ORIGINAL PAPER

A molecular dynamics study of a miRNA:mRNA interaction

**Giulia Paciello · Andrea Acquaviva · Elisa Ficarra ·
Marco Agostino Deriu · Enrico Macii**

Abstract In this paper we present a methodology to evaluate the binding free energy of a miRNA:mRNA complex through molecular dynamics (MD)–thermodynamic integration (TI) simulations. We applied our method to the *Caenorhabditis elegans* let-7 miRNA:lin-41 mRNA complex—a validated miRNA:mRNA interaction—in order to estimate the energetic stability of the structure. To make the miRNA:mRNA simulation possible and realistic, the methodology introduces specific solutions to overcome some of the general challenges of nucleic acid simulations and binding free energy computations that have been discussed widely in many previous research reports. The main features of the proposed methodology are: (1) positioning of the restraints imposed on the simulations in order to guarantee complex stability; (2) optimal sampling of the phase space to achieve satisfactory accuracy in the binding energy value; (3) determination of a suitable trade-off between computational costs and accuracy of binding free energy computation by the assessment of the scalability characteristics of the parallel simulations required for the TI. The experiments carried out demonstrate that MD simulations are a viable strategy for the study of miRNA binding characteristics, opening the way to the development of new computational target prediction methods based on three-dimensional structure information.

G. Paciello (✉) · A. Acquaviva · E. Ficarra · E. Macii
Department of Control and Computer Engineering,
Politecnico di Torino,
C.so Duca degli Abruzzi 24,
10129 Turin, Italy
e-mail: giulia.paciello@polito.it

M. A. Deriu
Department of Mechanics, Politecnico di Torino,
C.so Duca degli Abruzzi 24,
10129 Turin, Italy

Introduction

MicroRNAs (miRNAs) are small (approximately 22 nucleotides) noncoding RNAs in plants and animals [1] that are remarkably similar in size to small interfering RNAs (siRNA) [14]. MiRNAs are produced by the enzyme Dicer from endogenous stem-loop RNA molecules, and are implicated in the control of several biological processes such as differentiation, cell proliferation and developmental timing [2, 13, 15]. Over the past decade, miRNAs have emerged as fundamental molecules involved in the regulation of eukaryotic gene expression because of their function in controlling mRNA translation and degradation. After base-pairing with their target mRNA strands, miRNAs are able to silence cytoplasmic mRNA by triggering endonuclease cleavage, promoting translation repression or accelerating mRNA decapping [4]. These small RNA sequences carry out their activities within specific ribonucleoprotein (RNP) complexes, ensuring that each nucleic acid strand is located in the correct cellular compartment together with the appropriate complement of proteins it needs to perform its biochemical function [14].

The strategic role played by miRNAs in numerous biological processes has made their target prediction an interesting field of research that has been investigated widely in recent years. Almost all the existing miRNAs target prediction algorithms consider similar general principles based on secondary structure analyses such as the base pairing pattern, comparative sequence analysis to check conservation, the examination of multiple target sites and thermodynamic analysis of the miRNA:mRNA hybrid [3], and share the

common problem of predicting a far from negligible number of false positives.

To date, secondary structure has been used because of the lack of availability of three-dimensional (3D) configurations of miRNA:mRNA complexes, the large computational effort required, and the challenges typical of molecular simulations of nucleic acids in general because of their instability. Nowadays, 3D structure information of these complexes is starting to become available [13], which, coupled with increasing computational power, makes it possible to perform molecular dynamics (MD) simulations of miRNA: mRNA interactions to explore their binding characteristics.

In this work, we present a study of the energetic stability of a specific, experimentally validated interaction [7]: the interaction between the *Caenorhabditis elegans* let-7 miRNA fragment and its complementary site *LCS*2 in the 3′-untranslated region (UTR) of the lin-41 mRNA [13], by calculating its binding free energy via MD–thermodynamic integration (TI) simulations. The contribution of this work is two-fold. First, we devise a methodology for the computation of free energy of this interesting class of RNA strands, thus reducing the well known challenges of nucleic acids molecule simulations [5, 8, 12]. The methodology includes: (1) positioning of the restraints imposed on the simulations in order to guarantee complex stability; (2) optimal sampling of the phase space to achieve a satisfactory accuracy in the binding energy value; (3) determination of a suitable trade-off between computational costs and accuracy of binding free energy computation by the assessment of the scalability characteristics of the parallel simulations required for the TI. Second, thanks to the MD simulations, we highlight the relevance of 3D structure information on free energy computation. In particular, we observed a stronger pairing between the RNA strands compared to that estimated by means of tools based on secondary structure. We believe that this work represents a step forward in the refinement of miRNA target stability evaluation, and thus in the deeper understanding of their regulatory effects, opening the field to strategies for the integration of 3D structure information into miRNA target prediction tools.

The rest of the paper is organized as follows. The Materials and methods section describes the structure under investigation, explaining the modifications brought to the initial molecular configuration, and presents the simulation conditions and the restraints adopted to obtain complex stability, and then illustrates the technique used to perform the binding free energy calculation for the specific miRNA: mRNA interaction. Results and discussion presents the results achieved by adopting the proposed method, discussing the choices made to perform the simulations and evaluating the actual computational costs of the technique. The Conclusion summarizes the findings and considers potential further contributions of the research.

Materials and methods

System set-up

In this work we performed three types of simulations on three different molecular complexes. The main simulation was performed on the let-7 miRNA:lin-41 mRNA [13] to compute the binding free energy. Two companion simulations were carried out on the Aa- Ago bound to 22-mer complex [6], opportunely modified, to determine a realistic configuration of the restraints to be imposed during the main simulation, i.e., with and without Ago protein residues, respectively. The NMR atomic structure of the miRNA:mRNA complex under investigation was taken from the Protein Data Bank (PDB; accession code 2JXV). It is a 33 nt RNA stem-loop model with an internal asymmetric loop, designed to represent the *C. elegans* let-7 miRNA fragment bound to its complementary site *LCS*2 in the 3′-UTR of the lin-41 mRNA. The original structure present in the PDB contained two strands linked with a GAAA tetraloop as well as other base pair differences [13] compared with the real RNA sequences interacting in nature [7].

The secondary structure of the model is reported in Fig. 1.

The original atomic structure (2JXV, [13]) contained a GAAA tetraloop connecting the two strands, which we removed in order to better model the interaction between miRNA and mRNA likely to occur under physiological conditions [7], in which the two RNA fragments are different entities not linked by other sequences of nucleotides like the external GAAA tetraloop, but interact only through hydrogen bonds. For this reason, the four nucleotides of the GAAA tetraloop were excluded in the system investigated. Therefore, the atomic structure of the RNA complex was composed by two chains, made up of 14 and 15 nucleotides, modeling miRNA and mRNA, respectively. The new secondary structure is shown in Fig. 2.

The modified complex was placed in a periodic cubic box that was subsequently filled with about 12,000 TIP4P water molecules to avoid distortions of the structure typical of in vacuo nucleic acid simulations [5]. The negative net charge of the system was neutralized with 27 sodium ions placed randomly in the simulation box. The coordinates were
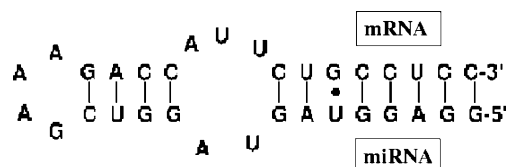

Fig. 1 Let-7 miRNA:lin-41 mRNA initial model. The *dot* represents a wobble base pair
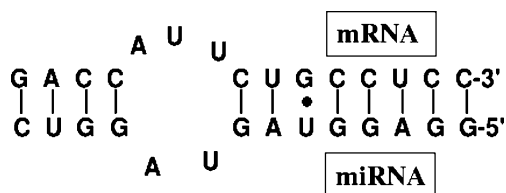
Fig. 2 let-7 miRNA:lin-41 mRNA modified model. The *dot* represents a wobble base pair

optimized via 3 ps steepest descent energy minimization, after which an all atom position restrained simulation for 5 ps under a constant force of 1,000 kJ mol^{-1}nm^{-2} was carried out.

The first type of companion simulations were run on the Aa-Ago bound to the 22-mer self-complementary structure downloaded from the PDB (accession code 2F8S) [6]. Also in this case the initial atomic structure, composed of 704 Ago protein residues and 44 RNA residues, was adjusted in order to make the length of the two chains of this complex equal to those of the miRNA:mRNA structure of interest. The first eight nucleotides from the 5′-end of the strand bound to Ago, considered the 'guide strand' since it presents a specific sequence and structure capable of guaranteeing the interaction with the protein complex [6], were removed with the purpose of reproducing the length of the lin-41 miRNA—part of the model under investigation. In the same manner, the first seven nucleotides from the 3′-end of the second chain, considered the 'passenger strand' [6], were deleted to emulate the 15-nt long let-7 mRNA fragment under study. In this manner, the structure obtained, characterized by a chain length equal to that of miRNA:mRNA complex, allowed us to evaluate the suitability of the restraints that we applied to the structure, as explained in more detail below.

The complex was then solvated in a cubic box with about 28,046 TIP4P water molecules. The system negative net charge was counterbalanced with 210 sodium ions placed randomly in the simulation box. Even in this case, atom positions were optimized via 3 ps steepest descent and then restrained for 5 ps under a constant force of 1,000 kJmol^{-1}nm^{-2} and then restrained for 5 ps under a constant force of 1,000 kJmol^{-1}nm^{-2}. The complex was subsequently modified by deleting Ago protein residues to perform the second type of companion simulations, once again to justify the restraints imposed. The two chains so obtained were solvated in a cubic box with 6,097 TIP4P water molecules and the negative net charge counterbalanced with 29 sodium ions. The molecule was then treated following the setup above described for the other two models. All simulations were performed using the ffamber94 Force Field [20] integrated in the Gromacs 4.0.5 distribution [21, 22].

Molecular dynamics simulations

A preliminary set of MD simulations was conducted to check the conformational stability of the molecular system and to select the best simulation parameters and conditions to adopt in the subsequent free energy simulations. During these MD runs, the temperature was maintained constant at 310 K by means of temperature coupling based on a velocity rescaling with a stochastic term. All bonds were kept rigid thanks to the LINCS algorithm. Non-bonded interactions were calculated using a neighbor list that was updated every 5 time steps with a grid scheme. The cut-off distance for short-range interactions was set at 1.2 nm. Fast particle-mesh Ewald (PME) electrostatics in all the three dimensions were used to calculate long-range electrostatic forces between atom pairs inside a 1.2 nm cut-off distance [8].

The interpolation order for PME was set as cubic, with the relative strength of Ewald-shifted direct potential at a Coulomb distance at 10^{-5}. The Lennard-Jones potential was imposed as normal out to Van der Waals switch radius set at 1.0 nm; after this distance it was switched off to reach zero at 1.1 nm. The center of mass translation was removed. A leap-frog algorithm was used for integrating Newton's equations of motion. Trajectory structures for analysis were saved at 0.1 ps intervals from a 5 ns long MD simulation. Due to the not negligible helix shape modifications investigated in the past in several studies [9], the first and last nucleotides of both RNA chains were restrained with a
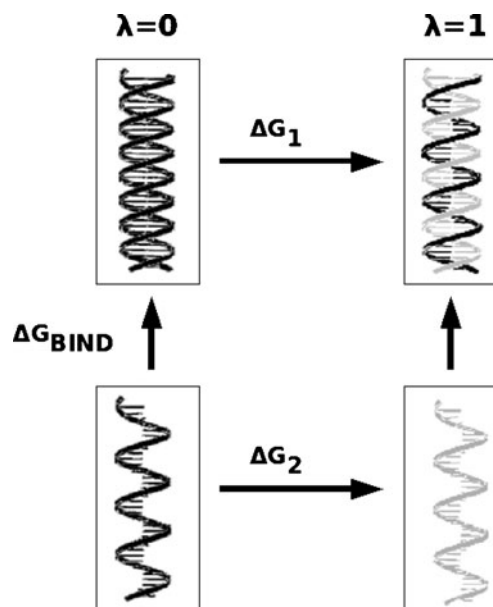

Fig. 3 Thermodynamic cycle used in binding free energy calculations. Both branches of the cycle must be considered in order to obtain the free energy of binding between the ligand, mRNA, and the receptor miRNA
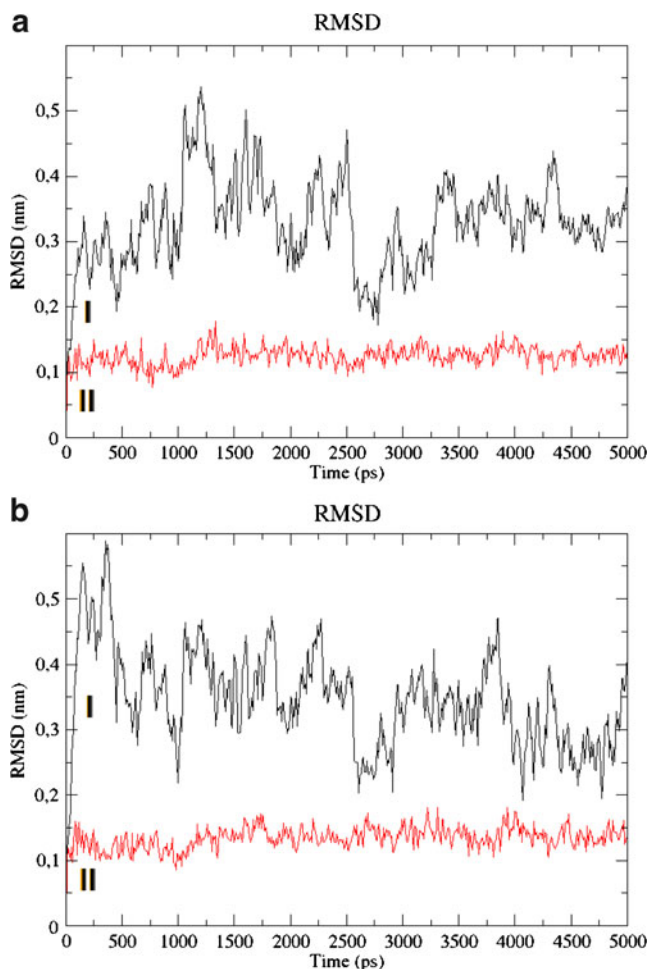
Fig. 4 B-factor analysis for let-7 miRNA and lin-41 mRNA. **a** B-factors in the molecular dynamics (MD) trajectory structures from the initial molecular configuration during the unrestrained (*trace I*) and restrained (*trace II*) simulation for Chain A. **b** The same representation for Chain B. The *vertical black lines* separate the first and last nucleotides of both chains, which were restrained during the simulation, from the rest of the strand

force constant of 1,000 kJmol^{-1}nm^{-2}. Trajectory structures were saved every 0.1 ps.

Two runs of companion MD simulations were performed. The first run on the complete siRNA-Ago complex and the second on the same molecule reduced as illustrated before removing Ago protein. Trajectory structures were saved at intervals of 2 ps from a 1 ns long MD simulation. The latter was performed under analogous conditions.

Free energy calculations

The TI approach was used to calculate the binding free energy [10, 23]. The TI technique allows calculation of the free energy difference ΔG between two states of a system, A described by a Hamiltonian $H(\lambda=0)=H_A$ and B defined by a Hamiltonian $H(\lambda=1)=H_B$, by changing the interaction

parameters that define the Hamiltonian H as a function of a coupling parameter λ from A to B state as shown in Eq. 1.

$$\Delta G_{BA} = G_B - G_A = \int_0^1 \frac{dG(\lambda)}{d\lambda} d\lambda = \int_0^1 \left\langle \frac{dH(\lambda)}{d\lambda} \right\rangle_\lambda d\lambda$$

(1)

The thermodynamic cycle implemented to achieve the desired binding free energy value is reported in Fig. 3 [11]. Taking into account the scheme in Fig. 3, the energy spent in binding let-7 miRNA to its target lin-41 mRNA was obtained by performing two main simulations with TI: one for the complex let-7 miRNA:lin-41 mRNA and the other for the ligand lin-41 mRNA only. In the first, the system was mutated from an A state, in which miRNA and mRNA were bound together, to a B state in which only the miRNA was present. The second instead evaluated the free
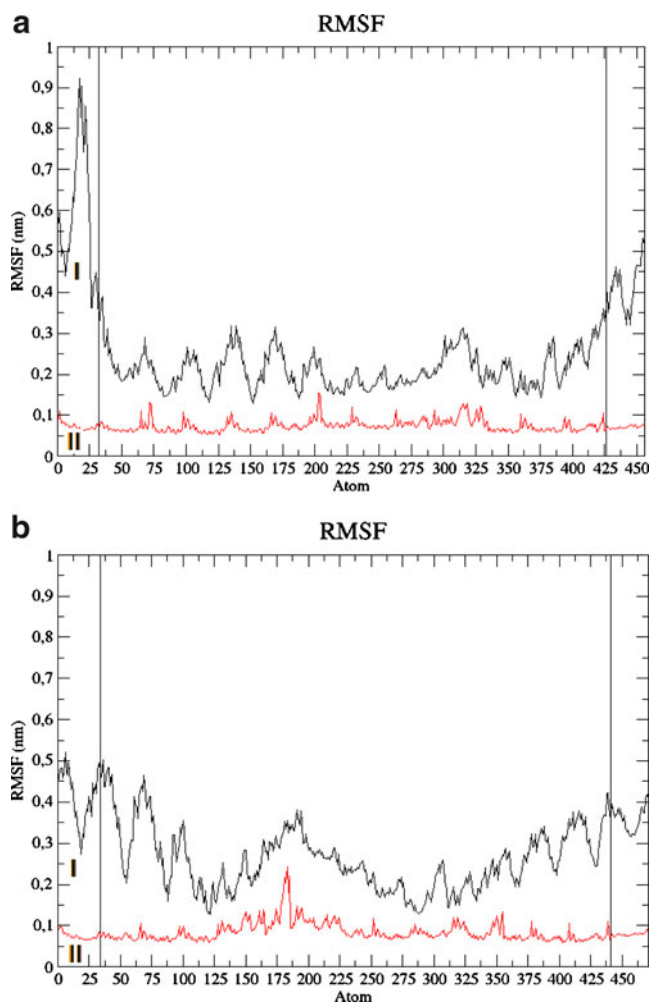

Fig. 5 Root mean square deviation (RMSD) values for let-7 miRNA and lin-41 mRNA. **a** RMSD in the MD trajectory structures from the initial molecular configuration during the unrestrained (*trace I*) and restrained (*trace II*) simulation for Chain A. **b** Same representation for Chain B
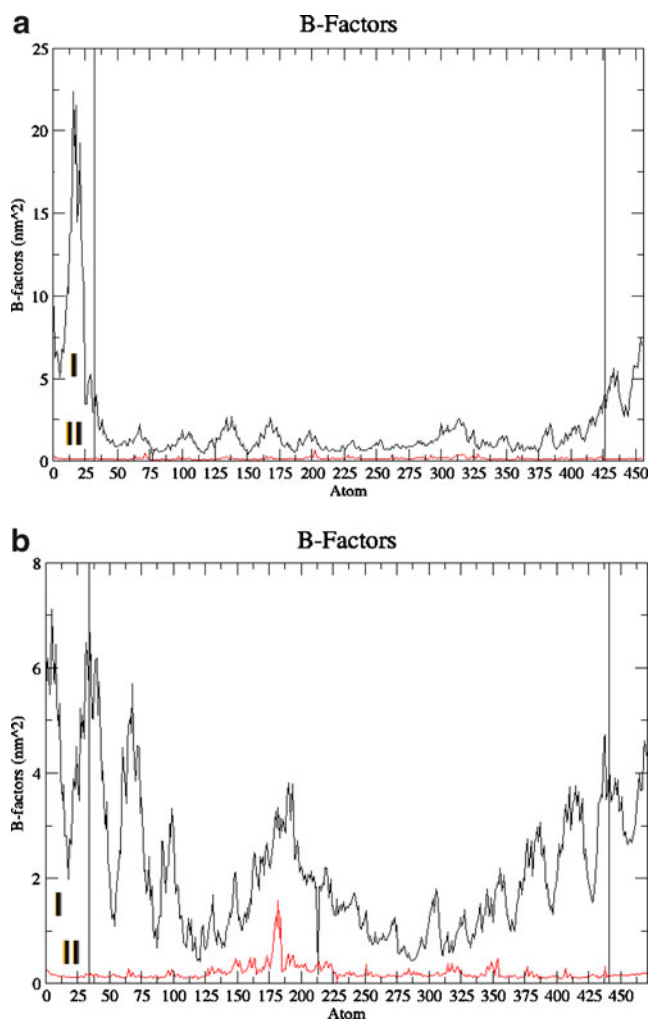
Fig. 6 Root mean square fluctuation (RMSF) values for let-7 miRNA and lin-41 mRNA. **a** RMSF in the MD trajectory structures from the initial molecular configuration during the unrestrained (*trace I*) and the restrained (*trace II*) simulation for Chain A. **b** Same representation for Chain B. The *vertical black lines* separate the first and last nucleotides of both chains, which will be restrained during the simulation, from the rest of the strand

energy of formation of the ligand mRNA, performing a mutation from an *A* state in which mRNA was present to a *B* state in which it was absent. Mutations from the *A* to *B* state were achieved by gradually switching off the non-bonded interactions from the initial to the final state. Atom type for the *B* state was set as an atom type characterized by null values for Lennard-Jones σ and ε parameters. In this manner Lennard-Jones energies were progressively reduced from the *A* state to reach zero in the *B* state. Setting even atomic charges for the *B* state to zero, all the non-bonded energies were gradually excluded from the *A* to *B* state. It is worth nothing that, during the simulation involving the let-7 miRNA:lin-41 mRNA complex, in order to implement the conditions described by the thermodynamic cycle, the *B* state was edited for the mRNA only, whereas the parameters set for

the miRNA *B* state were the same as those of the *A* state. At any intermediate state the Hamiltonian of the system was defined by Eq. 2.

$$H(\lambda) = (1 - \lambda)H_A + \lambda H_B \qquad (2)$$

Regarding Eq. 2, the ensemble averages of the derivative of the Hamiltonian were calculated, for both the complex and for the single mRNA strand, via 1 ns long MD simulations. The binding free energy was obtained using 13 different numerical lambda values, from 11 to 60, in the interval between 0 and 1. All the simulations were conducted by imposing a constant pressure of 1 atm with Rahman-Parrinello barostat and incorporating both soft-core potentials. With the purpose of ensuring the stability of the complex, the first and last nucleotides of each strand were restrained as indicated above.

The averages of the derivatives obtained from the simulations were then integrated, giving rise to one energy value for the first simulation involving the miRNA:mRNA complex (ΔG_1 in Fig. 3) and one for the second involving only the lin-41 mRNA (ΔG_2 in Fig. 3). The binding free energy was then calculated by subtracting the first quantity from the second as illustrated in Fig. 3.

Results and discussion

Molecular dynamics

MD simulations without TI were performed initially to test the stability of the let-7:lin-41 complex, which is known to depend strictly on several parameters. Indeed, it is due to these parameters that, in the past, MD simulations of nucleic acids
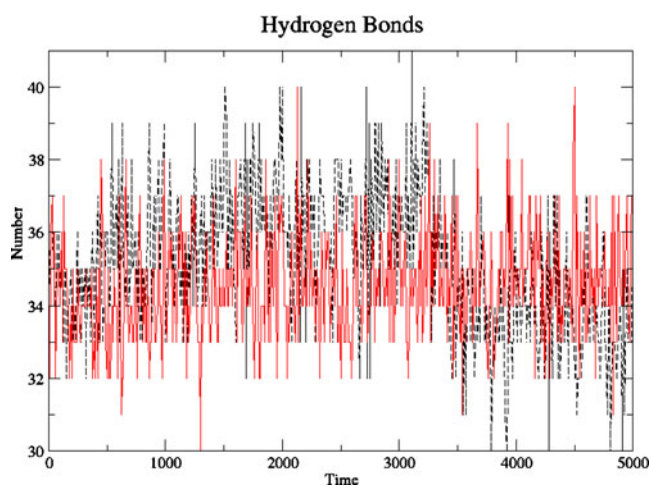

Fig. 7 Hydrogen bonds during MD simulations. *Broken black line* Trend of hydrogen bonds formed during the unrestrained simulation, *continuous red line* number of hydrogen bonds during the restrained simulation

have lagged behind protein simulations. In particular, in comparison to proteins, nucleic acids have a non-globular structure, are remarkable charged, and their flexibility is influenced by their secondary structure [5, 8, 12].

The MD simulations on the two chains that are part of the complex were analyzed in term of root mean square deviation (RMSD) and root mean square fluctuation (RMSF). Moreover, as an average of the overall simulation, B-factors [26] were calculated since they also reflect simulation accuracy (Fig. 4a,b). To further consider the stability of the interaction between miRNA and mRNA throughout the dynamic simulation, the number of hydrogen bonds between the chains every 0.1 ps of simulation was also evaluated. In this manner, we were able to point out base pair opening or closing phenomena affecting the structure during the run.

Concerning let-7 miRNA, called 'Chain A', an average RMSD value of 0.33 nm was obtained (trace 'I' in Fig. 5a), with a maximum RMSD value around 0.54 nm at roughly 1,200 ps of simulation time. The lin-41 mRNA, called 'Chain B' presents instead a RMSD mean value of about 0.35 nm, with remarkable deviation with respect to the initial trajectory in the first part of the run, which reaches its maximum value of about 0.63 nm at 350 ps of simulation (trace 'I' of Fig. 5b).

With the aim of assessing the contributions of individual atoms to the RMSD trend just presented, the root mean square fluctuation (RMSF) in the MD trajectory structures from the initial molecular configuration was calculated for both chain A and chain B.

Looking at traces 'I' in Fig. 6a,b, it can be seen that the maximum fluctuations are achieved by both the chains in connection with the atoms located at the beginning and the end of the strands, and that considerable deviations are present also in the areas characterized by the internal tetraloop: for Chain A in the interval between atom number 265 and 327, and for Chain B between atom number 130 and 220. This behavior is in agreement with that reported in literature for similar structures [19].

Figure 7 shows the number of hydrogen bonds throughout the overall simulation for the restrained simulation (red continuous line) and the unrestrained (broken black line). At the beginning of the dynamics, there were 33 H-bonds but this number is not maintained constant throughout the run. In the unrestrained dynamics, the mean value of hydrogen bonds increases slightly (mean value = 35); this increase could be related to the initial base pair opening and closing activity responsible for the loss of stability of the complex as pointed out by Wang et al. [24].

In order to maintain the original helix shape during the run, while at the same time modifying the natural dynamics of the complex as little as possible, restraints were imposed only in those positions characterized by higher RMSF

values with the exception of the tetraloop region, simply to avoid imposing an unrealistic reduction in flexibility on the complex. So, about 30 atoms at the beginning and end of both chains were restrained: since, in Amber Force Field, a nucleotide is composed of about 30 atoms during the following simulations, the atomic fluctuations of the first and last nucleotides of both chains were limited. Analyzing line 'II' of Fig. 6a,b, a significant reduction in the average RMSF values for both chains can be noted. In particular, concerning chain A, the mean RMSF after imposition of the restraints is equal to 0.07 nm, less than half that calculated for the unrestrained simulation (0.25 nm), whereas for the second strand, the value changes from 0.2 nm without restraints to 0.08 nm with restraints. Analogous considerations can be made for the RMSD, which is reduced (trace 'II' in Fig. 5a,b). The RMSD average value for Chain A is

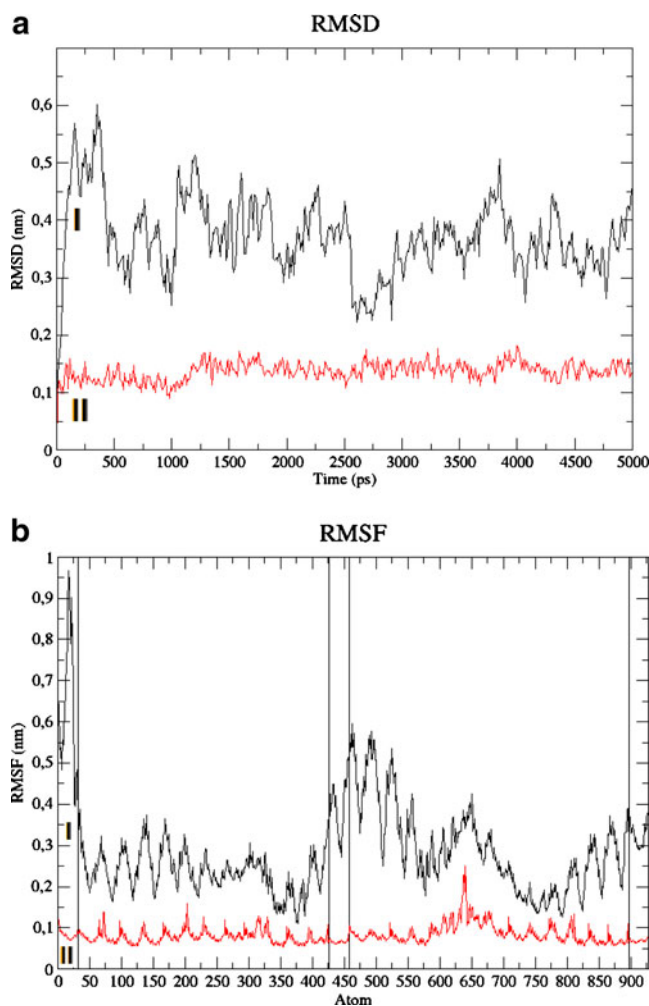**Fig. 8** RMSD and RMSF values for let-7 miRNA:lin-41 mRNA. **a** RMSD in the MD trajectory structures from the initial molecular configuration during the unrestrained (*trace I*) and restrained (*trace II*) simulation. **b** Same representation for RMSF. The *vertical black lines* separate the first and last nucleotides of both chains, which will be restrained during the simulation, from the rest of the strand

now 0.125 nm whereas for chain B it is 0.13 nm. Traces 'I' and 'II' in Fig. 8a show the RMSD values for the whole molecular structure achieved during the unrestrained and restrained simulations, respectively: after the imposition of restraints, the maximum RMSD value achieved is about 0.18 nm after 4,000 ps of run. Traces 'I' and 'II' of Fig. 8b show the RMSF values for the whole molecular structure during the unrestrained and restrained simulations, respectively: in the first case, the average value is about 0.28 nm, whereas it is 0.08 nm in the latter.

By imposing restraints, the original helix shape (Fig. 9a) is better maintained during the run, as can be noted from the images in Fig. 9d,e, in comparison to what happens for the unrestrained simulation (Fig. 9b,c).

The four restraints applied to the first and last nucleotides of the chains can be justified by considering the structure and the biology that govern both miRNA and mRNA functionality. In relation to the first strand, in nature, after several enzyme-dependent modifications, miRNAs are incorporated into 550 kDa RNP complexes, from within which they carry out their activities [14]. The main component of these complexes is a bilobal protein from the *Ago* family [6, 16, 17]; in *C. elegans*, for example, it is the RDE-1 protein [15]. As shown

in previous studies on a ternary complex of wild-type *Thermus thermophilus* Argonaute [17, 25], and recently investigated in depth by Wang and coworkers [24], the site of binding between an RNA fragment and the Ago protein is attributed to the extension of the complementarities between the RNA bound to the protein and its passenger strand.

In particular, pairing beyond the seed segment antici-pates the release of the 3′-end of the guide strand that was bound to the protein [17], whereas the 5′-end remains bound to the protein. By hypothesizing a similar interaction between miRNA, RNP and mRNA, since for let-7 miRNA: lin-41 mRNA a base pairing beyond the seed sequence was pointed out in the literature [7], the imposed restraints allows us to reproduce the conditions found in nature, in which the miRNP binds the complementary mRNA sequences located on a long RNA fragment in a stable way. In particular, as sketched and summarized in Fig. 10, the restraints applied to the mRNA fragment model the effect of more than 6,000 nucleotides that in nature surround the LCS2 fragment on the lin41 mRNA, whereas for the miRNA, the restraint applied to the 3′-end model the effect of the RNP complex bound to the opposite end of the miRNA. On the other hand, the restraints applied to the
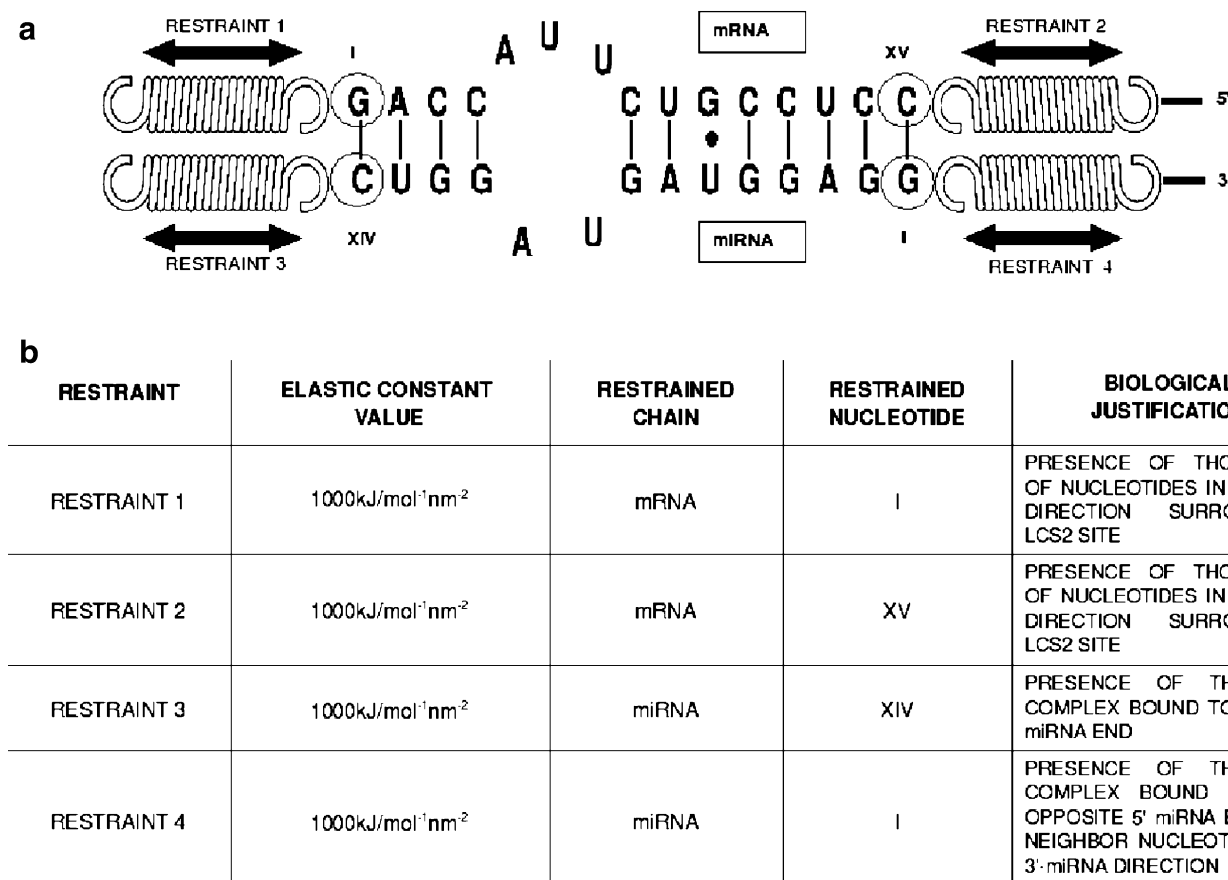

RESTRAINT	ELASTIC CONSTANT VALUE	RESTRAINED CHAIN	RESTRAINED NUCLEOTIDE	BIOLOGICAL JUSTIFICATION
RESTRAINT 1	1000kJ/mol⁻¹nm⁻²	mRNA	I	PRESENCE OF THOUSANDS OF NUCLEOTIDES IN 3' mRNA DIRECTION SURROUNDING LCS2 SITE
RESTRAINT 2	1000kJ/mol⁻¹nm⁻²	mRNA	XV	PRESENCE OF THOUSANDS OF NUCLEOTIDES IN 5' mRNA DIRECTION SURROUNDING LCS2 SITE
RESTRAINT 3	1000kJ/mol⁻¹nm⁻²	miRNA	XIV	PRESENCE OF THE RNP COMPLEX BOUND TO THE 5' miRNA END
RESTRAINT 4	1000kJ/mol⁻¹nm⁻²	miRNA	I	PRESENCE OF THE RNP COMPLEX BOUND TO THE OPPOSITE 5' miRNA END AND NEIGHBOR NUCLEOTIDES IN 3'-miRNA DIRECTION

Fig. 9 Comparison of helix shape modifications. **a** Initial molecular configuration of the complex; **b**, **c** atomic structure during the unrestrained simulation at 1 ns (**b**) and 5 ns (**c**) of run; **c**, **e** molecular complex during the restrained simulation at 1 ns (**d**) and 5 ns (**e**) of run

Fig. 10 Model and explanation of the molecular system during the restrained simulation. **a** Scheme of the restrained molecular complex, **b** summary of the restraints applied
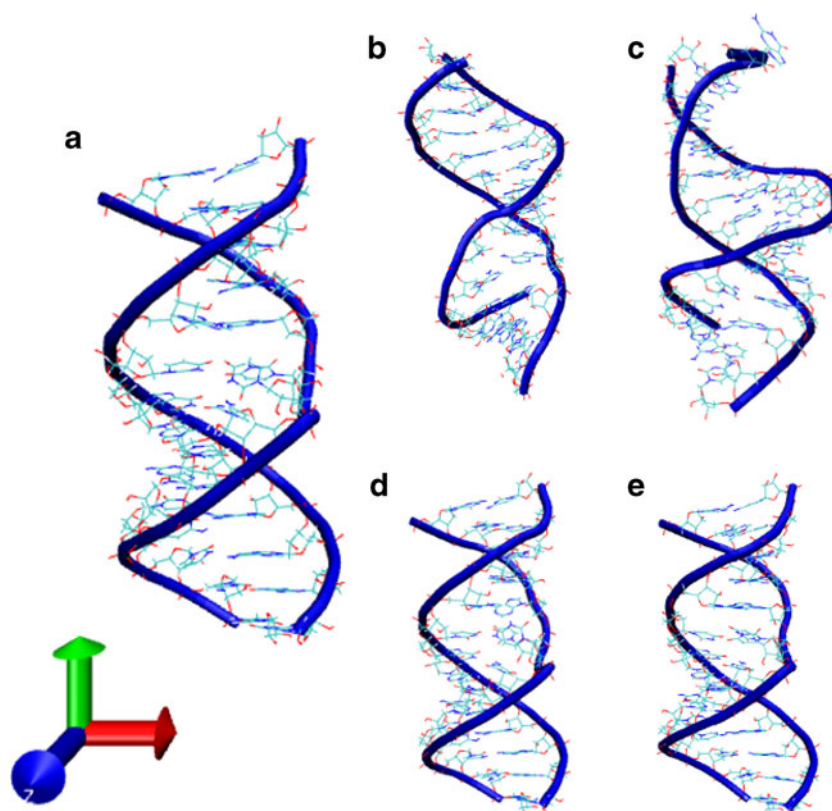
5′-miRNA model both the effect imposed by the protein bound to the RNA fragment and also the neighboring nucleotides in the 3′ direction.

The assumptions adopted when restraints were applied to the miRNA fragment were considered plausible, taking into account not only previous biological research [14], but also other molecular structures considered in the next two simulations that were modified, as stated above, in order to better mould the complex under investigation. The first MD simulation was performed on the whole Ago-siRNA complex. No discernable helix shape modifications or base pairs opening were highlighted in this trajectory. The average RMSD value for the chain bound to the complex, previously referred to as 'Chain A', is here in fact equal to 0.252 nm, whereas for the other, 'Chain B', has an RMSD value of 0.314, even in the absence of restraints. The same discussion can be applied to RMSF values, which are 0.140 for Chain A and 0.179 for Chain B. In order to verify if this more stable siRNA behavior was related to Ago bound to the 'guide' strand, another

simulation was run on the complex in the absence of protein residues. It is worth nothing that removing Ago protein residues led to a remarkable increase in the deviation of the trajectory structures from the initial chains' structure (by an order of magnitude). Since the secondary structure of the chains was the same in the last two simulations performed, it is possible to assert a role for Ago in chain stabilization, as assumed when restraints were adopted for let-7 miRNA.

Free energy calculations

The free energy of binding values calculated were obtained by starting with the lambda number set to 11 and increasing it progressively by increments of four values until (for a fixed lambda number), the percentage error of the combined averages of the derivative of the Hamiltonian between two successive lambda values was less than a threshold set at 4% with respect to the total variation between $\lambda=0$ and $\lambda=1$. The choice of this threshold was obtained as a trade-off between

Table 1 Average percentage errors between successive lambda values

λ values	11	15	19	23	27	31	35	39	43	47	51	56	60
$\%\varepsilon(\Delta G_1)$	11.59	8.28	6.45	5.27	4.46	3.86	3.41	3.05	2.76	2.52	2.25	2.07	1.8
$\%\varepsilon(\Delta G_2)$	11.63	8.31	6.40	5.29	4.47	3.88	3.42	3.06	2.77	2.53	2.32	2.09	1.91
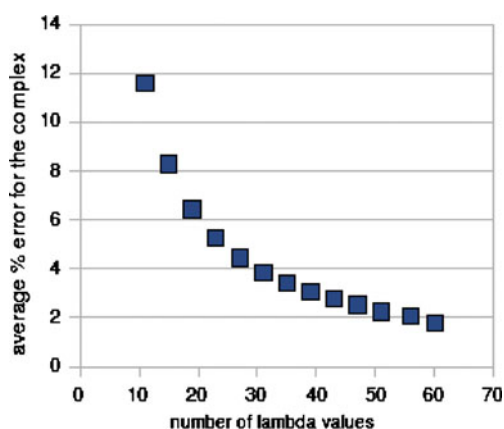
Fig. 11 Trend of the average percentage error between successive lambda values

Fig. 12 Trend of the percentage error on the binding free energy values

the need for good sampling of the phase space and the demands of limited computational time. An average time of 32 h is required for a single energetic lambda value calculation using 4+4 Intel(R) Core(TM) i7 CPUs 920 @ 2.67GHz.

To reach the desired accuracy with a minimum number of simulations, the number of lambda values considered was increased progressively by selective sampling. That is, by adding new lambda values in those intervals characterized by a higher percentage error with respect to the variation of the ensemble averages of the derivative of the Hamiltonian from $\lambda=0$ to $\lambda=1$. In this manner we, obtained a progressive improvement in the energetic information, as we will show in the following.

The selective sampling procedure started from 11 λ values, separated by a constant gap in the interval from $\lambda=0$ and $\lambda=1$. Under these uniform sampling conditions, a very high variation in the ensemble averages of the derivative of the Hamiltonian was seen close to the extreme values of the interval (i.e., 0 and 1). For instance, the energetic value at $\lambda=0.1$ shows a relative variation from that in $\lambda=0$ of more than 87% with respect to the variation in the whole interval between $\lambda=0$ and $\lambda=1$. A similar behavior was observed around $\lambda=1$. For this reason, we increased incrementally the number of lambda values used to perform the free energy of binding calculations around the extreme points. The same strategy was then applied successively in the central regions of the interval.
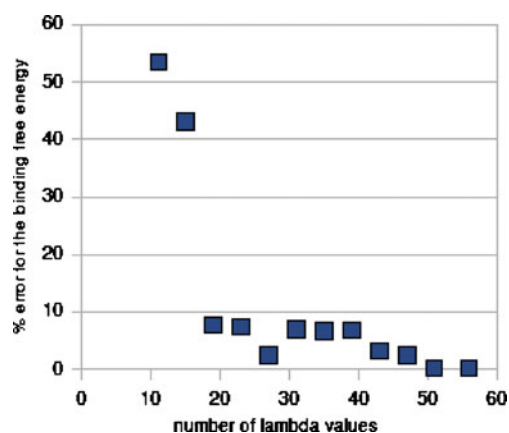
Following this strategy, the criterion of the error threshold was satisfied across the whole interval for both ligand and complex with 60 lambda values. Using this sampling conditions, the free energy for the branch relative to the complex was equal to 4,043.57 kcal/mol, whereas for the complex it was 3,939.32 kcal/mol. As such, the final binding free energy obtained was −104.25 kcal/mol. Table 1 lists the average percentage errors between two successive lambdas for the two branches of the thermodynamic cycle. Figure 11 shows the plot relative to the complex. Note that the curve for the ligand is very similar. Looking at Fig. 11, it is noteworthy that, starting from 51 lambda values, the shape of the error curve clearly reaches asymptotic behavior with percentage errors around 2%.

This aspect was further investigated by considering the accuracy of the free energy results obtained for both branches of the thermodynamic cycle and for the final binding value.

Once the desired sampling is achieved, characterized by a maximum percentage error of 4% between two successive lambda values for the single branches, the accuracy of the energetic results obtained was in fact analyzed by considering the percentage error on the binding free energy value found for the two branches of the thermodynamic cycle. As in the previous case, we report the percentage error for an increasing number of lambda values. Remember that the relative error is computed with respect to the error achieved using 60 lambda values, for which the threshold error criterion is satisfied.

Table 2 Average percentage error on the free energy values calculated

λ value	11	15	19	23	27	31	35	39	43	47	51	56
$\%\varepsilon(\Delta G_1)$	9.05	5.02	0.16	0.058	0.021	0.028	0.020	0.011	0.029	0.019	0.0053	0.0047
$\%\varepsilon(\Delta G_2)$	10.70	6.29	0.38	0.26	0.11	2.67	0.20	0.90	0.055	0.046	0.0012	0.0047
$\%\varepsilon(\Delta G_{BIND})$	53.14	43.08	7.86	7.54	2.50	7.06	6.67	6.87	3.20	2.50	.25	.25

Table 3 Central processing unit (CPU) statistics

Number of parallel processes	2	4	8	12	16	24	32	63
Average simulation time (s)	382	465	555	819	1,081	1,638	2,201	4,332
Average % CPU utilization for processes (%)	99.81	99.73	99.57	69.68	52.20	34.52	29.04	11.55

Table 2 lists the calculated errors, whereas Fig. 12 shows the trend of the error for the final free energy of binding value. It can be seen that very good accuracy is reached already with 19 lambda, and a further improvement is obtained with 51 lambda. In particular, from 51 lambda values, even if the maximum percentage error between two successive lambda is slightly higher than the threshold, the accuracy of the result is almost constant.

By comparing the results of free energy of binding obtained using 60 lambda values with that predicted by the Server Vienna [18], we observe a higher absolute value of the energy computed by MD simulations (i.e., a lower negative energy). The free energy of hybridization calculated by Vienna is in fact equal to −18.87 kcal/mol. This highlights the more stable interaction between the strands. The reason for this difference could be due to the different strategy used to calculate the property of interest. Vienna indeed obtains the binding free energy by considering only secondary structure, thus neglecting the 3D configuration of the interaction.

We believe that by considering also the dynamics of the complex and its spatial arrangement, it is possible to give a more realistic prediction of the stability of the binding, determined by the number and strength of hydrogen bonds established between the chains. Variations in atomic disposition or movement lead to variations in the strength of these hydrogen bonds. Future work will aim to evaluate the binding free energy for other validated miRNA targets in order to further investigate the presence of a correlation between particular atomic configurations and stable binding.

Scalability analysis

In the following, we report the characterization of the stability of binding free energy computations on a Symmetric Multi-processing (SMP) architecture, namely a 4+4 Intel(R) Core (TM) i7 CPUs 920 @ 2.67 GHz machine. The simulations for various lambdas were implemented as independent processes. In order to assess scalability properties, the experiments were conducted using the set up used for the MD-TI simulations discussed previously, but with a length reduced to 5 ps. By profiling short and long simulations, we observed the same characteristics in terms of system resource utilization (i.e., percentage of memory and cache accesses). As such, we decided we could safely use shorted simulations for the scalability exploration.

We ran 2, 4, 8, 12, 16, 24, 32 and 64 parallel simulations during which the average percentage CPU user utilization for the processes was recorded, in the first three cases every 48 s, and in the last every 78 s, by taking advantage of the *Sysstat* utilities designed for Debian 5 Lenny, a collection of performance monitoring tools for Linux and, in particular, of the command *pidstat*, aimed at reporting statistics for Linux tasks. We will report here only the CPU user utilization statistics obtained, because the system utilization for the processes was found to be negligible for these kinds of simulations. Table 3 reports the average percentage CPU utilization for the processes, and the mean simulation times obtained by increasing the number of parallel processes executed.

As shown by trace 'I' in Fig. 13, the real trend of the average CPU utilization for the processes is very different from the ideal situation represented by Fig. 13, trace 'II'. Trace 'II' represents an ideal case obtained by assuming a perfect independence between processes, i.e., simulation time is independent of the number of lambdas. In this case, even if we increase the number of simulations, and thus the workload imposed on the system (but also the accuracy of the free energy computation), the simulation time remains constant until the number of CPUs is larger or equal to the number of processes. Then it increases in a linear way.
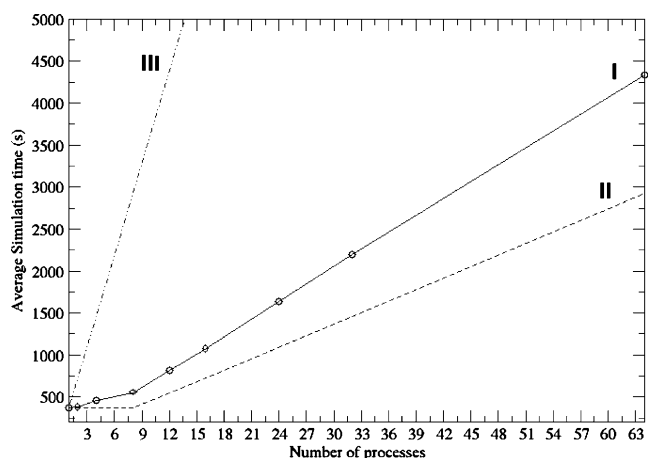

Fig. 13 Trend of simulation time with an increase of running processes. *Traces I* and *II* are the real and the ideal trend for simulation time obtained by running an increasing number of parallel processes, respectively, whereas *trace III* is that relative to an increasing number of serial processes. The *circles* on Trace 'I' mark the number of processes for which tests have been made

Trace 'I' is far from the ideal behavior represented by trace 'II', as the simulation time increases as a function of lambda. This is due mainly to two effects. The first effect is the impact of memory congestion. Indeed, even if the processes do not communicate or use shared memory regions, there is a considerable contention for access to the L2 cache and the main memory. This contention increases with the number of processes. The second effect is the context switch and process migration overhead. The latter arises when the number of processes exceeds the number of available CPUs (i.e., when the number of lambdas is larger than 8 in the considered architecture). Indeed, it can be observed that trace 'I' is characterized by two different trends. The first is from 1 to 8 lambdas, where only the memory contention contributes. The second, from 8 to 64 lambda, is where the context switch and process migration overheads are also present, leading to an increasing slope of the curve. In Table 3, the percentage CPU utilization for each simulation is reported. The utilization is almost 100% before 8 lambda and decreases after 8 lambda, where each process has to share the CPU with others.

However, it is worth noting that, besides the above discussion, parallelization is unquestionably more advantageous with respect to performing serial simulations, whose simulation times are reported in Trace 'III' of Fig. 13. Overall, combining the accuracy results shown in Fig. 12 with the computation costs reported in Fig. 13, we can conclude that a suitable trade-off between computation and accuracy can be found with 19 lambda, which is the value for which we have a big reduction in the free energy computation error from 43.08% to 7.86% (see Table 2).

Conclusion

In this paper, we have presented a methodology to estimate the binding free energy of miRNA:mRNA interaction via MD simulations. The methodology has been designed to face the various challenges of nucleic acid simulations and binding free energy computations. First, the positioning of restraints was determined in order to overcome stability problems due to the remarkable folding activity and the base pair opening of the complex. As a result, we were finally able to perform MD simulations under stable conditions.

A second challenge was the accuracy of the TI steps (i.e., the number of lambda values for sampling of the derivative of the Hamiltonian). In this regard, we studied the scalability of the TI to determine a reasonable trade-off between computation and accuracy. We have shown that, to achieve a maximum percentage error between successive lambdas values (for a fixed number of lambda) lower than 4%, we need to perform 60 parallel simulations (i.e., corresponding to 60 lambda). On the other hand, a good trade-off between computation and accuracy is already reached using 19 parallel simulations.

This methodology has allowed us to understand the implications of 3D structure for miRNA–mRNA interactions in more depth. That is, the free energy of binding calculated for the let-7 miRNA:lin-41 mRNA interaction proved the stronger pairing between the RNA strands compared to that estimated by means of the Vienna Server [18]. The energy deviation, about 85 kcal/mol, confirms that 3D atomic arrangements can play a fundamental role in the recognition of miRNA targets, and opens the field to strategies for the integration of this information into miRNA target prediction tools.

References

1. Triboulet R, Gregory RI (2010) Pumilio turns on microRNA function. Nat Cell Biology 12(10):928–929
2. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116(2):281–297
3. Min H, Yoon S (2010) Got target?: computational methods for microRNA target prediction and their extension. Exp Mol Med 42 (4):233–244
4. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R (2010) Control of translation and mRNA degradation by miRNAs and siRNAs. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
5. Cheatham TE III (2004) Simulation and modelling of nucleic acid structure, dynamics and interactions. Curr Opin Struct Biol 14 (3):360–367
6. Yuan Y-R, Pei Y, Chen H-Y, Tuschi T, Patel DJ (2006) A potential protein–RNA recognition event along the RISC-loading pathway from the structure of A. aeolicus Argonaute with externally bound siRNA. Structure 14(10):1557–1565
7. Vella MC, Reinert K, Slack FJ (2004) Architecture of a validate microRNA::target interaction. Chem Biol 11:1619–1623
8. Norberg J, Nilsson L (2002) Molecular dynamics applied to nucleic acids. Acc Chem Res 35:465–472
9. Pan Y, Mackerell AD Jr (2003) Altered structural fluctuations in duplex RNA versus DNA: a conformational switch involving base pair opening. Nucleic Acids Res 31(24):7131–7140
10. Knight JL, Brooks CL III (2009) λ-dynamics free energy simulation methods. J Comput Chem 30(11):1692–1700
11. Lawrenz M, Baron R, McCammon JA (2009) Independent-trajectories thermodynamic-integration free-energy changes for biomolecular system: determinants of N5HN1 avian influenza virus neuraminidase inhibition by Peramivir. J Chem Theory Comput 5 (4):1106–1116
12. MacKerell AD Jr, Nilsson L (2008) Molecular dynamics simulations of nucleic acid-protein complexes. Curr Opin Struct Biol 18(2):194–199
13. Cevec M, Thibaudeau C, Plavec J (2008) Solution structure of a let-7 miRNA:lin41 mRNA complex from C. elegans. Nucleic Acids Res 36:2330–2337
14. Schwarz DS, Zamore PD (2002) Why do miRNAs live in the miRNP? Genes Dev 16:1025–1031
15. Tomari Y, Zamore PD (2005) Perspective: machines for RNAi. Genes Dev 19(5):517–529
16. Tanaka Hall TM (2005) Structure and function of Argonaute proteins. Structure 13:1403–1408
17. rWang Y, Juranek A, Li H, Sheng G, Tuschl T, Patel DJ (2008) Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature 456:921–926
18. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–3431

19. DePaul AJ, Thompson EJ, Patel SS, Haldeman K, Sorin EJ (2010) Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics. Nucleic Acids Res 38:4856–4867
20. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197
21. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317
22. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718
23. Chipot C (2005) Free energy calculations in biological systems. How useful are they in practice? In: Leimkuhler B, Chipot C, Elber R, Laaksonen A, Mark AE, Schlick T, Schütte C, Skeel R (eds) New algorithms for macromolecular simulation, vol 49. Springer, Berlin, pp 183–209
24. Wang Y, Li Y, Ma Z, Yang W, Chunzhi A (2010) Mechanism of microRNA-target interaction: molecular dynamics simulations and thermodynamic analysis. PLoS Comput Biol 6(7):e1000866. doi:10.1371/journal.pcbi.1000866
25. Wang Y, Juranek S, Li H, Sheng G, Wardle GS, Tuschl T, Patel DJ (2009) Nucleation, propagation and cleavage of target RNA in Ago silencing complexes. Nature 461:754–756
26. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. Fold Des 2(3):173–181