PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

Physical Chemistry Chemical Physics
CREATED USING THE RSC LaTeX PCCP ARTICLE TEMPLATE – SEE www.rsc.org/electronicfiles FOR DETAILS
ARTICLE TYPE www. rsc.org/xxxxxx | XXXXXXXX

Genetic algorithms coupled with quantum mechanics for refinement of force fields for RNA simulation: a case study of glycosidic torsions in the canonical ribonucleosides. †

Rodrigo B. Kato,[a†] Frederico T. Silva,[b‡] Gisele L. Pappa,[a¶] and Jadson C. Belchior*[b¶]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

We report the use of genetic algorithms (GA) as a method to refine force field parameters in order to determine RNA energy. Quantum-mechanical (QM) calculations are carried out for the isolated canonical ribonucleosides (adenosine, guanosine, cytidine and uridine) that are taken as reference data. In this particular study, the dihedral and electrostatic energies are reparametrized in order to test the proposed approach, i.e, GA coupled with QM calculations. Overall, RMSE comparison with recent published results for ribonucleosides energies shows an improvement, on average, of 50%. Finally, the new reparametrized potential energy function is used to determine the spatial structure of RNA (PDB code 1r4h) that was not taken into account in the parametrization process. This structure was improved about 82% comparably with previously published results.

1 Introduction

It is well known that molecular dynamics (MD) techniques can provide accurate data for analyzing properties of a specified system[1]. However, this analysis is strongly dependent on the level of precision of the force field describing the system. The precision of the force field is inverse proportional to the size of the system as well as the amount of electrons taken into account[2]. Usually one needs to also consider correlation effects, which increase the demand for computational time.

The most effective way of obtaining the energy of a system is based on QM calculations, and several methods with different levels of precision are available[3,4]. Accordingly, almost all of them are highly computational time consuming, especially for larger systems. This computational time demand can be reduced by using empirical potentials (force field). In this case, the procedure is generally based on the optimization of mathematical functions with several terms that contribute to the total energy, including physical contributions, such as attractive and repulsive terms (intermolecular forces), and others that determine the internal energy contribution (intramolecular forces). The development of empirical models to describe the system is an efficient alternative to provide a way to determine energies and structures by proposing force fields.

Potential energy calculations are common in the literature and, in general, one tries to improve the force field by a reparametrization process of all or part of the terms that contribute to the total energy. For example, CHARMM (all19, all22, all27) and Amber (ff94, ff96, ff98, ff99) force fields are commonly reparametrized[3,5–9]. However, due to the approximation of pairwise additive empirical potentials taken into account to build up force fields, such a task usually decreases the precision of the results. Among others, Zgarbová *et. al.*[10] stated that this imprecision is beyond the capabilities of simple force field approximations. Hence, they have concluded that it is not surprising that different force fields often provide remarkably different descriptions of the same structure, a phenomenon known as "force field dependent polymorphism"[4,11–14].

Among other works[3,5–9], Zgarbová *et. al.*[10] recently reported a reparameterization of the glycosidic torsion (χ) as defined by Cornell *et. al.*[5] force field applied to RNA molecules. Although their work improved the description of the *syn* region and the *syn-anti* balance as well as enhance MD simulations of various RNA structures, the results showed deviations from QM calculations.

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

[a] *Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte*
[b] *Department of Chemistry, Universidade Federal de Minas Gerais, Belo Horizonte*
* *Author to whom correspondence should be addressed.*
Email: jadson@ufmg.br

This is quite common on the use of empirical force fields. Therefore, it seems that studies to improve parametrization are important in order to better compare semiclassical results against experimental or QM calculations. Particularly, Banás *et. al.*[15] recently showed tests using a reparametrized force field, and deviations for A-DNA and B-DNA structures using MD were observed.

Techniques usually applied to optimize force fields are based on gradient methods such as quasi-Newton approach[16] and Levenberg-Maquardt method[16,17]. For example, Yildirim *et. al.*[18] used steep descent method and all of the techniques aforementioned showed to be accurate in comparison with quantum results. More recently other heuristic approaches have been proposed and tested to refine parameters of force fields. For example, Sakae and Okamoto[19] applied Monte Carlo and simulated annealing to calculate the parameters of force field applied to proteins.

Alternative approaches for optimization processes have been recently proposed for parametrizing force fields for nucleic acids[20–22]. Similarly, we have recently applied GA coupled with QM calculations to determine structures and energies of the small systems, with particular attention to clusters of the alloys and compounds[23,24]. The GA method is commonly used for the calculation of the geometrical structures and energies of the molecules[25,26]. The GA applied in this paper is similar of Wang and Kollman[20] applied in hydrocarnos, but in our work it is applied to RNA canonical nucleosides and on their studies it was applied to hydrocarbons. For this reason the comparison with their results is not appropriated. We refined only the parameters of the dihedral term and afterwards an improvement was done by adding to the refinement process the electrostatic term which has not been previously reported in the literature as carried out in the present study.

Ab initio methods are widely applied to determine electronic structures of molecules. For example, the HF/6-31G(d) level of QM theory has been applied to determine the electrostatic potentials for studying hydrocarbon molecules, whereas B3LYP functional is used to investigate the conformational landscapes of the canonical and modified nucleosides[27–33] and also the pairs of the DNA/RNA bases[34–37]. In the present study we have used PBE functional[38,39] combined with 6-311++G (3df, 3pd) basis set for the comparison of our results with the theoretical data obtained at the same level of QM theory in the work[10].

Aiming to improve the precision of force field glycosidic torsion profile, this work proposes a heuristic solution based on genetic algorithms[40–42] to better estimate the reparametrization of force field to dihedral term. A comparison of our refined results with previous publication as well as with PDB (Protein Data Bank)[43] is finally analyzed.

2 Methods

Molecular Mechanics Calculations

The force field used in our calculations is represented by the sum of several energy contributions, namely, the bond stretching (E_{bond}), angle bending (E_{angle}), dihedral (E_{dih}), nonbonded electrostatic (E_{elst}) and van der Waals (E_{vdW}) terms and it is given by[10,44]

$$E_T = E_{bond} + E_{angle} + E_{dih} + E_{elst} + E_{vdW} \tag{1}$$

According to Lankas *et. al.*[45] the glycosidic torsion (χ) is one of the most relevant bond in nucleic acids. Therefore, the proposed approach based on genetic algorithms focused on the optimization of the dihedral term given by

$$E_{dih} = \sum_{all\ torsions} \sum_{n}^{n_{max}} \frac{V_n}{2}[1 + \cos(n\varphi - \gamma)], \tag{2}$$

where n is the periodicity of the torsion, n_{max} is the maximum periodicity of each torsion[5], V_n is the rotational barrier, φ is the dihedral angle, and γ is the phase angle. As in reference [10], the actual analysis has also taken into account the contribution of the electrostatic term (E_{elst}). This term plays an important role for describing force fields[46] as also recently pointed out by Yao *et. al.*[44]. Due to this importance we have also considered these charges but taken them from QM calculations at the optimized torsion energy conditions. Accordingly, in the parametrization process of the torsion energy two types of refinement were carried out. In one case we have considered only the dihedral term and all others energies (Eq. 1) were kept constant during the refinement process, namely GAN. In the second analysis, namely GAW, the electrostatic energy was considered to refine the parameters of dihedral term. The charge electronic density of electrostatic energy was calculated by the software GAMESS[47]. For these charges an average was taken due to several rotations of all rotated structures (8 angles) used in our calculations. The average density charge is then calculated as

$$\bar{q}_i = \frac{1}{N}\sum_{0}^{360} q_i, \tag{3}$$

where N defines the degree steps (in the present work are 8 angles) and q_i is the individual charge of each atom i. Therefore, taking the average charge ($\bar{q}_i$) one can provide an adequate estimative for the charge density. In this case the electrostatic energy term is written according to

$$E_{elst} = \sum_{i=1}^{N-1}\sum_{j>i}^{N} \frac{\bar{q}_i \bar{q}_j}{\varepsilon r_{ij}} \tag{4}$$

where $\bar{q}_i$ and $\bar{q}_j$ are the charges between two atoms, r_{ij} is the distance between them and ε is the dielectric constant.

In order to carry out the molecular mechanics we used the well established Tinker software[48]. This software is widely used in the literature, and examples of studies based on its calculations are described elsewhere[49–51].

Model System

The methodology used to optimize the dihedral term of the glycosidic torsion (χ) parameter of the force field, is showed in Figure 1. We applied two approaches to optimize the parameters used in the electrostatic and dihedral (specifically the glycosidic torsion) energies of the force field in order to describe RNA canonical nucleosides, as previously explained. The procedure in both cases are carried out based on a set of QM calculations to obtain a reference data (energy and structure), which are used as a template for the reparametrization of the force field using other methods, in the actual case, a GA approach.

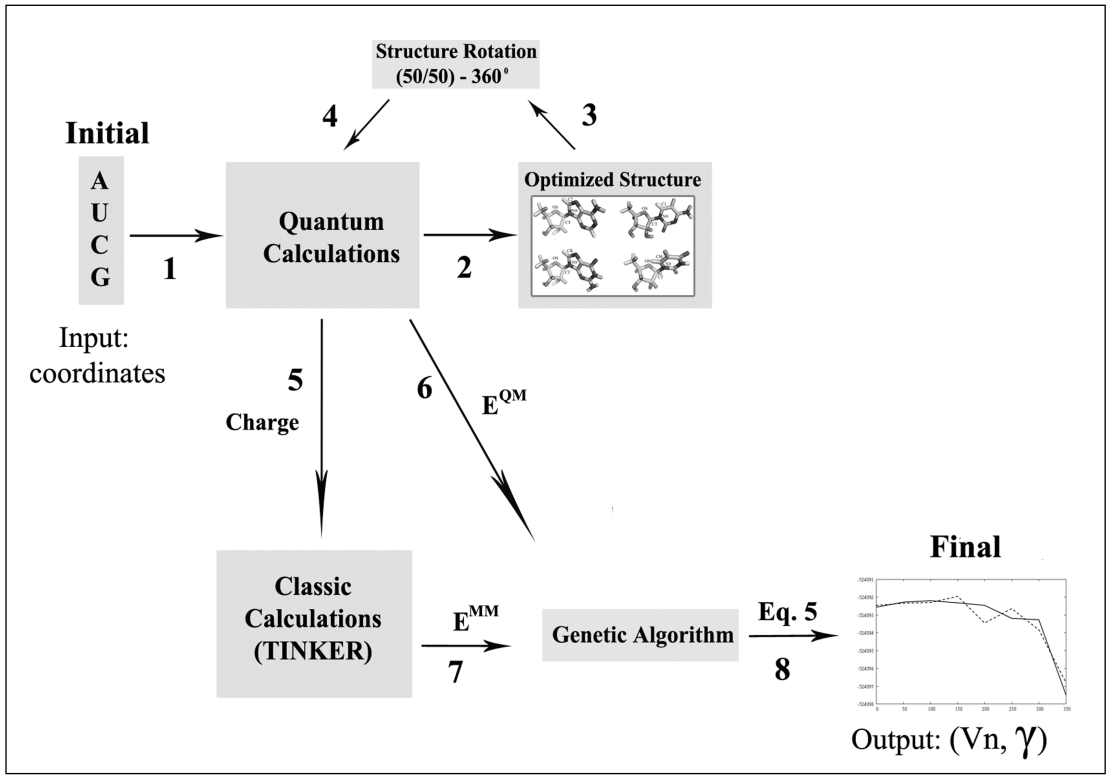

Fig. 1 Scheme (8 steps) of the proposed methodology (GAW), where A is adenosine, U is uridine, C is cytidine, G is guanosine that are defined by their coordinates.

The inputs of QM calculations are the isolated structures of each RNA nucleosides (Figure 2). In the next step these structures are optimized using QM calculations, performed by the software GAMESS. The outputs are: an optimized structure of nucleo-

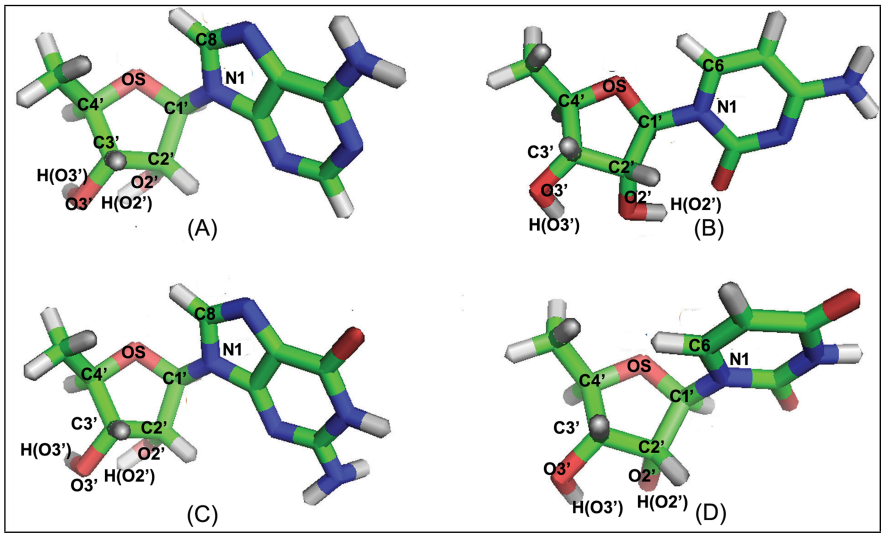

Fig. 2 Nucleosides used as input for the calculations. (A) Adenosine, (B) Cytidine, (C) Guanosine and (D) Uridine. The glycosidic torsion (χ) is defined by the OS-C1'-N1-C8 for adenosine, by the OS-C1'-N1-C6 for cytidine, by the OS-C1'-N1-C8 for guanosine and by the OS-C1'-N1-C6 for uridine.

sides (for the GAN and GAW), a set of charges for the involved atoms (only for GAW) and the quantum energy (for the GAN and GAW). These results are used to refine the parameters of force field (GAN: only dihedral parameters and GAW: electrostatic and dihedral parameters). The next step uses a standard genetic algorithm to optmize the dihedral energy term. The parametrization of this term can be defined according to

$$\min_{w} \parallel E_i^{QM} - E_i^{MM}(w) \parallel^2, \tag{5}$$

where E^{QM} and E^{MM} are quantum and molecular mechanics energies, respectively; and w is defined as

$$\begin{aligned} w_{GAN} &= \{V_n, \gamma\} \\ w_{GAW} &= \{V_n, \gamma; q_i^Q; q_j^Q\}, \end{aligned} \tag{6}$$

where for w_{GAW} the charges q_i^Q and q_j^Q are obtained from QM calculations at the lowest energy optimization for the torsion structure and are used in Eq. 4.

Obtaining the Torsion Profiles

In order to obtain the torsion profiles some constraints are also used similar to reference [10]. The first is to hold the sugar puckering close C3'-endo (C1'-C2'-C3'-C4'), i.e., for rA, rU, rG, and rC at -34.7, -38.9, -39.6, and -39.2 degrees, respectively. Secondly, the C1'-C2'-O2'-H(O2') and C3'-C2'-O2'-H(O2') torsion angles were fixed at -60 and -120 degrees, respectively, to prevent the intramolecular H-bonding between the O2' hydroxyl group and the base and also between the O2' and O3' hydroxyl groups[52–55].

In order to provide preliminary tests for the proposed GA methodology we have considered small number of calculations for the dihedral energy parameters due to a large computing demand. In the present case the dihedral angles of the molecule were rotated in steps of 50 degrees in the glycosidic torsion (χ) for the range of 0 to 360 degrees. If our proposed procedure can improve the results (structures and energies) then one can expect similar behaviour also for other angles. Previous calculations[10] have performed larger number of points using smaller steps (10 degrees). However, the main aim in the present analysis is concentrated with the efficiency of the proposed approach.

The GA applied in GAN and GAW receives as reference data seven rotated optimized structures and one without rotation for each nucleoside. The algorithm considers the same geometry for both, molecular mechanics and QM calculations. This procedure is similar to previous analysis[10] that applied the Monte Carlo method to carried out their refinement. The analysis is performed using RMSE as given by

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{[(E_{i-1}^{QM} - E_i^{QM}) - (E_{i-1}^{MM} - E_i^{MM})]^2}{n}}, \quad (7)$$

where i is the rotation, n is total number of rotations, E^{QM} and E^{MM} as previously defined.

The force field energy (E^{MM}) is calculated using Amber force field (ff99χOL), which uses the parameters returned by the GA. Different rotation energies (0 to 360 degrees with 50 degrees steps) are compared with quantum energies (E^{QM}) for each nucleoside (adenosine, guanosine, cytidine and uridine) calculated from GAMESS.

QM calculations

The relevancy of the comparison of the results obtained by gas-phase quantum-chemical calculations with the experimental data in the condensed phase characteristic for the biomolecular systems can be explained by the insignificant influence of the stacking, sugar-phosphate backbone and surrounding environment on the energetic characteristics of the base pairs that can be neglected in the first approximation[56–63]. This speculation enables us to provide the QM modeling in vacuum approximation[64–67].

The choice of quantum methods is based on works published in the literature. Few parametrizations based on Amber force field (ff94[5], ff98[68] and ff99[7]) have considered the Hatree-Fock method. In the latter case the electrons interact without considering electron correlation. This fact generates great consequences in the interpretation of wave functions[69]. Hence, recent studies focused on more effective methods. For example, the latest improvement over the CHARMM force field has optimized molecular geometries based on MP2/6-31+G* and RI-MP2/cc-pVTZ calculations. In turn, the calculations are used to improve the dihedral potential for the Amber force field[10,70,71].

There are several functionals better than PBE/6-311++G (3df, 3pd) that is used in our work. For example, Walker *et. al.* applied M06-2x functionals and compared the results against several other functionals at the DFT level[72]. However, in order to be able to compare correctly our results with those data obtained by Zgarbová and co-workers[10], we have performed our calculations using the same level of QM theory (PBE).

In the present study the same level of theory (DFT) is used, i.e, the PBE functional and the 6-31++G (3df, 3pd) basis set as recently also considered in recent publications[2,10]. The latter work showed good agreement compared against quantum data.

Dihedral Term - GA Optimization

GAs are methods based on Darwin's principles of evolution and survival of the fittest. They are well-known for performing a global search while analyzing multiple solutions of the search space at each iteration. These algorithms work with individuals, where each individual represents a solution to the problem to be solved. In our case, an individual is a vector of 8 positions, as illustrated in Figure 3, representing the terms V_n and γ, previously defined in Eq. 2.

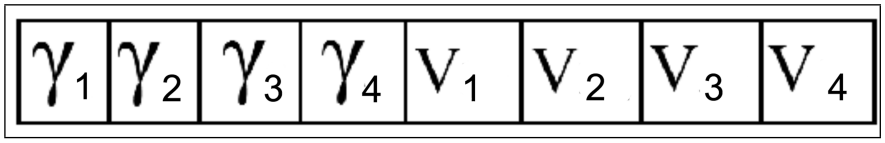

Fig. 3 Representation of individual in the GA.

The standard GA search procedure is shown in Figure 4. First, an initial population of individuals is randomly generated, following a normal distribution with average 0 and standard deviation defined as 1. Next, individuals are evaluated according to how well they solve the problem at hand. As our individuals are dihedral parameters, they are evaluated according to how far the molecular mechanics energy E^{MM} (obtained by Tinker with the GA optimized parameters) approaches to the quantum energy E^{QM} (obtained by GAMESS). The RMSE (as previously defined in Eq. 7) is taken as the fitness function defined by the GA

approach. The lower the RMSE, better the force field approximation is performed. Therefore, the GA optimizies the parameters (one step GAN and the other GAW) in order to minimize the RMSE.

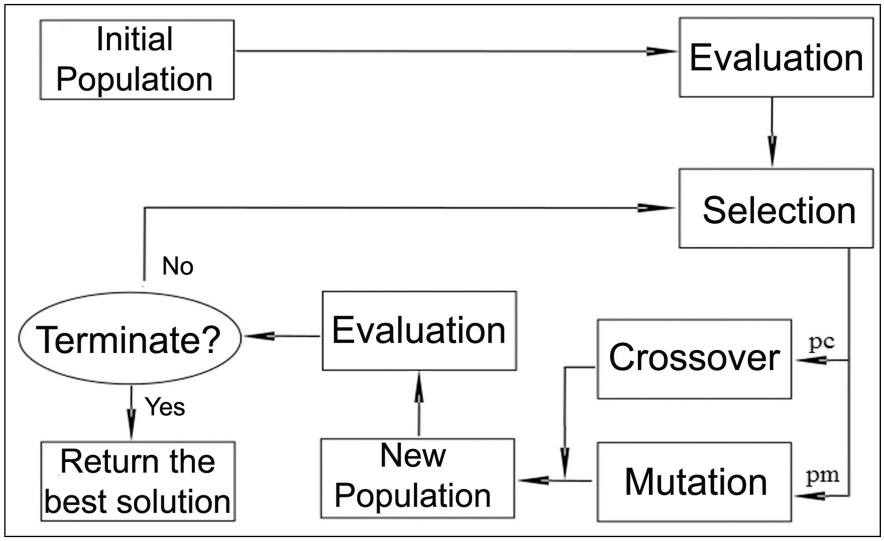

Fig. 4 Flowchart of a standard GA, where p_c and p_m are crossover and mutation probabilities, respectively.

Following the GA structure shown in Figure 4, after evaluation, individuals are selected to undergo crossover and mutation operations according to the probabilities namely, p_c (crossover probability) and p_m (mutation probability). The tournament selection method is used in the present GA[41]. In this procedure, k individuals are randomly selected from the population, and compete with each other for a chance to undergo genetic operations. The individual with the best fitness (in our case, lowest RMSE), is selected. Note that, in this case, individuals with higher fitness have greater probability of surviving, and this probability varies according to the size of tournament (k) selected.

Individuals selected through consecutive tournaments can be modified using crossover or mutation before being inserted into the new population. Crossover takes two individuals, and its main objective is to exchange material between them, creating two children. Here we used uniform crossover, that generates a mask to determine the positions of the two individuals that are exchanged, as shown in Figure 5.

Mutation, in turn, is performed over a single individual, and its objective is to cause a small random modification into the individual, in order to leave a local stagnation in the search region (local optimum). The mutation used here changes the parameter at a randomly selected position by +(-) 0.02.

All individuals generated by the above operators are inserted into a new population, together with the best of all individuals, which is preserved from one generation to the next (procedure known as elitism). This process goes on until a stopping criteria is met, which is usually a minimum error or maximum number of generations.

Note that the algorithm uses a set of parameters, including number of individuals, number of generations, crossover and mutation probabilities and tournament size. An appropriate choice of these parameters is crucial for a good coverage of the search space. Number of individuals and generations are parameters highly dependent on the size of the search space, determined by the problem understudied. Here we tested population sizes of 50, 100, 150, 200 and 250, and generations of 200, 500, 800 and 1200. It was revealed that among all tested by us combinations, the best results, i.e. the minimum value of RMSE, were obtained at 200 individuals and 800 generations with sufficient convergence.

Crossover and mutation probabilities have more standard values, with high crossover and low mutation rates. We tested crossover probabilities of 0.9, 0.95 and 0.99, and mutation rates of 0.01, 0.1, 0.2, 0.25, 0.3 and 0.5. All possible combinations were tested, and the best results obtained are 0.95 and 0.3 for crossover and mutation, respectively. For tournament size, values of 6, 8, 10, 20 and 30 were tested, and 8 produced the best results. All parameter tested were run five times, and the best results selected.

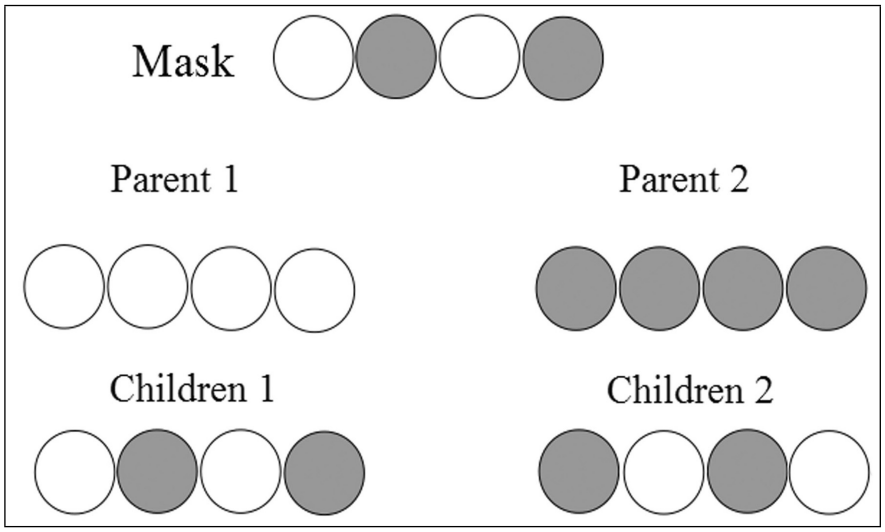

Fig. 5 Representation of uniform crossover operation.

3 Results and Discussion

The Optimization

The dihedral energy parameter optimization is done by a GA, which uses as reference the nucleoside structure and its energy provided by QM calculations. The GA optimizes the dihedral parameters (see Eq. 2) and the refined process is evaluated according to Eq. 7 for each nucleoside (adenosine, guanosine, cytidine and uridine).

The dihedral parameters obtained via GAN and GAW for the glycosidic torsions (defined in Figure 2) are compared in Table 1 with those reported in reference [10].

As GA is a heuristics method, it was executed 5 times in order to demostrate that the algorithm did not find an appropriate solution by chance. As the variance of the results is small, we report the parameters that produced the smallest RMSE (Eq. 7) calculated using different nucleoside rotations. The computational time of each GA generation for both tests is, on average, 60 seconds. Each generation evaluated a set of 200 different parameters, and 800 generations are executed. Hence, only one complete "GA experiment" runs for about 13 hours in a computer with 2GB of RAM memory and one i7 processor. In the actual work a set of 140 "GA experiments" were carried out. This corresponds to 76 days of computing. The QM calculation runs for about 5 days in CENAPAD (Regional Center of the National System for High Computing Performance) using one node with two quadcore processors and 32GB of RAM. For this study a set of 32 calculations were performed, corresponding for about 60 days using 12 nodes.

Figure 6 presents the evolution of the individuals along the generations in terms of RMSE that is calculated by the differences of E^{QM} and E^{MM} for different nucleosides rotations. The RMSE is the measure of evaluation adopted in our GA analysis and it is used to ranking individuals evaluated by the GA. This RMSE is also used to select the individuals for crossover. From now on, we refer to this measure simply as RMSE. These results represent the best individual and the evolution of the RMSE improved along the generations. Note that, for guanosine, the data stabilize faster, while for cytidine the improvements happen until the end of the search process. The average RMSE is an indicative of search convergence, as at this point all individuals (the set of optimized parameters) produce very similar values of fitness. In the present analysis convergence starts around 300 generations.

Comparison with Literature

The previous[10] RMSE reported and the results calculated by the GAN and GAW approaches used in the present work are compared in Figure 7. Our results using the GAN procedure (without the electrostatic term) show that the RMSE values produced a small improvement compared to the results reported in reference [10]. On the other hand, the GAW procedure was able to greatly improve the RMSE and hence, reduced the relative global energy in 37% for adenosine, 67% for guanosine, 21% for cytidine

Table 1 Parameters found by the genetic algorithm for the dihedral term

Nucleoside	Torsion χ	n^a	GAN		*GAW*[d]		Reference [10]	
			$Vn/2^b$	γ^c	$Vn/2^b$	γ^c	$Vn/2^b$	γ^c
Adenosine	OS-C1'-N1-C8*	1	0.9310	88.44	0.9660	69.00	0.6956	68.79
	(OS-CT-N9-C2)**	2	1.0650	6.00	0.9465	15.89	1.0740	15.64
		3	2.9778	301.18	0.0597	24.95	0.4575	171.68
		4	1.7601	147.11	4.0000	308.85	0.3092	19.09
Guanosine	OS-C1'-N1-C8*	1	0.9660	69.00	0.5327	50.60	0.7051	74.76
	(OS-CT-N9-CK)**	2	1.0037	95.35	0.0002	0.80	1.0655	6.23
		3	3.1155	301.80	0.0001	1.96	0.4427	168.65
		4	3.6411	308.70	0.3919	3.91	0.2560	3.97
Cytidine	OS-C1'-N1-C6*	1	1.6639	158.00	0.0092	0.87	1.2251	146.99
	(OS-CT-N9-C1)**	2	3.1230	301.24	2.9787	277.70	1.6346	16.48
		3	3.5517	308.40	3.5213	278.05	0.9375	185.88
		4	0.1332	12.90	0.1253	10.03	0.3103	32.16
Uridine	OS-C1'-N1-C6*	1	1.3911	132.10	1.9634	186.52	1.0251	149.88
	(OS-CT-N9-CM)**	2	1.5847	150.50	1.4182	137.21	1.7488	16.76
		3	1.9126	175.30	1.0670	69.31	0.5815	179.35
		4	1.1374	85.60	1.4675	96.34	0.3515	16.00

[a] The periodicity of the torsion χ.
[b] Magnitude of the rotational barrier in kcal/mol.
[c] Phase off set in deg.
[d] New electrostatic charge parameters are in the supporting information.
* See Figure 2 for details.
** atoms type in ff99χOL.

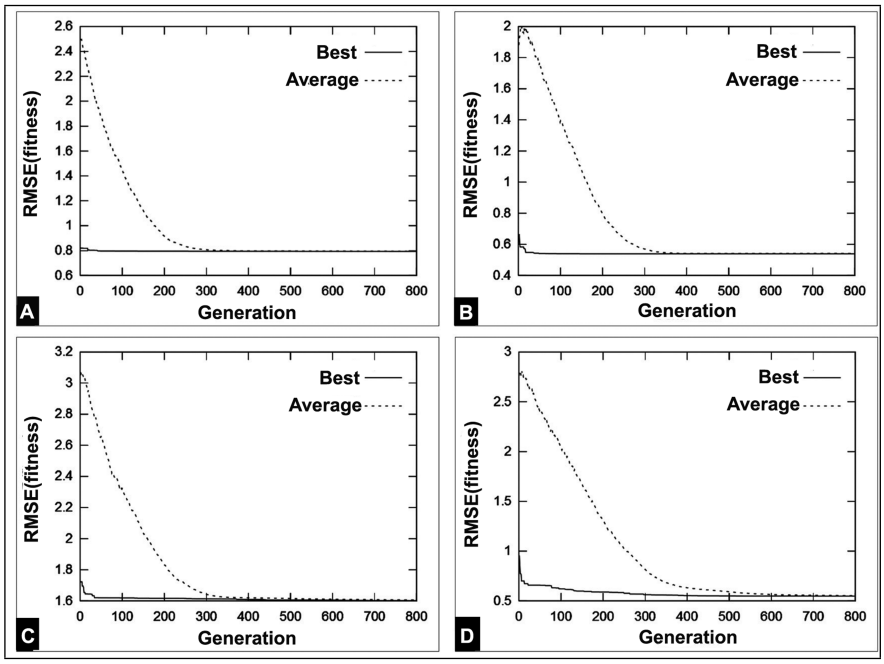

Fig. 6 RMSE (Eq. 7) for different nucleosides rotations of the individuals evolved by the GA, by comparing the best individual and the average population for: (A) Adenosine; (B) Guanosine; (C) Cytidine and (D) Uridine.

and 72% for uridine when compared against data from reference [10]. These results show that the method of parametrization performed by GAW may be more effective, especially for uridine.

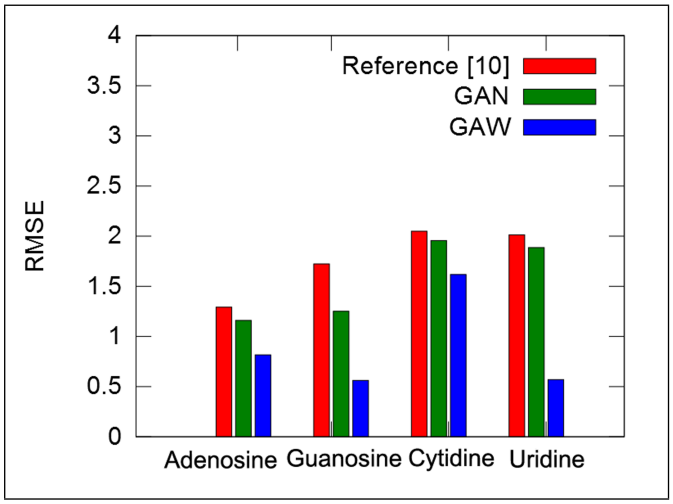

Fig. 7 RMSE analyzes of nucliosides of RNA between the reference [10] and methodologies proposed (GAN and GAW).

Figure 8 shows that the total potential energy that represents the classical energy for all nucleosides using GAW is closer to the quantum results if compared to the reference [10] or the GAN method. The actual study shows an improvement in the energy calculation based on GAW optimization. A re-scale procedure previous applied[73] for this type of analysis was also adopted in order to compare published results and our calculations against quantum results taken as reference data.

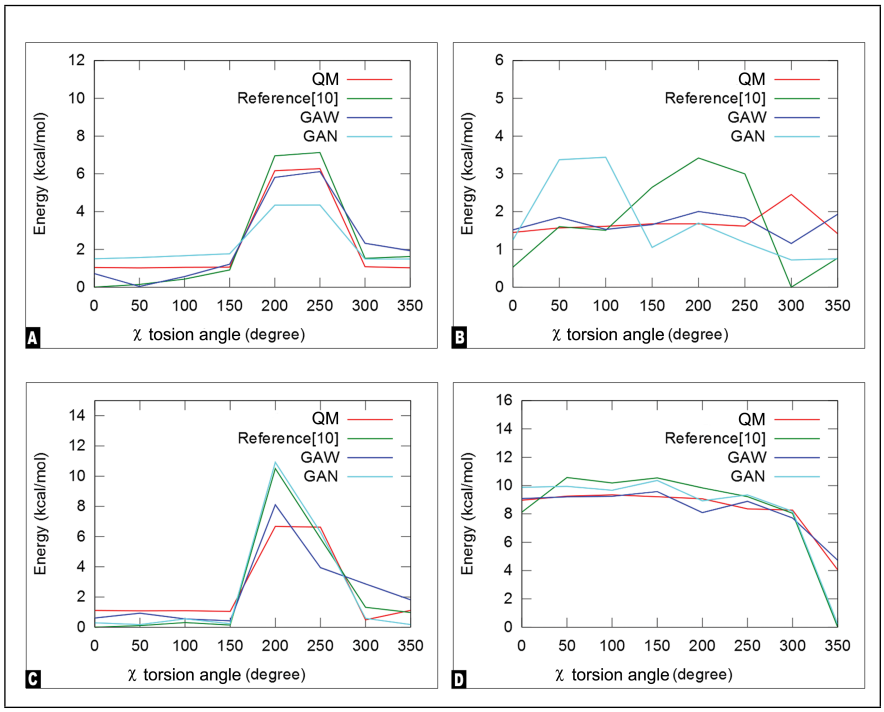

Fig. 8 Comparison of E^{QM} with the E^{MM} previously reported in reference [10] and the E^{MM} obtained in the present study (GAN and GAW) for (a) Adenosine, (b) Guanosine, (c) Cytidine and (d) Uridine.

As already pointed out, the best results were obtained for uridine. Observing the curves in Figure 8 (D), the shape of the E^{MM} data using GAW (Eq. 1) is closer to the E^{QM} result. The parametrization shown in reference [10] has a high level of precision. However, our new approach shows an improvement of such parametrization. The consequence of small deviations usually observed in force field is pointed out by Banás *et. al.*[15] as a justification for the underestimation of base pairing energies. As shown in this figure, the proposed methodology reduced this problem, providing the RNA structure closer to the data assumed as reference (QM calculations).

RNA Crystal Simulation using New Parameters

In order to better validate the proposed procedure, we selected the RNA molecule from PDB to demonstrate that the refined parameters can be applied to optimize the representation of RNA structure. Spatial structure of this molecule was obtained using NMR spectroscopy and it contains ten nucleotides. It was used in the structural analysis of the IIIc domain of GB human virus[74]. Due to its relatively small size, this structure provides a reasonable computer simulation with no difficulties for testing analyzes. In addition, if our proposed method of using GA to refine the dihedral and electrostatic parameters is effective (for instance, a different molecule that was not considered in the refine process), then we believe that the application of this method to other terms (Eq. 1) can even better improve the force field, at least for nucleic acids of RNA.

We compare the structure obtained using the parameters reported in reference [10] and the results calculated here against the experimental data (PDB structure). Tinker[48] was used to perform the molecular mechanics calculations for both sets of parameters. In order to simulate similar conditions to those applied in the experiment 1r4h, a water box was used to solvate the molecule (dimensions 41 x 41 x 41 Å with 333 molecules of water for one 1r4h). The simulation was performed using the software DICE[75], and the total charges was neutralized with Na^{2+} ions[76].

A fundamental point to perform the analysis above described is to select and calculate an appropriate measure of similarity between different structures[77]. Here we choose Root Mean Square Deviation (RMSD), a measure widely used in the literature to represent the root mean square deviation of the atoms[77] and it is calculated as

$$RMSD = \sqrt{\sum_{n}^{i=1} \frac{(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2}{n}}, \quad (8)$$

where x, y and z are the experimental data coordinates, x', y' and z' are the coordinates found by Tinker and n defines the number of atoms in the molecule. As pointed out in reference [78], RSMD is a good indicator of the precision of the atomic coordinates. The lower its value, the better is the precision.

The results of RMSD considering the structure generated with GAW and GAN parameters and reference [10] are compared in Table 2. One observes that the RMSD (for GAW) is reduced in one order of magnitude with respect to previous publication [10]. Although, Zgarbová *et. al.* [10] have a good value of RMSD, we were able to improve this value in 13% with GAN and 78% with GAW.

Table 2 Comparison of RMSD data for the RNA molecule (PDB code 1r4h)

	RMSD[a]
Reference [10]	1.50
GAN	1.30
GAW	0.33

[a] Values set in Å.

As observed in Figure 9, when the electrostatic term is taken into account (GAW) in the parametrization process one obtains a better theoretical structure for RNA molecule (PDB code 1r4h) that is much closer to the experimental PDB data. This is mainly due to the hydrogen bonds and hydrophobic effects that are fundamental for the stability of the helix [15]. As expected, the structures are greatly affected by base-pairing interaction.

The structures were aligned using the software Pymol [79]. In order to provide easier visualization, the water molecules making solvation are removed. The results show an improvement of the structure using our new parametrization when compared against experimental data.

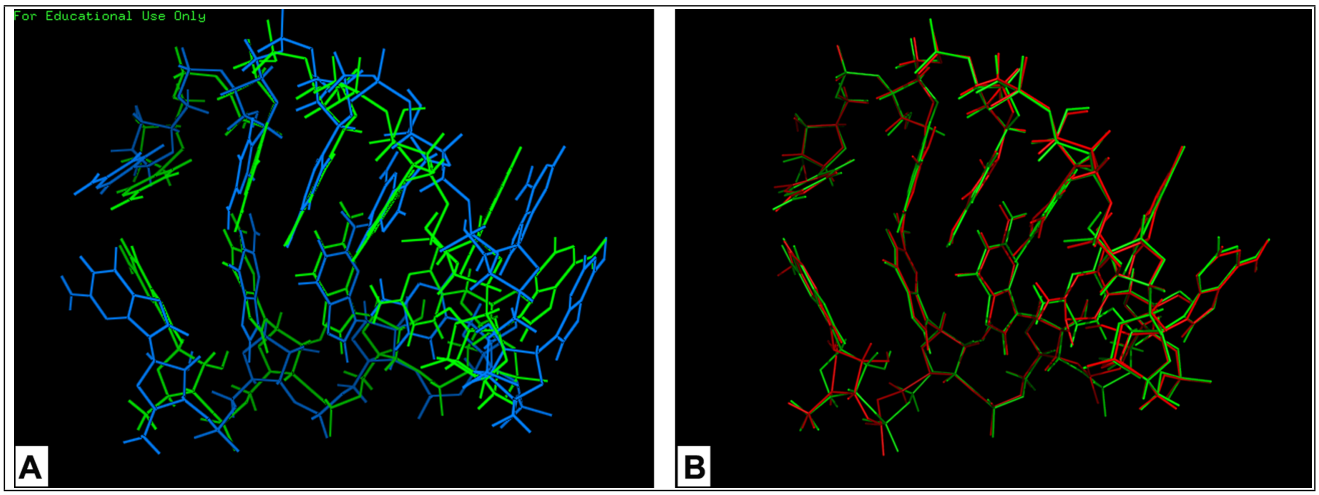

Fig. 9 Comparison of the structures obtained by the two methodologies proposed in this work (GAN and GAW) against the experimental data. (A) Optimization generated with parameters found in GAN (blue) and experimental data (green). (B) Optimization generated with parameters found by the GAW (red) and experimental data (green).

A Final Refinement in GAW

As can be observed in Figure 10, almost the whole structure is well predicted when compared against RNA structure (PDB code 1r4h). However, for the sugar pucker of the guanine (marked with a square box) one observes deviations. Particularly, the greatest errors (RMSD) compared against the PDB data arise due to the oxygen atom of this sugar pucker. The bond length of carbon (C1') and oxygen of this part of the structure was not well optimized. In the actual case it is 0.33 Å. One can therefore address the following question: why only in this sugar pucker the superposition was not good enough?

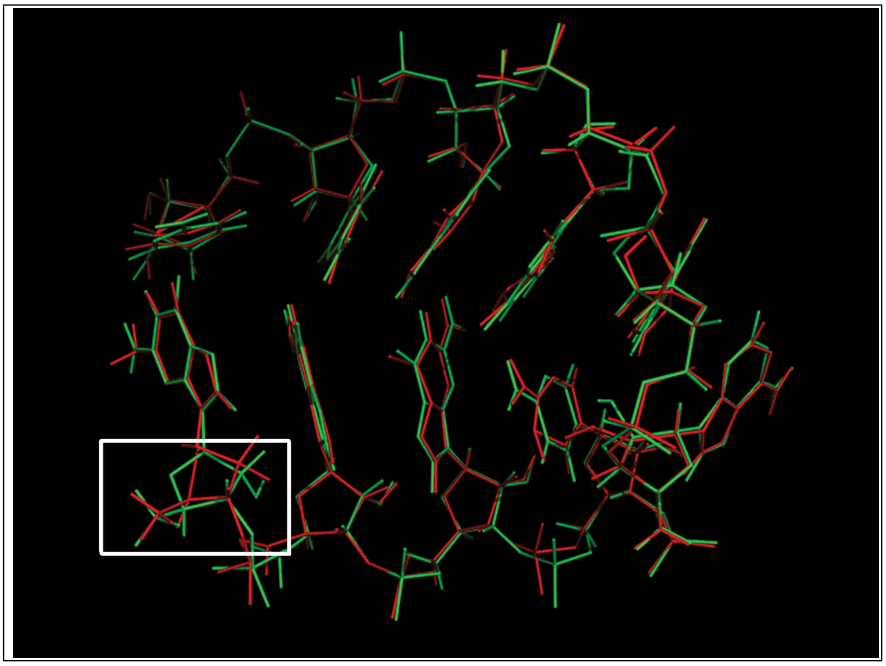

Fig. 10 Comparison of the RNA structure obtained by the proposed methodology using GAW (red) and the experimental data (PDB code 1r4h) (green).

In order to answer the above question we have isolated this part of the structure (see Figure 11A). As can be seen the oxygen bond is not well predicted and this is part of the reason of a higher RMSD error for the whole structure. When one verifies the hydrogen bonds (Figure 11B) in the whole structure, one realizes that the isolated sugar (square box) does not have this type of interaction. Therefore, by analyzing the same type of interaction for cytosine (both are at the extremity of RNA and does not have the hydrogen bond) this error is not observed and its average dihedral energy is higher to that from reference [10]. Our average value of V_n found using GA methodology is 1.6586 kcal/mol against 1.0269 kcal/mol obtained in reference [10]. Accordingly, one needs to analyze the balance between RMSE and the torsion barrier (V_n). Actually, one can verify in Table 1 that for the GAW procedure the average optimization dihedral parameter is 0.2312 kcal/mol while from reference [10] it is 0.6173 kcal/mol. The greatest discrepancy arises from the two small optimized results in our calculations (V_2 and V_3). This may be the reason of our worse results for the guanine shown in Figure 10. In order to improve the guanine structure in the methodology GAW, we carried out the analysis based on the balance between RMSE and the average values of V_n. The advantage of our multiple GA solutions was used in order to rank all possible solutions. Based on the latter quality of our GA results we used the third optimum optimized data (see Table 3) from our ranking GAW output. As can be seen in Table 3, with this new data (all V_n were modified) the new average dihedral parameter (0.5042 kcal/mol) is now closer to that obtained in reference [10] (0.6173 kcal/mol).

The new higher precise alignment is reflected in the RMSD values shown in Table 4. Note that the values are further improved (0.27 Å against 0.33 Å in the first GAW parametrization). This final result shows a better agreement of about 82% (over 1Å) using GAW-New in comparison with the results obtained in reference [10].

The new plot of the RNA molecule compared against the experimental data (green) is shown in Figure 12. It is compared to the structure from reference [10] (Figure 12A) against the PDB structure and our improved structure against the PDB experimental

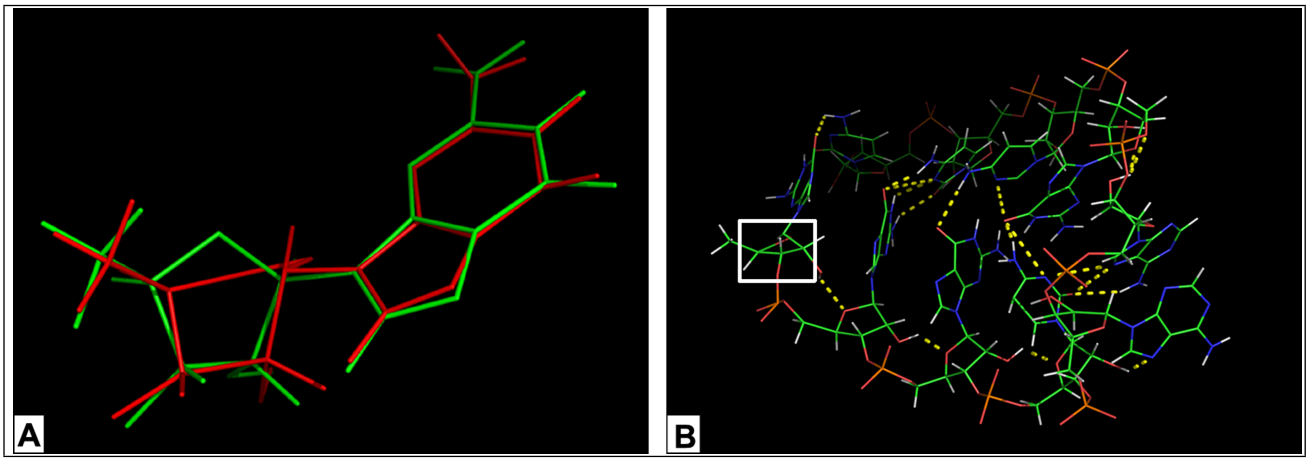

Fig. 11 The problem with Guanine. (A) Alignment of the guanine marked with a square in RNA 1r4h (Figure 10) against the experimental data and (B) Hydrogen bonds in 1r4h shown in dotted lines.

Table 3 New parameters selected from the GAW for guanine

Nucleoside	Torsion χ	n^a	GAW-New		GAW	
			$Vn/2^b$	γ^c	$Vn/2^b$	γ^c
Guanosine	OS-C1'-N1-C8*	1	1.1954	113.57	0.5327	50.60
	(OS-CT-N9-CK)**	2	0.0996	5.76	0.0002	0.80
		3	0.0004	13.11	0.0001	1.96
		4	0.7215	4.27	0.3919	3.91

[a] The periodicity of the torsion χ.
[b] Magnitude of the rotational barrier in kcal/mol.
[c] Phase off set in deg.
* See Figure 2 for details.
** atoms type in ff99χOL.

Table 4 Root Mean Square Deviation analyzes for the RNA molecule (PDB code 1r4h) with the new parameters of guanosine.

	$RMSD^a$
Reference [10]	1.50
GAW	0.33
GAW - New	0.27

[a] Values set in Å.

data (Figure 12B). As observed, the GAW-New alignment is closer to the experimental data than the results obtained in reference [10]. This new result is based on a much better optimization of the force field based on GA.

In addition, one observes that the problem with the guanine (shown in Figure 12) was solved. One can also observe that the rings of the nucleotides bases are better represented than the structure skeleton considering the final refinement. This happens because the algorithm focused its optimization in the glycosidic torsions (χ), improving the RMSD value with respect to the experimental data with computer-generated structures and reducing the parameter value from 1.5 to 0.27 Å.

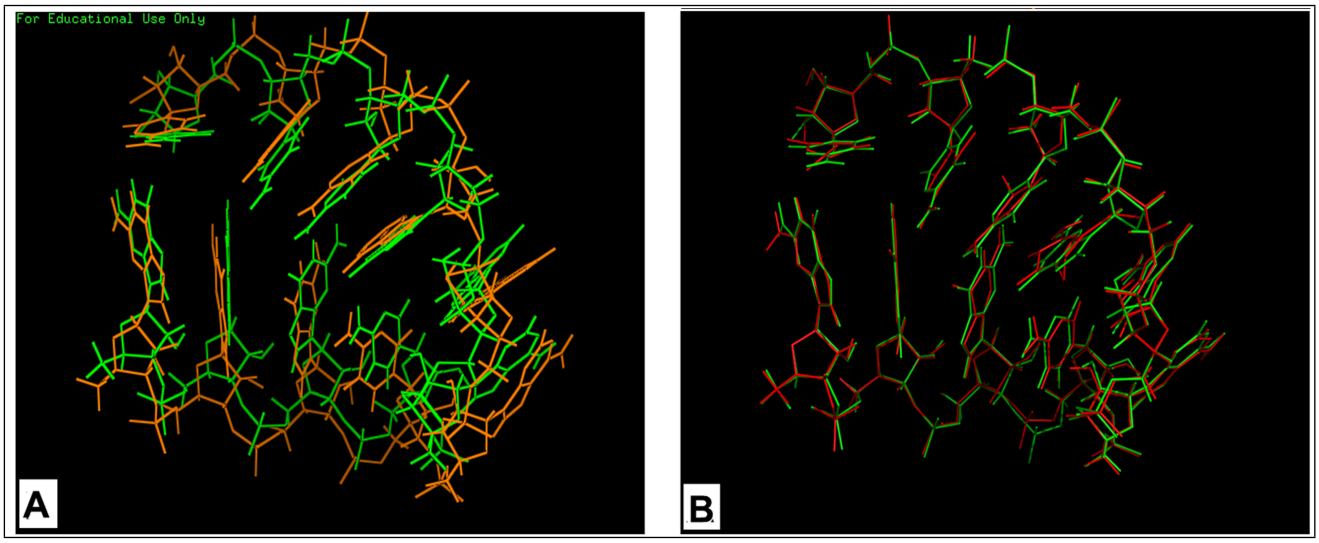

Fig. 12 Comparison of the structures obtained by the two optimization and experimental data. (A) Optimization generated with parameters found in reference [10] (orange) and experimental data (green). (B) Optimization generated with parameters found by the GAW-New (red) and experimental data (green).

4 Conclusions

In summary, the present study has proposed the use QM calculations as references for a genetic algorithm to optimize the dihedral term parameters of the glycosidic torsion (χ) for better accuracy in representation of RNA spatial structure. As previously stated in reference [45] this is the most relevant conformational parameter in nucleic acids studies. In the present study the parameters of the Amber force field (ff99) were better determined when compared against those published by Zgarbová *et. al.* [10].

The optimization using GAW method (refinement of electrostatic and dihedral parameters) produced better results demonstrating the importance of base pairing interaction. This effect is fundamental for the stability of the RNA double helix [15]. Accordingly, a good parametrization of the electrostatic and dihedral term is important to get quantitative representation of the molecules (closer to experimental data).

The objective of the methodology was to further reduce the error of the computational calculations involved in the parametrization of the force field with particular attention to nucleosides dihedral energies. The new parametrization reduced, on average, the RMSE calculated with the differences of QM and MM energies in 50%. As demonstrated in the present work, the parameters found led to a better structural representation of the RNA molecule (PDB code 1r4h) when compared to parametrizations previously reported [10]. It is important to emphasize that the present approach was able to go further and predict with a reasonable level of accuracy an structure observed experimentally by NMR spectroscopy that was not taken into account in the refinement process.

Finally, the proposed approach of using GA coupled with QM calculations demonstrated an improvement in the refinement process to the electrostatic and dihedral parameters. The method can be also applied to refine parameters of other terms of the force field. In addition, it is demonstrated that the actual methodology can provide an efficient manner to represent via computer simulation a molecule structure closer to the experimental data. The representation can provide many advantages for others studies, for example, we can predict several functions of RNA that are not well understood. For example, by using homology (the same structure has the same functions) approaches.

Acknowledgement

We thank CENAPADs (UFRS and UFMG) for help us in the QM calculations and Dr. Lucas Bleicher for discussions in the final refinement section. We also acknowledge the financial support the Brazilian National Council of Research (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do estado de Minas Gerais

(FAPEMIG).

References

1 M. A. Ditzler, M. Otyepka, J. Sponer and N. G. Walter, *Accounts of Chemical Research*, 2010, **43**, 40–47.
2 J. Sponer, A. Mldek, J. E. Sponer, D. Svozil, M. Zgarbová, P. Banás, P. Jurecka and M. Otyepka, *Physical Chemistry Chemical Physics*, 2012, **14**, 15257–15277.
3 A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton and M. Orozco, *Biophysical Journal*, 2007, **92**, 3817–3829.
4 P. Auffinger and E. Westhof, *Current Opinion in Structural Biology*, 1998, **8**, 227–236.
5 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *Journal of the American Chemical Society*, 1995, **117**, 5179–5197.
6 N. Foloppe and A. D. MacKerell, *Journal of Computational Chemistry*, 2000, **21**, 86–104.
7 J. M. Wang, P. Cieplak and P. A. Kollman, *Journal of Computational Chemistry*, 2000, **21**, 1049–1074.
8 D. Bosch, N. Foloppe, N. Pastor, L. Pardo and M. Campillo, *Biophysical Journal*, 2007, **92**, 3817–3829.
9 N. Foloppe and A. D. MacKerell, *Journal Physical Chemistry*, 1998, **34**, 6669–6678.
10 M. Zgarbová, M. Otyepka, J. Sponer, A. Mládek, P. Banás, T. E. Cheatham, III and P. Jurecka, *Journal Chememical Theory and Computatonal*, 2011, **7**, 2886–2902.
11 S. Y. Reddy, F. Leclerc and M. Karplus, *Biophysical Journal*, 2003, **84**, 1421–1449.
12 I. Besseova, M. Otyepka, K. Reblova and J. Sponer, *Physical Chemistry Chemical Physics*, 2009, **11**, 10701–10711.
13 N. J. Deng and P. Cieplak, *Biophysical Journal*, 2010, **98**, 627–636.
14 C. G. Ricci, A. S. C. de Andrade, M. Mottin and P. A. Netz, *Biophysical Journal*, 2010, **98**, 627–636.
15 P. Banás, A. Mladék, M. Otyepka, M. Zgarbová, P. Jurecka, D. Svozil, F. Lankas and J. Sponer, *Journal Chememical Theory and Computatonal*, 2012, **8**, 2448–2460.
16 Y. B. Cheng and J. K. Sykulski, *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 1996, **9**, 59–69.
17 M. Buchvarova and P. I. Y. Velinov, *Advances in Space Research*, 2010, **45**, 1026–1034.
18 I. Yildirim, H. A. Stern, S. D. Kennedy, J. D. Tubbs and D. H. Turner, *Journal of Chemical Theory and Computation*, 2010, **6**, 1520–1531.
19 Y. Sakae and Y. Okamoto, *Journal of Theoretical and Computational Chemistry*, 2004, **3**, 339358.
20 J. Wang and P. A. Kollman, *Journal of Computational Chemistry*, 2001, **22**, 1219–1228.
21 J. Solomon, P. Chung, D. Srivastava and E. Darve, *Computational Materials Science*, 2014, **81**, 453–465.
22 J. Kästner, J. M. Carr, T. W. Keal, W. Thiel, A. Wander and P. Sherwood, *Journal Physical Chemistry*, 2009, **113**, 11856–11865.
23 M. X. Silva, B. R. L. Galvão and J. C. Belchior, *Physical Chemistry Chemical Physics*, 2014, **16**, 8895–8904.
24 D. D. C. Rodrigues, A. M. Nascimento, H. A. Duarte and J. C. Belchior, *Journal Chemical Physics*, 2008, **349**, 91–97.
25 F. F. Guimarães, J. C. Belchior, R. L. Johnston and C. Roberts, *Physical Chemistry Chemical Physics*, 2002, **19**, 8327–8333.
26 M. Böyükata, E. Borges, J. C. Belchior and J. P. Braga, *Physical Chemistry Chemical Physics*, 2002, **19**, 8327–8333.
27 Y. P. Yurenko, R. O. Zhurakivsky, M. Ghomi, S. P. Samijlenko and D. M. Hovorun, *Journal Physical Chemistry B*, 2007, **111**, 6263–6271.
28 Y. P. Yurenko, R. O. Zhurakivsky, M. Ghomi, S. P. Samijlenko and D. M. Hovorun, *Journal Physical Chemistry B*, 2007, **111**, 9655–9663.
29 Y. P. Yurenko, R. O. Zhurakivsky, M. Ghomi, S. P. Samijlenko and D. M. Hovorun, *Journal Physical Chemistry B*, 2008, **112**, 1240–1250.
30 A. G. Ponomareva, Y. P. Yurenko, R. O. Zhurakivsky, T. Mourik and D. M. Hovorun, *Journal Physical Chemistry B*, 2012, **14**, 6787–6795.
31 T. Y. Nikolaienko, L. A. Bulavin and D. M. Hovorun, *Physical Chemistry Chemical Physics*, 2012, **14**, 7441–7447.
32 A. G. Ponomareva, Y. P. Yurenko, R. O. Zhurakivsky, T. Mourik and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2014, **32**, 730–740.
33 O. O. Brovarets, R. O. Zhurakivsky and H. D. M, *Biopolymers and Cell*, 2010, **26**, 398–405.
34 O. O. Brovarets, Y. P. Yurenko and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2014, **32**, 993–1022.
35 O. O. Brovarets, *Journal of Molecular Modeling*, 2013, **19**, 4223–4237.
36 O. O. Brovarets', Y. P. Yurenko and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamics*, 2014, 1–29.
37 O. O. Brovarets and D. M. Hovorun, *Physical Chemistry Chemical Physics*, 2013, **15**, 20091–20104.
38 P. J. P, K. Burke and M. Ernzerhof, *Physical Review Letters*, 1996, **77**, 3865–3868.
39 P. J. P, K. Burke and M. Ernzerhof, *Physical Review Letters*, 1997, **78**, 1396.
40 A. H. Wright, Foundations of Genetic Algorithms, 1991, pp. 205–218.
41 T. Back, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*, Oxford University Press, Oxford, UK, 1996.
42 A. A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*, Springer, Gernany, 2002.
43 *An Information Portal to Biological Macromolecular Structures*, 2014.
44 X. X. Yao, C. G. Ji, D. Q. Xie and J. Z. H. Zhang, *Journal of Computational Chemistry*, 2013, **34**, 1136–1142.
45 F. Lankas, N. Spackova, M. Moakher, P. Enkhbayar and J. Sponer, *Nucleic Acids Research*, 2010, **38**, 3414–3422.
46 H. Margenau and N. R. Kestner, *Int. Series of Mono. In Natural Phy.: Theory of Intermolecular Forces*, Pergamon Press, New York, 1971.
47 M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, *Journal of Computational Chemistry*, 1993, **14**, 1347–1363.
48 S. Rubenstein, C. Kundrot, S. Huston, M. Dudek, Y. Kong, R. Hart, M. Hodsdon, R. Pappu, W. Mooij, G. Loeffler, M. Vorobieva, N. Sokolova, P. Bagossi, P. Ren, A. Carlsson, A. Kutepov, A. Grossfield, M. Schnieders, D. Gohara and T. Darden.
49 P. H. Shah and R. C. Batra, *Computational Materials Science*, 2014, **83**, 349–361.

50 A. Zhong, X. Jiang, Y. Hu and C. Du, *Journal of the Society of Leather Technologists and Chemists*, 2013, **97**, 121–124.
51 S. Saito, K. Ohno, T. Suzuki and H. Sakuraba, *Molecular Genetics and Metabolism*, 2012, **105**, 244–248.
52 R. O. Zhurakivsky and D. M. Hovorun, *Physics of Alive*, 2007, **15**, 91–106.
53 R. O. Zhurakivsky and D. M. Hovorun, *Physics of Alive*, 2007, **15**, 24–34.
54 R. O. Zhurakivsky and D. M. Hovorun, *Ukrainica Bioorganica Acta*, 2007, **5**, 41–51.
55 R. O. Zhurakivsky and D. M. Hovorun, *Bioppolymers and Cell*, 2008, **24**, 142–157.
56 V. Zoete and M. Meuwly, *Journal Chemical Physics*, 2004, **121**, 4377–4388.
57 O. O. Brovarets, Y. P. Yurenko, I. Y. Dubey and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2012, **29**, 1101–1109.
58 O. O. Brovarets and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2015, **33**, 28–55.
59 O. O. Brovarets' and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamics*, 2014, 1–21.
60 O. O. Brovarets and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2014, **32**, 127–154.
61 O. O. Brovarets and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2014, **32**, 1474–1499.
62 O. O. Brovarets and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamic*, 2013, **31**, 913–936.
63 O. O. Brovarets' and D. M. Hovorun, *Molecular Physics*, 2014, 1–14.
64 O. O. Brovarets and D. M. Hovorun, *Physical Chemistry Chemical Physics*, 2014, **16**, 15886–15899.
65 O. O. Brovarets and D. M. Hovorun, *Physical Chemistry Chemical Physics*, 2014, **16**, 9074–9085.
66 O. O. Brovarets, R. O. Zhurakivsky and D. M. Hovorun, *Physical Chemistry Chemical Physics*, 2014, **16**, 3715–3725.
67 O. O. Brovarets', R. O. Zhurakivsky and D. M. Hovorun, *Journal of Biomolecular Structure and Dynamics*, 2014, 1–16.
68 T. E. Cheatham, P. Cieplak and P. A. Kollman, *Journal of Biomolecular Structure and Dynamic*, 1999, **16**, 845–862.
69 C. Froese−Fischer, *The Hartree−Fock Method for Atoms*, Wiley, New York, 1997.
70 B. S. M and H. Zhang, *IEEE Transactions on Evolutionary Computation*, 1999, **3**, 1–21.
71 P.-P. Zhou and W.-Y. Qiu, *Journal Physical Chemistry*, 2009, **113**, 10306–10320.
72 M. Walker, H. A. J, A. Sen and C. E. Dessent, *Journal of Physical Chemistry A*, 2013, **17**, 12590–12600.
73 O. Guvench and A. D. MacKerell, *Journal Molecular Modeling*, 2008, **14**, 667–679.
74 R. Rijnbrand, V. Thiviyanathanb, K. Kaluarachchib, S. M. Lemona and D. G. Gorensteinb, *Journal of Molecular Biology*, 2004, **343**, 805–817.
75 K. Coutinho, Symposium in Memory of Michael C. Zerner, 2000.
76 J. Aqvist, *Journal Physical Chemistry*, 1990, **94**, 8021–8024.
77 A. M. Lesk, *Introdução à Bioinformática*, Artmed, São Paulo, 2005.
78 P. E. Bourne and H. Weissig, *Structural bioinformatics*, John Wiley & Sons, New Jersey, 2003.
79 L. Schrödinger, PyMOL The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.