Scuola Internazionale di Studi Superiori Avanzati

Corso di dottorato in Fisica e Chimica dei Sistemi Biologici

Studying the dynamical properties of small RNA molecules with computational techniques

Supervisor:
Prof. Giovanni Bussi

PhD candidate:
Giovanni Pinamonti

Anno Accademico 2015-2016

Abstract

The role of ribonucleic acid (RNA) in molecular biology is shifting from a mere messenger between DNA (deoxyribonucleic acid) and proteins to an important player in many cellular activities. The central role of RNA molecules calls for a precise characterization of their structural and dynamical properties. Nowadays, experiments can be efficiently complemented by computational approaches.

This thesis deals with the study of the dynamical properties of small RNA molecules, exploiting various computational techniques. Specifically we investigate two different complementary methods, elastic network models (ENMs) and Markov state models (MSMs).

ENMs are valuable and efficient tools for characterizing the collective internal dynamics of biomolecules. We evaluate their performance by comparing their predictions with the results of atomistic molecular dynamics (MD) simulations and selective 2′-hydroxyl analyzed by primer extension (SHAPE) experiments. We identify the optimal parameters that should be adopted when putting into use such models.

MSMs are tools that allow to probe long-term molecular kinetics based on short-time MD simulations. We make use of MSMs and MD simulations to measure the kinetics and the timescale of the stacking-unstacking motion for a collection of short RNA oligonucleotides, comparing the results with previously published relaxation experiments. We then move to the study of the process of the fraying of the terminal base pair in a helix, characterizing the different involved pathways and the sequence dependence of the process timescale.

Acknowledgements

First of all I have to thank Giovanni, by far the best PhD supervisor that I've ever had. Then my "boss-doc", Sandro, that guided my through the first years of my PhD, with infinite patience and sincere friendship. I also would like to thank all the SBP group: the professors, the post-docs, the students, and all the past members, because, in the last four years in SISSA, I've never felt surrounded by colleagues, but by friends. A special mention to my office mates (quelle del quarto piano): I'll miss our pointless discussions on the blackboard and the endless code debugging.

I'm grateful to Frank Noè and Fabian Paul, along with all the CMB group, that welcomed me and helped me during my months in Berlin. I want to thank Professor Doug Turner, for sharing his scientific enthusiasm with me and engaging in a lot of interesting and fruitful discussions.

Per finire voglio ringraziare la mia famiglia: i miei genitori che mi hanno sempre sostenuto e incoraggiato durante i miei lunghi anni di studio, e poi i fratelli, i nonni, gli zii e i cugini. Un grazie anche a tutti gli amici, i compagni di sport, di feste, di "nerdate", di casa, di camminate, di viaggi. Sono anche tutti questi piccoli dettagli che hanno contribuito a farmi arrivare fino a qui.

E per ultima ovviamente ringrazio Alice, per avermi supportato e sopportato, sempre.

Contents

Introduction

Since the statement of the "central dogma of biology" in the '50s, ribonucleic acid (RNA) began its journey in the world of biology as a simple messenger between DNA (deoxyribonucleic acid) and proteins. Today many things have changed. Over the past three decades, discoveries have unearthed a surprising complex world of RNA molecules that seems to rival proteins in terms of three dimensional structures and biochemical functions (for a review on the topic see for example [Morris and Mattick, 2014]). The study of the conformational changes that RNA molecules undergo and their dynamical features is a fundamental step to understand the complex biological mechanisms in which these molecules are involved. Nowadays different experimental techniques can help in elucidating the structure, from the single nucleotide resolution of chemical probing experiments, to the finer scales revealed by cryo-EM, NMR and X-ray crystallography. By contrast, to elucidate fine details of the dynamics and the kinetics of these molecules is more challenging, and the amount of information that can be obtained with methods like single molecule pulling experiments, FRET, or ensemble relaxation experiments, is limited.

For this reason, computer simulations are a fundamental tool to study the functional dynamics of RNAs. Indeed, they allow to accurately model and interpret experimental data, allowing one to access to a vast quantity of fine-detailed information, otherwise unavailable. Accurate predictions can be achieved with the help of accurate, atomistic molecular dynamics (MD) simulations (see Šponer et al. [2014] for a recent review on the topic). Given the number of atoms that has to be modeled in these calculations, the simulation of biomolecules with MD is a very challenging task, and even the simulation of a few nanoseconds of dynamics usually requires hours of CPU time, while even the fastest relevant biological processes take place in time scales on the order of microseconds.

Different methods have been developed in order to circumvent this obstacle. For example, calculation of thermodynamic quantities of interest can be sped up by reducing the complexity of the model by coarse-graining the representation of the system (see, e.g., Dawson et al. [2016]), or by introducing artificial forces that guides the system through the energetic barriers separating different configurations (see Bernardi et al. [2015] for a recent review).

Several efforts are being spent towards the development of coarse-grained approaches capable of striking a good balance between having a simplified (and computationally efficient) description of the structure and interactions of a molecule and being able to capture its salient kinetic and thermodynamic features. It is important to note that, viable coarse-grained models are valued not only because they are amenable to extensive numerical characterization, but precisely because their simplified formulation can offer a valuable insight into the main physico-chemical mechanisms that affect the behavior and properties of a given biomolecule.

Because of their transparent and simple formulations elastic network models (ENMs) have proved very valuable to characterize, with a minimal computational effort, the collective internal dynamics of proteins and enzymes starting from the sole knowledge of their structures. The increasing evidence that, as for proteins, also the biological functionality of RNAs is often linked to their innate internal dynamics, poses the question of whether ENM approaches can be successfully extended to these biomolecules.

In the first part of this thesis, we tackle this still-largely unexplored issue by considering various possible families of ENMs for RNAs and assess their validity and predictive capabilities by comparison against extensive MD simulations and selective 2'-hydroxyl analyzed by primer extension (SHAPE) experimental data. We observe that the best ENM performance is attained when each nucleotide is represented by a specific combination of three ENM centroids (sugar-base-phosphate, or SBP). We also use ENM representations to estimate the entropic contributions to the free-energy of formation of tertiary structure elements of an adenine riboswitch.

While coarse-grained models as ENMs are extremely valuable tools it is sometimes necessary to take advantage of the more precise predictions available with the help of accurate, atomistic MD simulations. A possible route to bypass the aforementioned sampling problem of MD is to exploit the the constant increase in the availability of parallel computational resources. It is currently feasible to produce tens to hundreds of independent MD trajectories, that can be combined to reach timescales that start to have a relevance in biological processes. The analysis of the multiple conformational changes that occur during such long trajectories is however a challenging task. One of the most successful approaches is to make use of the mathematical framework of Markov State Models (MSMs) (see Pande et al. [2010] and Chodera and Noé [2014]). This methods have been shown to be extremely successful in analyzing MD trajectories of many biomolecular systems.

The MSM framework can be employed to elucidate the kinetics of conformational changes that are strictly tied to the sophisticated cellular processes played by RNA. These dynamics are often regulated by different signals such as ligand concentration, temperature or pH, and may take place over different timescales, from picoseconds to minutes [Mustoe et al., 2014]. We decided to focus on the two key interactions in the formation of RNA 3D structures, namely base pairing and base stacking.

In the second part of this thesis we investigate the kinetics of stacking and pairing in RNA. By analyzing atomistic MD simulations with MSM we reveal that the main relaxation modes of RNA oligonucleotides consist in transition between alternative folded states and that the kinetic properties predicted by the current RNA amber99 force-field are consistent with the results of previously reported relaxation experiments. Moreover we present a novel combination of methods to construct MSMs and apply it to unravel the kinetics of the opening, or fraying, of an RNA double helix, which is the first step in many biologically relevant transitions.

Outline of the thesis

This thesis is divided in two parts, related to the two complementary and independent topics of elastic network models (ENM) (Part I) and Markov state models (MSM) (Part II), both applied to the study of the dynamical properties of RNA molecules.

Part I is organized as follows:

- Chapter 1 contains a brief introduction on the topic of ENM.
- Chapters 2 and 3 report the results of a comparison between the predictions of ENM and the results of molecular dynamics (MD) simulations and selective 2 -hydroxyl acylation analyzed by primer extension (SHAPE) experiments. This work has been published in [Pinamonti et al., 2015];
- Chapter 4 summarizes preliminary results on the applicability of ENM to estimate the vibrational entropy of RNA molecules.

Part II is organized as follows:

- Chapter 5 introduces the formalism and the basic theory behind MSM;
- In Chapter 6 MSMs are applied to the study of the kinetic properties of oligonucleotides. The content of this chapter is part of a paper currently in preparation [Pinamonti et al., 2016];

- In Chapter 7 we present and test an alternative approach to the construction of MSMs, that combines the methods of Buchete and Hummer [2008] and d'Errico et al. [2016]. The application of this method on the unzipping of the terminal base pair of a RNA double helix is described in detail in Chapter 8. The results discussed in these two chapters will be also collected in a future paper.

Part I

Elastic Network Models

Chapter 1

Theory of ENM

Understanding the functional dynamics of ribonucleic acid (RNA) molecules is extremely important due to the increasing number of known biological roles that they play. Numerical simulations that exploit accurate atomistic models are becoming more and more important in this sense, helped by the constant growth in computational resources available (see, e.g., Colizzi and Bussi [2012]; Chen and García [2013]; Kührovaá et al. [2013]; Yildirim et al. [2013]; Musiani et al. [2014]; Pan et al. [2014]; Šponer et al. [2014]). Nevertheless, the cost, in terms of time and energy consumption, is often a limiting factor for this kind of studies. For this reason less informative but faster coarse-grained representations are often the only avenue to the investigations of conformational dynamics in big and complex molecules. Elastic network models (ENM) are interesting candidates in this sense. They are simple models with a small number of tunable parameters, and, thanks to the simplicity of their potential energy form, their dynamical properties can be obtained without the need for an explicit time evolution of the equation of motion.

Such models were originally motivated by the seminal work of Tirion [1996] who showed that the low-energy structural fluctuations of globular proteins obtained by atomistic molecular dynamics (MD) simulations could be reliably reproduced by replacing the detailed inter-atomic force field by spring-like, harmonic interactions. This remarkable fact was rationalized *a posteriori* in terms of the generally collective and large-scale character that low-energy fluctuations have in proteins, which makes them amenable to be captured with models that are oblivious of the small-scale atomistic details. This observation, in turn, prompted the further development of simplified harmonic models where not only intra-molecular interactions, but the structural descriptions themselves were simplified by representing each amino acid with only few interaction centers, or centroids [Hinsen, 1998; Bahar et al., 1997; Atilgan et al., 2001].

Compared to the well-established case of proteins, the development and application of ENM to RNAs is still relatively unexplored. In fact, starting from seminal work of Bahar and Jernigan [1998], it is only recently that it has been tackled systematically [Setny and Zacharias, 2013; Zimmermann and Jernigan, 2014].

In this chapter we will introduce the general theory of ENM, that will be used in Chapters 2 and 3 when benchmarking the performance of ENM against MD simulations and SHAPE experiments.

1.1 Quadratic potential

Elastic or Gaussian models are coarse-grained representations able to capture the essential dynamical space of a macromolecule with a minimal computational cost. The basic assumption of such models is that the vibrations of a molecule (represented as a set of interaction centers, or beads) around its minimal-energy structure are sufficiently small. Therefore it is possible to expand its Hamiltonian up to the second order in terms of the deviation around the minimum. Another assumption of the ENM is that the second derivative of the potential energy can be written as a sum of two-body terms. With these simplifications, the resulting potential energy is equivalent to the one of a set of beads connected with harmonic springs with elastic constants k_{ij}:

$$V(\boldsymbol{r}_i, \boldsymbol{r}_j) = \frac{1}{2}k_{ij}(|\boldsymbol{r}_i - \boldsymbol{r}_j| - \tilde{d}_{ij})^2 = \frac{1}{2}k_{ij}(d_{ij} - \tilde{d}_{ij})^2 \tag{1.1}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are the positions of beads i and j, $\tilde{d}_{ij}$ is the distance between them in a reference structure.

Figure 1.1 shows a representation of the ENM of an RNA molecules (The sarcin ricin domain (SRD) from *E.coli* 23S rRNA), compared with the atomistic representation.

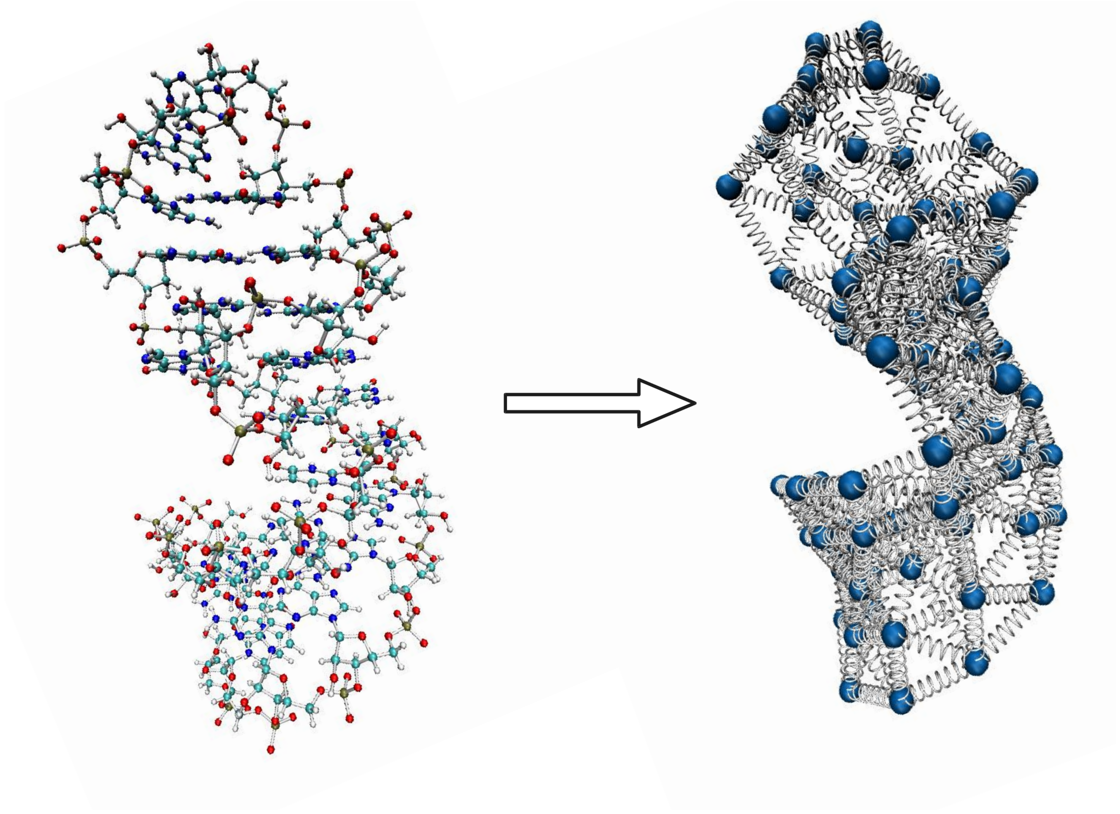


Figure 1.1: Representation of the ribosome sarcin-ricin domain with full atomistic detail (left) compared with an ENM (right). The ENM shown uses only one bead per nucleotide.

Expanding the potential energy we have

$$V(\boldsymbol{d}_{\boldsymbol{ij}}) \approx V(\tilde{\boldsymbol{d}}_{ij}) + \frac{k_{ij}}{2} \sum_{\mu\nu} \frac{\tilde{d}_{ij}^{\mu} \tilde{d}_{ij}^{\nu}}{\tilde{d}_{ij}^{2}} \delta d_{ij}^{\mu} \delta d_{ij}^{\nu} \tag{1.2}$$

where $\boldsymbol{d}_{ij} = \tilde{\boldsymbol{d}}_{ij} + \delta \boldsymbol{d}_{ij}$, μ and ν denote the Cartesian coordinates x, y, z. Eq. 1.2

can be rewritten in a simpler form as

$$V(d_{ij}) = V(\tilde{d_{ij}}) + U, \tag{1.3}$$

$$U(\delta r_{i,\mu}, \delta r_{j,\nu}) = \delta r_{i,\mu} M_{ij,\mu\nu} \delta r_{j,\nu} \tag{1.4}$$

where $\boldsymbol{M}$ is a $3N \times 3N$ symmetric matrix proportional to the Hessian of U, and $\delta \boldsymbol{r_i}$ is the deviation of bead i from its position in the reference structure.

The values of the parameters k_{ij} can be defined in different ways [Bastolla, 2014]. In this work, we considered a sharp cutoff model, in which the elastic constants are simply given by

$$k_{ij} = \begin{cases} -k & \text{if} \quad \tilde{d}_{ij} < R_C \\ 0 & \text{if} \quad \tilde{d}_{ij} \geq R_C \end{cases} \tag{1.5}$$

Here R_C is a cut-off radius, which defines the range of interaction between the beads.

1.2 Choice of the beads

The two key ingredients of ENMs are the choice of the beads and their range of interaction (R_C). Both of them must be tuned in order to obtain a model that optimally reproduces the dynamical features of the studied biomolecules. It is of course of primary importance that the choice for these two parameters is universal among a family of biomolecules.

Although the original ENM considered all heavy atoms as beads [Tirion, 1996], it was later shown that a simple coarse grain representation is able to capture the main dynamical features of a protein [Hinsen, 1998; Atilgan et al., 2001; Delarue and Sanejouand, 2002; Micheletti et al., 2004]. In their simplest formulation, ENMs incorporate harmonic interactions between pairs of C^{α} centroids while a two-centroid description, (for both the main and side-chains) appear to be optimally suited to capture pairwise mechanical correlations consistently with MD simulations [Micheletti et al., 2004; Micheletti, 2013; Fuglebakk et al., 2013].

By comparison with proteins, the development and application of ENMs aimed at nucleic acids is still relatively unexplored. Bahar and Jernigan [1998] first applied ENMs to the conformational dynamics of a transfer RNA using a model with two beads per nucleotide. Several authors further simplified this model using a single bead placed on the phosphorus atom [Tama et al., 2003; Wang et al., 2004; Van Wynsberghe and Cui, 2005; Wang and Jernigan, 2005; Fulle and Gohlke, 2008; Kurkcuoglu et al., 2009; Zimmermann and Jernigan, 2014]. This choice seems to rise from the analogy with the C_αs on the backbone of proteins, but has never been justified in a rigorous way. More recently, an extensive search on all the atoms on the backbone [Setny and Zacharias, 2013], suggested that the best candidates to host a single ENM bead is the the sugar ring in the backbone. A successive work by Zimmermann and Jernigan [2014] shown that a model with one bead in each heavy atom is the best choice in order to reproduce the variations in an ensemble composed by different experimental structures of the same molecule. An intermediate choice between one bead per nucleotide and an all-atom model has been adopted by Delarue and Sanejouand [2002] and Yang et al. [2006] that considered three beads for each nucleotide, representing the phosphate group, the ribose ring, and the nucleobase.

1.3 How to benchmark ENM against atomistic simulations

It is useful to compare the predictions of ENM with atomistic MD simulations of selected systems, in order to validate their applicability, optimally tune the model's parameters, and compare the performance of different alternative elastic models.

In order to extract from an MD simulation information that can be directly compared with ENM it is useful to consider the covariance matrix of the displacements of each atom/bead. The covariance matrix is defined as

$$C^{\mathrm{MD}}_{ij,\mu\nu} = \langle \delta r_{i,\mu} \delta r_{j,\nu} \rangle \tag{1.6}$$

$$\delta r_{i,\mu} = (r_{i,\mu} - \langle r_{i,\mu} \rangle) \tag{1.7}$$

after optimal superposition on a reference structure. The covariance matrix can be equivalently rewritten as

$$C^{\mathrm{MD}}_{ij,\mu\nu} = \sum_{\alpha} \lambda_\alpha v^\alpha_{i,\mu} v^\alpha_{j,\nu} \tag{1.8}$$

Where λ_α and $\boldsymbol{v}^\alpha$ are the eigenvalues and the eigenvectors of C^{MD}, respectively. The first (i.e. associated to the largest eigenvalues) eigenvectors of this matrix are often referred to as principal components (PC) of motion of the system. The corresponding eigenvalues represent the amplitudes of the motion of the system projected on that eigenvector.

In the case of the ENM, it is easy to show that the covariance matrix of the system is given by

$$C^{\mathrm{ENM}}_{ij,\mu\nu} = k_B T \sum_{\alpha=1}^{N-6} \frac{1}{\sigma^\alpha} w^\alpha_{i,\mu} w^\alpha_{j,\nu} \tag{1.9}$$

where σ^α and $\boldsymbol{w}^\alpha$ are the eigenvalues and the eigenvectors of the matrix M, defined in Eq. 1.4. The sum is here performed excluding the six eigenvectors with null eigenvalues, which correspond to the translational and rotational degrees of freedom. The eigenvectors and eigenvalues of $\boldsymbol{C}^{ENM}$ represent the normal modes of the elastic network and correspond to the PCs predicted by that model, that can be directly compared with those obtained from MD simulations.

Several ways to compare the prediction of ENM with MD trajectories have been developed and proposed (see for example Fuglebakk et al. [2013]). Here we summarize some of them.

Correlation between fluctuations

In the ENM framework, the mean square fluctuation (MSF) of each bead is obtained from the covariance matrix as

$$\mathrm{MSF}_i = \langle \delta r_i^2 \rangle = \sum_{\mu=1}^{3} C_{ii,\mu\mu} \tag{1.10}$$

Analogous fluctuations can be obtained from the MD simulation. The two MSFs profiles can then be compared by means of the Pearson correlation coefficient, *R*. It is worth remarking that the MSF are known to be correlated with the inverse of the number of nearest neighbors [Halle, 2002]. In this sense a realistic prediction of these amplitudes is a condition which should be easily satisfied by any meaningful ENM.

Overlap between principal components eigenspaces

In order to take into account all the relevant information embedded in the correlation matrix one should consider a quantity that relates the similarity of its PCs.

Given two eigenspaces identified by their set of eigenvectors $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ and the corresponding eigenvalues λ_i, γ_i, one can define the root mean square inner product (RMSIP) (cit.)

$$\mathrm{RMSIP} = \sqrt{\frac{1}{n} \sum_{i,j=1}^{n} (\boldsymbol{v}_i \cdot \boldsymbol{w}_j)^2} \tag{1.11}$$

Although this measure is often used to compare ENMs and MD, it bears an arbitrariness in the choice of the number of relevant modes to compare, *n*, which is usually taken to be equal to 10. The root weighted square inner product (RWSIP) was defined to overcome this difficulty [Carnevale et al., 2007], and its expression is given by

$$\mathrm{RWSIP} = \sqrt{\frac{\sum_{i,j=1}^{3N} \lambda_i \gamma_j (\boldsymbol{v}_i \cdot \boldsymbol{w}_j)^2}{\sum_{i=1}^{3N} \lambda_i \gamma_i}} \tag{1.12}$$

In the following chapter we will present our findings using the latter measure, although the RMSIP was shown to give equivalent results.

1.4 | Effective Interaction Matrix

When comparing different ENMs one must consider only the degrees of freedom in common between the models. To achieve this, it is necessary to compute the effective interaction between the degrees of freedom (i.e. beads) of interest [Zen et al., 2008; Micheletti, 2013]. Let us consider a system of N degrees of freedom, and two subsystems, a and b. The interaction matrix of the total system can be written as

$$M = \left(\begin{array}{c|c} M_a & W \\ \hline W^T & M_b \end{array}\right) \tag{1.13}$$

Where M_a and M_b are the interaction matrices of the two subsystems, while W represent the interactions between them. Now, if we are interested only in subsystem a, the effective interaction matrix governing its dynamics will be given by

$$M_a^{\mathrm{eff}} = M_a - W M_b^{-1} W^T \tag{1.14}$$

If the two subsystems are connected by a sufficient number of springs, M_b does not have null eigenvalues and can be straightforwardly inverted. For a detailed derivation of this equation see [Zen et al., 2008]. Using this effective matrix one can compute the fluctuations relative to the subsystem considered, as well as the corresponding PCs of motion.

Chapter 2

Testing ENM performance against atomistic MD simulations

In this chapter we present an exhaustive and rigorous study regarding the applicability of elastic network models (ENMs) on ribonucleic acids (RNA), comparing the predictions of the model with atomistic molecular dynamics (MD) simulations. This kind of comparison has never been performed systematically. In Setny and Zacharias [2013] only the atomistic dynamics of one simple RNA double strand was studied and in one older work [Van Wynsberghe and Cui, 2005] a comparison between ENMs and MD was performed only for a limited time-scale. Nevertheless, it is extremely interesting to compare ENM with MD because simulations can give insights on the dynamics of a molecule with an extremely fine level of detail, enabling to explore and fully understand the accuracy of ENM.

In this study, we took in exam different RNA molecules, each bearing different kind of secondary and tertiary structures, in order to investigate the applicability of ENMs on an arbitrary ribonucleic system. We also focused our attention on the comparison between different possible choices for the beads over which the ENM should be constructed in a ribonucleic system. Towards the goal of identifying the most suitable RNA ENM, we went beyond the single-centroid representation and assessed the performance of an enlarged family of ENMs where we consider several single- or multi-centroid alternative representations for ENMs as well as their cutoff interaction distance.

With our approach, that optimally complements the insight offered by previous studies, we established that the best compromise between dealing with the minimal number of degrees of freedom and yet have an accurate description of the internal dynamics, is offered by a three-centroid representation.

The main content of this chapter has been published in Pinamonti et al. [2015].

2.1 Details of the MD simulations

We performed atomistic MD simulations on four different RNA molecules (Figure 2.1). These systems were chosen so as to cover a variety of size and structural complexity and yet be amenable to extensive simulations, as detailed in Table 2.1. The RNA duplex $\substack{\texttt{GAGUGCUC}\\\texttt{CUCGUGAG}}$ with two central G-U wobble base pairs was taken from an NMR model [Chen et al., 2000]. As a second system, we considered the sarcin ricin domain (SRD) from *E.coli* 23S rRNA, which consists of a GAGA tetraloop, a flexible region with a G-bulge and a duplex region [Correll et al., 2003]. The U nucleobase at the 5′ terminal was excised from the high resolution crystal structure. Additionally, we performed MD simulations on two more complex molecules: the hammered ribozyme [Scott et al., 1996] and the *add* adenine riboswitch [Serganov et al., 2004]. Both systems are composed of three stems linked by a three-way junction. In the *add* riboswitch, two hairpins are joined by a kissing loop interaction. A schematic representation of the secondary structures is shown in Fig. 2.1. Except for the RNA duplex, all the other systems were previously studied by means of MD simulations

	PDB code	chain length	simulation time (μs)
Duplex	1EKA	16	1.0
Sarcin-ricin domain	1Q9A	25	0.9
Hammerhead ribozyme	301D	41	0.25
add Riboswitch	1Y26	71	0.25

Table 2.1: RNA dataset: details and length of MD simulations.

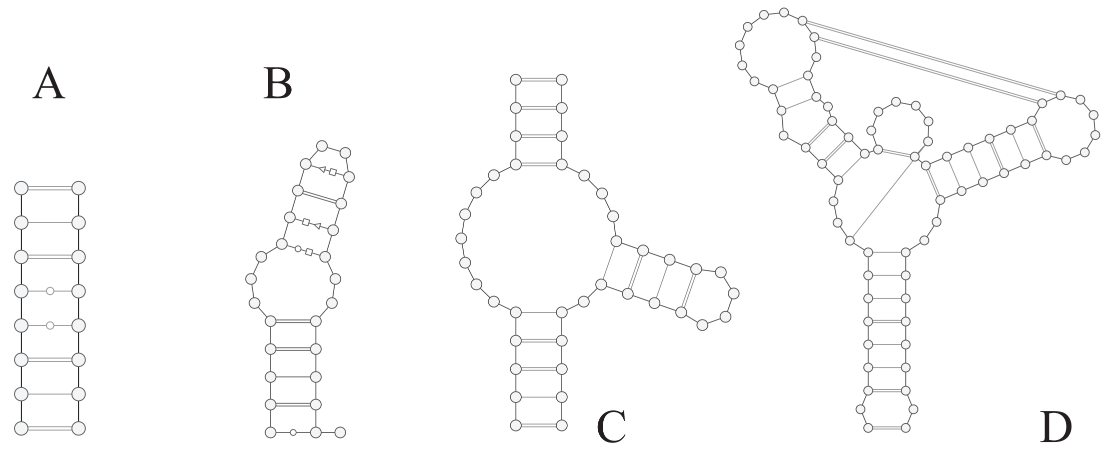


Figure 2.1: Secondary structures of the four molecules studied: A) eight-base pairs duplex; B) sarcin-ricin domain; C) hammerhead ribozyme. D) *add* adenine riboswitch;

[Špačková and Šponer, 2006; Van Wynsberghe and Cui, 2005; Priyakumar and MacKerell, 2010; Gong et al., 2011; Allnér et al., 2013; Di Palma et al., 2013, 2015].

All MD simulations were performed using GROMACS 4.6.7 [Pronk et al., 2013] with the AMBER99 force field [Hornak et al., 2006] including parmbsc0 [Pérez et al., 2007] and χ_{OL3} [Banáš et al., 2010] corrections. GROMACS parameters can be found at `http://github.com/srnas/ff`. The trajectories were obtained in the isothermal-isobaric ensemble ($T = 300$ K, $P = 1$ atm) with stochastic velocity rescaling [Bussi et al., 2007] and Berendsen barostat [Berendsen et al., 1984]. Long range electrostatics were treated using particle-mesh-Ewald summation [Darden et al., 1993]. The equations of motion were integrated with a 2 fs time step. All bond lengths were constrained using the LINCS algorithm [Hess et al., 1997]. Na^+ ions were added in the box in order to neutralize the charge, and additional Cl^- and Na^+ at a concentration of 0.1 M. AMBER-adapted parameters were used for Na^+ [Aaqvist, 1990] and Cl^- [Dang, 1995]. The adenine ligand bound to the *add* riboswitch was parametrized using the general Amber force field (gaff) [Case et al., 2004] and partial charges were assigned as discussed in reference [Di Palma et al., 2013]. The analyses of the hammerhead ribozyme and of the *add* riboswitch trajectories were performed after discarding the first 10 ns and 5 ns, respectively.

2.2 Details of the ENM

We investigated different ENM representations considering all possible combinations based on the use of one or more interaction centers representing the three chemical groups of each nucleotide: the sugar, the base and phosphate (in short S, B and P, respectively). Each group is represented by a specific atom, namely C1′ for the sugar, C2 for the base and P for the phosphate group. This selection follows from the customary coarse-graining choices previous adopted in various contexts [Hyeon and Thirumalai, 2005] including elastic networks [Delarue and Sanejouand, 2002; Yang et al., 2006; Setny and Zacharias, 2013; Zimmermann and Jernigan, 2014].

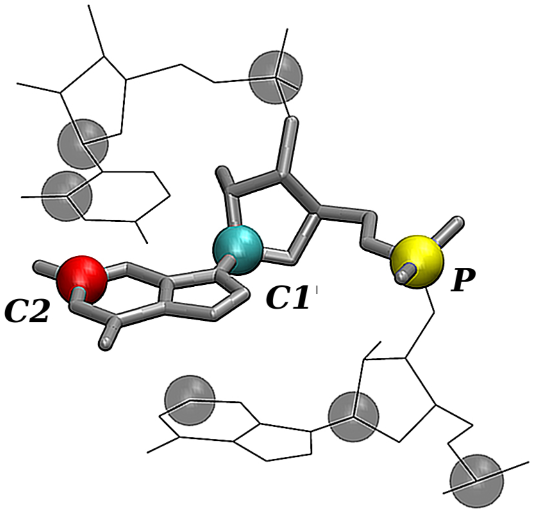


Figure 2.2: Schematic representation of the beads used to construct the ENM. The three atom used as centroids are the C2 carbon in the base (red), the C1′ carbon in the sugar ring (cyan) and the P atom in the phosphorous group (yellow).

The second crucial parameter in the ENM construction is the cutoff radius R_C. This should be chosen large enough so that the resulting Hessian matrix has only six null eigenvalues. Since the optimal value could depend on the precise choice of the beads, we systematically evaluated the agreement with MD trajectories for a large range of possible values of R_C. For each model the interaction cutoff distance, R_c, is varied in the $3-30$ Å range with 1 Å increments so as to assess the dependence of the predictions on the degree of connectivity of the elastic network.

For each RNA dataset entry, the reference structure for ENM calculations is set equal to the centroid structure of the associated MD trajectory. This is the conformer with the lowest average mean square distance from all MD-sampled structures after an optimal rigid structural alignment [Kabsch, 1976]. In the case of the *add* riboswitch, the adenine ligand atoms are included in the ENM calculation. In most practical applications, elastic models are built on the experimental structures. We test our calculations on each molecule considering both choices, and found no significant difference between the results.

The consistency of ENM and MD simulations was then assessed by computing both the correlation coefficient between the mean square fluctuations (MSF) of each bead and the root weighted square inner product (RWSIP) for the essential dynamical spaces.

Reference Models

The statistical significance of both the MSF correlation and the RWSIP is assessed by using two terms of reference. The first one is given by the degree of consistency of the MSF or RWSIP for first and second halves of the atomistic MD trajectories. This sets, in practice, an upper-limit for very significant correlations of the observables. The second one is the degree of consistency of the random elastic network (RNM) of Setny and Zacharias [2013] with the reference MD simulations. This is a fully-connected elastic network where where all pairs of beads interact harmonically though, for each pair, the spring constant is randomly picked from the $[0,1]$ uniform distribution. Because this null ENM does not encode properties of the target molecule in any meaningful way, it provides a practical lower bound for significant correlations between ENMs and MD simulations.

2.3 Results of the comparison

In this section we compare the fluctuations predicted by different ENMs with those obtained from MD. by means of the similarity measures described in Chapter 3. To keep the comparison as simple and transparent as possible, each measure was computed separately for the S, B and P interaction centers. For multi-center ENMs this required the calculation of the effective interaction matrix (Eq. 1.14). Using as a reference the experimental structure in place of the MD centroid introduces only minor differences in the results, see Fig. A.3. Each measure was then averaged over the four systems in Table 2.1 (see Fig. A.4 for non-averaged values). The results, shown in Fig. 2.3, are profiled as a function of the elastic network interaction cutoff distance, R_c. The smallest physically-viable value for R_c, that is the abscissa of the left-most point of the curves, is the minimum value ensuring that the ENM zero-energy modes exclusively correspond to the six roto-translational modes.

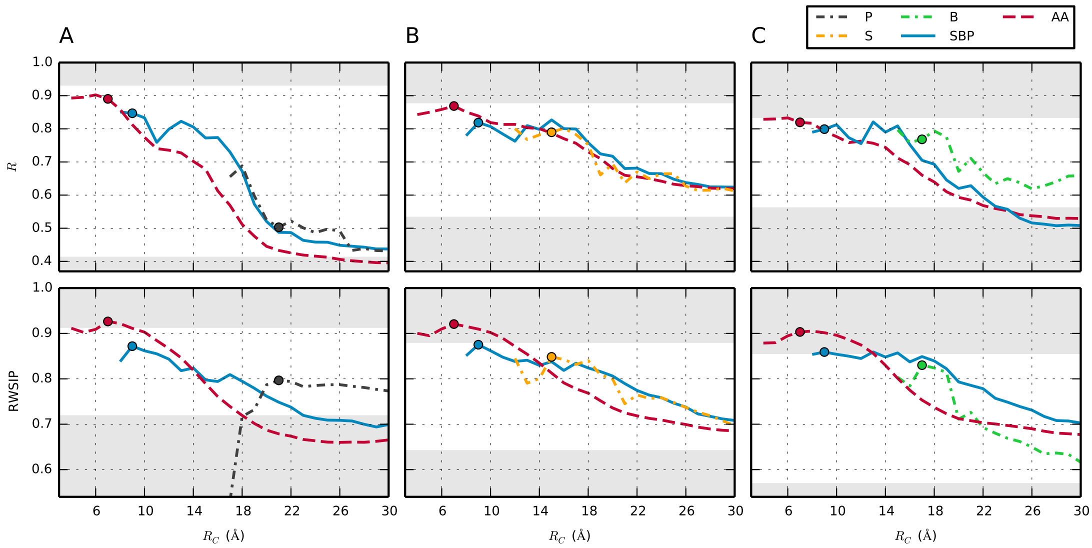


Figure 2.3: Agreement between MD simulations and ENM for different radii of cutoff. Correlation between MSF (upper panels), and RWSIP (lower panels). Values at the optimal cutoff values are represented by circles. A: phosphate beads; B: sugar beads; C: nucleobase beads. The gray regions correspond to values below the random-network model or above the MD self-agreement.

The main feature emerging from Fig. 2.3 is that, across the various models, the highest consistency with MD is attained when R_c is marginally larger than its smallest physically-viable value. It is also noted that the minimum value of R_c varies significantly across the models: for the AA model, which is the most detailed ENM, it is as low as 4 Å, while for the single-bead ones it is often larger than 10 Å. The MSF and RWSIP accord both decrease systematically as R_c is increased starting at the optimal value. This fact, which to our knowledge has not been reported before, can be rationalized *a posteriori* by considering that upon increasing R_c, one endows the network with harmonic couplings among nucleotides that are too far apart to be in direct physical interaction, and this brings about a degradation in model performance.

Furthermore, it is noted that the detailed, but also computationally more onerous, AA model is consistently in better accord with MD data than any of the coarse-grained ENMs. For this model, the degree of ENM-MD consistency is practically as high as the internal MD consistency at the optimal value $R_c \approx 7$ Å, or even higher in some cases. As a general trend, we notice that the accord between MD and ENMs decreases for coarser models (see also Fig. A.5 for models including two beads per nucleotide). Importantly, the AA and SBP models perform well not only on average but for each considered structure, whereas the performance of models with fewer interactions centers is less consistent across the repertoire of RNA molecules, see Fig. A.4. For all models, considering the optimal value of R_c both MSF and RWSIP accord are significantly higher than for the null model, indicating that all the ENMs are overall capable to capture the salient physical interactions of the system.

It is important to mention here that in the MD simulation of the duplex we observed a fraying event at time $\approx$ 670 ns (see Fig. 2.4), followed by a re-zipping into the native structure. o As a matter of fact, fraying events are expected at RNA termini on the μs time-scale covered by our simulations [Zgarbová et al., 2014]. In spite of the fact that these events are clearly out of the linear perturbation regime where one would expect ENM to properly predict fluctuations, the correlation between MD and ENM is reasonably high. By removing from the analysis the highly fluctuating terminal base pairs, the correlation is further improved (Fig. A.6).

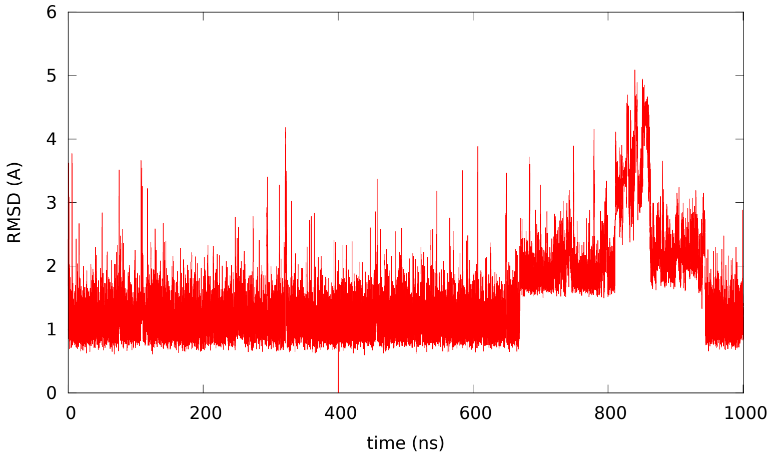


Figure 2.4: RMSD of the duplex along the 1 µs trajectory, computed with respect of the centroid frame. The increase in the RMSD visible at $t \simeq 670$ ns correspond to a base fraying event.

In Table 2.2 we summarize all the results for the optimal cut-off radius, determined as the radius that maximizes the RWSIP. The last column of the table reports the average number of neighbors of a bead, that is the number of other beads at distance smaller than R_c from it.

ENM	C1′	C2	P	others	best R_c (Å)	number of neighbors
P			✓		20	15.3
S	✓				15	9.9
B		✓			17	14.8
SP	✓		✓		19	30.4
BP		✓	✓		18	29.9
SB	✓	✓			11	15.4
SBP	✓	✓	✓		9	12.0
AA	✓	✓	✓	✓	7	52.9

Table 2.2: Summary of the tested ENMs. For each model, the adopted beads are marked. AA include all heavy atoms. Values of the cutoff radius (R_c) that maximize the RWSIP and average number of neighbors are also shown.

2.4 Effect of ion concentration

Ribonucleic acids are charged molecules, and electrostatic interactions play a major role, both in their structural stability and in their functional and dynamical behavior. It is an open question how and if elastic models are able to capture such a feature [Zimmermann and Jernigan, 2014]. The importance of long-range interactions can be a key difference between ribonucleic acids and proteins, and this should be, in principle, taken into account when developing an ENM suitable for RNA molecules. With this goal in mind, we conducted MD simulations at different Na^+/Cl^- concentrations and with different ion parameterizations. We then performed PCA on each trajectory, considering only the motions of atoms that we used as beads in the coarse-grained ENMs, namely C2, C1′ and P for each nucleotide. The resulting eigenspaces are then compared by mean of the RWSIP. The results of this study are summarized in Tab. 2.3. We notice that in our simulations with standard AMBER ions we did not observe any ion-crystallization event [Auffinger et al., 2007]. For a maximum robustness we tested the alternative ion parametrization by Joung and Cheatham III [2008], obtaining very similar results.

Molecule	0.0 M	0.1 M	0.5 M	1.0 M
Duplex	0.938	0.998	0.991	0.990
SRD	0.983	0.983	0.982	0.993

Table 2.3: RWSIP between 100 ns trajectories at different NaCl concentrations and a 500 ns trajectory at 0.1 M. Except for one case, all the values of RWSIP are comparable with the value obtained comparing the first 100 ns with the rest of the 0.1 M trajectory.

From Tab. 2.3 can see that there is no significant difference between the principal components of motion for systems simulated at different ion concentrations. This finding is in agreement the study of Virtanen et al. [2014], and suggests that the dynamic of RNA molecules is not sensitive to changes in concentrations, at least for what concern the time-scale spanned by our simulations. This result justifies the use of ENM with no explicit dependence on the ionic strength. It is however important to note that our test was limited to monovalent cations. The treatment of divalent cations is known to be very challenging because of force-field limitations and sampling difficulties.

2.5 Discussion

ENMs are simple but powerful models that enable to study and characterize the global dynamics of macromolecular complexes. In this chapter we applied ENM to RNA molecules with different, non-trivial structural elements such as GU Wobble pairs, non-canonical base pairs, bulges, junctions and pseudo knots. Two criteria in the choice of the molecules to treat, have been a) a limited size in order to be able to efficiently produce statistically significant MD trajectories and b) a stable tertiary structure, so that the Gaussian approximation of the ENM can be, at least in principle, applied. The accuracy of different ENMs was tested with respect to the agreement with atomistic MD simulations in explicit solvent.

The results presented in this work show that, in general, the fluctuations and the normal modes predicted by ENMs for nucleic acid systems are consistent with the PC computed by atomistic MD simulations. This is in agreement and confirms the general results of previous studies, that compared ENMs with experimentally determined ensembles [Setny and Zacharias, 2013; Zimmermann and Jernigan, 2014]. In this work we probed the nature of this agreement, systematically comparing different models, with different level of coarse-graining.

Among the models with one bead in each nucleotides we found that the best candidate is a model with a bead in the sugar ring (the C1′ atom, in our case). This fact was already pointed out in Setny and Zacharias [2013], where the authors performed a systematic search for the best bead position among all the atoms in the backbone. We complemented those findings by analyzing a wider range of RNA motifs and enlarging the search for the optimal representative atom to the nucleobase.

We note that the model with a single bead on the C2 atom of the base (B model) reproduces structural fluctuations less accurately than the S model and the optimal interaction cutoff is more dependent on the specific molecule. This can affect the general transferability of the model to different RNA molecules.

These shortcomings are even more evident for the P model, which present the worse performance when compared to other models. In Fig. 2.3 we can observe that the average performance of this model at its optimal cutoff value is significantly lower that the ones reached by different, more accurate models. Moreover, the cutoff radius that maximizes the RWSIP is different from the one that gives the best agreement for the fluctuations. As we can see from Fig. A.4, on the structure 1EKA the model seems to fail to predict meaningful fluctuations for values of R_C greater than 18 Å.

Moving on to two-beads models, we observe that ENMs employing beads both in the bases and in the backbone (SB, BP) perform systematically better than any single-bead model with only a modest increase in the computational complexity. SB and BP models also outperforms the SP model. We also stress that being able to reproduce the fluctuations of the bases is by itself an advantage because their functional role is of primary importance in nucleic acids and their dynamics can affect different aspects of the behavior of RNA molecules (see, e.g., Refs. [Colizzi and Bussi, 2012; Zgarbová et al., 2014; Gendron et al., 2001])

Increasing the number of beads featured in the ENM models (see also Fig. A.1 for 5/6-beads model), improves the agreement with MD, consistently with what had been observed for proteins [Fiorucci and Zacharias, 2010]. The AA-model showed to be the best ENM among the ones we tested. Fig. 2.3 shows this model reach an high level of accuracy, often comparable with the MD simulation self-agreement. This finding is in agreement with the recent work of Zimmermann and Jernigan [2014]. We focused our attention on this model, as well as on the the SBP model, which seems to give a net increase in the overall performance with respect to the single-bead models, at the cost of a small increase in the computational complexity (Fig. 2.3). It is also worth noting that the optimal radius of interaction for the SBP model (9 Å) is close to the cutoff value usually employed for constructing ENMs on protein systems [Micheletti et al., 2004; Fuglebakk et al., 2013], a fact that can be convenient when considering RNA-protein complexes.

In conclusion, ENMs were here compared systematically with fully atomistic MD simulations. We found that, in spite of their simplistic nature, the three-center model (SBP) and AA elastic networks are capable of properly reproduce MD fluctuations. Of these two accurate ENMs, the three-center model (SBP) provides an ideal compromise between accuracy and computational complexity, given that retaining the full atomistic detail when modeling large structures, such as the ribosome and other macromolecular RNA/protein complexes, can be computationally extremely demanding.

Chapter 3

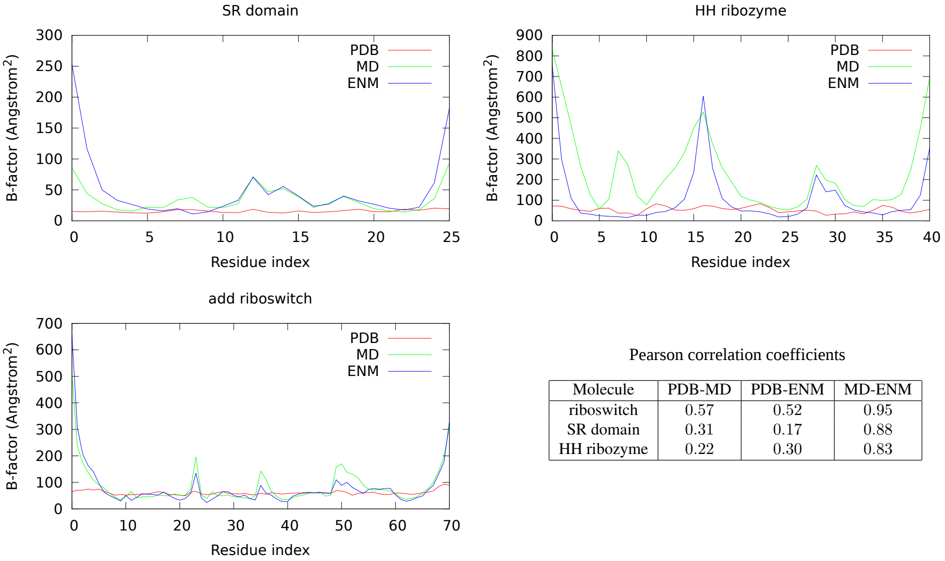
Comparing atomistic fluctuations with SHAPE reactivity

In this chapter, we propose and test a procedure to compare fluctuations with selective 2 -hydroxyl acylation analyzed by primer extension (SHAPE) experiments [Merino et al., 2005; Wilkinson et al., 2006; Weeks and Mauger, 2011].

SHAPE is a chemical probing technique that makes use of reagents that preferentially flexible regions of the ribonucleic acid (RNA) backbone. The reagents used for this experiments (two commonly employed examples are N-methylisotoic anhydride (NMIA) and 1-methyl-7-nitroisatoic anhydride (1M7)) react with the 2′-hydroxyl group forming a complex that inhibits the action of reverse transcriptase, so that a comparison of retrotranscribed DNA fragments enables a quantification of SHAPE reactivity at nucleotide level.

SHAPE reactivity is empirically known to correlate with base dynamics and sugar pucker flexibility at the nucleotide level [McGinnis et al., 2012]. For this reason it is, in principle, a good candidate for validating predictions of RNA internal dynamics. Recently, Kirmizialtin et al. [2015] have proposed a link between fluctuations of selected torsional angles and SHAPE reactivity and used SHAPE data as an input to improve the accuracy of force-field terms in an atomistic structure-based (Go-like) model. However, no other attempt of using SHAPE reactivity measurements to assess the predictive accuracy of three-dimensional coarse-grained models or atomistic molecular dynamics simulations, as been performed in the past.

We first set out to analyze the MD simulations so as to identify the local fluctuations that best correlates with SHAPE data, then we test the accuracy of elastic network models (ENMs) in reproducing such fluctuations, and we finally compare directly the ENM predictions with the SHAPE experimental results. A related comparison based on B-factor profiles, which are commonly used to validate ENM predictions (albeit with known limitations [Fuglebakk et al., 2013]), is provided in Fig. 3.1.



Pearson correlation coefficients

Molecule	PDB-MD	PDB-ENM	MD-ENM
riboswitch	0.57	0.52	0.95
SR domain	0.31	0.17	0.88
HH ribozyme	0.22	0.30	0.83

Figure 3.1: Comparison between the B-factors relative to the C1′ atoms, predicted by ENM, computed from MD simulations and extracted from the PDB files. The correlation between the experimental B-factors and the values predicted by ENM and MD are significantly lower than the correlation between ENM and MD. This can be explained considering that the experimental B-factors include many effects, such as crystal contacts and lattice defects, and thus are not guaranteed to provide a reliable account of the local amplitude of motion for a molecule in solution [Fuglebakk et al., 2013].

The main content of this chapter has been published in Pinamonti et al. [2015].

3.1 Comparison between MD fluctuations and SHAPE

SHAPE experiments have been proved to be able to capture the flexibility of an RNA chain at nucleotide resolution [Merino et al., 2005]. However, there is no standard way to obtain SHAPE reactivity directly from an MD simulations or from an ENM. We here considered several possible proxies for SHAPE reactivity, namely: i) the variance of the distance between selected pairs of beads and ii) the variance of the angle between selected triplets of beads. The latter approach was inspired by a similar analysis performed in Ref. [Soukup and Breaker, 1999] to predict in line probing experiments.

We first compared MD simulation of the *add* riboswitch described in Chapter 3 with SHAPE data reported in Ref. [Hajdin et al., 2013]. The fluctuations of these quantities were computed for each nucleotide, using PLUMED [Tribello et al., 2014]. Afterwards, we computed the correlation coefficient with the corresponding SHAPE reactivities. Figure 3.2 shows the correlation of the resulting fluctuations with SHAPE data. In order to avoid possible bias due to the correlation definition we computed both the Pearson correlation coefficient and the Kendall rank correlation coefficient. As we can see from Figure 3.2, the quantity that better reproduces experimental data is the fluctuation of the distance between consecutive C2s ($R = 0.78$). This is remarkable, since the SHAPE reaction does not explicitly involve the nucleobases. These fluctuations are shown, as a function of the residue index, in Figure 3.5.

This result can be interpreted by considering that most of the structural constraints in RNA originates from base-base interactions, and fluctuations in base-base distance are required for backbone flexibility. The fluctuations of the angle O2′-P-O5′ instead showed a poor correlation with experimental SHAPE data ($R = 0.05$). We notice here that the value of this angle has been shown to correlate with RNA stability related to in-line attack [Soukup and Breaker, 1999], and its fluctuations were recently used in the SHAPE-FIT approach to optimize the parameters of a structure-based force-field using experimental SHAPE reactivities [Kirmizialtin et al., 2015].

We also observe that the fluctuations of the distance between consecutive C2 atoms could be correlated with ribose mobility, which in turn depends

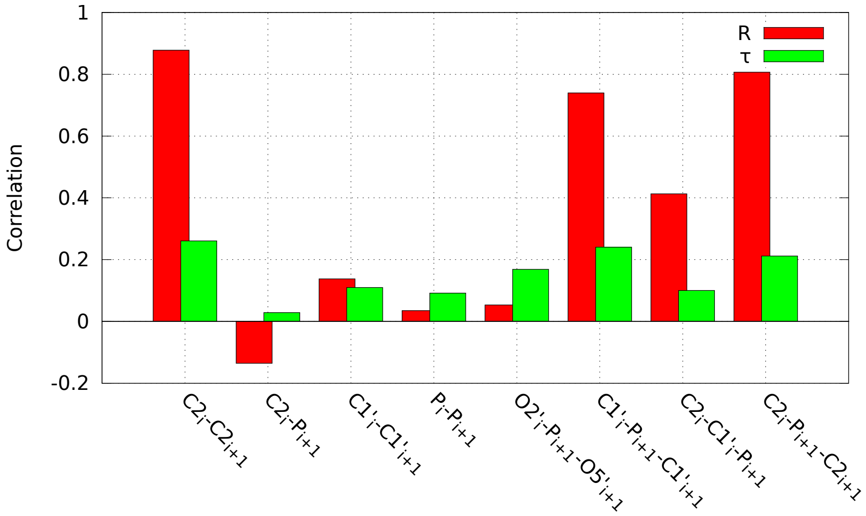


Figure 3.2: Comparison between SHAPE reactivities and distance/angle fluctuations, computed from an atomistic MD simulation. The SHAPE reactivities relative to the i-th residues are compared with the fluctuations between atoms belonging to both the considered residue and the one following in the chain. The two profiles are then compared by mean of the Pearson linear correlation coefficient (R) and the Kendall rank correlation coefficient (τ).

on sugar pucker [Altona and Sundaralingam, 1972, 1973]. Interestingly, C2′-endo conformations have been shown to be overrepresented among highly reactive residues in the ribosome [McGinnis et al., 2012]. An histogram of C2-C2 distances for selected sugar puckers is shown in Fig. 3.3, indicating that C2′-endo conformations correspond to a larger variability of the C2-C2 distance.

In conclusion, although the scope of the present SHAPE profiles comparison could be affected by the limited accuracy or precision of both experimental and MD-generated data, the obtained results suggest that a good structural determinant for SHAPE reactivity is arguably provided by base-base distance fluctuations.

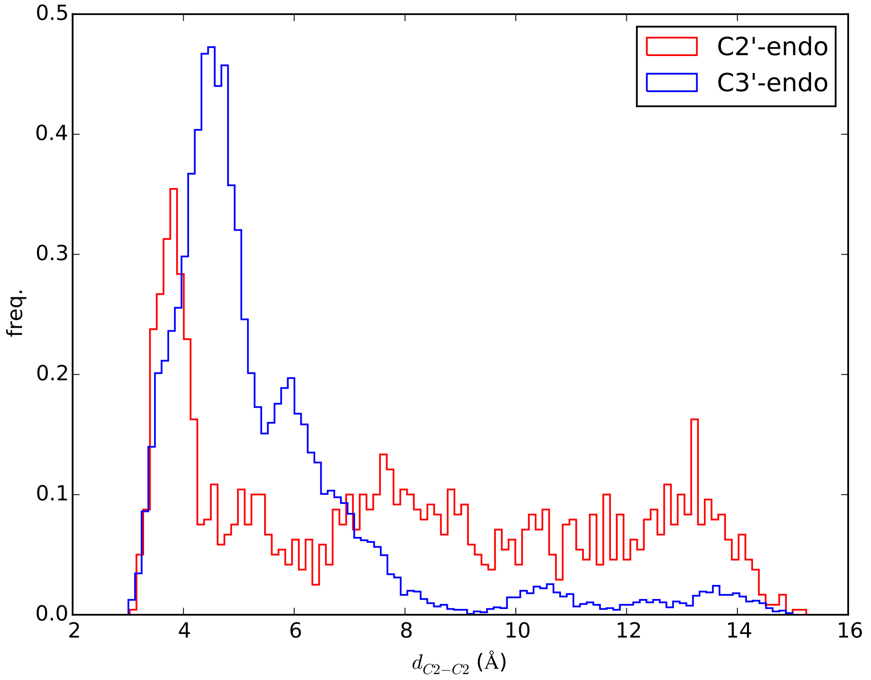


Figure 3.3: Distribution of the C2-C2 distances recorded during the MD simulation of the *add* riboswitch for residues in C2′-endo/ C3′ conformation. The pucker conformation was determined from the pseudorotation phase P, computed using the baRNAba analysis tool (`http://github.com/srnas/barnaba`). We identified C2′-endo conformation with values of P between 100 and 250 degrees.

3.2 | Fluctuations from ENM

In the ENM framework, the variance of the distance between two beads can be directly obtained from the covariance matrix in the linear perturbation regime as

$$\sigma^2_{d_{ij}} = \sum_{\alpha,\beta=1}^{3} \frac{\tilde{d}^{\alpha}_{ij}\tilde{d}^{\beta}_{ij}}{\tilde{d}^2}(C^{\alpha\beta}_{ii} + C^{\alpha\beta}_{jj} - C^{\alpha\beta}_{ij} - C^{\alpha\beta}_{ji}) \tag{3.1}$$

Where $\tilde{d}^{\mu}_{ij}$ is the μth cartesian coordinates of the reference distance between bead i and j.

Using eq. 3.1 we quantified to what extent ENMs are able to reproduce distance fluctuations at the nucleotide level. This test complements the assessment made using MSF and RWSIP, which mostly depends on the agreement of large scale motions and does not imply a good performance in the prediction of local fluctuations. This comparison is presented in Figure 3.4 where the ENM-MD Pearson correlation coefficients for each considered ENM are summarized, for each of the different ENMs and RNA molecules taken into exam in Chapter 3.

We remark here that the duplex (1EKA) is undergoing a base fraying, so that MD exhibits very large fluctuations at one terminus (see Fig. 2.4). The overall accord between MD and ENM is moderately good, although significantly worse than the accord with the large scale motions presented before. Overall, it is seen that the both the SBP model and the AA models provide the best agreement. We thus again focused our attention on these two models.

3.3 | Comparison between ENM and SHAPE

We compared the predictions of ENM with the SHAPE data for two different molecules. The *add* riboswitch, already considered in Chapter 3, and the *thiM* riboswitch (PDB code: 2GDI). The SHAPE data were taken from Hajdin et al. [2013]. As we can see from Fig. 3.5 the prediction of ENM are in qualitative agreement with the SHAPE data. In particular, high SHAPE reactivity

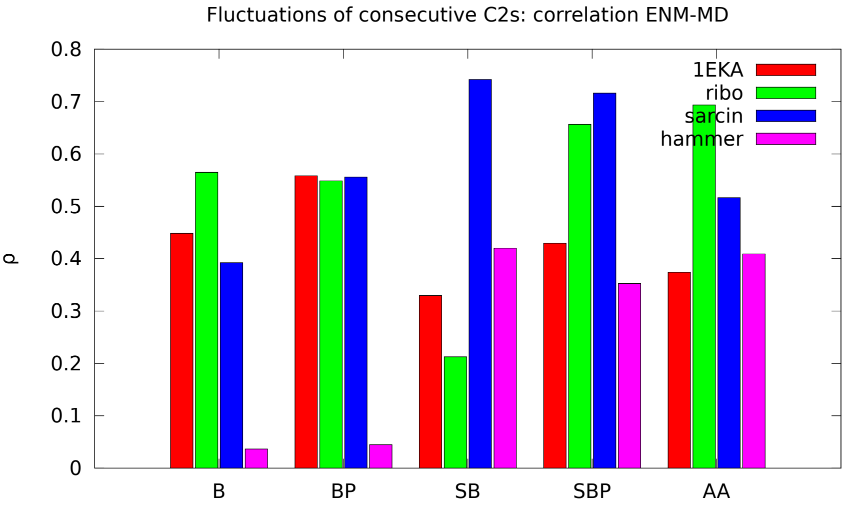


Figure 3.4: Pearson correlation coefficient R, computed between the fluctuations of the distance between consecutive C2s, from the MD simulation and from the different ENMs.

in the loop and junction regions correspond to highly fluctuating beads, both for the *add* and *thiM* riboswitch. We notice that this agreement goes beyond the mere identification of the residues involved in Watson-Crick or wobble pairings [Hajdin et al., 2013], as there appear several unpaired bases with a low SHAPE reactivity. This feature seems to be often correctly reproduced by the C2-C2 fluctuations profile. By visual inspection, it can be seen that non-reactive, non-paired bases often engage non- Watson-Crick base pairs as well as stacking interactions, as shown in Fig. 3.6. The Pearson correlation coefficients are summarized in Table 3.1. In this case too, it is found that the AA ENM performs better than the SBP ENM which, nevertheless, is much less demanding computationally because of its simpler formulation. We notice that

Molecule	SBP	AA	MD
add	0.64	0.76	0.88
thiM	0.37	0.59	-

Table 3.1: Pearson correlation coefficients between C2-C2 fluctuations predicted by ENM/MD and SHAPE reactivities.

using the Pearson coefficient as a measure of similarity we assume that the

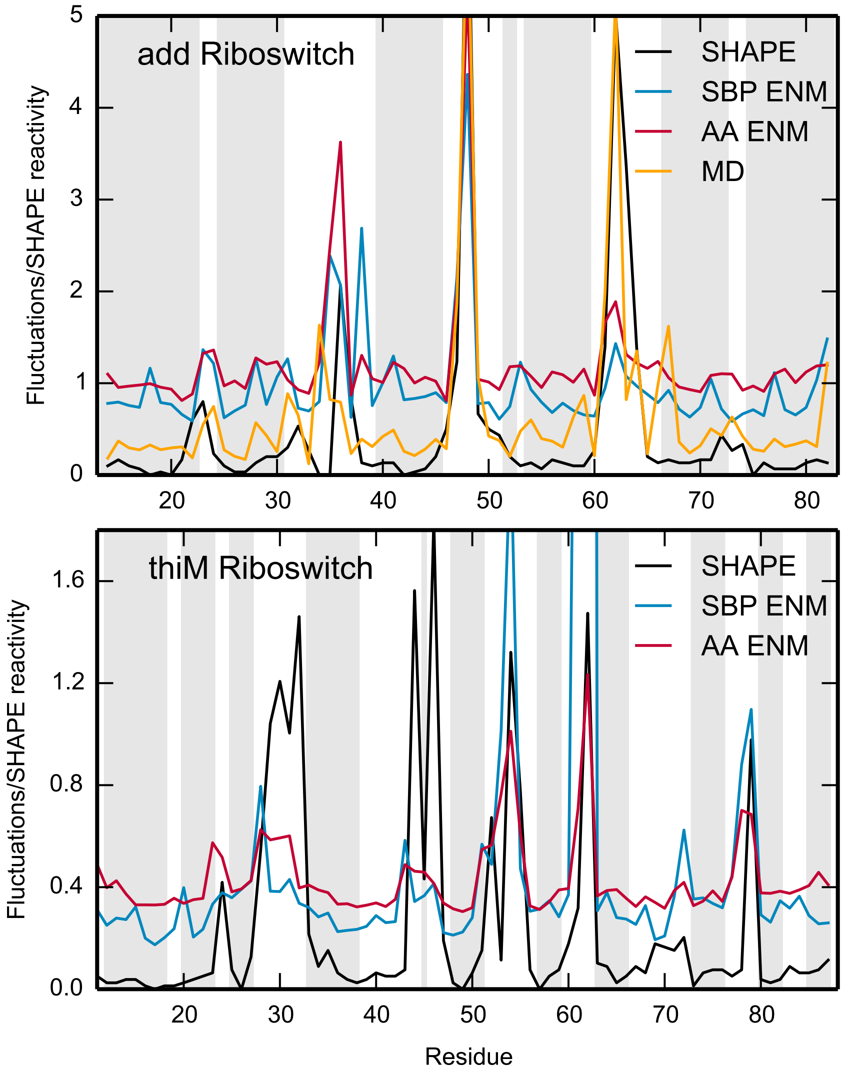


Figure 3.5: Comparison of the flexibility of the *add* riboswitch (upper panel) and the *thiM* riboswitch (lower panel), computed from the fluctuations of C2-C2 distanced from ENM (blue) and measured in SHAPE experiment (red). The fluctuations C2-C2 computed from the MD simulation are shown for the *add* riboswitch (green line, upper panel).

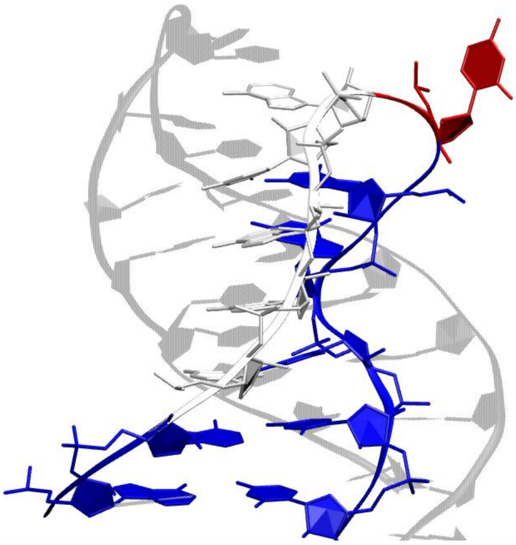


Figure 3.6: Example of nucleotides that do not form pairing interaction and have either low or high SHAPE reactivity. Residues 31 to 35 from *add* riboswitch, enlightened in white, have low SHAPE reactivity (< 0.5); they do not form Watson-Crick or Wobble pairs, but they are stabilized by stacking interactions. Residue 36, in red, presents an high reactivity ($\simeq 2.07$).

C2-C2 fluctuations are linearly correlated with SHAPE reactivity.

3.4 Conclusion

The results presented in the previous sections of this chapter shown that fluctuations obtained from MD and ENM fluctuations can be compared with experimental SHAPE data. SHAPE is a relatively new technique and a standard way to connect SHAPE reactivities with molecular fluctuations has not been reported yet.

As a first step of the analysis we considered various observables computed from atomistic MD simulations against SHAPE data, and established that the relative fluctuations of consecutive nucleobases provide a viable proxy for SHAPE data. Our comparative analysis showed that such fluctuations can be captured well using the SBP ENM, and to an even better extent with the AA ENM. Possibly, this is a step in the direction of defining a model able to directly correlate three- dimensional structures with SHAPE reactivities. This task is notoriously challenging, partly due to the difficulties of identifying from a priori considerations structural or dynamical observables that correlate significantly with SHAPE data.

Interestingly, both the ENMs are completely independent from the dihedral potentials and thus should not be directly affected by the pucker conformation of the ribose. The fact that they can provide a reasonable estimate of the backbone flexibility as measured by SHAPE reactivity suggests that the backbone flexibility is mostly hindered by the mobility of the bases.

Chapter 4

Entropy from ENM: a test on the *add* riboswitch

Entropic contributions to the free energy are fundamental in many biomolecular processes such as folding, binding, and conformational changes. The task of estimating the entropy from a molecular dynamic (MD) simulation is far from trivial since it involves sampling a large conformational space. Standard methods for free-energy calculation requires to collect MD samples along a sequence of point connecting the initial and final states of interest. This requires a large amount of computation in order to sample several transitions.

An alternative approach is to estimate directly the entropy for a given state of a macromolecule, which can be summed to the enthalpy if the free energy is needed. Unfortunately this task is far from trivial. Many different methods have been developed to perform this estimation. A few example are methods based on the covariance matrix of atomic coordinates [Andricioaei and Karplus, 2001] or forces [Hensen et al., 2014], or on computing the probability distribution based on the density observed in the simulation [Hnizdo et al., 2007, 2008; Fogolari et al., 2015]. Elastic network models (ENMs) have been used to estimate the change in conformational entropy upon binding in protein complexes, in order to refine the predictions of scoring functions [Zimmermann et al., 2012; Zamuner, 2015].

Here we report our tests on the reliability of ENM-derived entropy on a complex ribonucleic acid (RNA) molecule, the *add* adenine riboswitch, already introduced in the previous chapters. We focus in particular on the entropic contribution to the change in free energy upon the opening of the so-called "kissing loop", a tertiary interaction between the P2 and the P3 stems. We performed four MD simulations constraining the distance between the centers of mass of the terminal loops of the two stems, in order to evaluate the free energy as a function of the distance between the loops. The results obtained can be compared with the estimates performed with umbrella sampling simulations [Di Palma et al., 2015].

4.1 Methods and results

We considered the *add* adenine riboswitch in its Apo and Holo form. In order to evaluate the free-energy dependence on the loop-loop distance we enforced a restraint in the distance between the centers of mass of the residues in the two kissing loops.

For each form we perform four different MD simulations, with different values of the distance d ($d = 12.5$, 20, 30, 34 nm). Each simulation was 40 ns long. The starting structures were taken from the published by Di Palma et al. [2015].

The average total enthalpy, U, as well as its statistical error, were computed using the "g_energy" routine of GROMACS. Fig. 4.1 reports the average U for each trajectory.

The entropy for an elastic network whose dynamics is described by the interaction matrix M, is given by:

$$S_{\text{ENM}} = -\frac{1}{2} k_B \log(\det(M)) + const. \tag{4.1}$$

We are focusing only on the terms in the definition of S_{ENM} that will change when the riboswitch changes is conformation. We recall that in the definition of M the elastic constant of the springs enters as an arbitrary factor, k. This leads to a term $-\frac{1}{2}k_B(3N-6)\log(k)$ in the entropy.Since the value of k is not expected to change considerably when considering different conformations of

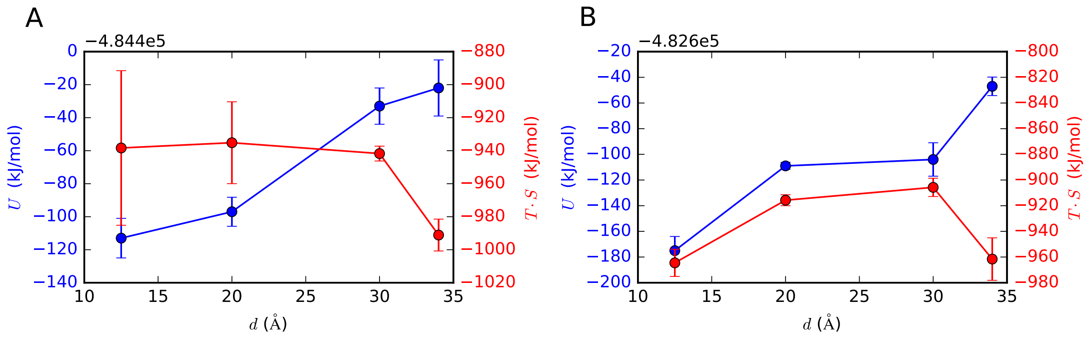


Figure 4.1: Total energy (blue) and entropy (red) computed from the MD simulations, for the Holo (panel A) and the Apo (panel B) forms.

the same molecule we ignored this term when computing entropy differences. Nevertheless this approximation may be a possible source of imprecision for the model.

The entropy was computed from each of the 40 ns MD trajectories considering a three-bead SBP elastic model. The ENM was constructed using, as reference structures, one frame every 100 ps (401 frames in total), and the entropy was computed using eq. 4.1 for each of this frames and taking the total average for a given loop-loop distance. The error on S was computed considering the standard deviation of the means of 10 different block (of 40 frames each). Fig. 4.1 reports the average S_{ENM} for each trajectory, as well as the corresponding standard error.

The free energy is computed from the enthalpy and the ENM-entropy, as $F = U - TS$, with $T = 300$ K. Results are shown in figure 4.2.

4.2 Discussion

Statistical errors reported in Fig. 4.2 are large. This is expected since the riboswitch considered is a complex system, and the calculation of average potential energy from a simulation is known to require extensive sampling in order to sample the phase space of all the molecules of both solute and solvent.

The entropy estimation could likely be improved considering a finer coarse

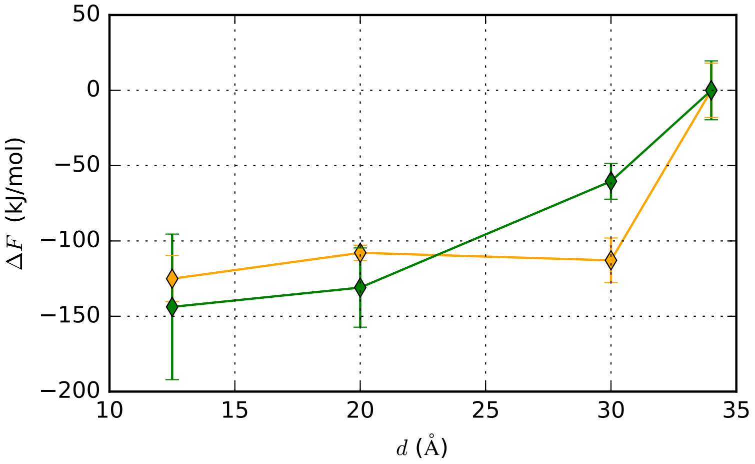


Figure 4.2: Free energy as a function of the loop-loop distance, as computed from the MD using the entropy estimated from the SBP-ENM, for the *add* riboswitch in its Apo (orange) and Holo (green) form.

graining, namely the AA-ENM model previously described, but the higher complexity of the model corresponds to a larger computational cost.

The contribution from water entropy is not taken into account by the ENM. This may be an major shortcoming considering that the stacking interactions are known to be heavily mediated by hydrophobic effects.

Comparison with umbrella sampling estimations

The free energy estimated here can be compared with the results obtained by Di Palma et al. [2015]. via umbrella sampling simulations. The values of ΔG between the open and closed kissing loops obtained here are in the same order of magnitude. Nevertheless there are consistent quantitative differences between our results and the umbrella sampling ones. This may be due to different factors, such as insufficient statistics, oversimplifications of the three-beads ENM, or the neglection of solvent contribution.

Part II

Markov State Models

Chapter 5

Basic theory of Markov state models

Here we briefly introduce the topic of Markov state models (MSMs) of molecular kinetics. MSMs are powerful tools that enable extraction of relevant kinetic information from multiple MD simulations. See Pande et al. [2010]; Noé and Fischer [2008]; Chodera and Noé [2014] for a brief introduction to this topic, or Bowman et al. [2013]; Prinz et al. [2011] for a more detailed discussion. Here we summarize the basic concepts which are relevant for the present work.

5.1 Estimation

The idea behind a MSM is to reduce the complexity of an MD simulation by dividing the phase space into discrete microstates (e.g. clustering the frames of the trajectory). Let's consider a trajectory of N configurations stored at a fixed time interval Δt:

$$X = \{\boldsymbol{x}(t=0), \boldsymbol{x}(t=\Delta t), ..., \boldsymbol{x}(t=(N-1)\Delta t)\} \tag{5.1}$$

Since each configuration can be assigned to a microstate, s_i, the trajectory can be transformed into a sequence of indexes $\{s_1, s_2, ..., s_N\}$. It is then possible to compute the transition matrix, $\boldsymbol{T}(\tau)$, whose elements, $T_{ij}(\tau)$, represent the probability that the system, starting from microstate s_i, will transition to microstate s_j, after a time τ:

$$T_{ij} = \mathrm{P}\left[\boldsymbol{x}(t+\tau) \in s_i | \boldsymbol{x}(t) \in s_j\right] \tag{5.2}$$

This matrix is sufficient to describe the dynamics of the system if the probabilities of transition are only determined by the actual state of the system, and do not depend on its previous history, i.e. the system is Markovian.

The Markovian assumption implies that the probabilities of being in microstate s_i propagates as

$$p_i(t+\tau) = \sum_{j=1}^{n} T_{ij}(\tau) p_j(t) \tag{5.3}$$

or in matrix form:

$$\boldsymbol{p}(t+\tau) = \boldsymbol{T}(\tau)\boldsymbol{p}(t) \tag{5.4}$$

We define the count matrix $\boldsymbol{S}$

$$S_{ij} = \text{“number of transitions from } s_j \text{ to } s_i \text{ observed from MD”} \tag{5.5}$$

The likelihood of having a certain transition matrix $\boldsymbol{T}$ given the observed transitions $\boldsymbol{S}$ is given by

$$\prod_{ij} T_{ij}^{S_{ij}} \tag{5.6}$$

By taking the logarithm and maximizing with respect to the elements T_{ij} it can be shown that the maximum likelihood estimator for the transition matrix $\boldsymbol{T}$ is given by

$$T_{ij} = \frac{C_{ij}}{\sum_i C_{ij}} \tag{5.7}$$

If our MD simulation satisfy detailed balance, also this estimated matrix $\boldsymbol{T}$ will satisfy it, in the limit of infinite statistics. However, if the amount of simulation data available is finite, the matrix $\boldsymbol{S}$ will be non symmetric and $\boldsymbol{T}$ will break detailed balance. In order to obtain a maximum likelihood estimator for a reversible transition matrix $\boldsymbol{T}$ one has to maximize Eq. 5.6 with the detailed balance constraint $T_{ij}\pi_j = T_{ji}\pi_i$. This problem can be solve iteratively, using a quadratic optimization method as explained in Prinz et al. [2011].

5.2 Analysis

Relevant information about the kinetics of the system can be extracted from the eigenspectrum of $\boldsymbol{T}$. Specifically, given the conditions of ergodicity and detailed balance the eigenvalues λ_i of $\boldsymbol{T}$ will be real and satisfy $1 = \lambda_0 > |\lambda_j|$ with $j > 1$ [Hohmann and Deuflhard, 2012]. The left eigenvector associated to λ_0 will correspond to the stationary distribution $\boldsymbol{\pi}$.

The matrix $\boldsymbol{T}$ will be non-symmetric. Assuming detailed balance we can define the symmetric matrix:

$$C_{ij} = \sqrt{\pi_i} T_{ij} \frac{1}{\sqrt{\pi_j}} \tag{5.8}$$

The left and right eigenvectors, $\boldsymbol{r}$ and $\boldsymbol{l}$, of the matrix $\boldsymbol{T}$, and the eigenvectors, $\boldsymbol{v}$ of $\boldsymbol{C}$ will be related by

$$\frac{1}{\sqrt{\pi_i}} l_i^\alpha = \sqrt{\pi_i} r_i^\alpha = v_i^\alpha \tag{5.9}$$

where i is the index of microstates, α the index of the eigenvectors. The vectors $\boldsymbol{v}$ can be chosen to be orthonormal so that $\boldsymbol{v}^\alpha \cdot \boldsymbol{v}^\beta = \delta_{\alpha\beta}$, which implies that the left and right eigenvectors of $\boldsymbol{T}$ are related by $\boldsymbol{l}^\alpha \cdot \boldsymbol{r}^\beta = \delta_{\alpha\beta}$.

By performing a spectral analysis of matrix $\boldsymbol{T}$ we can decompose the dynamics of the system into independent processes, each represented by the ith eigenvector of $\boldsymbol{T}$, for $i > 0$. The relaxation timescales of such processes can be computed from the eigenvalues λ_i of $\boldsymbol{T}$ as

$$t_i = -\frac{\tau}{\ln |\lambda_i|} \tag{5.10}$$

The eigenvectors associated with the eigenvalues smaller than 1 will be associated to different independent processes of the system. Each eigenvector represents a certain slow transition (or process) occurring (possibly multiple times) in the simulation. In particular, regions of the phase-space where the eigenvector has an opposite sign are the regions connected by the slow process.

Test the Markovianity

A simple test of the Markovianity of the system is given by the convergence of the implied timescales as a function of τ [Swope et al., 2004]. This can be easily explained given that exact Markovianity implies that

$$\boldsymbol{T}(k\tau) = \boldsymbol{T}(\tau)^k \tag{5.11}$$

which in turn means that $\lambda_i(k\tau) = \lambda_i(\tau)^k$. By inserting this in Eq. 5.10 we obtain

$$t_i(k\tau) = -\frac{k\tau}{\log(\lambda_i(k\tau))} = -\frac{\tau}{\log(\lambda_i(\tau))} = t_i(\tau). \tag{5.12}$$

In a real application the discretized dynamics is not exactly Markovian and thus MSMs constructed at different lag times will not satisfy this equation exactly. Nevertheless it is reasonable to expect that this equation should be approximately satisfy at least for the slowest implied timescales.

A different way to test the Markovianity of a MSM is to check the validity of Eq. 5.11 by examining the evolution of transition probabilities between given metastable sets of microstates. By comparing $P[A \to B, \boldsymbol{T}(k\tau)]$ with $P[A \to B, \boldsymbol{T}(\tau)^k]$, where A,B are two metastable regions of interest, one can asses if the prediction of the MSM built at lag-time τ can reproduce the probability of transitions observed at a larger value of lag-time $k\tau$. This is a more stringent test than the convergence of the implied timescales. This test is usually referred to as “Chapman-Kolmogorov test” [Prinz et al., 2011].

Systematic error induced by the discretization

The source of systematic error is the discretization of the phase space into a finite number of microstates. This step breaks the Markovianity of the system, and modeling the system as a Markov chain causes deviation from the true dynamics. Nevertheless it has been shown [Prinz et al., 2011] that this deviation can be reduced in two ways. Increasing the lag-time τ, and using a finer and finer discretization. It can be proved that, in the limit of infinite statistics, the discretization error can be made arbitrarily small by following any of those two steps. In practice, when dealing with real finite-length simulations, there are limitations in this sense. Namely, increasing the lag-time will decrease the time-sensitivity of the model, and increasing the number of microstates in order to construct a finer discretization will cause overfitting problems [McGibbon and Pande, 2015].

Statistical error

Estimating the statistical error associated with the predictions of a MSM is not a trivial task. In principle the uncertainty of any interesting physical quantity derived from a MSM directly derives from the statistical error in the estimation of the elements of the count matrix $\boldsymbol{S}$. However, the fact that many of the quantities of interest are highly non linear with respect to the elements S_{ij} makes it extremely challenging to give an unbiased and reliable estimation for this uncertainty. A practical and general way of solving this problem is to draw samples of random transition matrices from the posterior distribution of possible transition matrices given a fixed count matrix.This can be done efficiently using Markov chain Montecarlo (MCMC) sampling of transition matrix, as explained by Trendelkamp-Schroer et al. [2015]. The statistical uncertainty of any observable can then be directly computed as its standard deviation over the set of sampled transition matrices.

TICA

Time-lagged independent component analysis (TICA) [Molgedey and Schuster, 1994] is a technique used to reduce the dimensionality of the initial data set of MD trajectories before proceeding with the discretization of the phase space and the construction of a MSM [Pérez-Hernández et al., 2013; Schwantes and Pande, 2013].

The TICA method requires to compute the time-lagged covariance matrix

$$C_{ij}(\tau) = \langle r_i(t+\tau) r_j(t) \rangle_t \tag{5.13}$$

and the instantaneous covariance matrix

$$C_{ij}(0) = \langle r_i(t) r_j(t) \rangle \tag{5.14}$$

The next step is to solve the generalized eigenvalue problem

$$\boldsymbol{C}(\tau)\boldsymbol{U} = \boldsymbol{C}(0)\Lambda \tag{5.15}$$

where Λ is a diagonal matrix whose diagonal elements are the generalized eigenvalues, λ_α, and $\boldsymbol{U}$ is a matrix whose columns are the generalized eigenvectors, $\boldsymbol{u}^\alpha$. Finally, the initial coordinates can be projected on the slowest time-lagged independent components (TICs).

It has been shown that TICA is an optimal way of reducing the dimensionality of the input data prior to the construction of a MSM [Pérez-Hernández et al., 2013]. This dimensionality reduction can be further improved by using the kinetic map projection proposed by Noé and Clementi [2015], which consists in projecting the input data on the space defined by the rescaled eigenvectors $\tilde{\boldsymbol{u}}^\alpha = \lambda_\alpha \boldsymbol{u}^\alpha$.

Coarse-graining a MSM

In order to obtain a good MSM the number of microstates need to be large enough, so as to reduce the discretization error. This usually leads, for medium sized biomolecules, to MSM with 10^2-10^4 states. This makes visualization and intuitive analysis of the model hard.

There exist several methods that enable to overcome this problem by exploiting the kinetic information provided by an MSM to construct an even coarser representation of the system, lumping the MSM microstates into a few, metastable macrostates. The most standard approach is to use Perron-cluster cluster analysis (PCCA) [Schütte et al., 1999], a method that exploits the sign structure of the eigenvectors to define the optimal metastable partition of the MSM microstates. Today, more advanced versions of the method, PCCA+ [Deuflhard and Weber, 2005] and PCCA++ [Röblitz and Weber, 2013], can be used to assign to each microstate a probability of being a member of a certain metastable macrostate.

Another possibility is to reduce the complexity of the model by constructing a hidden Markov models (HMM) of the kinetics, as introduced by Noé et al. [2013]. The idea is that the system is modeled as a Markov chain between hidden metastable macrostates. These states are not directly observable but are measured looking at the microstate which at every step is extracted from a distribution probability that depends on the hidden state.

Chapter 6

Kinetic properties of RNA oligonucleotides

ribonucleic acid (RNA) stability depends on a large variety of interactions, including stacking, hydrogen bonding, and interactions with water and ions [Bloomfield et al., 2000]. In vacuum, stacking interactions arise from complex interactions between aromatic rings [Hobza and Šponer, 1999]. However, in biological environment these interactions are heavily mediated by water. Dinucleotides and short oligonucleotides are perfect models to study stacking in RNA. While the equilibrium properties have been extensively characterized by NMR measurements [Vokacova et al., 2009; Olsthoorn et al., 1982; Ezra et al., 1977; Lee et al., 1976; Lee and Tinoco, 1980; Lee, 1983; Condon et al., 2015; Yildirim et al., 2011; Tubbs et al., 2013], their kinetics have been only studied in a limited number of temperature-jump (T-jump) experiments [Pörschke, 1976, 1978; Dewey and Turner, 1979].

Molecular dynamics (MD) provides a tool that can be used to characterize in detail the kinetics of these systems, and provide insightful fine-detailed information that can complement experimental measurements.

In this chapter we present a systematic analysis of the kinetic processes for RNA oligonucleotides, as predicted by MD simulations. We used Markov state models (MSMs) and hidden Markov models (HMMs), to provide a complete description of the transitions characterized by the slowest relaxation times. We studied a number of dinucleoside monophosphates, a trinucleotide (AAA), and a tetranucleotide (AAAA) so as to characterize the dependence of kinetics on length and sequence. Results are compared with available experiments. Whereas some of the reported transitions correspond to known artifacts of the current force field, our results can explain the overall trends. Importantly, we suggest that measured autocorrelation times may not be directly associated to transitions between helix and coil structures but to transitions between kinetic traps characterized by different stacking patterns.

The results presented in this chapter are part of a paper currently in preparation [Pinamonti et al., 2016].

6.1 | Methods

Molecular dynamics simulations

MD simulations were run with different salt concentrations, ionic strength, sequence, and oligonucleotide length. The dinucleotides and trinucleotides simulations were performed using GROMACS 4.6.7 [Pronk et al., 2013]. The tetranucleotide simulation was run using AMBER [Case et al., 2014]. We used amber99 force-field parameters [Hornak et al., 2006] with parmbsc0 [Pérez et al., 2007] and χOL3 [Banáš et al., 2010] corrections. Simulations were run at different temperatures using the stochastic velocity rescaling thermostat [Bussi et al., 2007] and the Parrinello-Rahman barostat [Parrinello and Rahman, 1981]. RNA molecules were solvated in explicit water (TIP3P parameters [Jorgensen et al., 1983]), adding Na^+ counterions to neutralize the RNA charge, plus additional NaCl to reach the nominal concentration. Details of all simulations are reported in Tab. 6.1

Markov state models

In order to analyze the trajectories produced from the MD we considered the following set of coordinates:

Sequence	T (K)	Na$^+$ (M)	N. traj.	Total length (µs)	Stride (ps)	TICA lagtime (ns)	TICA dimensions	N. of microstates	MSM lagtime (ns)	Frac. of active states
CC	277	1.0	4	9.6	10	1.0	10	400	0.5	0.93
AC	277	1.0	4	9.7	10	1.0	10	400	0.5	0.94
CA	277	1.0	4	9.1	10	1.0	10	400	0.5	0.95
AA	277	1.0	4	8.9	10	1.0	10	400	0.5	0.99
CC	300	1.0	8	7.0	10	1.0	10	400	0.5	0.94
AC	300	1.0	8	7.0	10	1.0	10	400	0.5	0.93
CA	300	1.0	8	6.6	10	1.0	10	400	0.5	0.95
AA	300	1.0	16	7.0	10	1.0	10	400	0.5	1.0
AAA	300	0.1	17	57.0	100	5.0	19	100	5.0	1.0
AAAA	275	0.13	4	35	100	1.0	44	400	20.0	0.99

Table 6.1: Details of the MD simulations and of the MSM.

1. G-vectors (4D vectors connecting the nucleobases ring centers, as described by Bottaro et al. [2014])

2. Backbone dihedrals

3. Sugar ring torsional angles

4. Glycosidic torsional angles

The dimensionality of the input data was then reduced using time-lagged independent components analysis (TICA) as described in Pérez-Hernández et al. [2013]; Schwantes and Pande [2013]. Data were projected on the slowest TICs using a kinetic map projection [Noé and Clementi, 2015] and then discretized using a k-means clustering algorithm [MacQueen, 1967]. A lag-time τ was used to construct MSMs that approximate the dynamics of the discretized systems. Statistical uncertainties were estimated by means of the Markov chain Monte Carlo (MCMC) sampling of transition matrices from the posterior distribution described in Trendelkamp-Schroer et al. [2015]. All details and parameters used in the MSMs construction are reported in Tab. 6.1. The

MSM construction and analysis was performed using the software PyEMMA 2.2 [Scherer et al., 2015].

Combined discretization of dinucleotides trajectories

Since the dinucleotide systems share the same number of residues and the same backbone, the number of coordinates is the same for all of them. This can be exploited to perform TICA on a virtual trajectory obtained merging all the individual trajectories of the dinucleotides. We discretized the merged trajectories using k-mean clustering. For each dinucleotide system, we then built a separate MSM.

Analysis of the kinetics

From the eigenvectors and the eigenvalues of the transition matrix of an MSM we can obtain detailed information about the slow processes occurring during the simulations, as well as a precise estimation of their predicted timescales.

The eigenvectors of the different dinucleotides' MSMs were then compared using an appropriate measure of similarity. Since the active sets of different MSMs is different we first mapped all the eigenvectors, $\boldsymbol{\psi}$, to a common 400-dimensional space, defining

$$\tilde{\psi}_i = \{ \begin{matrix} \sqrt{\pi_i}\psi_i & \text{if } i \in A \\ 0 & \text{otherwise} \end{matrix} \tag{6.1}$$

Here, i is the index of the microstate, A is the set of active microstates, $\boldsymbol{\pi}$ is the stationary distribution of the MSM considered. We then compute the similarity between two eigenvectors, $\tilde{\boldsymbol{\psi}}^{\alpha}$, $\tilde{\boldsymbol{\psi}}^{\beta}$, from different MSMs as the square of their scalar product, $(\tilde{\boldsymbol{\psi}}^{\alpha} \cdot \tilde{\boldsymbol{\psi}}^{\beta})^2$. We also used kernel principal components analysis (KPCA) [Schölkopf et al., 1997] to project the first three eigenvectors of the eight dinucleotides' MSMs on a 2-D surface, in order to visually group similar processes from different MSMs. As kernel definition we used $\Phi(\tilde{\boldsymbol{\psi}}) = \tilde{\boldsymbol{\psi}} \otimes \tilde{\boldsymbol{\psi}}$, where $\otimes$ denotes the outer product. This is invariant for changes in sign of $\tilde{\boldsymbol{\psi}}$. This analysis was possible since the MSMs share a common set of microstates, given that the clustering was performed on the joint set of MD data of all dinucleotide systems.

As a further analysis of the dinucleotides' slow processes, we computed the correlations of these eigenvectors with all the dihedral angles of the dinucleotides. The variables with the highest correlation coefficient with a given eigenvector should be the best suited to describe the correspondent transition (as explained in Pérez-Hernández et al. [2013]). To avoid ambiguities due to the periodicity of dihedrals we compute the correlation between eigenvector $\boldsymbol{\psi}$ and torsion θ as $\max_{\eta}[\mathrm{corr}(\boldsymbol{\psi}, \cos(\theta+\eta))]$, that is shifting the angle by a phase that maximizes the correlation.

The major non-bonded interaction in short oligonucleotides is the stacking interaction between consecutive nucleobases. In order to study this we used the stacking definition proposed in Condon et al. [2015], that takes into account 1) the distance between the centers of mass of the two nucleobases, 2) the angle defined by the distance vector between the two centers of mass and the vector normal to the first base plane, 3) the angle between the two vectors normal to the two bases' planes. These quantities are combined in a score, s, that goes from -2 to $+2$. Nucleotides are considered stacked if $s > 1$, unstacked otherwise.

To further simplify the tri- and tetra-nucleotide models (and analyze their features) we used the kinetic information from the MSMs to lump the microstates into a few metastable macrostates. This was done using a hidden Markov model (HMM), as described in Noé et al. [2013]. The resulting metastable states were then analyzed by looking at the distributions of selected observables (dihedrals, distances between key atoms, G-vectors [Bottaro et al., 2014], and stacking score [Condon et al., 2015] between bases).

Comparison with relaxation experiments

MSM predictions can be compared with relaxation experiments that probe the kinetics of biomolecules. An exhaustive explanation of the theory behind this comparison is given by Noé et al. [2011]; Buchete and Hummer [2008]. Here we will briefly summarize the key concepts.

Consider a system described by a MSM with n microstates and transition matrix $\boldsymbol{T}$. In a typical relaxation experiment a perturbation of the thermodynamic state of the system (e.g. a change in temperature) results in the starting distribution, $\boldsymbol{\pi}_0$ becoming out of equilibrium. The system then relaxes to its new equilibrium distribution. The relaxation process is monitored by measuring the evolution of an observable A, which is a suitable function of the state of the system. The time-evolution of A during the relaxation process is given by

$$A(t) = A_{eq} + \sum_{i=2}^{n} \exp\left(-\frac{t}{t_i}\right)\gamma_i \tag{6.2}$$

Where A_{eq} is the value of A at the final equilibrium, and γ_i is the amplitude of the ith decay process, which in general depends both on the shape of $\boldsymbol{\pi}_0$ and on the nature of the observable A. The decay constant of the ith process, t_i, is given by the ith implied timescale of the transition matrix governing the system's dynamics.

Calculation of the amplitudes, γ_i, requires accurate knowledge of the initial state of the system. When this information is not available, the relaxation time can be approximated by the autocorrelation time of $A(t)$, which is given by

$$\tau_{corr}(A) = \sum_{i=2}^{n} t_i c_i \tag{6.3}$$

where the amplitudes, c_i, are closely related with the factors γ_i. See Noé et al. [2011] for a more detailed derivation.

6.2 Results

Dinucleotides CC, AC, CA, AA

We here report the kinetic analysis performed on all the dinucleotides. Trajectories for all the investigated dinucleotides were merged together and analyzed with a single TICA. The complex phase space of the different dinucleotides can then be conveniently projected on the 2-D surface defined by the first two TICs

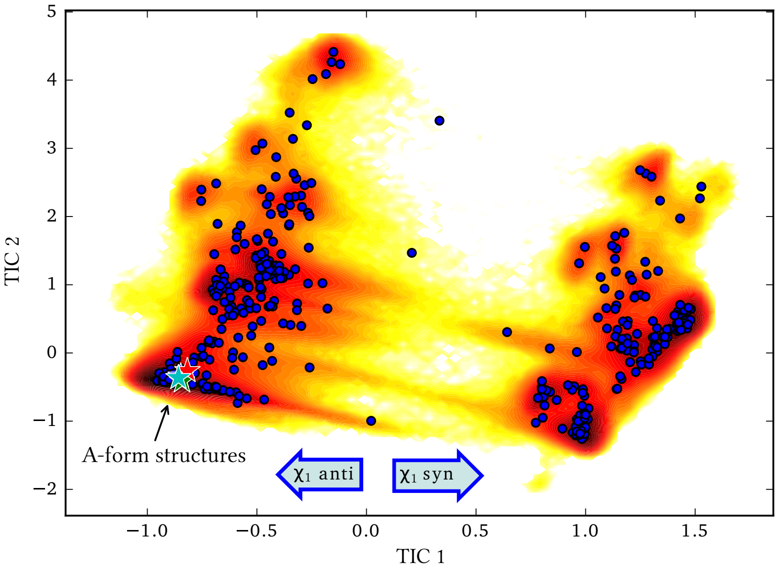


Figure 6.1: 2-D histogram of the joint MD data of the four dinucleotides, projected on the first two TICs; blue circles represent the centers of the microstates obtained from the k-means clustering. The native A-form structures are indicated with stars.

(see Fig. 6.1).

An initial analysis of the TICA components and the trajectories shows that the 1st TIC classifies the structures based on the value of the torsional angle χ_1 relative to the rotation of the glycosidic bond of the 5′ nucleobase (*anti*=negative values, *syn*=positive values). This suggests this isomerization is the slowest kinetic process in dinucleotides.

We then constructed a MSM for each of the investigated systems. The convergence of the MSMs was validated by monitoring the convergence of the implied timescales as a function of the lagtime (see Fig. B.1). In Tab. 6.2 we report the slowest timescales of the nine resulting MSMs at the chosen lag-time of 0.5 ns.

The four dinucleotides exhibit very different timescales. In particular, for $T = 277$ K the largest timescales for CC and CA are in the order of 200-300 ns, whereas for AA and AC it is around 40 ns. The situation is analogous at $T = 300$ K, but the timescales are shorter, in agreement with expectations for higher temperatures.

Fig. B.2 shows the first three eigenvectors for each of the dinucleotides′

Sequence	t_1 (ns)	t_2 (ns)	t_3 (ns)
$T = 277$ K			
CC	397 ± 48	28 ± 3	11 ± 2
AC	36 ± 18	31 ± 1	12 ± 1
CA	341 ± 16	86 ± 4	33 ± 8
AA	40 ± 6	33 ± 1	18 ± 0.4
$T = 300$ K			
CC	83 ± 4	9 ± 0.5	5 ± 0.1
AC	12 ± 0.2	6 ± 1.2	6 ± 0.4
CA	77 ± 4	16 ± 0.4	12 ± 1
AA	14 ± 0.2	11 ± 1	42 ± 0.8

Table 6.2: Implied timescales of the first three eigenvectors for the eight dinucleotides systems as predicted by the MSM built at $\tau = 0.5$ ns.

MSM, projected on the first two TICs. Since there are a large number of eigenvectors, it is convenient to exploit the fact that some of them share common features and define groups of similar processes occurring in different dinucleotides. In order to do this we evaluate the similarity of two eigenvectors using the square of their scalar product. A table summarizing the similarity between the eigenvectors relative to all systems is reported in Fig. B.3. The KPCA algorithm was then used to project them on a 2-D plane where we can easily identify clusters of similar processes (Fig. 6.2). Using the information from the 2-D projection shown in Fig. 6.2 and looking at the correlations between each eigenvector and the dihedrals angles (See Fig. 6.3), it is possible to identify five groups of eigenvectors that share similar features between them and are separate from the main group (labeled as A) by the KPCA.

1. *Group A* This group collects together all the eigenvectors that are not classified in other groups by the KPCA.

2. *Group B* These eigenvectors represent the flipping of the χ_1 torsion. This process is extremely slow (200-300 ns at $T = 277$) when this nucleobase is a cytosine (CC and CA), while it is much faster (< 20 ns) when the base is an adenine (AA and AC).

3. *Group C* These processes are related to the rotation of the dihedral χ_2.

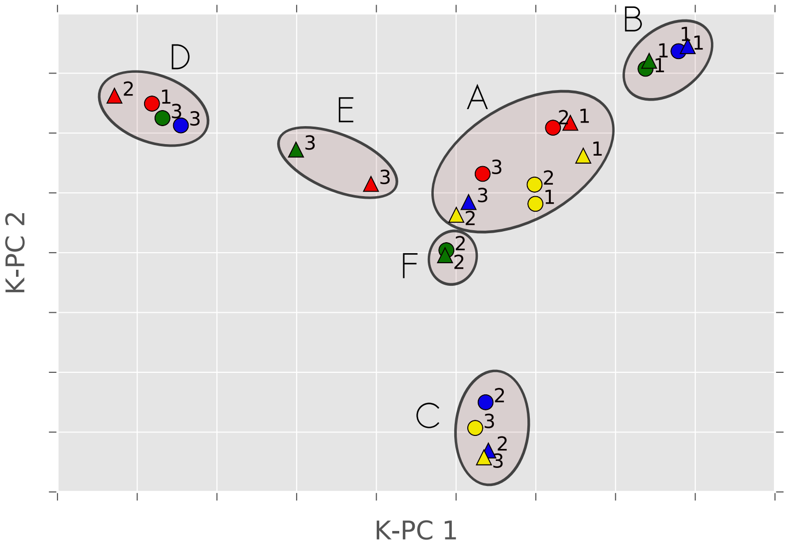


Figure 6.2: First three eigenvectors of each of the eight dinucleotides' MSMs, projected on the plane defined by the first two directions identified by kPCA. Numbers indicate eigenvectors' indexes. Colors indicate the sequence: CC (blue); AC (yellow); CA (green); AA (red). Shapes indicate the simulation conditions: $T = 277$ K (circle), $T = 300$ K (triangle).

Cytosines at the 3′ end show a much faster dynamics (~30-40 ns at $T = 277$) than those at the 5′ end.

4. *Group D, E, F* These processes are instead linked to the formation of specific structures. The conformation of the backbone in these structures is the same found in RNA Z-helices. They are in general characterized by γ_2 in *trans* conformation ($\gamma_2 > 150°$ or $\gamma_2 < -150°$), and a low distance (< 0.4 nm) between the O4′ atom on the sugar ring of the 5′-end nucleotide and the center of mass of the 3′ base [D'Ascenzo et al., 2016]. These three groups represent the formation of Z-motifs, that differ in the orientation of the χ_1,χ_2 glycosidic torsion or in the pathway of the process.

Trinucleotide AAA

We here report the MSM obtained for the AAA trinucleotide. The MSM was validated with the implied timescales test (see Fig. B.4). The MSM of the trin-

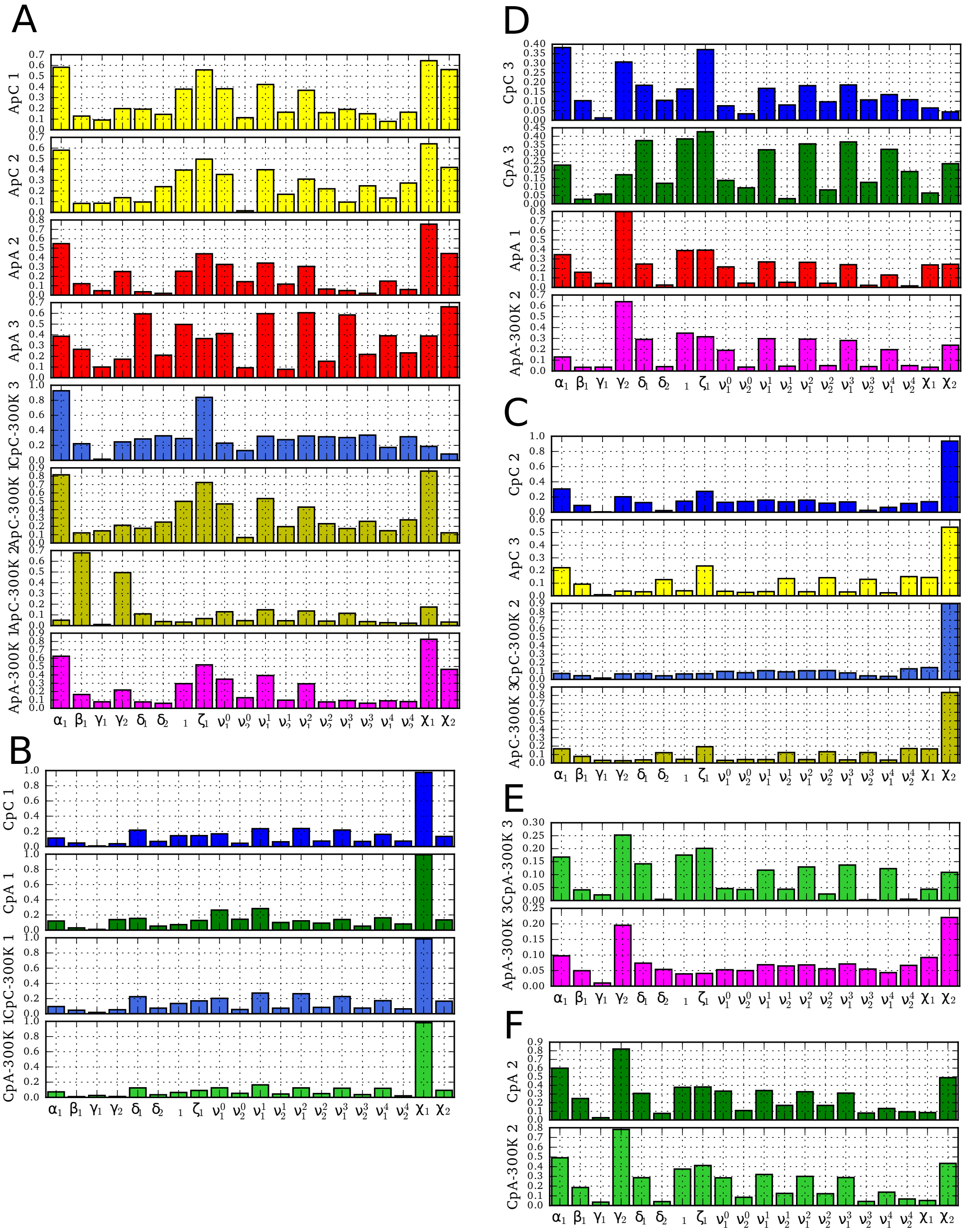


Figure 6.3: Correlation between the first three eigenvectors of the eight dinucleotide systems and each of the torsional angles. Angles have been shifted by a phase in order to maximize the correlation as define in the Methods section of main text.

	MSM	HMM
t_1	212 ± 8	295 ± 95
t_2	42 ± 2	41 ± 4
t_3	41 ± 2	39 ± 4
t_4	27 ± 2	-

Table 6.3: Implied timescales (in ns) associated to the slowest processes, for the MSM and the HMM of the adenine trinucleotide.

ucleotide AAA identified a very slow process ($t = 213 \pm 9$ ns). The fastest processes are dominated by two timescales around 40 ns (see Tab. 6.3). In order to gain further insight on the nature of the first three slow processes identified by the MSM, we coarse-grained the microstates space into 3 metastable sets, using an HMM. A schematic representation of the HMM is shown in Fig. 6.4. A first observation about the HMM is that state #2 corresponds to a particularly stable state. The transition in and out of this state has a very large timescale (200 to 300 ns). The equilibrium populations of the four states is reported in Fig. 6.4.

In order to understand the nature of these four states we analyzed the distribution of key observables (angles, distances, and stacking score) in the different HMM states, see Fig. B.5. From this analysis we discovered that state #2 corresponds to an intercalated structure, in which base A3 stacks between bases A1 and A2. State #3 corresponds to the native state, with a single A-form helix conformation, having all χ torsions in *anti* conformation. State #0 acts as an intermediate state, often visited by the system before transitioning to state #2. In state #1 the sequence of stacking interaction is analogous to state #3, while the main difference lies in the orientation of base A2, that in state #1 corresponds to a *syn* conformation of the torsion χ_2.

We also noticed that a significant fraction of the structures corresponding to state #0 present an A1-A3 stacking. This state is nevertheless well separated kinetically from #2. In fact, while in state #2 the stacking of A2-A3 occurs simultaneously with the stacking of A1-A3, in the intermediate state #0 the two stacking interactions are never formed together.

We also observed a recurrent hydrogen bond forming between the non-bridging oxygen of the phosphate group of base 2 and the 5′ hydrogen of

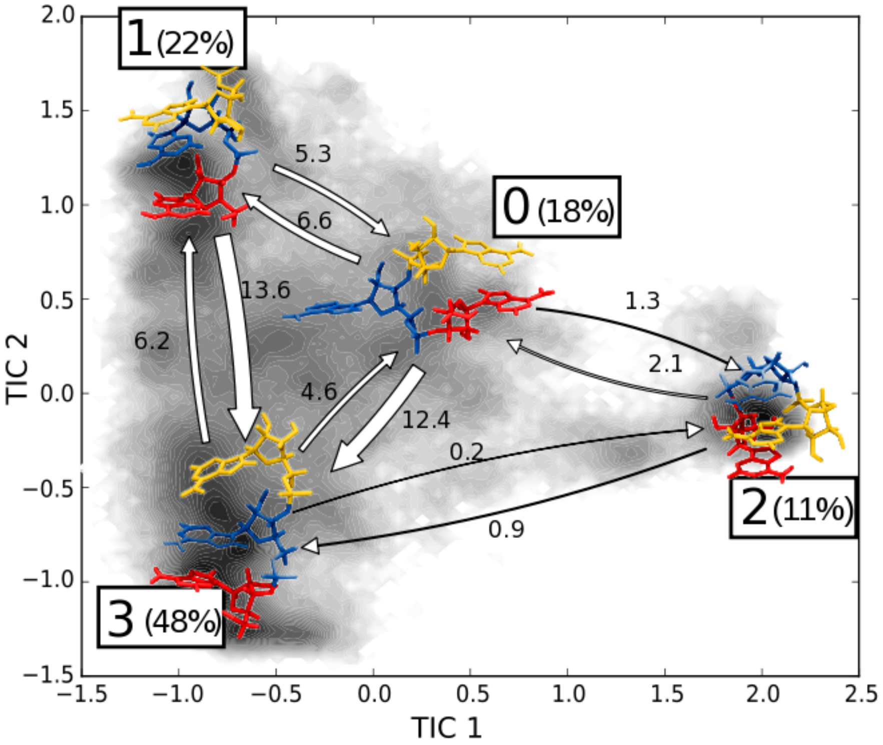


Figure 6.4: Schematic representation of the 4-state HMM of AAA. A1 (red), A2 (blue), A3 (yellow). Percentages indicate the equilibrium population of each state; the width of the arrows is proportional to the transition rate between the states which are also indicated in μs^{-1} units. Shading indicates the distribution of the simulation data on the TICA plane.

base 1, in the intercalated structures. Fig. B.5 shows the distribution of the distance between these two atoms in the four metastable states. The formation of this hydrogen bond is clearly a fundamental step in the formation of the intercalated structures.

The distribution of the dihedral angles, particularly the couple α_{i+1},ζ_i, is also informative, as shown in Fig. B.5:

1. states #1 and #3 are characterized by α_2 and $\zeta_1 < 0$
2. state #2 is characterized by α_2 and $\zeta_1 > 0$
3. state #0 has α_2, ζ_1, α_3, and $\zeta_2 > 0$

State #0 and #2 are also distinguished by the value of the angle χ_1 (*syn* in #0, *high-anti* in #2). The distributions of the three γ dihedrals do not vary significantly between the four metastable states, and we can exclude the presence of kinetically stable Z-motifs, in contrast with what was observed for the AA dinucleotide.

Tetranucleotide AAAA

The analysis of the MSM of the AAAA tetranucleotide follows a scheme similar to that described for the trinucleotide. The complexity and the number of available conformations grow exponentially with the number of bases in an oligonucleotide. For this reason it is particularly challenging to sample all the relevant conformational space for a tetranucleotide using only plain MD [Bergonzo et al., 2015; Gil-Ley et al., 2016]. In fact, even if our simulations have lengths of several microseconds, many transitions are observed only once. This reflects on the quality of the MSM, as it can be seen from the implied timescales plot (see Fig. B.6), and leads to extremely large statistical uncertainties.

Nevertheless it is possible to qualitatively compare the predictions of the MSM for AAAA with those described above for shorter oligonucleotides.

The first two implied timescales exhibited by the system (see Tab. 6.4) are in the microseconds range (3.1 ± 1.1 and 1.3 ± 0.6 µs), and are associated with the formation of two different intercalated structures, analogous to the ones

	MSM	HMM
t_1	3072 ± 1123	4685 ± 3646
t_2	1296 ± 567	781 ± 226
t_3	631 ± 371	364 ± 110

Table 6.4: Implied timescales (in ns) associated to the slowest processes, for the MSM and the HMM of the adenine tetranucleotide.

described for the trinucleotide. Again, to simplify the model we built an HMM, coarse-graining the MSM into 4 metastable macrostates. Fig. 6.5 shows a schematic representation of the HMM projected on the first two TICs. Also in this case the TICA identifies the formation of the intercalated structures (State #3) as the slowest process.

Two of the resulting states (#1 and #2) display a canonical stacking pattern whereas the other two states (#3 and #0) are characterized by the stacking of non-consecutive bases (see Fig. B.7). Specifically, state #2 contains the canonical A-form helix, whereas in state #1 base A3 is flipped to *syn* conformation. States #3 and #0 instead are distinguished by their stacking pattern, and they share the same features reported for trinucleotides, that is, α and ζ in $g+$ conformation and the presence of stabilizing hydrogen bonds with non-bridging oxygens. State #3 contains intercalated structures analogous to the one reported in previous works [Bergonzo et al., 2015; Gil-Ley et al., 2016; Condon et al., 2015], while state #0 presents A2 and A4 flipped out and stacked on each other. It is reasonable to expect other combinations of base orientation and stackings to arise when increasing the sampling.

Unfortunately the large errors in the HMM timescales make it difficult to discriminate quantitatively the different processes, and to clearly assign an implied timescale to each of them.

Comparison with Temperature-jump experiments

The timescales predicted by our MSMs can be compared with relaxation times measured using T-jump experiments in Pörschke [1978]. A proper comparison should follow the procedure explained by Noé et al. [2011], where the relaxation of an experimental observable can be decomposed in exponential

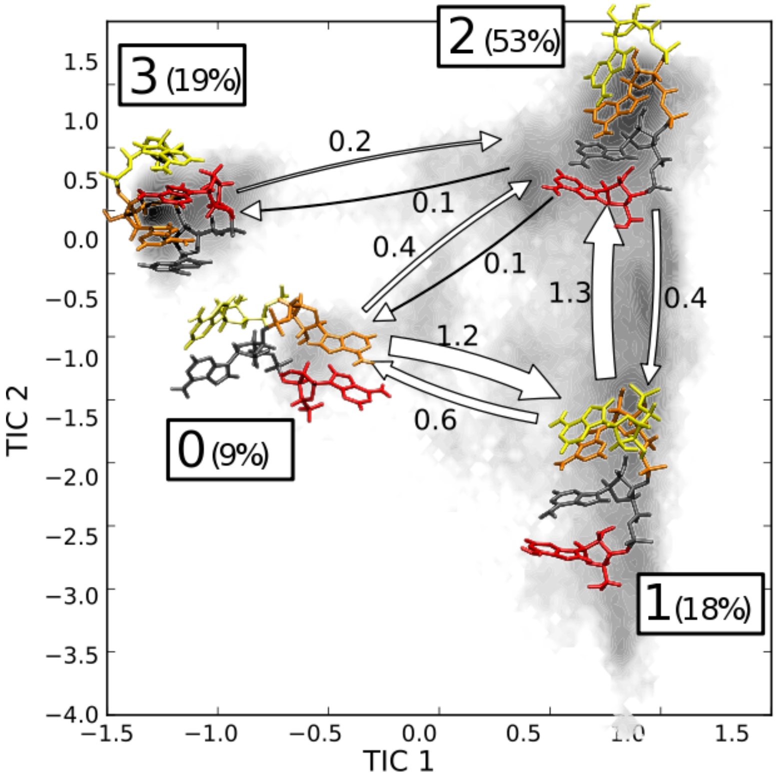


Figure 6.5: Schematic representation of the 4-state HMM of AAAA. A1 (red), A2 (blue), A3 (orange), A4 (yellow). Percentages indicates the equilibrium population of each state; the width of the arrows is proportional to the transition rate between the states which are also indicated in μs^{-1} units. Shading indicates the distribution of the simulation data on the TICA plane.

contributions coming from each MSM eigenvector. This requires knowledge of the experimental observable. In Pörschke [1978] the relaxation is measured with UV absorption. We modeled this using the stacking score proposed by Condon et al. [2015]. To estimate the relaxation rate without the need of further assumptions on the equilibrium distribution of the systems prior to T-jump, we computed the autocorrelation time of the stacking score.

The results of this calculation are reported in Tab. 6.5, along with the experimental relaxation times measured in Pörschke [1978]. For AAA and AAAA we report the autocorrelation time (Eq. 6.3) relative to the slowest stacking interaction (A1-A2 for AAA, A1-A3 for AAAA). We excluded from the calculations the contributions of the formation of intercalated/non-canonical structures, since the increased stability of such structures is a known limitation of the current force field [Bergonzo et al., 2015; Yildirim et al., 2011; Gil-Ley et al., 2016; Condon et al., 2015]. This was done using by Eq. 6.3 and setting to zero the amplitudes, γ_i, of the relative processes. We notice that the contributions of the slow modes of CC and CA to the stacking score kinetics are extremely small, since the autocorrelation time is almost ten times shorter than the associated timescale. The values of τ_{corr} predicted for CC, CA, AAA and AAAA are in good agreement with the experimental relaxation times. On the other hand, the values obtained for AA and AC are significantly shorter than the experimental values.

The T-jump relaxations at 297 K reported by Pörschke [1978] also included a long relaxation time of 600-900 ns for A_2, A_3, A_4, A_5, and A_{14} in 1 M Na^+ when the transition was probed with > 280 nm light. In the same study, relaxation times of 200 ± 20 ns and 700 ± 140 ns were reported for poly(A) in 0.2 M Na^+. These experiments were conducted with a cable discharge temperature jump apparatus where up to 200 kV/cm is transiently applied to the sample. An independent study using a laser induced temperature jump of poly(A) in 0.2 M Na^+, $T = 298$ K, however, reported only a 270 ± 70 ns relaxation at 285 nm [Dewey and Turner, 1979]. None of the MD simulations of A_2 and A_3 generated a timescale longer than 300 ns, It is therefore possible that the high electric field in the cable discharge experiments somehow affected the RNA, leading to the appearance of an artifactual relaxation process.

Molecule	τ (ns)	Exp. (ns) [Pörschke, 1978]
$T = 277$ K		
CC	24 ± 2	30 ± 6
AC	9 ± 1	42 ± 8
CA	25 ± 1	30 ± 6
AA	9.1 ± 0.2	50 ± 10
AAAA	$171^{\mathrm{a,b}} \pm 26$	-
$T = 300$ K		
CC	7.2 ± 0.4	-
AC	3.44 ± 0.03	-
CA	4.8 ± 0.1	-
AA	3.36 ± 0.04	$29^{\mathrm{c}} \pm 6$
AAA	25.8 ± 0.4	$45^{\mathrm{c}} \pm 9$
AAAA	-	$270^{\mathrm{c,d}} \pm 27$ (20%)
	-	$47^{\mathrm{c,d}} \pm 5$ (80%)

Table 6.5: Autocorrelation times of the stacking score predicted by the different MSMs, compared with the experimental relaxation times (1 M Na$^+$). [a] Simulations performed at 275 K. [b] Value for A1-A3 a stack not observed by NMR [Condon et al., 2015]; values for A1-A2, A2-A3, and A3-A4 are 78, 81, and 92 ns, respectively. [c] Experiments performed at 297 K. [d] Pörschke [1978] reports two relaxation times, with the relative amplitudes shown in brackets.

6.3 Discussion

What can we learn from this analysis about the kinetic properties of oligonucleotides?

The slow implied timescales observed for CC and CA are one order of magnitude longer than the experimental relaxation times (Tab. 6.2). These timescales are related to the transition from *anti* to *syn* of the cytosine at the 5′ end. An explanation for this inconsistency may be found in the inaccuracy of the force-field, which is a known limitation in the field of MD simulations of RNA. *Syn* cytosines are rare in non-catalytic RNAs [Sokoloski et al., 2011] and may be over represented in simulations.

The slowest timescales observed in AAA and AAAA are related to the formation of kinetically stable intercalated structures. These are in contrast both with the values for the relaxations times obtained in Pörschke [1978] and with NMR data for AAAA [Condon et al., 2015]. These two inconsistencies suggest that this metastable structure is an artifact of the simulation, likely caused by an imperfect parametrization of the force field. Analysis of the intercalated states revealed some structural details that seem to play a crucial role in stabilizing these structures. In particular these are 1) the formation of a hydrogen bond between the non-bridging oxygen of the phosphate group of one nucleotide and a hydroxyl from another nucleotide [Condon et al., 2015] , 2) the transition from negative to positive of the torsional angles α_{i+1},ζ_i (see Fig. B.7). We propose that this information must be kept in mind when trying to modify the parameters to improve force-fields accuracy. In particular it has been shown that tuning the parametrization of the dihedrals α and ζ significantly improves agreement with NOE data for several tetranucleotides [Gil-Ley et al., 2016].

Once the unphysical structures and transitions have been removed from the MSM, it is possible to use the remaining eigenvalues and eigenvectors to estimate the experimental relaxation times. It is important to recall that for an appropriate comparison with experimental data it is necessary to define an observable that is proportional to the measured intensity. We here used the stacking score defined by Condon et al. [2015]. The autocorrelation time of this score is reported in Tab. 6.5 and can be directly compared with experiments. This gives good agreement. Considering experimental error, the largest difference between autocorrelation times and experimental relaxation rates may be as small as 6-fold, corresponding to a difference of 1 kcal/mol in activation free energy at 300 K.

In general, the predicted relaxation time in all the considered systems is not determined by the rate of the helix ↔ coil transition. It is instead related to the rate of transitions between different structures, stabilized by stacking or other kinds of interactions. Examples of this are the Z-motifs in dinucleotides, or helices with flipped nucleotides in longer sequences. This suggests that the timescales obtained from relaxation experiments of oligonucleotides may be due to transitions between different "folded states", rather than between stacked, native structure and random coil. That is, these relaxation rates are dominated by the presence of various "kinetic traps", i.e. kinetically stable structures where the system may get stuck for a relatively long time before being able to reach its minimum free-energy conformation.

The work presented in this chapter demonstrates how MD simulations and MSMs can be used to provide deeper interpretation of experimental measurements of the kinetics of RNA folding. While most experimental methods report weighted averages for an ensemble of structures, MD follows transitions of a single molecule. If the simulations are converged, then the results from MSM analysis can be compared to experimental measurements. Although current RNA force fields do not accurately predict structural ensembles [Bergonzo et al., 2015; Condon et al., 2015], the results presented here suggest that the latest amber force field can reproduce the order of magnitude of T-jump relaxations [Dewey and Turner, 1979; Pörschke, 1978] measured for short oligonucleotides. Most current tests of force fields use structural data from NMR and x-ray diffraction as benchmarks. The results presented above indicate that comparison to experimental kinetic data can provide new benchmarks in the future. Moreover, when MD simulations accurately predict structures, they can generate more detailed interpretations of experiments and suggest new experiments to test hypotheses.

Chapter 7

Density-peak clustering applied to core-set MSMs

In this chapter we present an alternative approach to construct a Markov state model that describes the dynamics of a biomolecular system, starting from atomistic MD simulations. We make use of the unsupervised density peak (UDP) clustering algorithm, introduced by Rodriguez and Laio [2014] and further developed by d'Errico et al. [2016]. We combine this algorithm with time-lagged independent component analysis (TICA) [Molgedey and Schuster, 1994] in order to define the microstates of the system and we next define the transition probabilities between them using a "core-set approach" [Buchete and Hummer, 2008]. In the next sections we explain the basics concept of these methods, present their application to MSMs and finally test the applicability of this novel approach on several complex RNA molecules. In the next chapter we focus on the results obtained applying this method to the topic of RNA base fraying.

7.1 Core-set MSM

In the previous chapters we have reviewed the basics theory behind MSMs, and successfully applied this methodology to the study of RNA oligonucleotides. We learned that the key steps in the construction of a solid MSM involve the fine partition of the full data space into microstates. After this is done, the time evolution of the system is observed, and a transition is counted when the system, after time τ, has cross the border between two states.

An alternative approach to compute the transition probabilities between given microstates, called "transition-based-assignment" (TBA) or "coring", has been proposed by Buchete and Hummer [2008]. This method bears similarities to the concept of milestoning [Faradjian and Elber, 2004], which has in turn been developed and applied to the framework of rate-matrix-based MSM [Schütte et al., 2011] and transition-matrix-based MSMs [Sarich et al., 2013]. Here we will not discuss the differences between these alternatives nor compare their results, limiting our discussion to the method of Buchete and Hummer [2008].

The idea of the coring approach is to define a collections of "core sets", i.e. metastable regions of the phase space, which are not required to be in contact between them.

The important requirement is that each of this core regions is associated with a different metastable state of the system. This means that the system that just left a core region will return back to it more often than transitioning to another core. Under the assumption that the internal relaxation in these states is faster than the rate of transitions between them, the dynamics of the system can be approximated as a discrete Markov process between these states.

In order to properly estimate the transition probability between our states we follow the procedure originally proposed by Buchete and Hummer [2008]. Suppose that the system is in the core region of state A, C_A, at time t. We define a transition from state A to a second state B to occur only when the system reaches its core region C_B. Then the system will be considered in state B until it goes back to C_A, or reaches a third core region, independently of how many times it exits and re-enters in C_B before reaching a new state. This procedure is visually illustrated in Fig. 7.1.

The fundamental step of this approach is to start with a good definition of metastable core sets. This requirement is usually in contrast with the fact that, when studying the dynamics of a complex biomolecule, no prior knowledge of the free-energy landscape of the system is available. Therefore, in order

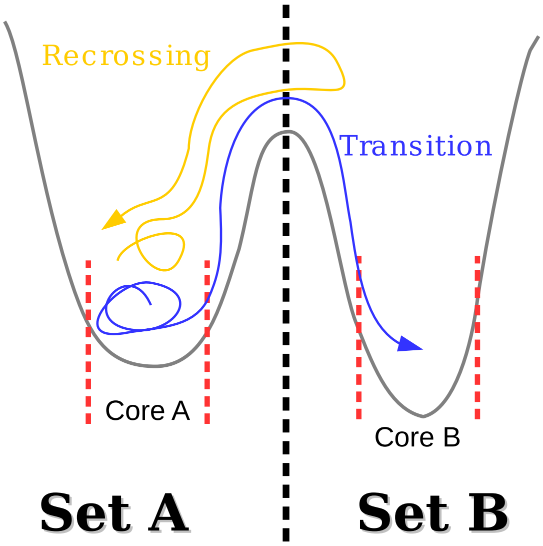


Figure 7.1: Graphical example of transitions in the core set approach, on a 1-D two-wells potential. The yellow trajectory exits from the core region A but does not reach the core region B, and thus is not counted as a transition A→B. The blue trajectory, instead, is considered as a proper transition since it reaches the core region B.

to successfully apply this method it is necessary to extract these information from the simulation data, preprocessing the trajectories in order to identify different states and define realistic core regions. A smart way to do this is to make use of a density-based clustering algorithm to separate the MD data set into a collection of clusters and identify the core regions of these clusters as the regions with higher density. In the following section we will introduce a recently published clustering algorithm that is optimally suited to address this problems.

7.2 Unsupervised density peak clustering algorithm

In this section we briefly report the methods published by d'Errico et al. [2016] and Rodriguez and Laio [2014].

An unbiased density estimator: LOCk-NN

Estimating local densities in complex data sets is a complex task, that often is a key step of data analysis in different scientific fields. Here we report the theory behind the adaptive and parameter-free density estimator LOCk-NN [d'Errico et al., 2016], a generalization of the k-NN algorithm [Silverman, 1986].

Let's consider a large data-set of points in a d-dimensional set. The density around a certain point i can be estimated with the k-NN estimator as

$$\rho_k = \frac{k}{r_k} \tag{7.1}$$

where k is a parameter that defines the number of nearest neighbors to considered, v_k is the volume of an hypersphere in d dimensions with radius r_k equal to the distance between point i and its kth nearest neighbor. The selection of parameter k is a critical step in the application of k-NN algorithm to complex data set, and can be a challenging task since many real-world data set may present highly-inhomogeneous densities.

A viable solution is to select a different value of k for each point i. In general the density will be constant in a neighborhood of a point i, while dishomogeneities will arise when considering points further and further apart. Therefore, the optimal k is the maximum value for which ρ is truly constant. This can be done by means of log-likelihood techniques by comparing the probability distribution of the volumes of the hyperspherical shells enclosed between successive neighbors of i with the exact distribution in the uniform case. The details of the algorithm are reported in d'Errico et al. [2016], where the LOCk-NN algorithm is shown to be able to accurately estimate both the density and the associated statistical uncertainty of points in complex highly-dimensional data sets.

Density peak clustering

After the density ρ_i and the associated statistical error ϵ_i have been computed for each point i in the data set of interest, using the LOCk-NN algorithm, this information can be used to partition the data set into separated clusters. This can be done with the density peak algorithm, that proceeds as follows.

The first step is to compute the following quantity:

$$g_i = \prod_{l \neq i} \int_{-\infty}^{+\infty} dx N_x(\rho_l, \epsilon^2_{\rho_l}) \int_{x}^{+\infty} N_y(\rho_i, \epsilon^2_{\rho_i}) \tag{7.2}$$

which is, the product of the probabilities that the density of point i is higher than the density of any other point l. $N_x(a, \sigma)$ represents the Gaussian distribution with average a and variance σ. The cluster centers are defined as the local maxima of g_i, i.e. points that have the highest probability of being surrounded by points with a lower density. The second step is to compute $\delta_i = \min_{j:g_j>g_i} r_{ij}$, which is the distance to the nearest point with higher g. A point is identified as a center only if $\delta_i > r_{\hat{k}_i}$, that is, a data point is a center only if all its $\hat{k}$ neighbors have a value of g lower than g_i. After all the centers have been identified all points are assigned, in order of decreasing g, to the same cluster of the nearest point with higher g. As final step of the clustering a merging is performed between clusters that are not separated in a statistically meaningful sense, i.e. if the density at the border between them is comparable, considering statistical uncertainty, with the peak density at their centers.

7.3 Comparison with standard MSMs

Methods

In order to test the viability of our approach that combines the core-set approach with UDP clustering, we analyzed MD simulations of four different RNA molecules, and compare the results with the ones of a standard full-state MSM. As RNA systems we considered the adenine dinucleotide and trinucleotides, already described in Chapter 7, together with the duplexes $\substack{\texttt{5'-ACGC}\\ \texttt{3'-UGCG}}$ and $\substack{\texttt{5'-UUGCG}\\ \texttt{3'-AACGC}}$. Details of the MD simulations of the adenine oligonucleotides are reported in Chapter 6. Since the main purpose of the MD simulation of the RNA duplexes was to study the dynamics of the fraying of the A-U terminal, we constrained the distances between the heavy atoms involved in the

Molecule	N. traj.	Total length (μs)	Stride (ps)	TICA lagtime (ns)	TICA dimensions	k-means centers	UDP clusters	N. of data subsets
AA	4	19.6	10	1.0	7	500	31	4
AAA	17	57.0	100	10.0	14	100	58	6
4-bp duplex	105	50.1	100	5.0	23	500	63	5
5-bp duplex	56	112.1	100	50.0	27	500	81	4

Table 7.1: Details of the MD simulations and of the MSM.

hydrogen bonds corresponding to the G-C pairs using an harmonic potential. We select for the present analysis only a subset of the MD trajectories of the 4-bp duplex in order to remove unphysical artifactual structures. More details about this selection will be given in the next chapter.

The RNA duplexes were solvated in a truncated dodecahedral box filled with TIP3P water molecules [Jorgensen et al., 1983]. Na^+ and Cl^- ions were added to the simulation box in order to neutralize the total charge and reach a concentration of 0.1 M. All MD simulations were performed using GROMACS 4.6.7 [Pronk et al., 2013] with the AMBER99 force field [Hornak et al., 2006] including parmbsc0 [Pérez et al., 2007] and χ_{OL3} [Banáš et al., 2010] corrections. GROMACS parameters can be found at `http://github.com/srnas/ff`. AMBER-adapted parameters were used for Na^+ [Aaqvist, 1990] and Cl^- [Dang, 1995], The trajectories were obtained in the isothermal-isobaric ensemble ($T = 300$ K, $P = 1$ atm) with stochastic velocity rescaling [Bussi et al., 2007] and Parrinello-Rahman barostat [Parrinello and Rahman, 1981]. Long range electrostatics were treated using particle-mesh-Ewald summation [Darden et al., 1993]. The equations of motion were integrated with a 2 fs time step. All bond lengths were constrained using the LINCS algorithm [Hess et al., 1997]. Further details about the simulations are reported in Tab. 7.1.

In order to analyze the trajectories produced from the MD we considered the following set of coordinates: 1) G-vectors, 2) backbone dihedrals, 3) sugar-ring torsional angles, 4) glycosidic torsional angles. The dimensionality of the input data was reduced using time-lagged independent components analysis (TICA) as described in Pérez-Hernández et al. [2013]. The data were projected on the time-independent components using the kinetic map projection proposed by Noé and Clementi [2015]. The lag-time used for TICA and the number of components considered is reported in Tab. 7.1.

As a benchmark for standard MSM we discretized the TICA projected space using a k-means clustering algorithm, and successively follow the standard MSM approach, using the software pyEMMA 2.2 [Scherer et al., 2015]. The number of k-means centers used in the analysis is reported in Tab. 7.1.

The UDP clustering was performed as described by d'Errico et al. [2016] using home-built python scripts. This requires no input parameter other than the intrinsic dimensionality of the data set, which was estimated using the 2-NN method, recently developed by Facco and Laio [2016]. The UDP algorithm gives as output both the local density ρ_i at each data point and its cluster assignment. The number of clusters obtained is reported in Tab. 7.1. We defined a point i to be in a core if

$$\frac{\rho_i}{\rho_{max}} > e^{-1} \tag{7.3}$$

where ρ_{max} is the maximum density of the cluster to which point i is assigned.

In order to have a proper statistical analysis of the performance of different models, we performed a cross-validation analysis as described by McGibbon and Pande [2015]. This consist in subdividing the available simulation data into k disjointed sets, construct a MSM excluding the subset l from the input data and then test it by evaluating a performance score on the subset l. The average of the performance score over the k possible choices of the test subset of data l, will give an unbiased estimation of the general perfomance of the model, taking into account any systematic error due to overfitting of the data. McGibbon and Pande [2015] proposed an optimal performance score for MSMs, which is the generalized matrix Rayleigh quotient (GMRQ), previously introduced by Noé and Nuske [2013].

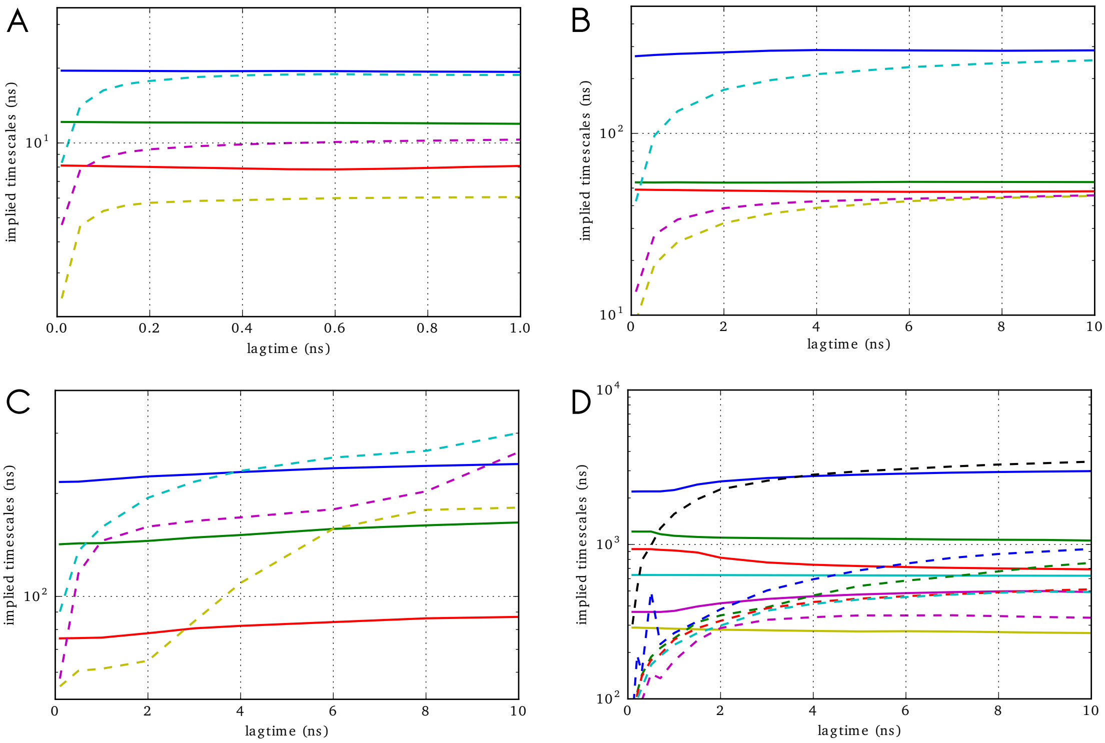


Figure 7.2: Implied timescales dependency on the lag time for MSMs built using different clustering algorithm: UDP (continuous line), k-means (dashed line). The four panels show the results for the four molecules considered: adenine dinucleotide (A), adenine trinucleotide (B), 4-bp duplex (C), 5-bp duplex (D).

Results of the comparison

Fig. 7.2 shows the behavior of the implied timescales relative to the slowest processes of the MSMs built with both the standard and the core-set approach, for the four RNA molecules. The core-set MSM exhibits a virtually perfect convergence of the implied timescales even for very small values of the lagtime τ. This is a convenient feature of the core-set approach, as it makes possible to deal with MSM with a high time resolution.

In contrast, the implied timescales from the full-state MSM approach reach convergence only for relatively large values of τ for the two oligonucleotide systems (0.2 ns for the dinucleotides; 5 ns for the trinucleotide). When dealing with the higher complexity of the two unzipping duplexes the k-means-based MSM fails to reach properly converged implied timescales for lagtimes in the

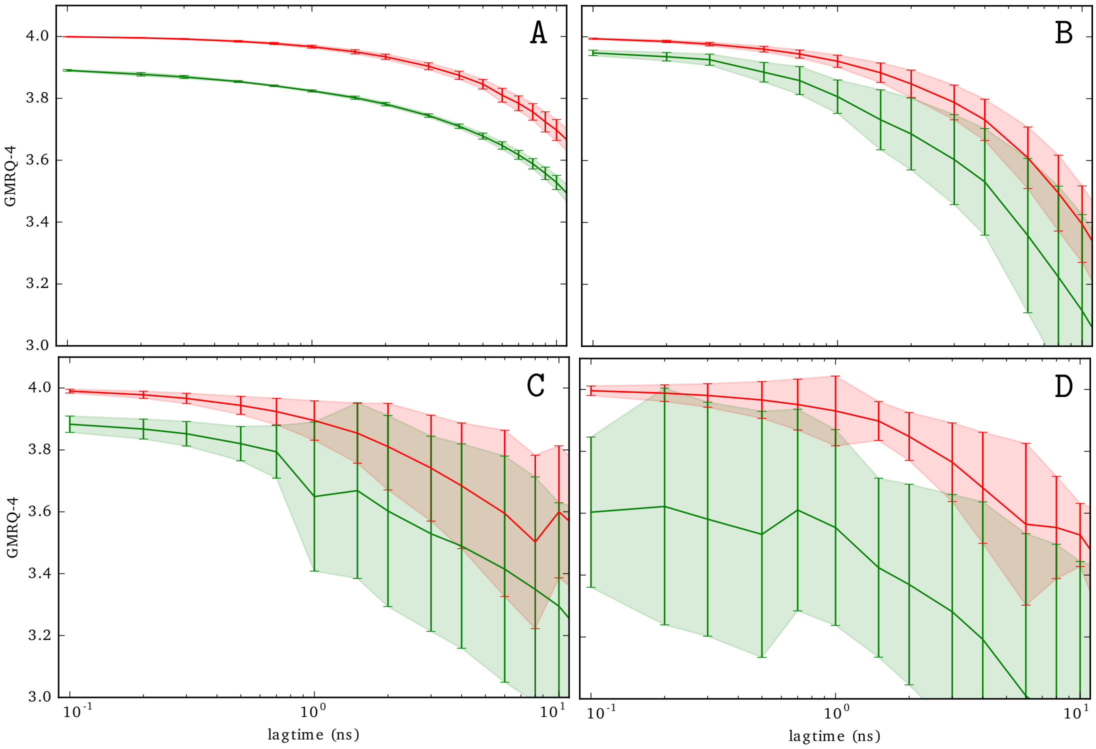


Figure 7.3: Cross validation of the MSM built with different clustering algorithms: UDP (red) k-means (green). Panel A: adenine dinucleotide; panel B: adenine trinucleotide; panel C: 4-bp duplex; panel D: 5-bp duplex.

range $<$ 10 ns. We notice that in the AA dinucleotide, where convergence is reached for relatively small values of τ, the timescales obtained with the full-state MSM are sensibly smaller than the ones obtained with the "UDP+coring" approach. This is an index of a MSM of lower quality as explained by Noé and Nuske [2013] and Nüske et al. [2014].

In order to properly assess the statistical significance of the improvement in the MSM accuracy due to the "UDP+coring" approach we performed a cross-validation using the GMRQ as described by McGibbon and Pande [2015]. Fig. 7.3 shows the results of this cross-validation. The value of the GMRQ decreases when the MSM lagtime increases, as expected [McGibbon and Pande, 2015]. We remark that since the eigenvalues of an MSM are expected to depend on the lag time, even in the situation of a perfectly Markovian system, it make no sense to compare the values of GMRQ for different values of τ. Nevertheless, the GMRQ computed for the “UDP+coring” MSMs is always larger than the one obtained for k-means MSMs. This difference is particularly significant in the well-converged dinucleotide system. The statistical uncertainty become larger when increasing the lagtime and the complexity of the studied system.

Discussion

From the results presented in the previous section, we shown that a MSM built using a combination of UDP clustering and core-set milestoning is able to reproduce the kinetics of complex RNA biomolecules with a level of accuracy equal or greater than a standard full-state MSM based on k-means clustering.

One of the main advantages of the core-set MSM approach is to enable an extremely fine time resolution. This is fundamental for an accurate estimation of kinetic quantities such as mean first passage time, or to the identification of transition pathways between different states of the system.

Other convenient features are: 1) the relatively small number of clusters required in order to have a good convergence of the implied timescales, which translate in a simpler model, easier to analyze and visualize; 2) the unsupervised nature of the UDP clustering, that automatically determines an optimal clusterization for complex high-dimensional and non-homogeneous data sets, without requiring to tune any input parameter, or manually perform a hierarchical cluterizations. This last feature makes this approach extremely easy and quick to implement.

In the next chapter we will exploit this approach to analyze more deeply the kinetics of unzipping of the terminal base pair of an RNA duplex.

Chapter 8

Kinetics of base fraying

In this chapter we report a study on the kinetics of fraying of the terminal base pair in an RNA double helix. In particular we performed a computational analysis using 1) atomistic molecular dynamics (MD) to simulate the dynamics of a small RNA duplex and 2) Markov state models (MSM) to extract kinetic information from the MD.

The work presented in this chapter has been mainly performed during a three weeks visit in the Computational Molecular Biology group at the Freie Universität Berlin, thanks to the financial contribution of the European Molecular Biology Organization.

8.1 Introduction

The phenomenon of base fraying consists in the breaking of the pairing and stacking interactions of the bases at one terminus of an RNA (or DNA) double helix. This phenomenon plays crucial a role in the stability of double helices and it is the first step of the opening of a duplex, which can be either spontaneous or driven by nucleic acid processing enzymes. Base fraying has been characterized in terms of free energy by Colizzi and Bussi [2012] and it has been recently studied by means of computational techniques by Zgarbová et al. [2014]. However, a quantitative characterization of the kinetic of the process, which is fundamental to understand the time-scales involved, is still missing.

In particular Colizzi and Bussi [2012] examined the two possible pathways in which the 5′ or the 3′ terminal base is the first to unstack from the rest of the helix. A significant free-energy difference between the two intermediates was found, suggesting a preference for the dangling 3′ pathway. This hypothesis is supported by experimental evidence such as ultrafast spectroscopy [Liu et al., 2008], analysis of ribosome x-ray structure [Mohan et al., 2009], and contribution of 5′/3′ dangling ends to duplex stability [Turner et al., 1988]. However, to give a final answer to this question, from a theoretical point of view, it is necessary to study the kinetics of the process, understanding the different rates associated with the two possible pathways.

8.2 | Molecular dynamics

The simulated systems consist in a $\genfrac{}{}{0pt}{}{\texttt{5'-ACGC}}{\texttt{3'-UGCG}}$ duplex. Since we decided to focus the study on the fraying of the A-U terminal pair, we constrained the distances between the heavy atoms involved in the hydrogen bonds corresponding to the G-C pairs. Details of the simulations are given in Section 7.3.

From an initial analysis of the MD trajectories we observed the formation of several ladder-like structures [Banáš et al., 2010] (fig. 8.1). These unphysical structures are a known artifact present in MD simulations of RNA duplexes, and are likely caused by the constraints on the base pairs distances Bergonzo et al. [2015].

In order to exclude these unphysical structures from the analysis we proceed in this way:

1. for each frame, compute the root mean square deviation (RMSD) after optimal alignment of all the atoms in the three constrained G-C pairs, with respect to the canonical double helix (fig. 8.1);

2. remove all the frames with RMSD$>$ 0.275 nm;

3. keep only the continuous chunks of trajectories with length $>$ 100 ns.

This procedure lead to a final set of 105 trajectories, with a total length of 51 µs.

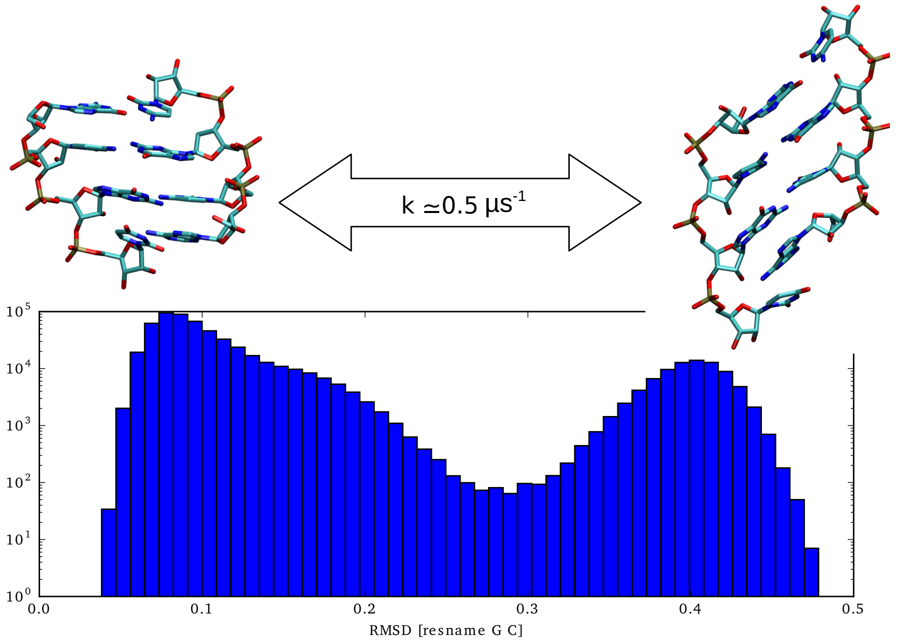


Figure 8.1: Histogram of the RMSD computed only on the three G-C base pairs, in logarithmic scale; the A-form helix and the ladder-like structure are shown above, along with the transition rate between the two states.

8.3 Markov state model

In order to analyze the trajectories produced from the MD we considered the following set of coordinates: 1) G-vectors; 2) backbone dihedrals; 3) sugar ring torsional angles; 4) glycosidic torsional angles. The dimensionality of the input data was then reduced using TICA [Pérez-Hernández et al., 2013] with a lag-time of 5 ns. Data were projected on the 23 slowest TICs using a kinetic map projection [Noé and Clementi, 2015], and then discretized using the UDP clustering [Rodriguez and Laio, 2014; d'Errico et al., 2016]. A lag-time $\tau = 100$ ps was used to construct MSMs that approximate the dynamics of the discretized systems. The quality of the Markovian approximation was tested 1) looking at the convergence of the implied timescales predicted by the MSM for increasing values of τ (see figure 8.2A), 2) performing a Chapman-Kolmogorov test, as described in [Prinz et al., 2011] (see figure 8.2B).

The timescales predicted by the MSM show a small gap between the 3rd

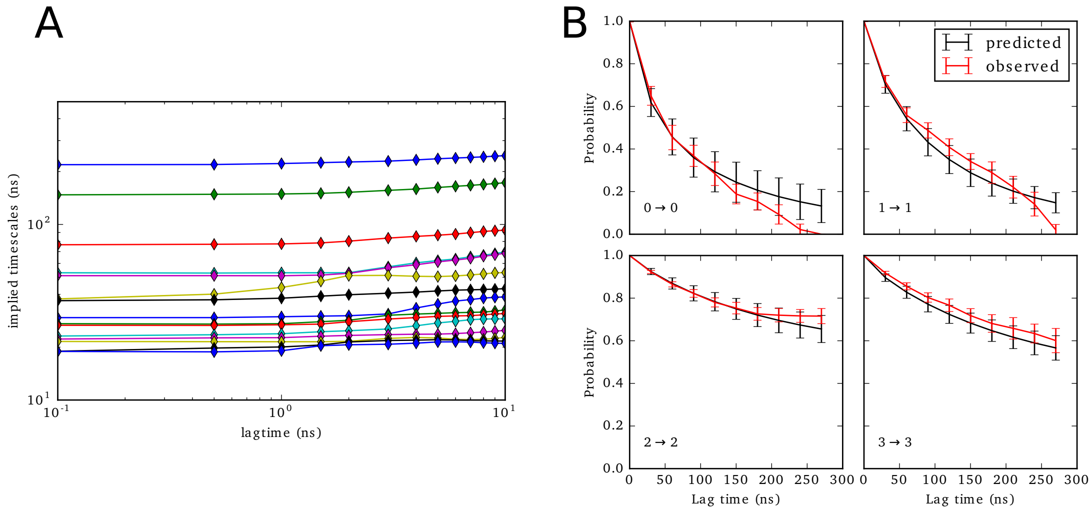


Figure 8.2: Test of Markov approximation; A: implied timescales predicted by MSM as a function of lagtime; B: comparison of the residence probability for each metastable macrostate as predicted propagating the transition probability of the MSM, built at $\tau = 0.1$ ns, and the actual transitions observed from the simulation.

and the 4th one. This allows to lump the microstates into a few metastable macrostates, using an hidden markov model (HMM) [Noé et al., 2013]. Fig. 8.3 shows the structure of the terminal base pair in the four macrostates. The macrostates 2 and 3 are distinguished by the orientation of the uracil base at the 3′ terminus (U3′); in state 3 the U3′ base is in the canonical orientation, forming a Watson-Crick base pair with the adenine at the 5′ terminus (A5′), while in state 2 it is flipped upside down, forming a non-canonical base-pair with U3′. In states 0 and 1 base A5′ is almost completely solvated, without any stacking with G2 nor pairing with U3′. States 0,1 are again separated by the orientation of U3′. We can therefore describe the slow processes in the system, from slower to faster, in this way: the slowest process is the flipping of U3′ base while A5′ is stacked on the remaining part of the duplex (220 ns), followed by the unstacking of A5′ (150 ns), then followed by the flipping of U3′ base while A5′ is unstacked (80 ns). This analysis tells us that the state with A5′ open is a kinetically stable state. On the other hand we do not identify a metastable state with U3′ open, nor one with both bases unstacked.

In order to study the opening of the two terminal A/U bases individually

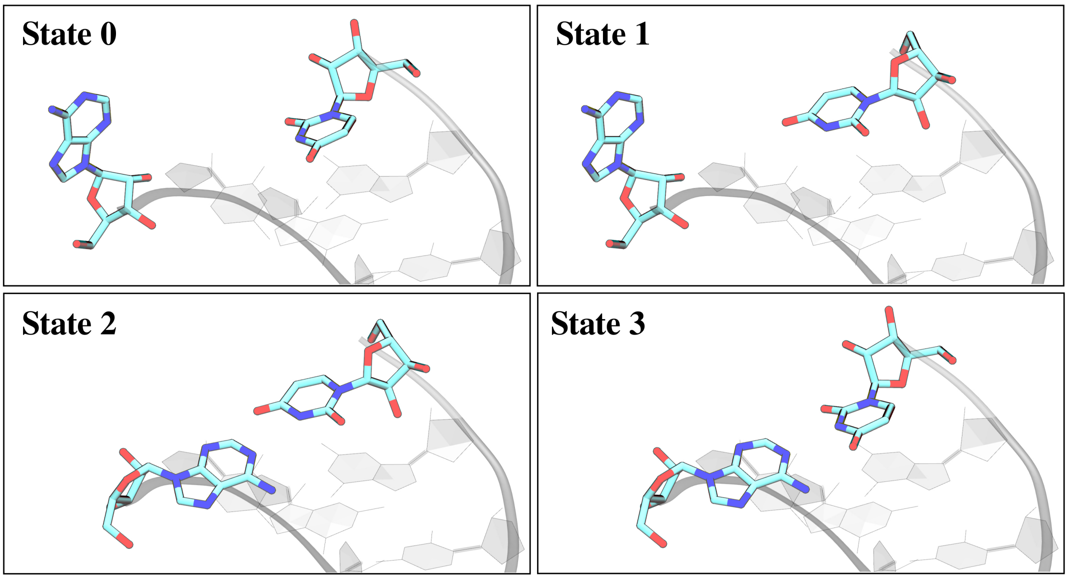


Figure 8.3: Representation of the structure of the terminal bases in the four metastable states identified by the HMM.

we quantitatively measured their stacking interaction with the adjacent G/C base, using the stacking definition proposed in [Condon et al., 2015].

We already observed that structures in which the terminal base-pair is completely open are not metastable since they are not distinguished by the kinetic coarse-graining. For this reason we further classify the clusters based on the fraction of stacked structures that they contain (see figure 8.4). For each microstate we define the stacking interaction of one of the two terminal bases to be broken if the fraction of stacked structures is < 0.5. Microstates where both stacking interactions are formed correspond to a closed terminal pair, while the structures with both stacking broken correspond to completely open terminals.

After dividing the microstates in these four groups (see figure 8.4) it is possible, using transition path theory, as described by Noé et al. [2009], to compute the the mean first passage time (MFPT) from the closed group of states (C) to the open one (O), and viceversa, as well as the fraction flux from $C \rightarrow O$ passing through group $5pO$ (A5′ open) or $3pO$ (U3′ open). The results of this analysis are shown in figure 8.4. We can see that the fraction of flux going through $3pO$ is significantly smaller than the fraction of flux passing

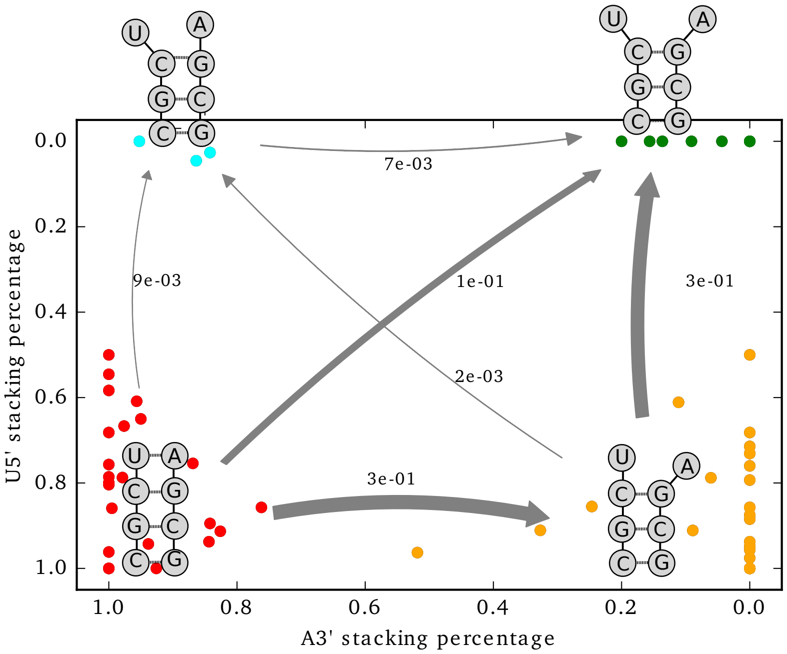


Figure 8.4: Representation of the probability flux from C to O; circles represent the microstates, with x and y coordinates equal to the fraction of stacked structures they contain; colors indicate if the microstate is considered closed (red), open (green), A5′ open (yellow), or U3′ open (cyan); arrows′ width is proportional to the net flux (indicated in μs^{-1}).

through 5*pO*, while only a negligible amount of flux is jumping directly from *C* to *O*. The MFPT from *C* to *O* is 5 μs.

8.4 Discussion

The MSM built on extensive MD simulations of the $\substack{\texttt{5'-ACGC}\\\texttt{3'-UGCG}}$ duplex showed to be able to approximate the full atomistic dynamics of the system with a satisfactory of accuracy (see fig. 8.2). Such MSM identifies no kinetically stable state corresponding to a fully frayed terminal. The metastable states instead identify structures with flipped U5′ base and unstacked A5′ base.

We then focused on the mechanism associated to RNA unzipping. We observe that the preferred pathway for the system is the one in which the A5′ base breaks its stacking and pairing interactions first, while the U3′ base opening is the last step of the reaction. This confirms the mechanism proposed by previous computational studies [Colizzi and Bussi, 2012], where it was shown that the free energy increase associated to opening one of the bases of the terminal pair is smaller when opening the 5′ base. We remark that the previous work by Colizzi and Bussi [2012] was limited to an energetic analysis. The present work confirms the proposed mechanism by means of a complete kinetic analysis.

Concluding, we obtained a MSM able to accurately reproduce the kinetic properties of the terminal base pair of an RNA double helix. This makes possible to obtain a robust estimation of rates and MFPT between folded and frayed structures, to identify metastable states and their relative equilibrium populations, as well as the unzipping pathways followed by the system.

Conclusions and perspectives

The importance of ribonucleic acid (RNA) in the field of biology is constantly growing, as researchers discover more and more roles played by non-coding RNAs in the life of cells. Characterizing the functional dynamics of RNA molecules represents one of the avenues that are being most actively pursued by molecular biologists. This thesis investigated the topic of computational modeling of the dynamical properties of small RNA molecules. The research presented focused both on atomistic molecular dynamics (MD) simulations and coarse-grained models.

In particular, we reported the application on RNA molecules of a class of very simple coarse-grained representations, elastic network models (ENMs) [Pinamonti et al., 2015]. We studied the applicability of such models benchmarking their predictions against accurate atomistic molecular dynamics simulations, comparing different possible models and identifying the optimal parameters to describe such systems. In addition, we tested the robustness of our findings, comparing ENM predictions with RNA chain flexibility measured by SHAPE experiments (Selective 2'-hydroxyl acylation analyzed by primer extension, see Wilkinson et al. [2006]). Finally, we investigated the precision of the vibrational entropy computed from ENM on the *add* riboswitch.

In the second part of the thesis we dealt with the application of Markov state models (MSMs, see Bowman et al. [2013]) to the analysis of MD simulations of RNA molecules, using them to study the kinetics of the fundamental interactions of such molecules, namely base stacking [Pinamonti et al., 2016] and base pairing. Concurrently we examined a novel recipe for the construction of reliable core-set MSMs, based on the unsupervised density-peak clustering developed by d'Errico et al. [2016].

The work reported in this thesis may provide a starting point for different applications in the field of computational modeling of RNA systems. Possible routes for future research include the application of ENM to analyze the flexibility of larger and more complex RNA molecules, and a comparison with a larger pool of experimental data. The analysis of the vibrational entropy could be improved with a more deep investigation, by employing more detailed ENMs and including different systems in the comparison.

On the MSM side, our analysis of RNA-helix unzipping can be easily extended by examining different RNA sequences in order to highlight the sequence dependence of the process, as well as by adding to the picture the fraying of the following base-pairs. The computational predictions of the opening/closing rates could also be compared with single-molecule experiments with optical tweezers, and this comparison could be further improved by introducing a pulling force on the terminal base pairs during the simulation.

MSM analysis of more complex RNA systems is also an appealing development. A feasible process to be investigated is the so called "RNA strand invasion", the process by which one strand of an RNA double helix is replaced by an equivalent strand. Due to the equivalence of the sequences of the original and the replacing strand, the free-energy difference between the initial and the final states of this process is zero. The key feature of this process is thus determined by its kinetics. An even more complex future task would be to apply the MSM approach in the study of the conformational changes in one of the riboswitches that have been shown to operate in a kinetic regime, rather than at the thermodynamic equilibrium [Lemay et al., 2011].

Appendix A

ENM supplementary data

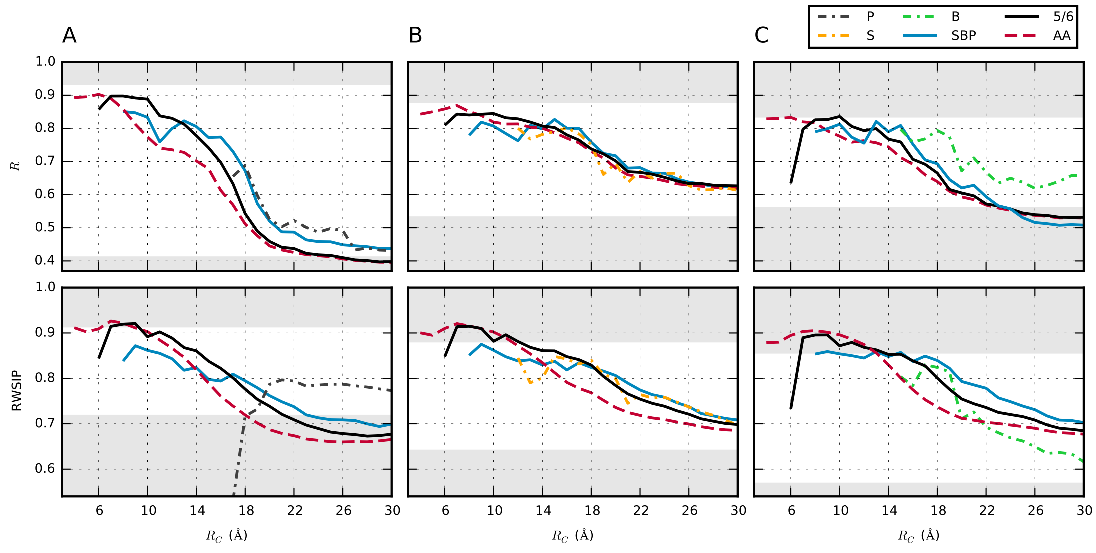


Figure A.1: Comparison between the MD simulations and a ENM constructed considering 5 or 6 beads per nucleotide: P, C1′, C4′, C2, C5 for pyrimidines; the same plus C8 for purines. Also the 1-,3-beads models and the all-atom model are shown for comparison. Fluctuations′ correlation (upper panels) and RWSIP (lower panels) are computed considering the beads in the base (A), sugar (B) and phosphate (C), and are averaged over the four molecules considered.

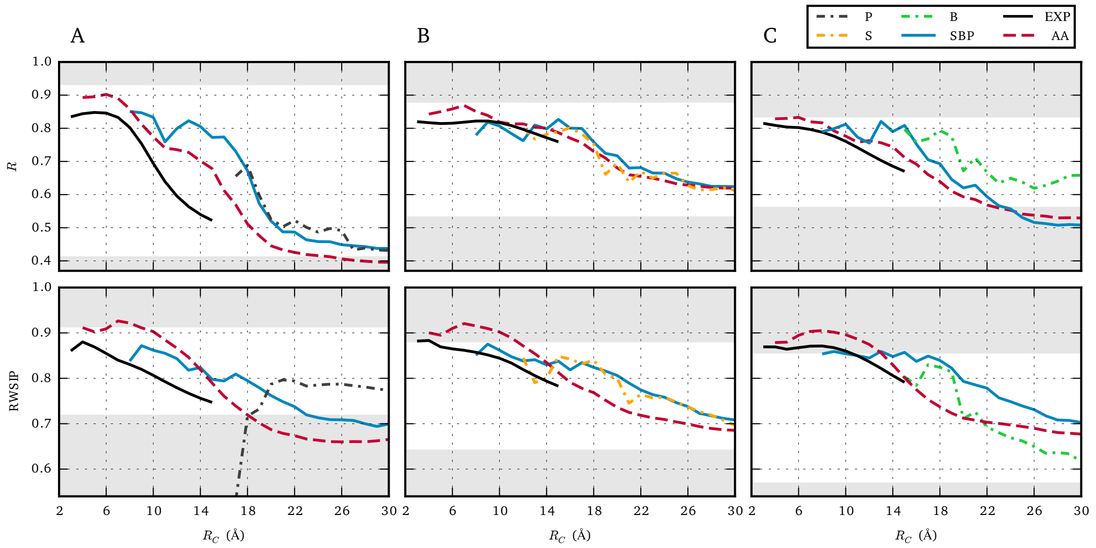


Figure A.2: Comparison between the MD simulations and a 3-beads SBP ENM with space-dependent elastic constant. The elastic constant is given by $k_{ij} = \exp(-(d_{ij}/R_C)^2)$. The 3-beads SBP model with sharp cutoff is shown for comparison. Fluctuations' correlation (upper panels) and RWSIP (lower panels) are computed considering the beads in the base (A), sugar (B) and phosphate (C), and are averaged over the four molecules considered.

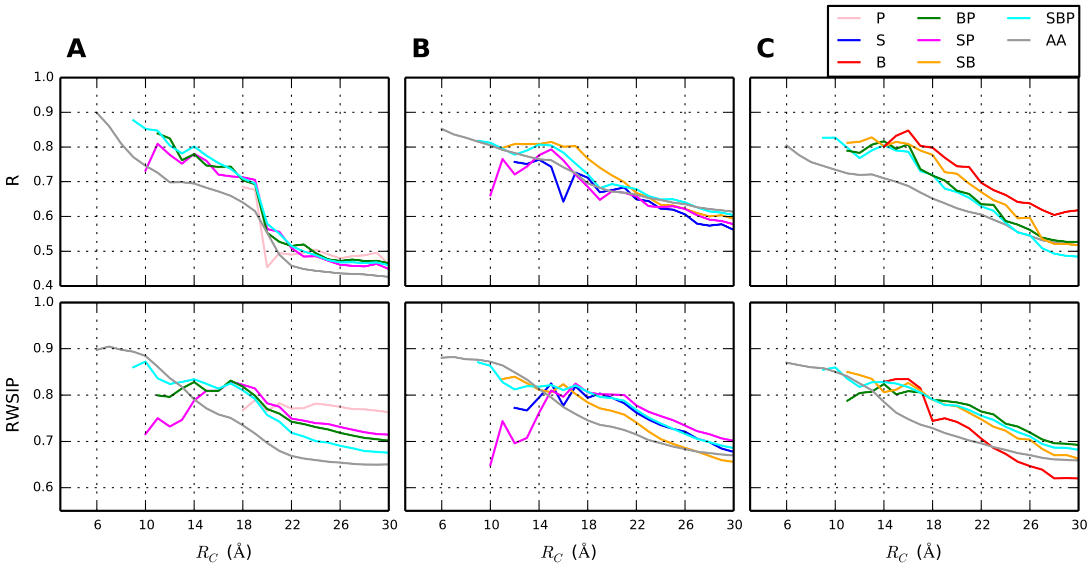


Figure A.3: Comparison between the MD simulations and the ENMs using as references the experimental structures. Fluctuations' correlation (upper panels) and RWSIP (lower panels) are computed considering the beads in the base (A), sugar (B) and phosphate (C), and are averaged over the four molecules considered.

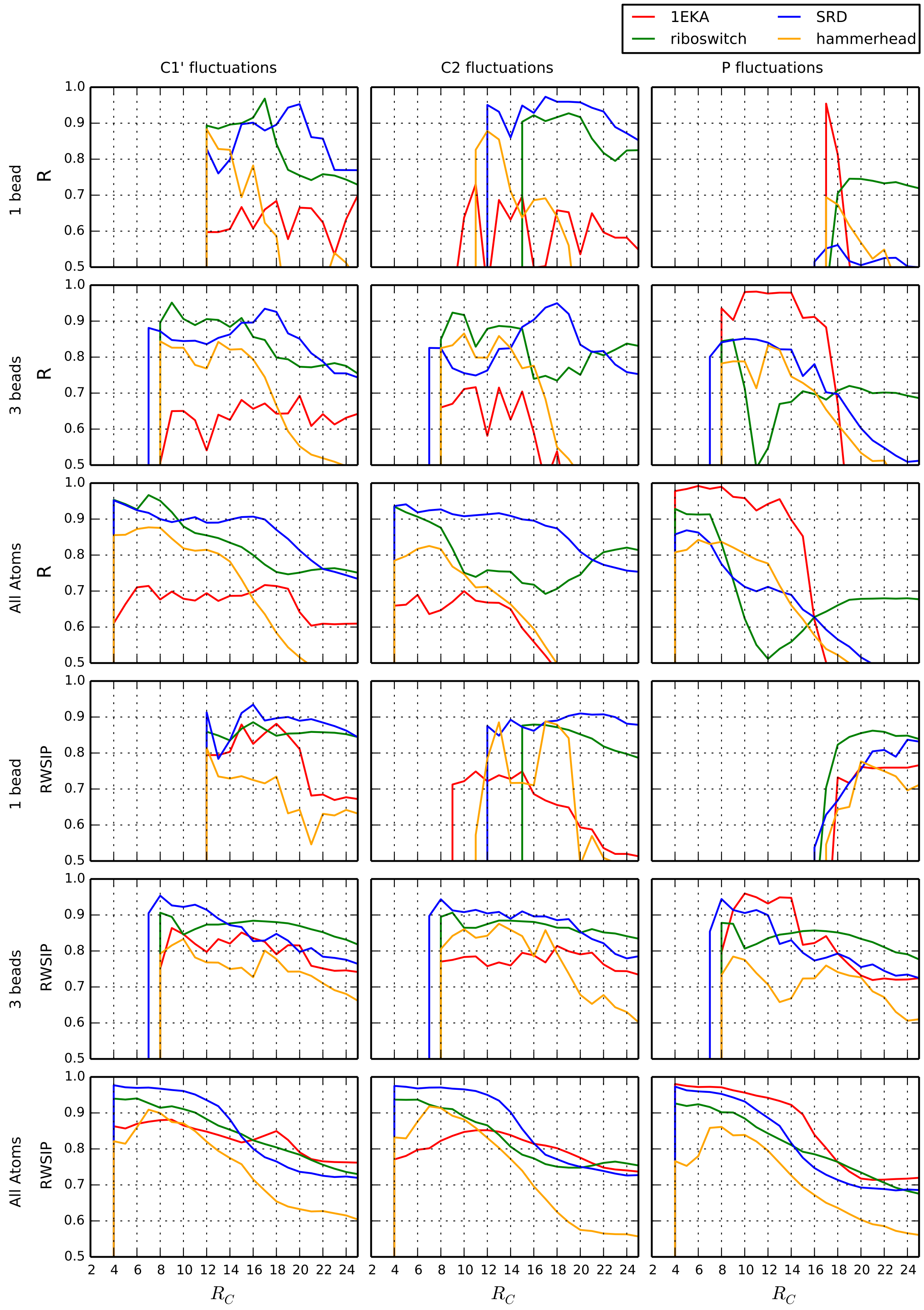


Figure A.4: Agreement between MD simulations and ENM for different radii of cutoff. Correlation between MSF (upper panels), and RWSIP (lower panels). The results are shown separately for the 4 different molecules for the 1 bead, 3-beads SBP and for the all atom model, as labeled. Left: phosphate beads; middle: sugar beads; right: nucleobase beads.

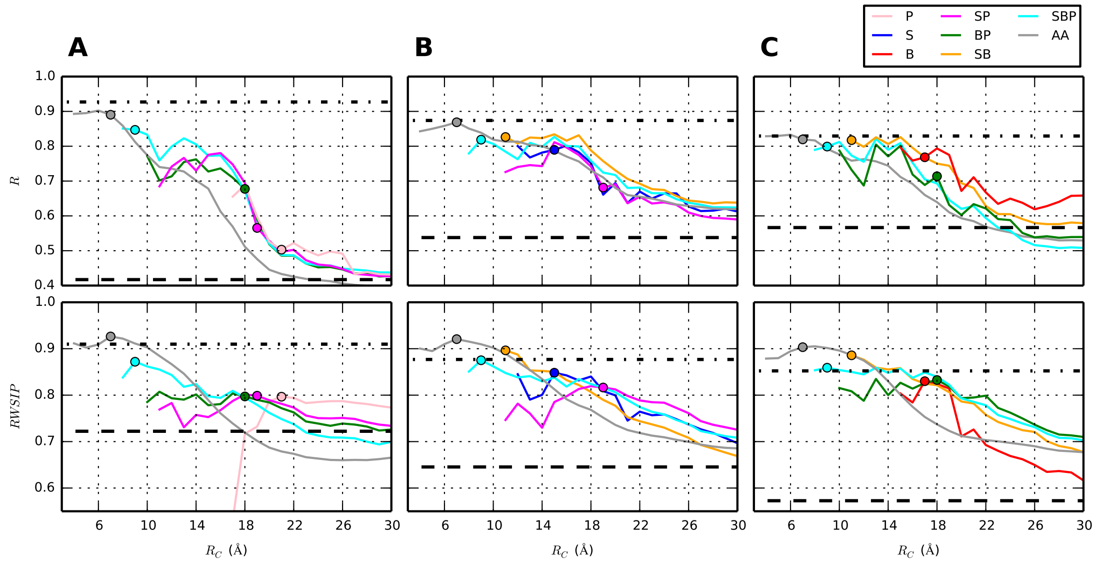


Figure A.5: Agreement between MD and ENM. Comparison of all the considered models with 1,2,3 beads per nucleotide as well as the all-atoms model. Values at the optimal cutoff values are represented by circles. Fluctuations' correlation (up) and RWSIP (down) are computed considering the beads in base (A), sugar (B) and phosphate (C), and are averaged over the four molecules studied.

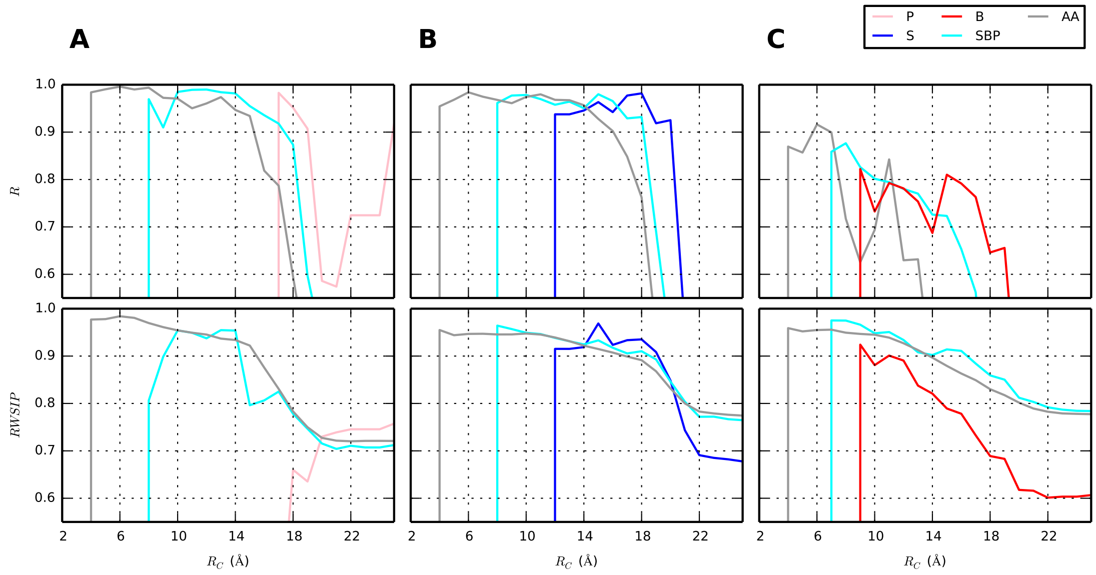


Figure A.6: Agreement between MD and ENM on the duplex molecule excluding the terminal residues from the analysis. Fluctuations' correlation (up) and RWSIP (down) are computed considering the beads in base (A), sugar (B) and phosphate (C).

Appendix B

MSM supplementary data

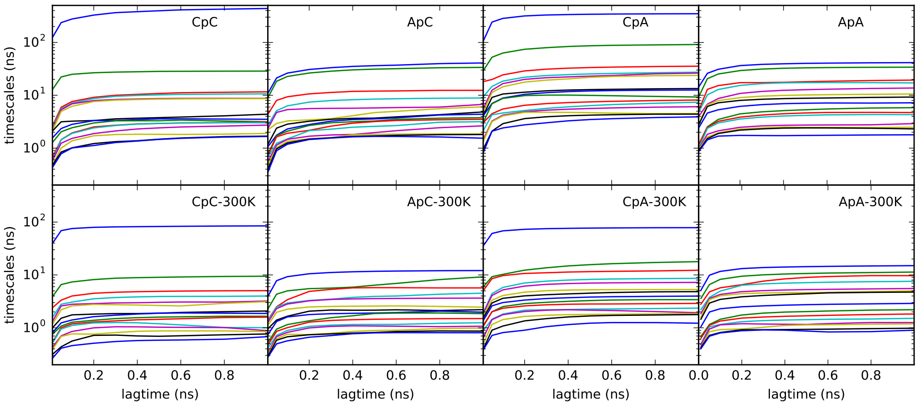


Figure B.1: Convergence of the implied timescales of the eight dinucleotide systems as a function of the MSM lagtime.

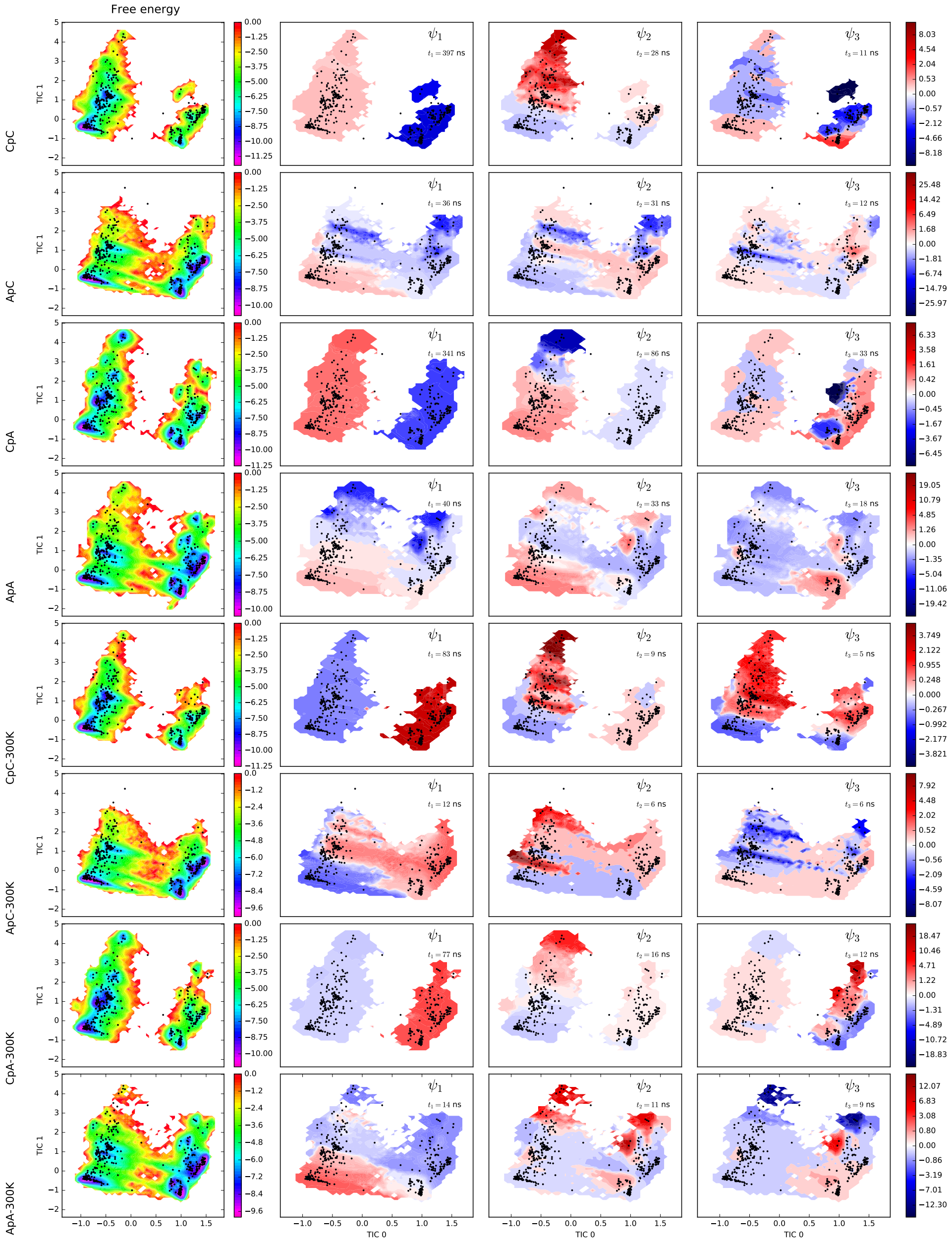
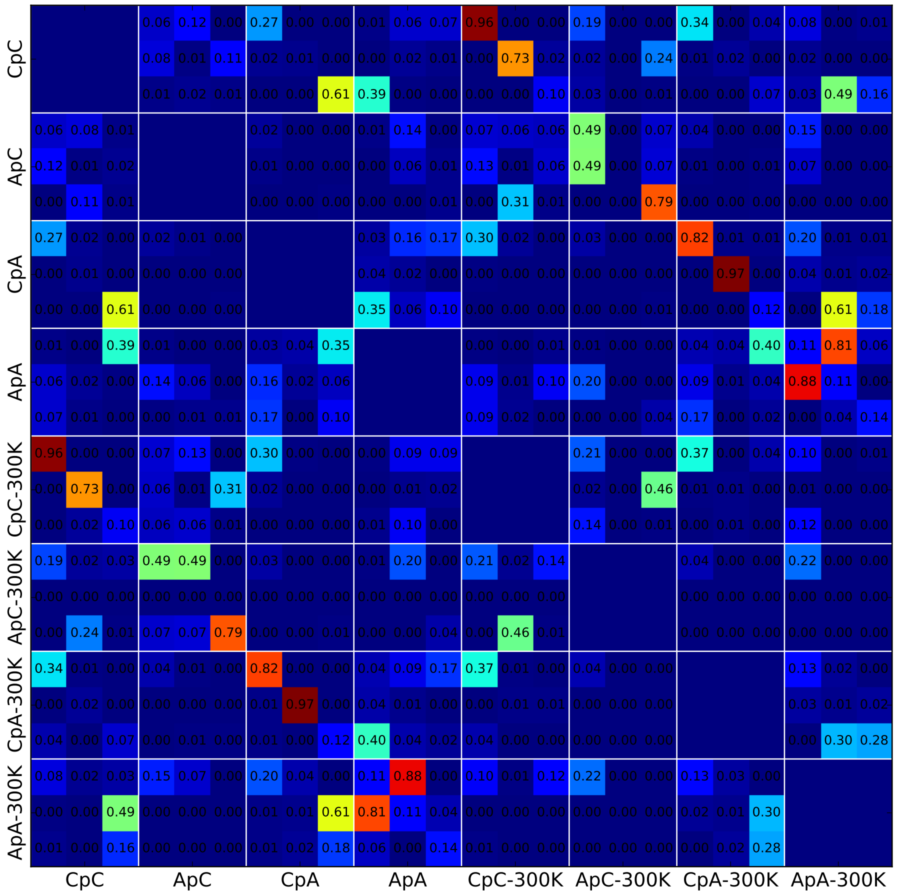


Figure B.2: Graphical representation of the first three eigenvectors of each of the eight dinucleotides systems.



	CpC			ApC			CpA			ApA			CpC-300K			ApC-300K			CpA-300K			ApA-300K		
CpC				0.06	0.12	0.00	0.27	0.00	0.00	0.01	0.06	0.07	0.96	0.00	0.00	0.19	0.00	0.00	0.34	0.00	0.04	0.08	0.00	0.01
				0.08	0.01	0.11	0.02	0.01	0.00	0.00	0.02	0.01	0.00	0.73	0.02	0.02	0.00	0.24	0.01	0.02	0.00	0.02	0.00	0.00
				0.01	0.02	0.01	0.00	0.00	0.61	0.39	0.00	0.00	0.00	0.00	0.10	0.03	0.00	0.01	0.00	0.00	0.07	0.03	0.49	0.16
ApC	0.06	0.08	0.01				0.02	0.00	0.00	0.01	0.14	0.00	0.07	0.06	0.06	0.49	0.00	0.07	0.04	0.00	0.00	0.15	0.00	0.00
	0.12	0.01	0.02				0.01	0.00	0.00	0.00	0.06	0.01	0.13	0.01	0.06	0.49	0.00	0.07	0.01	0.00	0.01	0.07	0.00	0.00
	0.00	0.11	0.01				0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.31	0.01	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00
CpA	0.27	0.02	0.00	0.02	0.01	0.00				0.03	0.16	0.17	0.30	0.02	0.00	0.03	0.00	0.00	0.82	0.01	0.01	0.20	0.01	0.01
	0.00	0.01	0.00	0.00	0.00	0.00				0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.04	0.01	0.02
	0.00	0.00	0.61	0.00	0.00	0.00				0.35	0.06	0.10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.12	0.00	0.61	0.18
ApA	0.01	0.00	0.39	0.01	0.00	0.00	0.03	0.04	0.35				0.00	0.00	0.01	0.01	0.00	0.00	0.04	0.04	0.40	0.11	0.81	0.06
	0.06	0.02	0.00	0.14	0.06	0.00	0.16	0.02	0.06				0.09	0.01	0.10	0.20	0.00	0.00	0.09	0.01	0.04	0.88	0.11	0.00
	0.07	0.01	0.00	0.00	0.01	0.01	0.17	0.00	0.10				0.09	0.02	0.00	0.00	0.00	0.04	0.17	0.00	0.02	0.00	0.04	0.14
CpC-300K	0.96	0.00	0.00	0.07	0.13	0.00	0.30	0.00	0.00	0.00	0.09	0.09				0.21	0.00	0.00	0.37	0.00	0.04	0.10	0.00	0.01
	0.00	0.73	0.00	0.06	0.01	0.31	0.02	0.00	0.00	0.00	0.01	0.02				0.02	0.00	0.46	0.01	0.01	0.00	0.01	0.00	0.00
	0.00	0.02	0.10	0.06	0.06	0.01	0.00	0.00	0.00	0.01	0.10	0.00				0.14	0.00	0.01	0.00	0.01	0.00	0.12	0.00	0.00
ApC-300K	0.19	0.02	0.03	0.49	0.49	0.00	0.03	0.00	0.00	0.01	0.20	0.00	0.21	0.02	0.14				0.04	0.00	0.00	0.22	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.24	0.01	0.07	0.07	0.79	0.00	0.00	0.01	0.00	0.00	0.04	0.00	0.46	0.01				0.00	0.00	0.00	0.00	0.00	0.00
CpA-300K	0.34	0.01	0.00	0.04	0.01	0.00	0.82	0.00	0.00	0.04	0.09	0.17	0.37	0.01	0.00	0.04	0.00	0.00				0.13	0.02	0.00
	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.97	0.00	0.04	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00				0.03	0.01	0.02
	0.04	0.00	0.07	0.00	0.01	0.00	0.01	0.00	0.12	0.40	0.04	0.02	0.04	0.00	0.00	0.00	0.00	0.00				0.00	0.30	0.28
ApA-300K	0.08	0.02	0.03	0.15	0.07	0.00	0.20	0.04	0.00	0.11	0.88	0.00	0.10	0.01	0.12	0.22	0.00	0.00	0.13	0.03	0.00			
	0.00	0.00	0.49	0.00	0.00	0.00	0.01	0.01	0.61	0.81	0.11	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.30			
	0.01	0.00	0.16	0.00	0.00	0.00	0.01	0.02	0.18	0.06	0.00	0.14	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.28			

Figure B.3: Similarity between the first three eigenvectors of the eight dinucleotides systems. Numbers refers to the square scalar product between eigenvectors, as define in the Method section of main text.

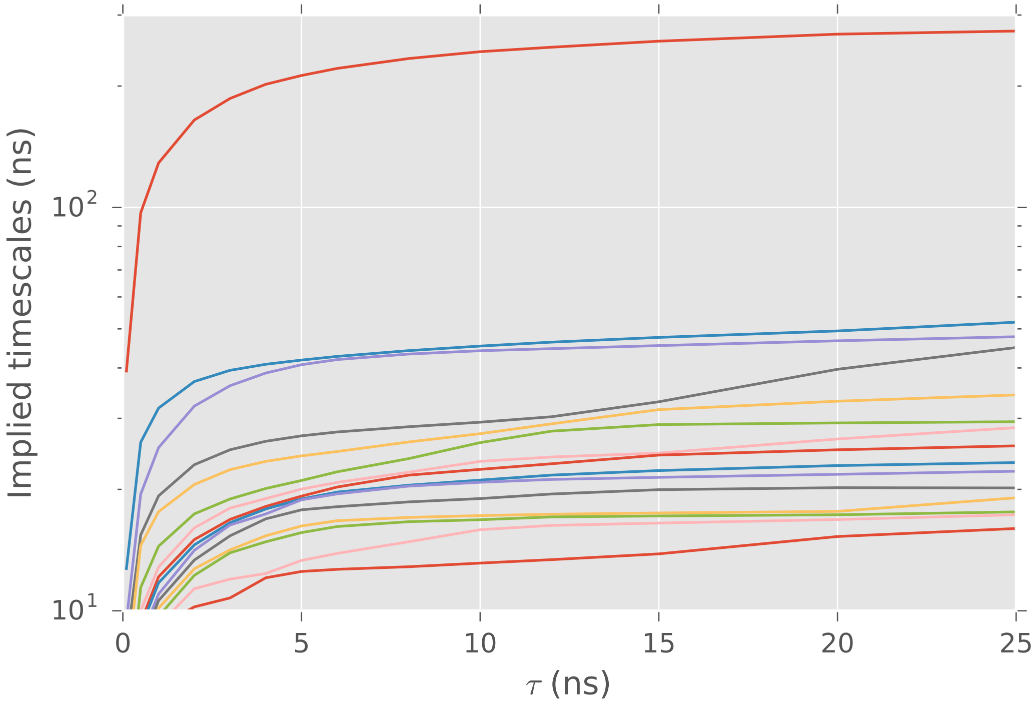


Figure B.4: Convergence of the implied timescales of the adenine trinucleotide system, as a funcion of the MSM lagtime.

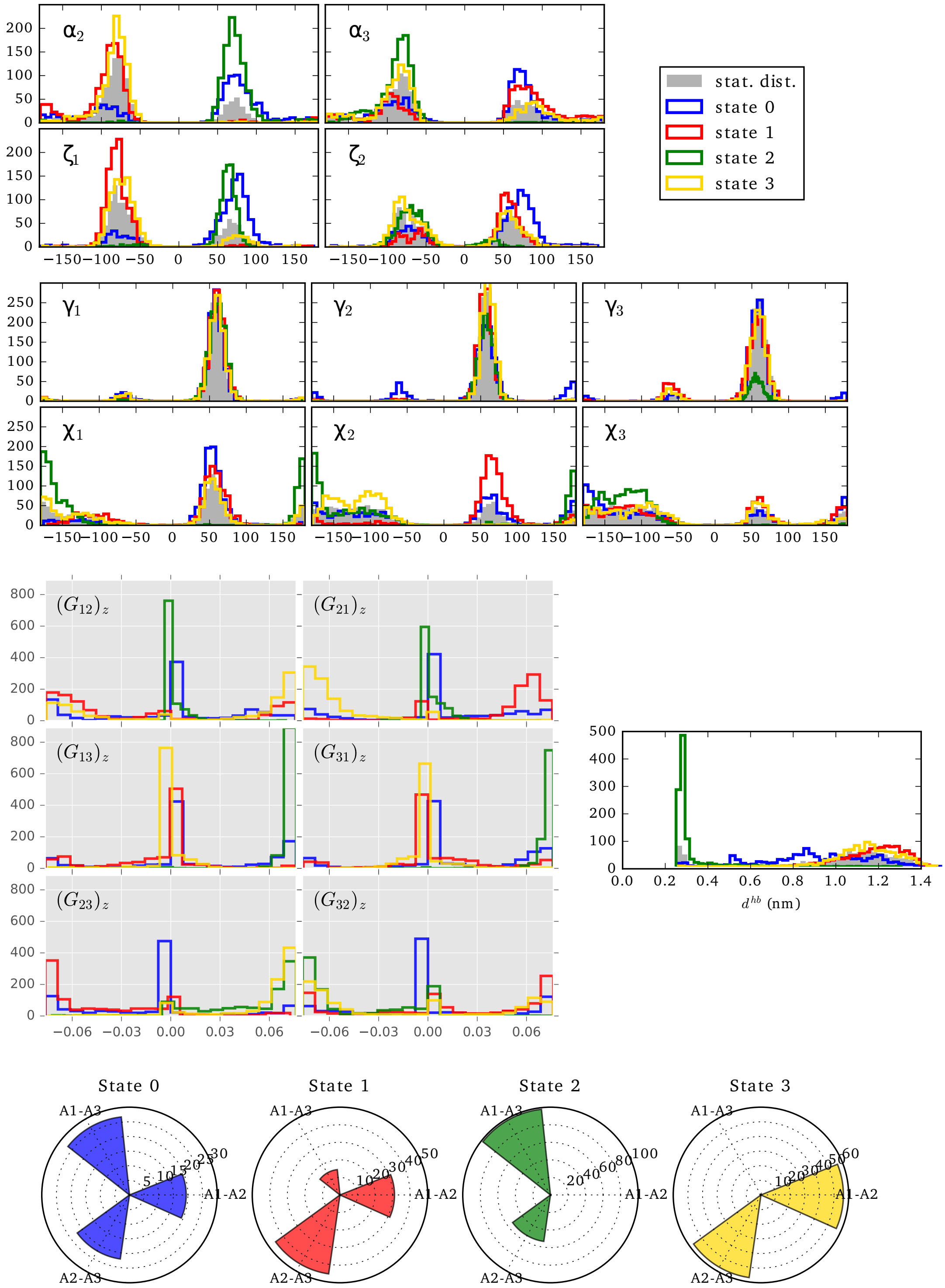


Figure B.5: Distributions of torsional angles, distances, and G-vectors, in the different metastable states identified by the HMM of AAA.

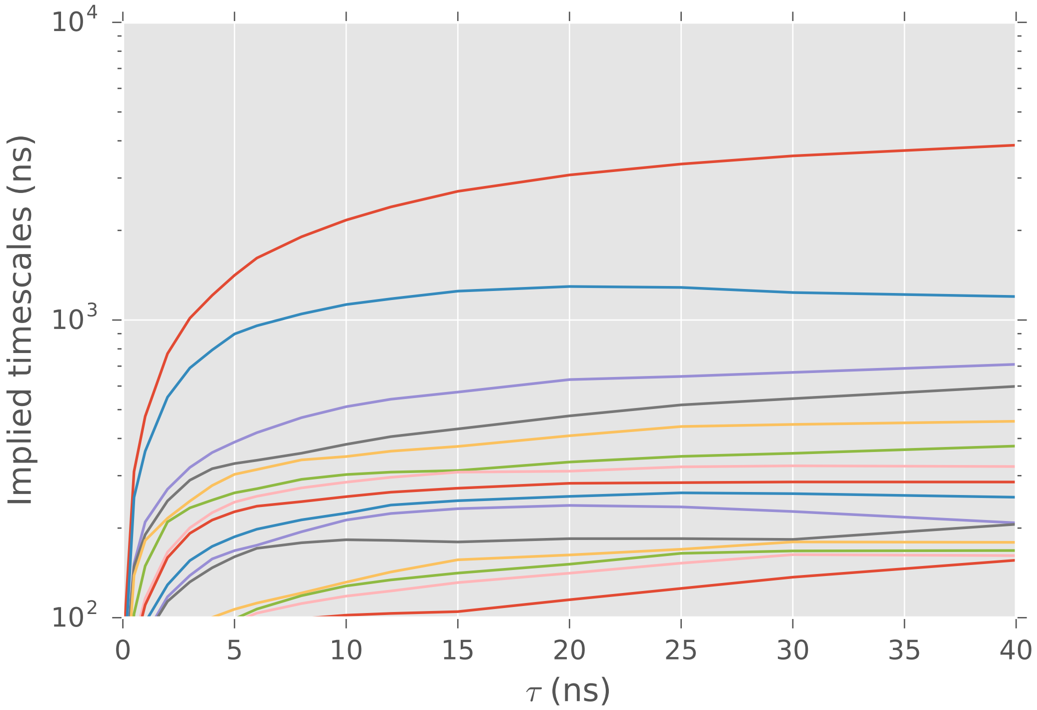


Figure B.6: Convergence of the implied timescales of the adenine tetranucleotide as a function of the MSM lagtime.

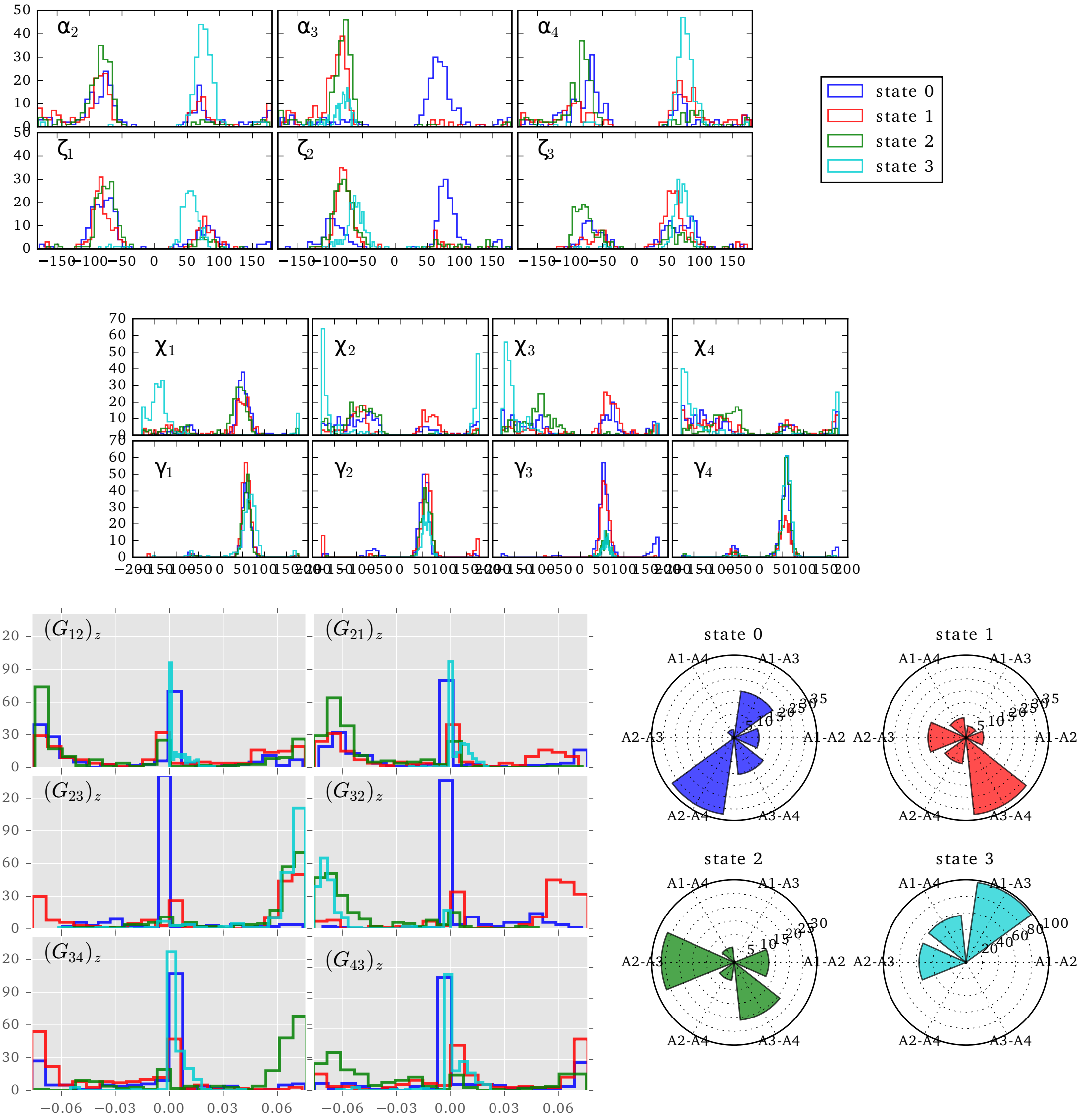


Figure B.7: Distributiontions of torsional angles, atom distances, and G-vectors, in the different metastable states identified by the HMM of AAAA.

Bibliography

Aaqvist, J. [1990], 'Ion-water interaction potentials derived from free energy perturbation simulations', *The Journal of Physical Chemistry* **94**(21), 8021–8024.

Allnér, O., Nilsson, L. and Villa, A. [2013], 'Loop–loop interaction in an adenine-sensing riboswitch: A molecular dynamics study', *RNA* **19**(7), 916–926.

Altona, C. T. and Sundaralingam, M. [1972], 'Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation', *Journal of the American Chemical Society* **94**(23), 8205–8212.

Altona, C. T. and Sundaralingam, M. [1973], 'Conformational analysis of the sugar ring in nucleosides and nucleotides. improved method for the interpretation of proton magnetic resonance coupling constants', *Journal of the American Chemical Society* **95**(7), 2333–2344.

Andricioaei, I. and Karplus, M. [2001], 'On the calculation of entropy from covariance matrices of the atomic fluctuations', *The Journal of Chemical Physics* **115**(14), 6289–6292.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O. and Bahar, I. [2001], 'Anisotropy of fluctuation dynamics of proteins with an elastic network model', *Biophysical Journal* **80**(1), 505–515.

Auffinger, P., Cheatham, T. E. and Vaiana, A. C. [2007], 'Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue?', *Journal of Chemical Theory and Computation* **3**(5), 1851–1859.

Bahar, I., Atilgan, A. R. and Erman, B. [1997], 'Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential', *Folding and Design* **2**(3), 173–181.

Bahar, I. and Jernigan, R. L. [1998], 'Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms', *Journal of Molecular Biology* **281**(5), 871–884.

Banáš, P., Hollas, D., Zgarbová, M., Jurečka, P., Orozco, M., Cheatham III, T. E., Šponer, J. and Otyepka, M. [2010], 'Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins', *Journal of Chemical Theory and Computation* **6**(12), 3836–3849.

Bastolla, U. [2014], 'Computing protein dynamics from protein structure with elastic network models', *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4**(5), 488–503.

Berendsen, H. J., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. and Haak, J. [1984], 'Molecular dynamics with coupling to an external bath', *The Journal of Chemical Physics* **81**(8), 3684–3690.

Bergonzo, C., Henriksen, N. M., Roe, D. R. and Cheatham, T. E. [2015], 'Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields', *RNA* **21**(9), 1578–1590.

Bernardi, R. C., Melo, M. C. and Schulten, K. [2015], 'Enhanced sampling techniques in molecular dynamics simulations of biological systems', *Biochimica et Biophysica Acta (BBA)-General Subjects* **1850**(5), 872–877.

Bloomfield, V., Crothers, D. and Tinoco, I. [2000], *Nucleic Acids: Structures, Properties, and Functions*, University Science Books.
URL: *https://books.google.it/books?id=Qfg028GkovEC*

Bottaro, S., Di Palma, F. and Bussi, G. [2014], 'The role of nucleobase interactions in RNA structure and dynamics', *Nucleic Acids Research* **42**(21), 13306.

Bowman, G. R., Pande, V. S. and Noé, F. [2013], *An introduction to markov state models and their application to long timescale molecular simulation*, Vol. 797, Springer Science & Business Media.

Buchete, N.-V. and Hummer, G. [2008], ‘Coarse master equations for peptide folding dynamics’, *The Journal of Physical Chemistry B* **112**(19), 6057–6069.

Bussi, G., Donadio, D. and Parrinello, M. [2007], ‘Canonical sampling through velocity rescaling’, *The Journal of Chemical Physics* **126**(1), 014101.

Carnevale, V., Pontiggia, F. and Micheletti, C. [2007], ‘Structural and dynamical alignment of enzymes with partial structural similarity’, *Journal of Physics: Condensed Matter* **19**(28), 285206–14.

Case, D. A., Darden, T. A., T E Cheatham, I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Wang, B., Pearlman, D. A. et al. [2004], ‘AMBER 8. 2004’.

Case, D., Babin, V., Berryman, J., Betz, R., Cai, Q., Cerutti, D., Cheatham Iii, T., Darden, T., Duke, R. and Gohlke, H. [2014], ‘AMBER 14’.
URL: *http://ambermd.org/*

Chen, A. A. and García, A. E. [2013], ‘High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations’, *Proceedings of the National Academy of Sciences of the United States of America* **110**(42), 16820–16825.

Chen, X., McDowell, J. A., Kierzek, R., Krugh, T. R. and Turner, D. H. [2000], ‘Nuclear magnetic resonance spectroscopy and molecular modeling reveal that different hydrogen bonding patterns are possible for G·U pairs: One hydrogen bond for each G·U pair in r(GGC$\underline{\text{GU}}$GCC)$_2$ and two for each G·U pair in r(GAG$\underline{\text{UG}}$CUC)$_2$’, *Biochemistry* **39**(30), 8970–8982.

Chodera, J. D. and Noé, F. [2014], ‘Markov state models of biomolecular conformational dynamics’, *Current Opinion in Structural Biology* **25**, 135–144.

Colizzi, F. and Bussi, G. [2012], ‘RNA unwinding from reweighted pulling simulations’, *Journal of the American Chemical Society* **134**(11), 5173–5179.

Condon, D. E., Kennedy, S. D., Mort, B. C., Kierzek, R., Yildirim, I. and Turner, D. H. [2015], 'Stacking in RNA: NMR of four tetramers benchmark molecular dynamics', *Journal of Chemical Theory and Computation* **11**(6), 2729–2742.

Correll, C. C., Beneken, J., Plantinga, M. J., Lubbers, M. and Chan, Y.-L. [2003], 'The common and the distinctive features of the bulged-G motif based on a 1.04 å resolution RNA structure', *Nucleic Acids Research* **31**(23), 6806–6818.

Dang, L. X. [1995], 'Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: a molecular dynamics study', *Journal of the American Chemical Society* **117**(26), 6954–6960.

Darden, T., York, D. and Pedersen, L. [1993], 'Particle mesh Ewald: An N log (N) method for Ewald sums in large systems', *The Journal of Chemical Physics* **98**(12), 10089–10092.

D'Ascenzo, L., Leonarski, F., Vicens, Q. and Auffinger, P. [2016], ''Z-DNA like' fragments in RNA: a recurring structural motif with implications for folding, RNA/protein recognition and immune response', *Nucleic Acids Research* **44**(12), 5944–5956.

Dawson, W. K., Maciejczyk, M., Jankowska, E. J. and Bujnicki, J. M. [2016], 'Coarse-grained modeling of RNA 3D structure', *Methods* **103**, 138–156.

Delarue, M. and Sanejouand, Y.-H. [2002], 'Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model', *Journal of Molecular Biology* **320**(5), 1011–1024.

d'Errico, M., Laio, A., Facco, E. and Rodriguez, A. [2016], 'Fully automated clustering by accurate non-parametric density estimation', *submitted* .

Deuflhard, P. and Weber, M. [2005], 'Robust Perron cluster analysis in conformation dynamics', *Linear Algebra and its Applications* **398**, 161–184.

Dewey, T. and Turner, D. H. [1979], 'Laser temperature-jump study of stacking in adenylic acid polymers', *Biochemistry* **18**(26), 5757–5762.

Di Palma, F., Bottaro, S. and Bussi, G. [2015], 'Kissing loop interaction in adenine riboswitch: insights from umbrella sampling simulations', *BMC bioinformatics* **16**(9), 1.

Di Palma, F., Colizzi, F. and Bussi, G. [2013], 'Ligand-induced stabilization of the aptamer terminal helix in the *add* adenine riboswitch', *RNA* **19**(11), 1517–1524.

Ezra, F. S., Lee, C.-H., Kondo, N. S., Danyluk, S. S. and Sarma, R. H. [1977], 'Conformational properties of purine-pyrimidine and pyrimidine-purine dinucleoside monophosphates', *Biochemistry* **16**(9), 1977–1987.

Facco, E. and Laio, A. [2016], 'A two nearest-neighbours estimator for intrinsic dimensionality', *in preparation* .

Faradjian, A. K. and Elber, R. [2004], 'Computing time scales from reaction coordinates by milestoning', *The Journal of Chemical Physics* **120**(23), 10880–10889.

Fiorucci, S. and Zacharias, M. [2010], 'Binding site prediction and improved scoring during flexible protein–protein docking with ATTRACT', *Proteins: Structure, Function, and Bioinformatics* **78**(15), 3131–3139.

Fogolari, F., Corazza, A., Fortuna, S., Soler, M. A., VanSchouwen, B., Brancolini, G., Corni, S., Melacini, G. and Esposito, G. [2015], 'Distance-based configurational entropy of proteins from molecular dynamics simulations', *PLoS one* **10**(7), e0132356.

Fuglebakk, E., Reuter, N. and Hinsen, K. [2013], 'Evaluation of protein elastic network models based on an analysis of collective motions', *Journal of Chemical Theory and Computation* **9**(12), 5618–5628.

Fulle, S. and Gohlke, H. [2008], 'Analyzing the flexibility of RNA structures by constraint counting', *Biophysical Journal* **94**(11), 4202–4219.

Gendron, P., Lemieux, S. and Major, F. [2001], 'Quantitative analysis of nucleic acid three-dimensional structures', *Journal of Molecular Biology* **308**(5), 919–936.

Gil-Ley, A., Bottaro, S. and Bussi, G. [2016], 'Empirical corrections to the amber RNA force field with target metadynamics', *Journal of Chemical Theory and Computation* **12**(6), 2790–2798.

Gong, Z., Zhao, Y., Chen, C. and Xiao, Y. [2011], 'Role of ligand binding in structural organization of add A-riboswitch aptamer: A molecular dynamics simulation', *Journal of Biomolecular Structure and Dynamics* **29**(2), 403–416.

Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H. and Weeks, K. M. [2013], 'Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots', *Proceedings of the National Academy of Sciences of the United States of America* **110**(14), 5498–5503.

Halle, B. [2002], 'Flexibility and packing in proteins', *Proceedings of the National Academy of Sciences of the United States of America* **99**(3), 1274–1279.

Hensen, U., Gräter, F. and Henchman, R. H. [2014], 'Macromolecular entropy can be accurately computed from force', *Journal of Chemical Theory and Computation* **10**(11), 4777–4781.

Hess, B., Bekker, H., Berendsen, H. J. and Fraaije, J. G. [1997], 'LINCS: a linear constraint solver for molecular simulations', *Journal of Computational Chemistry* **18**(12), 1463–1472.

Hinsen, K. [1998], 'Analysis of domain motions by approximate normal mode calculations.', *Proteins* **33**(3), 417–429.

Hnizdo, V., Darian, E., Fedorowicz, A., Demchuk, E., Li, S. and Singh, H. [2007], 'Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules', *Journal of Computational Chemistry* **28**(3), 655–668.

Hnizdo, V., Tan, J., Killian, B. J. and Gilson, M. K. [2008], 'Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods', *Journal of Computational Chemistry* **29**(10), 1605–1614.

Hobza, P. and Šponer, J. [1999], 'Structure, energetics, and dynamics of the nucleic acid base pairs: nonempirical ab initio calculations', *Chemical Reviews* **99**(11), 3247–3276.

Hohmann, A. and Deuflhard, P. [2012], *Numerical analysis in modern scientific computing: an introduction*, Vol. 43, Springer Science & Business Media.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. and Simmerling, C. [2006], 'Comparison of multiple Amber force fields and development of improved protein backbone parameters', *Proteins* **65**(3), 712–725.

Hyeon, C. and Thirumalai, D. [2005], 'Mechanical unfolding of RNA hairpins', *Proceedings of the National Academy of Sciences of the United States of America* **102**(19), 6789–6794.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. and Klein, M. L. [1983], 'Comparison of simple potential functions for simulating liquid water', *The Journal of Chemical Physics* **79**(2), 926–935.

Joung, I. S. and Cheatham III, T. E. [2008], 'Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations', *The Journal of Physical Chemistry B* **112**(30), 9020–9041.

Kabsch, W. [1976], 'A solution for the best rotation to relate two sets of vectors', *Acta Crystall. A-Crys.* **32**(5), 922–923.

Kirmizialtin, S., Hennelly, S. P., Schug, A., Onuchic, J. N. and Sanbonmatsu, K. Y. [2015], 'Integrating molecular dynamics simulations with chemical probing experiments using SHAPE-FIT', *Methods in Enzymology* **553**, 215–234.

Kührovaá, P., Banáš, P., Best, R. B., Šponer, J. and Otyepka, M. [2013], 'Computer folding of RNA tetraloops? are we there yet?', *Journal of Chemical Theory and Computation* **9**(4), 2115–2125.

Kurkcuoglu, O., Kurkcuoglu, Z., Doruker, P. and Jernigan, R. L. [2009], 'Collective dynamics of the ribosomal tunnel revealed by elastic network modeling', *Proteins* **75**(4), 837–845.

Lee, C.-H. [1983], 'Conformational studies of 13 trinucleoside bisphosphates by 360-MHz 1H-NMR spectroscopy', *European Journal of Biochemistry* **137**(1-2), 347–356.

Lee, C.-H., Ezra, F. S., Kondo, N. S., Sarma, R. H. and Danyluk, S. S. [1976], 'Conformational properties of dinucleoside monophosphates in solution: dipurines and dipyrimidines', *Biochemistry* **15**(16), 3627–3639.

Lee, C.-H. and Tinoco, I. [1980], 'Conformation studies of 13 trinucleoside diphosphates by 360 mhz pmr spectroscopy. a bulged base conformation: I. base protons and H1′ protons', *Biophysical Chemistry* **11**(2), 283–294.

Lemay, J.-F., Desnoyers, G., Blouin, S., Heppell, B., Bastet, L., St-Pierre, P., Massé, E. and Lafontaine, D. A. [2011], 'Comparative study between transcriptionally-and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms', *PLoS Genetics* **7**(1), e1001278.

Liu, J. D., Zhao, L. and Xia, T. [2008], 'The dynamic structural basis of differential enhancement of conformational stability by 5′-and 3′-dangling ends in RNA', *Biochemistry* **47**(22), 5962–5975.

MacQueen, J. [1967], Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA., pp. 281–297.

McGibbon, R. T. and Pande, V. S. [2015], 'Variational cross-validation of slow dynamical modes in molecular kinetics', *The Journal of Chemical Physics* **142**(12), 124105.

McGinnis, J. L., Dunkle, J. A., Cate, J. H. and Weeks, K. M. [2012], 'The mechanisms of RNA SHAPE chemistry', *Journal of the American Chemical Society* **134**(15), 6617–6624.

Merino, E. J., Wilkinson, K. A., Coughlan, J. L. and Weeks, K. M. [2005], 'RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (shape)', *Journal of the American Chemical Society* **127**(12), 4223–4231.

Micheletti, C. [2013], 'Comparing proteins by their internal dynamics: Exploring structure–function relationships beyond static structural alignments', *Phys. Life Rev.* **10**(1), 1–26.

Micheletti, C., Carloni, P. and Maritan, A. [2004], 'Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models', *Proteins* **55**(3), 635–645.

Mohan, S., Hsiao, C., VanDeusen, H., Gallagher, R., Krohn, E., Kalahar, B., Wartell, R. M. and Williams, L. D. [2009], 'Mechanism of RNA double helix-propagation at atomic resolution', *The Journal of Physical Chemistry B* **113**(9), 2614–2623.

Molgedey, L. and Schuster, H. G. [1994], 'Separation of a mixture of independent signals using time delayed correlations', *Physical Review Letters* **72**(23), 3634.

Morris, K. V. and Mattick, J. S. [2014], 'The rise of regulatory RNA', *Nature Reviews. Genetics* **15**(6), 423.

Musiani, F., Rossetti, G., Capece, L., Gerger, T. M., Micheletti, C., Varani, G. and Carloni, P. [2014], 'Molecular dynamics simulations identify time scale of conformational changes responsible for conformational selection in molecular recognition of HIV-1 transactivation responsive RNA', *Journal of the American Chemical Society* **136**(44), 15631–15637.

Mustoe, A. M., Brooks, C. L. and Al-Hashimi, H. M. [2014], 'Hierarchy of RNA functional dynamics', *Annual Review of Biochemistry* **83**, 441–466.

Noé, F. and Clementi, C. [2015], 'Kinetic distance and kinetic maps from molecular dynamics simulation', *Journal of Chemical Theory and Computation* **11**(10), 5002–5011.

Noé, F., Doose, S., Daidone, I., Löllmann, M., Sauer, M., Chodera, J. D. and Smith, J. C. [2011], 'Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments', *Proceedings of the National Academy of Sciences of the United States of America* **108**(12), 4822–4827.

Noé, F. and Fischer, S. [2008], 'Transition networks for modeling the kinetics of conformational change in macromolecules', *Current Opinion in Structural Biology* **18**(2), 154–162.

Noé, F. and Nuske, F. [2013], 'A variational approach to modeling slow processes in stochastic dynamical systems', *Multiscale Modeling and Simulation* **11**(2), 635–655.

Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. and Weikl, T. R. [2009], 'Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations', *Proceedings of the National Academy of Sciences of the United States of America* **106**(45), 19011–19016.

Noé, F., Wu, H., Prinz, J.-H. and Plattner, N. [2013], 'Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules', *The Journal of Chemical Physics* **139**(18), 184114.

Nüske, F., Keller, B. G., Pérez-Hernández, G., Mey, A. S. and Noé, F. [2014], 'Variational approach to molecular kinetics', *Journal of Chemical Theory and Computation* **10**(4), 1739–1752.

Olsthoorn, C. S., Doornbos, J., Leeuw, H. P. and Altona, C. [1982], 'Influence of the 2′-hydroxyl group and of 6-N-methylation on the conformation of adenine dinucleoside monophosphates in solution', *European Journal of Biochemistry* **125**(2), 367–382.

Pan, F., Roland, C. and Sagui, C. [2014], 'Ion distributions around left-and right-handed DNA and RNA duplexes: a comparative study', *Nucleic Acids Research* **42**(22), 13981–13996.

Pande, V. S., Beauchamp, K. and Bowman, G. R. [2010], 'Everything you wanted to know about Markov state models but were afraid to ask', *Methods* **52**(1), 99–105.

Parrinello, M. and Rahman, A. [1981], 'Polymorphic transitions in single crystals: A new molecular dynamics method', *Journal of Applied Physics* **52**(12), 7182–7190.

Pérez, A., Marchán, I., Svozil, D., Šponer, J., Cheatham III, T. E., Laughton, C. A. and Orozco, M. [2007], 'Refinement of the AMBER force field for nucleic acids: Improving the description of α γ conformers', *Biophysical Journal* **92**(11), 3817–3829.

Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. and Noé, F. [2013], 'Identification of slow molecular order parameters for Markov model construction', *The Journal of Chemical Physics* **139**(1), 015102.

Pinamonti, G., Bottaro, S., Micheletti, C. and Bussi, G. [2015], 'Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments', *Nucleic Acids Research* **43**(15), 7260–7269.

Pinamonti, G., Zhao, J., Condon, D. E., Paul, F., Noé, F., Turner, D. H. and Bussi, G. [2016], 'Predicting the kinetics of RNA oligonucleotides using Markov state models', *submitted* .

Pörschke, D. [1976], 'Cable temperature jump apparatus with improved sensitivity and time resolution', *Review of Scientific Instruments* **47**(11), 1363–1365.

Pörschke, D. [1978], 'Molecular states in single-stranded adenylate chains by relaxation analysis', *Biopolymers* **17**(2), 315–323.

Prinz, J.-H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schütte, C. and Noé, F. [2011], 'Markov models of molecular kinetics: Generation and validation', *The Journal of Chemical Physics* **134**(17), 174105.

Priyakumar, U. D. and MacKerell, A. D. [2010], 'Role of the adenine ligand on the stabilization of the secondary and tertiary interactions in the adenine riboswitch', *Journal of Molecular Biology* **396**(5), 1422–1438.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B. and Lindhal, E. [2013], 'GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit', *Bioinformatics* **29**(7), 845–854.

Röblitz, S. and Weber, M. [2013], 'Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification', *Advances in Data Analysis and Classification* **7**(2), 147–179.

Rodriguez, A. and Laio, A. [2014], 'Clustering by fast search and find of density peaks', *Science* **344**(6191), 1492–1496.

Sarich, M., Banisch, R., Hartmann, C. and Schütte, C. [2013], 'Markov state models for rare events in molecular dynamics', *Entropy* **16**(1), 258–286.

Scherer, M. K., Trendelkamp-Schroer, B., Paul, F., Perez-Hernandez, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J.-H. and Noé, F. [2015], 'PyEMMA 2: A software package for estimation, validation, and analysis of Markov models', *Journal of Chemical Theory and Computation* **11**(11), 5525–5542.

Schölkopf, B., Smola, A. and Müller, K.-R. [1997], *Kernel principal component analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 583–588.
URL: *http://dx.doi.org/10.1007/BFb0020217*

Schütte, C., Fischer, A., Huisinga, W. and Deuflhard, P. [1999], 'A direct approach to conformational dynamics based on hybrid Monte Carlo', *Journal of Computational Physics* **151**(1), 146–168.

Schütte, C., Noé, F., Lu, J., Sarich, M. and Vanden-Eijnden, E. [2011], 'Markov state models based on milestoning', *The Journal of Chemical Physics* **134**(20), 204105.

Schwantes, C. R. and Pande, V. S. [2013], 'Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9', *Journal of Chemical Theory and Computation* **9**(4), 2000–2009.

Scott, W. G., Murray, J. B., Arnold, J. R., Stoddard, B. L. and Klug, A. [1996], 'Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme', *Science* **274**(5295), 2065–2069.

Serganov, A., Yuan, Y.-R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R. and Patel, D. J. [2004], 'Structural basis for discriminative regulation of gene expression by adenine-and guanine-sensing mRNAs', *Chemistry and Biology* **11**(12), 1729–1741.

Setny, P. and Zacharias, M. [2013], 'Elastic network models of nucleic acids flexibility', *Journal of Chemical Theory and Computation* **9**(12), 5460–5470.

Silverman, B. W. [1986], *Density estimation for statistics and data analysis*, Vol. 26, CRC press.

Sokoloski, J. E., Godfrey, S. A., Dombrowski, S. E. and Bevilacqua, P. C. [2011], 'Prevalence of *syn* nucleobases in the active sites of functional RNAs', *RNA* **17**(10), 1775–1787.

Soukup, G. A. and Breaker, R. R. [1999], 'Relationship between internucleotide linkage geometry and the stability of RNA', *RNA* **5**(10), 1308–1325.

Špačková, N. and Šponer, J. [2006], 'Molecular dynamics simulations of sarcin–ricin rRNA motif', *Nucleic Acids Research* **34**(2), 697–708.

Šponer, J., Banáš, P., Jurečka, P., Zgarbová, M., Kührová, P., Havrila, M., Krepl, M., Stadlbauer, P. and Otyepka, M. [2014], 'Molecular dynamics simulations of nucleic acids. from tetranucleotides to the ribosome', *Journal of Physical Chemistry Letters* **5**(10), 1771–1782.

Swope, W. C., Pitera, J. W. and Suits, F. [2004], 'Describing protein folding kinetics by molecular dynamics simulations. 1. theory', *The Journal of Physical Chemistry B* **108**(21), 6571–6581.

Tama, F., Valle, M., Frank, J. and Brooks, C. L. [2003], 'Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy', *Proceedings of the National Academy of Sciences of the United States of America* **100**(16), 9319–9323.

Tirion, M. M. [1996], 'Large amplitude elastic motions in proteins from a single-parameter, atomic analysis', *Physical Review Letters* **77**(9), 1905.

Trendelkamp-Schroer, B., Wu, H., Paul, F. and Noé, F. [2015], 'Estimation and uncertainty of reversible Markov models', *The Journal of Chemical Physics* **143**(17), 174101.

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. and Bussi, G. [2014], 'PLUMED 2: New feathers for an old bird', *Computer Physics Communications* **185**(2), 604–613.

Tubbs, J. D., Condon, D. E., Kennedy, S. D., Hauser, M., Bevilacqua, P. C. and Turner, D. H. [2013], 'The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics', *Biochemistry* **52**(6), 996–1010.

Turner, D. H., Sugimoto, N. and Freier, S. M. [1988], 'RNA structure prediction', *Annual Review of Biophysics and Biophysical Chemistry* **17**(1), 167–192.

Van Wynsberghe, A. W. and Cui, Q. [2005], 'Comparison of mode analyses at different resolutions applied to nucleic acid systems', *Biophysical Journal* **89**(5), 2939–2949.

Virtanen, J. J., Sosnick, T. R. and Freed, K. F. [2014], 'Ionic strength independence of charge distributions in solvation of biomolecules', *The Journal of Chemical Physics* **141**(22), 22D503.

Vokacova, Z., Budesínský, M., Rosenberg, I., Schneider, B., Šponer, J. and Sychrovsky, V. [2009], 'Structure and dynamics of the ApA, ApC, CpA, and CpC RNA dinucleoside monophosphates resolved with NMR scalar spin-spin couplings', *The Journal of Physical Chemistry B* **113**(4), 1182–1191.

Wang, Y. and Jernigan, R. L. [2005], 'Comparison of tRNA motions in the free and ribosomal bound structures', *Biophysical Journal* **89**(5), 3399–3409.

Wang, Y., Rader, A., Bahar, I. and Jernigan, R. L. [2004], 'Global ribosome motions revealed with elastic network model', *Journal of Structural Biology* **147**(3), 302–314.

Weeks, K. M. and Mauger, D. M. [2011], 'Exploring RNA structural codes with SHAPE chemistry', *Accounts of Chemical Research* **44**(12), 1280–1291.

Wilkinson, K. A., Merino, E. J. and Weeks, K. M. [2006], 'Selective 2ʹ-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution', *Nature protocols* **1**(3), 1610–1616.

Yang, L.-W., Rader, A., Liu, X., Jursa, C. J., Chen, S. C., Karimi, H. A. and Bahar, I. [2006], 'oGNM: online computation of structural dynamics using the Gaussian network model', *Nucleic Acids Research* **34**(suppl 2), W24–W31.

Yildirim, I., Park, H., Disney, M. D. and Schatz, G. C. [2013], 'A dynamic structural model of expanded RNA CAG repeats: a refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations', *Journal of the American Chemical Society* **135**(9), 3528–3538.

Yildirim, I., Stern, H. A., Tubbs, J. D., Kennedy, S. D. and Turner, D. H. [2011], 'Benchmarking AMBER force fields for RNA: comparisons to NMR spectra for single-stranded r(GACC) are improved by revised χ torsions', *The Journal of Physical Chemistry B* **115**(29), 9261–9270.

Zamuner, S. [2015], Local sampling and statistical potentials for scoring protein structures, PhD thesis, University of Padova. `http://paduaresearch.cab.unipd.it/7915/1/stefano_zamuner_thesis.pdf`.

Zen, A., Carnevale, V., Lesk, A. M. and Micheletti, C. [2008], 'Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families', *Protein Science* **17**(5), 918–929.

Zgarbová, M., Otyepka, M., Šponer, J., Lankas, F. and Jurečka, P. [2014], 'Base pair fraying in molecular dynamics simulations of DNA and RNA', *Journal of Chemical Theory and Computation* **10**(8), 3177–3189.

Zimmermann, M. T. and Jernigan, R. L. [2014], ‘Elastic network models capture the motions apparent within ensembles of RNA structures’, *RNA* **20**(6), 792–804.

Zimmermann, M. T., Leelananda, S. P., Kloczkowski, A. and Jernigan, R. L. [2012], ‘Combining statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses’, *The Journal of Physical Chemistry B* **116**(23), 6725–6731.