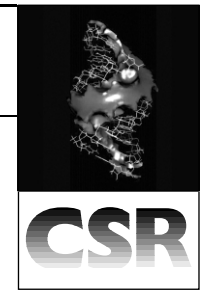
Theoretical methods for the simulation of nucleic acids



Modesto Orozco,*[a] Alberto Pérez,[a] Agnes Noy[a] and F. Javier Luque*[b]

[a] *Institut de Recerca Biomèdica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona E-08028, SpainDepartament de Bioquímica i Biologia Molecular, Facultat de Química, Universitat de Barcelona, Martí i Franquès 1, Barcelona E-08028, Spain*
[b] *Departament de Farmàcia, Unitat Fisicoquímica, Universitat de Barcelona, Avgda Diagonal 643, Barcelona E-08028, Spain*

Received 25th February 2003
First published as an Advance Article on the web 11th July 2003

Different theoretical methods for the description of nucleic acid structures are reviewed. Firstly, we introduce the concept of classical force-field in the context of nucleic acid structures, discussing their accuracy. We then examine theoretical approaches to the description of nucleic acids based on: i) a rigid or quasi-rigid description of the molecule, ii) molecular mechanics optimization, and iii) molecular dynamics. Special emphasis is made ion current state of the art molecular dynamics simulations of nucleic acids structures.

Introduction

Genetics, biochemistry and molecular biology, as understood at the beginning of the XXI century, would be impossible without the knowledge of the structure of DNA, which justifies the transmission of the genetic information and explains how this information can be translated into simple instructions to the cellular machinery.

Watson and Crick, in the middle of the fifties, reported the first reliable model of the double helix,[1] which was undoubtedly

Modesto Orozco was born in Barcelona, Spain in 1962. He received his BA degree in Chemistry from the Universitat Autònoma de Barcelona in 1985 and his PhD from the same University in 1990. He joined the Departament de Bioquímica i Biologia Molecular of the Universitat de Barcelona in 1986 as assistant professor. He is now full professor of Biochemistry at the same department and group leader at the Institut de Recerca Biomèdica at the Parc Científic de Barcelona (IRBB-PCB). He has received different national and international scientific awards, is the authors of almost 200 papers, and is a member of the editorial board of Journal of Computational Chemistry and of Theoretical Chemistry Accounts. His research interests are the simulation of systems of biochemical impact, with special emphasis in nucleic acids, and the development of methods for the study of molecular interactions and solvent effects.

Alberto Pérez was born in Barcelona, Spain in 1980. He received his BA degree in Chemistry from the Universitat de Barcelona in 2002. He is now working for his PhD at the Universitat de Barcelona under the supervision of Professors Orozco and Luque. His research interests are analysis of 3-D databases of nucleic acids, and the simulation of the dynamics of polynucleotides.

Agnes Noy was born in Barcelona, Spain in 1979. She received her BA degree in Biochemistry from the Universitat de Barcelona in 2002. She is now working for her PhD at the Universitat de Barcelona under the supervision of Professors Orozco and Luque. Her research interests are the simulation of DNA and RNA by molecular dynamics simulations.

F. Javier Luque was born in Barcelona, Spain in 1962. He received his BA degree in Chemistry from the Universitat Autònoma de Barcelona in 1985 and his PhD from the same University in 1989. He joined the Departament de Física-Química of the Universitat de Barcelona in 1986, where he is now Professor of Physical Chemistry. He has received different scientific awards, is the author of almost 200 papers, and is a member of the editorial board of Theoretical Chemistry Accounts. His research interest are the theoretical representation of solvation effects in molecular structure and chemical reactivity of biochemical systems.

Modesto Orozco

Alberto Pérez

Agnes Noy

F. Javier Luque

 DOI: 10.1039/b207226m

one of the most important scientific discoveries of the past century. They arrived at this model based on scarce and rather rough experimental data, which would be meaningless without the intelligent use of molecular modelling made by these two scientists. Using wires and bolts, they built up structural models following two rules: i) the models should be consistent with the physical rules governing the structure of molecules, and ii) they should explain all the available experimental information. The same basic rules are applied today, almost 60 years later, in molecular simulations of nucleic acids.

Since the seminal work of Watson and Crick, simulation has been one of the most powerful tools for gaining insight into the nucleic acids. A simple reference analysis using Idealibrary web server (www.idealibrary.com) for the period Jan 2000 to Sep 2002 shows around 9000 articles on nucleic acids containing the keywords molecular dynamics, simulation or theoretical study in their headers. Just considering the keywords DNA and molecular dynamics, more than 2000 hits are detected in this short period of time. Clearly, molecular simulation of nucleic acids is now a mature field, and modelling techniques are powerful tools for the description of nucleic acids. This paper tries to summarize some of the most recent advances in this field, with a special emphasis on the theoretical framework underlying molecular simulations of nucleic acids. It is not our purpose to perform an exhaustive revision of the current literature, but to provide very general principles to understand the strengths and weaknesses of simulations in nucleic acids, especially for non-expert readers. More detailed reviews can be found elsewhere.[2–9]

The classical approach

Simulation of nucleic acids requires a functional to connect the coordinates of a system with its energy. This could be done *a priori* using quantum mechanics (QM), which would allow the calculation of the structure, dynamics and reactivity of nucleic acids based on first principles. Unfortunately, QM techniques are computationally very demanding when applied to large flexible systems like hydrated nucleic acids. Thus, despite the recent advances in the development of efficient codes for *ab initio* molecular dynamics,[10] the use of QM techniques in the analysis of nucleic acids is rare, mostly limited to static studies of reactivity or to the detailed description of small fragments.

Classical approaches assume that the nucleic acids can be represented using Newton's laws and simple equations (the force-field) relating the nuclear structure of the system with its energy. The level of accuracy in the representation of nucleic acids and in the force-field leads to a variety of simulation methods, which are oriented to the study of different aspects of the nucleic acids. Thus, there is interest in the simulation of macroscopic properties of very large segments of DNA, the electrophoretic behaviour of nucleic acids, or its ability to wrap around large proteins. But there is also interest in the atomic-detailed study of the interactions occurring in a short piece of DNA. Following ideas developed by Olson, Zhurkin[2,6] and Lavery's,[4] those methods can be classified in three categories: i) ideal-elastic (macroscopic), ii) mesoscopic (intermediate), and iii) microscopic models.

Macroscopic (ideal-elastic) models

Many aspects of the structure and functionality of nucleic acids, such as the chromosome packing, superhelicity, hydrodynamic and electrophoresis behaviour or the packing of nucleic acids in viral capsids, are related to their macroscopic polymeric nature.[2] All these processes have two main characteristics: i) they imply very large fragments of nucleic acids, and ii) the mechanism involved in those processes should be very general and independent of the particular characteristics of the nucleic acid sequence. Therefore, the interest lies in the representation of properties related to the general polymeric properties of nucleic acids, and not to specific sequence-dependent structural details.

Macroscopic elastic models assume that the nucleic acid is a flexible ideal rod, whose properties can be represented using principles of macroscopic mechanics.[2] The model can be enriched to include experimental data, like that derived from electron microscopy, enzymatic footprinting or chemical cross-linking experiments. When these models are used, deformations of nucleic acids in any direction of the space or any part of the rod are equally probable. This makes it possible, for instance, to represent large pieces of DNA or even the entire ribosomal RNAs,[11] but only with a low resolution, thus neglecting fine details related to environmental or sequence effects.

Mesoscopic (intermediate) models

The concept "mesoscopic" is very popular in physics, but not commonly used in chemistry.[12] It refers to the diffuse interface between microscopic and macroscopic descriptions of a system, *i.e.* the study of those systems too big to be treated at the microscopic level, but too complex to be represented at the macroscopic level. In the case of nucleic acids, mesoscopic studies refers to the representation of very large pieces of DNA, where sequence or environment effects makes unsuitable the use of the ideal-elastic rod models[2,6,13,14]

These models divide the polynucleotide into small rod elements, often named "beads". Each bead is considered a rigid entity, while the connections between "beads" are flexible and allow the nucleic acid to adapt to external deformation forces. The size of one "bead" can vary depending on the type of problem under analysis (from a single base pair for medium-sized oligonucleotides to several kilobases for an entire chromosome). The deformation energy is computed using elastic potentials, which distinguish between different type of deformations. A typical elastic potential is shown in equations 1–3,[13] where ß, τ refer to bending and twisting angles and l to stretching distance. The constants B, C and S define the rigidity of a nucleic acid fragment in front of deformation of bending, twisting and stretching. The subscript "0" denotes the equilibrium values for bending, twisting and stretching of a given nucleic acid fragment.

$$E_{\text{bending}} = 0.5(\text{B}/l_0)(\beta - \beta_0)^2 \quad (1)$$

$$E_{\text{twisting}} = 0.5(\text{C}/l_0)(\tau - \tau_0)^2 \quad (2)$$

$$E_{\text{stretching}} = 0.5(\text{S}/l_0)(l - l_0)^2 \quad (3)$$

The force-field parameters can be derived from the analysis of macroscopic properties of known DNA sequences and from the inspection of high resolution structures of short DNA fragments. For instance, Figure 1 represents the roll *versus* twist distribution of all the A- and B-type DNA sequences deposited in the December 2002 release of the Protein Data Bank (PDB). Some relevant features can be noted in Figure 1. First, only a fraction of roll/twist combinations are allowed in the DNA duplex. Second, these combinations are specific of the helical (A or B) family. Third, significant differences exist in the roll/twist distribution depending on the nature of the base pair dimer. Thus, the center of the distribution for B-type DNA changes from −3.7/36.5 (roll/twist; degrees) for purine–pyrimidine steps to 2.9/35.8 for pyrimidine–purine steps, indicating a different intrinsic equilibrium geometry for these two types of sequences. Interestingly, not only the center of the distributions, but also their shapes depend on the helical family and the sequence. For example, the ratio between the number of

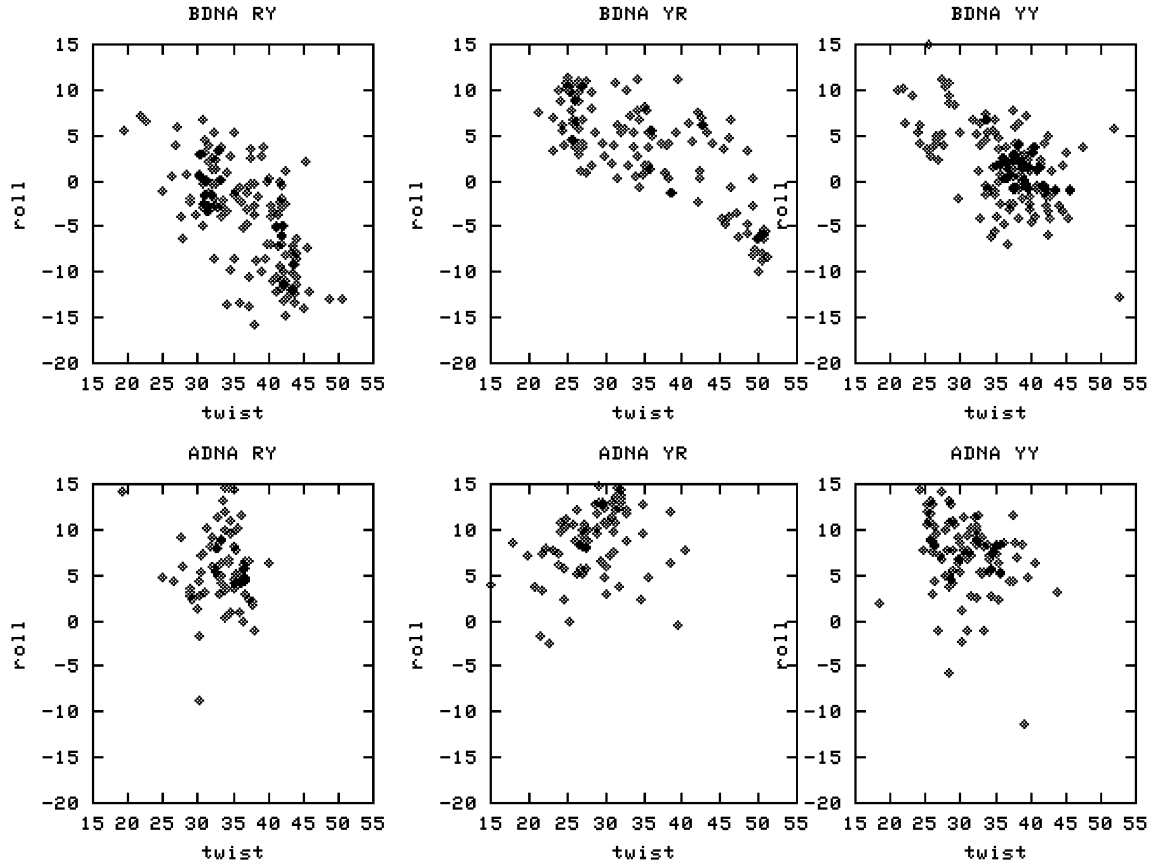


Fig. 1 Roll/Twist distributions of the structures found in PDB of A- and B- type DNA. Base pairs at the extremes of the helices are removed from the study. The rest are clustered in purine–pyrimidines (RY), pyrimidine–purine (YR), and pyrimidine–pyrimidine (YY) steps. All values are in degrees.

different "visited" microstates (defined in 2 × 2 degree bins in Figure 1) and the total number of visited microstates for a given sequence in PDB (number of points in Figure 1) show different values for B- (around 0.44) and A- (around 0.35) type DNAs, suggesting different roll/twist flexibility between A- and B-forms of DNA. The same calculation show also that for B-DNA duplexes pyrimidine–pyrimidine tracks show less variability in roll/twist values than pyrimidine–purine tracks, suggesting a different intrinsic flexibility in both types of sequences. Clearly, this type of result can be manipulated to derive "knowledge-based" force-fields, which in turn can be exploited in mesoscopic studies of the DNA.

The parametrization of mesoscopic models from crystal data permits the connection of the microscopic features of polynucleotides with their macroscopic properties. However, caution is necessary regarding the quality of the final parametrized model. First, the number of high resolution nucleic acid structures is still limited and clearly insufficient to derive statistically significant parameters for all base pair steps. Second, flexibility and deformability of DNA are dynamic concepts, which might not be properly reflected in the structures deposited in the PDB. Thus, an interesting alternative to the parametrization of mesoscopic models relies on the use of molecular dynamics (MD) simulations performed by using an atomic-level representation of the nucleic acids.[6,15] These simulations provide maps of visited conformations similar to those derived from the analysis of the PDB. This is illustrated in Figure 2, which contains the roll/twist values visited in a 6 ns MD simulation of a 12-mer RNA molecule. The MD-based parametrization permits the inclusion of an unlimited number of structures for the fitting, though in practice it will depend on the computational capabilities. In particular, current MD simulations (typically 1–5 ns) make it difficult to visualize slow conformational transitions, which can be important in understanding the flexibility of nucleic acids.[15] Moreover, the results can be affected by the physical force-field, which has been previously parametrized from experimental and quantum mechanical data. In contrast, this approach allows us to introduce the dynamic properties derived from the analysis of the trajectory in the parametrization process.

Microscopic models

Many functional aspects of the nucleic acids depend on fine sequence and structural details, whose analysis requires an atomic or *quasi*-atomic level of description. All the classical microscopic models are based on the calculation of the molecular energy for a given nuclear configuration using a force-field. The differences between the methods are found in: i) *the representation of the nucleic acid (the degrees of freedom considered)*, ii) *the force-field*, and iii*) the post-processing of the energy information derived from the force-field*.

Level of representation of the nucleic acid

Most microscopic calculations pursue: i) to obtain an optimised set of coordinates of the nucleic acid, ii) to sample the configurational space of the system, or iii) to determine the impact of external stimulus on the structure of the system. In general microscopic methods consider all or almost all the atoms of the system (some hydrogen atoms might not be explicitly treated). However, the degrees of freedom explicitly explored during the calculation can vary depending on the "level of representation of the nucleic acid", it being possible to distinguish between *atomic resolution models* and *collective variables models*.

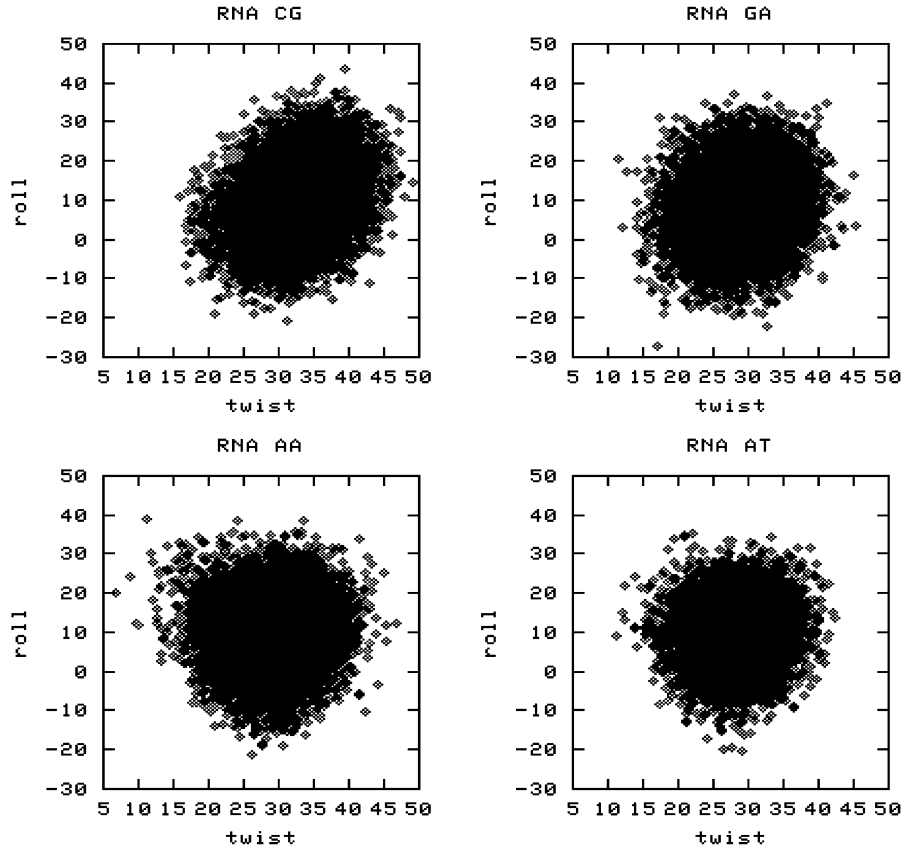


Fig. 2 Roll/twist distributions of different base steps of the r(CGCGAAUUCGCG)$_2$ obtained in a 6 ns trajectory in water. MD simulation was carried out using the PME protocol for long range effects and the AMBER-99 force-field.

Atomic resolution models. These do not assume any previous knowledge on the behaviour of nucleic acids other than that implicitly assumed in the development of the force-field. Accordingly, all the atoms of the system are free to move and their individual trajectories are in principle just restrained by the interactions with the rest of the atoms. In practice, some degrees of freedom, typically bond distances, are kept fixed to improve the efficiency of the simulations (see below). These models typically work in Cartesian coordinates and are intended to study fine structural details (angle distortions in the sugars, out of the plane bending of nucleobases, base opening, rearrangements in hydrogen bond pattern, *etc*.) of relatively short (typically less than 15-mer) nucleic acid structures. In most cases the solvent (water and counterions) is explicitly represented at the same level of accuracy as the nucleic acid. Typically, these models are coupled to MD algorithms (see below).

Collective variable models. These models are specific to the study of nucleic acid structures. They exploit previous knowledge on the behaviour of these molecules to reduce the number of degrees of freedom. For example, they generally ignore degrees of freedom related to all bond lengths and most angles, taking advantage of the fact that bonds and angles are very close to their optimum values in most structures. Additional reduction in the degrees of freedom arises from the use of pseudorotation model, by the introduction of helical restrictions, or by imposing the planarity in the bases or base pairs. Clearly, the reduction in the degrees of freedom largely simplifies the configurational space, but at the expense of a loss in information. In their most simplified version, collective variable approaches use simple ideal helical models of DNA, where the structure is represented by means of a series of translational and rotational parameters defining the position of bases and base-pairs with respect to the helical axis, and the neighbour bases (or base-pairs). Helical parameters have been very useful for the description of canonical families of DNA and RNA, and computer programs like Curves,[16] NewHelix[17] and 3DNA[18] are routinely use to characterize nucleic acid structures obtained from X-ray, NMR or atomic-detailed simulations. Similar algorithms can be use to generate starting conformations for nucleic acid structures. Unfortunately, pure helical representations tend to neglect local distortions and may not be appropriate to describe irregular nucleic acids, where the definition of helical axis or base pairs is unclear.

An example of a successful collective variable approach is the JUNMA (*junction minimisation of nucleic acids*) program developed by R. Lavery and coworkers.[19] JUNMA breaks the structure into 3′monophosphate nucleotides cutting at the O5′–C5′ bond. The nucleotides are then positioned in the space with respect to a common helical axis using three rotational (inclination, tip and twist) and three translational (Xdisp, Ydisp and Rise) parameters. The nucleotides are not treated as rigid entities, since a set of variable parameters (a few bond angles and most torsions in the phosphoribose backbone) is defined. This set is divided into two blocks: the first represents the independent variables (those considered as explicit degrees of freedom), and the second (the "dependent" set) include all the other geometrical parameters that vary owing to changes in the first set of parameters. Independent variables include the phosphodiester (C3′–O3′ and O3′–P) and glycosidic rotations, three bond angles in the ribose and the C1′–C2′ and C2′–C3′ rotations (the sugar is disconnected in the C4′–O4′ bond). Changes in this reduced set of independent variables are accompanied by concerted movements in the "dependent" flexible variables guided by harmonic restrains, which guarantee the maintenance of "closure" conditions (C4′–O4′ and C5′–O5′ bond lengths and a few bond angles). Summing up, JUNMA reduces the degrees of freedom to six helix-related translations and rotations plus five dihedral angles and three bond angles per nucleotide. Other features of the program allow the introduction of additional restrains to maintain strong

hydrogen bonds (following many known H-bond patterns) or stacking interactions between selected bases.[20]

Similar approaches have been followed by Zhurkin and Olson's groups,[21,22] who used a helical reference system where six independent pair parameters (Propeller Twist, Buckle, Opening, Shear, Stretch and Stagger), six step parameters (Twist, Roll, Tilt, Shift, Slide and Rise), and the phase and glycosidic angles are used to describe the DNA duplex geometry. The method has been used to study the structure and deformability of large pieces of duplex DNA,[21] and can be incorporated in both molecular mechanics and Monte Carlo algorithms (see below).

Collective variables models are often applied in the context of rigid-body or molecular mechanics optimizations, though some authors have developed Monte Carlo[21] or MD-adapted codes (see below). The solvent is typically considered at an implicit level by means of effective dielectric functions which mimic the screening effect of water and counterions, and by the reduction of the phosphate charges. Recently, efforts have been made to treat explicitly counterions and water environments in collective variables calculations (see below).

Microscopic force-fields

The force-field is a classical expression for the molecular Hamiltonian, which connects the structure the system with its potential energy. Many microscopic force-fields have been used for the study of nucleic acids structure, and it is not our purpose to revise all of them here. We will just summarize a few of the characteristics of the most popular force-fields for current microscopic simulations of nucleic acids.

A typical force-field computes the potential energy using a general equation similar to that shown in equation 4, where E_{str} and E_{bnd} stands for the stretching and bending energies (see eqns. 5 and 6), E_{tor} is used to represent the energy profile of rotations around chemical bonds (see eqn 7), E_{nb} stands for the non-bonded interaction energies, and E_{other} accounts for any other type of interactions included in the calculations, such as improper torsions (in united-atom force-fields) experimental or symmetry restrains, external potentials, *etc*.

$$E = E_{str} + E_{bnd} + E_{tor} + E_{nb} + E_{other} \quad (4)$$

$$E_{str} = \sum_{bonds} K_{str}(l - l_o)^2 \quad (5)$$

$$E_{bnd} = \sum_{angles} K_{ang}(\Theta - \Theta_o)^2 \quad (6)$$

where K_{str} and K_{bnd} stands for stretching and bending constants, and l_0 and Θ_0 are equilibrium bond lengths and angles.

$$E_{tor} = \sum_{tor}\sum_{n=1}^{3}\frac{V_n}{2}(1 + \cos n\Phi - \gamma) \quad (7)$$

where n stands for the periodicity of the Fourier term, Φ is the torsion angle, γ is the phase angle and V_n is the torsional barrier.

Force-fields used exclusively in collective variable calculations ignore stretching terms and many bending and torsion contributions. Force-fields used in the atomic-level representation of nucleic acids might also neglect part or all of the stretching contributions, but include all bending and torsion terms into the calculation of molecular energy.

Non-bonded interactions typically include a van der Waals term (E_{vw}), which accounts for dispersion–repulsion interactions between atoms, and a Coulombic term, which represents the electrostatic interactions (E_{ele}) between atomic charges. Some force-fields include terms specific for hydrogen-bond interactions. Finally, the newest generation of force-fields also include terms related to polarization contributions, which are expected to improve the representation of highly charged systems like nucleic acids, but they are still experimental and very costly, which explains their limited impact on the field.

The van der Waals term can be represented following different formalisms, the simple Lennard–Jones formalism (see eqn. 8) being the most popular. The parameters A_{ij} and C_{ij} (eq. 8) are typically obtained from atomic van der Waals parameters using geometric or arithmetic combination rules. As noted above, some force-fields include specific formalisms for hydrogen-bonded systems. Thus, old versions of AMBER force-field replaced the normal r^{-12}–r^{-6} expression by a softer r^{-12}–r^{-10} formalism in hydrogen-bond contacts, and the same procedure is used in Zhurkin's force-field.[21–23] The FLEX force-field[24] introduces a more complex formalism to deal with hydrogen-bonds, which include different van der Waals parameters for normal and hydrogen-bond interactions, as well as a directional term which ensures the linearity of H-bond interactions (see eqn. 9).

$$E_{vw} = \sum_{i,j}\left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{C_{ij}}{R_{ij}^{6}}\right) \quad (8)$$

where the sum extends for all non-bonded pairs.

$$E_{H\text{-}bond} = \sum_{k,l}\cos\mu\left(\frac{B_{kl}}{R_{kl}^{12}} - \frac{D_{kl}}{R_{kl}^{6}}\right) + (1 - \cos\mu)\left(\frac{A_{kl}}{R_{kl}^{12}} - \frac{C_{kl}}{R_{kl}^{6}}\right) \quad (9)$$

where the sum extends for all non-bonded pairs involved in possible H-bond interactions. B_{kl} and C_{kl} are H-bond van der Waals parameters, and μ is a directional angle formed by the bonds X–H and Y···H in a X–H···Y hydrogen bond.

Electrostatics is the key issue in the representation of nucleic acids. All force fields use a simple Coulombic expression like that shown in eqn. 10, where the charges are typically (but not always) located at atomic nuclei. The difference between force-fields stems from the nature of the point charges used to represent the charge distribution, and the treatment of the dielectric screening. The first generation of force-fields used charges obtained from low-level QM calculations and very primitive charge population methods. Often, these charges were scaled to reproduce experimental or QM data. The newest force-fields, in their non-polarized versions, always use charges derived from HF/6-31G(d) wavefunctions. The popularity of this medium-level QM calculation conforms with its well known tendency to overestimate polarity, thus mimicking the polarizing effect of water. Different strategies have been developed to derive point charges from the HF/6-31G(d) wavefunction. In our experience, the (R)ESP method,[25,26] used for example in AMBER,[27,28] is especially simple and powerful.

$$E_{ele} = \sum_{m,n}\frac{Q_m Q_n}{\varepsilon(R_{mn})R_{mn}} \quad (10)$$

where the sum extends for all non-bonded pairs, and ε is the dielectric constant.

A remarkable difference between force-fields stems from the description of dielectric response. The use of dielectric functions in microscopic simulations is always uncomfortable, since "dielectrics" is a macroscopic concept which in principle should be captured intrinsically by simulation. Unfortunately, this is only true if the system is heavily solvated, and if proper sampling of the solvent configurational space is obtained. In principle, this implies the inclusion of tens of thousands of extra degrees of freedom in molecular dynamics or Monte Carlo calculations, leading then to a dramatic increase in the cost of the simulation.

Force-field calculations performed in the context of the collective variables approach, like Zhurkin's method,[21,23]

DUPLEX[29] or FLEX,[24] typically use effective dielectric functions, which are complemented by a scaling down (by 0.5 or 0.25) of the charge at the phosphate to mimic the counterion environment. Many dielectric functions have been developed to simulate the screening effect of water on charge–charge interactions. The simplest one, still used in preliminary optimizations of nucleic acids, is the linear relationship, with scaling factors (EPS) ranging from 1 to 4 (eqn. 11). Debye–Hückel screening functions (eqn. 12) have also been used in early simulations. However, the newest dielectric formalisms use sigmoidal functions, which better simulate the rapid increase in dielectric response at a certain interatomic distance, and the stabilization of such a response at large distances. Examples of these functions include the popular Hyngerty's expression (eqns. 13–14; [30]) and the Mehler–Solmayer functional (eq. 15; [31]). In our experience, both functionals satisfactorily mimic the electrostatic potential derived from the rigorous Poisson–Boltzman approach. Furthermore, when used in standard force-field optimization routines, these functions reproduce many characteristics of nucleic acids,[21–24] but obviously cannot reproduce fine details of solvent environment. For a more complete explanation on effective dielectric constant, see reference 32

$$\varepsilon(R_{ij}) = EPS \cdot R_{ij} \tag{11}$$

$$\varepsilon(R_{ij}) = \varepsilon_\infty \exp\left(\frac{-R_{ij}}{D}\right) \tag{12}$$

where ε_∞ is the bulk dielectric of the solvent, and D is the Debye length constant

$$\varepsilon(R_{ij}) = \Im(\varepsilon_\infty)\left[RS^2 + 2RS + 2\right]\exp(-RS) \tag{13}$$

where R is an empirical constant set to 0.356 in Hyngerty's original work and to 0.16 in a further refinement by Lavery's group.[24] The permittivity function ($\mathcal{F}(\varepsilon_\infty)$) is determined as shown in eqn. 14

$$\Im(\varepsilon_\infty) = \varepsilon_\infty - \frac{\varepsilon_\infty - 1}{2} \tag{14}$$

$$\varepsilon(R_{ij}) = E + \frac{F}{1 + u\exp\left(-\lambda F R_{ij}\right)} \tag{15}$$

where in the original parametrization $E = -8.5525$, $F = \varepsilon_\infty - E$, $\lambda = 0.003627$ and $u = 7.7839$.

Other effective dielectric functions and parameters have been suggested, such as a very steep sigmoidal function recently developed by Hingerty and Olson, which is optimized to reproduce the electrostatic properties of B-type DNA duplexes,[29] and that is used in a newly reparametrized version of DUPLEX for simulations of DNA in the absence of explicit waters and counterions.[29]

Traditionally, force-fields used in atomic-level representations of nucleic acids, like AMBER[27,28] or CHARMM,[33,34] introduce explicit representations of the solvent, avoiding then the need to use effective dielectric constants ($\varepsilon(R_{ij}) = 1$ in eqn. 10). TIP3P[35] and SPC[36] models of water are the most commonly used for nucleic acid simulations. More recent and accurate models (including many-site potentials and/or polarized potentials) have not been able to replace these two venerable models, which in our experience reproduce most of the properties of water which are of interest in simulations of nucleic acids.[32]

The use of explicit solvent representations presents many advantages, especially when force-field calculations are coupled to MD algorithms. However, the extra computational effort can impede the simulation of large fragments of DNA or limit the length of the MD simulations. This has stimulated the development of intermediate methods based on continuum solvent theories, which reproduce both the solute–solvent interaction and the screening effect of solvent on intra-solute interactions for a large range of systems.[32] The most popular one has been the Generalized-Born/solvent accessible surface (GB/SA) approach, originally developed by Clark and Still,[37] and now available in slightly different implementations (see ref. 32 for discussion). The GB/SA approach assumes that the solvation free energy is determined as a combination of steric and electrostatic effects (eqn. 16). Steric contributions (including cavitation and dispersion terms) are typically represented by means of an empirical linear relationship with the solvent accessible surface using either universal or atom-specific surface tension parameters (ξ in eqn. 17). The electrostatic term is computed using an empirical generalization of the Born's equation (see eqns. 18–20). GB equations are of empirical nature, but guarantee that in the limit of large and short distances Born's and Bell–Onsager's models of solvation for monopoles and dipoles are fulfilled.[32,37]

$$\Delta G_{\text{sol}} = \Delta G_{\text{ster}} + \Delta G_{\text{ele}} \tag{16}$$

$$\Delta G_{\text{ster}} = \sum_k \xi_k SAS_k \tag{17}$$

$$\Delta G_{\text{ele}} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon_\infty}\right)\sum_{m,n}\frac{Q_i Q_j}{\Gamma_{\text{GB}}} \tag{18}$$

where the empirical GB screening function is computed as:

$$\Gamma_{\text{GB}} = \sqrt{R_{ij}^2 + \alpha_i\alpha_j \exp(-D_{ij})} \tag{19}$$

with

$$D_{ij} = \frac{R_{ij}^2}{d\alpha_i\alpha_j} \tag{20}$$

where α stands for the Born's radii determining the average distance from a given atom to the solvent, and the scaling constant d is typically set to 4,[37] though other values have been suggested.[32]

The bottleneck of a GB calculation is the determination of the Born radii, since they depend not only on the intrinsic van der Waals radii of atoms, but also on the position of all the other atoms of the system. The numerical methods proposed initially to compute the Born radii were accurate, but very slow, and have been replaced for approximate methods, like those developed by Hawkins *et al.*[38] or Still's groups,[39] which increase the efficiency of the calculation with just a small loss in the quality of the results.

Case and coworkers[40] have reported extensions of the GB equation which include Debye–Hückel corrections to deal with salt effects based on the substitution of the dielectric factor (see eqns 20 and 21) by a term dependent on the Debye–Hückel screening constant (D in eqn. 21). However, despite its potential interest the use of Case's approach has not became popular yet.

$$\left(1 - \frac{1}{\varepsilon_\infty}\right) \rightarrow \left(1 - \frac{e^{-\lambda}}{\varepsilon_\infty}\right) \tag{21}$$

where

$$\gamma = 0.7D\Gamma_{\text{GB}}$$

GB/SA methods have become very popular owing to the balance between computational efficiency and accuracy.[37–41] However, few words of caution seem necessary since GB us a semi-empirical approach to solving the Poisson–Boltzman equation[33,42] whose use implies a large range of assumptions on the nature of the solvent as a dielectric continuum. Furthermore, the current implementations of GB/SA routines in MD algorithms neglect of friction effects makes unrealistic the time scale of solute movements. In our experience, this can lead to artefacts in large MD simulations of nucleic acids when the GB/

SA approach is used. In summary, GB/SA is one of the most powerful approaches to simulate nucleic acids in physiological environments. However, caution and common sense is necessary to evaluate the goodness of the methodology to deal with a given problem.

Accuracy of atomic-level force-fields. We can distinguish between force-fields specific for nucleic acids and those of general use. The first ones are typically used in the context of collective variable models, and their parametrization was mainly performed using known structural data of nucleic acids. The general-purpose force-fields were typically developed in the context of atomic-level representation of nucleic acids, and were designed to represent both nucleic acids and proteins. These force-fields were mainly parametrized from experimental and QM data on small systems (nucleotides and peptides) and tested by calculation of medium-sized macromolecules.

Among the variety of force-fields, CHARMM and AMBER are clearly the most popular, probably because of historical reasons. Calculations performed with the first versions of both force-fields yielded reasonable representations of proteins, but rather poor ones for nucleic acids. Part of these problems were not strictly due to the force-field, but to other technical details of the simulation, particularly the treatment of long-range electrostatic effects. In 1995 both Kolman's and Karplus's groups developed a second version of their force-fields (CHARM-22;[42] AMBER-95[27]), which combined with suitable methods to treat long-range electrostatic effects (like the Particle Mesh Ewald approach; PME[43]) reproduced quite accurately many static properties of nucleic acids, and yielded stable MD trajectories in physiological conditions without constraints (see refs. 5,7,8). However, these force-fields present several shortcomings. For example, AMBER-95 underestimates the average twist of DNA, and has problems in reproducing the nucleoside conformation. On the other hand, CHARMM-22 was too A-philic, and drives DNA structures to unrealistic conformations. Refinements of AMBER-95 dihedral terms by Cheathman *et al*.[28] and CHARMM-22 by Mackerell and coworkers[33,34] led to AMBER-99 and to CHARMM-27, respectively. Independently, using a hybrid strategy and mixing CHARMM-22 and AMBER-95 parameters, Langley derived the BMS force-field,[44] which can be considered another refined version of CHARMM.

The three latest force-fields (AMBER-99, CHARMM-27 and BMS) provide accurate representations of standard DNA and RNA structures. The root-mean square deviation (RMSD) of the simulated nucleic acids with respect to experimental structures is small, the dihedral distributions are correct, and the average helical parameters are also reasonably close to the accepted experimental values.[8,27,28,33,34,44] Several details of sequence-dependent conformational properties of DNAs are also well reproduced (see ref. 8 for discussion). Interestingly, coupling of these force-fields to MD algorithms allow the representation of some basic transitions between nucleic acid conformations, like the A $\rightarrow$ B transition of duplex DNA in water[45] or the same transition in parallel DNA triplexes.[46] Very impressively, Beveridge's group recently showed[47] that MD calculations using one of these force-fields (AMBER-95) reproduced crude NMR data (NOESY volumes and dihedral angles) with good accuracy. In fact, the theoretical models generated by the MD simulation using AMBER-95 reproduce NMR spectra better than X-ray or fibber diffraction structures.

However, despite their success, some caution is still necessary with the latest generation of force-fields, since the refinement process which ensures the quality of simulated structures might *bias* the conformational space accessible to the nucleic acid. This not only might generate problems in the simulation of non-standard nucleic acid structures, but also in the general representation of nucleic acids flexibility. The later issue was pointed out by Cheatham and coworkers,[8] who detected that the refined AMBER-99 force-field has poorer sampling ability than AMBER-95, it being unable to reproduce (on the ns time scale) some transitions and conformational movements which were well represented by AMBER-95. For a detailed discussion of the characteristics of the different force-fields, we address the reader to the original papers[26,27,32,33,43] and to refs. 5, 7, 8 and 48.

Choice of one force-field over the other conforms in most cases to tradition, to the possibility to obtain new parameters, and to knowledge of the force-field characteristics. BMS seems to be the most rigid force-field, but that which better reproduces experimental structures of duplex DNA and RNA, and is excellent in reproducing A$\leftrightarrow$B conformational changes. AMBER-99 still underestimates twist in canonical DNA duplexes compared with X-ray data, and has problems to reproduce B$\leftrightarrow$A transitions, but seems to be a quite well balanced force-field. Finally, CHARMM-27 shows similar characteristics to AMBER-99 (something expected considering the similarity of the parametrization process), but improves the average twist of the DNA duplexes at the expense of a reduction in the flexibility of the system.[8] Clearly, all these conclusions are obtained from the analysis of standard nucleic acids. For non-standard structures, comparison with experiment, and accordingly benchmarking of the force-fields is more difficult. In the last years, our group has been using AMBER-95 and AMBER-99 for the simulation of many anomalous nucleic acid structures (including triplexes, tetraplexes, PNA-hybrids, mutated DNAs, RNA·DNA hybrids *etc.*), and we feel in general satisfied with the performance of these force-fields.

Though the ability to reproduce experimental structures of nucleic acids is a great success of current force-fields, further efforts are necessary to check the quality of the force-field by independent tests. In this area, it is worth mentioning the studies performed by Hobza and coworkers comparing QM estimates of nucleobase-nucleobase (including stacking and H-bonds complexes) interaction energies with force-field values.[49] For the force-fields available at that time, the best results were obtained with AMBER-95/99,[50] which shows a surprisingly good ability to reproduce MP2 data. Our group has also studied numerous (normal and anomalous) nucleic acid structures containing modified nucleobases, which were not present in the original AMBER-95/99 force-field, like 8-aminopurines, difluorotoluene, uracil derivatives, minor tautomers of purines and pyrimidines, isoguanine, oxanosine, thioguanine and many other non standard purines and pyrimidines. Force-field parameters for the modified nucleobases were typically tested by i) comparison of force-field estimates of nucleobase-nucleobase interaction energies with MP2 and B3LYP calculations, ii) comparison of force-field and B3LYP estimates of the energy of selected water–nucleobase complexes, and iii) comparison of force-field and MST/6-31G(d) estimates of hydration free energy. In all cases we have found that the parametrization procedure used in AMBER is powerful and robust, and the fitted parameters accurately reproduced QM data.

As an additional test on the quality of AMBER-95/99 force-field, we examined the ability of simple force-field calculations to predict the DNA conformation as a stable arrangement for the constituent nucleobases. For this purpose, we first defined potential energy maps corresponding to the interaction of two nucleobase pairs with all the helical parameters set to the standard B-type ones, except two of them (see Figure 3) which were systematically varied within a reasonable range. In parallel, we analysed the B-DNA structures deposited in the December 2002 PDB release using Olson's 3DNA program[18] to characterize the helical parameters for different base-pair steps. Figure 3 compares the results of both energy and database analysis performed here. Two types of useful information appear: i) within the steric constraints imposed by the helical backbone, the local geometry of the many base-pair steps is

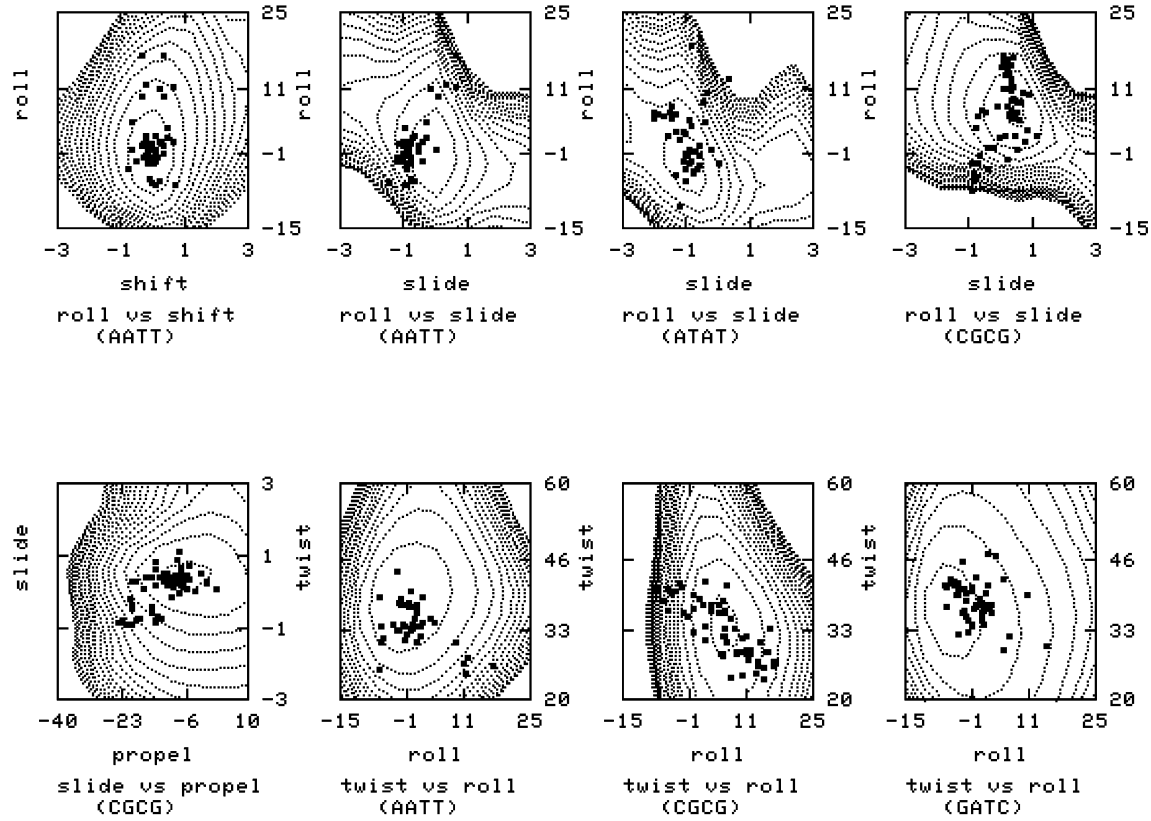


Fig. 3 Selection of some base-base interaction energy energy maps of different base pair dimers (contours are displayed each 1.5 kcal/mol). The maps are obtained in this work by changing two helical parameters (X and Y) while the rest are set at equilibrium values. The projection of the B-DNA type structures in Dec 2002 release of PDB on the X|Y map are displayed as square symbols in the map.

dominated by nucleobase–nucleobase interactions, *i.e.*, many properties of DNA can be represented at the base-pair dimer level (as done in mesoscopic models), and ii) AMBER-95/99 force-field seems able to define with accuracy the most stable geometries for the pairs of nucleobases.

In summary, atomic-level force-fields provide good representations of nucleic acid structures in dilute aqueous solutions in the presence of Na^+ or K^+ as counterions. Clearly, some improvement in the parameters might be possible if new high quality experimental data is introduced in the parametrization. However, in our opinion, current force-fields are close to convergence, and no further major improvements appear evident unless the general formalism of the force-field is changed. In fact, further refinements guided by an over-concern in reproducing known experimental structures might lead to unbalanced force-fields unable to reproduce structures or properties not considered in the calibration. In our experience the most relevant shortcomings of current force-fields involve the representation of interactions with bivalent cations like Ca^{2+} or Mg^{2+}, which strongly influence the charge distribution. The limitation of current force-fields to deal properly with these two cations is especially disappointing considering their physiological importance and their occurrence in X-ray structures. In the near future, we can expect problems in force-field related to the neglect of polarization and charge-transfer effects in simulations of nucleic acids regarding the introduction of cationic species of Pt, Zn, Ag, Au, or Cd. Either the generalization of QM/MM methods or the use of new force-fields including polarization and charge-transfer terms seems necessary.

Post-processing of force-field calculations

The output of a force-field calculation is the energy associated with a given nuclear configuration. This information can be used directly or post-processed using different algorithms. According to the post-processing we can then divide force-field based methods in four categories: i) rigid calculations, ii) molecular mechanics, iii) Monte Carlo (MC), and iv) molecular dynamics (MD).

Rigid calculations. The simplest calculations are those based on a rigid structure, which can be useful to obtain average representations of the intrinsic reactivity of nucleic acids. Single point calculations, which combine the non-bonded part of force-fields with continuum representations of the solvent,[32] provide very pictorial representations of the intrinsic ability of nucleic acids to interact with a given probe molecule. For example, Figure 4 shows the regions where a Na^+ molecule ("the probe molecule") is most likely to interact with a 12-mer RNA. Despite its static nature, this information is very useful to describe the ability of nucleic acids to interact with small cationic groups like minor groove binders. For example, it was exploited to predict possible protein–triplex DNA interactions,[46] and to design chemical changes that stabilize parallel triplexes.[46,50]

Molecular mechanics (MM). These methods use gradient techniques to find conformations of nucleic acids which minimize their potential energy. MM methods have been widely used due to their formal simplicity and reduced computational cost. Unfortunately, they present several shortcomings: i) they do not provide a dynamic picture of the system, and ii) for large systems optimisations are typically trapped in local energy minima close to the starting configurations, but far from the absolute minimum. In our opinion, MM calculations should be considered only as a preliminary step to Monte Carlo or MD simulations. More powerful are MM calculations in the context of collective variable representations of nucleic acids, where

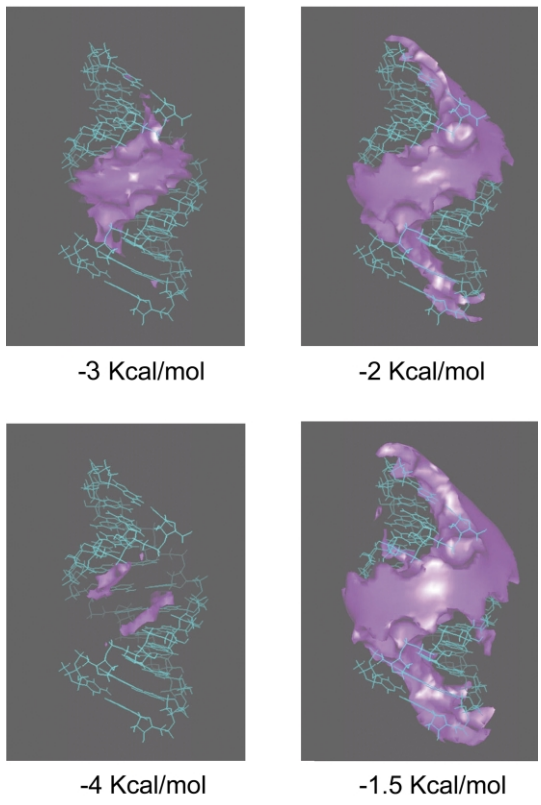


Fig. 4 Molecular interaction potential maps for the RNA–Na^+ interaction. Structure of RNA was obtained by averaging a 6 ns MD simulation.

only a reduced number of significant degrees of freedom are considered, and where the solvent is implicitly included.

Interesting MM calculations in the context of collective variable models have been reported (see ref. 4 for a recent reviews) by Olson's and Lavery's groups.[51,53] For example, Olson and coworkers analyzed the energy profile of DNA and RNA under the action of stretching and compression forces. The extreme plasticity of nucleic acid duplexes was reflected in compression/stretching factors of $\frac{1}{2}$ or 2 without a dramatic decrease in stability.[22,52] Using similar techniques, the same group also investigated the DNA bending induced by asymmetric phosphate neutralization,[53] a process which can have a dramatic impact in the formation of nucleosomes, and other DNA–protein complexes. Using JUNMA methodology, Lavery and coworkers examined a large variety of deformations in DNA.[4,51] Within this context, they introduced the concept of "lexide", which has been implemented in the ADAPT-JUNMA method.[19,51] A "lexide" is defined as a hybrid nucleobase, which contains different percentages of each natural nucleobase. To define a lexide in a given position of the DNA, the four nucleobases have to be superimposed along their C1′–N1/N9 glycosidic bonds. The four nucleobases do not see each other, and their interactions with the rest of the DNA are modulated by a coefficient ranging from 0 to 1. For example, a "lexide" defined by coefficients $C_A = 1$, $C_G = 0$, $C_T = 0$, $C_C = 0$ will be equal to an adenine, while a "lexide" defined by coefficients $C_A = 0.5$, $C_G = 0.5$, $C_T = 0$, $C_C = 0$, will be equal to a generic purine (see eqn. 22). The energy of a DNA containing different "lexides" will be computed as shown in equation 22.

$$E = E_{\text{intra}} + \sum_{i,j}^{no-X} E_{ij} + \sum_{i \in no-X} \sum_{k \in X}^{k<i} \sum_{n=1}^{4} c_{nk} E_{nki} + \sum_{k,l \in X} \sum_{n=1}^{4} c_{nk} c_{nl} E_{nkl} \quad (22)$$

where E_{intra} stands for the intramolecular term, the second term accounts for all interactions involving normal nucleobases, the third term accounts for the interactions between the lexides (X) and normal nucleobases and the last term for lexide–lexide interactions.

The use of "lexides" in collective-variable MM algorithms allows the performance of very fast calculations of combinational studies of sequence effects on DNA structure. For example, let us assume that two conformations exist for a given DNA, the normal one (N) and a distorted one (D), and we wish to study how different sequences favour/disfavour the N→D transition. This will be almost impossible with usual methodologies if the number of studied sequences is large. However, the same calculation is simple in JUNMA-ADAPT framework, since only one optimization (with lexides in all the positions of interest) is performed for each (N, D) state. Deconvolution of the computed energy using eqn. 22 will then provide almost instantaneous estimates of how well a given sequence will adapt to the N→D transition.

Monte Carlo (MC). This approach is typically applied within the traditional Metropolis implementation, and can be considered as the simplest alternative to obtain an average picture of the configurational space accessible to molecules. The method generates a Markov's chain of configurations, where a new configuration {X} is randomly generated from a previous one {X_0} and accepted or rejected based only on the relative stability of configurations {X} and {X_0}. The generation of a new configuration is made by perturbation of the previous one, the perturbation being fitted to obtain a reasonable acceptance ratio (typically around 40%). To improve the efficiency in sampling, MC methods work in the internal space, and the degrees of freedom considered are directly controlled by the user, which makes MC methods well suited for the collective variables approach.

MC methods became popular years ago as a cheap strategy to study DNA interactions. In these studies, the DNA was considered to be a rigid molecule surrounded by one of several interacting particles free to move around it. Other authors, including Zhurkin and coworkers, have used Metropolis MC techniques to study the influence of thermal fluctuations in the bending of A-track sequences of DNA,[21] but in general the use of MC to study flexible nucleic acids is rare. The reason is probably related to the intrinsic problems of these techniques to deal with conformational changes in long solvated polymers, where the internal coordinates are intercorrelated in a very complex way. Different authors (for a review see ref. 4) have suggested possible solutions to this problem by using implicit solvent models and the collective variables approach, and future popularisation of the use of MC techniques in DNA can be expected.

Molecular dynamics. MD is probably the most popular computational strategy for the study of flexible nucleic acids. The method is based on the integration of the equations of Newton's equations of motion. A MD calculation starts with a set of initial coordinates (derived from experimental data or modelling) and velocities (typically generated randomly at a given temperature). The force-field determines the potential energy and the forces acting on the system, and Newton's second law is then used to determine accelerations on each particle. Numerical integration of these accelerations provides a set of new velocities and positions, which are collected to build up a trajectory. Typically, time steps of 0.5–2 fs are used for integration of Newton's laws of motion, which implies that 1 ns of MD trajectories need around 10^6 calculations of energies and forces acting on the system. Current MD protocols include algorithms to fix the temperature and the pressure, allowing then the simulation of nucleic acids under conditions close to the physiological ones.

MD is a technique that naturally works on Cartesian coordinates, and accordingly it is typically applied with atomic-level representation of nucleic acids. Bond lengths (all or just those involving hydrogens) are the only degrees of freedom which are typically frozen in MD simulations. Most MD

simulations of nucleic acids are done using explicit solvent representations including thousands of water molecules and periodic boundary conditions (PBC). Ions (typically Na^+ and Cl^-) are introduced to neutralize the system and simulate a given ionic strength.

The MD simulation of nucleic acids is especially complex due to the strong interactions between charged particles. In fact, until 1995 unrestrained MD simulations of DNA duplexes in solution led to unfolded structures after a few hundreds picoseconds irrespective of the force-field used in the calculation. The situation improved dramatically with the introduction of efficient variants of the Ewald summation technique (like the PME method[43]) to account for long-range electrostatic effects. Clearly, when PME-PBC conditions are used, artefactual periodicity is introduced in the simulation of diluted nucleic acids, which *a priori* might *bias* the simulations. However, in our experience if large boxes of water are used to solvate nucleic acids, the PME-PBC treatment has not a dramatic influence in the trajectories.

Current state-of-the-art PME-PBC MD simulations of medium-size nucleic acids (like a 12-mer duplex DNA) cover around 5–10 ns of unrestricted trajectory. This time scale is large enough as to sample conformational space around minima, or to detect fast transitions, as for the A→B transitions in duplex and triplex DNA in water.[45] However, many conformational transitions occur in a longer time scale, and cannot be captured by current MD simulations. These slow movements include not only folding/refolding of DNA duplexes or the local opening a duplex, but also subtle changes like some sugar re-puckerings, most breathing movements, or the Na^+ reorganization around DNA. Caution is then necessary to evaluate the ability of MD simulations to provide correct picture of a conformational transition in nucleic acids.

In our opinion, incomplete sampling is the most important source of uncertainties in current MD simulations of nucleic acids. Extension of trajectories can be obtained by three sources: i) increase in the speed of the force/energy calculation, ii) use of simpler Hamiltonians or systems, and iii) increase in the integration time step. The increase in the power of computers has been crucial for the generalization in the use of MD codes. In our own group, the improvement in computer power explains a 10-fold increase in the length of the trajectories in 5 years. Very recently, the popularisation of Lynux-clusters has stimulated the parallelization of codes, leading to more efficient programs. The existence of these new computer platforms explains also the tendency to replace one single large trajectory by many smaller trajectories. This strategy might be successful to analyse systems in equilibrium, but caution is necessary if this approach is used to analyse, for example, conformational transitions.

The reduction in the number of particles and degrees of freedom of the system is typically obtained by using simplified solvent models. A first possibility is the use of microsolvation conditions combined with some damping of the phosphate–phosphate repulsion, but in general this is not advisable, at least for atomic-detailed MD simulations of nucleic acids, since very distorted structures are obtained (T. Cheatham private communication). A more popular strategy is to represent the solvent by using a continuum model, which avoids the need to introduce explicit solvent molecules. Several codes including AMBER and CHARMM incorporate efficient implementations of the GB/SA model (see above). As noted before, we just underline that these methods can be accurate enough in some cases, but in general we recommend caution in the use of this methodology, which is far from being a black-box.

The use of collective variables models in MD simulation raises many technical problems, but might allow *a priori* some improvement in the efficiency of sampling the conformational space. This has been exploited by Mazur, who using the Internal Coordinate method[54–57] in conjunction with AMBER-95/99 force-fields studied different structural and conformational aspects of the duplex DNA. Using this technique, combined with the microsolvation scheme, reduction of phosphate charges and use of a distance-dependent dielectric constant, Mazur obtained stable trajectories of Dickerson's dodecamer [54] and reproduces the A-track induced bending in DNA.[55] Recently,[56] a version of the method which incorporates Ewald summation technique has been developed (the periodic system considered here is a water drop containing the DNA placed in a rectangular box surrounded by vacuum) and used to study the Na^+ atmosphere around DNA. Clearly, the great advantage of Mazur's method is the possibility to reduce the degrees of freedom of the system, freezing those involved in the fastest vibrations, and allowing then the use of a very large integration step (10 fs). Once again some caution is necessary, since the stability of a trajectory does not guarantee its quality, but it can reflect a poor ability to sample the conformational space.

Analysis of the MD trajectories. The analysis of the trajectories is a key issue in a MD simulation. For systems in evolution the trajectory shows the movement from one unstable state to a stable one. For systems in equilibrium, the trajectory represents a Boltzman's sampling of the configurational space accessible to the molecule, which can be used to derive statistical descriptors of the system. The approaches developed to analyse the trajectories can be classified in four major categories: i) average structural data, ii) dynamic structural information, iii) interaction profiles, and iv) stability (free energy) information.

Average structural data. This is obtained by analyzing the MD-averaged structure, which is derived by averaging the Cartesian coordinates of the nucleic acid along a stable portion of the trajectory and further refining by MM algorithms. The MD average structure can be analyzed for any specific structural detail (intra- or intermolecular distances, backbone dihedrals, H-bonds, solvent accessible surface, radii of gyration,…), or simply compared with available experimental data. MD-averaged structures are typically analyzed using standard programs such as Curves,[16] NewHelix[17] or 3DNA,[18] which provide a very complete set of helical parameters defining the structure of the nucleic acid according to accepted rules.

Dynamic structural information. This is obtained from the analysis of thousands of the structures collected at constant intervals (typically 1–5 ps) during a stable part of the trajectory, which are analyzed using similar codes to those considered for the MD-averaged structure. Thus, information is gained not only of the average value of a geometrical parameter, but also of its time evolution and flexibility. All this information can be processed to characterize conformational changes or molecular flexibility. In some cases, comparison with experimental *B*-factors (eqn. 23) can be performed, but in other cases no direct comparison with experimental data is possible, and caution must be taken, since the flexibility detected in MD simulation can be influenced by the force-field and subtle simulation details.

$$B = (8\pi^2/3) < \Delta r^2 > \qquad (23)$$

where Δr are the atomic fluctuations with respect to the average position.

Recently, essential dynamics has been used to perform a more quantitative description of the motions of nucleic acids. The technique determines the motions of a structure that explains most of the variance detected during the trajectory. Technically, this implies the diagonalization of the covariance matrix leading to a set of $3N$ (N = number of atoms in the system) eigenvalues and eigenvectors. One eigenvalue represents the percentage of variance explained by the corresponding eigenvector (v_i). By using harmonic approximations, the eigenvalues can be translated into frequencies, which indicate the softness of a given essential motion. The eigenvectors can be

considered to be 3*N* dimensional vectors representing the nature of the essential motion. The eigenvectors of one trajectory can be compared with those of another trajectory, deriving quantitative measures of the similarity between the essential motions of two independent trajectories (see eq. 24).

$$\gamma_{AB} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}(\nu_i^A \bullet \nu_j^B)^2 \qquad (24)$$

where *n* stands for the minimum number of eigenvectors which explain more than a given threshold of the variance (typically we use 90% in simulations of nucleic acids) of the trajectories (A and B).

Essential dynamics calculations are very useful for determining MD trajectories. However, caution is needed in the analysis, since the technique is very sensitive to the extension of the trajectory and to numerical errors in the calculations, and can neglect important local distortions in favour of general, but perhaps less relevant movements. It is very important to keep in mind that essential motions are detected only if they occur in the MD trajectory, but slow motions might be difficult to detect in current "state-of-the-art" simulations (5–10 ns). Systematic studies performed in our group dividing large trajectories of DNA into smaller non-overlapped subtrajectories found similarity indexes around 0.7–0.9 when two equal portions of the same trajectory are compared, indicating a non-negligible dependence of the results on the initial conditions and numerical errors in the simulation. Additional sources of error exist in the definition of a common reference system for the trajectories, and on the elimination of translational and rotational degrees of freedom of the molecule. Caution and common sense is then necessary for a reasonable use of the very powerful essential dynamics tool.

Following the philosophy of essential dynamics, Lankas *et al.*[58] have developed an interesting method to derive elastic properties of standard nucleic acids from extended nucleic acid calculations following ideas previous developed previously by Olson and coworkers.[59] The method projects the Cartesian coordinates collected along the dynamics in a reduced set of 4 internal coordinates, which are derived from Curves calculations, and represents several common deformations of the DNA duplex. The elastic energy (defined as the minimum work necessary to deform the fragment out of equilibrium) is determined from the oscillations of these four variables around their equilibrium values using a complex equation. The force-constants defining the different types of elasticity are obtained by inversion of the covariance matrix obtained in the reduced four-dimensional space. The method has been successfully applied to derive sequence-dependent elastic properties of DNA,[58] and might have a large impact in the derivation of mesoscopic models for DNA simulation. The method is clear and simple to implement, and provides a very intuitive picture of nucleic acid deformability, a phenomena difficult to characterize by other techniques. However, it relies on a simplified picture of nucleic acid deformability, whose validity for non-standard nucleic acids has not been yet demonstrated. Evidence is also needed on the dependence of the results on the length of the trajectories, since we might expect that the DNA would appear more rigid in short than in large trajectories.

The dynamic analysis of a trajectory is not limited to the macromolecule, but can be performed also for solvent molecules. In this field, radial or cylindrical distribution functions (see eqn. 25) have been used to obtain information about the solvent distribution around the nucleic acids. More powerful are solvent density maps (see Figure 5), since in this case the local solvent density is projected in regular grids, without assumption of any particular geometrical dependence (eqn. 26). Note that in the limit of a rigid molecule, the density maps allow us to obtain free energies of transfer of a solvent molecule from a random position in a pure solvent to a given position around the nucleic

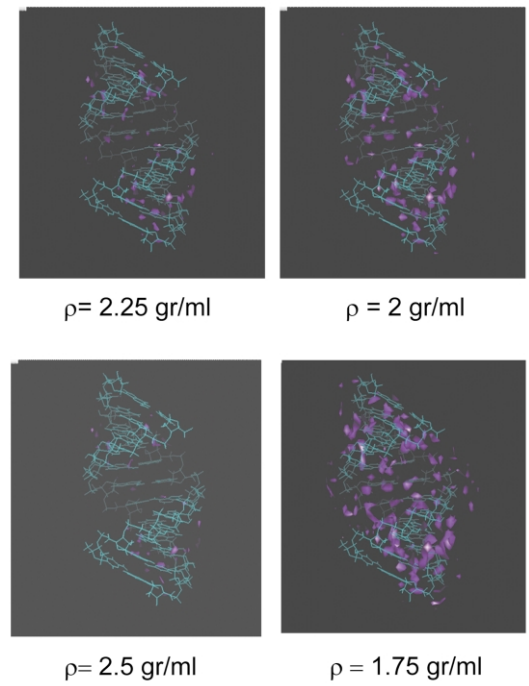


Fig. 5 Water density maps corresponding to a RNA dodecamer. Water density data are obtained from 6 ns MD samplings (see text).

acid (see eqn. 27). The criticisms of the technique are mostly related to the use a common grid for the entire trajectory (typically defined from the MD-averaged structure), which for flexible molecules can lead to an artefactual smoothing of the solvent density maps. The use of grids defined in internal coordinates allows the grid to follow macromolecular movements, leading then to more realistic solvation maps.[60] Unfortunately, the use of grids in internal coordinates has not yet become popular in the field.

$$df = \frac{\langle N_i \rangle_{r+dr}}{\rho \delta V_{r,r+dr}} \qquad (25)$$

where ρ is the density of solvent, and $\langle N_i \rangle$ is the population of solvent molecules *i* found during the dynamics in the volume element (δV) defined by radial distances *r* and *r* + d*r*.

$$dm_k = \frac{\langle N_i \rangle_k}{\rho \Delta V_k} \qquad (26)$$

where *k* stands for a grid element of grid volume ΔV_k

$$\Delta G_{s \to k} = -kT \ln(dm)_k \qquad (27)$$

Stability calculations. The information derived from the MD simulation of a nucleic acid or a nucleic acid complex can be processed to estimate the absolute or relative stability of the system. This is done by assuming a partition in the free energy of the system (see eqn. 28) in intramolecular free energy (here complexes are considered as a supermolecule), and the solvation free energy. The intramolecular free energy is in turn divided into intramolecular enthalpy and entropy (see eqn. 29). The first is obtained directly from energy calculations using the ensemble of structures collected in the trajectory, and the same force-field considered in the MD simulations (see equation 29).

$$G = G_{intra} + G_{solv} \qquad (28)$$

$$G_{intra} = H_{intra} + TS_{intra} \approx \langle E_{intra} \rangle + TS_{intra} \qquad (29)$$

The intramolecular entropy can be derived using quasi-harmonic methods like those derived by Karplus's group (see

eqn. 30 and ref. 61) or Schlitter (see eqn. 31 and ref. 62), which are based on the diagonalization of the covariance matrix and on the assumption of the harmonic oscillator for each macromolecular vibration. We have used quite systematically the Schlitter's method to derive intramolecular entropies, finding a reasonably good ability to describe the entropy of normal nucleic acids. In our hands, similar results are found with the method developed by Andricioaei and Karplus.

The possibility of obtaining entropies from MD simulations is very attractive, since entropy has been always considered the most elusive magnitude in molecular simulations. However, we cannot ignore the large number of approximations implicitly assumed. First, the harmonic treatment of vibrations might be valid only for structures fluctuating around a single minima, but not for molecules displaying a complex dynamics. Second, the entropy reflects the extension of the configurational space accessible to a macromolecule, and accordingly it is very sensitive to the length of the trajectory.[63] This is clear in Figure 6 which represents the entropy of a RNA dodecamer obtained

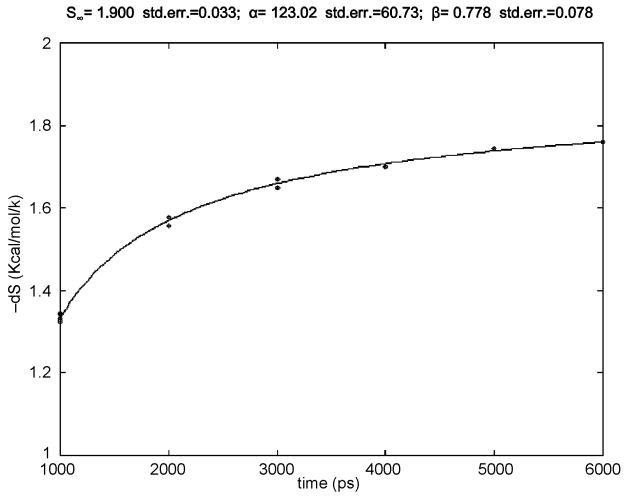


Fig. 6 Entropy estimates of a RNA dodecamer obtained from Schlitter's treatment and considering samplings obtained in windows of different size along a common trajectory.

when Schlitter's method is applied to windows of 1, 2,…6 ns of a common trajectory. Fortunately, the relationship between entropy and the length of the trajectory can be fitted in most cases to an exponential relation like that found in eqn. 32,[63] which allows us to obtain from finite simulations the entropy expected for an infinite simulation S_∞. The use of S_∞ has many advantages with respect to individual entropy estimates, but in our experience non negligible numerical uncertainties are still expected for current trajectories of nucleic acids.

$$S = k\sum_i \frac{\alpha_i}{e^{\alpha_i}-1} - \ln(1-e^{-\alpha_i}) \tag{30}$$

$$S = 0.5k\sum_i \ln\left(1+\frac{e^2}{\alpha^2}\right) \tag{31}$$

where $\alpha_i = \hbar\omega_i/kT$; ω being the eigenvalues (in frequency units) obtained by diagonalization of the mass-weigthed covariance matrix (see explanation of essential dynamics). The sum extends to all the non-trivial vibrations of the system.

$$S(t) = S_\infty - \frac{a}{t^b} \tag{32}$$

where *a* and *b* are fitted parameters

The calculation of the solvation free energy is performed using continuum models applied to the ensemble of structures collected along the trajectory (see eqn. 32). The GB/SA method has been used by different authors to obtain the free energy of solvation, but in our experience more consistent results are obtained with methods based on the numerical solution of the Poisson–Boltzman equation (PB/SA; for a discussion see ref. 32). However, irrespective of the method used to compute solvation free energies, the intrinsic shortcomings of continuum methods to describe solvation in complex polianionic systems cannot be ignored.[32]

$$G_{\text{solv}} = \langle G_{\text{solv}} \rangle \tag{33}$$

The use of equation 28 has became very popular for the analysis of trajectories obtained with explicit solvent representation since it was suggested by Kollman and coworkers,[64] because it provides a very intuitive description of the stability of the nucleic acids or their complexes. The intrinsic errors related to the use of equation 28 are those derived from the simplifications used for the computation of each term, and from the numerical errors in the averaging originated from the limited length of the trajectories. Strategies to reduce the noise in the calculation of molecular free energy other than the "pure force" solution of the increase in the length of the simulation must be developed to improve the ability of MM/PB-SA to describe small changes in stability. For structures with common-repetitive sequences, our group uses a strategy based on the parallel calculation of trajectories of oligonucleotides of different sizes placed in the conformations of interest. The total free energies follow a perfect linear relationship (r^2 = 1.0000 in all the cases) with the length of the oligonucleotide. The corresponding regression equations (one for each structural model considered) can then be used to derive with high statistical quality estimates of the difference in stability between two structural models.[64,65]

A more rigorous approach to compute free energy estimates in nucleic acids relies in the use of classical statistical mechanics. Among the different methods developed to compute free energy differences between two states, free energy perturbation (FEP) and thermodynamic integration (TI) have become the two most popular ones in the field. Both FEP and TI compute the free energy difference between two states by mutation, *i.e.*, one state is moved slowly to the other by a physical or unphysical reversible way. In most cases, this is achieved by coupling the system Hamiltonian to a variable λ, which changes from 0 (state A) to 1 (state B), as shown in equation 34. The free energy associated to the change is computed as displayed in equations 35 (TI) and 36 (FEP). The two states might refer, for example, to two different conformations of a molecule, the bound and the unbound form of a complex, *etc*. Furthermore, in conjunction with suitable themodynamic cycles, TI and FEP can be used to compute the impact of a given chemical change in the stability of a nucleic acid, or to determine how a chemical change in the structure of a ligand can alter the binding affinity to the nucleic acid (see Figure 7).

$$H_\lambda = (1-\lambda)H_A + \lambda H_B \tag{34}$$

$$\Delta G^{A\to B} = \sum_\lambda^{1-\Delta\lambda}\left[\int_\lambda^{\lambda+\Delta\lambda} \langle \partial E_{\lambda'}/\partial\lambda' \rangle_{\lambda'} d\lambda\right] \tag{35}$$

$$\Delta G^{A\to B} = -\sum_{\lambda=0}^{1-\Delta\lambda} kT \ln\left\langle \exp\left[(E_\lambda - E_{\lambda+\Delta\lambda})/kT\right]\right\rangle_\lambda \tag{36}$$

Equations 35 and 36 are formally exact provided the sampling at λ is correct, and the perturbations $\Delta\lambda$ are small enough to guarantee a smooth reversible pathway between states A and B. In practice, equations 34–36 are useful only when states A and B are close enough. For example, FEP and TI are very accurate to determine the impact on the stability of a nucleic acid structure of transitions between related purines or pyrimidines,[50,65,66] but they will be more noisy to study

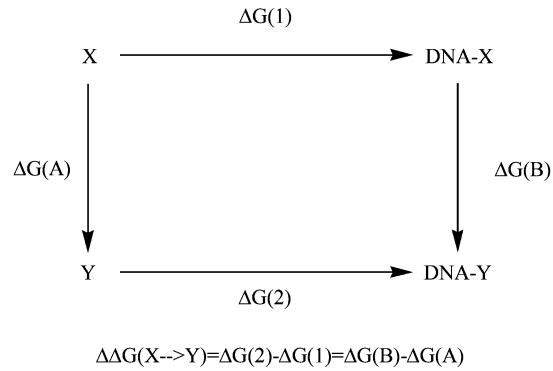


Fig. 7 Example of a thermodynamic cycle used to compute the difference in free energy of binding to the DNA of two drugs X and Y. The magnitude of interest is the $\Delta G(1)-\Delta G(2)$, while that computed is $\Delta G(B)-\Delta G(A)$

purine↔pyrimidine transitions. Similarly, these methods will be accurate to determine the difference in binding free energy between two very similar drugs, but not between two very different molecules. FEP and TI have also been used to study some transitions of nucleic acids,[67–69] but these techniques will provide accurate results only when there is a clear reaction coordinate, and when the potential energy surface is smooth. Intense research effort is being put into the development of more powerful techniques for the calculation of free energy differences based on the same principles that give rise to equations 35 and 36. Among them, methods like adaptative umbrella sampling (A-US[70]), or the related weighted histogram method (WHAM[71]) appear very promising and have been already used to study some conformational movements of the DNA.[72]

Current areas of research in MD simulations of nucleic acids

MD and related techniques are being used by a large number of research groups to study properties of nucleic acids. It is out of the scope of this paper to comment on all the relevant publication in this field, and the reader is addressed to recent reviews[3,5,7,8,64] for comprehensive reviews of MD simulations of nucleic acids. We will limit ourselves to noting the big areas of current use of MD for the analysis of nucleic acids, trying to remark on the type of information that is derived from these simulations.

Structural studies. The ability of current MD simulations to reproduce the structure of standard nucleic acids is good, but this is not surprising, since structural information on canonical nucleic acids structures has been (directly or indirectly) used in the parametrization process. More surprising is the ability of the technique to reproduce subtle structural properties like minor groove narrowing in A-tracks of B-type DNA or the bending in DNA induced by phosphate neutralization (see refs. 5,7 and 8). Even more remarkable is the astonishing capability of MD to describe anomalous nucleic acid structures, which were not considered in any way during the parametrization process: H-type RNA pseudoknots in viruses,[73] PNA-hybrids,[74] duplex RNAs,[75] DNA·RNA hybrids,[74] ribozymes,[75] tRNA,[76] i-DNA,[77] G-DNA,[78] and triplexes,[47] as well as other structures not experimentally characterized before the simulation like the Hoogsteen and the reverse Watson Crick parallel DNA duplexes.[67] In several cases the MD simulations were able to reproduce properly the structural properties of nucleic acids, even when the starting configurations were incorrect. In other cases, where the incorrect fold was found to be a metastable conformation, the MD trajectories showed the higher stability of the correct *versus* the incorrect fold. Clearly, and despite their caveats (largely commented above), MD simulations are now one of the most powerful structural tools to study nucleic acid structures.

Analysis of the solvent environment. MD provides a microscopic picture of the solvent environment around nucleic acids. The technique has been used to study the first hydration shell and the Na^+ distribution around RNA and specially DNA duplexes. Any current MD simulation is able to reproduce the "spine of hydration" of B-type DNA, and the specific solvation patterns of duplex RNA.[3,5,7,8,74,79,80] The relaxation time of Na^+ is quite large, and in our own experience, memory of the initial position of the Na^+ might exist after 15 ns of MD simulation, generating equilibration problems with the simulation. However, even with this problem, MD simulations have been useful to qualitatively describe the preferred regions for residence of Na^+. For example, the temporary residence of Na^+ ions inside the minor groove of B-type DNA, which has been subject to an intense debate among crystallographic groups in recent years, was in fact anticipated years before in unrestrained MD simulations by Beveridge's group.[81]

Analysis of the impact of chemical modifications in DNA. MD constitutes a fast alternative to experimental techniques to study the impact of mutations on the structure of nucleic acids, provided that structural changes induced by chemical alterations of the nucleotides are small. Examples of chemically-modified nucleotides studied by MD simulations included DNA methylation at cytosines,[82] DNAs containing benzo[*a*]pyrene–adenine adducts,[83] photodamaged DNAs,[84] DNA containing oxanosine,[85] inosine,[86] or DNA containing apolar isosters of the nucleobases.[87] The impact in DNA of some alterations in the backbone, including the introduction of phosphoramidates,[88] peptide nucleic acids[89] and other modifications have been also studied.

Dynamic properties of nucleic acids. The demonstration that MD was able to represent the A↔B conformational transition opened the possibility of using MD simulations to reproduce the dynamic properties of nucleic acids. Despite the shortcomings derived from the short time scale of the MD trajectories, many interesting dynamic features have been analysed, such as the breathing of natural or non-natural base pairs,[87] the dynamic properties of bendable sequences,[86] the elastic properties of different sequences of DNA,[57] or the deformability of DNA or RNA in the presence of proteins.[68,69] The shortcomings derived from the limited extension of current simulations were partially overcome by using model systems showing faster transitions, by the use of essential dynamic procedures, or by considering samplings biased following WHAM, TI, MD or US protocols.

Complexes of nucleic acids with other molecules. MD simulations have been used to describe the properties of nucleic acid–drug complexes, including both minor groove binders and intercalators.[90–92] Most calculations were carried out for canonical duplexes of DNA, but simulations of RNA–drug complexes and of complexes between anomalous DNA structures (such as the G-DNA) and small drugs have been also published. Interestingly, MD trajectories not only reproduced accurately structural details of the complexes, but also distinguished between possible binding modes or predicted binding free energies. A similar explosion in the use of MD has occurred for the study of complexes between nucleic acids and proteins.

Traditionally these studies were focused on the description of subtle structural characteristics, but with the improvement in the methods for calculation of free energies several authors have explored the stability of proteins and nucleic acids complexes.[92–96]

The future

Rationalization of the evolution of a field in the past is always easier than the prediction of its future evolution. However, few general trends for the future might be suggested. It is expected that the capabilities of computers will continue to grow and their prices will continue to fall, which is expected to lead to: i) the use of more accurate simulation systems and simulation protocols, and ii) the consideration of bigger, more realistic systems for simulation. In turn, we can expect methodological developments to extend the validity of simulation techniques to systems where current empirical force-fields do not lead to correct results.

The target of modelling studies in the nucleic acid field might change in part in the near future. More effort will be directed to the study of higher order structures of nucleic acids, and on the interaction between nucleic acids and proteins, trying to understand the molecular mechanisms which control nucleic acid function. The modeller should be ready to work with large and flexible macromolecular systems including systems as large as the ribosome. The anomalous forms of nucleic acids will focus more attention due to their biotechnological and biomedical interest, and to the difficulties of studying these systems with experimental techniques. Clearly, integration of low resolution mesoscopic techniques with atomic-detailed calculations is going to be necessary to cover the vast range of problems expected for the near future.

An important change in the field will appear as a consequence of the evolution of experimental techniques in the biochemistry laboratory. Genomic and proteomic techniques, which are now universally present in the laboratory provide massive amounts of data, which should be processed using fast coarse-grain theoretical methods. Recent works by Lavery's group, which used the "lexide" approach with rigid geometries to scan the ability of a vast amount of sequences to interact with a give protein, or the recent development by Beveridge's group of a Hidden Markov's Model trained from sequence and MD data to predict regions in the DNA with propensity to bind the CAP protein[97] are excellent examples of this convergence between modelling and bioinformatics. Clearly, an exciting future is coming for our community.

Acknowledgements

This review would have been impossible without the help of many distinguished colleagues, who provided comments, suggestions, and in some cases access to unpublished material. Among them, we must cite C. Laughton, W. K. Olson, J. Sponer, D. Beveridge, T. Cheatham, R. Lavery, V. Zhurkin, L. Pardo and C. J. Cramer. We want to emphasize discussions about microsolvation schemes with Professor Cheatham, and with Professor Laughton on the use of GB/SA for MD simulations of nucleic acids. M. O. thanks A. López for support and encouragement. The Catalan Supercomputer Center (CESCA) and the Centre de Parallelisme de Barcelona (CEPBA) are acknowledged for providing access to computer resources. This work has been supported by the Spanish Ministry of Science and Technology (SAF2002–4282 and PM99–0046). A. Pérez is a fellow of the Catalan Science Organization (CIRIT), and A. Noy of the Spanish Ministry of Science and Technology (MCyT).

References

1 J. D. Watson and F. H. C. Crick, *Nature*, 1953, **171**, 737.
2 W. K. Olson, *Curr. Opin. Struct. Biol.*, 1996, **6**, 242.
3 P. Auffinger and E. Westhof, *Curr. Opin. Struct. Biol.*, 1998, **8**, 227.
4 I. Lafontaine and R. Lavery, *Curr. Opin. Struct. Biol.*, 1999, **9**, 170.
5 D. L. Beveridge and K. J. McConnell, *Curr. Opin. Struct. Biol.*, 2000, **10**, 182.
6 W. K. Olson and V. B. Zhurkin, *Curr. Opin. Struct. Biol.*, 2000, **10**, 286.
7 T. E. Cheatham and P. A. Kollman, *Ann. Rev. Phys. Chem.*, 2000, **51**, 435.
8 T. E. Cheatham and M. A. Young, *Biopolymers*, 2001, **56**, 232.
9 E. Giudice and R. Lavery, *Acc. Chem. Res.*, 2002, **35**, 350.
10 P. Carloni, U. Rothlisberger and M. Parrinello, *Acc. Chem. Res.*, 2002, **35**, 455.
11 A. Malhotra, R. K. Tan and S. C. Harvey, *Biophys. J.*, 1994, **66**, 1777.
12 R. B. Laughlin, D. Pines, J. Schmalian, B. P. Stojkovic and P. Wolynes, *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 32.
13 N. Bruant, D. Flatters, R. Lavery and D. Genest, *Biophys. J.*, 1999, **77**, 2366.
14 A. Matsumoto and W. K. Olson, *Biophys. J.*, 2002, **83**, 22.
15 J. A. D. Wattis, S. Harris, C. R. Grindon and C. A. Laughton, *Phys. Rev. E.*, 2001, **63**, 61903.
16 R. Lavery and H. Sklenar, *J. Biomol. Struct. Dyn.*, 1989, **6**, 655.
17 R. E. Dickerson, NewHelix Program. University of California at Los Angeles. 1992.
18 X.-J. Lu, Z. Shakked and W. K. Olson, *J. Mol. Biol.*, 2000, **300**, 819; 3 DNA Program, Rutgers University, 2001.
19 R. Lavery, K. Zakrzewska and H. Sklenar, *Comput. Phys. Commun.*, 1995, **91**, 135.
20 S. C. Harvey, C. Wang, S. Teletchea and R. Lavery, *J. Comput. Chem.*, 2003, **24**, 1.
21 V. B. Zhurkin, N. B. Ulyanov, A. A. Gorin and R. L. Jernigan, *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 7046.
22 K. M. Kosikov, A. A. Gorin, V. B. Zhurkin and W. K. Olson, *J. Mol. Biol.*, 1999, **289**, 1301.
23 V. B. Zhurkin, V. I. Poltev and V. L. Florent'ev, *Mol. Biol. USSR (Engl. Ed.)*, 1981, **14**, 882.
24 R. Lavery, K. Zakrzewska and A. Pullman, *J. Comput.Chem.*, 1984, **5**, 363.
25 F. A. Momany, *J. Chem. Phys.*, 1978, **82**, 592.
26 C. I. Bayly, P. Cieplak, W. D. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269.
27 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Fergurson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **117**, 5179.
28 T. E. Cheatham, P. Cieplak and P. A. Kollman, *J. Biomol. Struct. Dyn.*, 1999, **16**, 845.
29 L. Wang, B. E. Hingerty, A. R. Srinavasan, W. K. Olson and S. Broyde, *Biophys. J.*, 2002, **83**, 382.
30 B. E. Hingerty, R. H. Ritchie, T. L. Ferrel and J. E. Turner, *Biopolymers*, 1985, **24**, 427.
31 E. L. Mehler and T. Solmajer, *Protein Eng.*, 1991, **4**, 903.
32 M. Orozco and F. J. Luque, *Chem. Rev.*, 2000, **100**, 4187.
33 N. Foloppe and A. D. Mackerell, *J. Comput. Chem.*, 2000, **21**, 86.
34 A. D. Mackerell and N. K. Banavali, *J. Comput. Chem.*, 2000, **21**, 105.
35 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L Klein, *J. Chem. Phys.*, 1983, **79**, 926.
36 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren and J. Hermans, in *Intermolecular Forces*. B. Pullman (ed). Reidel, Dordrecht, 1981, p. 331.
37 W. C. Still, A. Tempcyck, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.*, 1990, **112**, 6127.
38 G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *Chem. Phys. Lett.*, 1995, **246**, 122.
39 D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, *J. Phys. Chem. A.*, 1997, **101**, 3003.
40 J. Srinavasan, M. W. Trevathan, P. Beroza and D. Case, *Theor. Chem. Acc.*, 1999, **101**, 426.
41 A. Onufriev, D. A. Case and D. Bashford, *J. Comput. Chem.*, 2002, **23**, 1297.

42 A. D. McKerell, J. Wiorkiewicz-Kuczera and M. Karplus, *J. Am. Chem. Soc.*, 1995, **117**, 11946.
43 T. A. Darden, D. M. York and L. G. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089.
44 D. R. Langley, *J. Biomol. Struct. Dyn.*, 1998, **16**, 487.
45 T. E. Cheatham and P. A. Kollman, *J. Mol. Biol.*, 1996, **259**, 434.
46 G. C. Shields, C. A. Laughton and M. Orozco, *J. Am. Chem. Soc.*, 1997, **119**, 7463.
47 H. Arhanari, K. J. McConnell, R. Beger, M. A. Young, D. L. Beveridge and P. H. Bolton, *Biopolymers*, 2003, **68**, 3.
48 M. Feig and B. M. Pettitt, *Biophys. J.*, 1998, **75**, 134.
49 P. Hobza, M. Kabelac, J. Sponer, P. Mejzlik and J. Vondrasek, *J. Comput. Chem.*, 1997, **18**, 1136.
50 R. Güimil, E. Ferrer, N. J. Macias, R. Eritja and M. Orozco, *Nucleic Acids Res.*, 1999, **27**, 1991.
51 I. Lafontaine and R. Lavery, *Biophys. J.*, 2000, **79**, 680.
52 V. Katrich, C. Bustamante and W. K. Olson, *J. Mol. Biol.*, 2000, **295**, 29.
53 K. M. Kosikov, A. A. Gorin, X.-J. Lu, W. K. Olson and G. S. Manning, *J. Am. Chem. Soc.*, 2002, **124**, 4838.
54 A. K. Mazur, *J. Am. Chem. Soc.*, 1998, **120**, 10928.
55 A. K. Mazur, *J. Am. Chem. Soc.*, 2000, **122**, 12778.
56 A. K. Mazur, *J. Am. Chem. Soc.*, 2002, **124**, 14707.
57 A. K. Mazur, *J. Am. Chem. Soc.*, 2003, **125**, 7849.
58 F. Lankas, J. Sponer, P. Hobza and J. Langowski, *J. Mol. Biol.*, 2000, **299**, 695.
59 W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock and V. B. Zhurkin, *Proc. Natl. Acad. Sci., USA*, 1998, **95**, 11163.
60 R. Radner and P. Kollman, *J. Comput. Aided Mol. Des.*, 1998, **12**, 215.
61 I. Andricioaei and M. Karplus, *J. Chem. Phys.*, 2001, **115**, 6289.
62 J. Schliter, *Chem. Phys. Lett.*, 1993, **215**, 617.
63 S. Harris, E. Gavathiotis, M. S. Searle, M. Orozco and C. A. Laughton, *J. Am. Chem. Soc.*, 2001, **123**, 12658.
64 W. Wang, O. Donini, C. M. Reyes and P. A. Kollman, *Annu. Rev. Biophys. Biomol. Struct.*, 2001, **30**, 211.
65 E. Cubero, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2001, **123**, 12018.
66 E. Cubero, A. Aviñó, B. G. de la Torre, M. Frieden, R. Eritja, F. J. Luque, C. González and M. Orozco, *J. Am. Chem. Soc.*, 2002, **124**, 3133.
67 E. Cubero, C. A. Laughton, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2000, **122**, 6891.
68 L. Pardo, N. Pastor and H. Weinstein, *Biophys. J.*, 1998, **74**, 2191.
69 L. Pardo, N. Pastor and H. Weinstein, *Biophys. J*, 1998, **75**, 2411.
70 C. Bartles and M. Karplus, *J. Comput. Chem.*, 1997, **18**, 1450.
71 S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. Rosenberg, *J. Comput. Chem.*, 1992, **13**, 1011.
72 P. Varnai and R. Lavery, *J. Am. Chem. Soc.*, 2002, **124**, 7172.
73 K. Csaszar, N. Spackova, R. Stefl, J. Sponer and N. B. Leontis, *J. Mol. Biol.*, 2001, **313**, 1073.
74 R. Soliva, E. Sherer, F. J. Luque, C. A. Laughton and M. Orozco, *J. Am. Chem. Soc.*, 2000, **122**, 5997.
75 T. Cheatham and P. A. Kollman, *J. Am. Chem. Soc.*, 1997, **119**, 4805.
76 T. Hermann, P. Auffinger, W. G. Scott and E. Westhof, *Nucleic Acids Res.*, 1997, **25**, 3421.
77 M. C. Nagan, S. S. Kerimo, K. Musier-Forsyth and C. J. Cramer, *J. Am. Chem. Soc.*, 1999, **121**, 7310.
78 N. Spackova, I. Berger, M. Egli and J. Sponer, *J. Am. Chem. Soc.*, 1998, **120**, 6417.
79 N. Spackova, I. Berger and J. Sponer, *J. Am. Chem. Soc.*, 1999, **121**, 5519.
80 M. Feig and B. M. Pettitt, *J. Mol. Biol.*, 1999, **286**, 1075.
81 M. A. Young, B. Jayaram and D. L. Beveridge, *J. Am. Chem. Soc.*, 1997, **119**, 59.
82 S. Derreumaux, M. Chaoui, G. Tevanian and S. Fermandjian, *Nucleic Acids Res.*, 2001, **29**, 2314.
83 S. Yan, R. Shapiro, N. E. Geacintov and S. Broyde, *J. Am. Chem. Soc.*, 2001, **123**, 7054.
84 T. I. Spector, T. E. Cheathman and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **119**, 7095.
85 B. Hernández, R. Soliva, F. J. Luque and M. Orozco, *Nucleic Acids Res.*, 2000, **28**, 4873.
86 E. Sherer, S. A. Harris, R. Soliva, M. Orozco and C. A. Laughton, *J. Am. Chem. Soc.*, 1999, **121**, 5981.
87 E. Cubero, E. C. Sherer, F. Javier Luque, M. Orozco and C. A. Laughton, *J. Am. Chem. Soc.*, 1999, **121**, 8653.
88 J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case, *J. Am. Chem. Soc.*, 1998, **120**, 9401.
89 G. Shields, C. Laughton and M. Orozco, *J. Am. Chem. Soc.*, 1998, **120**, 5895.
90 S. B. Singh and P. A. Kollman, *J. Am. Chem. Soc.*, 2001, **123**, 8902.
91 B. Wellenzohn, R. H. Winger, A. Hallbrucker, E. Mayer and K. R. Liedl, *J. Am. Chem. Soc.*, 2000, **122**, 3927.
92 H. Han, D. R. Langley, A. Rangan and L. H. Hurley, *J. Am. Chem. Soc.*, 2001, **123**, 8902.
93 D. R. Langley, T. W. Doyle and D. L. Beveridge, *J. Am. Chem. Soc.*, 1991, **113**, 4395.
94 B. Jayaram, K. J. McConell, S. B. Dixit and D. L. Beveridge, *J. Comput. Phys.*, 1999, **151**, 333.
95 D. M. Blakaj, K. J. McConnell, D. L. Beveridge and A. M. Baranger, *J. Am. Chem. Soc.*, 2001, **123**, 2548.
96 B. Jayaram, K. J. McConnell, S. B. Dixit, A. Das and D. L. Beveridge, *J. Comput. Chem.*, 2002, **23**, 1.
97 K. M. Thayer and D. L. Beveridge, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 8642.