significantly increased. However, such MD simulations are not straightforward and great care has to be exerted during the setup stage in order to choose the appropriate MD package, force fields and ionic conditions. Furthermore, the choice and a correct evaluation of the main characteristics of the starting structure are primordial for the generation of informative and reliable MD trajectories since experimental structures are not void of inaccuracies and errors. The aim of this review is to provide, through numerous examples, practical guidelines for the setup of MD simulations, the choice of ionic conditions and the detection and correction of experimental inaccuracies in order to start MD simulations of nucleic acid systems under the best auspices.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

With the steady growth of the number of new RNA crystallographic (∼960 by July 2008) and NMR (∼414) structures of sizes ranging from dinucleotides to the about 15000 atoms large ribosomal units [1], there is a clear demand for going beyond strict structural aspects and exploring some of the dynamical features of biomolecules that are associated with folding, catalytic and other essential molecular recognition process. In this respect, single molecule approaches based on fluorescence and force measurement techniques have proven to be extremely useful and have allowed to collect dynamic information at a single base pair step resolution (better than 3.4 Å) on some RNA motors that are involved in transcription, replication or translation steps [2–5].

From the theoretical point of view, single molecule approaches are the rule and not the exception [6] since computational limitations still hinder empirical methods such as molecular dynamics (MD) simulation techniques to investigate the dynamics of more than one biomolecule at a time although some attempts to develop methods allowing parallel sampling of the configurational space through multiple molecular dynamics (MMD) techniques were presented [7–9].

It is the scope of this review to guide the readers setting up efficient simulation protocols and help them avoiding some of the pitfalls they may encounter during such complex theoretical undertakings. The MD techniques described in this review are not really specific (with a few exceptions) to RNA systems and can be used for setting up MD simulations of other biomolecular systems.

This review integrates and significantly extends the content of two less accessible reviews published as book chapters [10,11]. Historical aspects related to molecular dynamics simulations of RNA systems can be found in [12]. Other reviews dealing with molecular dynamics aspects of RNA [6,13] and other biopolymers [14–16] should be of interest to the reader.

2. MD methods

Force field based simulations make use of an analytical empirical potential energy function and a set of parameters to describe the physical properties of a given biomolecular system. In MD simulations, a “possible” time evolution of the chosen system starting from an initial set of coordinates and velocities is calculated by integrating Newton’s classical equations of motion over a predefined timeframe. Descriptions of the most popular integration algorithms can be found in [17–20]. Algorithms designed to accelerate the MD simulation process by using multiple time step integrators are currently under development [21–24].

3. Simulation setups

3.1. Choosing a starting structure

The choice of an appropriate starting structure is a determining step for generating reliable and informative data through MD simulations and one must take great care in checking all its features. Such starting structures are mainly derived from (i) model-built
systems, (ii) low- to high resolution X-ray and (iii) NMR data, or (iv) a combination of pertinent experimentally derived and/or model-built molecular fragments. NMR and crystallographic structures of nucleic acids can be retrieved from the NDB (Nucleic acid DataBase; http://ndbserver.rutgers.edu and the PDB (Protein Data Bank; http://www.rcsb.org/pdb/) [25,26].

### 3.1. Model-built structures

When no experimental structures are available, one is often tempted to use molecular modeling techniques to build a starting model that is subsequently used for initiating MD runs. Continuous progress in our understanding of the rules that preside to the stability and function of complex RNA systems allows us to build globally “correct” structural models of large RNA’s [27] by using appropriate software’s like “Ribosome Builder” [28] or PARADISE (http://paradise-ibmc.u-strasbg.fr/), a drastic extension of the “Sequence to Structure” tool (S2S; [29]). For an overview, see also [30]. However, at the atomic level, these models are generally far from being precise enough to be successfully used as starting structures for MD runs [31]. Nonetheless, since the degree of precision of such model-built structures is strongly correlated to the complexity of the investigated system, models of small RNA fragments such as double helices or stem-loops can still be successfully used for initiating MD runs [32].

#### 3.1.2. X-ray structures

X-ray structures are, by far, the most reliable source of three-dimensional structures. Those with resolution below 2.0 Å provide generally consistent information about the structure of the first and, sometimes, the second solvation shell including important ion binding features. Structures with resolution above 3.0 Å are often relatively inaccurate (at the atomic level) and should be approached with caution.

#### 3.1.3. NMR structures

Due to the lower overall precision of these structures compared to those derived from X-ray diffraction experiments, NMR structures are mostly chosen when no X-ray data are available. Unfortunately, such methods are generally unable to provide data related to the solvation of the simulated structure and especially of its ion binding features.

#### 3.1.4. Mixed structures

Sometimes, when no experimental structures are available, one could be tempted to rely on combining experimental datasets with model-build fragments. This is, nowadays, common practice in three-dimensional modeling activities [27,29,30]. Yet, such mixed structures might comprise very imprecise regions that could severely impact ensuing MD simulation results.

### 3.2. Adding hydrogen atoms

Hydrogen atoms are an integral part of each functional biomolecular system. Yet, only ultra-high resolution X-ray and neutron diffraction structures provide information on the location of hydrogen atoms belonging to the solute and the solvent [33,34]. Unfortunately, such structures are scarce (see the ‘Hydrogen and Hydration DataBase’ or HHDB; http://hhdb01.tokai-sc.jaea.go.jp/HHDB/) [35]. For instance, no neutron diffraction structures of an RNA fragment are currently available. Yet, since MD simulations take into account all atoms present in a given biomolecular system, it is essential, before starting a simulation, to add hydrogen atoms by using: (i) the tools included in available MD packages; (ii) external web services such as MolProbity (http://kinemage.biochem.duke.edu/ [36]); or (iii) visualization programs such as PyMOL (http://pymol.sourceforge.net/) or VMD (http://www.ks.uiuc.edu/Research/vmd/). Note the existence of a revival for “United atom force fields” in which carbon atoms and their bound hydrogen’s are merged into one fictive atom for the need of coarse-grained simulations [37,38]. Adding hydrogen’s is also an essential step for being able to precisely detect steric clashes that might be present in starting structures. This process is straightforward in the vast majority of occurrences with the notable exception of the hydrogen atom belonging to the RNA 2’-OH hydroxyl group. Stereochemical rules delimiting three favored orientations for the 2’-OH group have been derived from MD simulations and can serve as a guide for inferring statistically favored hydrogen atom locations [39].

### 3.3. Checking the starting structures

As stated above, low-resolution X-ray, NMR and model-built structures are often not precise enough at the atomic level to initiate reliable MD runs. Yet, even structures derived from high quality crystallography might conceal various types of errors or inaccuracies that can compromise the reliability of data extrapolated from MD runs. Some of these issues will be detailed next.

#### 3.3.1. Protonation

In specific structural contexts, nucleic acid bases such as adenines or cytosines are protonated. Unfortunately, X-ray data rarely provide direct evidence for their protonation states and the modeller has to derive them from available structural data by using its chemical intuition based on suitable stereochemical knowledge. This may lead to locally inaccurate starting structures. For instance, in a 500 ps long MD simulation of a tRNA [36] anticodon stem-loop based on crystallographic data, a reordering of its tertiary structure was observed [40]. This reordering was first attributed to the absence of Mg²⁺ cations in the model. Later on, it was proposed that an adenine residue belonging to the base triple Ψ13-G22-A46 was protonated and that this difference induced the calculated structural deviations [41]. Nowadays, the prevalence of the participation of protonated residues in the acid–base catalytic mechanism attributed to various ribozymes seems clearly established [42,43] and, thus, it is important to understand how the structural context might affect the pKa of such key residues [44,45] considering that local pKa changes might sometimes induce unforeseen structural transitions. When it is not possible to determine the most appropriate protonation state of a given chemical group, simulations of both, charged and neutral residues, might help solve the issue. Besides nucleic acid residues, protonation states of amino acids and drugs like aminoglycosides [46] have to be clearly defined before initiating an MD run [47]. Note that numerous pKa prediction methods have been developed for ligands and amino acids [48].

Different experimental pH conditions can also alter the structure of a given molecular fragment. For example, the NMR structure of a tRNA [36] anticodon stem-loop obtained at an acidic pH is different from those extracted from tRNA crystal structures essentially because of the formation of a wobble A–C pair [49] instead of a more standard (in this context) bifurcated A–C pair [50]. Thus, care must be taken in choosing a starting structure that reflects the experimental conditions that one wants to simulate.

#### 3.3.2. Checking biomolecular conformations

The correct orientation of biopolymeric residues is sometimes difficult to derive from experimental data. This is true for the HIS, ASN and GLU amino acids [51]. Hence, some groups have designed automated methods based on the generation of “contact maps” (see, e.g. MolProbity [36]; http://kinemage.biochem.duke.edu) for detecting incorrect or at least suspicious orientations of various kinds of residues and, subsequently, suggesting alternative conformations. In the case of HIS residues, the correct orientation
can generally be deduced from simple stereochemical considerations involving the number and type of hydrogen bonds that link these residues to their environment. These methods can also be used to detect misorientations of modified nucleic acid residues like pseudouridines (Ψ) that are present in a large number of wild type RNA structures [52]. In this respect, the PDB proposes the special “ASX” and “GLX” residue names in case the experimental data are not accurate enough for unambiguously allowing distinguishing between the ASP/ASN and the GLU/GLN residues (http://wwrcsb.org/pdb/). Similarly, “UNK” and “A” stand for undetermined residues and atoms, respectively. The code “A” applies generally to the terminal atoms of the ASP and GLU side chains and to the ring atoms of the HIS residues but should also be used for solvent atoms when there is some ambiguity related to their nature. Starting an MD simulation with “wrongly” oriented residues may have serious consequences on the generated trajectories and on ensuing interpretations.

Contact map based methods are also very useful for detecting steric clashes that may be present in experimental structures. For nucleic acids, the respective position of the bases is usually well determined while for the sugar–phosphate backbone, there are often too many degrees of freedom relative to the number of observable data and steric clashes (≥0.4 Å overlap) are not uncommon when hydrogen atoms are taken into account [36,53,54]. Thus, it is important to be aware of the location of such steric clashes and to try to minimize them before starting an MD run. The newly developed RNABC program (http://www.cs.unc.edu/~xwang/RNABC) has precisely for aim to reduce all-atom steric clashes in RNA backbones by proposing alternative conformations [55].

A MD simulation of an RNA/antibiotic complex provides another example of the implications the use of a locally imprecise starting structure can have. A small experimental difference in the resolution of the second binding site of this symmetrical structure led to a detectable dynamical divergence between the two sites that could subsequently be restored by a slight stereochemical adjustment [47].

Evidently, in order to appraise the quality of an RNA structure, it is critical to be aware of the manifold stereochemical rules that preside to their folding, structural stability and functions. These rules comprise knowledge of the various patterns nucleobases [56,57] can form when inserted into higher order structural motifs [27] without forgetting the large underlying structural diversity of the RNA backbone [58–60]. Some statistical data related to the structure of RNA base pairs and their associated backbones are accessible through the SwS web service (http://www-ibmc.u-strasbg.fr/arn/sws.html) [61].

### 3.3.3. Checking solvent density maps

X-ray crystallographic experiments generate electron density maps. Yet, such density maps are void of information related to the nature or chemical type of the atoms at the origins of the diffraction patterns. Hence, crystallographers have to make choices when interpreting density maps. From the attribution point of view, fitting a biopolymer in such maps is relatively easy since one knows beforehand the connectivity of the crystallized molecules. On the other side, it is much more difficult to associate solvent atom types (oxygen atoms of water molecules, cations, anions, . . .) with regions of high electron density and various types of misinterpretations/misattributions can take place. In this respect, a survey of NDB structures revealed that it is not unusual for water molecules to be assigned to sites with large electron densities (above those expected for a localized water molecule). Such large densities are rather indicative of the presence of a heavy metal ion [62]. Lighter metals such as Na⁺ and Mg²⁺ have diffraction properties close to those of water molecules and are particularly difficult to detect on the sole basis of electron density intensities. Fortunately, coordination properties of Mg²⁺ cations and oxygen atoms of water molecules are very different. Usually, six and four electronegative atoms with average contact distances close to 2.1 and 2.6 Å surround Mg²⁺ cations and Ow atoms, respectively. On the other hand, Na⁺ cation and Ow atom coordination distances are similar. Thus, Na⁺ cations are very difficult to associate with specific electron density patterns and it is highly probable that an undefined proportion of water molecules found in crystal structures correspond in reality to undetected Na⁺ cations when these ions are present in the crystallization liquors.

Again, a good knowledge of the coordination properties of water molecules and of the various metals found in the vicinity of nucleic acid structures [63,64] is essential if one wants to avoid using locally incorrect experimental structures. Among others, it is certainly important to be familiar with the fact that anions can intrude into the first solvation shell of nucleic acids. The lack of this key structural knowledge prompted some crystallographers to assign Mg²⁺ cations instead of SO₄²⁻ anions to regions of high electron density [65,66]. In this respect, it is good to note that the Protein Data Bank (PDB) and the Nucleic acid DataBase currently allow the documented deposition of corrected crystallographic structures. For example, the 1JJM and 1JJN structures from the PDB were replaced by the ‘reinterpreted’ 2BBR and 2BBS structures.

Statistical data related to the structure of the first solvation shell around the most common nucleic acid base pairs should contribute to increase our understanding of critical water and ion binding features to nucleic acids (SwS; http://www-ibmc.u-strasbg.fr/arn/sws.html) [61].

### 3.3.4. Are experimental ionic conditions always appropriate?

Experimental structures and especially X-ray structures are, as stated above, often used for initiating MD simulations in aqueous solution. Very frequently such structures include details about the binding of divalent (Mg²⁺, Mn²⁺, Sr²⁺, . . .) cations. Yet, for the purpose of obtaining well diffracting crystals, crystallographers are sometimes inclined to raise the concentration of divalent cations in the mother liquors. Henceforth, many of these experimentally detected cations might not be present and, consequently, functionally relevant in “in vivo” conditions. Thus, the modeler needs to be critical about such experimental information. For instance, five Mg²⁺ cations have been assigned to high electron density regions in the crystal structure of the 5S rRNA loop E motif while other experimental data point to the fact that only one or two Mg²⁺ cations are needed for stabilizing this structural fragment [67,68]. A further example is provided by an hydrolase crystal structure where four Mg²⁺ cations were located close to the active site while it has been inferred through different experimental techniques that only two Mg²⁺ cations were needed for catalysis [69]. The authors concluded that the additional divalent cations detected around the catalytic site might be the result of crystallization artifacts.

### 3.4. Choosing the environment (crystal, liquid, . . .) and ion types

Simulations of biomolecules can be run in various types of environments. Liquid phase is most commonly selected since this environment is close to that used in “in vitro” (NMR) experiments where the biomolecules are highly diluted. Liquid phase conditions are not necessarily close to “in vivo” conditions in which biomolecules are found to evolve in a heavily crowded setting [70]. Crystalline conditions have been chosen in a few occurrences. For RNA, to the best of our knowledge, only the r(GpC) and r(ApU) dinucleotides [71], a few RNA duplexes [72,73], and the full length hammerhead ribozyme [74] crystal cells have been simulated.
The ions used to neutralize the nucleic acid solutes can be of various types (Na\(^+\), K\(^+\), NH\(_4\)+, Mg\(^{2+}\), Mn\(^{2+}\), ...) and different ionic conditions ranging from a number of ions sufficient to neutralize the negative charge carried by the polyanionic backbone to a number of ions that would appropriately represent solutions containing 0.1–1.0 M of added salt can be selected. In the last case co-ions (Cl\(^-\), SO\(_4\)\(^{2-}\), ...) must be placed in the simulation box as well. Note that anions can, in specific conditions and in contrast to common believes, break into nucleic acid first solvation shells [65] and should be included in the modeling scheme.

3.5. Setting the box size and placing the ions

3.5.1. Box size

For simulations of the crystal phase, the box size and shape is determined by the size and shape of the crystal cell. In solution, one can choose between a rectangular parallelepiped, a hexagonal or a truncated octahedron box depending on the solutes shape. The size of the box extends generally by 8–12 Å around the solute, ensuring three to four solvation layers in each direction.

3.5.2. Monovalent ions

Monovalent ions (Na\(^+\), K\(^+\), NH\(_4\)+, Cl\(^-\), ...) are then positioned in the simulation box by replacing the water molecules with the lowest (for cations) and highest (for anions) electrostatic potential [32]. If one wants to avoid biases due to the initial placement of ions, it is also essential to set the following limits: ions should not be too close to the solute and other ions. For this reason, a minimal approach distance of 5 Å is often chosen. In this manner, at the beginning of an MD run, no ions are found in the grooves of the nucleic acid and no contact ion pairs are present.

3.5.3. Divalent ions

The limited length of MD runs does not allow to simulate the dehydration of Mg\(^{2+}\) cations and the formation of Mg\(^{2+}\). O direct contacts that are often observed in crystal structures since such process take place on the microsecond timescale [75,76]. Thus, unless one wants only to investigate the diffusion process of the Mg\(_{2}\)(H\(_2\)O)\(_{6}\)\(^{2+}\) cations, divalent cations cannot be placed in the simulation box as described above. Such systems can only be realistically simulated if the divalent cations occupy their crystallographic location right at the beginning of the MD run [67,77]. Subsequently, during the equilibration phase, water molecules complete the hydration shell of the divalent cations (Mg\(^{2+}\), Mn\(^{2+}\), Ca\(^{2+}\), ...). Hence, it is rarely necessary to take into account crystallographic water positions. An even more realistic and complete set takes into considerations divalent (Mg\(^{2+}\)) and monovalent (K\(^+\)) cations as well as anions (Cl\(^-\)) as in a study of the SS loop E motif [68].

3.5.4. Minimal salt conditions

From an historical point of view, most MD simulations of nucleic acids were performed in “minimal salt” conditions implying the placement of a number of monovalent cations just sufficient for neutralizing the charge carried by the negatively charged nucleic acid backbone. Yet, in “in vivo” and in “in vitro” conditions, cations and anions are present in excess and there is no further justification for using minimal salt conditions if one is careful in choosing appropriate empirical potentials for the ionic particles [77,78]. Furthermore, using a finite ionic concentration (from 0.1 to 1.0 M and above) allows a much better sampling of the configurational space accessible to these solvent particles.

3.6. Choosing the MD packages and force fields

3.6.1. MD packages

Several MD packages like AMBER (http://amber.scripps.edu/ [79]), CHARMM (http://www.charmm.org/ [80]), GROMACS (http://www.gromacs.org/ [24,81]), GROMOS (http://www.igc.ethz.ch/gromos/) and NAMD (http://www.ks.uiuc.edu/research/namd [83]) are currently available for simulating biomolecular systems. Some of them are associated with their own force field and most of them allow the use of force fields developed by other groups. Package choice is a very subjective matter and is mostly dictated by the lab or group history. One should be attentive to the fact that some desired features may not be present in all of these packages.

3.6.2. Force Fields

AMBER [84–86], CHARMM [72,73,87] and OPLS [88,89] force fields are most commonly used for simulating the dynamics of nucleic acids and proteins. These force fields evolve constantly under the pressure of newly uncovered artifacts that forces the developers to adjust their parameters. Hence, it is essential to be aware of their limitations and to use these force fields in their applicability range. They are generally able to reproduce quite well most of the important features of biological systems. However, they differ in some specific aspects. An early example is associated with structural deviations that were observed when DNA duplexes where simulated by using the AMBER and the CHARMM force fields. It was shown that AMBER and CHARMM favored the B and A-DNA form, respectively, [90]. This led to the development of the concept of “force field dependent polymorphism” [91] that manifests itself in the theoretical conformational space associated with each force field. These issues have been addressed in subsequent versions of the AMBER and CHARMM force fields. Hence, it is strongly advisable to always seek the latest version of the chosen MD package and associated force fields. Although, it can safely be assumed that non-specific or even future force field will be able to capture all the features of a real biomolecular system, novel polarizable force fields will probably increase the applicability range of current MD simulations at the expense of a larger computational cost [92]. These force fields are still at an early stage of their development and will have first to be thoroughly tested on small systems before using them for simulating large solvated biomolecular assemblies [93,94].

3.6.3. Modified nucleotide and ligand parameterization

Most empirical force fields are equipped with parameters for modeling all canonical RNA, DNA and proteic residues, but rarely for modeling modified residues and organic ligands. Thus, new parameter sets have to be developed. Although some automated procedures for the development of force field parameter sets have been developed, based on fitting to both experimental and quantum mechanical data sets, force field parameterization remains a matter of experts. A classical force field parameterization review has been written by Dinur and Hagler [95]. Fortunately, some tools and methods for generating reasonable to high quality force field parameters are available. Among them, it is worth mentioning the AMBER module Antechamber which greatly simplifies the generation of new parameters and the Automated Frequency Matching Methodology (AFFM) developed for the CHARMM users but extendable to all other parameter sets [96]. AFFM allows development of parameter sets for small to medium sized molecules using high quality quantum mechanical calculations as reference data. The REDDB database [97] for computing and retrieving RESP [98] and ESP atomic charges as well as force field libraries has also been made available to facilitate the derivation of atomic partial charge sets from electrostatic potential calculations (http://q4md-force-fieldtools.org/REDBB/index.php). In addition to those for nucleic acids and amino acids, some force fields have been developed for specific classes of molecules such as carbohydrates (OPLS-AA [99], GLYCAM [100] http://glycam.ccr.columbia.edu/); sulfates and sulfamates [101], polyphosphates [102] or small organic molecules...
(GAFF [103]. Ion parameters compatible with the chosen force field can easily be retrieved from the literature [77,78] and polarizable force fields are being developed for monovalent [104] and divalent cations [105,106]. Lastly, it is worth mentioning a study that developed a set of parameters for the 107 known naturally occurring modified nucleotides [107] and thio-substituted nucleotides [108].

3.6.4. Water models
Among available water models, the TIP3P (Transferable Inter-atomic Potential–3 Point [109]) model is generally associated with the AMBER force field. The more computationally “expensive” four (TIP4P [110]) and five (TIP5P [111,112]) point models have been developed and tested on various systems. SPC (Single Point Charge [113]), SPC/E (SPC/Extended [114]) and SPC/S (SPC like [115]) models are also available. For DNA, it has been shown that the use of TIP3P and SPC/E water models resulted in comparable hydration patterns although the water densities associated with the water model with the highest diffusion rate (TIP3P) appear to be more blurred than those calculated with the SPC/E model that computed diffusion rates in better agreement with experimental data [116]. Besides, a MD simulation of a protein in a crystal environment has led to the conclusion that results obtained with the SPC/E model are in much better agreement with neutron-scattering data than those collected with the TIP3P model [117]. Thus, a particular water model can significantly alter MD simulation results and more studies are needed to precisely evaluate their influence [118,119].

TIP water models have been adapted to the use of the Ewald summation method for calculating the long-range electrostatic interactions and are called TIP3Pew [120], TIP4Pew [121] and TIP5Pew [122]. Some flexible water models were also developed [123] but their use remains marginal.

3.6.5. Non-classified intermolecular interactions and related force field issues
Current force fields take into account a large part of what is known about intermolecular interactions that comprises for the largest part the vast field of regular (N/O-H…O/N) hydrogen bonds that are major actors in the formation, stabilization and reactivity of biomolecular systems [124–126]. Yet, the constant growth in number and variety of biomolecular structures generates its share of surprises by providing examples of unclassified intermolecular interactions. Some of them like O/N-H…π hydrogen bonds [127,128] or the weaker C-H…O/N/S=π hydrogen bonds [129,130] are, like regular hydrogen bonds, to be found in almost all biomolecular system. However, their nature as weak hydrogen bonds was debated for many decades before being entrusted by most structural scientists to belong to the enlarged hydrogen bond family. Although they have not been parameterized for such a purpose, current force fields can capture some if not all of the features of these hydrogen bonds [131,132].

Other “exotic” interactions involve interactions with aromatic rings [133] such as are cation/π interactions [134,135], carbohydrate/aromatic interactions [136,137] or aromatic/aromatic interactions [138]. Besides, stacking interactions in nucleic acids are not yet well understood [139,140]. One has also to take into account interactions involving atoms seldomly found in biological systems like halogen atoms, but that are more and more integrated into synthetic drugs [141] and that play a major role in mammal metabolism through thyroid hormone related systems [142]. In short, carbon bound halogen atoms have the potential to form C–X…O/N/S (X=Cl, Br, I) interactions that can compete in strength with regular hydrogen bonds and displace conformational equilibria [143]. No empirical force field is currently able to accurately model such interactions that have their origin in the high polarizability of the electronic cloud of the larger halogen atoms. Hence, it is advised either to create patches for available force fields or to refrain simulating such systems when no appropriate parameters are available.

3.7. Treatment of electrostatic interactions
The ensemble of parameters in the above referenced force fields is usually considered to be the most important factor in determining the quality of a MD trajectory. Yet, if the empirical equation that models the intermolecular forces at play in biomolecular systems is not accurate enough or if one of its terms is not correctly evaluated, even the most precisely developed set of parameters cannot realistically account for the dynamics of these systems. Indeed, numerous studies have revealed that severe artifacts emerge when the long-range electrostatic contributions are neglected. In other words, when only a part of the Coulomb term \( \frac{q_i q_j}{4 \pi \varepsilon_0 r_{ij}} \) is estimated [144–146]. It has been shown that an accurate treatment of the long-range electrostatic interactions is of paramount importance for generating realistic trajectories of nucleic acid systems. The particle mesh Ewald (PME) summation method [147,148] has proven to lead to very stable nucleic acid trajectories [146,149] and is the method of choice that is currently used for simulating macromolecular systems and especially highly charged nucleic acid systems. Despite the popularity of Ewald summation methods, alternative techniques for the treatment of long-range electrostatic interactions such as shifted truncation methods are still in use [150,151]. Other promising methods are currently being developed but have not yet been used for MD simulations of RNA systems [152].

The program default parameters are generally adequate for using the PME method. These parameters are such that a cubic interpolation scheme and a \( 10^{-7} \text{ Å} \) tolerance for the direct space sum cutoff are chosen. To speed-up the fast Fourier transform in the calculation of the reciprocal sum, the size of the PME charge grid is chosen to be an integer power of 2, 3, and 5 and to be slightly larger than the size of the periodic box, in order to lead to a grid spacing of \( \approx 1 \text{ Å} \) or less.

3.8. Other simulation parameters

3.8.1. Thermodynamic ensemble
MD simulations are sometimes conducted in a microcanonical ensemble or \( (N,V,E) \) for constant number of particles, volume and energy in which energy conservation is naturally satisfied when the classical equation of motion are used. In this ensemble, there is no need to couple the microscopic system to macroscopic thermodynamic properties such as pressure and temperature [150]. Yet, the isothermal-isobaric or \( (N,P,T) \) ensemble is most commonly chosen for several practical reasons, one of them being that a good control of the density of the system is required during the equilibration and possible subsequent large conformational rearrangements where “holes” or regions of negative water density can be created in the solvent box if its volume is kept constant. Working in the grand canonical ensemble in which the number of particle of the system is allowed to change is by no mean straightforward [153]. “Experimental” techniques combining MD simulations with grand canonical Monte Carlo methods have been developed in order to compute ligand/receptor binding free energies [154].

3.8.2. Temperature and pressure
In order to maintain a constant temperature and pressure, the Berendsen temperature coupling scheme and an isotropic molecule based constant pressure scaling scheme with a time constant of 0.2–0.5 ps for both are often used [155]. Although different coupling schemes can be found in the literature, it is often appropriate to stick to the default values proposed by the chosen MD package.
The target temperature and pressure values are usually set to 1 atm and 298 K or 25 °C, also called "room temperature". It is interesting to note that the choice of this temperature target is dictated by the fact that most force field parameters were derived from experimental data obtained at "room" temperature. Nonetheless, simulations using different target temperatures (e.g., 5.25 and 37 °C) have been used to unravel some important temperature dependent differences in the solvation of RNA systems [156]. Simulations of a hammerhead crystal cell were computed at 100 K in order to remain close to the experimental conditions [74].

3.8.3. Shake, time steps and update of non-bonded pair list

A tolerance of 0.0005 Å is generally used for the SHAKE algorithm [157] that allows to "safely" use a 2 ps time step (instead of 1 ps) by artificially freezing the most rapid vibrational motions that occur in biomolecular systems like those associated with C-H, O-H, N-H, … elongations. A strong tolerance value has been found to reduce the so-called "flying ice cube" phenomena [158,159] that involves an increase of the center-of-mass velocity of the entire system and a concomitant cooling of its internal motions. This effect results from the fact that small energy draining occurs when the non-bonded pair list is not updated at every time step and when constant pressure and temperature algorithms are used. Indeed, in order to speed-up the calculation, a list of interacting atom is only calculated once every nth step (usually every 10 steps). To limit the manifestation of this "flying ice cube syndrome", GROMACS removes any center-of-mass motion [81].

3.9. Equilibration

The main aim of equilibration procedures is to alleviate tensions created in the system during the early building stages. These tensions would, if neglected, lead to unrealistic trajectories. In the following, a typical equilibration protocol used by us will be described.

First, 500 steps of steepest descent minimization are applied to the entire system with periodic boundary conditions in the (N,V,T) ensemble. This is followed by 25 ps of simulation in the same thermodynamic ensemble where only the water molecules and the hydrogen atoms of the solute are allowed to move (the heavy atoms of the solute and the ions are frozen by using the BELLY option in AMBER). During the next 25 ps, the constraints placed on the monovalent ions are released so that they can start to equilibrate around the RNA system while the BELLY option is still used to freeze the atoms of the solute with the exception of the hydrogen atoms. Then, several rounds of 50 ps MD are performed in the (N,P,T) ensemble during which positional constraints of 10, 5, 2, 1, 0.5, 0.1, 0.01, and 0.001 kcal mol\(^{-1}\) Å\(^2\) are applied to the heavy atoms of the solute over 50 ps long sequences, yielding a partially constrained 450 ps MD trajectory. Using an (N,P,T) ensemble allows also for the density value of the system to be able to converge. Then the production phase can take place. However, in order to allow a better sampling of the conformational space by the solute and the ions, the next 0.5 to 1.0 ns are generally included in the equilibration phase.

The length of the equilibration phase can be extended or shortened at will. However, it has to be noted that the phase during which constraints are applied ensures that the solvent (water and ions) is appropriately equilibrated in the vicinity of the solute but also throughout the simulation box. The part during which no constraints are applied allows then for the solute to relax in the presence of a relatively well-equilibrated solvent environment. Other equilibration protocols start at low temperature (50 K); then, the temperature is raised to the target temperature at discrete time intervals [160]. An apparent advantage of the former procedure is that the sampling of the conformational space by the solvent is conducted at the target temperature and, hence, probably more efficient.

3.10. Sampling

3.10.1. How long should a simulation be?

Ideally, MD simulations should be long enough to sample all the conformational transitions occurring in a particular biomolecular system. Given present and predictable computational means, this goal will certainly not be achieved in the next decade. It is, thus, impossible to address in the near future biochemical processes with very long relaxation times by using classical MD simulation techniques with explicit solvent representations. Up to now, the longest MD simulations of RNA systems rarely overcome the 100 ns time scale. Nevertheless, many processes associated with limited conformational perturbations lie within reach of present methods. Among them, those that involve the binding of the smallest ligands to nucleic acids (monovalent ions and water molecules [113,32,161]) or the orientation of the 2'-OH hydroxyl groups [39], among many others, can trustfully be evaluated on the current nanosecond time scale.

3.10.2. When to stop a simulation?

Another important and often overlooked issue is: when to stop a MD simulation or when does it start to be a waste of time to continue sampling? Indisputably, a very large number of approximations are included in MD simulations. These approximations are, among others, related to: (i) imprecision in the force field parameters; (ii) incomplete evaluation of the intermolecular forces; (iii) neglect of polarization and charge transfer effects; and (iv) an incomplete knowledge of the starting conditions. This last point by itself can easily lead to the generation of partially or totally unrealistic simulations. For example, an instability of the tertiary core structure of the yeast tRNA\(^{Aosp}\) molecule characterized by a reordering of several base triples has been observed during a 500 ps MD simulation [40]. This has first been attributed to the absence of Mg\(^{2+}\) ions in the model. Later on, it was proposed that an adenine involved in the above-mentioned base triples is protonated [41]. Probably both effects led to the calculated transitions and further sampling would, in this case, not have contributed to a better understanding of the structure and the dynamics of this molecule. Hence, there is a certain risk for long MD simulations to oversample regions of the "configurational space" specific to the used force field and MD parameters. These explored theoretical "configurational space" regions may not or only partially overlap with the "true" configurational space explored by the "real" system. A further example is provided by the formation of salt crystals in undersaturated conditions resulting from the use of inappropriate ionic parameters [77].

3.10.3. Multiple molecular dynamics (MMD) simulations

An alternative to long MD simulations consists in generating an ensemble of several "shorter" trajectories by using MMD simulation technique. This method introduces slight perturbations (such as different initial velocities) in the starting conditions and, thus, exploits the chaotic nature of MD simulations [162] in order to generate an ensemble of several uncorrelated trajectories [7–9,131,163–165]. Indeed, multiple trajectories generated from similar but slightly different starting conditions may provide, on a statistical basis, more information than a single long trajectory. A parallel view consists of generating and comparing several trajectories obtained by starting from different initial configurations [9]. In short, MMD methods allow a statistical evaluation of MD simulation that cannot be achieved from the analysis of single MD trajectories [146].

Some alternatives to MMD techniques exist. For instance, simulations of entire crystal cells containing more than one image of the
same molecule can generate multiple trajectories of a molecular fragment [74]. Furthermore, constructing a model system containing several times the same structural element can also allow to significantly enhance conformational space sampling [32,47]. In a MD simulation of a crystallographically determined RNA/antibiotic complex containing two identical binding sites, a slight dissymmetry between them led to a detectable diverging dynamical behavior that could be corrected by a single small (0.2 Å) contraction of a water binding site in the starting structure [47].

3.10.4. Other sampling methods
Since efficient sampling of the conformational space of biomolecular systems is highly desirable but complicated by slow barrier crossing on the rugged energy landscape of these systems, numerous “enhanced” sampling methodologies were developed in the past years. These methods are based on molecular dynamics simulation techniques but do not reproduce the dynamics of a real system. Hence, they will not be described here and the interested reader might check following reviews [166–168].

4. Validation

4.1. Evaluating the quality of MD trajectories

Before trying to extract information from MD simulations, it is crucial to evaluate the quality of the generated trajectories by checking their internal consistency and confronting them to all available experimental data [16,169,170].

4.1.1. Consistency checks

Internal consistency checks are related to the detection of unphysical behaviors such as those resulting from the use of truncation methods for the evaluation of the electrostatic interactions [144,145], the “flying ice cube” syndrome [158,159] or the artefactual aggregation of monovalent ions in diluted aqueous solution [77]. When such issues come fore, it becomes clear that the generated trajectories should no longer be used to derive biologically relevant information. Instead, the focus of the study has to shift towards the understanding and correction of the detected artifacts. By doing so, the importance of the neglected contributions is generally brought to light. For instance, the stabilizing role of “hydration forces” could be assessed by comparing trajectories issued from simulations including or discarding long-range electrostatic interactions [66,91,131] and the structural role of water molecules could be appraised by comparing “in vacuo” and “in aquo” MD simulations [66,171]. Consequently, “in silico” simulations in which important components of natural systems are switched “on” and “off” illustrate the effects linked to the presence or absence of associated interatomic interactions.

4.1.2. Comparison with experimental data

Confrontation with available experimental data (that are most frequently of a structural type) is in all cases mandatory. For instance, if important tertiary interactions (Watson–Crick base pairing, …) present in the starting structure break during the simulation over short time scales it is probable that the simulation protocols are not adequate and the reasons underlying such behavior must be uncovered [91,145]. Early simulations performed under “in vacuo” conditions or with truncation of the electrostatic interactions where disruptions of important tertiary interactions were observed illustrate clearly this point [145]. Moreover, if the structure of the investigated system diverges significantly from the initial structure, in part or entirely, during the time course of the simulation, it must be taken into account that this observation might result from force field inaccuracies. Only comparisons with experimental data on the same or related systems can help resolving such issues [172].

4.1.3. Force field comparison and sensitivity analysis

Comparing MD runs generated by using different force fields is a difficult task since, when results differ from one force field to another, it is not always obvious to get a clue about which parameter set is the most appropriate unless evident flaws are detectable in one ensemble of trajectories (ion aggregation for instance [77]), or direct comparison with experimental data discards one of them. In a study that had for aim to compare ion parameters for NaCl simulations, the authors reported large differences in computed radial distribution functions and concluded that the calculated uncertainty reflects our incomplete experimental knowledge of the structural properties of ionic solutions at finite molarity [173]. Other studies aimed at comparing force fields and associated water models with sometimes-contradictory conclusions [87,90,118,119,174–176]. Since small changes in parameters might lead to large calculated differences it would be desirable to develop sensitivity analysis tools [116,177,178]. However, in a setting in which non-linear processes are at play, it represents a very tricky if not impossible task.

4.1.4. Visualization

Visual checks are absolutely mandatory and can be performed by using the programs that are delivered with each MD package or by using programs such as VMD (Visual Molecular Dynamics; http://www.ks.uiuc.edu/Research/vmd/) that can read most of the available MD outputs. In rare cases, the Open BABEL utility (http://openbabel.sourceforge.net/) is needed in order to convert data written in different formats. PyMOL is also a very efficient and complete visualization tool for experimental structures (http://www.pymol.org). Note that visualization is required not only for analyzing the structure of the solute in order to get a first glimpse of the conformational transitions that might occur during the MD run but to evaluate the structural characteristics of its environment.

4.2. Convergence issues

Afterwards, the convergence of the simulation with respect of the investigated properties has to be assessed. For this, it is necessary for the simulation length to exceed the average relaxation time of the investigated property which is not necessarily the longest relaxation time of the system that can largely exceed the currently accessible nanosecond time scales [169]. For example, it is possible to estimate on a statistical basis the residence times of water molecules (between 10 and 500 ps) around a regular RNA helix from nanosecond MD simulations, although the statistics will obviously be less good for the water molecules with the longest residence times [32,161]. For less regular RNA systems, water molecules may be trapped into specific pockets for much longer times and only lower bounds for the residence times can be obtained [179]. A similar problem is associated with the estimation of the binding properties of monovalent and divalent ions. While it is feasible to estimate the exchange time of monovalent ions located in RNA grooves from nanosecond MD simulations [68], it is impossible on such time scales to estimate the same values for divalent ions [67]. It should be noted that blindly seeking for convergence is not advised. Sometimes, it is possible, by visual inspection or by using other property screening tools, to detect during the course of a MD simulation flaws that would inform one to stop the simulation before any kind of convergence is achieved.
4.3. Solvent analysis

Solvent analysis is an important aspect of the evaluation of MD trajectories [66] and is related to the characterization of water and ion binding sites, to the estimation of their residence times and to their role in association processes such as drug binding [47,180]. Although the solvent is generally not considered as being part of the solute, numerous studies point to the opposite and indicate that the solvent (ions and water) is an integral part of nucleic acid structures. Several reviews have already addressed these issues for nucleic acid systems [179,181–184] and we will not discuss them in further details. The SwS web service (http://www-ibmc.u-strasbg.fr/arn/sws/html) [61] provides a statistical overview of the solvation of nucleic acid base pairs that is useful for better understanding crystal structures and analyzing MD results.

5. Hopes, wishes and desires

It is our hope that these guidelines will facilitate the generation of future MD simulations of RNA systems by helping solve critical technical issues while keeping an eye on aspects related to the starting structures that are all too often neglected or overlooked by the modelers. Indeed, MD simulations are only as accurate as are the chosen starting structures, simulation packages and force fields. Herewith, it is our wish to arouse the critical spirit of modelers and to increase their awareness when analyzing results from MD trajectory since MD simulations will always generate their load of inaccuracies due to the use of empirical models and should not be over interpreted. It is finally our desire to see blossom more powerful MD techniques over the ashes of those that we are currently working with. The most interesting techniques and data are yet to come.

Acknowledgments

We wish to thank all the authors of RNA molecular simulations that shared their knowledge with us. Y. Hashem was supported by a grant from the French ministry of research (MRT).

References


