Geometric and statistical methods for the analysis and prediction of structural interactions between biomolecules

Julie Bernauer

▶ **To cite this version:**

Julie Bernauer. Geometric and statistical methods for the analysis and prediction of structural interactions between biomolecules. Bioinformatics [q-bio.QM]. Université Paris-Sud XI, 2015. tel-01136261

HAL Id: tel-01136261

https://theses.hal.science/tel-01136261v1

Submitted on 26 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Méthodes géométriques et statistiques pour l'analyse et la prédiction des interactions structurales de biomolécules

Habilitation à Diriger des Recherches

(Spécialité Informatique)

Université Paris-Sud 11

présentée et soutenue publiquement le 13 janvier 2015

Julie Bernauer

Président :	Philippe Dague	Professeur, Université Paris Sud, LRI/LaDHAC
Rapporteurs :	Patrice Koehl	Professor, University of California, Davis, Department of Computer Science
	Erik Lindahl	Professor, KTH Royal Institute of Technology & Stockholm University, Department of Biochemistry and Biophysics
	Dave Ritchie	Directeur de Recherche, Inria Nancy Grand Est, LORIA/Orpailleur
Examinateurs :	Johanne Cohen	Chargée de Recherche CNRS, HDR, LRI/Galac
	Erik Lindahl	Professor, KTH Royal Institute of Technology & Stockholm University, Department of Biochemistry and Biophysics
	Dave Ritchie	Directeur de Recherche, Inria Nancy Grand Est, LORIA/Orpailleur
	Céline Rouveirol	Professeure, Université Paris Nord, LIPN
	Jean-Marc Steyaert	Professeur, École Polytechnique, LIX/BioInfo
	Alain Viari	Directeur de Recherche, Directeur Scientifique Adjoint *Santé, Biologie et Planète*, Inria
Marraine :	Christine Froidevaux	Professeure, Université Paris Sud, LRI/BioInfo

Équipe projet AMIB
Inria Saclay Île-de-France, LIX École Polytechnique – CNRS UMR 7161
1 rue Honoré d'Estienne d'Orves, Bâtiment Alan Turing, Campus de l'École Polytechnique
91120 Palaiseau, Cedex France
Tél.: +33 (0)1 72 92 59 00 – Fax : +33 (0)1 74 85 42 42
http://www.inria.fr/centre/saclay

Acknowledgments

My scientific research career started in 2001 when I began my Ph.D. studies under the supervision of Anne Poupon at LEBS in Gif-sur-Yvette. Anne trusted me with a bioinformatics project and her support was essential to most of my career. I am also thankful to the people would trusted me at the time and made possible for a chemistry student to pursue biology studies while teaching computer science. In particular, I am grateful to Jean-Marc Steyaert, Hubert Comon and also Joël Janin. Not only did Joël welcome me to his institute but he also introduced me to the docking community. I am also extremely thankful to Herman van Tilbeurgh, the Yeast Structural Genomics group members, IBBMC and LEBS people for having created a great work environment for my Ph.D. studies. At this time, being a computer science teaching assistant at Orsay also lead me to meet people who would have an essential part in the professional choices I would make in the future. Jérôme Azé introduced me to machine learning and our collaboration has been ongoing ever since. The interactions triggered by the bioinformatics community Christine Froidevaux set up in Orsay were decisive in my pursuing research in computer science for biology.

I also had the chance to meet Michael Levitt in 2003. This life changing encounter got me to move to the US in 2006, where I spent a lot of time ever since and will certainly continue to do so. I met very valuable collaborators and friends during that time. I would especially like to thank Dahlia Weiss, Xuhui Huang, Adelene Sim, Peter Minary, Sam Flores, Tanya Raschke, Bick Do for supporting me in doing science but also being great people to work with.

After my post-doctoral stay at Stanford, I joined Inria in 2007, where I met incredibly talented people from which I learned a lot. I am thankful to Frederic Cazals for having welcomed me and also to Sébastien Loriot, Monique Teillaud, Olivier Devillers and Sylvain Pion for being great people to interact with. I then had to move back to the Paris area and I am extremely thankful to Christine Eisenbeis, Serge Steer and Alain Viari for trusting me and for having made that transition possible. My warmest thanks go to Mireille Regnier who gave me the opportunity to join her group and the people at Inria Saclay with whom I really enjoyed working on a daily basis. I also have to thank the people from LIX and DIX who contributed to make the last four years I spent at Polytechnique a wonderful time. I also thank Henry van den Bedem for having welcomed me at SLAC and made this last year in the Bay Area a great experience. My thanks also to Johanne Cohen who was a essential collaborator and support in the last two years.

I benefited from the help of Amélie Héliou, Alice Héliou, Adelene Sim, Rasmus Fonseca and Kilian Cavalotti for the proofreading of this manuscript and I am especially grateful to them. All this would not have been possible without all of you. Thanks a million!

Julie Bernauer
November 6th, 2014.

A human must turn information into intelligence or knowledge. We've tended to forget that no computer will ever ask a new question.
Grace Hopper in *The Wit and Wisdom of Grace Hopper* by Philip Schieber, 1987.

It is a mistake to think you can solve any major problems just with potatoes.
Douglas Adams, *Life, the Universe and Everything*, 1982.

CONTENTS

Contents

CHAPTER 1

INTRODUCTION

A Context: modeling of biological structures

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. To understand the function of a biological macromolecule, their interactions should be understood from a structural viewpoint. When they interact, biomolecules adopt different conformations: their structure is distorted upon binding. Their function is also determined by the structure they adopt. Protein and nucleic acids differ from polypeptides and polynucleotides by their spatial organization. Being able to understand how those molecules adopt these specific, almost unique structures has been a big challenge for the last two decades, and will remain to be in the near future. It involves two processes: folding and docking.

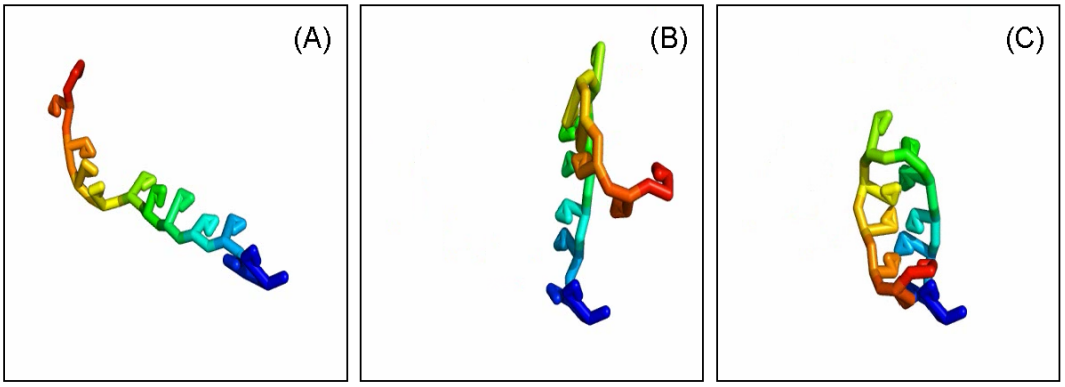


FIGURE 1.1 : Different snapshots of an in silico folding experiment using a coarse-grained (5pt, see Section 1 in Chapter 2) knowledge-based (KB) potential with base planes shown as triangles: (A) extended starting structure, (B) intermediate conformation, (C) final conformation. The final conformation shows that the base stacking handling of the potential is incorrect as the bases should be facing each other and perpendicular to the hairpin backbone.

Folding can be described as understanding the way biomolecules adopt certain configurations (see Figure 1.1). Docking is the study of how the molecules interact with each other. Ideally if both these processes were to be understood, we could predict how molecules fold, move and interact and have a strong impact not only on therapeutic studies [Zha08, Rit08] but also on nanotechnology [Guo10]. Recently, with the development of molecular systems biology aiming to integrate different levels of information, the structural study of protein and nucleic acid assemblies is even more critical. Indeed, structural analysis of biological macromolecules could provide a better understanding on the molecular processes and machinery occurring in the cell.

B Protein, RNA and complex structures

Proteins are biological macromolecules made of linear chains of amino acids bound together by peptide bonds. In physiological conditions, a protein folds to a specific compact 3D structure. This folding process depends on the interactions, not only between the different amino acids but also between the amino acids and the solvent. These interactions are still not yet well understood and the phenomenon is so complex that the folding process cannot be exhaustively described.
RNA molecules are made of linear chains of nucleic acids. Many RNA molecules are non-coding, and instead have regulatory functions in the cell. Like proteins, these structured RNA fold into specific 3D structures in order to carry out their function. The folding of RNA molecules is dependent on the interactions between its nucleic acids, the solvent and – as RNA molecules are charged – ions.

Protein and nucleic acid structures are characterized by different levels of organization (see Figure 1.2):

- the primary structure, or sequence, is the linear sequence of amino acids (encoded by a 20-letter alphabet) or nucleotides (encoded by a 4-letter alphabet);
- the secondary structure is resulting from short range interactions (mainly hydrogen bonds between atoms). For proteins, these interactions only depend on the nature of the lateral chains of the amino acids: some segments of the protein are made of amino acids showing a periodic pattern in the successive dihedral angles. For RNAs, these interactions are also dependent on the nature of the nucleic acids (and thus side-chains) and also on the type of pairing they can make;
- the tertiary structure is the functional form of a folded proteic or nucleic chain. It is the result of the arrangement of secondary structures into a specific topology/shape;
- the quaternary structure, which includes complexes, is the association of several amino acids or nucleic acids chains (identical or not).

The first protein structure (of myoglobin) was solved experimentally in 1960 by X-ray crystallography [KDS+60]. As of today, the *Protein Data Bank* (PDB) [BBB+00, BWF+00], the data bank for all biomolecular structures, contains more than 103 000 files, of which 92 000 are structures that were solved by X-ray crystallography and 10 000 were solved by Nuclear Magnetic Resonance (NMR). While the PDB contains more than 96 000 protein structures, less than 3 000 structures of nucleic acids can be found as these are much more difficult to solve experimentally.

C Experimental and computational challenges

The development of structural genomics projects in the 2000s has contributed to the rationalization of the structure solving process. The rate of structure determination has drastically increased. A decade ago,

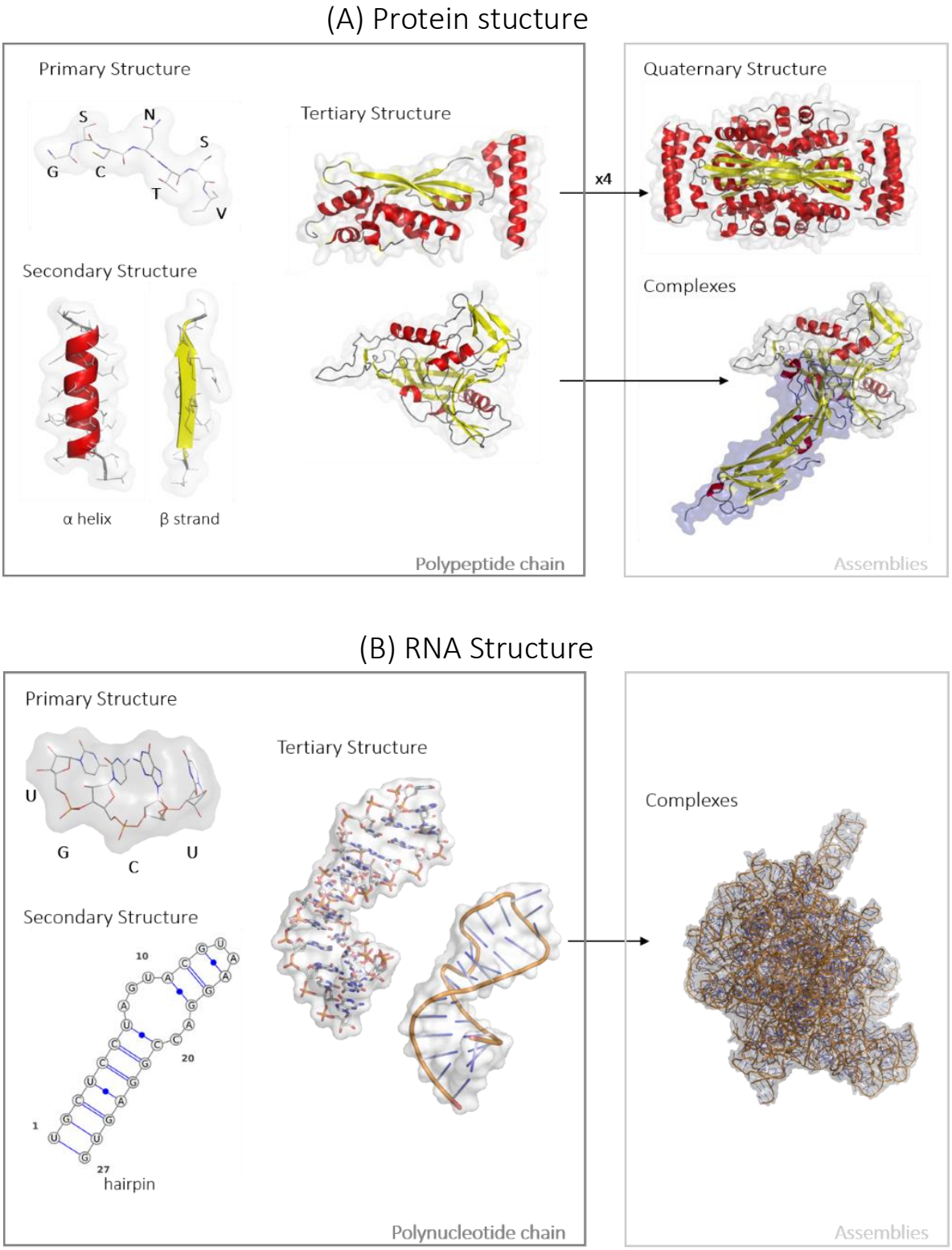


FIGURE 1.2 : Different levels of structure: (A) protein, (B) RNA.

the PDB contained 33 000 structures: 28 000 had been solved by crystallography and 5 000 by NMR. While other techniques can be used, sometimes leading to different but nonetheless interesting structural results, crystallography is still the most successful technique.

The experimental challenges and bottlenecks in solving the stucture of a biomolecule are numerous (See Figure 1.3). For crystallography, a diffracting crystal is required which is a lengthy and limiting step. NMR is limited to relatively small molecules and requires very pure sample solutions. The structural genomics initiatives have increased the yield of structure solving, for instance by providing integrated laboratory management systems [POU+05]. The cost of solving a structure has dropped [CB06], but still neither all structures can nor will be solved with current technologies. Modeling is needed more than ever and the data acquired in the last decade can be of great help.

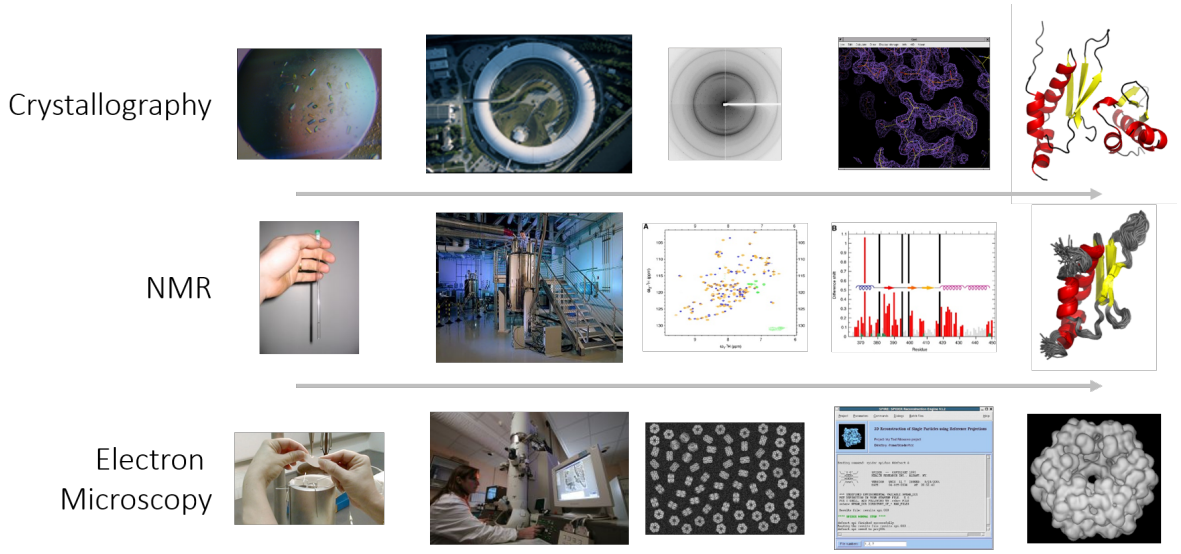


FIGURE 1.3 : Structure solving: steps for the three main techniques. From left to right for each experimental technique: sample, equipment, raw data, data treatment, result.

X-ray crystallography provides a static image of a molecule or a complex. NMR provides various conformations of the same structure. Other techniques, such as electron microscopy, provide a view of an envelope (See Figure 1.3). The complementarity of these techniques is obvious and computational structural biology is of the essence to combine and interpret the data. The algorithms and techniques to model static protein structures have been studied extensively (See [Les13, Koe10, Zha08]) and will not be detailed here.

The function of the biomolecules relies on their dynamic nature which cannot be accessed through experimental techniques in general. To address the issue of studying the dynamics of a macromolecule, one has to explore the biologically possible spatial configurations. The two most common techniques currently used in computational structural biology are Molecular Dynamics (MD) and Monte Carlo techniques (MC). In MD computer simulations, the time evolution of a set of interacting atoms is followed by integrating their equations of motion (classical mechanics) [Lin08, Lin15]. Unlike Monte Carlo techniques, it is deterministic for a given set of starting atomic velocities. MD is a statistical mechanics technique, made to obtain a set of configurations called a statistical ensemble. Physical quantities are averages over configurations in a certain statistical ensemble. MD provides such a good ensemble as a trajectory. Once a suitable configuration has been found, it can be refined, i.e. a minimization procedure is applied to get closer to the biological native structure. Those techniques require the evaluation of a

potential or force-field, which for computational biology are often empirical. They mainly consist of a summation of bonded forces associated with chemical bonds, bond angles, and bond dihedrals, and non-bonded forces associated with van der Waals forces and electrostatic charges. They are also primarily based on simple distance measurements. Moreover, the evaluation of the potential for each atom is the computation time bottleneck of those techniques. Details can be found in [DB03, Sat10, Lin08, Lin15].

To study interactions at the structural level, the classical strategy is very similar to the one for structure dynamics. Indeed, upon interaction, biomolecules get distorted and there may be some important configuration distance between the so-called *bound* and *unbound* configurations (i.e. found in the complex and in free form). Docking basically contains two parts: exploration and scoring. The exploration phase consists of generating all the possible binding conformations between two partners. A scoring phase is then applied to evaluate these conformations and find the one which is the most biologically relevant (or close enough solutions for biological applications) as shown on Figure 1.4. Exploration is performed by exhaustive geometric search with a lot of different techniques (see [VK09, Rit08] for reviews). During exploration, due to computational time constraints, a simple filtering is performed (surface area filtering or coarse-grained interaction potential evaluation). After exploration, a large scale scoring procedure may involve more expensive techniques: physics-based energy potentials are usually applied. Some newer techniques also involve both knowledge-based potentials and machine learning procedures. Once a small number of putative conformations have been selected, they should be refined in order to get correct atomic contacts when the experimental studies require a detailed knowledge of the interaction. Flexibility has to be taken into account in all these steps, and as the CAPRI experiment results have shown, this is still a major issue [LW10, JHM+03, Jan10, Rit08].

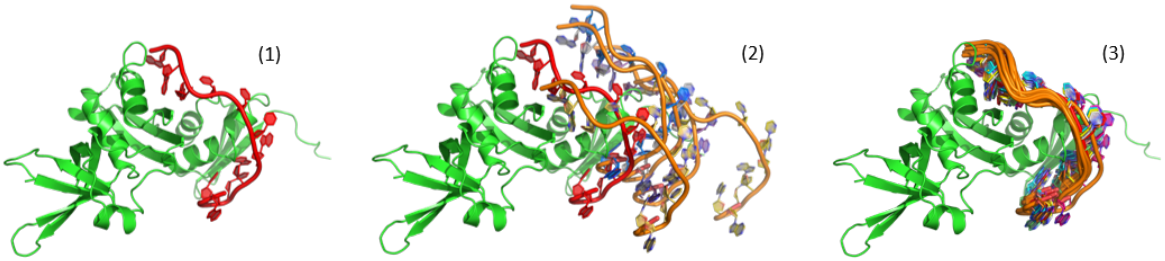


FIGURE 1.4 : Example of a docking procedure: the *Mycobacterium tuberculosis* NusA-RNA complex (PDB id 2ASB). (1) Native structure. (2) Conformations obtained by a standard Rosettadock perturbation run. (3) Close to native conformations selected by an accurate scoring function.

Simulations to address the structural dynamics or interactions of macromolecules are known to be computationally expensive. The main reason is that atomic detail is needed to accurately represent the chemical processes. While efficient algorithms can be found, the size of the biomolecules and their assemblies is often the limiting factor. Recent advances in multi-scale modeling have shown that the computational time can be greatly reduced by accessing atomic detail only when needed. Developing novel algorithms, potentials and multi-scale modeling techniques are thus the main directions for computational structural biology.

D Objective

The purpose of my research is to explore algorithms and computer science techniques for the efficient 3D modeling of biomolecules. Ideally this would lead to a generic computational framework that could be

used for therapeutic and nanotechnology applications. For that purpose, I addressed different aspects of computer science for structural biology: geometry, machine learning, statistics and robotics. While the different studies might seem opportunistic, depending on the type of molecules or applications, there is no doubt that they can be combined to address biological problems involving a wide range of molecules and molecular machineries. Each step of the modeling must focus on an optimized strategy at a specific scale as sought for in the different studies.

In the following, I will first present several coarse-grained models, including the Voronoi model for protein-protein complexes which allowed for a good description and scoring of binary interactions. I will then detail different machine learning strategies we developed for scoring biological 3D complexes, including protein-protein and protein-RNA interactions. A knowledge-based[1] scoring function for RNA molecule conformation selection will then be presented. Built from reliable statistics, this knowledge-based potential is differentiable and has been shown to hold great promises for RNA modeling. Finally, RNA dynamics modeling from a robotics-inspired technique will be detailed. This inherently multi-scale approach being very generic, it can be used to address a wide range of biological problems.

[1] In the computational structural biology community, *knowledge-based* is used for potentials or methods that are derived from statistics or measurements on data. It thus should not be confused with the classical definition of knowledge-based systems in computer science.

Contents

CHAPTER

2

GEOMETRIC MODELS AND SUPERVISED LEARNING FOR PROTEIN-PROTEIN AND PROTEIN-RNA DOCKING: THE KEY TO EFFICIENT PREDICTIONS?

Main articles from the chapter

[GGFAB14] **Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities**, PLoS One, 2014

[FBS+12] **Multiscale modeling of macromolecular biosystems**, Brief Bioinform, 2012

[BBAP11] **A collaborative filtering approach for protein-protein docking scoring functions**, PLoS One, 2011

[FWS+11] **Community-wide assessment of protein-interface modeling suggests improvements to design methodology**, J Mol Biol, 2011

[LCB10] **ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules**, Bioinformatics, 2010

[BBAP09] **Comparing Voronoi and Laguerre tessellations in the protein-protein docking context**, ISVD, 2009

[BBR+08] **DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions**, Bioinformatics, 2008

[BAJP07] **A new protein-protein docking scoring function based on interface residue properties**, Bioinformatics, 2007

[BPAJ05] **A docking analysis of the statistical physics of protein-protein recognition**, Phys Biol, 2005

A Geometric coarse-grained models and geometric constructions

1 Coarse-grained models

In 2013, the Nobel prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel for the *development of multiscale models for complex chemical systems*. This award shed light on computational structural biology and computer science in general. In particular, it emphasized the development of computational techniques and the importance of computer *dry* science for applied *wet* science, mainly chemistry and biology.

At the core of this Nobel prize are multi-scale models which appeared in the '70s [LW75, Lev76]. The principle of multi-scale models and simplified representation is that atomic accuracy and quantum mechanics are not always required to obtain good explicative and predictive models: coarse-grained representations and newtonian-inspired physics might be sufficient for most of the simulation. While we might discuss this shift from complex quantum mechanics models to classical physics later, in the light of the *data* era (see Chapter 5), these models have been extremely successful so far and cover a wide range of applications [FBS+12, Toz05]

The idea behind coarse-grained models is simple. Using a simplified representation, one can perform computations that would be intractable otherwise: the fine representation level is added at specific stages of the simulation when detailing is essential. A simple example is the five-point (*5pt*) representation for RNA (see Figure 2.1). Instead of using all the atoms for each nucleotide, only a subset of atoms is taken into consideration for the computation. The whole topology has to be redefined but this model accounts for both the location and the orientation of nucleotides (see Figure 1.1 for an experiment using this representation). Three-point (*3pt*) and one-point (*1pt*) representations are defined the same way.

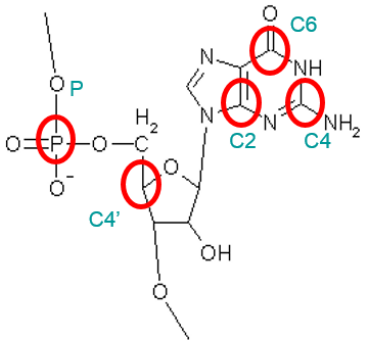


FIGURE 2.1 : Five-point (5pt) representation for RNA: example of a Guanine. Only the P, C4', C2, C4 and C6 atoms are taken into account for the computation. This reduces the computational complexity while keeping the information on the location and orientation of the nucleotide base.

I had the chance to work on several coarse-grained models: from multi-point representations (see Chapter 3) to Voronoi cells described in this section. With Samuel C. Flores, Xuhui Huang, Seokmin Shin and Ruhong Zhou, I organized two sessions at the Pacific Symposium on Biocomputing [BFH+11, FBH+10] which demonstrated a wide range of applications for multi-scale strategies and uses: protein structure modeling, prediction and dynamics; RNA modeling; docking, assembly prediction and aggregates modeling. We reviewed these applications in [FBS+12].

The algorithms and techniques developed for my research all contain coarse-grained models and/or multi-scale representations:

– for docking (this section), each residue is represented as a Voronoi cell [BAJP07, BBR+08, BPAJ05, BBAP11, BBAP09, FWS+11], or a as a combination of one point and all-atom representation [GGFAB14];
– Knowledge-based (KB) potentials use a combination of five-point (5pt), three-point (3pt) and one-point (1pt) representations [BHSL11, SSLB12] (see Chapter 3);
– our kinematics inspired models [FPBv14] is based on an all-atom representation with rigid groups involving sets of atoms (see Chapter 4).

In the following, I will detail the Voronoi diagram and its derivative, the Laguerre diagram.

2 Geometric constructions: the Voronoi and Laguerre diagrams[2]

2.1 Use cases in biology

While the first use of the Voronoi cells was described by John Snow in 1854 for modeling the cholera epidemic in London [Joh06], it was not mathematically formalized until the original article from Georgy Voronoi was published in 1908 [Vor08]. The Voronoi diagram has been widely used more recently in structural biology (see [Pou04] for a review) and for protein structure analysis and predictions in particular [ASJ+02, EFL96, GC96, HGC94, PRW96, SCM+00, SB10, MK11, KK12, LMK13, MK13]. We introduced coarse-grained Voronoi models for protein 3D interaction prediction [BAJP07, BBR+08, BPAJ05, BBAP11, FWS+11]. In our studies, a machine learning strategy is developed for the prediction but other scoring methods now also involve Voronoi models [KZT12]. In parallel, Frédéric Cazals developed atomic models of protein interface using Voronoi cells for structural analysis [LC10, BGNC09]. We collaborated on defining a software framework for the efficient construction of Voronoi and geometry based models for biomolecules upon my joining Inria in Sophia Antipolis from 2007 to 2009. ESBTL[3] (Easy Structural Biology Template Library) [LCB10] is the result of this collaboration which allows for binding of the renowned CGAL library[4]. The performance of ESBTL rendered obsolete previously available naive construction software for proteins [DSJ+05], by benefiting from the optimized CGAL implementations.

2.2 The Voronoi diagram

Let $E = \{p_1, \ldots, p_n\}$ be a set of points in $\mathbb{R}^d$ called sites. To each site p_i, is associated a Voronoi cell (region) $V(p_i)$ made of the points of $\mathbb{R}^d$ closer to p_i than any other site in E:

$$V(p_i) = \{x \in \mathbb{R}^d : \|x - p_i\| \leq \|x - p_j\|, \forall j \leq n\} \tag{2.1}$$

Let Π_{ij} be the bisecting plane[5] of p_i and p_j, and π_{ij}^i the half-spaces delimited by Π_{ij} containing p_i. $V(p_i)$ is the intersection of the half-spaces $\pi_{ij}^i, \forall j \neq i$, i.e.:

$$V(p_i) = \bigcap_{j \neq i} \pi_{ij}^i \tag{2.2}$$

[2] For this section, the reader is invited to refer to the reference book of Boissonnat, Yvinec and Brönnimann [BYB98].
[3] `http://esbtl.sf.net`
[4] `http://www.cgal.org`
[5] To avoid any confusion, in the following, we use, for objects in $\mathbb{R}^d$, the vocabulary of objects in 3D. The illustrating figures will however describe the 2D case for simplicity.

This intersection contains the point p_i and is not empty. $V(p_i)$ is a convex polyhedron, possibly not bounded. The Voronoi diagram of E is the set of all Voronoi cells and their facets (Figure 2.2a).

All points in $\mathbb{R}^d$ belong to at least one Voronoi cell: the diagram is a *partition* of $\mathbb{R}^d$. If a point belongs to $k \geq 1$ cells, it belongs to the *facet* of the diagram shared by k cells. Such a point is closest to the k sites of E than any other point.

2.3 The Delaunay triangulation

Let E be a set of n points of $\mathbb{R}^d$. A *triangulation* of E is a set of tetrahedra having for vertices (corners) the points of E satisfying:

- the intersection of two tetrahedra is either empty or a facet shared by the two tetrahedra,
- the set of the vertices of the tetrahedra coincides with E,
- the tetrahedra are a subdivision of the convex hull of E.

The power of a point relatively to a sphere σ of center c and radius r is the real number:

$$\sigma(x) = (x - c)^2 - r^2 \tag{2.3}$$

The sphere σ is defined by the set of points x such as: $\sigma(x) = 0$. A sphere σ is said to *include* a point y if the inside of the sphere (ball) contains the point, which is equivalent to $\sigma(y) < 0$.

Let E be a set of n points $p_1, \ldots, p_n$ of $\mathbb{R}^d$. The *Delaunay triangulation* of E is a triangulation of E where all tetrahedra can be circumscribed by a sphere which does not include any of the points p_i (Figure 2.2b). A *Delaunay sphere* is the circumscribed sphere of a tetrahedron of a Delaunay triangulation and a *Delaunay ball* is the ball delimited by such a sphere.

2.4 Properties

In the following, as all points are in a general position (i.e. no more than four points are cospherical and points are not colinear), the Delaunay tessellation and the Voronoi diagram are unique and the tetrahedra are not flat. The Delaunay tessellation is the dual of the Voronoi diagram (Figure 2.2c): a vertex in the Delaunay tessellation corresponds to a facet in the Voronoi diagram.

2.5 The Laguerre diagram

The so-called Laguerre diagram in the French scientific literature is also known as the power diagram [Aur87] in computational geometry. For the following, we will use Laguerre diagram as this name was mainly used in biophysics related applications.

Let $E = \sigma_1, \ldots, \sigma_n$ be a finite set of spheres in $\mathbb{R}^d$. We denote c_i the center of σ_i and r_i its radius. To each σ_i is associated the cell (region) $L(\sigma_i)$ made of the points of $\mathbb{R}^d$ whose power relatively to σ_i (see Equation 2.3) is smaller than its power to any other sphere of E:

$$L(\sigma_i) = \{x \in \mathbb{R}^d : \sigma_i(x) \leq \sigma_j(x), \forall j \leq n\} \tag{2.4}$$

The set of points having the same power relatively to two spheres σ_i and σ_j is a plane, denoted ρ_{ij} and called *radical plane* of σ_i and σ_j. ρ_{ij} is orthogonal to the line joining the centers of σ_i and σ_j. Let

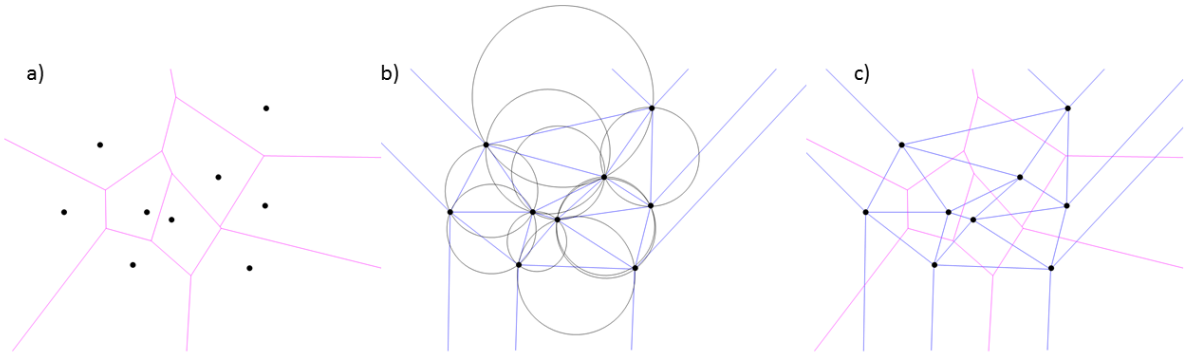


FIGURE 2.2 : The Voronoi diagram and Delaunay tessellation. a) The Voronoi diagram. b) The Delaunay triangulation showing the circumscribed circles. c) Superposition showing the Voronoi diagram is the dual of the Delaunay triangulation.

ϱ_{ij}^{i} denote the half space delimited by ρ_{ij} whose points have a smaller power relatively to σ_i than to σ_j. $L(\sigma_i)$ is the intersection of the half-spaces ϱ_{ij}^{i}, $\forall j \neq i$. If this intersection is not empty, it is a convex polyhedron possibly unbounded. Non-empty $L(\sigma_i)$ are called Laguerre cells (or regions).

The Laguerre diagram is the set of Laguerre cells of E and their facets (Figure 2.3). It is possible for a sphere σ_i not to be fully contained in its Laguerre cells.

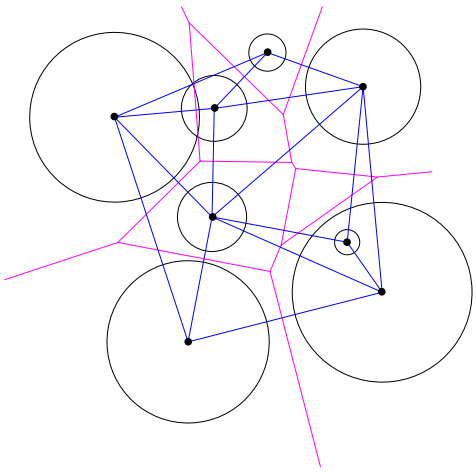


FIGURE 2.3 : Laguerre diagram (pink) and its dual, the regular triangulation (blue). On this drawing, one of the spheres (the smallest) is not located in its Laguerre cell. It can also happen that a Laguerre cell associated to a sphere "does not exist" .

When all sphere radii are equal, the Laguerre diagram of the spheres is the same as the Voronoi diagram of their centers. Exactly like in the case of the Voronoi diagram, a dual triangulation can be defined for the Laguerre diagram. This triangulation is called the regular triangulation (Figure 2.3).

2.6 Voronoi and Laguerre cells constructions for amino and nucleic acids

The Voronoi or Laguerre diagram of a set of sites can be seen as a partition of the space into *influence zones* of these sites: the cells. There are many ways to construct the Voronoi and Laguerre diagrams for proteins. Our coarse-grained model used one site per residue or nucleotide. Well chosen, this site can accommodate for the side-chain to be mainly located inside the cell and accurately represent the packing. Figure 2.4 illustrates such a cell for phenylalanine residue in a complex.

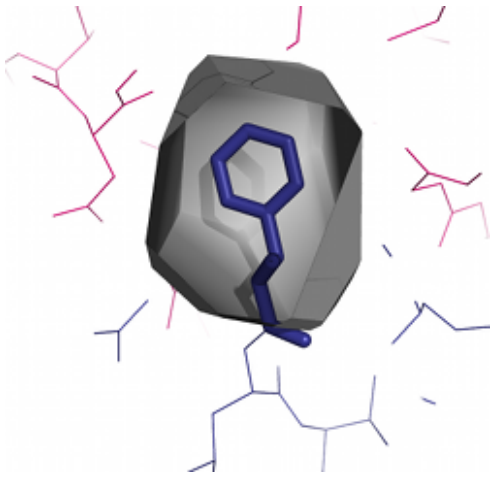


FIGURE 2.4 : Voronoi cell of an amino acid in a protein complex structure. The side chain of the phenylalanine (shown in stick representation) is located inside the cell for a one site per residue model.

Many descriptors of the structure can be derived from such a construction, including the distance between two Voronoi (Laguerre) neighbors, i.e. sites sharing a Voronoi (Laguerre) facet. In our docking studies, these distances and the volumes of the Voronoi cells were shown to be great descriptors for proteins [BAJP07, BBR+08, BPAJ05]. While both the Voronoi and Laguerre constructions lead to different measurements, their performance for prediction in docking are similar [BBAP09]. Defining the same type of model for nucleotides has appeared to be slightly harder [GG14] but is promising. Using multiple sites per nucleotide might help accommodating for large side chain movements, including sugar puckering (described in Chapter 4) for example.

B Predicting protein-protein complexes: supervised learning for docking

1 The docking problem

Knowing the structure of two putative protein partners (or having accurate enough models), the docking problem aims at predicting the structure of the complex (see Figure 2.5). Various algorithms and techniques have been used to perform exploration by exhaustive geometric search such as: grid searches, Fourier correlation techniques, random searches, spherical harmonics, geometric hashing, shape complementarity etc. On the millions of putative configurations generated, the scoring procedure is applied. Exploration and scoring are intertwined steps and usually a simple scoring is applied at the exploration stage for selection and a detailed scoring scheme is applied after the generation. Once several suitable configurations have been selected, an additionnal refinement scheme is performed [JHM+03, LW10, Rit08].

The exploration step is inherently computationally expensive: when no biological information on the

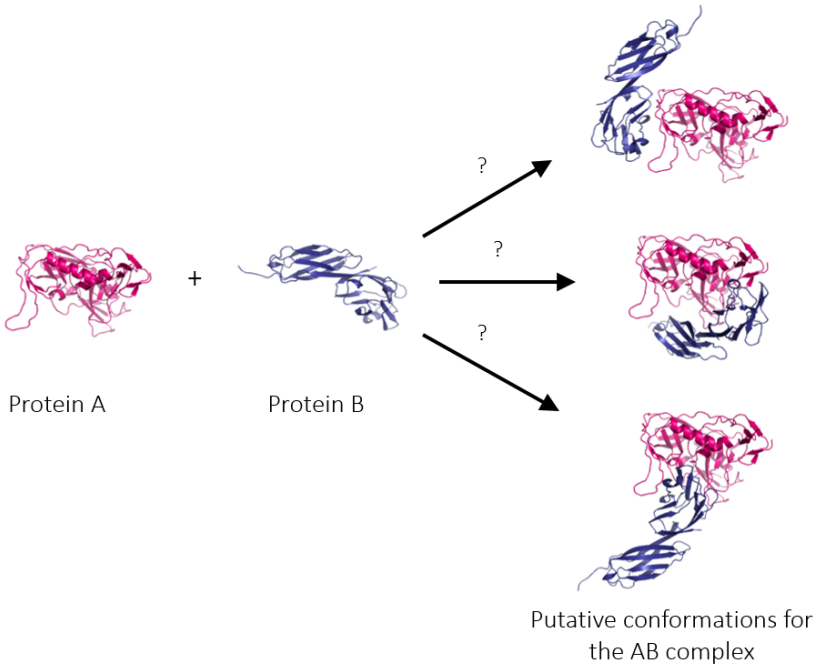


FIGURE 2.5 : Example of a docking problem: predicting the structure of protein A, the HIV envelope glycoprotein gp120 (pink) to protein B, the CD4 receptor (blue). The exploration step will provide a large number of conformations that will further be scored.

interacting regions, called *epitopes*, is known, several millions of conformations have to be generated. Most of the docking software resort to coarse-grained models for the first tentative stage, in the same way we developed a docking strategy based on Voronoi models. Exploration and scoring phases are performed iteratively from a coarse-grained level to atomic accuracy. Initiated during my Ph.D., this work on protein-protein complexes was further pursued in the Ph.D. thesis of T.Bourquard [Bou09]. We recently extended it to protein-nucleic acid complexes during the Ph.D. of Adrien Guilhot [GG14]. While Voronoi models can be used for exploration, we mainly focused on the coarse-grained scoring step which was fast and reliable enough to obtain good results in the CAPRI context (see C).

From a computer science perspective, scoring can be seen as "fishing out" a set of good *near-native* structures from a large number of *decoys*, making this problem a perfect case for prediction by supervised learning techniques.

2 Supervised learning and prediction[6]

2.1 Principle

Inferring a function from labeled training data in machine learning is called supervised learning. Knowing a set of labeled examples, the purpose is to best predict the label of new unseen instances.

Let $\mathcal{X}$ be the set of *examples* (or data) and $\mathcal{Y}$ the set of *labels* (also called *classes*) that can be associated with the examples. In the research presented here, only binary labels were used $\mathcal{Y} = \{+1, -1\}$

[6]For this section, the reader is invited to refer to the reference books of Hastie, Tibshirani and Friedman [HTF09] and/or Cornuéjols and Miclet [CM11] .

(also written $\{+, -\}$).

Data can be of two types:

- labeled data. In general very few data of this kind are available, as knowing the label of each piece of data is often difficult and/or expensive (e.g.: obtaining the crystal structure of a protein-protein complex). Such data are used to learn a model to predict the labels of the new unseen examples;
- unlabeled data. In general easy to obtain.

The set of labeled data is called the *training set* denoted $\mathcal{A} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $\boldsymbol{x}_i \in \mathbb{R}$ and $y_i \in \mathcal{Y}, \forall i \in \{1, \ldots, n\}$. $\boldsymbol{x}$ is a vector of dimension d where each dimension is one of the *features* of the example $\boldsymbol{x}$.

A whole *learning* procedure usually consists in three stages:

- learning a model so as to best predict the training data;
- evaluating the model on a subset of data extracted from the training set;
- testing the model on a dataset not contained in the training set.

For the evaluation of the model performance not to be biased, the data should not have been already used for the training. Two techniques are usually used: *cross-validation* and *leave-one-out*.

Cross-validation (CV) is performed by partitioning the training set $\mathcal{A}$ in k non-overlapping parts. The training is performed on the union of $k-1$ parts and the evaluation is done on the remaining part. The process is repeated k times so that all the examples in $\mathcal{A}$ are used once for the test and $k-1$ times for learning. The choice of k depends on the size of the dataset. $k = 3$ or $k = 10$ are common values.

Leave-one-out (LOO) evaluation is a generalization of cross-validation with $k = n$. For each example, a model is learned from the whole dataset from which the test example has been removed. This protocol is mainly used when few data are available or when CV would lead to remove too many data from the training set. When having a large dataset, the computational cost of this technique is very high as n models have to be learnt.

A wide range of supervised learning algorithms is available but no single learning algorithm works best on all supervised learning problems (this is often called the *No free lunch theorem*). Several issues have to be considered in supervised learning, including the famous *bias–variance dilemma*. The reader is invited to refer to [HTF09] and/or [CM11] for details.

2.2 Evaluation criteria

Upon applying a predictive model on a dataset, for each label, one can measure the number of examples correctly associated to their label and the number of examples incorrectly associated to their label. This information is provided by the *confusion matrix*. For binary classification, the confusion matrix can be found in Figure 2.6.

Global criteria From the confusion matrix, many evaluation criteria can be computed. Some are detailed below.

- The *precision* or *Positive Predictive Value (PPV)* $P = \frac{TP}{TP+FP}$ represents the percentage of correct predictions associated to the positive class (it can also be defined for the negative class).
- The *recall* $R = \frac{TP}{TP+FN}$ represents the percentage of positive examples correctly predicted as positive (exactly like precision, it can be defined for the negative class).

		Actual	
		+	-
Predicted	+	TP	FP
	-	FN	TN

FIGURE 2.6 : Confusion matrix with TP True Positives (hit), FP False Positives (false alarm), FN False Negatives (miss) and TN True Negatives (correct rejection). This definition for the binary case (two classes) can be extended for n classes (False Positives and False Negatives definitions being extended too).

- The $\boldsymbol{F}_{score}(\beta) = \frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$ aggregates in one measure precision and recall. The β parameter allows for weighting the precision relatively to the recall. When $\beta < 1$, precision is more important; when $\beta > 1$, recall is. $\beta = 1$ gives equals importance to precision and recall: β is often set to 1.
- The *accuracy* $Acc = \frac{TP+TN}{TP+TN+FP+FN}$ is used to evaluate the global performance of a classifier by reporting the percentage of correct predictions independently of the class.
- The *sensivity* or *True Positive Rate (TPR)* $Se = \frac{TP}{TP+FN}$ is equal to the recall of the positive class examples. This measure originating from signal processing is widely used in biomedical applications.
- The *specificity* or *True Negative Rate* $Sp = \frac{TN}{FP+TN}$ is the recall of the negative class examples. Also originated from signal processing, it is used in biomedical applications in combination with the sensitivity for detection tests (using $1 - Sp$).

Other criteria are also widely used for medical applications and tests, often in the light of the prevalence of a condition, such as the *False Discovery Rate (FDR)* or the *False Omission Rate (FOR)*. All these measures provide a global evaluation of the performance of a classifier, as a single value is used to assess the behavior on the whole dataset.

Local criteria Using only the precision and recall (or other global criteria) to assess classifier performance is very limited when classifiers have to be compared or assessed for each of their prediction. Novel measures were defined in the 2000s to overcome this shortcoming [LHZ03b, FF03], in particular to:

- compare different classifiers or assess different settings of the same classifier,
- evaluate classifiers by providing a score to each prediction. This score, that can be seen as a confidence level for the prediction, allows for the ordering of the predictions and provides more information than a label.

The recall vs. precision curve is an example of the first category as it can be used to easily compare different classifiers.

In our studies, we focused on the second category, so as to obtain a quantitative value for the confidence associated to each prediction or a direct probability for an example to belong to a class. For this purpose, we used the *Receiver Operating Characteristic (ROC)* curve [Met78, VC06] to visualize the sensitivity/specificity trade-off. Figure 2.7 shows a ROC curve we obtained in [BAJP07] for the protein

docking problem.

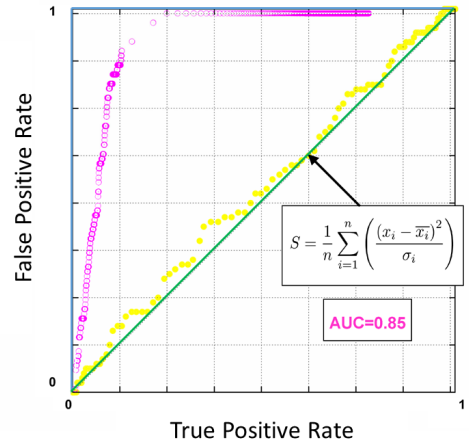


FIGURE 2.7 : Example of a ROC curve. For an ideal classifier, the ROC curve is made of two line segments: (0,0) to (0,1) for the perfectly ordered positive examples and (0,1) to (1,1) for the negative examples whose scores are all lower than the scores of the positive examples (blue). A non-discriminating classifier will be close to the diagonal (green). The ROC curve shows the results we obtained for two docking scoring functions: a simple function by using the square deviation which shows poor performance (yellow) and a better one using a genetic algorithm (magenta).

Upon visual inspection of the ROC curves, it is easy to assess the relative performance of one or many classifiers. To numerically compare several classifiers, so as to find the "best" one, the *Area Under the ROC curve (AUC)* can be used. It was shown to be a better measure than accuracy [LHZ03a, LHZ03b]. Many classifiers have been adapted to optimize the AUC, including Support Vector Machines (SVM) [Rak04] and the ROGER genetic algorithm [ALS04] that we used and describe below.

2.3 Algorithms

Logistic regression This first model is often the simplest used to test whether supervised learning can be used for a biological problem. This simple model had been widely used for bioinformatics, including for protein structure prediction [MGHT99, MGHT02]. One of the advantages for using such a model for biology is that the relative influence of each feature is known, often leading to interesting biological interpretation.

In logistic regression, the probability to observe a positive example is evaluated as:

$$P(\boldsymbol{x}) = 1/\left[1 + \exp\left(-w_o - \sum_i w_i \boldsymbol{x}_i\right)\right] \tag{2.5}$$

where w_i is the weight of the feature i of the input feature vector $\boldsymbol{x}$ and w_0 a global initial weight.

The vector of weights w is estimated by maximum likelihood on the training set. Logistic regression is a type of generalized linear model, which predicts variables with various types of probability distributions by fitting a linear predictor function to some sort of arbitrary transformation of the expected value of the variable. Logistic regression is also called perceptron, as it is equivalent to a single layer neural network. A full description of the model can be found in [HTF09].

ROGER: a genetic algorithm ROGER (ROc based Genetic learnER) is a genetic algorithm initially developed during the Ph.D. of my collaborator Jérôme Azé. The purpose of ROGER is to learn functions to order samples so as to optimize the ROC AUC corresponding to the ordering obtained. Functions have the form: $f(\boldsymbol{x}_i) = \sum_j w_j \times \boldsymbol{x}_i(j)$ where $\boldsymbol{x}_i(j)$ is the value of the j^{th} feature of the example $\boldsymbol{x}_i$. The algorithm finds weights w_j so that $\sum_i rank_f(\boldsymbol{x}_i) \times \mathbb{1}_{y_i=+1}$ is minimal where $rank_f(\boldsymbol{x}_i)$ is the rank of the example $\boldsymbol{x}_i$ provided by the function f, and $\mathbb{1}_{y_i=+1}$ is the indicator function equal to 1 when the class is $\boldsymbol{y}_i$ and 0 otherwise. A function maximizing the sum of the ranks of the positive examples also maximizes the AUC.

Aside from our studies, this algorithm was successfully used for various applications, including text mining [ARKS05a, ARKS05b] and prediction of cardio-vascular risks [SAL03, SLA03, ALS03].

Support vector machines From a set of binary labeled vectors, Support Vector Machines (SVMs) also train a classifier to be further used to label unseen examples [CST00, Sch97].

Input examples for the training $\{y_1, \ldots, y_n\}$ are projected in the feature space and the algorithm looks for a hyperplane to separate the positive and negative examples with the largest possible margin (see Figure 2.8). When the training set is not linearly separable, SVMs find a trade-off between good classification and large margin.

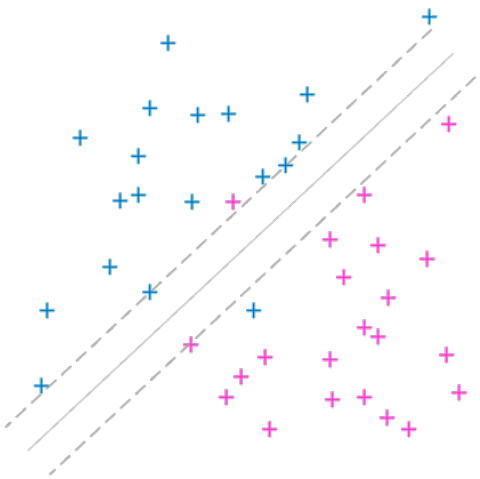


FIGURE 2.8 : Principle of data separation using a SVM. Positive examples (pink) and negative examples (blue) are separated by an hyperplane (grey line). Support vectors (grey dotted lines) correspond to the maximal margin.

SVMs can also perform non-linear classification using kernels. Instead of projecting in the feature space H, SVMs only address the feature space by computing the kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between two examples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, defined as:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i).\Phi(\boldsymbol{x}_j) \tag{2.6}$$

where Φ is the projection in the feature space H.

Common kernels include:

– the linear kernel :

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i.\boldsymbol{x}_j \tag{2.7}$$

– the polynomial kernel :

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i.\boldsymbol{x}_j + 1)^d \tag{2.8}$$

– the Gaussian kernel (also called *RBF* for *Radial Basic Function*) :

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{2.9}$$

where σ is the width of the kernel.

Optimizing kernel parameters is a hard problem. See [HTF09] for details.

Combining classifiers Combining different classifiers for the same problem might often lead to significant improvement as we saw in [BBAP11]. There are many ways to combine classifiers.

Ensemble learning consists in training different classifiers so as to combine them into a meta-classifier. Since its origins in the '90s, it developed significantly, mainly due to the increase in computing power. Ensemble learning techniques differ mainly by the aggregation techniques used to build the meta-classifier. Among the common techniques, one may cite *stacking* which modifies the feature vectors, *boosting* which alters the relative importance of the examples or *bagging* which builds different models by randomly drawing examples and aggregates them (by voting for example). Another interesting application to docking using bagging is available in [ABH+11]

To increase the performance of our docking scoring function, we used a different strategy inspired by recommender systems: collaborative filtering[7]. Several previous studies have shown that collaborative filtering (CF) can increase the accuracy of the prediction rate [SK09]. A CF recommender system contains a list of m users $U_1, U_2 \ldots, U_m$ and a list of n items $I_1, I_2, \ldots, I_n$. Each user rates each item. The final rating of each object is then defined by the ensemble of ratings received from each user. A key problem of collaborative filtering is how to combine and weight the preferences of the users. We used a simple voting scheme where the logarithm of the final score (rating) is the ratio of the positive (native structure) and negative (decoy) ratings. Positive and negative score are just taken as the sum of the votes weighted by the precision to have amplification. Details can be found in [BBAP11].

C Results and perspectives

1 Ground truth and experimental data

From the initial coarse-grained Voronoi construction detailed in my Ph.D. thesis, various models have been built and analyzed. A main difficulty in all these studies is to obtain a good dataset of native (positive) examples to consider as the *ground truth*. To discriminate between biological and crystallographic contacts, we benefited from a dataset manually extracted by biology experts [BBR+08]. For the protein-protein docking problem, we extracted and curated a reference set of protein-protein complexes from the PDB. Due to the growth rate of the PDB, this process had to be done several times to update the reference set [BAJP07, BBAP11]. The ESBTL software we implemented [LCB10] to be in strict accordance with the remediated PDB format (from 2007 onwards) has been key in simplifying this process. Despite the relatively large size of the PDB, the data it contains is very redundant and thus the obtained reference

[7] In this work, we use collaborative filtering in a general sense and do not restrict ourselves to the more narrow definition used in web recommender systems. For that reason, most of the commonly used systems will not be described here. The reader can refer to [RRS11] for details.

sets are small (less than 300 complexes). Protein-RNA complex structures are even harder to solve experimentally. The dataset we used [GGFAB14], obtained from the PRIDB [LWT+11] is also very small (120 complexes) for the same reason. Protein-protein and protein-RNA docking are biological problems that are very different from many biological applications for machine learning: few data are available and redundancy does not necessarily help in dataset curation.

2 Synthetic data generation

While the docking problem seems well suited for supervised learning, negative examples have to be computationally generated. For them to best mimic docking procedures and be used in a blind setting such as the CAPRI challenge, the negative examples have to be plausible: their predicted biophysical properties should be almost identical those of the native structures. The best way to obtain plausible negative examples is to run a docking program with a good biophysics-based function. We have been using many different software (Dock [WJ78], HADDOCK [DBB03], HEX [Rit03], RosettaDock [GMW+03],...), including our own coarse-grained procedure. This data generation step is often extremely computationally expensive (over several hundred thousands CPU hours for RosettaDock on the whole protein-RNA set). Furthermore, the pipeline to select the negative examples for the training can be sophisticated [GGFAB14]. This makes the whole procedure particularly heavy for beginners and might drive away students as a lot of practical and technical expertise is required in many different fields.

3 Voronoi construction and coarse-graining

The atomic Voronoi model for protein-protein complexes has been extensively studied at the atomic level but a correlation to the properties of the complexes is often hard to draw [BGNC09]. Our studies have shown that, with well-chosen sites, the coarse-grained Voronoi model can be as accurate as atomic methods (and even sometimes more). For the discrimination between biological and crystallographic contacts [BBR+08], we showed that accurate atomic features might not be needed and that physicochemical properties at the interface can sufficiently be encoded in a coarse-grained model. In [BPAJ05], we showed that the recognition phenomenon could be encoded by this coarse-grained model and our docking studies [BAJP07, BBAP11, BBAP09] in the light of the CAPRI context, have shown that a performance comparable to atomic models can be reached. While this model is performing well, it cannot provide high-resolution atomic solutions. To reach high-resolution, using an atomic Voronoi model might be ill-suited. One of the reasons the coarse-grained model is performing well for proteins is because of its ability to handle side-chains conformations at the interface. In an atomic model, a small conformational change from a side-chain would drastically modify the Voronoi diagram and its associated features, rendering the problem intractable. Another reason is that the atomic Voronoi model might encode more the chemistry of the atoms than the packing of the structure at the interface. As fine grained chemical features are relatively well handled by the biophysical force-fields available in docking software, an atomic Voronoi model would bring relatively little complementary information.

4 Supervised learning algorithms

SVMs have provided a great solution for discriminating between biological and crystallographic contacts for which having a training set was key and the influence of each feature was not a relevant information to the experimentalists [BBR+08]. For docking, despite resorting to a wide range of different types of algorithms including decision trees and rules [BBAP11], it is still not completely clear why the genetic

algorithm approach was the best one. It might be due to the poor diversity of the functional forms tested (mainly because of the small number of examples) but also to the fact that genetic algorithms can provide multiple solutions in very noisy environments. The protein-RNA case is an interesting example. In [GGFAB14], we used ROGER to perform logistic regression: it not only allowed for convergence but also improved RosettaDock default results.

Using the ROC curve and optimizing the AUC has been a strong point of our work. While the performance is good, the biological objective is somewhat different. Indeed in the CAPRI context (see section 5.2), only the best ten structures are assessed. The same applies to biology experiments where only a set of the top rank structures can be tested. A way to improve our results could thus be to optimize the initial slope of the ROC curve.

5 Relevance for biophysics and biology

5.1 The Random Energy Model (REM)

Neither coarse-grained models nor supervised learning can offer a physical description of biological phenomena. However, these models can be considered relevant when they provide access to solutions that could not be reached (because of theoretical or computational reasons). Based on the work of Joël Janin in 1996 [Jan96], we showed that protein–protein recognition can be described within the framework of the random energy model of statistical physics [Der81].

The score attributed to a docking model is taken to be an estimate of the energy E of the interaction between the two-component proteins in a particular state. The distribution of E is then analyzed as in [Jan96]: an energy spectrum is drawn by counting states with energies between E and $E + \mathrm{d}E$. If there are $m(E)$ such states, the entropy is:

$$S(E) = k_B \ln m(E) \tag{2.10}$$

where k_B is the Boltzmann constant.

The native state has an energy E_0 which taken to be zero for convenience. It is unique so that $S(E_0) = 0$. It is separated by an energy gap Δ from the non-native state of lowest energy.

The total number or states (including the native state) is:

$$N = 1 + \int_{\Delta}^{\infty} m(E)\,\mathrm{d}E \tag{2.11}$$

At thermodynamic equilibrium and temperature T, all states coexist and their relative abundance $n(E)$ follows Boltzmann's law:

$$n(E) = m(E) \exp\left(-\frac{E}{k_B T}\right) \tag{2.12}$$

The partition function Z is written:

$$Z = 1 + r = 1 + \int_{\Delta}^{\infty} n(E)\,\mathrm{d}E \tag{2.13}$$

The native state contributes 1 to Z, the non-native r.

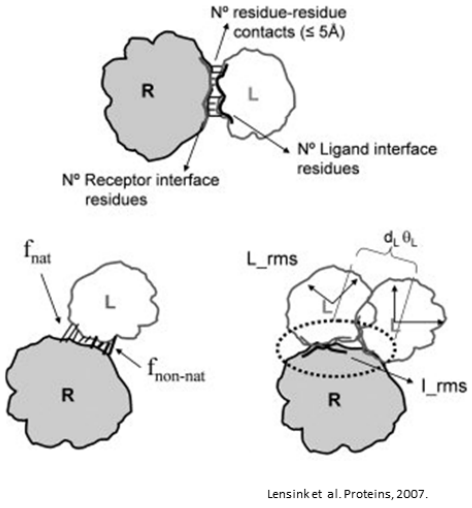
Specific recognition for complexes implies $r \ll 1$. The gap Δ should thus be large relatively to the thermal energy $k_B T$. The condition $r = \frac{1}{2}$ defines a temperature T_S that is called the *specificity*

transition temperature [Jan96]: below T_S, the native state is dominant; above T_S, non-native states take over. T_S can be obtained by drawing the tangent of the Entropy vs. Energy curve at the origin. The tangent at $E = \Delta$ defines another characteristic temperature of the system, its critical temperature T_C, also called *glass transition temperature* [Der81]. Below T_C, the only non-native states that compete with the native are those with energy near $E = \Delta$. Above T_C, many states of higher energy are populated. T_S and T_C can be calculated by fitting an analytical expression to the energy spectrum. The random energy model assumes a Gaussian distribution [Der81].

Taking the scores of the docking models to be interaction energies, we obtained the energy spectra for a large set of complexes and fit them to a Gaussian distribution, from which we derived physical parameters such as the glass transition temperature and the specificity transition temperature [BPAJ05].

5.2 Prediction results

The CAPRI challenge Together with CASP (see Chapter 3), the CAPRI challenge is a major community-wide blind assessment experiment in computational structural biology. Since the creation of CAPRI in 2001, the community has seen major improvements in protein-protein docking but has also extended the challenge to different types of related predictions: mutations [MFA+13], affinity [PJGPC+13], ion and water molecule prediction [KPX+13, LMB+14], or even design [FWS+11].



Ranking	Conditions based on CAPRI computed parameters
High ★★★	$f_{nat} \geq 0.5$ AND (L_rms ≤ 1.0 OR I_rms ≤ 1.0)
Medium ★★	($f_{nat} \geq 0.3$ AND $f_{nnat} < 0.5$) AND (L_rms ≤ 5.0 OR I_rms ≤ 2.0) OR $f_{nat} \geq 0.5$ AND (L_rms > 1.0 OR I_rms > 1.0)
Acceptable ★	($f_{nat} \geq 0.1$ AND $f_{nnat} < 0.3$) AND (L_rms ≤ 10.0 OR I_rms ≤ 4.0) OR $f_{nat} \geq 0.3$ AND (L_rms > 5.0 OR I_rms > 2.0)
Incorrect	$f_{nat} < 0.1$ OR (L_rms > 10.0 OR I_rms ≥ 4.0)

FIGURE 2.9 : CAPRI criteria to assess the quality of a prediction.

A usual mesure to assess the distance between two configurations is the *Root-Mean Square Deviation (RMSD)*. Once the two structures are superimposed, the RMSD can be computed as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{2.14}$$

where δ_i is the distance between the i^{th} pair of equivalent atoms. (N atoms are being considered for each conformation, as not all the atoms are necessarily taken into account).

The CAPRI scoring scheme [LMW07] is detailed on Figure 2.9. It uses a star-rating system based on four criteria:

- the fraction of native contacts included in the prediction f_{nat},
- the fraction of contacts in the prediction that are not in the native structure f_{nnat},
- the interface RMSD I_{RMSD} between the predicted and the native interface,
- the ligand RMSD L_{RMSD} between the predicted and the native ligands once the receptors have been superimposed.

The quality of the prediction (i.e. number of stars) is defined as:

- high ($\star\star\star$): $[f_{nat} \geq 0.5$ and $(I_{RMSD} \leq 1$ or $L_{RMSD} \leq 1)]$;
- medium ($\star\star$): $[(f_{nat} \geq 0.3$ and $f_{nnat} < 0.5)$ and $(L_{RMSD} \leq 5.0$ or $I_{RMSD} \leq 2.0)]$ or $[f_{nat} \geq 0.5$ and $(L_{RMSD} > 1.0$ or $I_{RMSD} > 1.0)]$;
- acceptable ($\star$): $[(f_{nat} \geq 0.1$ and $f_{nnat} < 0.3)$ and $(L_{RMSD} \leq 10.0$ or $I_{RMSD} \leq 4.0)]$ or $[f_{nat} \geq 0.3$ and $(L_{RMSD} > 5.0$ or $I_{RMSD} > 2.0)]$;
- incorrect in all other cases.

Each CAPRI target is made of two sessions: a full docking followed by a scoring session. For each session, participating groups are expected to provide ten ranked candidate solutions. The criteria are very stringent and acceptable solutions are often sufficient to perform experimental mutagenesis analyses.

Competitive results? In [BAJP07], we applied the Dock exploration procedure and our scoring model based on the Voronoi construction and the ROGER algorithm to ten targets from CAPRI rounds 3 to 6. We also evaluated the scoring model on HADDOCK conformations for five CAPRI targets (to which the HADDOCK group had participated). Our results showed that our predictive model ranks the solutions much more efficiently than the existing procedure (see Figure 2.10). The model is also independent of the algorithm used for the exploration phase. Following these interesting results, the CAPRI community decided to add a *scoring-only* phase to the challenge and a benchmark set has very recently been released [LW14].

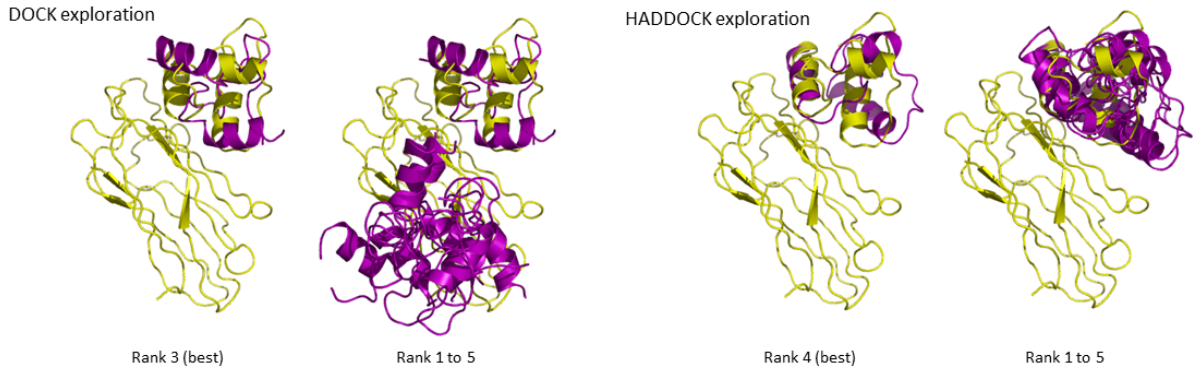


FIGURE 2.10 : Docking results on CAPRI target 11: the cohesin-dockerin complex. The left panel shows that no good solution is available from DOCK. The epitopes (interaction regions) are however detected and around half of the interface residues are predicted well for each epitope. The right panel shows that when reranking HADDOCK results, among a good set of samples, the solutions are correct. The best model can be detected (but is $\star$ only).

In [BBAP11], we used different classifiers to rank 12 CAPRI targets. The exploration is performed using a basic Voronoi exploration scheme. For ten targets out of twelve, we show that we can find a near-native conformation in the ten best solutions and for six targets, this solution is of high accuracy.

We also show that the scoring function substantially enriches the 100 best-ranked structure set. Detailed results are shown on Figure 2.11.

Groups	T22	T23	T25	T26	T27	T29	T32	T35	T37	T39	T40A	T40B
Wang	0	0	★★	0	★	0	0	★	★★	0	★★★	★★
Bonvin	★	0	★	-	★	★★	0	0	★	0	★★★	★★
Wolfson	-	-	★★	★★	★	0	0	0	★	0	★	0
Bates	-	-	-	-	★	★★	0	0	★★★	0	★★★	0
Weng	-	-	-	★★	★	★★	0	0	★★★	0	★★★	0
Fernandez-Recio	-	-	★★	-	★	★★★	0	0	0	0	0	0
Zhou	-	-	-	-	★	-	0	0	★★★	0	★★★	★★★
Haliloglu	-	-	-	-	-	-	-	-	★★	0	★★★	★★
Camacho	-	-	-	-	★★	★★	-	-	-	0	★★★	★★★
Takeda-Shitaka	-	-	-	0	0	0	0	0	-	0	★★★	★★
Vakser	-	-	-	-	-	-	★★	0	0	0	-	-
CF-GA	★★★[a]	★★★[a]	★	★★[a]	★★★[a]	★★	★★★	0	0	0	★	0

0: no acceptable or better solution found, - group has not participated, [a] f_{nat} evaluation

FIGURE 2.11 : Comparison of docking results with other CAPRI participating groups. For some targets, in particular when few experimental data is available, our methodology performs much better than the other groups. Our method however lacks the atomic accuracy when reranking high-accuracy solutions undergoing large conformational changes.

The community-wide experiment on design we participated in [FWS+11], showed that while machine learning techniques might have a bright future for protein-protein interaction prediction and design, the influence of the learning set is essential. It also showed that atomic accuracy might be needed for accurate design model selection. As of today, biophysics-based techniques are still leading for this task but the Schueler-Furman group has shown that supervised learning on atomic features could outperform these classical approaches.

We addressed the protein-RNA docking problem benefiting from our experience on protein-protein complexes during the Ph.D. of Adrien Guilhot [GG14]. In [GGFAB14] we developed a RosettaDock atomic scoring scheme using logistic regression and genetic algorithms. Not only is our model able to efficiently rank protein-RNA models while accounting for some flexibility, but it also outperformed the scoring scheme which was used by RosettaDock participants in CAPRI. Adrien Guilhot also tried to adapt the coarse-grained Voronoi model to RNA [GG14]. This happened to be a much harder problem. RNA side chains are much longer and for each nucleotide, one Voronoi site is not sufficient to account for the whole chain. Using two sites could solve this problem but the training set is so small that overlearning is an issue which is hard to overcome.

6 Perspectives

In ten years, we developed an efficient method for modeling and scoring protein-protein complexes. It is now performing well enough to address new biological applications. Our expertise was recently essential in analyzing and building a usable hAgo2-miRNA model [JSL+ed]. We also plan to use the strategy for specific models of therapeutic interest, e.g. antibody-antigen prediction.

Reaching the design stage will however require us to address flexibility and atomic models efficiently, in particular for RNA which undergoes large conformational changes and whose interactions depends

largely of ion and local electrostatics effects. The following chapters show two interesting approaches for this purpose.

Contents

CHAPTER

3

FROM BIOPHYSICS TO DATA: KNOWLEDGE-BASED POTENTIALS

Main articles from the chapter

[SSLB12] **Evaluating mixture models for building RNA knowledge-based potentials**, J Bioinform Comput Biol, 2012

[BHSL11] **Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation**, RNA, 2011

A Pioneer work and the RNA prediction challenge

1 Knowledge-based potentials for proteins

From the pioneer work of Anfinsen in the early '70s [Anf73], it has been known that the native structure of a protein usually has the lowest energy of all states. As a consequence, the development of a good free energy function would enable the prediction and assessment of protein structures [Dil85]. While the thorough sampling of the potential energy surface from a molecular mechanics force-field would in theory provide access to the free energy [BKP87], this is not possible in practice for computational reasons and inaccuracies in the energy functions. Another approach is to build a *statistical* or *knowledge-based (KB)* potential, i.e. a scoring function having a global minimum corresponding to the native structure and built from a sample of known native structures extracted from the PDB.

First studies used coarse-grained models and were based on residue types [TS76]. The first description of a distance-based potential was proposed by Sippl [Sip90] and since then, a large number of potentials have been developed and used for a wide range of applications, protein folding and the CASP challenge in particular (see [Zha08] for a review).

KB potentials are widely used in numerous applications because they are simple, efficient and accurate. They can however be difficult to derive depending on the experimental data available. In the following, I will describe the derivation of such potentials for RNAs and their performance in structure prediction.

2 RNA structure prediction

The function of non-coding RNA molecules is inherently linked to their 3D structures. RNA molecules, adopting various folds [GA06], are responsible for many biological functions in the cell. Some examples are detailed in Chapter 4. Being able to understand the way RNA folds would not only give us a good understanding of the relationship between structure and function, but also on evolution. Recent modelling initiatives have shown that the design of RNA structure is now within reach [DKB10, LKL+14] and RNA sequences can be designed to have specific biological functions or even perform specific tasks in nanomachines [Guo10]. Structure evaluation and modelling tools are thus even more needed than before.

From a word combinatorics perspective, the RNA sequence world seems much simpler than the protein world, the corresponding alphabet containing only four letters. This initially led researchers to believe that RNA structure prediction would be a much simpler problem than for proteins [TB99] as the structural diversity might be less, proteins having 20 different amino acids types. RNA structure prediction is however an extremely challenging task.

The folding process of RNA molecules is hierarchical [BRD99, TB99, BW97], allowing for simplification of the modelling process [SML12, SLM12]. An extended RNA first forms a stable secondary structure defined by base-pairing, then packs into a globular 3D shape. RNA secondary structure prediction has been widely studied [RHR+06, Zuk03, SYKB07, Mat06] and the challenge is now to determine how the local assembly of the bases affects the 3D structure. As shown on Figure 1.2, base interactions are mainly of two types: (i) base pairing, canonical (G-C and A-U) or non-canonical, and (ii) base stacking. Tertiary contacts, at the base of the 3D fold, can also be described with these two types. Numerous studies have provided classifications of base interactions [FMT+09, DB07, MARR03, SL05]. Stacking and pairing preferences are the basis of many recently developed RNA 3D structure prediction techniques, either in the form of fragment libraries from known RNA structures [PM08, DB07] or en-

ergy functions [FA10, DHT05, JRL+09]. While our understanding of base interactions has improved, in particular thanks to the classifications studies, it is still very hard to build and select the right 3D conformation corresponding to a specific secondary structure from a set of putative 3D structures called *decoys*.

Exactly like for the docking problem we presented in Chapter 2, energy functions based on coarse-grained models are often insufficient to extract good models based on a RMSD criterion with respect to the native structure [DB07] (see equation 2.14 in Chapter 2). Adding high-resolution terms greatly improved the prediction [DKB10]. The Rosetta package (FARNA [DB07], Rosie [LCC+13]) implements these fragment-based energy functions and allows for the reconstruction of small RNA motifs (the Rosie server[8] is limited to 23 nucleotide long RNAs) better than physics-based energy functions. It is based however on the parameterization of the weights of the various energy terms and it is unclear how this strategy will scale for larger RNAs.

Knowledge-based (KB) potentials were developed initially for protein structure prediction for the same reason (see Section 1). To derive such a potential, a training set of high-resolution, nonredundant structures is required. This set has to be carefully extracted and curated. Despite the availability of such potentials for proteins, the development of KB potentials for RNA has stalled mainly due to the smaller number of high-resolution RNA structures available.

3 Criteria for structural prediction

3.1 Energy vs. RMSD

Many criteria for assessing the quality of the prediction of a protein structure are available. Most were developed in the light of the CASP assesment procedure [CKF+09]. Despite its limitations, RMSD is still widely used for small structures. RMSD is a global criterion that cannot account for the differences in the quality of the prediction in different regions of the structures (see Chapter 2 equation 2.14). This is especially true for large structures and thus does not impact the RNA structures modelled here as they are of relatively modest size.

When assessing a large number of decoys, the visual inspection of the energy versus RMSD curve is extremely useful [BMB05]. With an ideal force-field, we would observe a funnel shape with a linear relationship between RMSD and energy in the close-to-native region. Figure 3.1 shows the difference between bad and good predictions, the latter having the characteristic funnel shape.

3.2 Native structure ranking

The native structure originates from the PDB and has been refined from experimental data using force-fields. As different force-fields lead to different minima, there is no guarantee that the native structure has a lower energy than good decoys. Refinement of near-native structure is thus a difficult problem [SL07].

In a prediction experiment, the native structure will not necessarily have the lowest energy. When the energies (or scores) of a large number of decoys are below the native structure, it often indicates that the energy selection:

- does not favor the native structure and thus will not be effective in ranking,
- leads to structures possibly containing a lot of atomic clashes,

[8] `http://rosie.rosettacommons.org/rna_redesign/`

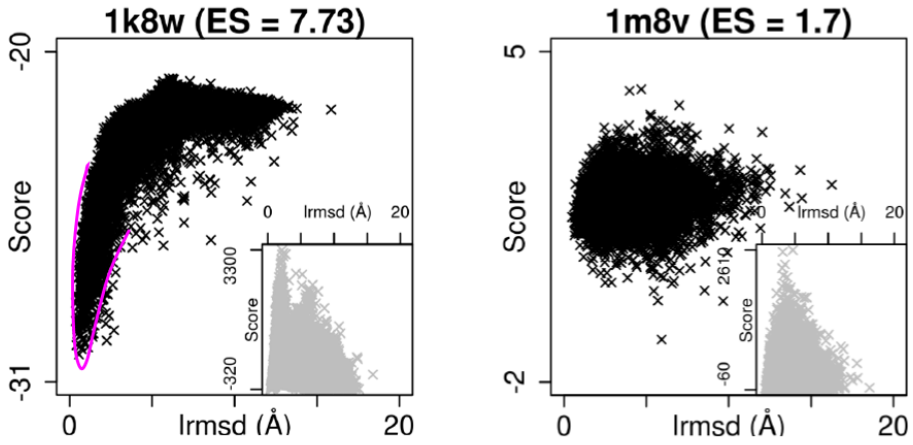


FIGURE 3.1 : Sample energy vs. RMSD plots for RNA docking (energy is in arbitrary units). The left panel shows a successful selection (in black) with a high enrichment score and the expected funnel shape (in pink). The right panel shows a "blob" shape corresponding to unsuccessful ranking (in black). Insets (in gray) show the ranking for a different scoring function which does not perform well for the same decoy sets. Enrichment scores are reported on top.

– favors uncommon interaction patterns at the residue/nucleotide level.

For such reasons, the number of structures having an energy lower than the native structure is often reported as a criterion. It quantitatively complements the energy vs. RMSD plot information.

3.3 Enrichment score

For another quantitative comparison between potentials, the Enrichment Score (ES) defined by Tsai et al. [TBM+03] is an interesting measure. It is defined as:

$$ES = \frac{|E_{top10\%} \cap R_{top10\%}|}{0.1 \times 0.1 \times N_{decoys}} \tag{3.1}$$

where:

– $E_{top10\%}$ corresponds to the set of structures in the best 10% of the energy range,
– $R_{top10\%}$ corresponds to the set of structures having their RMSD in the lowest 10% of the RMSD range,
– $|E_{top10\%} \cap R_{top10\%}|$ is the number of structures belonging to both $E_{top10\%}$ and $R_{top10\%}$.

For a linear scoring function $E_i = c \times R_i$, for each structure i and a constant c, this would give:

$$ES = \frac{|E_{top10\%} \cap R_{top10\%}|}{0.1 \times 0.1 \times N_{decoys}} = \frac{0.1 \times N_{decoys}}{0.1 \times 0.1 \times N_{decoys}} = 10 \tag{3.2}$$

In a random scoring case, we would have:

$$ES = \frac{|E_{top10\%} \cap R_{top10\%}|}{0.1 \times 0.1 \times N_{decoys}} = \frac{0.1 \times 0.1 \times N_{decoys}}{0.1 \times 0.1 \times N_{decoys}} = 1 \tag{3.3}$$

Hence, we have:

$$ES = \begin{cases} 10, \text{ perfect scoring} \\ 1, \text{ perfectly random} \\ < 1, \text{ bad scoring} \end{cases} \tag{3.4}$$

Clearly a good scoring function should have an ES between 1 and 10, the closer to 10 the better. Precisely defining what is a good scoring function is however not obvious. The ES is nevertheless a good criterion for comparing scoring functions and is in accordance with energy vs. RSMD plots (see Figure 3.1).

B RNA structural data and potential derivation

1 Dataset extraction and distance collection

Extracting a reference dataset for KB potentials is, like in the docking case (see Section 1 in Chapter 2), a time consuming process. The process is mainly automated but manual curation steps are necessary. The structures in the dataset should have very specific properties to limit the bias: (i) be of high-resolution (greater than 3.5Å in this case), (ii) contain unbound RNA where less than 5% of the nucleotides are non-standard or missing, (iii) have less than 20% sequence identity with another structure in the dataset, (iv) be the biologically active quaternary structure (symmetric chains have to be built if needed). The very stringent criteria led to less than a hundred structures. Despite its relative small size, this reference set is sufficient for distance-based potential building.

2 Building a potential

2.1 Formalism

Overview KB potentials are based on distance computation between atoms. For practical purposes, the pairwise preferences of atoms can be looked at through two different equivalent views: probabilities or free energies. The latter uses the Boltzmann formalism (see Section 5.1 in Chapter 2 and [BPAJ05] for a full description). It assumes an equilibrium distribution of atom-atom preferences for which: (i) the reference state might not correspond to any physical observation and (ii) the probability of observing a system in a given state must change with the temperature [Mou97]. The formalism will be briefly described below. The reader is invited to refer to [SM98] and [LS01] for details.

Conditional probabilities[9] For the following, we will divide the decoy set in two subsets: the set of conformations we will consider correct (i.e. near native conformations) C and the set of incorrect structures I. The set of *properties* (or features) of a structure is denoted $\{y_k\}$. In this specific case, this corresponds simply to the set of distances $\{d_{ab}^{ij}\}$ between atoms (or coarse-grained atoms/sites) i and j of types a and b respectively. The probability for a structure to be in the correct set C given its distances $\{d_{ab}^{ij}\}$ is denoted $P(C|\{d_{ab}^{ij}\})$. $P(d_{ab}^{ij}|C)$ is the probability of observing a distance d between atoms i and j of types a and b respectively. Subsequently, $P(d_{ab}^{ij})$ is the probability of observing such a distance in any structure.

[9] To simplify, the notations used in this chapter are slightly different from the previous chapter.

The chain rule for conditional probabilities gives [Sch94]:

$$P(C) \cdot P(d_{ab}^{ij}|C) = P(d_{ab}^{ij}) \cdot P(C|d_{ab}^{ij}) \tag{3.5}$$

where $P(C)$ is the probability of any structure picked at random to be in the correct set C. Making the approximation that all distances are independent from one another, the probabilities of observing the set of distances is the product of the probabilities of observing each individual distance, leading to:

$$P(\{d_{ab}^{ij}\}|C) = \prod_{ij} P(d_{ab}^{ij}|C) \text{ and } P(\{d_{ab}^{ij}\}) = \prod_{ij} P(d_{ab}^{ij}) \tag{3.6}$$

Consequently:

$$P(C|\{d_{ab}^{ij}\}) = P(C) \cdot \prod_{ij} \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \tag{3.7}$$

where, for a given sequence, $P(C)$ is a constant and will not be considered further. The following will thus not allow for the comparison of structures of different sequences.

The approximation used to write Equation 3.6 is obviously not true for 3D structures: if an atom A is close to an atom B and to an atom C, then B and C must be somewhat close. It is possible to write the probabilities so as to take this into account. The amount of data that would be required to obtain adequate statistics in practice would however be fairly larger that what we can get. It is also uncertain this would lead to better results.

Using the log form of Equation 3.7, we can build a score $S(\{d_{ab}^{ij}\})$ having a form similar to a potential of mean force [Kir36]:

$$S(\{d_{ab}^{ij}\}) = -\sum_{ij} \ln \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \quad \propto \quad -\ln P(C|\{d_{ab}^{ij}\}) \tag{3.8}$$

In the previous equation, $S(\{d_{ab}^{ij}\})$ can be computed for a given conformation by calculating all the distances between the pairs of atoms and summing up the probability log ratios assigned to each distance between a pair of atom types.

The distributions of $P(d_{ab}^{ij}|C)$ and $P(d_{ab}^{ij})$ have to be computed for all pairwise distance types in order to be able to use Equation 3.8. From the reference set of structures, all atom contacts can be computed for a particular distance *bin*. The probability of observing atom type a and atom type b in a distance bin centered on the distance d in a near-native conformation C, is:

$$P(d_{ab}^{ij}|C) = f(d_{ab}) = \frac{N(d_{ab})}{\sum_d N(d_{ab})} \tag{3.9}$$

where $N(d_{ab})$ is the number of observations of atom types a and b in distance bin centered on the distance d. $\sum_d N(d_{ab})$ is the number of a-b contacts for all distance bins. $f(d_{ab})$ denotes the distribution obtained from the reference set and represents the probabilities.

Potential of mean force The potential of mean force method for discriminating between correct and incorrect structures relies on three assumptions:

1. The total free energy of a molecule relative to some reference state ΔG_{tot} can be expressed as the sum of the relative free energies $\Delta G(R)$ of a number of individual contributions, R being the

value of a *reaction coordinate*. Considering the distance d between atoms i and j of type a and b respectively as a reaction coordinate, we have:

$$\Delta G_{tot} = \sum_{ij} \Delta G(d_{ab}^{ij}) \tag{3.10}$$

This formulation uses the approximation on the conditional probabilities described earlier: the enthalpy can reasonably be approximated by a pairwise sum of interactions but not the entropy for which contributions cannot be considered additive. Equation 3.10 was however shown to be valid for building empirical force fields [MH88].

2. The inverse Boltzmann law can be used to express the relative free energy of a particular interaction between any pair of atom types:

$$G(d_{ab}) = -kT \ln \frac{\rho(d_{ab}|C)}{\rho(d_{ab})} \tag{3.11}$$

where k is the Boltzmann constant, T is the absolute temperature, $\rho(d_{ab}|C)$ is the density of atom types a and b at a distance d in assessed structures and $\rho(d_{ab})$ is the same density in the reference state. Again, this requires that the distances between pairs of atoms are independent from the environment and also that the distribution of d_{ab} follows the Boltzmann distribution.

3. The native state is the lowest free energy conformation (this is the thermodynamic hypothesis described in [Anf73]). The substitution of Equation 3.11 in Equation 3.10 leads to:

$$G_{tot} = -kT \sum_{ij} \ln \frac{\rho(d_{ab}^{ij}|C)}{\rho(d_{ab}^{ij})} \tag{3.12}$$

which is similar to Equation 3.8, the densities replacing the probabilities.

Finding the lowest free energy structure is thus equivalent to finding the most probable structure in terms of Bayesian statistics.

2.2 Measurements and statistics

The previous description can be summarized by expressing the energy E of a given conformation as:

$$E = -kT \sum_{ij} \ln \frac{p_{obs}(d_{ij})}{p_{ref}(d_{ij})} \tag{3.13}$$

where T is the temperature (taken to be 300K) and k is the Boltzmann constant. $p_{obs}(d_{ij})$ and $p_{ref}(d_{ij})$ represent the observed and reference probabilities respectively for atom types i and j separated by a distance d_{ij}[10].

All the distances between all types of atoms are computed on a reference set[11]. As mentioned above, the bin size of the histogram is critical and has shown to lead to very different results, in particular when combined to spline fitting for differentiation [SL07]. To avoid this problem, we show it is possible to use *mixture models (MMs)* to estimate distance densities (reference probabilities) and thus obtain smooth analytically differentiable potential functions [SSLB12] that can be used in MD software such

[10] Previous use of atom types a and b has been dropped for simplicity.

[11] All computations and tests are performed in a leave-one-out validation setting as described in Section 2 in Chapter 2. This leads to computationally expensive experiments. To simplify, this will not be detailed here.

as Gromacs [PPS+13]. Each density is modeled as a mixture of univariate Gaussian distributions. This mixture has the general form:

$$p(d) = \sum_{i=1}^{N} w_i \mathcal{N}(\mu_i, \sigma_i^2) \tag{3.14}$$

with

$$\sum_{i=1}^{N} w_i = 1 \text{ and } w_i \geq 0 \, \forall \, 1 \leq i \leq N \tag{3.15}$$

Building MMs corresponds to estimating the parameters w_i, μ_i and σ_i^2 in order to maximize the quality of the approximation.

Many algorithms can be used to estimate the parameters of a MM, each having strong points and pitfalls for this kind of application. In [SSLB12] we review expectation-maximization (EM), Dirichlet process mixtures (DPMs) and kernel density estimation (KDE). We also address the simplification of mixture models using Bregman hard clustering and k-means based on the Kullback-Leibler divergence. We show that despite of being computationally expensive, the DPM approach we set up in [BHSL11] is the best to efficiently build potentials for RNA.

2.3 Low-count regions and distance cutoff

A common problem in computing the probabilities, is that the low-count regions (i.e. the distances closer to zero) are evidently poorly represented and might lead to numerical instabilities when deriving the potential. To overcome that difficulty, we show that it is possible to correct the potential in the low count region by using a linear approximation from the origin to the first descending inflexion point (first observed basin). A smooth truncation at a cutoff distance (usually around 14Å) is also performed by multiplying each potential by a negative sigmoid function. This allows for smooth potentials and performs better than the classical data treatment described in [SL07].

2.4 The reference state problem

In the previous equations, $P(d_{ab})$ can be seen as a property of the *reference state*: the probability of seeing a separation d between two atoms a and b in any possible structure. From a Bayesian statistics perspective, this is a *prior distribution*. This representation of our knowledge of distance distribution can be of various types. We used the simplest choice possible for the prior distribution, assuming that averaging over the different atom types in the reference set is an adequate representation of the random arrangements of atom types in any conformation. This leads to the following approximation of $P(d_{ab})$, the probability of finding atom types a and b in the distance bin centered on the distance d in any conformation (native or otherwise), as equal to $P(d)$, the probability of seeing any two atoms in the distance bin centered on the distance d:

$$P(d_{ab}) = P(d) = f(d) = \frac{\sum_{ab} N(d(d_{ab}))}{\sum_{d} \sum_{ab} N(d_{ab})} \tag{3.16}$$

where $\sum_{ab} N(d(d_{ab}))$ is the total number of contacts between all pairs of atom types in a particular distance bin centered on d. This latter approximation is clearly not true.

Several different studies have addressed the choice of the reference state. Some options include geometric filtering [ZGK06], an ideal gas reference state [ZZ02] or a quasi-chemical approximation [LS01],

which originates from the "uniform density" reference state defined by Sippl [Sip90]. Our studies used the latter with a composition-independent scale, i.e. the observed distances from all possible pairs are combined together to represent the reference state, as not much difference was shown in the discrimination performance.

C Outcome and limitations

1 Biological results

In [BHSL11], we show that the obtained potentials are very smooth and account for structural features of RNA molecules. Figure 3.2 shows an all-atom KB potential example. The analysis of the MM strategies and coarse-grained potential are detailed in [SSLB12]. Interestingly our approach, even in its coarse-grained flavor, outperformed the Rosetta model for high-resolution atomic structure. Figure 3.3 displays the results obtained for the GUAA tetraloop (PDB id 1MSY). Due to its use of templates, the Rosetta approach enforces strong base stacking, even in unpaired regions which often is not in agreement with experimental results. As we describe in Section 2 of Chapter 4, RNA loop structure and dynamics can be very diverse and favor very different configurations.

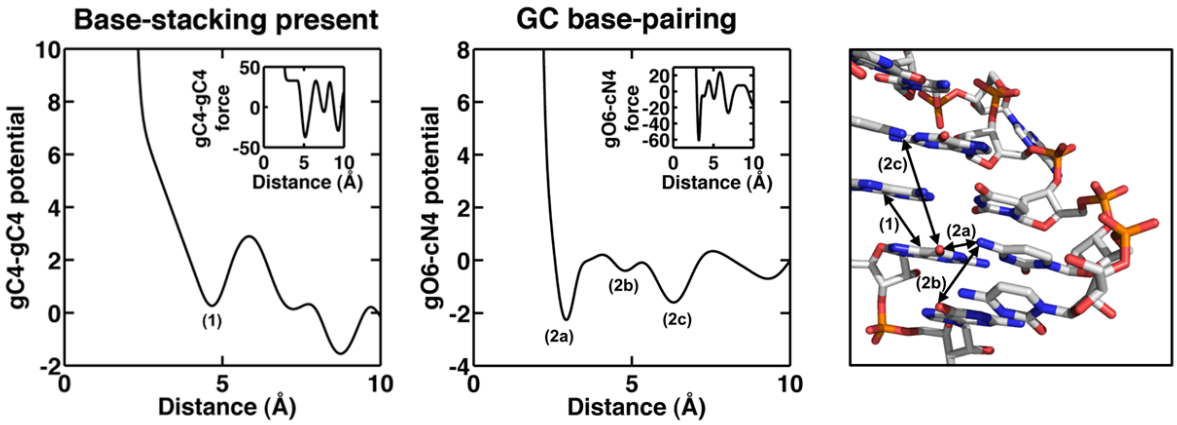


FIGURE 3.2 : Structural features captured by the all-atom KB potential for RNA. The plots (left) show the potentials for two specific atom pairs where base stacking and base pairing occur. The corresponding force is displayed in the inset. The distances highlighted on the plots are illustrated on the structure of the Rev element of HIV-1 (right).

2 KB potential derivation and applications

Our assessments of KB potentials derived from different MMs suggest that DPM modeling is an efficient approach to generate smooth, differentiable KB potentials of RNA that preserve important biological information. Applications of traditional KB potentials (derived from spline-fitting) in biological structural modeling have often been limited by excessive ruggedness of KB potentials. Our studies showed that the use of an appropriate MM (e.g. DPM) provides a less rugged KB potential with similar structure scoring properties. Hence the KB potential derived from DPM could certainly be more versatile than the traditional version, thereby allowing extensive and plausible applications in molecular modeling like

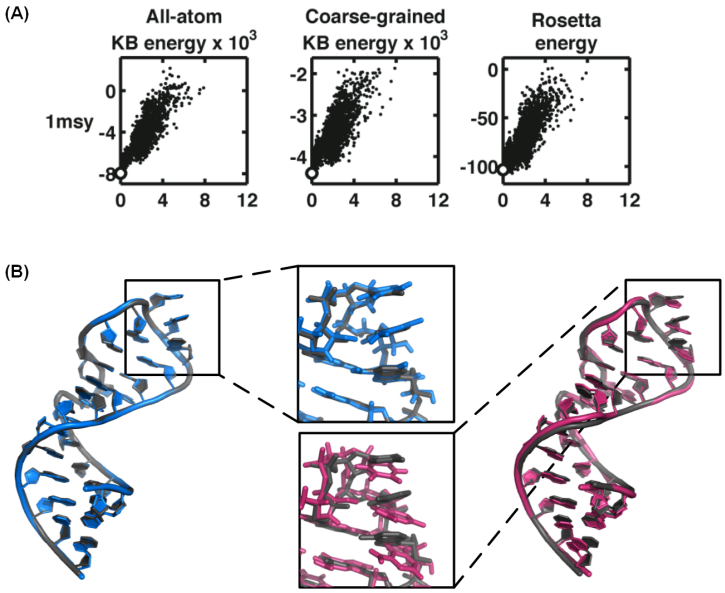


FIGURE 3.3 : Evaluation of the KB potential: comparison to the Rosetta approach for the GUAA tetraloop (PDB id 1MSY). Energy vs. RMSD curves are shown on panel A: all three scoring functions perform well and exhibit the funnel shaped pattern. Panel B shows the best scored KB structure (in blue) and the best scored Rosetta structure (in pink) relatively to the native structure (in black). The central inlets show close-up views of the tetraloop.

minimization and sampling. Figure 1.1 displays an example of a 5pt potential folding experiment for RNA.

Using KB potentials in this fashion is however controversial: as described above, many approximations are made and most of them are evidently wrong (see [FK98]). The theoretical basis of KB potentials can thus be questioned. The performance of KB potentials as scoring functions for filtering or sieving out bad solutions is however interesting, especially as it can be coarse-grained and derived for almost any type of measure. It can outperform template-based techniques which currently lead RNA structure prediction and can be used at any coarse-graining level, even coarser ones. While KB potentials are certainly not able to truly model the energy of a biological system and provide information on its free energy and reaction pathway, they are useful as good approximations of stable or metastable states.

A useful follow-up for structure prediction and docking techniques would be to integrate orientation-dependant parameters, both at the coarse-grained and atomic levels. Ideally, this would provide screening functions for prediction of large scale assemblies.

Contents

CHAPTER

4

A ROBOTICS-INSPIRED MODEL: INVERSE KINEMATICS SAMPLING FOR RNA

Main articles from the chapter

[FPBv14] **Characterizing RNA ensembles from NMR data with kinematic models**, Nucleic Acids Res, 2014

A Inverse kinematics and RNA dynamics

As stated in the previous chapters, noncoding ribonucleic acids (ncRNAs) mediate important cellular processes. Transfer RNA and ribosomal RNA are essential functional components in protein synthesis [NHB+00]. Short interfering RNAs (siRNAs) and microRNAs (miRNAs) are key in silencing the expression of specific genes in the cell and are thus interesting therapeutics targets [CWD09, DT04]. Riboswitches also regulate gene expression [TB05] and, like other functional RNAs, can be used in nanomedicine to silence cancer and infectious disease-specific genes [ZSG+11].

To interact with binding partners and perform their function [LV01, KAC+14], RNA molecules sample a wide range of conformations [LV01, KAC+14]. Characterizing the dynamics of RNA is very difficult as the native ensemble of biomolecules, i.e., the set of conformational states the molecule adopts in vivo, cannot be observed directly. Nuclear Magnetic Resonance (NMR) can probe the RNA conformational landscape at timescales ranging from picosecond to seconds or longer, often providing detailed evidence of dynamically interchanging, sparsely populated substates [RBF+11, BNE+11]. While being able to analyze NMR spectroscopy data guided by a conformational ensemble has long been recognized of great value [BVH+11, EBH+14], resolving motionally averaged NMR measurements into constituent, structural contributions remains extremely challenging [SBHH14].

Advanced molecular dynamics simulations can provide good insights on the conformational diversity of RNA [FSAHA09, BDMV13] but long trajectories with specialized force fields on dedicated supercomputers are required to adequately sample conformational space, limiting ensemble analyses to modestly-sized RNA molecules [SBAAH13].

In this chapter I describe a new research direction explored in collaboration with Rasmus Fonseca and Henry van den Bedem during the last two years. An efficient conformational sampling procedure, called Kino-Geometric Sampling for RNA (KGSrna) is presented. This model can report on ensembles of RNA molecular conformations orders of magnitude faster than MD simulations and has proven to be better at interpreting NMR results than Normal Mode Analysis (NMA). We also show in [FPBv14] that such sampling of 3D RNA models can recover the conformational landscape encoded by proton chemical shifts in solution. Combined with NMR residual dipolar coupling (RDC) measurements, our procedure can automatically select the size and weights of a small conformational ensemble that, provably, best agrees with the data.

The results can putatively guide interpretation of a wide range of experiments such as proton chemical shift [FHAAH13] or residual dipolar coupling data [BVH+11, FSAHA09], and complement insights obtained from single, averaged models [SCC+14, SLD+08], ensembles resulting from molecular dynamics, normal mode analysis [ZS00], Monte Carlo simulations [SLM12] or de novo tertiary structure prediction [RRPB11, DB07, SYKB07].

B KGSrna: a simple and efficient model

The Kino-Geometric Sampler for RNA (KGSrna) is an efficient conformational sampling procedure for RNA inspired from robotics. KGSrna represents an RNA molecule as a kinematic linkage, capitalizing on the tree-like structure of polynucleotides, with groups of atoms as links or rigid bodies and rotatable bonds as joints (See Figure 4.1a). In this representation, distance constraints such as non-covalent bonds create nested, closed loops or kinematic cycles (See Figure 4.1b). Degrees-of-freedom in a cycle demand carefully coordinated changes to avoid breaking the non-covalent bond, which greatly reduces the conformational flexibility [JRKT01, KGLK05, vLLD05, YDM+08, YZL12]. The reduced flexibility from

a network of nested kinematic cycles consequently deforms the biomolecule along preferred directions on the conformational landscape. Our procedure projects degrees-of-freedom onto a lower-dimensional subspace of conformation space, in which distance constraints are maintained (see Figure 4.1c).

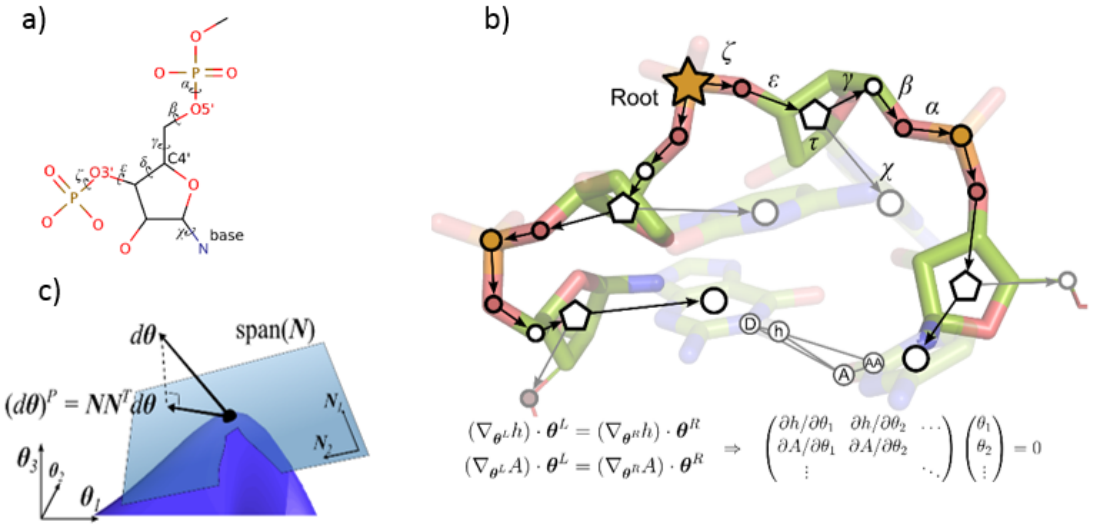


FIGURE 4.1 : Kinematic representation of RNA. a) A single nucleotide of RNA with its torsional degrees of freedom. b) Edges in the directed spanning tree encode n torsional degrees-of-freedom $\theta = (\theta_1, \cdots, \theta_n)$ and vertices (circles) encode rigid bodies. Pentagons represent riboses, which have an additional internal degree-of-freedom governing their conformation (puckering). The hydrogen bond h-A closes a kinematic cycle, and is one of m distance constraints in the molecule to be maintained. As the position of the hydrogen atom h changes through perturbation of dihedral angles in the left branch of the tree, the new position of h should be matched by appropriate changes in the right branch, i.e. $(\nabla_{\theta^L} h) \cdot \theta^L = (\nabla_{\theta^R} h) \cdot \theta^R$. Similarly, a change in position of heavy atom A from the right tree should be matched by changes in the left tree. These instantaneous distance constraints define the $6m \times n$ Jacobian matrix J . c) A schematic representation of the subspace of conformational space defined by the closure constraints. The subspace (blue surface) is highly nonlinear, but can be locally approximated by its tangent space, the null space of J (translucent blue plane).

In KGSrna, an RNA molecule is represented with rotatable, single bonds as degrees of freedom and groups of atoms as rigid bodies. Non-covalent bonds are distance constraints that create nested cycles (Figure 4.1b). We also integrated a differentiable parameterization of ribose conformations into the kinematic model. Our strategy rapidly generates representative ensembles of RNA molecular conformations and leads to excellent agreement with experimentally observed conformations.

1 Methodology

1.1 Overview

The purpose of KGSrna is to sample the native ensemble of RNA molecules starting from a single member of this ensemble. KGSrna takes as input an initial conformation, $\boldsymbol{q}_{\text{init}}$, and an exploration radius, $r_{\text{init}} \in \mathbb{R}$. First, a graph is constructed such that atoms are represented as vertices and covalent bonds and hydrogen bonds are edges (see Figure 4.2a). A minimal directed spanning tree is extracted from this graph and two conformational operators, the *null-space perturbation* and the *rebuild perturbation*,

are used to make conformational moves that never break any bond in the graph. KGSrna then grows a pool of conformations by repeatedly perturbing a seed conformation, $\boldsymbol{q}_{\text{seed}}$, selected among previously generated conformations in the pool (or $\boldsymbol{q}_{\text{init}}$).

1.2 Construction of the tree

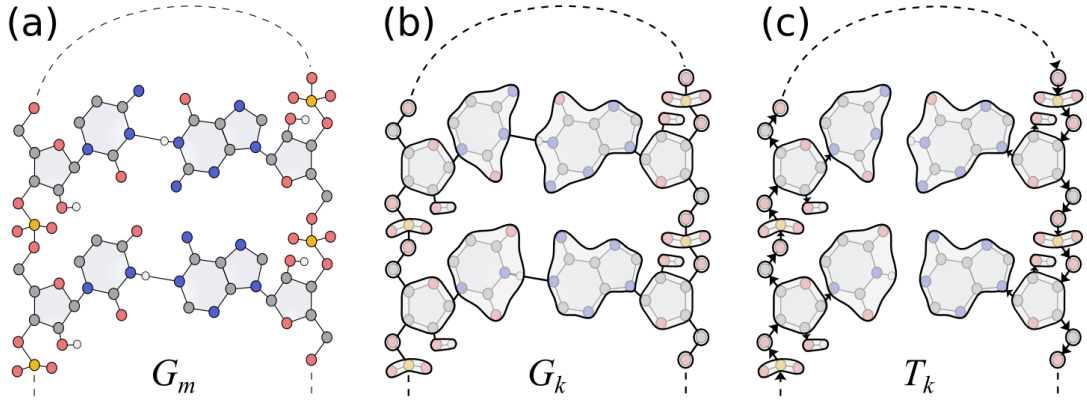


FIGURE 4.2 : Building the kinematic tree. a) The molecular graph (undirected), constructed from atoms and the covalent and hydrogen bond networks. b) The kinematic graph (undirected), constructed by edge-contracting all non-rotatable bonds in the molecular graph. c) The kinematic tree (directed), constructed by finding a spanning tree in the kinematic graph.

A graph $G_m = (V_m, E_m)$ is constructed such that V_m contains all atoms and E_m contains all covalent or hydrogen bonds (see Figure 4.2a). For RNA, we can use a simple model where only the hydrogen bonds A(N3)–U(H3) and G(H1)–C(N3) in canonical Watson-Crick (WC) base pairs are included as edges.

In a following stage, a compressed graph $G_k = (V_k, E_k)$ is constructed from G_m by repeatedly edge contracting members of E_m that correspond to: (i) partial double bonds, (ii) edges (u, v) where u or v has degree one, or (iii) edges in pentameric rings ((deoxy-)ribose in nucleic acids or proline in amino acids) (see Figure 4.2b). Each edge in E_k thus corresponds to a revolute joint, i.e. a rotating bond with 1 degree of freedom (DOF) and vertices in V_k correspond to collections of atoms that form rigid bodies.

In the final step, a rooted minimal spanning tree, $T_k = (V_k, E'_k)$, is constructed from G_k (see Figure 4.2c). Forward kinematics are defined as propagation of atom coordinate transformations from the root of T_k, along the direction of edges in E'_k. Constraints are defined as all edges in $C_k \equiv E_k \setminus E'_k$. As the two perturbation methods are approximations that can introduce small displacements of constraints, we assign a weight of 1 to covalent bonds, 2 to hydrogen bonds, and use Kruskal's algorithm for the spanning tree construction [LRSC01]. This guarantees that covalent bonds are favored over hydrogen bonds for inclusion in E'_k.

1.3 Modeling the conformational flexibility of pentameric rings

The flexibility of RNA is particularly dependent on conformational flexibility of ribose rings [LLW06], but directly perturbing a torsional angle in pentameric rings breaks the geometry of the ring. While pseudorotational angles [AS72] are frequently used to characterize ribose conformations, they are not convenient for a kinematic model as the equations mapping a pseudorotation angle to atom positions

are non-trivial. We therefore introduce a parameterization inspired by [HCSD05] from a continuous differentiable variable τ to the backbone δ angle (C5'-C4'-C3'-O3') so that the ideal geometry of the ribose is maintained (see Figure 4.3).

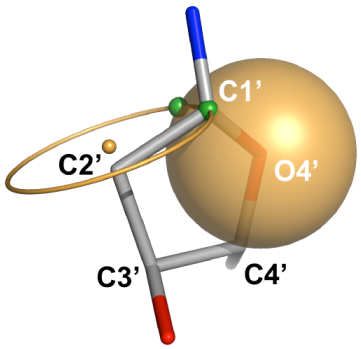


FIGURE 4.3 : Geometric characterization of ribose ring kinematics. The position of C1' is determined from an ideal O4'-C1' distance (yellow sphere), an ideal C1'-C2' distance and ideal C1'-C2'-C3' angle (yellow circle).

The positions of O4', C4', and C3' are determined by (torsional) DOFs higher in the kinematic tree. The position of C2' and the branch leaving C3' in the kinematic tree is determined from the C5'-C4'-C3'-O3' torsion, δ. Thus, only the remaining atom C1' needs to be placed. Positions of C1' with ideal C1'-C2' distance and C1'-C2'-C3' angle are represented by a circle (see Figure 4.3), centered on the C3'-C2' axis and having the C3'-C2' axis as its normal vector. Positions of C1' that have ideal C1'-O4' distance are represented by a sphere centered on O4'. The position of C1' is on either one of the intersections between the sphere and the circle, indicated by the variable $u \in \{-1, 1\}$

To avoid using the aforementioned variable u which is discontinuous, and δ which is limited by the ring geometry, we introduce the periodic and continuous variable τ, which uniquely specifies both δ and u. Since δ is restricted to move in the range $120° \pm A$ where A is typically $\approx 40°$, we set $\delta = 120° + A\cos\tau$. By defining $u = \text{sgn}(\sin\tau)$, the ribose conformation follows a continuous, differentiable, and periodic motion for $\tau \in \mathbb{R}$. The geometric equations describing the position and reference frame of C1' are all differentiable with respect to τ. This is essential as the inverse kinematics methods described in the following rely on taking position derivatives.

1.4 Null-space perturbations

The full conformation of a molecule is represented as a vector $\boldsymbol{q}$ containing values of all DOFs, both torsions and τ. To make a conformational move, we perform a so-called null-space projection of a random trial vector that ensures constraints stay together as described in [YZL12].

We use a constraint $c \in C_k$ with endpoints $\boldsymbol{a}$ and $\boldsymbol{b}$ and the paths L and R from each endpoint to their nearest common ancestor. Maintaining a constraint corresponds to maintaining the equations:

$$\boldsymbol{f}_L(\boldsymbol{a}, \boldsymbol{q}) = \boldsymbol{f}_R(\boldsymbol{a}, \boldsymbol{q}) \quad (4.1)$$

$$\boldsymbol{f}_L(\boldsymbol{b}, \boldsymbol{q}) = \boldsymbol{f}_R(\boldsymbol{b}, \boldsymbol{q}) \quad (4.2)$$

where $\boldsymbol{f}_L(\boldsymbol{x}, \boldsymbol{q})$ and $\boldsymbol{f}_R(\boldsymbol{x}, \boldsymbol{q})$ are the positions of $\boldsymbol{x}$ after applying forward kinematics of the DOFs in $\boldsymbol{q}$ along L and R respectively. We denote the subspace of conformations that satisfy these equations for all

constraints the *closure manifold*. The first-order approximation of these equations can be written:

$$\boldsymbol{J}d\boldsymbol{q} = \boldsymbol{0} \tag{4.3}$$

where $\boldsymbol{J}$ is a $6\,|C_k| \times n$ matrix containing partial derivatives of endpoints with relation to the n DOFs. Solutions to this equation are in the null-space of $\boldsymbol{J}$ which constitutes the tangent-space to the point $\boldsymbol{q}$ on the closure manifold. The right-singular vectors of the singular value decomposition $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ form a basis, $\boldsymbol{N}_J$ for the null-space of the Jacobian. As long as sufficiently small steps are taken in the null-space, it is possible to traverse any connected component of the closure manifold. A null-space perturbation of $\boldsymbol{q}_{\text{seed}}$ is therefore performed by finding a small random trial vector $\Delta\boldsymbol{q}$ and setting $\boldsymbol{q}_{\text{new}} \leftarrow \boldsymbol{q}_{\text{seed}} + \boldsymbol{N}_J\boldsymbol{N}_J^T\Delta\boldsymbol{q}$. Computing the singular value decomposition of the Jacobian dominates the running time so the Intel Math Kernel Library was used for its efficient parallel implementation of LAPACK.

1.5 Rebuild perturbations

The conformations of ribose rings change when performing null-space perturbations, but in general the changes are small enough that a full change from C3'-endo to C2'-endo is very rarely observed, even in flexible loop regions. As shifts from one ribose conformation to another are frequent and biologically important in RNA molecules [LW78], a rebuild perturbation was designed that can completely change a ribose conformation and rebuild the backbone so the conformation stays on the closure manifold.

A rebuild perturbation first picks a segment of two nucleotides neither of which are constrained by hydrogen bonds or aromatic stacking. It then disconnects the C4'-C5' bond at the 3' end of the segment, stores the positions of C4' and C5', and resamples the τ value of the two nucleotides, which breaks the C4'-C5' bond.

To reclose the broken bond we let $\boldsymbol{q}'$ denote the backbone DOFs in the segment (not including τ-angles) and let $\boldsymbol{e}$ denote the end-effector vector which points from the current positions of C4' and C5' to the stored ones. A first-order approximation to the problem of finding a vector $\boldsymbol{q}'$ that minimizes $|\boldsymbol{e}|$ can be written as:

$$\boldsymbol{J}d\boldsymbol{q}' \;=\; \boldsymbol{e} \tag{4.4}$$

where $\boldsymbol{J}$ is the $6 \times n'$ Jacobian matrix containing the derivatives of end-points with respect to the n' DOFs in $\boldsymbol{q}'$.

In general $\boldsymbol{J}$ is not invertible, so instead the pseudo-inverse, $\boldsymbol{J}^\dagger$ is used, which gives the least squares approximation solution to the above equation. The pseudo-inverse can be found from the singular value decomposition of $\boldsymbol{J}$: $\boldsymbol{J}^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{U}^T$ where $\boldsymbol{\Sigma}$ is a diagonal matrix with entries s_{ii} and $\boldsymbol{\Sigma}^\dagger$ is a diagonal matrix with entries $1/s_{ii}$ if $s_{ii} > 0$ and 0 otherwise. To reclose the C4'-C5' bond we therefore iteratively set $\boldsymbol{q}' \leftarrow \boldsymbol{q}' + 0.1 \cdot \boldsymbol{J}^\dagger\boldsymbol{e}$ until the distance between the original C4' and C5' atoms is less than 0.0001. Algorithms 4.1 and 4.2 describe the whole procedure.

Ribose conformations in experimental structures mainly fall in two distinct peaks corresponding to C2'-endo and C3'-endo. To mimic this behavior, τ-angles are sampled using a mixture of wrapped normal distributions. The following bimodal distribution (see also Figure 4.4) was obtained by fitting to the τ-angles of riboses taken from the high-resolution RNA dataset compiled for a previous study (see Chapter 3 and [BHSL11]):

$$P(\tau) = 0.6 \cdot N(\tau, 215^\circ, 12^\circ) + 0.4 \cdot N(\tau, 44^\circ, 17^\circ)$$

Only nucleotides that are not part of any base-pairing or stacking (as obtained by RNAView [YJL+03]) were included.

ALGORITHM 4.1: resampleSugar(segment)

resetDOFs ← all τ and χ in segment
resetValues ← Bimodal distribution for τ, random value for χ
recloseDOFs ← All backbone-torsions from P-C5' in the first residue to the C3'-O3' in the last
rigidDOFs ← All C2'-O2' bonds in segment
localRebuild(resetDOFs, resetValues, recloseDOFs, rigidDOFs)

ALGORITHM 4.2: localRebuild(resetDOFs, resetValues, recloseDOFs, rigidDOFs)

$E \leftarrow$ resetDOFs $\cup$ recloseDOFs $\cup$ rigidDOFs
$V \leftarrow$Vertices adjacent to E
$F \leftarrow$Covalent edge in E nearest the root of L
$B \leftarrow$edges in $L - E$ sharing exactly one end-vertex with edges in $E - F$
$P \leftarrow$Positions of endpoints of B
$T \leftarrow$Torsions of covalent edges in B
Freeze positions of all atoms except in V and B
Change DOFs in resetDOFs to values indicated in resetValues
$e' \leftarrow \infty$
$\sigma \leftarrow 0.1$
while $\left|e'\right| > 0.001$ **do**
  $e \leftarrow (P -$ Positions of endpoints of $B) \cdot \sigma$
  $J \leftarrow$ Jacobian(E, B)
  $J^\dagger \leftarrow$ PseudoInverse(J)
  $\Theta \leftarrow J^\dagger e$
  Change DOFs in recloseDOFs to values indicated in Θ
  if $|e| > 0.98 \cdot e'$ **then**
    $\sigma \leftarrow \sigma \cdot 0.1$
  end if
  $e' = e$
end while

1.6 Experimental design for validation

A *benchmark set* of sixty RNA molecules was compiled from the Biological Magnetic Resonance Bank (BMRB) [UAD+08] by downloading single-chain RNAs that contain more than 15 nucleotides and are solved with NMR spectroscopy. RNAs with high sequence similarity were removed so the edit-distance between the sequences of any pair was at least 5.

For each molecule in the benchmark set, the first NMR model is chosen as $\boldsymbol{q}_{\text{init}}$, and a pool of conformations are generated by repeatedly perturbing a seed conformation and placing the new conformation in the pool. The seed conformation is selected from the pool of existing conformations by picking a random non-empty interval of width $r_{\text{init}}/100$ between 0 and r_{init}. If there is more than one conformation in the pool whose distance to $\boldsymbol{q}_{\text{init}}$ falls within this interval, a completely random conformation is generated and the conformation nearest to the random structure is chosen as $\boldsymbol{q}_{\text{seed}}$. This guarantees that

samples in sparsely populated regions within the exploration radius are more likely to be chosen as seeds and that the sample population will distribute widely. A rebuild perturbation of two free nucleotides or a null-space perturbation is then performed at a 10/90 rate. A null-space perturbation can start from a seed generated by a rebuild perturbation or vice versa, allowing detailed exploration of remote parts of conformation space.

If a new conformation contains a clash between two atoms it is rejected and a new seed is chosen. An efficient grid-indexing method is used for clash detection by overlapping van der Waals radii [HO94]. The van der Waals radii were scaled by a factor 0.5.

The iMod toolkit [LBGC11] uses internal coordinates normal mode analysis (NMA) to explore conformational flexibility of biomolecular structures, for instance via vibrational analysis, pathway analysis, and Monte-Carlo sampling. The iMod Monte-Carlo sampling application was used for comparison with KGSrna and run with the default settings: heavy-atoms, 5 top eigenvectors, 1000 Monte-Carlo iterations per output structure, and a temperature of 300K.

2 Initial validation

To assess the performance of our model in representing RNA modes of deformation, we compared the distribution of our samples to the available NMR bundles. For this purpose, we performed sampling runs which all start from a single member of the NMR bundle and diffuse out to a predefined exploration radius. We define the *exploration width* as the ability of KGSrna to quickly diffuse away from the starting conformation and the *exploration accuracy* as the ability to sample conformations close to any biologically relevant member of the native ensemble. To evaluate the width and accuracy of the exploration we consider NMR models as representative members of the native ensemble and measure how close to KGSrna samples these members are, both in terms of local measures (τ-angle distributions) and in terms of full-chain measures (RMSD). KGSrna was used to generate 1,000 samples, starting from the first model of each of the sixty RNA structures in the benchmark set . The largest RMSD distance between any two models was used as the exploration radius for that molecule. The sampling is very fast as it took on average 6 minutes on an Intel Xeon E5-2670 CPU for molecules having around 60 nucleotides.

2.1 Broad and accurate atomic-scale sampling of the native ensemble

To assess the importance of the rebuilding procedure we evaluated the sampling with and without rebuild perturbations. Figure 4.4a illustrates distributions of the τ angle in KGSrna samples and NMR bundle structures for the Moloney MLV readthrough pseudoknot (PDB id 2LC8). Without any rebuilding step, KGSrna samples show a very narrow sampling in the geometrically constrained loop-region starting at nucleotide 40. With rebuilding enabled, the distributions of τ-angles widen significantly and all ribose conformations present in the NMR bundle are reproduced in the KGSrna sampling. When sampling without rebuilding, 9 out of the 196 nucleotides in the benchmark set that have both C3'-endo and C2'-endo conformations are fully recovered. When enabling rebuild perturbations all but four ribose conformations (98%) are recovered. These four are all in less common conformations such as O4'-endo or C1'-endo. Figure 4.5 shows the effects of KGSrna sampling with rebuilding on a δ/ϵ-plot.

Traditionally, ribose conformations are described using the pseudorotation angle P, which depends on all 5 torsions in the ribose ring [AS72]. Figure 4.4b shows the relationship between τ and P for all nucleotides in the benchmark set. While the two are not linearly related there is a monotonic relationship indicating that τ is as useful as P in characterizing ribose conformations in addition to being usable as a differentiable degree of freedom in a kinematic linkage.

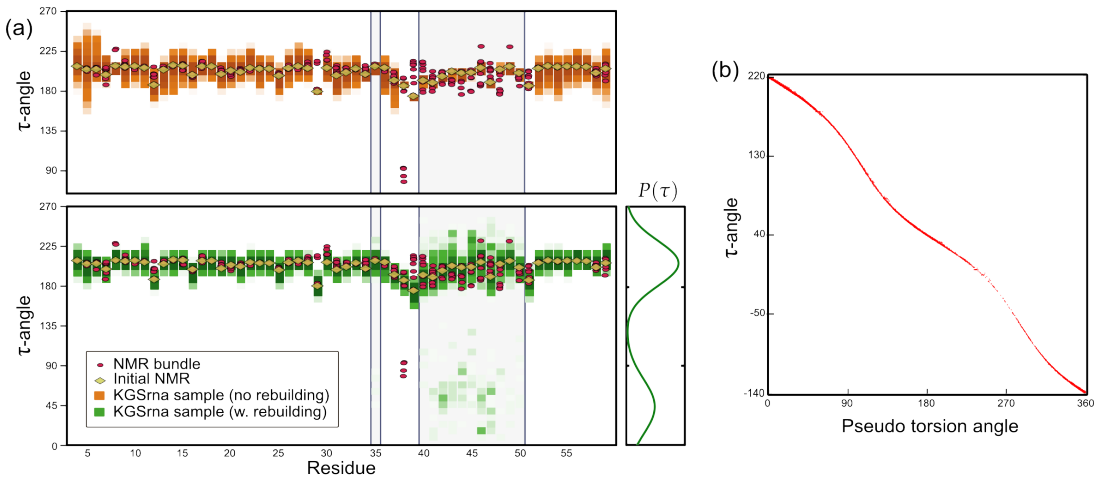


FIGURE 4.4 : KGS sampling illustrated by τ angle. a) Distributions of ribose conformations in KGS samples and in the NMR-bundle of MLV readthrough pseudoknot (2LC8). Ribose conformations of 1000 samples are displayed vertically as color-coded histograms with a bin-width of 1.8°. The top panel shows distributions without rebuilding steps and the bottom with rebuilding steps. Rebuild perturbations recover the full range of τ-angles in the NMR bundle for free nucleotides. The distribution from which τ-angles are sampled is shown on the right. The large peak corresponds to C3'-endo conformations and the smaller one to C2'-endo conformations. b) The relationship between the τ-angle and the pseudorotational angle introduced by Altona and Sundaralingam [AS72] for all nucleotides in the benchmark set. A monotonic relationship is observed indicating that τ is as expressive as P when characterizing ribose conformations, but can additionally be used as a differentiable DOF in a kinematic linkage.

2.2 Large scale deformations

We evaluated the performance of KGSrna in probing conformational states on whole-molecule scale using the RMSD of C4' coordinates after optimal superposition. Figure 4.6a shows the evolution of the minimum and maximum distance from each of the ten NMR bundle structures to the KGSrna sample of the Moloney MLV readthrough pseudoknot (PDB id 2LC8) as the sampling progresses. The sampling has expanded to the limits of the exploration radius after 400 samples. At this point it keeps populating the most sparsely populated region of the native ensemble. The minimum distance to each of the non-initial NMR bundle conformations quickly converges to approximately 2Å RMSD. Both these trends are consistent across the benchmark set with an average minimum RMSD of 1.2Å .

Regions of the molecule that are either constrained by tight sterics or by hydrogen bonds are difficult to deform, which is implicitly represented in the KGSrna model of flexibility. Figure 4.6b uses color-coding to highlight the regions of 2LC8 where the degrees of freedom show a particularly high variance. The base-paired regions that are tightly woven in a double helix show little flexibility while the unconstrained loop-region displays the highest degree of flexibility. Even though the O3'-terminal end (right-most side of Figure 4.6a) does not by itself display a large degree of flexibility it still moves over a large range as shown by the 25 randomly chosen overlaid KGSrna samples.

2.3 KGSrna as an alternative to NMA

The iMod Monte-Carlo application (iMC) is one of the state-of-the-art methods most directly comparable to KGSrna as it efficiently performs large conformational moves that reflect the major modes of

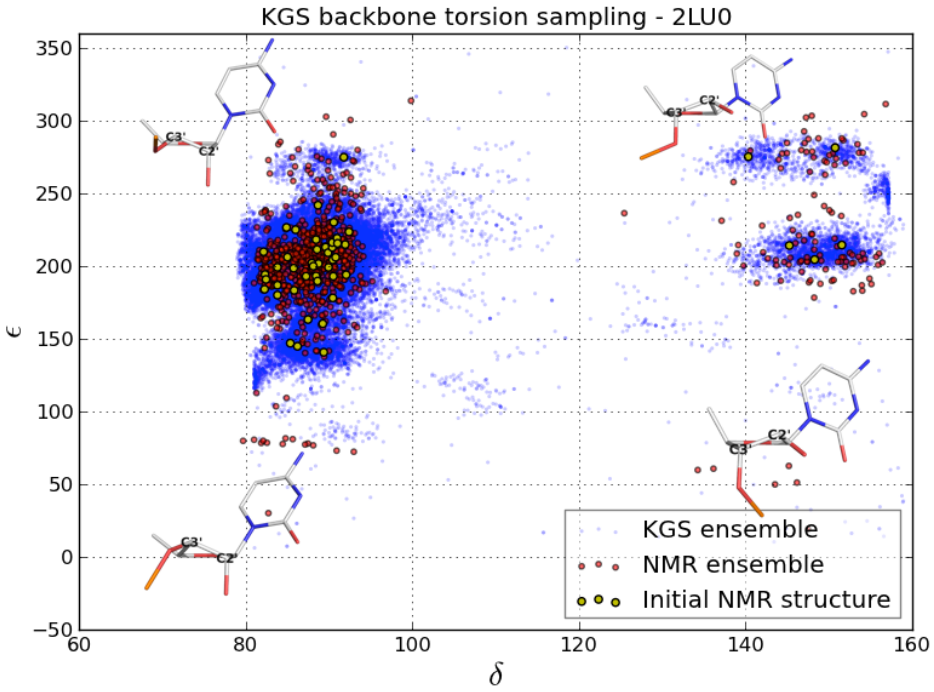


FIGURE 4.5 : Backbone $\delta - \epsilon$ torsional scatterplot of 1000 KGSrna samples of the S. cerevisiae group II intron (PDB id 2LU0). The left cluster usually correspond to C3'-endo and the right cluster to C2'-endo ribose conformations. KGSrna extensively samples both regions as well as intermediate ones. Richardson et al. [RSM+08] suggests that ribose conformations with $\epsilon < 155°$ corresponds to ribose conformations that have wrongly been fitted with C3'-endo conformations while they should have been C2'-endo. Interestingly, very few KGSrna samples lie in the region where $\epsilon < 100°$.

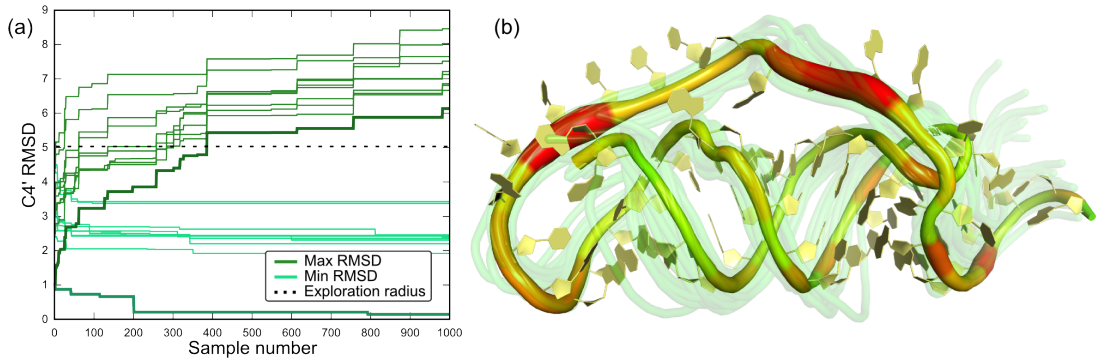


FIGURE 4.6 : Conformational exploration of KGSrna at molecular scale illustrated using the Moloney MLV readthrough pseudoknot (PDB id 2LC8). a) The evolution of smallest (lower bright-green curves) and largest (upper dark-green curves) RMSD as the sampling progresses. RMSD distances are measured to each of the 10 structures in the NMR bundle. The RMSD curves corresponding to the initial structure are indicated in bold. b) The conformation of the initial structure with 25 randomly chosen samples superposed. The color and thickness of the backbone indicates the degree of flexibility for nearby degrees of freedom. Very flexible regions are shown as thick and red-shifted while a region that remains rigid throughout all samples is thin and green.

deformation of biomolecules.

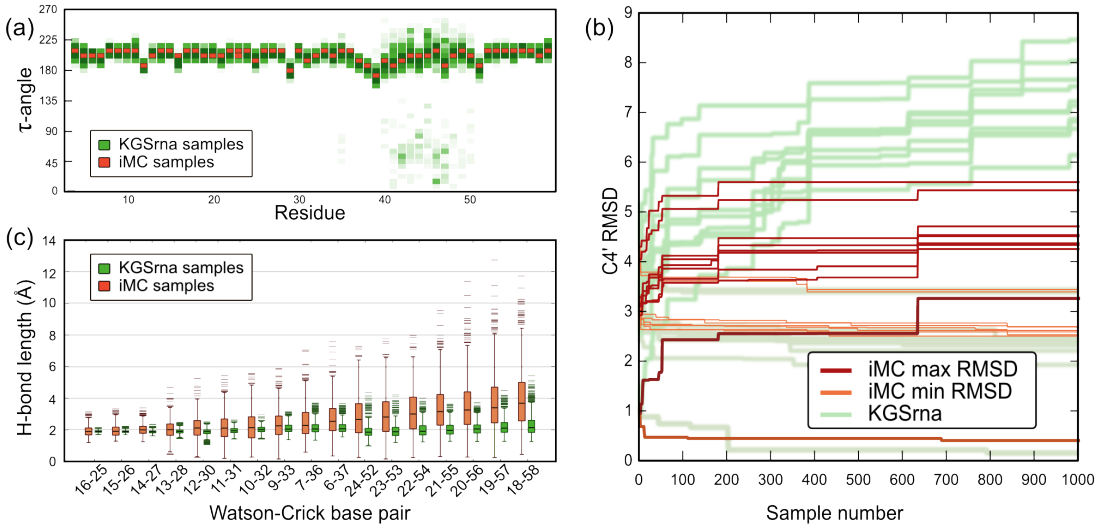


FIGURE 4.7 : a) Distributions of ribose conformations in 1000 iMC samples started from the same molecule and displayed on the same scale as KGSrna samples in Fig. 4.4. b) The evolution of minimum (light red curves) and maximum (dark red curves) C4' RMSD as the iMC sampling progresses. Minimum (resp. maximum) KGSrna curves are provided in light (resp. dark) green for reference. This panel is directly comparable to Fig. 4.6a. c) Distributions of hydrogen bond lengths in WC base pairs. The vast majority of samples generated by KGSrna has hydrogen bonds that fluctuate by less than 1Å. The same trend was observed over the rest of the benchmark set as well (data not shown).

Figures 4.7a and 4.7b show results of running iMC for 1,000 iterations on the Moloney MLV readthrough pseudoknot (PDB id 2LC8). While KGSrna is able to sample sugar conformations widely, the standard deviation of τ is less than 1° for all nucleotides in the iMC sample set. Furthermore, KGSrna samples widely and reaches the exploration radius of 5Å after 400 samples, while iMC has converged on 3.3Å after 1,000 samples. KGSrna generate structures closer than 2Å to an NMR bundle conformation while the best iMC conformation is just over 2.5Å from its nearest NMR bundle structure. This indicates both a broader exploration width and higher exploration accuracy of KGSrna compared to iMC.

Figure 4.7c shows distributions of hydrogen bond length in WC base pairs in the 1,000 samples from iMC and KGSrna respectively. The average standard deviation of hydrogen bond distances is 1.04Å for iMC base pairs which for most applications would constitute a full break of the bond. The standard deviation is only 0.33Å for KGSrna. The source of hydrogen bond fluctuations in KGSrna is primarily the null-space moves, where a relatively high step size causes the first-order approximations to introduce small deviations from the closure manifold.

C Extending biological results

1 Recovering proton chemical shifts

Chemical shifts are time-averaged measurements on conformational ensembles at sub-millisecond timescales [FHAAH13]. Non-exchangeable ^{1}H chemical shifts (CS) predicted directly from RNA three-dimensional structural models are generally in excellent agreement with those reported from experiments

in the BMRB [CHW01]. Experimental ^{1}H CS are widely available, are sensitive to conformational changes, and have aided in structurally characterizing conformational sub-states [SCC+14]. Researchers have combined measured ^{1}H CS for proteins with structure prediction algorithms that use a database of structural fragments to determine atomically detailed de novo conformations [SLD+08]. Das and co-workers recently established that proton chemical shifts can aid structure prediction algorithms in distinguishing decoys from a native state in RNAs [SCC+14].

In our work, we used ^{1}H CS on a benchmark set of three-dimensional RNA structures as time- and ensemble-averaged distributions over the conformational landscape. We examined the ability of KGSrna to sample native dynamical ensembles that recover sugar (H1') and nucleobase (H2, H5, H6, and H8) CS distributions for unconstrained (non-helical) and Watson-Crick paired (helical) regions. We used the program NUCHEMICS [CHW01] to predict ^{1}H CS from our three dimensional RNA structures.

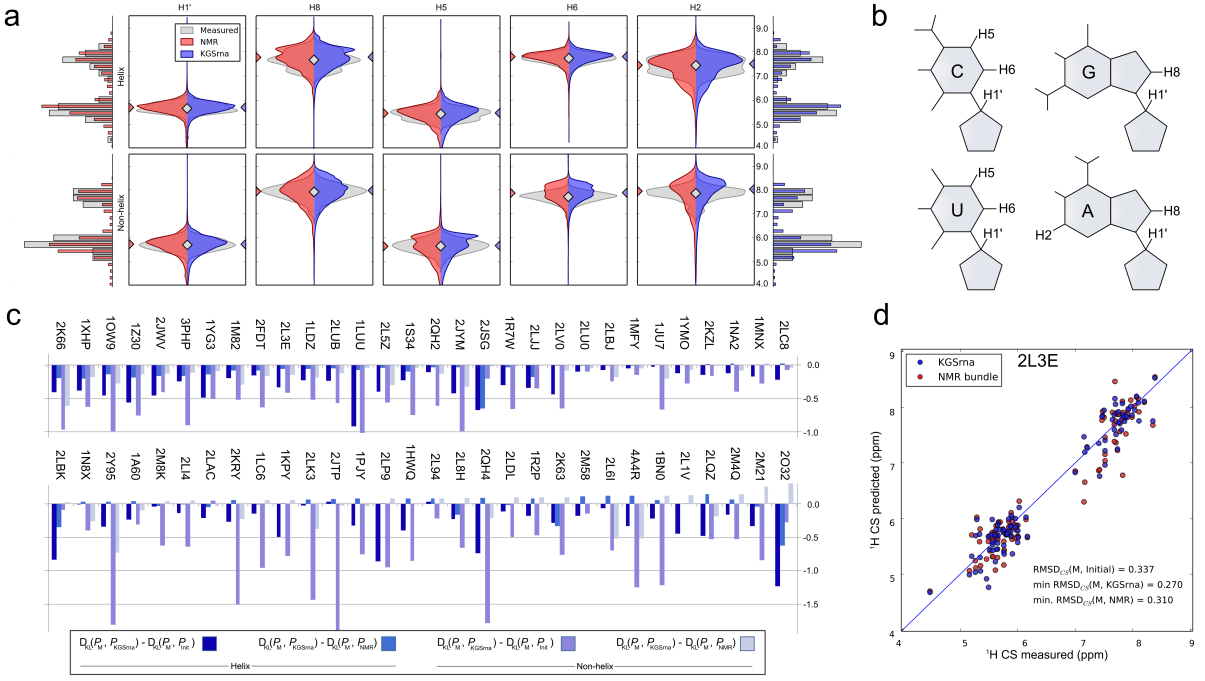


FIGURE 4.8 : Agreement between measured ^{1}H chemical shifts and those back calculated from KGSrna and NMR three-dimensional structures. a) Depicted chemical shift values are aggregated by proton type in helical (top) and non-helical regions (bottom). The discrete distributions were smoothed with a Gaussian kernel density estimator (bandwidth $n^{-0.2}$ where n is number of data points) for easier visualization. Measured values were taken from the BMRB and KGSrna samples and NMR bundle values were back calculated using NUCHEMICS. Marginal distributions are shown as histograms with bin-widths of 0.275ppm. b) Corresponding protons depicted on the chemical structure of each amino acid type. c) The symmetric Kullback-Leibler divergence indicates the degree of similarity of two distributions and is calculated for the marginal distributions of measured-to-KGSrna, measured-to-initial and measured-to-NMR. The differences between measured-to-KGSrna, measured-to-initial, and measured-to-NMR are shown in the bar plot. A negative value indicates better agreement of the ensemble of KGSrna 3D structures with measured values than its comparison 3D structures. d) Predicted ^{1}H chemical shifts calculated from the 3D structures from the KGSrna ensemble and the NMR bundle compared to the measured values of the 32nt P2a-J2a/b-P2b (helix-bulge-helix) of human telomerase RNA (PDB id 2L3E). The data points are expected to lie along a 45deg line if measured ^{1}H CS are accurately predicted.

KGSrna enables broad sampling to identify sparsely populated sub-states, while maintaining conformational distributions similar to those measured. Figure 4.8a shows the distribution of measured and predicted ^{1}H CS for helical (top row) and non-helical (bottom row) regions for each proton type over the whole benchmark set. Figure 4.8b shows the location of the probes. For helix backbone and base protons, the medians of the distributions are virtually identical. This suggests that, on average, our kinematic representation of RNA results in an unbiased exploration of the conformational landscape encoded in the measured proton chemical shifts.

For helical and non-helical regions, aggregate and individual sampling distributions of ^{1}H CS obtained with KGSrna are visually similar to the distributions obtained from experimental measurements. To further compare similarities between the measured chemical shift distributions P^M and the predicted distribution P^{KGSrna} we calculated the symmetrized Kullback-Leibler (KL) divergences $D_{KL}(P^M||P^{KGSrna})$ of P^{KGSrna} from P^M and compared those to the KL divergences $D_{KL}(P^M||P^{init})$ and $D_{KL}(P^M||P^{NMR})$ for our benchmark set (Figure 4.8c). The distributions P^{init} and P^{NMR} are the predicted distributions calculated from the first model of the NMR bundle only and the full NMR bundle. Both P^{KGSrna} and P^{NMR} deviate from P^M, in part owing to weighted motional averaging of the measured shifts. Similarities between the KGSrna predicted distributions and the measured distributions exceeded those of the predicted distributions from the first model. In 58 out of 60 of cases KGSrna improved agreement with the distribution of measured ^{1}H CS in non-helical regions (58 out of 60 for helical regions too) compared to the distribution calculated from the first model. The average KL divergence reduction was 33% (39% for helical regions). This suggests that KGSrna is able to diverge from the starting model, and explores beyond a local neighborhood of conformational space. In addition, in 70% of cases KGSrna improved agreement with the distribution of measured ^{1}H CS in non-helical regions (58% for helical regions) compared to the distribution calculated from the full NMR bundle. Predictions for non-helical regions were improved by our rebuilding procedure, conceivably resolving structural disorder inadequately represented by the NMR bundle [CHW01]. The similarities between predicted and measured distributions suggest that a simple kinematic model with constraints samples the conformational landscape according to the same distribution as RNA in solution.

We then asked how accurately just a single KGSrna sample could recover measured chemical shifts. The error between measured and predicted CS is attributable to measurement errors and systematic errors in prediction. Additionally, measured chemical shifts are a weighted motional average. We therefore regarded the NMR three-dimensional conformer that best agrees with measured chemical shifts as a benchmark of predictive value.

We calculated the RMSD ($RMSD_{CS}$) between the measured and predicted chemical shifts for all proton types for each three-dimensional model in the NMR bundles and in the KGSrna sample sets. The minimum $RMSD_{CS}$(M,KGSrna) ranges from 0.17 to 0.54 ppm (mean 0.30ppm) and the minimum $RMSD_{CS}$(M,NMR) from 0.16 to 0.53 ppm (mean 0.30ppm). A recent study observed a mean minimum weighted $RMSD_{CS}$ of 0.23ppm (ranging from 0.16 to 0.35ppm) for an ensemble of 8 000 conformers obtained from molecular dynamics simulations for four RNAs, but the proton chemical shifts were weighted to favor those that better agreed with measured values [FHAAH13]. In 80% of cases in our benchmark set the $RMSD_{CS}$ of the best KGSrna conformer is lower than that of the best conformer identified from the NMR bundle (Figure 4.8d). The average improvement over the starting model is 18%, and in some cases exceeds 40%. As proton chemical shifts can discriminate a native state, this result suggests that a simple kinematic representation yields a powerful conformational search algorithm.

2 Application: the HIV-1 TAR hairpin loop excited state

The 5'-end of the human immunodeficiency virus type-1 (HIV-1) transcript contains a 59-nucleotide trans-activation response element (TAR) stem-loop [HGWB04]. In the ground state, HIV-1 TAR binds human cyclin T1 and viral trans-activator protein Tat that activate and enhance transcription of the HIV-1 genome [AeKV96, DHM+08, LLXZ13, TBV+10]. The HIV-1 TAR apical hairpin loop plays a key role in binding Tat. Available structures for the HIV-1 TAR apical loop exhibit significant conformational differences, which indicate that the loop is highly flexible. However, a full atomic characterization of the structure and dynamics of the HIV-1 TAR hairpin loop remains elusive. Al-Hashimi and co-workers recently proposed a two state model (ground and excited state, GS and ES) of the apical HIV-1 TAR hairpin loop from NMR $R_{1\rho}$ relaxation dispersion measurements and mutagenesis. Their study suggested formation of a U31 G32 G33 G34 tetraloop in the ES, with a non-canonical closing base-pair C30-A35.

Residual dipolar couplings (RDCs) report the amplitude of motions that reorient C-H and N-H bond vectors on the sub-millisecond time-scale. Experimentally observed RDCs are a weighted average of all conformational substates.

To test if KGSrna can structurally characterize conformational substates of the HIV-1 TAR hairpin loop guided by RDC data, we calculated 20 000 samples each starting from models one to ten in the NMR bundle with PDB id 1ANR of free HIV-1 TAR. To enable structural characterization of the dynamics leading to the ES, we biased our sampling towards broad, non-specific conformational pairing of C30-A35 and U31-G34. A Metropolis criterion skewed the sample set to include favorable interactions of any charged hydrogen in base A with any hydrogen acceptor in base B. For each of the 200 000 samples, we back-calculated RDCs with the program PALES [Zwe08]. From each batch of 20 000, we then determined a weighted ensemble that optimally explained the experimentally observed residual dipolar couplings using a new constrained quadratic fit algorithm (rdcFit) that we adapted from an application we previously developed for X-ray crystallography applications (qFit) [vBY+13, vDLD09].

This procedure identified a ten-member, weighted ensemble from the sample set starting from model seven in the NMR bundle that agrees extremely well with experimentally observed RDC values (4.9a). The coefficient of determination between observed $^1D_{CH}$ values and those predicted from the weighted ensemble equals 0.98. The predicted values of the ensemble accurately reflect the mobility of riboses and nucleobases, with $^1D_{CH}$ small in magnitude indicating elevated mobility (4.9b). The RMSD between observed and predicted $^1D_{CH}$ values is 1.55Hz, below the experimental error of 2-4Hz [FSAHA09, SBAAH13].

Our ensemble characterizes disparities in mobility between nucleotides in exquisite atomic detail, consistent with the RDC data (Figure 4.9a inset). In our ensemble nucleobases U31, G32 and A35 are most mobile, with motions indicating looping in and out. Small magnitudes of U31 $^1D_{C6H6}$, G32 $^1D_{C8H8}$, and A35 $^1D_{C8H8}$ experimental values support this interpretation (Figure 4.9b). In the conformation of our ensemble most closely exhibiting features attributed to the GS, we confirmed the formation of a stabilizing cross-loop Watson-Crick bp C30-G34 [KOH+03] (Figure 4.9c, left). Experimental $^1D_{C8H8}$ data do not appear to support a similar large amplitude motion of the G34 base when adjusting from anti to syn. Instead, our ensemble suggests that G34 gently readjusts to accommodate A35 looping in.

To confirm this intermediate state towards the ES, we generated an additional 20 000 samples starting from this conformation, instructing KGSrna to further pair these nucleotides. In the model with most ideal hydrogen-bond geometry between these bases, we observe ribose conformations suggesting that C30 and A35 are adopting a C3'-endo conformation, continuing the A-form helical stem from bp C29-G36. To examine if the ES is kinetically accessible from this intermediate state, we started fifteen independent, 100ns molecular dynamics (MD) simulations. Consistent with the transient character of

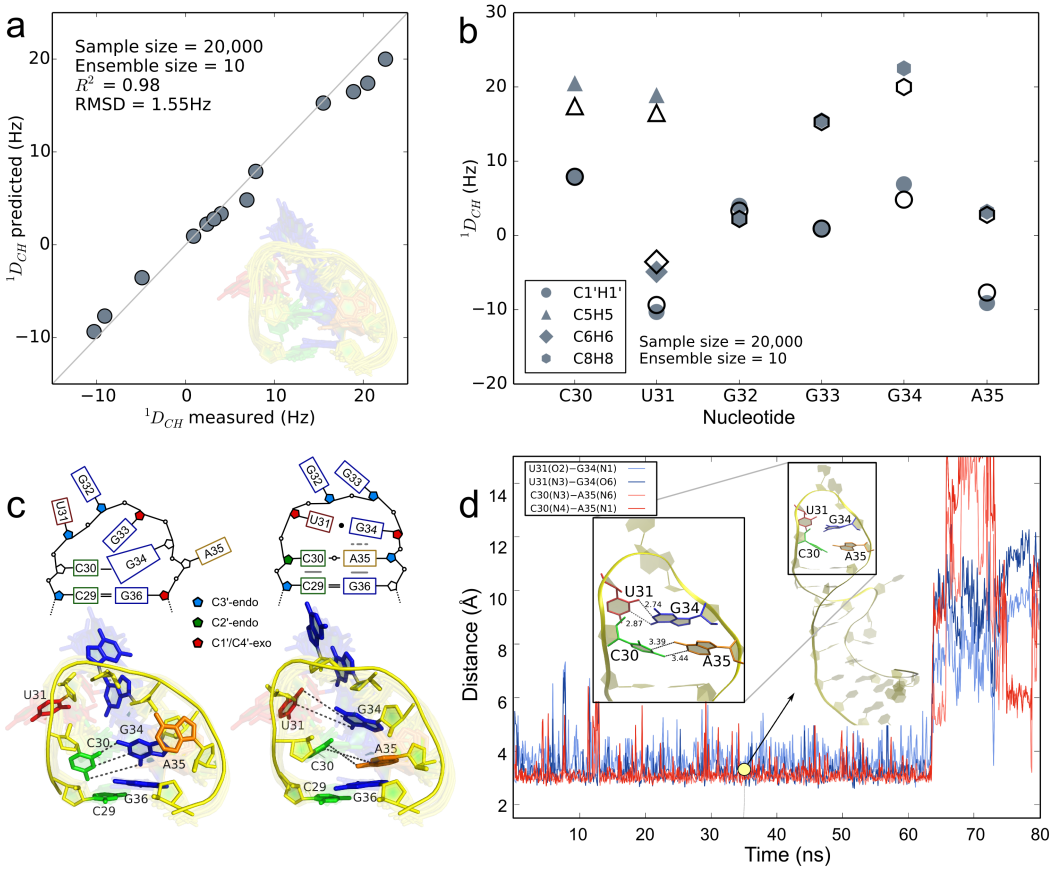


FIGURE 4.9 : Structural characterization of conformational substates of the apical Tat binding loop of HIV-1 TAR. a) Paired measured RDCs for apical loop nucleotides and those predicted from a 10-member weighted ensemble (inset) obtained from fitting 20,000 KGSrna samples to measured RDCs with a quadratic program. The data points are expected to lie along a 45deg line if measured RDCs are accurately predicted. The coefficient of determination for the predicted RDCs equals 0.98. b) Observed (solid symbols) and predicted (open symbols) RDCs for apical loop nucleotides. Smaller magnitudes for RDCs generally indicate more angular mobility in the bond vectors. c) Schematic of the GS (top panel, left) and the ES (top panel, right) corresponding to the three-dimensional structures closest to the GS and the ES in the ten-member ensemble. The bottom left panel shows the GS highlighted in the ensemble, with the other members translucent in the background. The bottom right panel shows the ES identified from biasing the sampling towards pairing C30-A35 and U31-G34. d) Time evolution of the hydrogen bond distances between reverse wobble pair C30-A35 (blue colors) and GU wobble pair U31-G34 (red colors) in the ES of HIV-1 TAR for 80ns of the molecular dynamics trajectory. The distances shown are between heavy donor and acceptor atoms, sampled every 100ps. Along the trajectory, the apical loop maintains a helical structure (inset at 35ns) until, at 65ns, pairing of U31-G34 and subsequently C30-A35 is disrupted.

the ES, the CA reverse wobble pair and GU wobble pair were maintained for 20-65ns in duration for four out of fifteen or 27% of the MD simulations. In the remaining simulations, pairings did not occur or were short-lived. Figure 4.9d shows the evolution of the distances of the hydrogen bonds between the CA pair and the GU pair for the MD simulation that maintained pairing for nearly 65ns. G34 forms additional hydrogen bonds G34(N2)—U31(O4') and G34(N1)—C30(O2') that stabilize the ES. The riboses of C30 and U31 adopt a C2'-endo conformation for the duration of U31-G34 pairing, after which C30 adopts a C3'-endo conformation. G32 and G33 intermittently stack during the simulation (4.9d inset). To our knowledge, this is the first time sustained and simultaneous pairing of C30-A35 and U31-G34 was observed in MD simulations of HIV-1 TAR.

D Outcome and future directions

MD simulations can provide detailed insight into the dynamics and spatiotemporal mechanisms of RNA molecules. Barriers in the free energy landscape often prevent these techniques from adequately sampling a representative set of conformational substates. By contrast, non-deterministic sampling algorithms coupled with simplified, knowledge-based potentials provide no information on dynamics but can broadly explore the conformational landscape [BHSL11, DB07]. Our analysis demonstrates that conformational ensembles of non-coding RNAs in solution are accessible from efficiently sampling coordinated changes in rotational degrees-of-freedom that preserve the hydrogen bonding network. Each member of a synthetic ensemble was approximated to within 2Å on average by a KGSrna sampled conformation on a benchmark set of sixty noncoding RNAs without relying on a force field. By contrast, an NMA-based sampling algorithm diffuses through the folded state at a slower rate, approximating each ensemble member with 25% less accuracy.

Hydrogen bonds and similar non-covalent constraints, like hydrophobic interactions, encode preferred pathways on the conformational landscape, enabling our procedure to efficiently probe the conformational diversity resulting from equilibrium fluctuations of the ensemble. The approach is generic, atomically detailed, mathematically well-founded, and makes minimal assumptions on the nature of atomic interactions. Combined with experimental data, it can provide insight into which substates are adopted. This procedure is easily adapted to DNA, and protein-protein or protein-nucleic acid complexes. It will certainly be of great help for improving docking experiments [VHK13].

Contents

CHAPTER 5

CONCLUSION AND PERSPECTIVES

A From biophysics to data: towards a multi-scale, multi-approach framework

1 The data challenge

The studies presented here all rely on an extensive extraction and curation of biological 3D data. Fortunately, the most valuable source of information, the PDB, is freely and conveniently available. The PDB format was however defined more than 40 years ago. Aside from a limited remediation [HFB+08], successive attempts at making the PDB available using a format more convenient for computational purposes have failed in being widely adopted and/or maintained in the long term [HFB+08, WIN+05]. The same applies to specialized databases and datasets which can be used as training sets and benchmarks. Docking benchmarks are being updated relatively frequently, but the existence of most of the datasets of added value is limited in time and suffers from a variety of uneven choices (sometimes erroneous).

The relatively quick and recent changes in experimental procedures, mainly the development of low-resolution techniques such as electron microscopy (EM) or small angle X-ray scattering (SAXS) requires to be able to handle information at a wide range of scales. The data format used are currently specific for each experimental technique and would really benefit from being better integrated in databases for computational analysis. Efforts must be made: a possible example could be the computer vision datasets being made available to the machine learning community.

The CAPRI community is essential for such advances to occur. The release of a scoring set in 2014 [LW14] was great for fostering the development of data-based techniques. The community should however trigger the early release of the similar datasets for novel comparisons such as prediction of

binding affinity, ion and water binding sites or effects of mutagenesis. This would allow for faster development of various models, and enhance process of computational structural biology.[12].

Dealing with data in structural biology is hard, in particular for students as they often need to extend their training to data curation and mining. While the size of the training data might seem reasonable with relation to modern problems, the extraction of this training data and the synthetic generation often require broad skills. The first set of skills is technical and required: the knowledge of computer clusters and large scale computing platforms. Also necessary is algorithmic thinking (albeit a deep, theoretical understanding is not necessary), and creativity to identify optimal algorithms for various problems like randomly drawing and sorting from large datasets.

Ideally with the current interest in data science classes and machine learning, most of these problems will be solved in a near future. This might not solve a more biological aspect of the problem: most of the leading techniques are currently being developed by structural biologists who were later trained in computer science. In computational structural biology, a large amount of training required to extract good features and build good models. It requires not only biology knowledge but also physics and chemistry. Gathering a group of students and people sharing these interests is hard and the culture gap is often an issue for research. To succeed, research teams have to be relatively large with close collaboration between group members.

2 Gathering techniques for an integrated model

The setting up of a model for computational structural biology requires a good understanding of the structures and other experimental data. This sharing of the different cultures is one of my favorite parts in the studies I presented here. The collaboration with Jérôme Azé in machine learning was essential for the scientific results but also for developing my interest in a computer science community I was not introduced to in my studies. I have also greatly benefited from fruitful collaborations in the fields of geometry and robotics.

On the structural biology side, I have been extremely lucky to benefit from the trust of experienced colleagues. While I did not present these studies, I had the chance to provide computational expertise for studies on different systems of biological interest: a laboratory management system [POU+05], the structure of a putative antibacterial target [GBS+06], a large RNA complex dynamics analysis [RSMS+14] and a dynamics model of a cancer target [JSL+ed]. These studies were essential for me in staying close to experimental problems and better understanding data. Ideally, provided we succeed in developing an integrated framework from the studies presented here, these kind of computational analyses will be made much more accessible to experimentalists and allow for better interpretation and prediction on how large biomolecule assemblies work.

The different studies related here rely on the development of a variety of computer science related techniques. They do however always invoke two concepts related to physics: sampling and optimization. In the work aforementioned, machine learning, statistics and kinematics were represented. Many other ways can be explored. In the perspectives below (see B), I will present the current attempts at further developing models to complement our current studies.

[12] An interesting example is the Kaggle platform (see `https://www.kaggle.com/`). A way to encourage interest from students and prospects could be to develop such a platform for difficult biology related academic experiments.

3 Biological interpretation: needs and limitations

A hard part in promoting the use of data for prediction in structural biology is often the difficulty in providing a physical model. In chapter 3, we saw that many hypotheses and approximations were made, most of them being obviously far from reality or from chemistry. In [BBR+08], while we build an accurate model, providing feedback on the biophysical phenomenon is hard. In [FPBv14], we claim that we can assess the dynamics of a RNA structure but have no way of proving this is true for all structures.

Overall, obtaining a biological explanation from a data-based model is hard. It would require an experimental validation. This validation might not be possible or we might not have the necessary expertise or means. Experiments required to validate computer generated models are also often different from experiments of interest in research. Recent initiatives such as FoldIt[13] or EteRNA[14] show that this is not necessarily out of reach, provided we are able to obtain a generic enough platform, for example one that allows for gaming or crowdsourcing.

In any case, should these validations be made in collaboration with experimentalists, it would undoubtedly be beneficial to both the structural biology and the computer science communities, triggering new research problems.

B Future directions: extending computer science models for biology

1 Prospective projects

The section below describes projects that I started addressing in the last year and that I think would be interesting in complementing the approaches I discussed before. In the longer term, the aforementioned models will be further pursued with the goal of developing a full docking procedure that can predict large assemblies, even at high-resolution. I want however to go on exploring new computer and data science approaches so as to complement and extend the framework.

2 Game theory sampling for RNA

This work, performed in collaboration with Johanne Cohen and her Ph.D. student Mélanie Boudard, addresses the sampling of 3D RNA molecules from a very coarse-grained perspective. We use a graph-based representation of RNA where each secondary structure element (SSE) is represented as one or a few nodes in a graph (see Figure 5.1). The nodes are linked by the covalent bond connections, and non-bonded interactions are represented using various types of KB potentials. Game theory is used to observe the system evolve and provide putative conformations: the nodes are the players of *sampling* games. Aside from proof-of-concept studies for RNA backbones [LQV+13] or different bioinformatics applications [BHW+14], game-theory approaches for RNA structure predictions have barely been studied.

Finding RNA conformations that satisfy structural and potential constraints can be seen as a local optimization problem where each secondary structure element (SSE) or *player* (or set of players) tries to maximize its *payoff* function (which is equivalent to minimizing its energy function). Game theory is a suitable tool for understanding systems in which the players have preferences for certain solutions. In these systems, the outcome of an action taken by one of the players depends not only on his own action

[13] `http://fold.it`
[14] `http://eterna.cmu.edu`

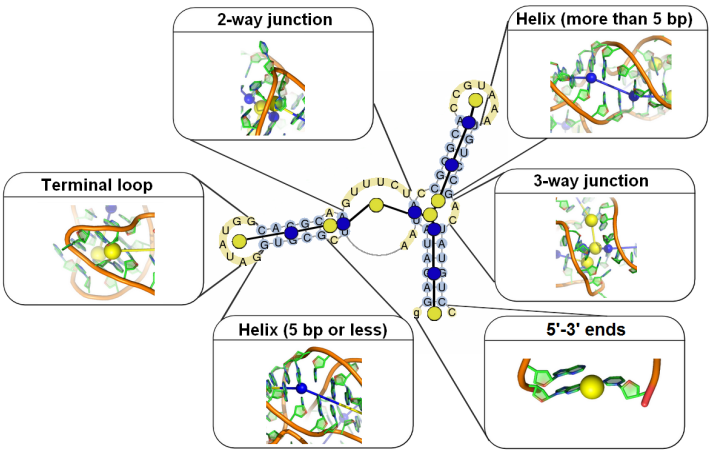


FIGURE 5.1 : Graph representation of the xpt-pbuX guanine riboswitch aptamer domain (PDB id 4FE5). Nodes (players) are build from the secondary structure representation where base-paired (respectively free) nucleotides are shown in blue (resp. yellow). Blue nodes correspond to helices and yellow nodes to junctions. Each element but the 3-way junction is represented by one node (shown as a sphere) taken to be the geometric center of the heavy atoms. The 3-way junction contains two nodes accounting respectively for helical stacking and branching.

but also on the actions taken by the others. In game theory: (i) the strategy set of an action is called *the set of pure actions* available, (ii) an action can be a distribution law and (iii) a *mixed strategy* is an assignment of a probability to each pure action. This allows a player to randomly select a pure action according to the mixed strategy.

An important result is the Nash theorem [Nas50]: every game with a finite number of players and a finite number of pure strategies has at most one Nash equilibrium in mixed strategies. The Nash equilibrium can be considered as a stable solution and can be interpreted as the states where the system will converge if the players are rational. Despite the fact that finding the Nash equilibrium is hard [DGP09], the corresponding tools can be used to understand the behavior of a such system and make predictions.

While there is no guarantee for this approach to be successful, a first way to obtain a Nash equilibrium is to use simple algorithms where each player selects a best response strategy evaluated relatively to the decisions taken by the other players. Even if a probabilistic version can be efficient, major improvements have been obtained by studying the dynamics of these algorithms [GK95].

Another approach is the *multi-armed bandit problem* [Rob85] where a gambler player faces a numerous number of times on different slot machines. At each step, it chooses the machine on which it will play and receives a gain. The goal is to maximize the total gain, i.e. the sum of gains received at each pull, taking into account the history of previous pulls. When gains depend on a fixed probability law and pulls are independent, the aim of the player is measured in terms of expected loss.

In our study we use regret minimization [LS82], where the expected loss after several play rounds is minimized. From a reference set of RNA structures we build a KB scoring framework that can be used in a lattice setting where the RNA secondary structure elements (SSEs) made of players evolve. We show that game theory strategies are efficient in sampling various confirmations efficiently (see Figure 5.2). This is especially true for relatively large molecules exhibiting complex substructures such as 3-way junctions. Our strategy is also fast, providing a valuable tool for conformational 3D structure search and

folding.

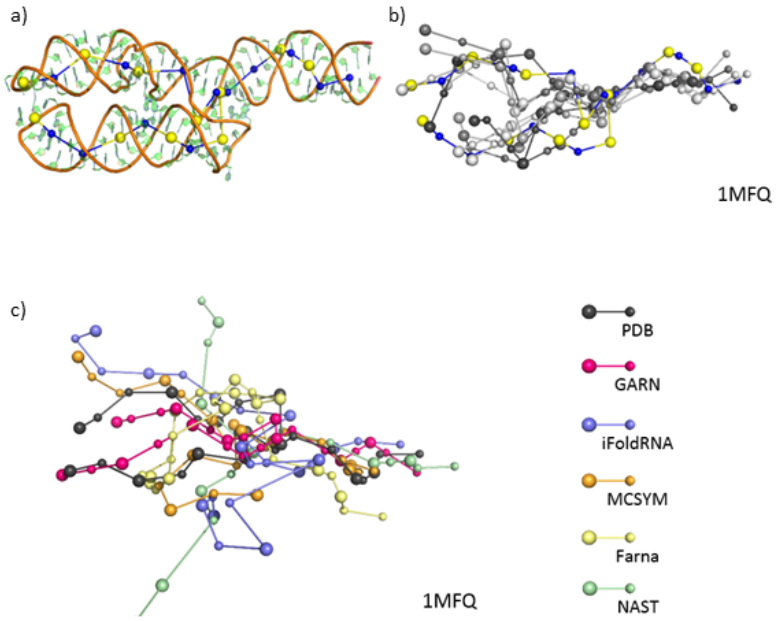


FIGURE 5.2 : Results of the game theory approach. a and b) Visualization of near-native samples for the *test set*. The native structure graph is superposed with the x-ray structure. Graph superposition show the native structure graph (in blue and yellow) represent the native X-ray structure well for each case. The closest to native graphs in grey (the darker the closer) superposed to the native structure show a good range of samples that could be used for reconstruction: the global shape of the molecule is recovered and the junction presents an interesting geometry. c) Comparison to other RNA structure prediction methods. Our best model (in pink) and the equivalent coarse-grained models obtained from other techniques (when available) are superposed to the native structure graph (in black). While not enforcing packing, the game-theory technique often provide the closest solution.

A first article is under review and we expect to extend the model to more complex problems, such as sampling of proteins and complexes. This project is part of the Ph.D. of Amélie Heliou I am co-supervising with Johanne Cohen.

3 Kinematics and clustering for docking

Protein-RNA complexes are especially difficult to predict and model because of the flexibility of RNA molecules as briefly seen in Chapter 2. Structure prediction and folding techniques, especially for RNA [DB07, DKB10, PM08, RRPB11], can be useful when trying to predict the conformation of the partners before docking. As mentioned in Chapter 3, resorting to molecular dynamics or normal modes which also have been used for docking [Rit08, VR12], the computation is very expensive and the analysis of the preliminary results might be tedious, forbidding such strategies in the CAPRI context or for the analysis of large biological systems. The protein-RNA benchmarks I and II [PCJGFR12, BCPB12] contain such examples of bound RNAs whose structure is significantly different from the unbound equivalent. To sample the bound structure, we use the Kino-Geometric Sampler (KGS) presented in Chapter 4 for protein and RNA molecules. Protein domain dynamics are shown to be accurately modelled by KGS

using an analysis of the H-bonding network.

We then attempt cross-docking from the obtained structures. To perform cross-docking, the starting conformations also have to be selected carefully [LSMD+13] so as to keep the search complexity reasonable. Many clustering techniques are used in experimental biology, molecular dynamics [BBL+11], and docking [KCVC05]. Most of these techniques use bin partitioning or centroid based clustering relying on the Voronoi diagram and often cannot account well for the density of the clusters. We show that Gaussian mixture models can be used to overcome that difficulty for biological configurations exactly like in building of knowledge-based potentials [BHSL11, SSLB12]. For several protein-RNA complexes, for which the unbound structure of both partners are known, we show that we can recover the properties of the bound structure from the sampling of the unbound structure with less than five cluster representative structures (see Figure 5.3).

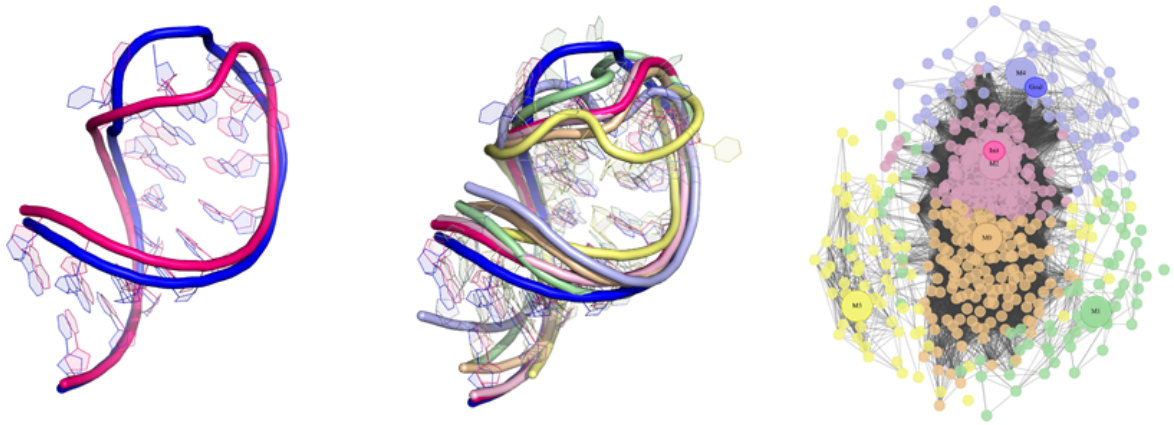


FIGURE 5.3 : Clustering RNA samples for docking: the SMAUG recognition element example (PDB id 2B6G). The two leftmost panels show the bound structure in pink, the unbound structure in blue and the five structures obtained by clustering. The right panel shows a multi-dimensional scaling style projection displaying clusters and centers using the same color scheme.

We then use the optimized version of RosettaDock for protein-RNA complexes Adrien Guilhot developed during his Ph.D. [GGFAB14] to perform different types of experiments. On an easy target where the bound structures of the partners does not differ very much from the unbound, we show that the docking results are very much enriched by the runs on cluster representatives. The extension to a full cross-docking run for a more intricate complex structure show that while still being computationally expensive because of the docking stage, the docking results are very much improved: while neither bound or unbound runs provide good solutions, the representative conformations provide interesting seeds for local refinement. The sampling strategy being extremely fast, efficient flexible docking is made possible and can be incorporated in different docking search methods. This work was performed by Amélie Heliou during her Master's internship and Rasmus Fonseca during his post-doctoral research. We are currently in the process of publishing this work and extending it to larger complexes of biological interest.

BIBLIOGRAPHY

[ABH+11] J. AZÉ, T. BOURQUARD, S. HAMEL, A. POUPON, and D. W. RITCHIE. Using Kendall-τ meta-bagging to improve protein-protein docking predictions. In *Pattern Recognition in Bioinformatics*, pages 284–295. Springer, 2011. 19

[AeKV96] F. ABOUL-ELA, J. KARN, and G. VARANI. « Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. ». *Nucleic Acids Res*, 24(20):3974–3981, 1996. 49

[ALS03] J. AZÉ, N. LUCAS, and M. SEBAG. « A new medical test for atherosclerosis detection: Geno ». *Proceedings of Discovery Challenge PKDD*, 2003. 18

[ALS04] J. AZÉ, N. LUCAS, and M. SEBAG. « A genetic roc-based classifier ». Technical report, LRI, Université Paris Sud, 2004. 16

[Anf73] C. B. ANFINSEN. « Principles that govern the folding of protein chains. ». *Science*, 181(4096):223–230, 1973. ..... 26, 31

[ARKS05a] J. AZÉ, M. ROCHE, Y. KODRATOFF, and M. SEBAG. « Preference Learning in Terminology Extraction: A ROC-based approach ». In *Proceedings of ASMDA'05 (Applied Stochastic Models and Data Analysis), Brest, France*, pages 209–219, 2005. 18

[ARKS05b] J. AZÉ, M. ROCHE, Y. KODRATOFF, and M. SEBAG. « Learning to order terms: supervised interestingness measures in terminology extraction ». *International Journal on Computational Intelligence*, 1(2):98–102, 2005. 18

[AS72] C. ALTONA and M. SUNDARALINGAM. « Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. ». *J Am Chem Soc*, 94(23):8205–8212, 1972. ..... 38, 43

[ASJ+02] B. ANGELOV, J. SADOC, R. JULLIEN, A. SOYER, J. MORNON, and J. CHOMILIER. « Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. ». *Proteins*, 49(4):446–56, 2002. 10

[Aur87] F. AURENHAMMER. « Power diagrams: properties, algorithms and applications ». *SIAM Journal on Computing*, 16(1):78–96, 1987. 12

[BAJP05] J. BERNAUER, J. AZÉ, J. JANIN, and A. POUPON. « Une nouvelle fonction de score pour l'amarrage protéine-protéine fondée sur les diagrammes de Voronoï ». In *Journées Ouvertes de Biologie Informatique Mathématiques*, 2005. 145

[BAJP07] J. BERNAUER, J. AZÉ, J. JANIN, and A. POUPON. « A new protein-protein docking scoring function based on interface residue properties. ». *Bioinformatics*, 23(5):555–562, 2007. . . . . 8, 10, 13, 16, 19, 20, 23, 145, 146

[BBAP09] T. BOURQUARD, J. BERNAUER, J. AZÉ, and A. POUPON. « Comparing Voronoi and Laguerre tessellations in the protein-protein docking context ». In *Sixth annual International Symposium on Voronoi Diagrams*, Copenhagen, Denmark, 2009. . . . 8, 10, 13, 20, 145

[BBAP11] T. BOURQUARD, J. BERNAUER, J. AZÉ, and A. POUPON. « A collaborative filtering approach for protein-protein docking scoring functions. ». *PLoS One*, 6(4):e18541, 2011. 8, 10, 19, 20, 23, 79, 147

[BBB+00] H. BERMAN, T. BHAT, P. BOURNE, Z. FENG, G. GILLILAND, H. WEISSIG, and J. WESTBROOK. « The Protein Data Bank and the challenge of structural genomics. ». *Nat Struct Biol*, 7 Suppl:957–9, 2000. 2

[BBL+11] K. A. BEAUCHAMP, G. R. BOWMAN, T. J. LANE, L. MAIBAUM, I. S. HAQUE, and V. S. PANDE. « MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. ». *J Chem Theory Comput*, 7(10):3412–3419, 2011. 58

[BBR+08] J. BERNAUER, R. P. BAHADUR, F. RODIER, J. JANIN, and A. POUPON. « DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. ». *Bioinformatics*, 24(5):652–658, 2008. . 8, 10, 13, 19, 20, 55, 147

[BCPB12] A. BARIK, N. C, M. P, and R. P. BAHADUR. « A protein-RNA docking benchmark (I): nonredundant cases. ». *Proteins*, 80(7):1866–1871, 2012. 58

[BDMV13] A. N. BORKAR, A. DE SIMONE, R. W. MONTALVAO, and M. VENDRUSCOLO. « A method of determining RNA conformational ensembles using structure-based calculations of residual dipolar couplings. ». *J Chem Phys*, 138(21):215103, 2013. 36

[BFH+11] J. BERNAUER, S. FLORES, X. HUANG, S. SHIN, and R. ZHOU. « MULTI-SCALE MODELLING OF BIOSYSTEMS: FROM MOLECULAR TO MESOCALE - Session Introduction. ». *Pacific Symposium on Biocomputing*, pages 177–80, 2011. 9

[BGNC09] B. BOUVIER, R. GRÜNBERG, M. NILGES, and F. CAZALS. « Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition. ». *Proteins*, 76(3):677–692, 2009. . . . . 10, 20

[BHSL11] J. BERNAUER, X. HUANG, A. Y. L. SIM, and M. LEVITT. « Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. ». *RNA*, 17(6):1066–1075, 2011. . . . . 10, 25, 32, 33, 40, 51, 58, 79, 147

[BHW+14] K. BOHL, S. HUMMERT, S. WERNER, D. BASANTA, A. DEUTSCH, S. SCHUSTER, G. THEISSEN, and A. SCHRÖTER. « Evolutionary game theory: molecules as players ». *Molecular BioSystems*, 2014. 55

[BKP87] C. L. BROOKS, M. KARPLUS, and B. M. PETTITT. *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*. J. Wiley, 1987. 26

[BMB05] P. BRADLEY, K. M. S. MISURA, and D. BAKER. « Toward high-resolution de novo structure prediction for small proteins. ». *Science*, 309(5742):1868–1871, 2005. 27

[BNE+11] J. R. BOTHE, E. N. NIKOLOVA, C. D. EICHHORN, J. CHUGH, A. L. HANSEN, and H. M. AL-HASHIMI. « Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. ». *Nat Methods*, 8(11):919–931, 2011. 36

[Bou09] T. BOURQUARD. « *Exploitation des algorithmes génétiques pour la prédiction de structure de complexe protéine-protéine* ». Theses, Université Paris Sud, 2009. 13

[BPAJ05] J. BERNAUER, A. POUPON, J. AZÉ, and J. JANIN. « A docking analysis of the statistical physics of protein-protein recognition. ». *Phys Biol*, 2(2):S17–S23, 2005. . 8, 10, 13, 20, 22, 29, 146

[BRD99] R. T. BATEY, R. P. RAMBO, and J. A. DOUDNA. « Tertiary motifs in RNA structure and folding ». *Angewandte Chemie International Edition*, 38(16):2326–2343, 1999. 26

[BVH+11] G. BOUVIGNIES, P. VALLURUPALLI, D. F. HANSEN, B. E. CORREIA, O. LANGE, A. BAH, R. M. VERNON, F. W. DAHLQUIST, D. BAKER, and L. E. KAY. « Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. ». *Nature*, 477(7362):111–114, 2011. 36

[BW97] P. BRION and E. WESTHOF. « Hierarchy and dynamics of RNA folding. ». *Annu Rev Biophys Biomol Struct*, 26:113–137, 1997. 26

[BWF+00] H. BERMAN, J. WESTBROOK, Z. FENG, G. GILLILAND, T. BHAT, H. WEISSIG, I. SHINDYALOV, and P. BOURNE. « The Protein Data Bank ». *Nucleic Acids Research*, 28:235–242, 2000. 2

[BYB98] J.-D. BOISSONNAT, M. YVINEC, and H. BRÖNNIMANN. *Algorithmic geometry*, volume 5. Cambridge University Press, 1998. 10

[CB06] J.-M. CHANDONIA and S. E. BRENNER. « The impact of structural genomics: expectations and outcomes. ». *Science*, 311(5759):347–351, 2006. 4

[CHW01] J. A. CROMSIGT, C. W. HILBERS, and S. S. WIJMENGA. « Prediction of proton chemical shifts in RNA. Their use in structure refinement and validation. ». *J Biomol NMR*, 21(1):11–29, 2001. . . . . 46, 48

[CKF+09] D. COZZETTO, A. KRYSHTAFOVYCH, K. FIDELIS, J. MOULT, B. ROST, and A. TRAMONTANO. « Evaluation of template-based models in CASP8 with standard measures. ». *Proteins*, 77 Suppl 9:18–28, 2009. 27

[CM11] A. CORNUÉJOLS and L. MICLET. *Apprentissage artificiel: concepts et algorithmes*. Editions Eyrolles, 2011. . . . . 14, 15

[CST00] N. CRISTIANI and J. SHAWE-TAYLOR. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. 18

[CWD09] T. A. COOPER, L. WAN, and G. DREYFUSS. « RNA and disease. ». *Cell*, 136(4):777–793, 2009. 36

[DB03] K. A. DILL and S. BROMBERG. *Molecular driving forces: statistical thermodynamics in chemistry and biology*. Garland Science, 2003. 5

[DB07] R. DAS and D. BAKER. « Automated de novo prediction of native-like RNA tertiary structures. ». *Proc Natl Acad Sci U S A*, 104(37):14664–14669, 2007. 26, 27, 36, 51, 58

[DBB03] C. DOMINGUEZ, R. BOELENS, and A. BONVIN. « HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. ». *J Am Chem Soc*, 125(7):1731–7, 2003. 20

[Der81] B. DERRIDA. « Random-energy model: An exactly solvable model of disordered systems ». *Physical Review B*, 24(5):2613, 1981. 21

[DGP09] C. DASKALAKIS, P. W. GOLDBERG, and C. H. PAPADIMITRIOU. « The Complexity of Computing a Nash Equilibrium ». *Commun. ACM*, 52(2):89–97, 2009. 56

[DHM+08] E. A. DETHOFF, A. L. HANSEN, C. MUSSELMAN, E. D. WATT, I. ANDRICIOAEI, and H. M. AL-HASHIMI. « Characterizing complex dynamics in the transactivation response element apical loop and motional correlations with the bulge by NMR, molecular dynamics, and mutagenesis. ». *Biophys J*, 95(8):3906–3915, 2008. 49

[DHT05] R. I. DIMA, C. HYEON, and D. THIRUMALAI. « Extracting stacking interaction parameters for RNA from the data set of native structures. ». *J Mol Biol*, 347(1):53–69, 2005. 26

[Dil85] K. A. DILL. « Theory for the folding and stability of globular proteins. ». *Biochemistry*, 24(6):1501–1509, 1985. 26

[DKB10] R. DAS, J. KARANICOLAS, and D. BAKER. « Atomic accuracy in predicting and designing noncanonical RNA structure. ». *Nat Methods*, 7(4):291–294, 2010. . . . . 26, 27, 58

[DSJ+05] F. DUPUIS, J. SADOC, R. JULLIEN, B. ANGELOV, and J. MORNON. « Voro3D: 3D Voronoi tessellations applied to protein structures. ». *Bioinformatics*, 21(8):1715–6, 2005. 10

[DT04] Y. DORSETT and T. TUSCHL. « siRNAs: applications in functional genomics and potential as therapeutics. ». *Nat Rev Drug Discov*, 3(4):318–329, 2004. 36

[EBH+14] P. S. EMANI, M. F. BARDARO, Jr, W. HUANG, S. ARAGON, G. VARANI, and G. P. DROBNY. « Elucidating molecular motion through structural and dynamic filters of energy-minimized conformer ensembles. ». *J Phys Chem B*, 118(7):1726–1742, 2014. 36

[EFL96] H. EDELSBRUNNER, M. FACELLO, and J. LIANG. « On the definition and the construction of pockets in macromolecules. ». *Pac Symp Biocomput*, pages 272–87, 1996. 10

[FA10] S. C. FLORES and R. B. ALTMAN. « Turning limited experimental information into 3D models of RNA. ». *RNA*, 16(9):1769–1778, 2010. 26

[FBH+10] S. FLORES, J. BERNAUER, X. HUANG, R. ZHOU, and S. SHIN. « MULTIRESOLUTION MODELING OF BIOLOGICAL MACROMOLECULES - Session Introduction. ». *Pacific Symposium on Biocomputing*, 15:201–204, 2010. 9

[FBS+12] S. C. FLORES, J. BERNAUER, S. SHIN, R. ZHOU, and X. HUANG. « Multiscale modeling of macromolecular biosystems. ». *Brief Bioinform*, 13(4):395–405, 2012. 8, 9, 79

[FF03] J. FÜRNKRANZ and P. A. FLACH. « An analysis of rule evaluation metrics ». In *International Conference on Machine Learning*. AAAI Press, 2003. 16

[FHAAH13] A. T. FRANK, S. HOROWITZ, I. ANDRICIOAEI, and H. M. AL-HASHIMI. « Utility of 1H NMR chemical shifts in determining RNA structure and dynamics. ». *J Phys Chem B*, 117(7):2045–2052, 2013. . . . . 36, 46, 48

[FK98] E. FURUICHI and P. KOEHL. « Influence of protein structure databases on the predictive power of statistical pair potentials. ». *Proteins*, 31(2):139–149, 1998. 34

[FMT+09] J. FRELLSEN, I. MOLTKE, M. THIIM, K. V. MARDIA, J. FERKINGHOFF-BORG, and T. HAMELRYCK. « A probabilistic model of RNA conformational space. ». *PLoS Comput Biol*, 5(6):e1000406, 2009. 26

[FPBv14] R. FONSECA, D. V. PACHOV, J. BERNAUER, and H. VAN DEN BEDEM. « Characterizing RNA ensembles from NMR data with kinematic models. ». *Nucleic Acids Res*, 42(15):9562–9572, 2014. . . . . 10, 35, 36, 55, 80, 148

[FSAHA09] A. T. FRANK, A. C. STELZER, H. M. AL-HASHIMI, and I. ANDRICIOAEI. « Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. ». *Nucleic Acids Res*, 37(11):3670–3679, 2009. . . . . 36, 49

[FWS+11] S. J. FLEISHMAN, T. A. WHITEHEAD, E.-M. STRAUCH, J. E. CORN, S. QIN, H.-X. ZHOU, J. C. MITCHELL, O. N. A. DEMERDASH, M. TAKEDA-SHITAKA, G. TERASHI, I. H. MOAL, X. LI, P. A. BATES, M. ZACHARIAS, H. PARK, J.-s. KO, H. LEE, C. SEOK, T. BOURQUARD, J. BERNAUER, A. POUPON, J. AZÉ, S. SONER, S. K. OVALI, P. OZBEK, N. B. TAL, T. HALILOGLU, H. HWANG, T. VREVEN, B. G. PIERCE, Z. WENG, L. PÉREZ-CANO, C. PONS, J. FERNÁNDEZ-RECIO, F. JIANG, F. YANG, X. GONG, L. CAO, X. XU, B. LIU, P. WANG, C. LI, C. WANG, C. H. ROBERT, M. GUHAROY, S. LIU, Y. HUANG, L. LI, D. GUO, Y. CHEN, Y. XIAO, N. LONDON, Z. ITZHAKI, O. SCHUELER-FURMAN, Y. INBAR, V. POTAPOV, M. COHEN, G. SCHREIBER, Y. TSUCHIYA, E. KANAMORI, D. M. STANDLEY, H. NAKAMURA, K. KINOSHITA, C. M. DRIGGERS, R. G. HALL, J. L. MORGAN, V. L. HSU, J. ZHAN, Y. YANG, Y. ZHOU, P. L. KASTRITIS, A. M. J. J. BONVIN, W. ZHANG, C. J. CAMACHO, K. P. KILAMBI, A. SIRCAR, J. J. GRAY, M. OHUE, N. UCHIKOGA, Y. MATSUZAKI, T. ISHIDA, Y. AKIYAMA, R. KHASHAN, S. BUSH, D. FOUCHES, A. TROPSHA, J. ESQUIVEL-RODRÍGUEZ, D. KIHARA, P. B. STRANGES, R. JACAK, B. KUHLMAN, S.-Y. HUANG, X. ZOU, S. J. WODAK, J. JANIN, and D. BAKER. « Community-wide assessment of protein-interface modeling suggests improvements to design methodology. ». *J Mol Biol*, 414(2):289–302, 2011. . . . . 8, 10, 22, 23

[GA06] R. GESTLAND and J. ATKINS. *The RNA World. The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor Laboratory Press, 2006. 26

[GBS+06] S. GRAZIANI, J. BERNAUER, S. SKOULOUBRIS, M. GRAILLE, C.-Z. ZHOU, C. MARCHAND, P. DECOTTIGNIES, H. VAN TILBEURGH, H. MYLLYKALLIO, and U. LIEBL. «

Catalytic mechanism and structure of viral flavin-dependent thymidylate synthase ThyX. ». *J Biol Chem*, 281(33):24048–24057, 2006. 54

[GC96] M. GERSTEIN and C. CHOTHIA. « Packing at the protein-water interface. ». *Proc Natl Acad Sci U S A*, 93(19):10167–72, 1996. 10

[GG14] A. GUILHOT-GAUDEFFROY. « *Modélisation et score de complexes protéine-ARN* ». Theses, Université Paris Sud, 2014. . . . . 13, 14, 24, 145

[GGFAB14] A. GUILHOT-GAUDEFFROY, C. FROIDEVAUX, J. AZÉ, and J. BERNAUER. « Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities. ». *PLoS One*, 9(9):e108928, 2014. . . . . 8, 10, 19, 20, 24, 58, 79, 147

[GK95] M. D. GRIGORIADIS and L. G. KHACHIYAN. « A sublinear-time randomized approximation algorithm for matrix games ». *Operations Research Letters*, 18(2):53–58, 1995. 56

[GMW+03] J. J. GRAY, S. MOUGHON, C. WANG, O. SCHUELER-FURMAN, B. KUHLMAN, C. A. ROHL, and D. BAKER. « Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. ». *J Mol Biol*, 331(1):281–299, 2003. 20

[Guo10] P. GUO. « The emerging field of RNA nanotechnology. ». *Nat Nanotechnol*, 5(12):833–842, 2010. . . . . 2, 26, 149

[HCSD05] B. K. HO, E. A. COUTSIAS, C. SEOK, and K. A. DILL. « The flexibility in the proline ring couples to the protein backbone. ». *Protein Sci*, 14(4):1011–1018, 2005. 38

[HFB+08] K. HENRICK, Z. FENG, W. F. BLUHM, D. DIMITROPOULOS, J. F. DORELEIJERS, S. DUTTA, J. L. FLIPPEN-ANDERSON, J. IONIDES, C. KAMADA, E. KRISSINEL, C. L. LAWSON, J. L. MARKLEY, H. NAKAMURA, R. NEWMAN, Y. SHIMIZU, J. SWAMINATHAN, S. VELANKAR, J. ORY, E. L. ULRICH, W. VRANKEN, J. WESTBROOK, R. YAMASHITA, H. YANG, J. YOUNG, M. YOUSUFUDDIN, and H. M. BERMAN. « Remediation of the protein data bank archive. ». *Nucleic Acids Res*, 36(Database issue):D426–D433, 2008. 53

[HGC94] Y. HARPAZ, M. GERSTEIN, and C. CHOTHIA. « Volume changes on protein folding. ». *Structure*, 2(7):641–9, 1994. 10

[HGWB04] H. HUTHOFF, F. GIRARD, S. S. WIJMENGA, and B. BERKHOUT. « Evidence for a base triple in the free HIV-1 TAR RNA. ». *RNA*, 10(3):412–423, 2004. 49

[HO94] D. HALPERIN and M. H. OVERMARS. « Spheres, molecules, and hidden surface removal ». In *Proceedings of the tenth annual symposium on Computational geometry*, pages 113–122. ACM, 1994. 42

[HTF09] T. HASTIE, R. TIBSHIRANI, and J. FRIEDMAN. *The elements of statistical learning*. Springer, second edition, 2009. . . . . 14, 15, 17, 19

[Jan96] J. JANIN. « Quantifying biological specificity: the statistical mechanics of molecular recognition. ». *Proteins*, 25(4):438–445, 1996. 21

[Jan10] J. JANIN. « Protein-protein docking tested in blind predictions: the CAPRI experiment. ». *Mol Biosyst*, 6(12):2351–2362, 2010. 5

[JHM+03] J. JANIN, K. HENRICK, J. MOULT, L. T. EYCK, M. J. E. STERNBERG, S. VAJDA, I. VAKSER, S. J. WODAK, and C. A. o. P. R. I. . « CAPRI: a Critical Assessment of PRedicted Interactions. ». *Proteins*, 52(1):2–9, 2003. . . . . 5, 13

[Joh06] S. JOHNSON. *The Ghost Map: The Story of London's Most Terrifying Epidemic–and How It Changed Science, Cities, and the Modern World*. Riverhead, 2006. 10

[JRKT01] D. J. JACOBS, A. J. RADER, L. A. KUHN, and M. F. THORPE. « Protein flexibility predictions using graph theory. ». *Proteins*, 44(2):150–165, 2001. 36

[JRL+09] M. A. JONIKAS, R. J. RADMER, A. LAEDERACH, R. DAS, S. PEARLMAN, D. HERSCHLAG, and R. B. ALTMAN. « Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. ». *RNA*, 15(2):189–199, 2009. 26

[JSL+ed] H. JIANG, F. K. SHEONG, Z. LIZHE, X. GAO, J. BERNAUER, and X. HUANG. « Markov State Models reveal a two-step mechanism of miRNA loading into the Human Argonaute protein: conformational selection followed by structural re-arrangement. ». submitted. 24, 54

[KAC+14] H. KIM, S. C. ABEYSIRIGUNAWARDEN, K. CHEN, M. MAYERLE, K. RAGUNATHAN, Z. LUTHEY-SCHULTEN, T. HA, and S. A. WOODSON. « Protein-guided RNA dynamics during early ribosome assembly. ». *Nature*, 506(7488):334–338, 2014. 36

[KCVC05] D. KOZAKOV, K. H. CLODFELTER, S. VAJDA, and C. J. CAMACHO. « Optimal clustering for detecting near-native conformations in protein docking. ». *Biophys J*, 89(2):867–875, 2005. 58

[KDS+60] J. KENDREW, R. DICKERSON, B. STANDBERG, R. HART, D. DAVIES, and D. PHILLIPS. « Structure of myoglobin. A three dimensional Fourier synthesis at 2Å resolution. ». *Nature*, 185:422–427, 1960. 2

[KGLK05] R. KOLODNY, L. GUIBAS, M. LEVITT, and P. KOEHL. « Inverse kinematics in biology: the protein loop closure problem ». *The International Journal of Robotics Research*, 24(2-3):151–163, 2005. 36

[Kir36] J. G. KIRKWOOD. « Statistical Mechanics of Liquid Solutions. ». *Chemical Reviews*, 19(3):275–307, 1936. 30

[KK12] J.-K. KIM and D.-S. KIM. « BetaSuperposer: superposition of protein surfaces using beta-shapes. ». *J Biomol Struct Dyn*, 30(6):684–700, 2012. 10

[Koe10] P. KOEHL. Protein Structure Prediction. In *Biomolecular applications of Biophysics*, pages 1–34. Humana press, New York, NY, 2010. 4

[KOH+03] T. KULINSKI, M. OLEJNICZAK, H. HUTHOFF, L. BIELECKI, K. PACHULSKA-WIECZOREK, A. T. DAS, B. BERKHOUT, and R. W. ADAMIAK. « The apical loop of the HIV-1 TAR RNA hairpin is stabilized by a cross-loop base pair. ». *J Biol Chem*, 278(40):38892–38901, 2003. 49

[KPX+13] K. P. KILAMBI, M. S. PACELLA, J. XU, J. W. LABONTE, J. R. PORTER, P. MUTHU, K. DREW, D. KURODA, O. SCHUELER-FURMAN, R. BONNEAU, and J. J. GRAY. « Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27. ». *Proteins*, 81(12):2201–2209, 2013. 22

[KZT12] R. KHASHAN, W. ZHENG, and A. TROPSHA. « Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. ». *Proteins*, 80(9):2207–2217, 2012. 10

[LBGC11] J. R. LOPÉZ-BLANCO, J. I. GARZÓN, and P. CHACÓN. « iMod: multipurpose normal mode analysis in internal coordinates. ». *Bioinformatics*, 27(20):2843–2850, 2011. . . 42

[LC10] S. LORIOT and F. CAZALS. « Modeling macro-molecular interfaces with Intervor. ». *Bioinformatics*, 26(7):964–965, 2010. 10

[LCB10] S. LORIOT, F. CAZALS, and J. BERNAUER. « ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules. ». *Bioinformatics*, 26(8):1127–1128, 2010. . . . . 8, 10, 19, 146

[LCC+13] S. LYSKOV, F.-C. CHOU, S. Ó. CONCHÚIR, B. S. DER, K. DREW, D. KURODA, J. XU, B. D. WEITZNER, P. D. RENFREW, P. SRIPAKDEEVONG, B. BORGO, J. J. HAVRANEK, B. KUHLMAN, T. KORTEMME, R. BONNEAU, J. J. GRAY, and R. DAS. « Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). ». *PLoS One*, 8(5):e63906, 2013. 27

[Les13] A. LESK. *Introduction to bioinformatics*. Oxford University Press, 2013. 4

[Lev76] M. LEVITT. « A simplified representation of protein conformations for rapid simulation of protein folding. ». *J Mol Biol*, 104(1):59–107, 1976. 9

[LHZ03a] C. X. LING, J. HUANG, and H. ZHANG. AUC: a better measure than accuracy in comparing learning algorithms. In *Advances in Artificial Intelligence*, pages 329–341. Springer, 2003. 16

[LHZ03b] C. X. LING, J. HUANG, and H. ZHANG. « AUC: a statistically consistent and more discriminating measure than accuracy ». In *IJCAI*, volume 3, pages 519–524, 2003. . 16

[Lin08] E. R. LINDAHL. « Molecular dynamics simulations. ». *Methods Mol Biol*, 443:3–23, 2008. . . . . 4, 5

[Lin15] E. LINDAHL. « Molecular dynamics simulations. ». *Methods Mol Biol*, 1215:3–26, 2015. 4, 5

[LKL+14] J. LEE, W. KLADWANG, M. LEE, D. CANTU, M. AZIZYAN, H. KIM, A. LIMPAECHER, S. YOON, A. TREUILLE, R. DAS, and E. N. A. P. . « RNA design rules from a massive open laboratory. ». *Proc Natl Acad Sci U S A*, 111(6):2122–2127, 2014. 26

[LLW06] N. B. LEONTIS, A. LESCOUTE, and E. WESTHOF. « The building blocks and motifs of RNA architecture. ». *Curr Opin Struct Biol*, 16(3):279–287, 2006. 38

[LLXZ13] H. Lu, Z. Li, Y. Xue, and Q. Zhou. « Viral-host interactions that control HIV-1 transcriptional elongation. ». *Chem Rev*, 113(11):8567–8582, 2013. 49

[LMB+14] M. F. Lensink, I. H. Moal, P. A. Bates, P. L. Kastritis, A. S. J. Melquiond, E. Karaca, C. Schmitz, M. van Dijk, A. M. J. J. Bonvin, M. Eisenstein, B. Jiménez-García, S. Grosdidier, A. Solernou, L. Pérez-Cano, C. Pallara, J. Fernández-Recio, J. Xu, P. Muthu, K. Praneeth Kilambi, J. J. Gray, S. Grudinin, G. Derevyanko, J. C. Mitchell, J. Wieting, E. Kanamori, Y. Tsuchiya, Y. Murakami, J. Sarmiento, D. M. Standley, M. Shirota, K. Kinoshita, H. Nakamura, M. Chavent, D. W. Ritchie, H. Park, J. Ko, H. Lee, C. Seok, Y. Shen, D. Kozakov, S. Vajda, P. J. Kundrotas, I. A. Vakser, B. G. Pierce, H. Hwang, T. Vreven, Z. Weng, I. Buch, E. Farkash, H. J. Wolfson, M. Zacharias, S. Qin, H.-X. Zhou, S.-Y. Huang, X. Zou, J. A. Wojdyla, C. Kleanthous, and S. J. Wodak. « Blind prediction of interfacial water positions in CAPRI. ». *Proteins*, 82(4):620–632, 2014. 22

[LMK13] J. Li, P. Mach, and P. Koehl. « Measuring the shapes of macromolecules - and why it matters. ». *Comput Struct Biotechnol J*, 8:e201309001, 2013. 10

[LMW07] M. F. Lensink, R. Méndez, and S. J. Wodak. « Docking and scoring protein complexes: CAPRI 3rd Edition. ». *Proteins*, 69(4):704–718, 2007. 22

[LQV+13] A. Lamiable, F. Quessette, S. Vial, D. Barth, and A. Denise. « An algorithmic game-theory approach for coarse-grain prediction of RNA 3D structure. ». *IEEE/ACM Trans Comput Biol Bioinform*, 10(1):193–199, 2013. 55

[LRSC01] C. E. Leiserson, R. L. Rivest, C. Stein, and T. H. Cormen. *Introduction to algorithms*. MIT Press, 2001. 38

[LS82] G. Loomes and R. Sugden. « Regret theory: An alternative theory of rational choice under uncertainty ». *Economic journal*, 92(368):805–824, 1982. 56

[LS01] H. Lu and J. Skolnick. « A distance-dependent atomic knowledge-based potential for improved protein structure selection. ». *Proteins*, 44(3):223–232, 2001. . . . . 29, 32

[LSMD+13] A. Lopes, S. Sacquin-Mora, V. Dimitrova, E. Laine, Y. Ponty, and A. Carbone. « Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. ». *PLoS Comput Biol*, 9(12):e1003369, 2013. 58

[LV01] N. Leulliot and G. Varani. « Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. ». *Biochemistry*, 40(27):7947–7956, 2001. 36

[LW75] M. Levitt and A. Warshel. « Computer simulation of protein folding. ». *Nature*, 253(5494):694–698, 1975. 9

[LW78] M. Levitt and A. Warshel. « Extreme conformational flexibility of the furanose ring in DNA and RNA ». *Journal of the American Chemical Society*, 100(9):2607–2613, 1978. 40

[LW10] M. F. Lensink and S. J. Wodak. « Docking and scoring protein interactions: CAPRI 2009. ». *Proteins*, 78(15):3073–3084, 2010. . . . . 5, 13

[LW14] M. F. LENSINK and S. J. WODAK. « Score_set: A CAPRI benchmark for scoring protein complexes. ». *Proteins*, 82(11):3163–3169, 2014. . . . . 23, 53

[LWT+11] B. A. LEWIS, R. R. WALIA, M. TERRIBILINI, J. FERGUSON, C. ZHENG, V. HONAVAR, and D. DOBBS. « PRIDB: a Protein-RNA interface database. ». *Nucleic Acids Res*, 39(Database issue):D277–D282, 2011. 19

[MARR03] L. J. W. MURRAY, W. B. ARENDALL, 3rd, D. C. RICHARDSON, and J. S. RICHARDSON. « RNA backbone is rotameric. ». *Proc Natl Acad Sci U S A*, 100(24):13904–13909, 2003. 26

[Mat06] D. H. MATHEWS. « Revolutions in RNA secondary structure prediction. ». *J Mol Biol*, 359(3):526–532, 2006. 26

[Met78] C. E. METZ. « Basic principles of ROC analysis ». In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978. 16

[MFA+13] R. MORETTI, S. J. FLEISHMAN, R. AGIUS, M. TORCHALA, P. A. BATES, P. L. KASTRITIS, J. P. G. L. M. RODRIGUES, M. TRELLET, A. M. J. J. BONVIN, M. CUI, M. ROOMAN, D. GILLIS, Y. DEHOUCK, I. MOAL, M. ROMERO-DURANA, L. PEREZ-CANO, C. PALLARA, B. JIMENEZ, J. FERNANDEZ-RECIO, S. FLORES, M. PACELLA, K. PRANEETH KILAMBI, J. J. GRAY, P. POPOV, S. GRUDININ, J. ESQUIVEL-RODRÍGUEZ, D. KIHARA, N. ZHAO, D. KORKIN, X. ZHU, O. N. A. DEMERDASH, J. C. MITCHELL, E. KANAMORI, Y. TSUCHIYA, H. NAKAMURA, H. LEE, H. PARK, C. SEOK, J. SARMIENTO, S. LIANG, S. TERAGUCHI, D. M. STANDLEY, H. SHIMOYAMA, G. TERASHI, M. TAKEDA-SHITAKA, M. IWADATE, H. UMEYAMA, D. BEGLOV, D. R. HALL, D. KOZAKOV, S. VAJDA, B. G. PIERCE, H. HWANG, T. VREVEN, Z. WENG, Y. HUANG, H. LI, X. YANG, X. JI, S. LIU, Y. XIAO, M. ZACHARIAS, S. QIN, H.-X. ZHOU, S.-Y. HUANG, X. ZOU, S. VELANKAR, J. JANIN, S. J. WODAK, and D. BAKER. « Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. ». *Proteins*, 81(11):1980–1987, 2013. 22

[MGHT99] M. MUCCHIELLI-GIORGI, S. HAZOUT, and P. TUFFERY. « PredAcc: prediction of solvent accessibility. ». *Bioinformatics*, 15(2):176–7, 1999. 17

[MGHT02] M. MUCCHIELLI-GIORGI, S. HAZOUT, and P. TUFFERY. « Predicting the disulfide bonding state of cysteines using protein descriptors. ». *Proteins*, 46(3):243–9, 2002. . 17

[MH88] J. A. MCCAMMON and S. C. HARVEY. *Dynamics of proteins and nucleic acids*. Cambridge University Press, 1988. 31

[MK11] P. MACH and P. KOEHL. « Geometric measures of large biomolecules: surface, volume, and pockets. ». *J Comput Chem*, 32(14):3023–3038, 2011. 10

[MK13] P. MACH and P. KOEHL. « An analytical method for computing atomic contact areas in biomolecules. ». *J Comput Chem*, 34(2):105–120, 2013. 10

[Mou97] J. MOULT. « Comparison of database potentials and molecular mechanics force fields. ». *Curr Opin Struct Biol*, 7(2):194–199, 1997. 29

[Nas50] J. F. NASH. « Equilibrium points in n-person games ». *Proc Natl Acad Sci U S A*, 36(1):48–49, 1950. 56

[NHB+00] P. NISSEN, J. HANSEN, N. BAN, P. B. MOORE, and T. A. STEITZ. « The structural basis of ribosome activity in peptide bond synthesis. ». *Science*, 289(5481):920–930, 2000. 36

[PCJGFR12] L. PÉREZ-CANO, B. JIMÉNEZ-GARCÍA, and J. FERNÁNDEZ-RECIO. « A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. ». *Proteins*, 80(7):1872–1882, 2012. 58

[PJGPC+13] C. PALLARA, B. JIMÉNEZ-GARCÍA, L. PÉREZ-CANO, M. ROMERO-DURANA, A. SOLERNOU, S. GROSDIDIER, C. PONS, I. H. MOAL, and J. FERNANDEZ-RECIO. « Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges. ». *Proteins*, 81(12):2192–2200, 2013. 22

[PM08] M. PARISIEN and F. MAJOR. « The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. ». *Nature*, 452(7183):51–55, 2008. 26, 58

[Pou04] A. POUPON. « Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. ». *Curr Opin Struct Biol*, 14(2):233–41, 2004. 10, 145

[POU+05] J. PRILUSKY, E. OUEILLET, N. ULRYCK, A. PAJON, J. BERNAUER, I. KRIMM, S. QUEVILLON-CHERUEL, N. LEULLIOT, M. GRAILLE, D. LIGER, L. TRÉSAUGUES, J. L. SUSSMAN, J. JANIN, H. VAN TILBEURGH, and A. POUPON. « HalX: an open-source LIMS (Laboratory Information Management System) for small- to large-scale laboratories. ». *Acta Crystallogr D Biol Crystallogr*, 61(Pt 6):671–678, 2005. 4, 54

[PPS+13] S. PRONK, S. PÁLL, R. SCHULZ, P. LARSSON, P. BJELKMAR, R. APOSTOLOV, M. R. SHIRTS, J. C. SMITH, P. M. KASSON, D. VAN DER SPOEL, B. HESS, and E. LINDAHL. « GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. ». *Bioinformatics*, 29(7):845–854, 2013. 31

[PRW96] J. PONTIUS, J. RICHELLE, and S. WODAK. « Deviations from standard atomic volumes as a quality measure for protein crystal structures. ». *J Mol Biol*, 264(1):121–36, 1996. 10

[Rak04] A. RAKOTOMAMONJY. « Optimizing area under ROC curves with SVMs ». In *ROCAI'04*, 2004. 16

[RBF+11] J. RINNENTHAL, J. BUCK, J. FERNER, A. WACKER, B. FÜRTIG, and H. SCHWALBE. « Mapping the landscape of RNA dynamics with NMR spectroscopy. ». *Acc Chem Res*, 44(12):1292–1301, 2011. 36

[RHR+06] J. REEDER, M. HÖCHSMANN, M. REHMSMEIER, B. VOSS, and R. GIEGERICH. « Beyond Mfold: recent advances in RNA bioinformatics. ». *J Biotechnol*, 124(1):41–55, 2006. 26

[Rit03] D. W. RITCHIE. « Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. ». *Proteins*, 52(1):98–106, 2003. 20

[Rit08] D. W. RITCHIE. « Recent progress and future directions in protein-protein docking. ». *Curr Protein Pept Sci*, 9(1):1–15, 2008. 2, 5, 13, 58

[Rob85] H. ROBBINS. Some Aspects of the Sequential Design of Experiments. In T. LAI and D. SIEGMUND, editors, *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985. 56

[RRPB11] M. ROTHER, K. ROTHER, T. PUTON, and J. M. BUJNICKI. « ModeRNA: a tool for comparative modeling of RNA 3D structure. ». *Nucleic Acids Res*, 39(10):4007–4022, 2011. . . . . 36, 58

[RRS11] F. RICCI, L. ROKACH, and B. SHAPIRA. *Introduction to recommender systems handbook*. Springer, 2011. 19

[RSM+08] J. S. RICHARDSON, B. SCHNEIDER, L. W. MURRAY, G. J. KAPRAL, R. M. IMMORMINO, J. J. HEADD, D. C. RICHARDSON, D. HAM, E. HERSHKOVITS, L. D. WILLIAMS, K. S. KEATING, A. M. PYLE, D. MICALLEF, J. WESTBROOK, H. M. BERMAN, and R. N. A. O. C. . « RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). ». *RNA*, 14(3):465–481, 2008. . . . 44

[RSMS+14] E. RANAEI-SIADAT, C. MÉRIGOUX, B. SEIJO, L. PONCHON, J.-M. SALIOU, J. BERNAUER, S. SANGLIER-CIANFÉRANI, F. DARDEL, P. VACHETTE, and S. NONIN-LECOMTE. « In vivo tmRNA protection by SmpB and pre-ribosome binding conformation in solution. ». *RNA*, 20(10):1607–1620, 2014. 54

[SAL03] M. SEBAG, J. AZÉ, and N. LUCAS. « Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning ». In *Proceedings of the Third IEEE International Conference on Data Mining, ICDM*, 2003. 18

[Sat10] A. SATOH. *Introduction to practice of molecular simulation: molecular dynamics, Monte Carlo, Brownian dynamics, Lattice Boltzmann and dissipative particle dynamics*. Elsevier, 2010. 5

[SB10] W. SHEFFLER and D. BAKER. « RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. ». *Protein Sci*, 19(10):1991–1995, 2010. 10

[SBAAH13] L. SALMON, G. BASCOM, I. ANDRICIOAEI, and H. M. AL-HASHIMI. « A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed. ». *J Am Chem Soc*, 135(14):5457–5466, 2013. . . . . 36, 49

[SBHH14] X. SHI, K. A. BEAUCHAMP, P. B. HARBURY, and D. HERSCHLAG. « From a structural average to the conformational ensemble of a DNA bulge. ». *Proc Natl Acad Sci U S A*, 111(15):E1473–E1480, 2014. 36

[SCC+14] P. SRIPAKDEEVONG, M. CEVEC, A. T. CHANG, M. C. ERAT, M. ZIEGELER, Q. ZHAO, G. E. FOX, X. GAO, S. D. KENNEDY, R. KIERZEK, E. P. NIKONOWICZ, H. SCHWALBE, R. K. O. SIGEL, D. H. TURNER, and R. DAS. « Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts. ». *Nat Methods*, 11(4):413–416, 2014. 36, 46

[Sch94] D. A. SCHUM. *The evidential foundations of probabilistic reasoning*. Northwestern University Press, 1994. 30

[Sch97] B. SCHÖLKOPF. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997. . . . 18

[SCM+00] A. SOYER, J. CHOMILIER, J. MORNON, R. JULLIEN, and J. SADOC. « Voronoi tessellation reveals the condensed matter character of folded proteins. ». *Phys Rev Lett*, 85(16):3532–5, 2000. 10

[Sip90] M. J. SIPPL. « Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. ». *J Mol Biol*, 213(4):859–883, 1990. . . . . 26, 33

[SK09] X. SU and T. M. KHOSHGOFTAAR. « A survey of collaborative filtering techniques ». *Advances in artificial intelligence*, 2009:4, 2009. 19

[SL05] M. T. SYKES and M. LEVITT. « Describing RNA structure by libraries of clustered nucleotide doublets. ». *J Mol Biol*, 351(1):26–38, 2005. 26

[SL07] C. M. SUMMA and M. LEVITT. « Near-native structure refinement using in vacuo energy minimization. ». *Proc Natl Acad Sci U S A*, 104(9):3177–3182, 2007. . . . . 27, 31, 32

[SLA03] M. SEBAG, N. LUCAS, and J. AZÉ. « ROC-based Evolutionary Learning: Application to Medical Data Mining ». In *Proceedings of the 6th International Conference on Artificial Evolution, EA 2003*, 2003. 18

[SLD+08] Y. SHEN, O. LANGE, F. DELAGLIO, P. ROSSI, J. M. ARAMINI, G. LIU, A. ELETSKY, Y. WU, K. K. SINGARAPU, A. LEMAK, A. IGNATCHENKO, C. H. ARROWSMITH, T. SZYPERSKI, G. T. MONTELIONE, D. BAKER, and A. BAX. « Consistent blind protein structure generation from NMR chemical shift data. ». *Proc Natl Acad Sci U S A*, 105(12):4685–4690, 2008. . . . . 36, 46

[SLM12] A. Y. L. SIM, M. LEVITT, and P. MINARY. « Modeling and design by hierarchical natural moves. ». *Proc Natl Acad Sci U S A*, 109(8):2890–2895, 2012. . . . . 26, 36

[SM98] R. SAMUDRALA and J. MOULT. « An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. ». *J Mol Biol*, 275(5):895–916, 1998. 29

[SML12] A. Y. L. SIM, P. MINARY, and M. LEVITT. « Modeling nucleic acids. ». *Curr Opin Struct Biol*, 22(3):273–278, 2012. 26

[SSLB12] A. Y. L. SIM, O. SCHWANDER, M. LEVITT, and J. BERNAUER. « Evaluating mixture models for building RNA knowledge-based potentials. ». *J Bioinform Comput Biol*, 10(2):1241010, 2012. . . . . 10, 25, 31, 32, 33, 58, 147

[SYKB07] B. A. SHAPIRO, Y. G. YINGLING, W. KASPRZAK, and E. BINDEWALD. « Bridging the gap in RNA structure prediction. ». *Curr Opin Struct Biol*, 17(2):157–165, 2007. 26, 36

[TB99] I. TINOCO, Jr and C. BUSTAMANTE. « How RNA folds. ». *J Mol Biol*, 293(2):271–281, 1999. 26

[TB05] B. J. TUCKER and R. R. BREAKER. « Riboswitches as versatile gene control elements. ». *Curr Opin Struct Biol*, 15(3):342–348, 2005. 36

[TBM+03] J. TSAI, R. BONNEAU, A. V. MOROZOV, B. KUHLMAN, C. A. ROHL, and D. BAKER. « An improved protein decoy set for testing energy functions for protein structure prediction. ». *Proteins*, 53(1):76–87, 2003. 28

[TBV+10] T. H. TAHIROV, N. D. BABAYEVA, K. VARZAVAND, J. J. COOPER, S. C. SEDORE, and D. H. PRICE. « Crystal structure of HIV-1 Tat complexed with human P-TEFb. ». *Nature*, 465(7299):747–751, 2010. 49

[Toz05] V. TOZZINI. « Coarse-grained models for proteins. ». *Curr Opin Struct Biol*, 15(2):144–150, 2005. 9

[TS76] S. TANAKA and H. A. SCHERAGA. « Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. ». *Macromolecules*, 9(6):945–950, 1976. 26

[UAD+08] E. L. ULRICH, H. AKUTSU, J. F. DORELEIJERS, Y. HARANO, Y. E. IOANNIDIS, J. LIN, M. LIVNY, S. MADING, D. MAZIUK, Z. MILLER, and OTHERS. « BioMagResBank ». *Nucleic acids research*, 36(suppl 1):D402–D408, 2008. 42

[vBY+13] H. VAN DEN BEDEM, G. BHABHA, K. YANG, P. E. WRIGHT, and J. S. FRASER. « Automated identification of functional dynamic contact networks from X-ray crystallography. ». *Nat Methods*, 10(9):896–902, 2013. 49

[VC06] M. VUK and T. CURK. « ROC curve, lift chart and calibration plot ». *Metodolo ski zvezki*, 3(1):89–108, 2006. 16

[vDLD09] H. VAN DEN BEDEM, A. DHANIK, J. C. LATOMBE, and A. M. DEACON. « Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. ». *Acta Crystallogr D Biol Crystallogr*, 65(Pt 10):1107–1117, 2009. 49

[VHK13] S. VAJDA, D. R. HALL, and D. KOZAKOV. « Sampling and scoring: a marriage made in heaven. ». *Proteins*, 81(11):1874–1884, 2013. 51

[VK09] S. VAJDA and D. KOZAKOV. « Convergence and combination of methods in protein-protein docking. ». *Curr Opin Struct Biol*, 19(2):164–170, 2009. 5

[vLLD05] H. VAN DEN BEDEM, I. LOTAN, J. C. LATOMBE, and A. M. DEACON. « Real-space protein-model completion: an inverse-kinematics approach. ». *Acta Crystallogr D Biol Crystallogr*, 61(Pt 1):2–13, 2005. 36

[Vor08] G. VORONOÏ. « Nouvelles applications des paramètres continus à la théorie des formes quadratiques. ». *Journal für die Reine und Angewandte Mathematik*, 133:97–178, 1908. 10

[VR12] V. VENKATRAMAN and D. W. RITCHIE. « Flexible protein docking refinement using pose-dependent normal mode analysis. ». *Proteins*, 80(9):2262–2274, 2012. 58

[WIN+05] J. WESTBROOK, N. ITO, H. NAKAMURA, K. HENRICK, and H. M. BERMAN. « PDBML: the representation of archival macromolecular structure data in XML. ». *Bioinformatics*, 21(7):988–992, 2005. 53

[WJ78] S. WODAK and J. JANIN. « Computer analysis of protein-protein interaction. ». *J Mol Biol*, 124(2):323–42, 1978. 20

[YDM+08] P. YAO, A. DHANIK, N. MARZ, R. PROPPER, C. KOU, G. LIU, H. VAN DEN BEDEM, J.-C. LATOMBE, I. HALPERIN-LANDSBERG, and R. B. ALTMAN. « Efficient algorithms to explore conformation spaces of flexible protein loops. ». *IEEE/ACM Trans Comput Biol Bioinform*, 5(4):534–545, 2008. 36

[YJL+03] H. YANG, F. JOSSINET, N. LEONTIS, L. CHEN, J. WESTBROOK, H. BERMAN, and E. WESTHOF. « Tools for the automatic identification and classification of RNA base pairs. ». *Nucleic Acids Res*, 31(13):3450–3460, 2003. 40

[YZL12] P. YAO, L. ZHANG, and J.-C. LATOMBE. « Sampling-based exploration of folded state of a protein under kinematic and geometric constraints. ». *Proteins*, 80(1):25–43, 2012. 36, 39

[ZGK06] A. ZOMORODIAN, L. GUIBAS, and P. KOEHL. « Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials ». *Computer Aided Geometric Design*, 23(6):531–544, 2006. 32

[Zha08] Y. ZHANG. « Progress and challenges in protein structure prediction. ». *Curr Opin Struct Biol*, 18(3):342–348, 2008. . . . . 2, 4, 26

[ZS00] M. ZACHARIAS and H. SKLENAR. « Conformational deformability of RNA: a harmonic mode analysis. ». *Biophys J*, 78(5):2528–2542, 2000. 36

[ZSG+11] J. ZHOU, Y. SHU, P. GUO, D. D. SMITH, and J. J. ROSSI. « Dual functional RNA nanoparticles containing phi29 motor pRNA and anti-gp120 aptamer for cell-type specific delivery and HIV-1 inhibition. ». *Methods*, 54(2):284–294, 2011. 36

[Zuk03] M. ZUKER. « Mfold web server for nucleic acid folding and hybridization prediction. ». *Nucleic Acids Res*, 31(13):3406–3415, 2003. 26

[Zwe08] M. ZWECKSTETTER. « NMR: prediction of molecular alignment from structure using the PALES software. ». *Nat Protoc*, 3(4):679–690, 2008. 49

[ZZ02] H. ZHOU and Y. ZHOU. « Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. ». *Protein Sci*, 11(11):2714–2726, 2002. 32

APPENDIX I

SELECTED PUBLICATIONS

SELECTION LIST

Review/Survey: Modeling of biomolecules

– **Multiscale modeling** [FBS+12] Page 93
Samuel C. FLORES, Julie BERNAUER, Seokmin SHIN, Ruhong ZHOU and Xuhui HUANG. « Multiscale modeling of macromolecular biosystems » In *Briefings in Bioinformatics*, 13(4):395-405, 2012. PMID:22228511

Docking

– **Protein-RNA docking** [GGFAB14] Page 104
Adrien GUILHOT-GAUDEFFROY, Christine FROIDEVAUX, Jérôme AZÉ, Julie BERNAUER. « Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities» In *PLoS One*, 9(9):e108928, 2014. PMID:25268579

– **Collaborative filtering for protein-protein docking** [BBAP11] Page 117
Thomas BOURQUARD, Julie BERNAUER, Jérôme AZÉ, Anne POUPON. « A collaborative filtering approach for protein-protein docking scoring functions» In *PLoS One*, 6(4):e18541, 2011. PMID:21526112

Knowledge-based potentials for RNA

– **Coarse-grained and all-atom KB potentials for RNA** [BHSL11] Page 128
Julie BERNAUER, Xuhui HUANG, Adelene Y.L. SIM, Michael LEVITT. « Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation» In *RNA*, 17(6):1066-75, 2011. PMID:21521828

Robotics-inspired modeling

– **RNA ensembles with kinematic models** [FPBv14] Page 141
Rasmus FONSECA, Dimitar V. PACHOV, Julie BERNAUER, Henry VAN DEN BEDEM. « Characterizing RNA ensembles from NMR data with kinematic models» In *Nucleic Acids Research*, 42(15):9562-72, 2014. PMID:25114056

MULTISCALE MODELING OF MACROMOLECULAR BIOSYSTEMS

Samuel C. FLORES, Julie BERNAUER J, Seokmin SHIN, Ruhong ZHOU and Xuhui HUANG. In *Briefings in Bioinformatics*, 13(4):395-405, 2012.

BRIEFINGS IN BIOINFORMATICS, VOL 13. NO 4. 395--405 doi:10.1093/bib/bbr077
Advance Access published on 6 January 2012

Multiscale modeling of macromolecular biosystems

Samuel C. Flores, Julie Bernauer*, Seokmin Shin, Ruhong Zhou and Xuhui Huang**

Submitted: 1st September 2011; Received (in revised form): 12th December 2011

Abstract

In this article, we review the recent progress in multiresolution modeling of structure and dynamics of protein, RNA and their complexes. Many approaches using both physics-based and knowledge-based potentials have been developed at multiple granularities to model both protein and RNA. Coarse graining can be achieved not only in the length, but also in the time domain using discrete time and discrete state kinetic network models. Models with different resolutions can be combined either in a sequential or parallel fashion. Similarly, the modeling of assemblies is also often achieved using multiple granularities. The progress shows that a multiresolution approach has considerable potential to continue extending the length and time scales of macromolecular modeling.

Keywords: *multiscale modeling; protein structure and dynamics; nucleic acid modeling; protein assemblies*

INTRODUCTION

Modeling the folding and assembly of biological macromolecules is challenging, since it is at large size and long time scales that require great computational cost. In response, many methods have been developed to reduce the computational cost by coarsening the granularity of the problem. Inevitably, the accuracy is lower for these methods and thus, in recent years, a consensus has emerged that one should work at multiple levels of resolution, simultaneously or in parallel [1] in a multiresolution approach.

At the coarsest granularity, an entire macromolecule can be treated as a single particle, whereas at the finest level each atom can be treated as a separate entity; indeed even higher levels of theory are possible [2]. There are many grainings in between, for example the kinematic or force unit may be a domain, a secondary structure element, a residue or chemical groups such as bases and ribose rings [3]. A simulation can be carried out at any level of resolution, or different levels can be used for the same problem either in serial or in parallel [1]. Furthermore, the force field and the kinematics can be treated at different levels of granularity [4], for example, secondary structure elements can be rigidified while interelement interactions can be treated at the all-atom level of resolution. The methods may consist of force fields [5, 6], sampling, structure prediction [5, 6] and dynamics tools [4], as well as novel algorithms.

Given the limited space, this mini-review is not aimed to be a complete review of this emerging field,

Corresponding author. Samuel C. Flores. Biomedical Center, Box 596, SE-751 24 Uppsala, Sweden. Tel: +46 0184714536; Fax: +46 18471530396; E-mail: samuelflores@gmail.com; Julie Bernauer. AMIB INRIA - Bioinformatique Laboratoire d'informatique (LIX), École Polytechnique 91128 Palaiseau Cedex, France. Tel: +33 (0)1 6933 4095; Fax: +33 (0)1 6933 4049; E-mail: julie.bernauer@inria.fr; Xuhui Huang. Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Tel: +852 23587363; Fax: +852 23587359; E-mail: xuhuihuang@ust.hk
*These authors contributed equally to this work.
Samuel Flores is an Assistant Professor at Uppsala University. He developed the MacroMoleculeBuilder internal coordinate modeling tool, and applied it to RNA structure prediction, ribosome dynamics and telomerase function.
Julie Bernauer is Senior Research Scientist in the AMIB project at INRIA. Her work focuses primarily on Structural Bioinformatics and specifically docking scoring functions and knowledge-based potential for molecular simulations.
Seokmin Shin is a Professor in Chemistry Department of Seoul National University. His research concerns the investigation of protein folding/misfolding, protein aggregation and peptide self-assembly by computer simulations.
Ruhong Zhou is Research Staff Scientist and Manager of Softmatter Theory & Simulation at IBM Research. His main research interests include protein folding, protein–protein interaction, confined water and nanoparticle–biomolecule interaction.
Xuhui Huang is an Assistant Professor in Chemistry Department of Hong Kong University of Science and Technology. His group is interested in modeling dynamics of conformational changes of biological macromolecules.

© The Author 2012. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 4. 395–405 doi:10.1093/bib/bbr077
Advance Access published on 6 January 2012

Multiscale modeling of macromolecular biosystems

Samuel C. Flores, Julie Bernauer*, Seokmin Shin, Ruhong Zhou and Xuhui Huang**

Submitted: 1st September 2011; Received (in revised form): 12th December 2011

Abstract

In this article, we review the recent progress in multiresolution modeling of structure and dynamics of protein, RNA and their complexes. Many approaches using both physics-based and knowledge-based potentials have been developed at multiple granularities to model both protein and RNA. Coarse graining can be achieved not only in the length, but also in the time domain using discrete time and discrete state kinetic network models. Models with different resolutions can be combined either in a sequential or parallel fashion. Similarly, the modeling of assemblies is also often achieved using multiple granularities. The progress shows that a multiresolution approach has considerable potential to continue extending the length and time scales of macromolecular modeling.

Keywords: *multiscale modeling; protein structure and dynamics; nucleic acid modeling; protein assemblies*

INTRODUCTION

Modeling the folding and assembly of biological macromolecules is challenging, since it is at large size and long time scales that require great computational cost. In response, many methods have been developed to reduce the computational cost by coarsening the granularity of the problem. Inevitably, the accuracy is lower for these methods and thus, in recent years, a consensus has emerged that one should work at multiple levels of resolution, simultaneously or in parallel [1] in a multiresolution approach.

At the coarsest granularity, an entire macromolecule can be treated as a single particle, whereas at the finest level each atom can be treated as a separate entity; indeed even higher levels of theory are possible [2]. There are many grainings in between, for example the kinematic or force unit may be a domain, a secondary structure element, a residue or chemical groups such as bases and ribose rings [3]. A simulation can be carried out at any level of resolution, or different levels can be used for the same problem either in serial or in parallel [1]. Furthermore, the force field and the kinematics can be treated at different levels of granularity [4], for example, secondary structure elements can be rigidified while interelement interactions can be treated at the all-atom level of resolution. The methods may consist of force fields [5, 6], sampling, structure prediction [5, 6] and dynamics tools [4], as well as novel algorithms.

Given the limited space, this mini-review is not aimed to be a complete review of this emerging field,

Corresponding author. Samuel C. Flores. Biomedical Center, Box 596, SE-751 24 Uppsala, Sweden. Tel: +46 0184714536; Fax: +46 18471530396; E-mail: samuelfloresc@gmail.com; Julie Bernauer. AMIB INRIA - Bioinformatique Laboratoire d'informatique (LIX), École Polytechnique 91128 Palaiseau Cedex, France. Tel: +33 (0)1 6933 4095; Fax: +33 (0)1 6933 4049; E-mail: julie.bernauer@inria.fr; Xuhui Huang. Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Tel: +852 23587363; Fax: +852 23587359; E-mail: xuhuihuang@ust.hk
*These authors contributed equally to this work.
Samuel Flores is an Assistant Professor at Uppsala University. He developed the MacroMoleculeBuilder internal coordinate modeling tool, and applied it to RNA structure prediction, ribosome dynamics and telomerase function.
Julie Bernauer is Senior Research Scientist in the AMIB project at INRIA. Her work focuses primarily on Structural Bioinformatics and specifically docking scoring functions and knowledge-based potential for molecular simulations.
Seokmin Shin is a Professor in Chemistry Department of Seoul National University. His research concerns the investigation of protein folding/misfolding, protein aggregation and peptide self-assembly by computer simulations.
Ruhong Zhou is Research Staff Scientist and Manager of Softmatter Theory & Simulation at IBM Research. His main research interests include protein folding, protein–protein interaction, confined water and nanoparticle–biomolecule interaction.
Xuhui Huang is an Assistant Professor in Chemistry Department of Hong Kong University of Science and Technology. His group is interested in modeling dynamics of conformational changes of biological macromolecules.

© The Author 2012. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

but instead to select a few specific examples in the following three main categories, some from our own studies, to illustrate the concept of multiscale modeling and its applications in macromolecular biological systems:

- protein structure and dynamics;
- strategies for the modeling of nucleic acids;
- interactions, complexes and assemblies.

MULTISCALE MODELING OF PROTEIN STRUCTURE AND DYNAMICS

Protein modeling is challenging because of the wide range of length and time spans. Modeling biological processes such as membrane pore assembly requires accurate description at both atomic/molecular scale (membrane protein) and mesoscopic/macroscopic scale (lipids) [7]. Furthermore, many biological processes of interest occur on a timescale (typically microseconds to milliseconds) that is much longer than what atomistic molecular dynamics (MD) simulations can reach routinely (typically hundreds of nanoseconds to microseconds). Therefore, it is computationally intractable to model complex protein systems in atomic detail, while also reaching biologically relevant timescales.

Multiscale approaches address the above challenge by coarse graining in either the length or the time domain. Simplifying protein representations in the length scale is a natural and popular solution to bridge the timescale gap [1, 8–10]. In such simulations, different length-scale resolutions can be mixed either in a sequential or parallel fashion [1]. The other solution for bridging the timescale gap is to construct models from short atomistic simulations to predict long timescale dynamics, effectively coarsening in the time domain. Discrete time and discrete state Markov State Model (MSM) is a method that has recently been developed to study microsecond or even millisecond protein-folding kinetics, by compiling many shorter (nanosecond-scale) atomistic simulations [11–18]. In this section, we will survey the multiscale modeling methods for proteins in both length and time domains.

Coarse graining in the length domain

Protein dynamics is governed by an underlying rugged free energy landscape. The number of local minima in this landscape increases exponentially with the system size. Therefore, sufficient sampling of the protein free energy surface using atomistic MD simulations is normally limited by computing power. Coarse-grained modeling provides an efficient way to alleviate the sampling problem by grouping multiple atoms into a single site. Coarse-grained models not only reduce the degrees of freedom of the system, but also smooth the potential energy surface [1, 19]. Thus, these coarse-grained simulations can run orders of magnitude faster than atomistic simulations.

Coarse-grained simulations can be parameterized based on information from atomistic simulations, knowledge from structural databases or even data from thermodynamic experiments [1]. In these simulations, different resolutions are employed in series rather than in parallel. Development of coarse-grained potentials for proteins can be traced back to the 1970s. In pioneering work by Levitt in 1975 [20], each protein residue was represented by three points: two for the main-chain and one for the side-chain. Since then, many protein coarse-grained models have been developed [8–10]. For example, Voth and Izvekov [21] developed the multiscale coarse-graining (MS-CG) potential, where force field parameters are extracted from atomistic MD simulations using a force matching procedure. Force field parameters obtained from this procedure are tabulated, and thus not restricted to any analytical functions. More recently, the Wu group [22] obtained parameters from both atomistic simulations and a coil library of high-resolution X-ray protein structures for their PACE coarse-grained force field [22]. The PACE force field has been shown to successfully fold small proteins [23].

Coarse-grained simulations can make the computations much more efficient. However, the fine-grained degrees of freedom also play important roles in many biological processes, so it is difficult to derive a single coarse-grained representation for the entire system that is both economical and accurate. For this reason, many research groups have recently combined fine-grained and coarse-grained representations in a single mixed-resolution simulation [1]. The hybrid Quantum Mechanics and Molecular Mechanics (QM/MM) simulations are one example of mixed-resolution simulations, where two resolutions are modeled in a single simulation system [8]. In another study using MS-CG method, Shi *et al.* [24] mixed all-atom and coarse-grained

model to simulate an ion channel embedded in the lipid bilayer. In their model, the ion channel was modeled at all-atoms resolution, whereas the lipid and water were modeled at coarse-grained resolution. They showed that their mixed-resolution simulations reproduce well the atomistic simulations, and are significantly faster. The Schulten group [25] has recently simulated the complex between the *lac* repressor protein (LacI) and a DNA segment using a mixed resolution model. In their simulation, the protein was described using the all-atom model, while the DNA loop connecting the two protein operators was described using a mathematical model mimicking a continuous elastic ribbon. Using this model, they revealed that the rotation of the head group is essential for the function of the LacI and further identified key residues that may lock LacI in a particular configuration.

Models at different resolutions can also be mixed in the framework of the replica exchange method (REM). Lwin and Luo [26] have developed a dual-resolution REM, where two replicas are simulated at high and low resolutions simultaneously. Exchanges of configurations are attempted periodically between these two replicas so that the low-resolution replica can help overcome entropic barriers in the energy landscape of the high-resolution model. Lyman *et al*. [27] have developed an REM where high-resolution, as well as low-resolution models exchange to accelerate the sampling of the high-resolution models. One of the major challenges for these methods is how to efficiently reconstruct the high-resolution configurations from low-resolution ones. Recently, Christen and van Gunsteren [28] proposed the multigraining algorithm to help the reconstruction process by describing the coarse-grained particles as virtual particles in the atomistic model. Moreover, they have defined a grain-level parameter λ to generate different replicas at intermediate levels of resolutions. They gradually change from atomistic model to the coarse-grained model; to keep a reasonable acceptance ratio, many intermediate levels are needed. More recently, Liu and Voth [19] proposed to relax the configuration at the coarse resolution before attempting the exchange. This 'smart' resolution REM scheme based on the smart walking method developed by Zhou and Berne [29] greatly increases the acceptance ratio so that only two replicas, one at the atomistic level and the other at the coarse-grained level are needed. This new method has been shown to quickly search the protein folded structure and also to approximately reproduce the Boltzmann distribution of the atomistic resolution model.

Coarse graining in the time domain

The biomolecular free energy landscape contains metastable free energy basins separated by free energy barriers. The presence of these metastable states in biomolecular dynamics has been suggested by various experiments. For example, using single-molecule FRET experiment, Zhuang *et al*. [30] observed four docked conformational states of distinct stabilities for a single hairpin ribozyme. In another relaxation dispersion NMR study, Mulder *et al*. [31] have identified two metastable states of the T4 lysozyme, and one state has around 2 kcal/mol higher free energy than the other one. If one can coarse grain conformational space into these metastable states, the fast motions within long-lived metastable states can then be further integrated out by coarse graining in the time domain. Discrete time and discrete state MSMs can automatically identify metastable states and calculate their equilibrium thermodynamics and kinetics. MSMs partition conformational space into a number of metastable states, and the resulting model has fast intrastate but slow interstate transitions. This separation of timescales can ensure that the model is Markovian at a discrete unit of Δt in time if Δt is longer than the fast intrastate relaxation time. Under this condition, the probability of a given state at time $t+\Delta t$ depends only on the state at time t. This allows MSMs built from short simulations to model long timescale dynamics. MSMs have been successfully applied to study protein conformational dynamics in a number of systems [11–18, 32].

MSMs are also inherently multiresolution due to the hierarchy of the free energy landscape. One can vary the resolution of an MSM by varying the degree of coarse graining in time as determined by the lag time. A short lag time will result in a high-resolution MSM with many metastable states. This high-resolution model will capture a large number of local free energy minima separated by small barriers. A long lag time will result in a low-resolution MSM with only a few states, and each of these states may contain multiple local free energy minima. Huang *et al*. [33] have developed the Super-level-set Hierarchical Clustering (SHC) algorithm that can

construct MSMs at multiple resolutions using hierarchical spectra clustering at different super-density levels. SHC is shown to be able to produce MSMs at different resolutions using different super-density level sets.

Different resolution models can also be mixed in a single MSM by integrating coarse-grain and atomistic subsystems. Kasson and Pande [34] have recently proposed the cross-graining algorithm to couple coarse-grained and atomistic simulations using MSMs. In this method, both the coarse-grained and atomistic subsystems are decomposed into discrete metastable states. Transition between a pair of states in the joint space is simply treated as the product of two subspace transitions by assuming that transitions in different subspaces are not correlated. This cross-graining method may provide a general way for simulating mixed-resolution systems such as membrane proteins.

In addition to the MSMs, there also exist other approaches of coarse-graining in the time domain. For example, Izaguirre *et al.* [35] have recently developed a new scheme to perform multiscale dynamics simulations using normal mode analysis. They first separate fast and slow modes. They then chose to propagate the dynamics only in the slow modes using a Langevin equation. This will allow as much as 500 times longer time step than that of atomistic simulations, and gain up to 200 times speed-up for protein simulations.

RNA MODELING

As explained, modeling of protein structure and dynamics has reached a degree of maturity, though considerable challenges remain. Computational modeling of RNA, on the other hand, has lagged behind proteins, due to the more recent appreciation of its importance, the greater experimental difficulty of crystallization, a smaller number of workers, the incompletely understood role of co-transcriptional [36] and hierarchical folding [37], and the theoretical issues of modeling its high charge, among other reasons. RNA workers have responded to the physicochemical differences between RNA and protein by developing computational tools that in many ways differ from those that work with proteins. In this section, we will discuss how dynamics and structure prediction are currently done in RNA. We then describe how these are being extended to greater time and length scales, with the objective of reaching the mesoscale, where the dynamics of molecules meets that of cells and even tissues.

We will encounter the recurring theme of multiresolution modeling. Following [1] what we did for proteins, we divide such methods into serial and parallel schemes. Serial methods include those in which RNA structure is solved by sampling from precompiled databases followed by evaluation [38, 39], training a coarse grained force field on such databases, then computing coarse-grained dynamics [40, 41], and potentially returning to fine resolution in a final step [42]. Parallel methods include those in which different molecules or regions are treated at different levels of kinematic resolution [3, 43], those in which the forces and kinematics are treated at different resolutions, and those in which time is coarse grained with multiple metastable states explored simultaneously [33].

RNA dynamics

The importance of RNA has only grown over the years, as its pervasive role in gene regulation has come to light. Computational methods for computing RNA dynamics have encountered roadblocks, which are not as important in the world of protein dynamics. Part of the challenge is that the RNA is a large, highly charged, very flexible molecule with a dearth of the distinctive surface features needed for recognition [44] and a propensity for kinetic traps. These issues in particular, challenge the widely acknowledged gold standard method of protein motion calculations, molecular dynamics. In response, the RNA community has experimented with multiple methods that restrict fine-grained calculations to selected regions. For instance, the Q program treats a small spherical region within a system using conventional MD, and surrounds this sphere with a restrained layer of water molecules [45]. This is useful for calculations (e.g. Free Energy Perturbation or Linear Interaction Energy) in a small region within a larger complex; the error in free energy change is on the order of 1 kcal/mol [45]. In an alternate approach, some workers decrease the granularity of the forces for the entire system [5, 40, 41]. For example, in Ref. [5] structural statistics are used to train a potential acting on a subset of atoms; this is useful for discriminating near-native structures from poorer quality decoys. Coarse-grained forces are typically knowledge-based (KB) [5] rather than physics-based. However, not all KB potentials are coarse-grained; some of the most

successful are atomistic [5, 6]. KB force fields have the advantage that they may not only be faster, but also circumvent theoretical roadblocks that often challenge physics-based methods [5, 6]. As with forces, the granularity of kinematics can be reduced either globally, as in [40, 41] where the velocities are only updated when pseudoatoms enter or leave each other's discrete neighborhoods, or in a region-dependent way [43]. Furthermore, the granularity of the kinematics can be the same as [40, 41] or different from [4] the granularity of the forces. The time integration methods are also more diverse, including variable time step [46] and collision-driven integrators [40, 47]. Last, it is possible to coarse-grain time and compute the probability of transitioning between metastable intermediates; for example in [33], many independent short-time trajectories of short RNA strand dynamics were connected to elucidate the statistical landscape of hairpin folding.

As mentioned, MD can be used for RNA modeling. It can even be used for large systems, famously including the ribosome [48]. However, the very long time scales require that such systems cannot be run unbiased until convergence [48]. One possibility for making unbiased physical simulation tractable is restricting the calculation to a sphere large enough to contain the region of interest, but not necessarily the entire solute molecule [45, 49]. Lastly, it is possible to effectively coarse-grain time using MSMs as described [33].

Many workers have abandoned physics-based force fields altogether in favor of KB force fields. However, these KB force fields have the disadvantage that having been trained on specific sets of observed phenomena, they cannot be expected to recapitulate phenomena not represented in training sets. Nevertheless, their saving grace is faster convergence, due to reduced degrees of freedom and/or force evaluations. One example of this approach is Discrete (DMD), which reduces complexity in three ways: (i) each residue is represented by three pseudoatoms, (ii) the potential is discontinuous, consisting of square wells and (iii) the time integration takes advantage of the resulting piecewise-constant particle velocities to update positions and velocities following a collision list [40, 47]. The Nucleic Acid Simulation Tool (NAST) reduces dimensionality even further, using a single pseudoatom to represent each residue; the user must provide base-pairing interactions to restrict the conformational search [41]. NAST can fold tRNA (within 8 Å RMSD of the native structure) and the P4P6 domain of the *Tetrahymena Group I Intron* (within 16 Å). MacroMoleculeBuilder (MMB) is more interactive; the force field consists of base-pairing interactions (of any type catalogued in [50]) and collision detecting spheres [46] (for preventing steric clashes), all specified by the user. Its internal coordinate framework [46] allows different regions of the molecule to be treated with different flexibility, e.g. any stretch of residues can be internally rigid for cost savings. It has been used to fold tRNA (within 9.6 Å) and P4P6 (within 10 Å) using biochemical and limited biophysical data [4]. It was also used to make a threaded model of a 200-nt ribozyme, coming within 4.6 Å of the native structure [3]. Last, it was able to easily model ribosomal translocation [43]. Due to its internal coordinate framework [51] it is particularly useful for modeling large complexes.

Thus, the state of the art in RNA dynamics has advanced significantly, with various methods found to reduce the problem space, degrees of freedom, forces and integration cost. These approaches have been useful for systems spanning a large size range, including the ribosome [43, 48, 49], ribozymes [3], tRNA [41] and small hairpins [33].

Structure prediction

Dynamics plays an important but not exclusive role in RNA structure prediction. There is no reported case of any RNA structure being solved by directly integrating the all-atoms equations of motion for the entire trajectory of folding, as has been done for proteins [52]. However, some structure prediction methods do use dynamics to minimize a coarse-grained potential, trained on structural data [40, 41]. These knowledge-based potentials can also be used at different resolutions to score the best high-resolution structures in an incremental way [5, 6]. In many cases, biochemical, biophysical and other specific nonstructural knowledge is used to restrain the problem [4, 41]. Other methods are not dynamical at all. A very popular approach is to assemble molecules using fragments drawn from structural databases, and evaluate these structures against a potential [39, 53]. Often, the structural sampling is done at a different resolution from the evaluation [39]. Fragment assembly is related to homology modeling [3, 54], which is becoming increasingly viable for RNA as the number of solved 3D structures increases.

The dynamical structure prediction codes include the mentioned DMD, NAST and MMB. DMD has successfully folded tRNA [40]. NAST [41] and MMB [4] have folded tRNA, as well as the P4P6 domain of the Tetrahymena Group I Intron. The latter two systems are the largest that have been solved using template-free dynamical methods. Threading can be done dynamically, as has been done for the entire Azoarcus Group I Intron [3].

Fragment assembly methods were introduced with MC-Sym [38], which has recapitulated numerous known structures [53] using a scoring function based on frequency of observation of fragments. Fragment Assembly of RNA (FARNA) similarly samples all-atom trinucleotide fragments from a structural database, and then evaluates the structures against a coarse-grained force field; it has predicted a number of small RNA structures *de novo* [55]. Like the dynamical methods, these have not yet been scaled to solve larger systems *de novo*. However, homology modeling can be done by fragment assembly, and this led to a model of the 16S subunit of the *Escherichia coli* ribosome [56].

COMPLEXES, ASSEMBLIES AND AGGREGATES

Having established protein- and RNA-specific modeling methods, we now go on to the analysis of large, possibly heterogeneous complexes. The function of a biomolecule largely depends on its interactions. Even if a large number of protein and nucleic acid structures are known, the structures of their assemblies remain mostly unknown. Modeling these assemblies is very complex as the number of degrees of freedom is large. Despite wide efforts and advanced techniques for studying protein and nucleic acid structures as described above, modeling assemblies and aggregates at different resolutions remains a challenge. Indeed upon interacting, the partners may undergo large conformational changes and the dynamics of such macromolecular machines are often intractable [57].

Docking is used to predict the structure of a complex when the individual structures of the components are known or can be modeled. Since the first description of a docking procedure in 1974, various techniques have been developed. They can be classified in two groups: rigid-body and flexible techniques. Their performances are evaluated in the community-wide experiment CAPRI (Critical Assessment of PRediction of Interactions) since 2001 [58]. It has shown interesting progress in the prediction of the interacting regions and in cases when flexibility is limited to small regions and changes [59].

Multiscale docking and assemblies

Protein–protein docking prediction techniques usually include three steps: finding putative complex conformations, scoring them to keep the most biologically relevant and refining the best scored structures [60, 61].

Finding suitable conformers involves 3-dimensional search and large conformational sampling. For very large assemblies, this cannot be easily achieved at the atomic level. Most of the protein–protein docking algorithms use coarse-grained representations for the initial sampling and scoring. To perform docking, the rigid-body procedures are widely used, however, it is also crucial to take flexibility into account for the partners.

Rigid body procedures imply finding putative candidates from the structures of the individual components, taken in their free (unbound) or complex (bound) form. They require an exhaustive spatial search for which many algorithms have been developed. Fast Fourier Transform (FFT) procedures are still commonly used [62–65], the level of representation being defined by the protein representation, such as those described for protein modeling (one, three or five pseudoatoms per residue mainly), or by the FFT grid size. The search is also done in direct space [66] or by geometric hashing [67], the protein models being coarse-grained. Aside from geometric hashing, these methods use algorithms that are computationally expensive and cannot deal with a fine grid size. The search is thus limited by the step size and may not lead to any usable results. The multiscale strategy involved is thus basic: reducing the grid/step size when external biological data or scoring functions provides information of putative epitopes. The strategy is successful in simple cases, for example protease-inhibitor complexes for which conformational changes are limited (e.g. a recent CAPRI example is the subtilisin Savinase—α-amylase subtilisin inhibitor BASI complex [59]).

Multilevel modeling is widely taken into account in scoring functions. A wide range of techniques are used, from data-driven docking, using conservation

or other experimental information [68], to machine learning techniques [69], physics-based force-fields [70] and statistical potentials [71]. Atomic details are often added after the scoring step to refine the prediction. The multilevel scoring step has shown to be a key part in the whole docking process [58]. As of today, scoring functions are still a bottleneck of docking procedures. They have yet proven able neither to predict binding affinities [72] nor to identify good conformations among a docking set in a blind setting [59].

Taking into account flexibility in a docking procedure is a very difficult task. Even if the flexibility is often limited to interface side chains [73], some complexes undergo large conformational changes as the docking benchmark shows [57]. Some docking procedures are able to deal with conformational changes and they make a great use of the different representation levels during the modeling. The RosettaDock suite is extremely well suited to this purpose. It can model backbone conformational changes using structural templates, model loops in free space and offer side-chain optimization either through a rotamer library or a well suited force field [74, 75]. Another approach is to use normal modes combined with different resolution levels to model the flexibility [76, 77]. In addition, geometric modeling has also been adapted [67, 78]. These methods mostly consider the molecules individually to model their flexibility and thus, cannot account for induced fit effects. They do provide some insights on the flexibility of the molecules but are often not accurate enough for describing the conformational changes involved in complex formation. For example, this sometimes leads to overly distorted models, such as the CAPRI results for various targets shows from Target 1 (HPR/HPR Kinase) [59, 79].

Multicomponent and symmetric docking can also be performed [67]. This is even a much harder problem but of great use when trying to fit experimentally obtained envelopes such as Electron Microscopy or Small Angle X-ray Scattering data. This problem is however, out of the scope of this review and will not be further discussed.

Due to its inherent complexity, protein–nucleic acid docking is lagging behind protein–protein docking. For example, some attempts have been made to predict RNA binding sites on proteins based on interaction propensity statistics combined with geometric calculations [80]. Some software suites such as PyDock [81] or HADDOCK [82] are also able to deal with protein–RNA and protein–DNA interaction prediction but benchmarking just recently appeared [83] and the CAPRI experiment conclusions show that the prediction techniques are not yet ready [59].

Despite a large number of published articles describing successful stories using multiscale procedures for various biological molecules of interest, automatic multiscale prediction with few or no biological external data is still limited. Over the recent years, the CAPRI experiment [59] and protein docking benchmark studies [57, 84] have shown that a satisfactory accuracy required for predicting interactomes [85] or binding affinity [72] has not yet been reached and results are still close to random when no or few external biological data is available. This may be due to the wide scale range the procedures have to accommodate, but also to the lack of efficient scoring functions both at the atomic and coarse-grained levels.

Aggregation

Protein aggregation and amyloid formation are key in the development of several diseases such as Alzheimer's, Parkinson's, Huntington's and Creutzfeldt–Jakob's. Computational modeling of protein aggregation has led to interesting insights on amyloid formation, such as the Ma-Nussinov-Tycko model for $\alpha\beta$ amyloid [86].

All-atom studies of such systems are mainly based on techniques described above for proteins. They make a great use of MD simulations including different representations of the systems. Replica Exchange Molecular Dynamics has allowed large simulations and normal mode-based simulations can account for conformation changes and description of the most stable state. These methods are well reviewed elsewhere [86, 87].

Interestingly, specific coarse-grained models have been developed for aggregation modeling [88]. The level of coarse-graining ranges from a few beads per residue to a cuboid per oligomer. Therefore, studies using various resolutions of models lead to the description of the aggregation phenomenon at a wide length and timescale.

Many existing coarse-grained models for aggregation are residue-based. For example, in the PRIME model, each residue is represented using four spheres (three for the backbone and one for the side chain) and $C\alpha$-$C\alpha$ distances are fixed using pseudo-bonds.

Simulations are made using Discontinuous Molecular Dynamics with hard sphere or square well potentials [89]. Another residue-based model proposed by Bellesia and Shea [90, 91], uses three points per residue and the backbone topology, adding a dihedral term to represent the flexibility and explicitly taking into account hydrogen-bonding and hydrophobic interactions for the peptide. This model is used in Langevin dynamic simulations. Another mid-resolution model used in Langevin dynamics simulations has been proposed by Caflisch and coworkers [92]. In this model, each peptide is described by four backbone beads and six side-chain beads. This model is not literally residue-based and partial charges are added to the backbone to represent the dipole. The flexibility is accounted for by a dihedral term populating a amyloid-protected state π and an amyloid competent state β.

Models with a resolution lower than one bead per residue are mainly used in Monte-Carlo (MC) simulations. The lattice model is made of eight connected beads on a lattice. Beads are tagged hydrophobic or polar and the extremities of the lattice are charged [93, 94]. Simulations using the lattice model are performed with local and global move sets. The tube model is a lower resolution model [95] that is used in MC or Discontinuous Molecular Dynamics simulations. Each peptide is represented by a tube, not accounting for the residue sequence. The tube thickness indicates the volume exclusion and bending stiffness. Hydrophobic interactions and hydrogen-bonding effects are modeled through geometric constraints. The lowest resolution model used is the cuboid [96]. In this model, a cuboid can be used to represent an extended or folded peptide or a small oligomer called a building block. Conformational changes of a building block are ignored. The interactions made by the cuboid are described by three parameters, corresponding to the three pairs of opposite sides of the cuboid. The parameters describe strong attraction between cuboids in the intrasheet hydrogen-bonding direction, weak attraction in the intersheet direction and repulsion in the direction parallel to the cuboid building block. Only single-building block moves are considered in the MC simulations using this model.

Most of these models are used for $\alpha\beta$ amyloid formation studies. Despite the various coarse-grain sizes used, these models can often not be connected in a fully multiscale procedure. This may lead to different results at different scales and precludes yet, a full description strategy of $\alpha\beta$ amyloid formation, for example, from atomic resolution dimer formation to large fiber analysis.

CONCLUSION

We reviewed how novel multiresolution approaches are making inroads in structure and dynamics of protein, RNA and complexes. Many new special- and general-purpose force fields and potentials have been developed, with different force and energy granularities. Structures have been solved using a variety of dynamical, minimization and Monte-Carlo approaches, often with kinematic or sampling granularity that differs from that of the corresponding potential or force field. Similarly for the prediction of assemblies, many geometric representations are in use, while kinematics and potentials can change granularities from stage to stage in a calculation.

Key Points

- Knowledge-based potentials and force fields are available at different granularities for different purposes.
- Coarse graining can be done not only in the length but also in the time domain.
- RNA modeling has peculiarities that are treated with special-purpose force fields and potentials, again at multiple granularities.
- The modeling of assemblies is often achieved by shifting kinematic or force granularity, and treating special regions such as interfaces at different flexibility.

FUNDING

The Hong Kong Research Grants Council (grants 661011 and HKUST2/CRF/10 to X.H.); INRIA Equipe Associee Program (GNAPI) (to J.B); National Research Foundation of Korea (grant nos 305-20100005 and 2010-0001630 to S.S.); IBM Blue Gene Science Program; Swedish Research Council [the eSSENCE project grant (essenceofscience.se) to S.F.]; the Swedish Foundation for International Cooperation in Research and Higher Education (stint.se) (travel funding to S.F.).

References

1. Ayton GS, Noid WG, Voth GA. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr Opin Struct Biol* 2007;**17**:192–8.

2. Svensson MHS, Froese RDJ, Matsubara T, *et al*. ONIOM: a multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)3)2 + H2 oxidative. *J Phys Chem* 1996;**100**:19357.
3. Flores S, Wan Y, Russell R, *et al*. Predicting RNA structure by multiple template homology modeling. *Pac Symp Biocomput* 2010;216–27.
4. Flores S, Altman R. Turning limited experimental information into 3D models of RNA. *RNA* 2010;**16**:1769–78.
5. Bernauer J, Huang X, Sim AY, *et al*. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* 2011;**17**:1066–75.
6. Capriotti E, Norambuena T, Marti-Renom MA, *et al*. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics* 2011;**27**: 1086–93.
7. Ayton GS, Voth GA. Systematic multiscale simulation of membrane protein systems. *Curr Opin Struct Biol* 2009;**19**: 138–44.
8. Sherwood P, Brooks BR, Sansom MS. Multiscale methods for macromolecular simulations. *Curr Opin Struct Biol* 2008; **18**:630–40.
9. Li W, Yoshii H, Hori N, *et al*. Multiscale methods for protein folding simulations. *Methods* 2010;**52**:106–14.
10. Tozzini V. Coarse-grained models for proteins. *Curr Opin Struct Biol* 2005;**15**:144–50.
11. Chodera JD, Singhal N, Pande VS, *et al*. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 2007;**126**:155101.
12. Bowman GR, Beauchamp KA, Boxer G, *et al*. Progress and challenges in the automated construction of Markov state models for full protein systems. *J Chem Phys* 2009;**131**: 124101.
13. Noe F, Schutte C, Vanden-Eijnden E, *et al*. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 2009; **106**:19011–6.
14. Morcos F, Chatterjee S, McClendon CL, *et al*. Modeling conformational ensembles of slow functional motions in Pin1-WW. *Plos Comput Biol* 2010;**6**:e1001015.
15. Voelz VA, Bowman GR, Beauchamp K, *et al*. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* 2010;**132**:1526–8.
16. Bowman GR, Voelz VA, Pande VS. Atomistic folding simulations of the five-helix bundle protein lambda(6-85). *J Am Chem Soc* 2011;**133**:664–7.
17. Bowman GR, Voelz VA, Pande VS. Taming the complexity of protein folding. *Curr Opin Struct Biol* 2011;**21**:4–11.
18. Silva DA, Bowman GR, Sosa-Peinado A, *et al*. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *Plos Comput Biol* 2011;**7**: e1002054.
19. Liu P, Voth GA. Smart resolution replica exchange: an efficient algorithm for exploring complex energy landscapes. *J Chem Phys* 2007;**126**:045106.
20. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* 1976;**104**:59–107.
21. Izvekov S, Voth GA. A multiscale coarse-graining method for biomolecular systems. *J Phys Chem B* 2005; **109**:2469–73.
22. Han W, Wan CK, Jiang F, *et al*. PACE force field for protein simulations. 1. Full parameterization of version 1 and verification. *J Chem Theory Comput* 2010;**6**:3373–89.
23. Han W, Wan CK, Wu YD. PACE force field for protein simulations. 2. Folding Simulations of peptides. *J Chem Theory Comput* 2010;**9**:3390–3402.
24. Shi Q, Izvekov S, Voth GA. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane-bound ion channel. *J Phys Chem B* 2006;**110**: 15045–8.
25. Villa E, Balaeff A, Schulten K. Structural dynamics of the lac repressor-DNA complex revealed by a multiscale simulation. *Proc Natl Acad Sci USA* 2005;**102**:6783–8.
26. Lwin TZ, Luo R. Overcoming entropic barrier with coupled sampling at dual resolutions. *J Chem Phys* 2005; **123**:194904.
27. Lyman E, Ytreberg FM, Zuckerman DM. Resolution exchange simulation. *Phys Rev Lett* 2006;**96**:028105.
28. Christen M, van Gunsteren WF. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *J Chem Phys* 2006;**124**: 154106.
29. Zhou RH, Berne BJ. Smart walking: A new method for Boltzmann sampling of protein conformations. *J Chem Phys* 1997;**107**:9185–96.
30. Zhuang X, Kim H, Pereira MJ, *et al*. Correlating structural dynamics and function in single ribozyme molecules. *Science* 2002;**296**:1473–6.
31. Mulder FA, Mittermaier A, Hon B, *et al*. Studying excited states of proteins by NMR spectroscopy. *Nat Struct Biol* 2001;**8**:932–5.
32. Zhuang W, Cui RZ, Silva DA, *et al*. Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the Markov state model approach. *J Phys Chem B* 2011;**115**:5415–24.
33. Huang X, Yao Y, Bowman GR, *et al*. Constructing multi-resolution markov state models (msms) to elucidate RNA hairpin folding mechanisms. *Pac Symp Biocomput* 2010; 228–39.
34. Kasson PM, Pande VS. "Cross-graining": efficient multi-scale simulation via markov state models. *Pac Symp Biocomput* 2010;260–8.
35. Izaguirre JA, Sweet CR, Pande VS. Multiscale dynamics of macromolecules using normal mode langevin. *Pac Symp Biocomput* 2010;240–51.
36. Pan T, Sosnick T. RNA folding during transcription. *Annu Rev Biophys Biomol Struct* 2006;**35**:161–75.
37. Brion P, Westhof E. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct* 1997;**26**: 113–37.
38. Major F, Gautheret D, Cedergren R. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc Natl Acad Sci USA* 1993;**90**: 9408–12.
39. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci USA* 2007;**104**:14664–9.

40. Ding F, Sharma S, Chalasani P, *et al*. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* 2008;**14**:1164–73.
41. Jonikas MA, Radmer RJ, Laederach A, *et al*. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* 2009;**15**:189–99.
42. Jonikas MA, Radmer RJ, Altman RB. Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics* 2009;**25**:3259–66.
43. Flores S, Altman R. Structural insights into pre-translocation ribosome motions. *Pac Symp Biocomput* 2011; **16**:205–16.
44. Ferre-D'Amare AR, Zhou K, Doudna JA. A general module for RNA crystallization. *J Mol Biol* 1998;**279**: 621–31.
45. Marelius J, Kolmodin K, Feierberg I, *et al*. Q: a molecular dynamics program for free energy calculations and empirical valence bond simulations in biomolecular systems. *J Mol Graph Model* 1998;**16**:213–25.
46. Flores S, Sherman M, Bruns C, *et al*. Fast flexible modeling of macromolecular structure using internal coordinates. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:1247–57.
47. Zhou Y, Karplus M. Folding thermodynamics of a model three-helix-bundle protein. *Proc Natl Acad Sci USA* 1997;**94**: 14429–32.
48. Sanbonmatsu KY, Joseph S, Tung CS. Simulating movement of tRNA into the ribosome during decoding. *Proc Natl Acad Sci USA* 2005;**102**:15854–9.
49. Sund J, Ander M, Aqvist J. Principles of stop-codon reading on the ribosome. *Nature* 2010;**465**:947–50.
50. Leontis NB, Stombaugh J, Westhof E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* 2002;**30**:3497–531.
51. Flores SC, Sherman MA, Bruns CM, *et al*. Fast flexible modeling of RNA structure using internal coordinates. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:1247–57.
52. Pande VS, Baker I, Chapman J, *et al*. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers* 2003;**68**: 91–109.
53. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 2008;**452**: 51–5.
54. Rother M, Rother K, Puton T, *et al*. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res* 2011;**39**:4007–22.
55. Das R, Kudaravalli M, Jonikas M, *et al*. Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci USA* 2008;**105**:4144–9.
56. Tung CS, Joseph S, Sanbonmatsu KY. All-atom homology model of the Escherichia coli 30S ribosomal subunit. *Nat Struct Biol* 2002;**9**:750–5.
57. Hwang H, Vreven T, Janin J, *et al*. Protein-protein docking benchmark version 4.0. *Proteins* 2010;**78**:3111–4.
58. Janin J. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst* 2010;**6**:2351–62.
59. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. *Proteins* 2010;**78**:3073–84.
60. Ritchie DW. Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci* 2008;**9**:1–15.
61. Vajda S, Kozakov D. Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol* 2009;**19**:164–70.
62. Eisenstein M, Ben-Shimon A, Frankenstein Z, *et al*. CAPRI targets T29-T42: proving ground for new docking procedures. *Proteins* 2010;**78**:3174–81.
63. Huang SY, Zou X. MDockPP: a hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins* 2010;**78**:3096–103.
64. Hwang H, Vreven T, Pierce BG, *et al*. Performance of ZDOCK and ZRANK in CAPRI rounds 13-19. *Proteins* 2010;**78**:3104–10.
65. Kozakov D, Hall DR, Beglov D, *et al*. Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. *Proteins* 2010;**78**:3124–30.
66. Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins* 2010;**78**:3131–9.
67. Mashiach E, Schneidman-Duhovny D, Peri A, *et al*. An integrated suite of fast docking algorithms. *Proteins* 2010; **78**:3197–204.
68. de Vries SJ, Melquiond AS, Kastritis PL, *et al*. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins* 2010;**78**:3242–9.
69. Bourquard T, Bernauer J, Aze J, *et al*. A collaborative filtering approach for protein-protein docking scoring functions. *PLoS One* 2011;**6**:e18541.
70. Li X, Moal IH, Bates PA. Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding. *Proteins* 2010;**78**: 3189–96.
71. Vreven T, Hwang H, Weng Z. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci* 2011;**20**:1576–86.
72. Fleishman SJ, Whitehead TA, Strauch EM, *et al*. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J Mol Biol* 2011;**414**:289–302.
73. Andrusier N, Mashiach E, Nussinov R, *et al*. Principles of flexible protein-protein docking. *Proteins* 2008;**73**: 271–89.
74. Movshovitz-Attias D, London N, Schueler-Furman O. On the use of structural templates for high-resolution docking. *Proteins* 2010;**78**:1939–49.
75. Sircar A, Chaudhury S, Kilambi KP, *et al*. A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins* 2010;**78**: 3115–23.
76. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. *Int J Mol Sci* 2010;**11**: 3623–48.
77. Zacharias M. Accounting for conformational changes during protein-protein docking. *Curr Opin Struct Biol* 2010; **20**:180–6.
78. Zhang Q, Sanner M, Olson AJ. Shape complementarity of protein-protein complexes at multiple resolutions. *Proteins* 2009;**75**:453–67.
79. Mendez R, Leplae R, De Maria L, *et al*. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* 2003;**52**:51–67.

80. Perez-Cano L, Fernandez-Recio J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins* 2009;**78**:25–35.
81. Perez-Cano L, Solernou A, Pons C, *et al*. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. *Pac Symp Biocomput* 2009;293–301.
82. van Dijk M, van Dijk AD, Hsu V, *et al*. Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res* 2006;**34**:3317–25.
83. van Dijk M, Bonvin AM. A protein-DNA docking benchmark. *Nucleic Acids Res* 2008;**36**:e88.
84. Kastritis PL, Moal IH, Hwang H, *et al*. A structure-based benchmark for protein-protein binding affinity. *Protein Sci* 2011;**20**:482–91.
85. Kastritis PL, Bonvin AM. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J Proteome Res* 2010;**9**:2216–25.
86. Ma B, Nussinov R. Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. *Curr Opin Chem Biol* 2006;**10**:445–2.
87. Straub JE, Thirumalai D. Principles governing oligomer formation in amyloidogenic peptides. *Curr Opin Struct Biol* 2010;**20**:187–95.
88. Wu C, Shea JE. Coarse-grained models for protein aggregation. *Curr Opin Struct Biol* 2011;**21**:209–20.
89. Cheon M, Chang I, Hall CK. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins* 2010;**78**:2950–60.
90. Bellesia G, Shea JE. Diversity of kinetic pathways in amyloid fibril formation. *J Chem Phys* 2009;**131**:111102.
91. Bellesia G, Shea JE. Effect of beta-sheet propensity on peptide aggregation. *J Chem Phys* 2009;**130**:145103.
92. Friedman R, Pellarin R, Caflisch A. Amyloid aggregation on lipid bilayers and its impact on membrane permeability. *J Mol Biol* 2009;**387**:407–15.
93. Li MS, Co NT, Reddy G, *et al*. Factors governing fibrillogenesis of polypeptide chains revealed by lattice models. *Phys Rev Lett* 2011;**105**:218101.
94. Li MS, Klimov DK, Straub JE, *et al*. Probing the mechanisms of fibril formation using lattice models. *J Chem Phys* 2008;**129**:175101.
95. Auer S, Trovato A, Vendruscolo M. A condensation-ordering mechanism in nanoparticle-catalyzed peptide aggregation. *Plos Comput Biol* 2009;**5**:e1000458.
96. Zhang J, Muthukumar M. Simulations of nucleation and elongation of amyloid fibrils. *J Chem Phys* 2009;**130**:035102.

PROTEIN-RNA COMPLEXES AND EFFICIENT AUTOMATIC DOCKING: EXPANDING ROSETTADOCK POSSIBILITIES

Adrien GUILHOT-GAUDEFFROY, Christine FROIDEVAUX, Jérôme AZÉ, Julie BERNAUER. In *PLoS One*, 9(9):e108928, 2014.

OPEN ACCESS Freely available online

PLOS ONE

Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities

Adrien Guilhot-Gaudeffroy[1,2,3], Christine Froidevaux[1,2], Jérôme Azé[1,2,4], Julie Bernauer[1,3]*

1 AMIB Project, Inria Saclay-Île de France, Palaiseau, France, **2** Laboratoire de Recherche en Informatique (LRI), CNRS UMR 8623, Université Paris-Sud, Orsay, France, **3** Laboratoire d'Informatique de l'Ecole Polytechnique (LIX), CNRS UMR 7161, Ecole Polytechnique, Palaiseau, France, **4** Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), CNRS UMR 5506, Université Montpellier 2, Montpellier, France

Abstract

Protein-RNA complexes provide a wide range of essential functions in the cell. Their atomic experimental structure solving, despite essential to the understanding of these functions, is often difficult and expensive. Docking approaches that have been developed for proteins are often challenging to adapt for RNA because of its inherent flexibility and the structural data available being relatively scarce. In this study we adapted the RosettaDock protocol for protein-RNA complexes both at the nucleotide and atomic levels. Using a genetic algorithm-based strategy, and a non-redundant protein-RNA dataset, we derived a RosettaDock scoring scheme able not only to discriminate but also score efficiently docking decoys. The approach proved to be both efficient and robust for generating and identifying suitable structures when applied to two protein-RNA docking benchmarks in both bound and unbound settings. It also compares well to existing strategies. This is the first approach that currently offers a multi-level optimized scoring approach integrated in a full docking suite, leading the way to adaptive fully flexible strategies.

Citation: Guilhot-Gaudeffroy A, Froidevaux C, Azé J, Bernauer J (2014) Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities. PLoS ONE 9(9): e108928. doi:10.1371/journal.pone.0108928

Editor: Jinn-Moon Yang, National Chiao Tung University, Taiwan

Received August 1, 2014; **Accepted** September 5, 2014; **Published** September 30, 2014

Copyright: © 2014 Guilhot-Gaudeffroy et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was granted access to the HPC resources of TGCC (Très Grand Centre de calcul du CEA - http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm) under the allocation t2013077065 made by GENCI (Grand Equipement National de Calcul Intensif - http://www.genci.fr). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: julie.bernauer@inria.fr

Introduction

Protein-RNA interactions often play a major role in the cell. They are involved in many processes such as replication, mRNA transcription or regulation of RNA levels and control the operation of key cellular machineries such as the RNA induced silencing complex (RISC). They are thus good candidates for therapeutic studies [1]. The variety of proteins able to bind RNA molecule is very large and covers a wide range of protein domains. This includes domains such as **RRM** and dsRIB which all show RNA binding activity and are well studied [2]. In the recent years, experimental techniques have shed the light on RNA and protein-RNA complexes. X-ray Crystallography [3] and NMR [4,5] have provided high-resolution structures offering insights into RNA function and binding activity and modes [6,7] but other experimental techniques have also allowed for the analysis of larger ensembles [8–10]. Single-molecule experiments can now provide high-resolution data [11] and the engineering of RNA binding molecule is with reach [12]. Despite the wide interest and advances in structural biology for RNA and protein-RNA complexes, the number of structures available in the PDB is relatively small (a few thousand for RNA molecules and around a thousand for protein-RNA complexes). And both the modelling and the prediction of protein-RNA interactions remain a challenge [13].

The structural modelling of large biomolecules and their interactions is a challenging task. A large number of methods for both predicting and evaluating the results have been developed [14–16] and the Critical Assessment of PRediction of Interactions (CAPRI http://capri.ebi.ac.uk) challenge [17] which allowed for an international blind prediction setting has shown that despite great progress, the methods available still rely on a great variety of biological data to be available [18] and the flexibility of the molecules remain a modelling and computational issue to overcome [19]. The techniques are however now able to integrate more data and predict better ion and water molecules which mediate the binding [20]. Binding affinity is not yet a predictable quantity but the originality and first results of the latest strategies is encouraging [21].

Protein-RNA complexes are especially difficult to predict and model for two reasons: the inherent flexibility of RNA molecules and the electrostatics driving the binding as the RNA molecule is negatively charged. Progress in RNA structure prediction and folding [22–26] allows to deal with flexibility but have yet to be fully multi-scale [27] and integrated in the docking processes. This can be done once the scoring function for protein-RNA are efficient enough and provide accurate conformation selection. Specially designed coarse-grained force-fields based on statistics [28–32] have shown great promises and coarse-grained versions for reducing the initial exploration phase of coarse-grained search are interesting [33,34]. The optimization is however often based

OPEN ACCESS Freely available online

PLOS ONE

Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities

Adrien Guilhot-Gaudeffroy[1,2,3], Christine Froidevaux[1,2], Jérôme Azé[1,2,4], Julie Bernauer[1,3]*

1 AMIB Project, Inria Saclay-Île de France, Palaiseau, France, **2** Laboratoire de Recherche en Informatique (LRI), CNRS UMR 8623, Université Paris-Sud, Orsay, France, **3** Laboratoire d'Informatique de l'École Polytechnique (LIX), CNRS UMR 7161, École Polytechnique, Palaiseau, France, **4** Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), CNRS UMR 5506, Université Montpellier 2, Montpellier, France

Abstract

Protein-RNA complexes provide a wide range of essential functions in the cell. Their atomic experimental structure solving, despite essential to the understanding of these functions, is often difficult and expensive. Docking approaches that have been developed for proteins are often challenging to adapt for RNA because of its inherent flexibility and the structural data available being relatively scarce. In this study we adapted the RosettaDock protocol for protein-RNA complexes both at the nucleotide and atomic levels. Using a genetic algorithm-based strategy, and a non-redundant protein-RNA dataset, we derived a RosettaDock scoring scheme able not only to discriminate but also score efficiently docking decoys. The approach proved to be both efficient and robust for generating and identifying suitable structures when applied to two protein-RNA docking benchmarks in both bound and unbound settings. It also compares well to existing strategies. This is the first approach that currently offers a multi-level optimized scoring approach integrated in a full docking suite, leading the way to adaptive fully flexible strategies.

Citation: Guilhot-Gaudeffroy A, Froidevaux C, Azé J, Bernauer J (2014) Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities. PLoS ONE 9(9): e108928. doi:10.1371/journal.pone.0108928

Editor: Jinn-Moon Yang, National Chiao Tung University, Taiwan

Received August 1, 2014; **Accepted** September 5, 2014; **Published** September 30, 2014

Copyright: © 2014 Guilhot-Gaudeffroy et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was granted access to the HPC resources of TGCC (Très Grand Centre de calcul du CEA - http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm) under the allocation t2013077065 made by GENCI (Grand Equipement National de Calcul Intensif - http://www.genci.fr). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: julie.bernauer@inria.fr

Introduction

Protein-RNA interactions often play a major role in the cell. They are involved in many processes such as replication, mRNA transcription or regulation of RNA levels and control the operation of key cellular machineries such as the RNA induced silencing complex (RISC). They are thus good candidates for therapeutic studies [1]. The variety of proteins able to bind RNA molecule is very large and covers a wide range of protein domains. This includes domains such as RRM and dsRDB which all show RNA binding activity and are well studied [2]. In the recent years, experimental techniques have shed the light on RNA and protein-RNA complexes. X-ray Crystallography [3] and NMR [4,5] have provided high-resolution structures offering insights into RNA function and binding activity and modes [6,7] but other experimental techniques have also allowed for the analysis of larger ensembles [8–10]. Single-molecule experiments can now provide high-resolution data [11] and the engineering of RNA binding molecule is with reach [12]. Despite the wide interest and advances in structural biology for RNA and protein-RNA complexes, the number of structures available in the PDB is relatively small (a few thousand for RNA molecules and around a thousand for protein-RNA complexes). And both the modelling and the prediction of protein-RNA interactions remain a challenge [13].

The structural modelling of large biomolecules and their interactions is a challenging task. A large number of methods for both predicting and evaluating the results have been developed [14–16] and the Critical Assessment of PRediction of Interactions (CAPRI http://capri.ebi.ac.uk) challenge [17] which allowed for an international blind prediction setting has shown that despite great progress, the methods available still rely on a great variety of biological data to be available [18] and the flexibility of the molecules remain a modelling and computational issue to overcome [19]. The techniques are however now able to integrate more data and predict better ion and water molecules which mediate the binding [20]. Binding affinity is not yet a predictable quantity but the originality and first results of the latest strategies is encouraging [21].

Protein-RNA complexes are especially difficult to predict and model for two reasons: the inherent flexibility of RNA molecules and the electrostatics driving the binding as the RNA molecule is negatively charged. Progress in RNA structure prediction and folding [22–26] allows to deal with flexibility but have yet to be fully multi-scale [27] and integrated in the docking processes. This can be done once the scoring function for protein-RNA are efficient enough and provide accurate conformation selection. Specially designed coarse-grained force-fields based on statistics [28–32] have shown great promises and coarse-grained versions for reducing the initial exploration phase of coarse-grained search are interesting [33,34]. The optimization is however often based

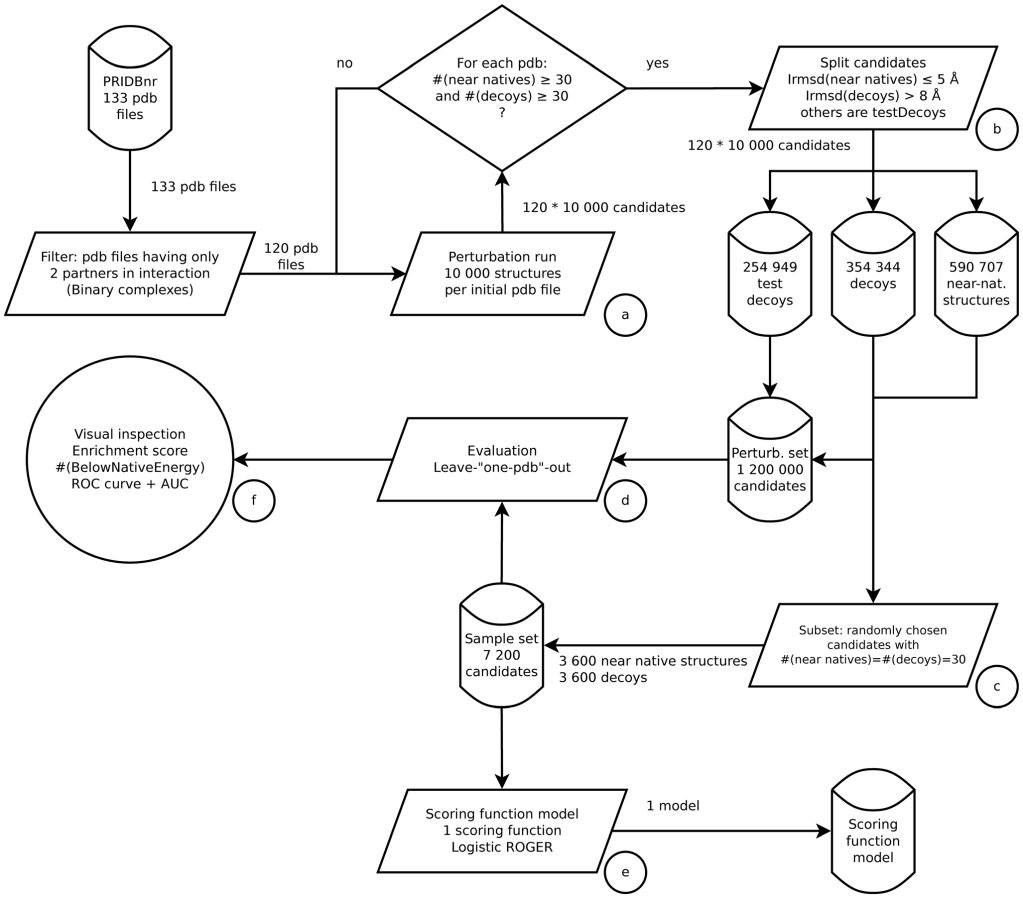


Figure 1. Flowchart of the machine learning strategy. The procedure is made of six steps: a) data processing using the non-redundant PRIDB to generate candidates, b) splitting of candidates into test decoys, decoys, near-native structures according to their Irmsd so as to define the perturbation set, c) definition of the sample set using 30 near-native structures and 30 decoys per native structure - randomly chosen, d) leave-one-pdb-out evaluation, e) scoring function learning using ROGER and f) result analysis.
doi:10.1371/journal.pone.0108928.g001

on relatively simple statistics measurements and rarely benefits from the variety of structural datasets recently made available to the community. The Protein-RNA interface database [35] offers high quality curated datasets for statistical analysis. Both available in a redundant and non-redundant version it allows for fine measurements on high-resolution structures. The three protein-RNA benchmarks available in the literature [36–38] also offer a great opportunity to assess and review high-resolution structures and predictions.

The availability of structural data is essential for machine learning based strategies for scoring in docking experiments. Various machine learning strategies have been developed in the past for protein-protein complexes [39–43] and have proven to be key in reranking and optimizing docking experiments for protein-protein complexes as the last CAPRI rounds has shown [44,45]. In this study, we use a machine-learning based strategy to optimize the well-known RosettaDock scoring function for high-resolution docking. RosettaDock is a leading edge protein-docking suite [46–48] which while being very versatile and widely used have been only seldom used for protein-RNA docking [28,49]. We first extended the RosettaDock low resolution model to RNA for both searching and scoring. We then used the Protein Interface Database [35] as reference dataset to generated near-native and plausible docking conformations. We then optimized the RosettaDock high-resolution scoring function using supervised machine learning. After cross-validation and carefully handling tests, we assessed the obtained protocol on the protein docking benchmarks I and II [36,38]. We show that the obtained RosettaDock RNA protocol performs better than in the previous attempts [49] in a semi-rigid body approach for both bound and unbound docking and can undoubtedly be used for successful protein-RNA predictions.

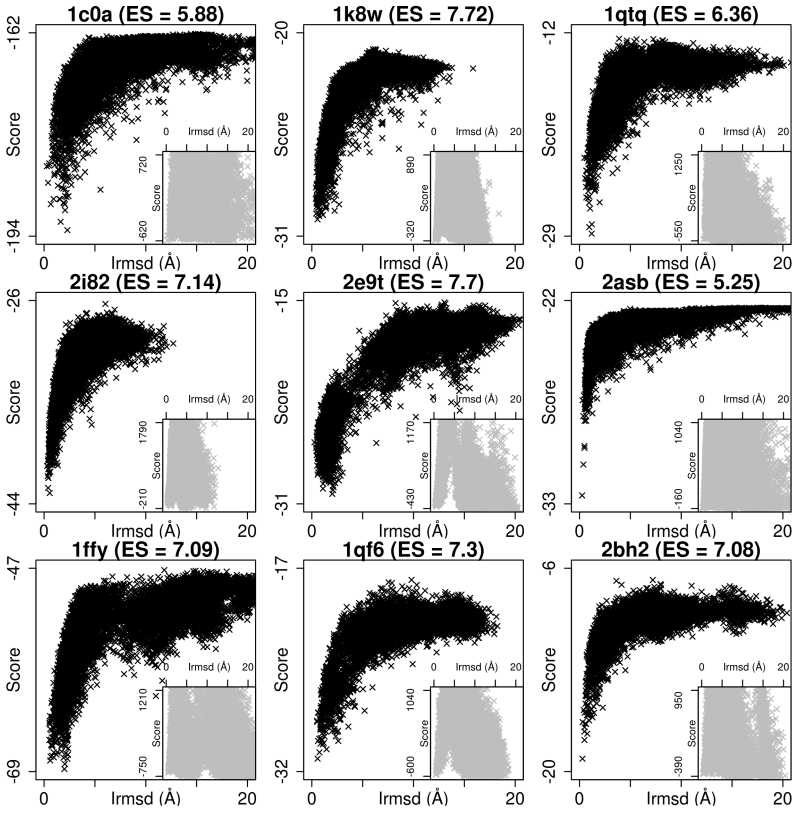


Figure 2. Energy vs Irmsd for 9 protein-RNA complexes. The 10,000 conformations evaluated for our optimized RosettaDock scoring function are shown in black. On each plot, the bottom left panel shows the equivalent non-optimized RosettaDock result.
doi:10.1371/journal.pone.0108928.g002

Materials and Methods

Protein-RNA complexes training and evaluation sets for RosettaDock

Protein-RNA native X-ray structures for learning were downloaded from the Protein-RNA Interface Database (PRIDB) [35]. The non-redundant PRIDB (RB199) contains 199 RNA chains extracted from the PDB in 2010. From the 134 complexes described in this set, we only kept the binary complexes: one protein and one RNA molecule. We also discarded complexes involving the ribosome because of their redundancy and to avoid biasing towards ribosome data but also to avoid computationally expensive procedures. The resulting native structure dataset from the PRIDB is made of 120 complexes (Table S1).

We also used the two protein-RNA benchmarks [36,38] as a validation set in bound and unbound (protein and RNA when available) settings. Among the 45 complexes contained in the Benchmark I [36], 11 complexes are not found in the PRIDB. Among the 106 complexes from the Benchmark II, we only kept the 76 complexes for which an unbound structure of the protein exists. Among these 76 complexes, 36 cannot be found in the PRIDB. After checking for overlap on the two benchmarks which were obtained using two different strategies, the resulting test set is made of 40 complexes. The list of complexes used in this study can be found in Table S2.

From all the native structures from both the PRIDB and the benchmarks, near-native and decoy conformations are generated using the Rosetta perturbation protocol [47]. For each pdb file, 10,000 perturbation conformations are to be obtained. Among these 10,000, to allow for correct learning, we want 30 near-native conformations whose Irmsd is smaller than 5 Å and 30 decoy conformations whose Irmsd is greater than 8 Å. Irmsd definition is taken from [14] and adapted to protein-RNA complexes by using the RNA backbone P atoms. For that purpose, the amplitude of the translation and the three rotations applied is chosen to follow a normal law of variance 1 and different expectations (small, regular and large). The regular setting is set to 3 Å for the translation and 8° for the rotations, the small (resp. large) setting is set to 1 Å (resp. 9 Å) for the translation and 4° (resp 27°) for the rotations. For each pdb file, the setting chosen is the smallest allowing for enough near-native and decoy conformation generation.

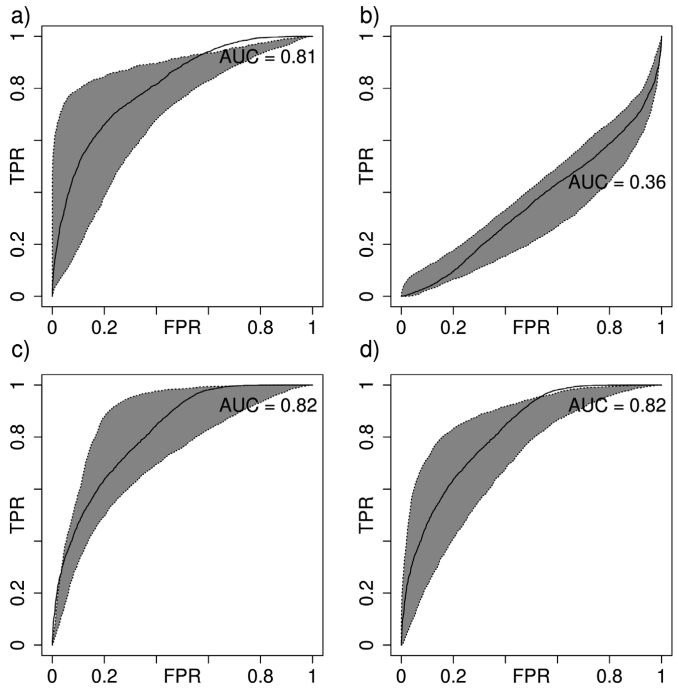


Figure 3. ROC Curves (True Positive Rate -TPR- vs. False Positive Rate -FPR-). (a) ROGER logistic scoring function, (b) Default RosettaDock score, (c) the whole protein-RNA benchmark I, (d) the whole protein-RNA benchmark II. The median ROC Area Under the Curve (AUC) is shown as a black line. The dotted lines delimiting the gray area correspond to the 1st and 3rd quartiles. Reported on the plots are ROC-AUC values for the median.
doi:10.1371/journal.pone.0108928.g003

RosettaDock protocol and scoring functions

The RosettaDock protocol is two-level docking search: low resolution and high resolution. The low resolution stage uses a coarse-grained representation of the partners to quickly sample the search space for candidates. The high resolution stage rebuilds the all-atom partners from the low resolution candidates to perform a refined atomic search possibly including rotamer search and loop optimization.

The low resolution scoring function uses the backbone of the molecule and one centroid per residue [47] and contains five weighted terms:

$$S_{Lowres} = w_{Contact}S_{Contact} + w_{Bump}S_{Bump} + w_{Env}S_{Env} + w_{Pair}S_{Pair} + w_{Align}S_{Align}$$

where $S_{Contact}$ represents the number of interface residues being defined by having a centroid less than 6 Å away from a centroid in the other partner; S_{Bump} is a distance-based penalty for steric clashes; S_{Env} defines the probability of finding a residue in a specific environment (buried/exposed and interface/non interface); S_{Pair} is a pair potential defining the propensity of residues to be found in interaction in given environments and S_{Align} is an optional term to match a specific alignment pattern (e.g. antibodies).

These five terms of the low resolution score can be computed for protein-RNA complexes in the same way they were for proteins. For RNA, the backbone is chosen to include the sugar ring and the centroid is taken to be the center of mass of the base. All the parameters for the low resolution scoring terms are computed on the PRIDB reference set.

The high resolution scoring function uses all the atoms of the molecules, including the hydrogen atoms, and is made of seven weighted terms:

$$S_{Highres} = w_{VdW}S_{VdW} + w_{Elec}S_{Elec} + w_{Solv}S_{Solv} + w_{Hbond}S_{Hbond} + w_{SASA}S_{SASA} + w_{Pair}S_{Pair} + w_{Rotamer}S_{Rotamer}$$

where S_{VdW} is a Van der Waals term (Lennard-Jones based), S_{Elec} is a Coulomb term, S_{Solv} a solvent term based on the Lazaridis-Karplus model, S_{Hbond} is a H-bond 10–12 potential term, S_{SASA} is the solvent accessible surface area term (often omitted), S_{Pair} is a pair potential defining the propensity of residues to be found in interaction in given environments and $S_{Rotamer}$ is a probability of finding a specific rotamer. Exactly like for the previous low resolution scores, all the terms can be computed for RNA. The rotamer term and loop optimization are switched off for RNA such as in [28] and in previous CAPRI runs containing RNA [49] for which the RosettaDock all-atom procedure was just used to

Table 1. Leave-one-pdb-out scoring statistics for nine protein-RNA complexes.

PDB code	Enrichment Score		Top10			Top100		# of near native	AUC	
	Default	Roger	Default	Roger	expected	Default	Roger		Default	Roger
1c0a	0.46	5.88	1	10	2.57	11	99	2568	30.41%	91.29%
1k8w	0.59	7.72	1	10	3.92	5	100	3916	35.28%	93.78%
1qtq	0.05	6.36	0	10	2.97	4	100	2971	30.37%	89.09%
2i82	3.45	7.14	6	10	6.91	60	100	6908	40.69%	84.12%
2e9t	0	7.7	0	10	1.69	0	100	1688	19.00%	99.80%
2asb	1.18	5.25	2	10	5.48	3	100	5475	45.69%	92.22%
1ffy	0.02	7.09	0	10	3.12	0	100	3121	42.32%	93.05%
1qf6	0.09	7.3	0	10	1.15	0	99	1154	23.09%	90.18%
2bh2	0.29	7.08	0	10	2.34	0	100	2340	32.26%	88.99%

Enrichment Score, 10 best energy candidates, 100 best energy candidates, number of near-native structures and Area Under the ROC Curve are reported for each native structure both using the non-optimized RosettaDock scoring function (Default) and our optimized scoring function (ROGER).

doi:10.1371/journal.pone.0108928.t001

refine the obtained conformation and RNA parameters were derived from protein data.

Low resolution weights

The low resolution representation for each residue/nucleotide is made of the backbone atoms and one pseudo-atom called centroid to represent the side-chain. For the residues, the location of the centroid is taken from RosettaDock (average over a reference set of PDB structures). For RNA nucleotides, the centroid is taken as the averaged position (See Figure S1). The low resolution scores are computed for RNA on the full PRIDB (more than a thousand structures). They represent counting statistics and are not optimized further.

High resolution scoring weights optimization strategies

We performed the optimization by supervised learning. To ensure an accurate learning phase, the perturbation was split in two categories for learning labelled near-native (Irmsd$<$5 Å) and decoy (Lrsmd$>$8 Å). The assessment was performed using slightly different categories so as to mimic the CAPRI context: near-native (Irmsd$<$5 Å) and non-native (Irmsd$\geq$5 Å). While these rmsd range are certainly not always likely to accurately represent a correct RNA binding mode, especially considering the variability in size of the RNA molecules, they represent a reachable goal not yet attained by the CAPRI community.

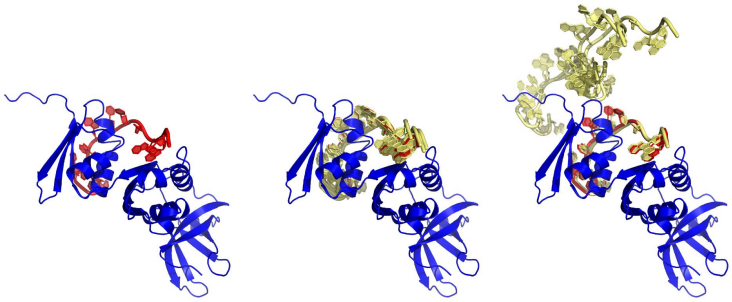
Weights for the all atom scoring function described above were optimized in the [0:1] interval within the ROC-based Genetic LearneR (ROGER) framework using logistic regression and Receiver Operating Characteristic (ROC) based genetic algorithm as previously described for protein-protein docking [40]. The optimization of the Area Under the ROC curve (ROC-AUC) is performed using 100,000 iterations with $\mu = 10$ and $\lambda = 80$.

The first evaluation of the whole scoring procedure is made using cross-validation and a leave-one-*pdb*-out approach. Inspired by the leave-one-out procedure in statistics, we previously used this strategy for machine learning of protein-protein docking scoring functions [40,41,43]. For a specific pdb file, all the native, near-native or decoy conformations, that were generated from this file, are removed from the learning set. The evaluation is then performed for this specific pdb file. The original set learning containing 120 complexes, the whole procedure is repeated 120 times. The set being non-redundant, like cross-validation, this computationally expensive process ensures that the result for a specific pdb file is not biased.

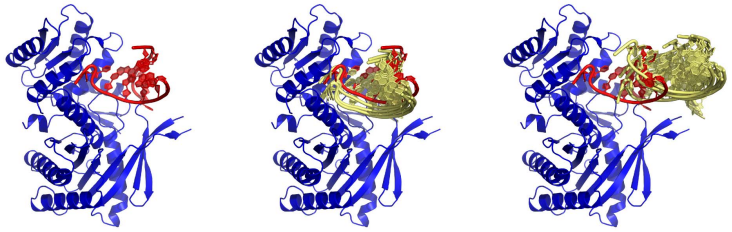
To also avoid biasing the samples towards a category while learning, learning is performed with 30 near-native and 30 decoy structures for each of the 120 pdb file leading to a total size of 7,200 structures for the learning set (3,600×2). Test is performed on the 10,000 candidates of each test pdb file. The global procedure flowchart is available in Figure 1.

Assessment

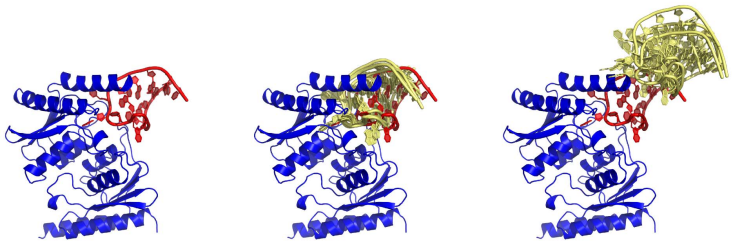
The learning procedure is initially assessed using standard machine learning criteria: analysis of the ROC curve, ROC-AUC in a cross-validation setting and precision for the top 10 structures. CAPRI/Critical Assessment of protein Structure Prediction (CASP) inspired biological criteria are used for the final assessment: Energy vs. Irmsd curve and Enrichment Score (ES). Interface root mean square deviation (Irmsd) is taken from Lensink et al. [50]. We adapted the Enrichment Score from Tsai et al. [51], and also used for RNA structure assessment [52,53]. The enrichment score is defined as: $ES = \frac{|E_{top10\%} \cap R_{top10\%}|}{0.1 \times 0.1 \times N_{candidates}}$ where $E_{top10\%}$ is the top 10% scoring and $R_{top10\%}$ the best 10% rmsd



a) M. tuberculosis transcription elongation protein (NusA) and a 11nt rRNA (2asb)



b) A. fulgidus Piwi protein and a 16nt siRNA duplex (2bgg)



c) S. aureus tRNA Adenosine Deaminase (TadA) and the tRNA-Arg2 anticodon stem-loop (2b3j)

Figure 4. 3D structures and predictions for three protein-RNA complexes (reference set). The protein is shown in blue, the native RNA in red and the RNA candidates in yellow. For each pdb example: (left) native structure, (middle) native structure superposed to the 5 best energy candidates from ROGER score and (right) native structure superposed to the 5 best energy decoys from RosettaDock default score.
doi:10.1371/journal.pone.0108928.g004

structures. By looking at the degree of overlap between the two categories, the enrichment score provides insight on how good the scoring is $ES<1$ corresponds to bad scoring, $ES = 1$corresponds to random scoring and $ES = 10$ is perfect scoring. Even if what can be considered good scoring is not obvious, the comparison of ES values between 1 and 10 provides good information on how well the strategy performs on different targets.

Results and Discussion

Native and near-native configurations are recovered

A data based docking procedure for protein-RNA complexes should first be able to recover the native and close-to-native states for a reference set of complexes. This is assessed by a careful cross-validation setting. In this study we assessed the performance of our learning procedure by plotting Energy vs. Irmsd and checking the enrichment scores of our procedure relatively to the Rosetta CAPRI default. Figure 2 shows detailed results for nine different complexes (the remaining plots can be found in Figure S2). Interestingly, while only one complex (2e9t) shows a funnel in the default Rosetta version, none of the others do. Funnels can be found however on all the optimized scoring function plots that correlate to a high Enrichment Score. While not all complexes in the dataset display such a good conformation selection, the optimized scoring always performs better than the default RosettaDock setting and seems suitable for prediction.

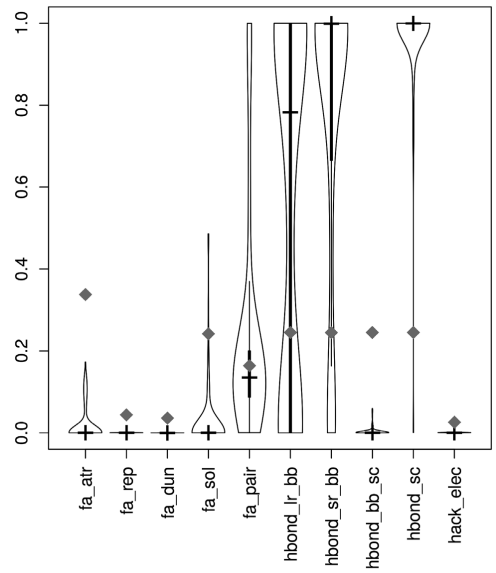


Figure 5. Violinplots of weights for the ROGER optimized RosettaDock scoring function. Default reference weights are shown in grey diamonds. fa_atr and fa_rep represent Lennard-Jones terms (attractive and repulsive). fa_dun corresponds to the internal energy of sidechain rotamers as derived from Dunbrack's; fa_sol is the Lazaridis Karplus solvation energy and fa_pair is the statistics based pair term, known to favour salt bridges for proteins. Remaining are H bond terms for long-range and short-range interactions for both backbone (bb) and side-chain (sc) terms. Last term (hack_elec) represents the empirical electrostatics contribution.
doi:10.1371/journal.pone.0108928.g005

The machine learning procedure was also assessed separately by plotting the ROC curves in the leave-one-*pdb*-out setting. Figure 3 (panels a and b) shows the ROC curves for the optimized and default RosettaDock scoring functions respectively. While the default strategy does not show any discrimination power, our optimized function performs very well. In particular, at the origin, the ROC curve is very steep. This is especially interesting as in the CAPRI challenge only 10 putative conformations can be submitted and in any experimental setting, not more than 100 can be easily tested. Table 1 reports the statistics for the previously mentioned complexes and confirms that a large number of near-native conformations can be found in the top10 and top100 conformations, making the optimized score suitable for prediction (Results on the whole reference set are available in S3). The ROC-AUC often shows larger improvements than the Enrichment Scores as the near-native category for the AUC is defined by a 5 Å threshold (the ES uses the top10% which is generally different than 5 Å).

The strategy was then further evaluated on the Benchmark I and Benchmark II protein-RNA complex structures in a bound setting. Figure 3 (panels c and d) shows the ROC curves for both benchmarks. The ROC-AUC confirms that the optimized scoring function performs well in a prediction setting and is robust to the biological diversity and flexibility encountered in both benchmarks.

Table 2. Scoring results on the unbound test set.

PDB code	Category	Prot/RNA	Enrichment		Top10		Top100		# near native	AUC	
			Default	Roger	Default	Roger	Expected	Default	Roger		Default	Roger
1m5o	U/B		0.43	4.75	0	4	4.79	3	66	4790	24.89%	79.37%
1qtq	U/U		0.00	6.09	0	10	2.798	0	98	2798	24.27%	90.99%
1wpu	U/B		1.60	3.15	9	10	8.723	71	100	8723	46.74%	70.37%
1yvp	U/B		0.93	0.20	6	10	9.362	61	100	9362	28.69%	80.89%
1zbh	U/U		0.88	0.00	5	10	9.834	82	100	9834	12.44%	96.07%
2ad9	U/B		0.00	6.06	0	0	0.001	0	0	1	1.65%	97.78%

Enrichment Score, 10 best energy candidates, 100 best energy candidates, number of near-native structures and Area Under the ROC Curve are reported for each native structure both using the non-optimized RosettaDock scoring function (Default) and our optimized scoring function (ROGER).
doi:10.1371/journal.pone.0108928.t002

Most of the best energy candidates are biologically relevant near-native candidates

The RosettaDock perturbation generation for the conformations ensures that the packing at the interface is relatively correct. Visual inspection shows that the conformations of best energy conformations are relevant from a biological perspective (interface area, contacts, clashes…). Figure 4 shows the 5 best energy candidates are very close to the native structure (bound setting). When various interface cavities are available for the docking (e.g. Figure 4b), the optimized function also clearly selects the right interface despite the atomic contacts being reasonable in both putative cavities. The default RosettaDock scoring function does select reasonably packed conformation but not always the right interface location.

Optimized weights and interface H-bonding network

In a bound setting, for protein-RNA, the relative influence of the parameters shows that the H-bond network is extremely important and must be maintained. Figure 5 shows the weights obtained for the RosettaDock scoring function by optimization. H-bond terms involving the backbone are high at short range but also at long range. Unsurprisingly the H-bonding terms of the side chains are extremely important both for single and double strand RNAs (data not shown). Except for the pair term, most of the other terms have a very small influence. Other than the putative H-bonding network, only the pair terms have some importance. This is in accordance with the previous pair scoring functions developed for protein-RNA docking [28]. The relative importance of the weights however has to be assessed keeping in mind the values of the terms cannot really be normalized in the same range. The Lennard-Jones terms not having influence might be due to the fact that the system is set up on perturbation decoys generated by RosettaDock. By definition these will have a relatively good packing and clashing or too distant conformation will be left out without having to use the scoring function. To ensure the biophysical interpretation of the sign of the weights was compatible with our results, we also tried to optimize the scoring function by allowing the weights in the $[-1;1]$ and in the $[-1;0]$ intervals [54]. This led to much less stable learning procedures and worse results. We also checked whether the structural nature of the RNA molecules (single-, double-strand, tRNA…) made a difference but could not find any remarkable pattern. Score being high-resolution in a bound setting, the atomic contacts are more significant than the overall shape criteria.

Benchmarking bound and unbound docking

The scoring function was then assessed in both bound and unbound (protein and RNA when available) settings. Perturbation runs were performed in a bound setting on the 40 complexes of the benchmarks not in the reference set. Only the 6 pdb files corresponding to median, 1st and 3rd quartile ROC performance were assessed in a full docking run unbound setting (for computational reasons). Results can be found in Table 2 and Table S4. As it was the case in a bound setting where results are consistent with the ones obtained on the reference set with cross-validation, the increase in performance for the unbound setting is also very clear. Results also show that AUC and enrichment score alone are not sufficient to evaluate the procedure and that the E vs. rmsd plots have to be checked as the rmsd distribution among the decoys can vary: while the enrichment score can be poor, the selection can be very good. The E vs. rmsd plots show very sharp funnels (Figures S3 and S4).These may contain two or three very sharp peaks corresponding to small changes in the residue rotamers and/or to the H-bonding network. All peaks do however correspond to native conformations in the CAPRI definition. For some case, the results stay poor: to improve these results flexibility of RNA should be taken into account so as to provide a wide range of small rmsd.

Limits

A current limit of our approach is the way RNA flexibility is handled. Handling RNA flexibility for RNA during docking is a very difficult task [13]. Thus, aside from hydrogen atoms and protein rotamers, flexibility is not well taken into account. This can however be handled by geometric sampling [55]. For small RNA molecules this lack of flexibility handling is a limitation that cannot allow for good results despite a good high-resolution scoring function as it calls for a preliminary sampling experiment. Modelling electrostatics is also a major issue when modelling RNA molecules: solvent and ions are often found at the interface and are still hard to predict [56]. In our reference set, the interaction between the mRNA binding domain of elongation factor SelB from E.coli in complex with SECIS RNA (PDB code 2pjp) is an example where the interface is mediated by sodium ions that our model does not take into account and for which we obtained very poor results (See Figure S5). While our approach could totally be adapted and used for protein-DNA complex prediction, providing the parameters are optimized on a suitable dataset, a similar effect where ions mediate the interaction would be seen. It is also unclear whether the changes and motifs occurring in the DNA double helix for binding could be well captured by this approach. In addition to limited flexibility treatment, this limits the current data based approaches.

Conclusions

Protein-RNA complexes are undoubtedly a real challenge for the design of good docking scoring functions. Using a well curated dataset and a well-designed optimization strategy, we show that we could set up of an efficient protein-docking scoring function that can be used in RosettaDock and that can perform better than the existing option in both bound and unbound settings. While scoring can be improved, the nature of RNA makes the prediction experiment still difficult. Electrostatics plays a large role in RNA interactions and ions have to be modelled. Like ours, the data based approaches are limited by the relatively small number of structures available to take ions into account carefully. RNA flexibility modelling for docking is then the next challenge: while some strategies allow for conformation sampling, selection of one or several putative bound states for large cross-docking experiments are still out of reach for both modelling and computational reasons.

Availability

The source code and files needed to modify RosettaDock 3.4 are available at: http://albios.saclay.inria.fr/rosettadockrna

Supporting Information

Figure S1 Model of a nucleic acid (uracile). The phosphate group and the sugar heavy atoms are depicted in gray: (a) coarse-grained level with the centroid atom in red and (b) full-atom level with the base atoms in blue. The centroid is the geometric center of the heavy atoms.
(TIFF)

Figure S2 Energy vs Irmsd for the whole reference dataset in a leave-one-pdb-out setting.
(PDF)

Figure S3 Energy vs Irmsd for the benchmark set in a bound setting. The 10,000 conformations evaluated for our optimized Rosetta scoring function are shown in black. On each plot, the bottom left panel shows the equivalent non-optimized Rosetta result.
(PDF)

Figure S4 Energy vs Irmsd for the unbound test set. The 10,000 conformations evaluated for our optimized Rosetta scoring function are shown in black. On each plot, the bottom left panel show the equivalent non-optimized Rosetta result.
(PDF)

Figure S5 Structure of the mRNA binding domain of elongation factor SelB from *E.coli* in complex with SECIS RNA (PDB code 2pjp). Mg2+ ions (shown in yellow) are located at the interface and mediate the interaction.
(TIFF)

Table S1 Protein-RNA complexes reference set from the PRIDB. The rightmost column indicates putative redundancy with the docking benchmarks. The *Type* column refers to the structural family of the RNA molecule: single strand RNA (ssRNA), double strand RNA or single-stranded RNA of helical/paired structure (dsRNA) or transfer RNA (tRNA).
(PDF)

Table S2 Protein-RNA complexes for the test set. For each complex the unbound column for protein and RNA reports the PDB code of the unbound structures when available. The difficulty codes are taken from [36,38].
(PDF)

Table S3 Leave-one-pdb-out scoring statistics for the reference dataset. Enrichment Score, 10 best energy candidates, 100 best energy candidates, number of near-native structures and Area Under the ROC Curve are reported for each native structure both using the non-optimized RosettaDock scoring function (Default) and our optimized scoring function (ROGER).
(PDF)

Table S4 Scoring results on the bound benchmark test set. Enrichment Score, 10 best energy candidates, 100 best energy candidates, number of near-native structures and Area Under the ROC Curve are reported for each native structure both using the non-optimized RosettaDock scoring function (Default) and our optimized scoring function (ROGER).
(PDF)

Acknowledgments

The authors thank Sid Chaudhury and Jeff Gray for their help with RosettaDock.

Author Contributions

Conceived and designed the experiments: AGG CF JA JB. Performed the experiments: AGG CF JA JB. Analyzed the data: AGG CF JA JB. Contributed reagents/materials/analysis tools: AGG CF JA JB. Wrote the paper: AGG CF JA JB.

References

1. Cooper TA, Wan L, Dreyfuss G (2009) RNA and disease. Cell 136: 777–793.
2. Clery A, Blatter M, Allain FH (2008) RNA recognition motifs: boring? Not quite. Curr Opin Struct Biol 18: 290–298.
3. Ke A, Doudna JA (2004) Crystallization of RNA and RNA-protein complexes. Methods 34: 408–414.
4. Scott LG, Hennig M (2008) RNA structure determination by NMR. Methods Mol Biol 452: 29–61.
5. Theimer CA, Smith NL, Khanna M (2012) NMR studies of protein-RNA interactions. Methods Mol Biol 831: 197–218.
6. Chen Y, Varani G (2005) Protein families and RNA recognition. FEBS J 272: 2088–2097.
7. Ellis JJ, Broom M, Jones S (2007) Protein-RNA interactions: structural analysis and functional classes. Proteins 66: 903–911.
8. Lipfert J, Doniach S (2007) Small-angle X-ray scattering from RNA, proteins, and protein complexes. Annu Rev Biophys Biomol Struct 36: 307–327.
9. Konig J, Zarnack K, Luscombe NM, Ule J (2011) Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet 13: 77–83.
10. Milek M, Wyler E, Landthaler M (2012) Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. Semin Cell Dev Biol 23: 206–212.
11. Zhou ZH (2008) Towards atomic resolution structural determination by single-particle cryo-electron microscopy. Curr Opin Struct Biol 18: 218–228.
12. Chen Y, Varani G (2013) Engineering RNA-binding proteins for biology. FEBS J 280: 3734–3754.
13. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM (2012) Computational methods for prediction of protein-RNA interactions. J Struct Biol 179: 261–268.
14. Mendez R, Leplae R, De Maria L, Wodak SJ (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 52: 51–67.
15. Vakser IA, Kundrotas P (2008) Predicting 3D structures of protein-protein complexes. Curr Pharm Biotechnol 9: 57–66.
16. Moreira IS, Fernandes PA, Ramos MJ (2010) Protein-protein docking dealing with the unknown. J Comput Chem 31: 317–342.
17. Janin J (2010) Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 6: 2351–2362.
18. de Vries SJ, Melquiond AS, Kastritis PL, Karaca E, Bordogna A, et al. (2010) Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. Proteins 78: 3242–3249.
19. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, et al. (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414: 289–302.
20. Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond AS, et al. (2013) Blind prediction of interfacial water positions in CAPRI. Proteins.
21. Lensink MF, Wodak SJ (2013) Docking, scoring, and affinity prediction in CAPRI. Proteins 81: 2082–2095.
22. Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci U S A 104: 14664–14669.
23. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7: 291–294.
24. Laing C, Schlick T (2010) Computational approaches to 3D modeling of RNA. J Phys Condens Matter 22: 283101.
25. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452: 51–55.
26. Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM (2011) RNA and protein 3D structure modeling: similarities and differences. J Mol Model 17: 2325–2336.
27. Flores SC, Bernauer J, Shin S, Zhou R, Huang X (2012) Multiscale modeling of macromolecular biosystems. Brief Bioinform 13: 395–405.
28. Chen Y, Kortemme T, Robertson T, Baker D, Varani G (2004) A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucleic Acids Res 32: 5147–5162.
29. Huang SY, Zou X (2014) A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucleic Acids Res.
30. Perez-Cano L, Solernou A, Pons C, Fernandez-Recio J (2010) Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac Symp Biocomput: 293–301.
31. Tuszynska I, Bujnicki JM (2011) DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinformatics 12: 348.
32. Zheng S, Robertson TA, Varani G (2007) A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. FEBS J 274: 6378–6391.
33. Li CH, Cao LB, Su JG, Yang YX, Wang CX (2012) A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins 80: 14–24.
34. Setny P, Zacharias M (2011) A coarse-grained force field for Protein-RNA docking. Nucleic Acids Res 39: 9118–9129.
35. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, et al. (2011) PRIDB: a Protein-RNA interface database. Nucleic Acids Res 39: D277–282.

36. Barik A, Nithin C, Manasa P, Bahadur RP (2012) A protein-RNA docking benchmark (I): nonredundant cases. Proteins 80: 1866–1871.
37. Huang SY, Zou X (2013) A nonredundant structure dataset for benchmarking protein-RNA computational docking. J Comput Chem 34: 311–318.
38. Perez-Cano L, Jimenez-Garcia B, Fernandez-Recio J (2012) A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. Proteins 80: 1872–1882.
39. Azé J, Bourquard T, Hamel S, Poupon A, Ritchie D (2011) Using Kendall-τ Meta-Bagging to Improve Protein-Protein Docking Predictions. In: Loog M, Wessels L, Reinders MT, Ridder D, editors. Pattern Recognition in Bioinformatics: Springer Berlin Heidelberg. pp. 284–295.
40. Bernauer J, Aze J, Janin J, Poupon A (2007) A new protein-protein docking scoring function based on interface residue properties. Bioinformatics 23: 555–562.
41. Bernauer J, Poupon A, Aze J, Janin J (2005) A docking analysis of the statistical physics of protein-protein recognition. Phys Biol 2: S17–23.
42. Bordner AJ, Gorin AA (2007) Protein docking using surface matching and supervised machine learning. Proteins 68: 488–502.
43. Bourquard T, Bernauer J, Aze J, Poupon A (2011) A collaborative filtering approach for protein-protein docking scoring functions. PLoS One 6: e18541.
44. Viswanath S, Ravikant DV, Elber R (2013) Improving ranking of models for protein complexes with side chain modeling and atomic potentials. Proteins 81: 592–606.
45. Zhu X, Ericksen SS, Demerdash ON, Mitchell JC (2013) Data-driven models for protein interaction and design. Proteins 81: 2221–2228.
46. Gray JJ (2006) High-resolution protein-protein docking. Curr Opin Struct Biol 16: 183–193.
47. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, et al. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 331: 281–299.
48. Kilambi KP, Pacella MS, Xu J, Labonte JW, Porter JR, et al. (2013) Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20–27. Proteins 81: 2201–2209.
49. Fleishman SJ, Corn JE, Strauch EM, Whitehead TA, Andre I, et al. (2010) Rosetta in CAPRI rounds 13–19. Proteins 78: 3212–3218.
50. Lensink MF, Mendez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 69: 704–718.
51. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, et al. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53: 76–87.
52. Bernauer J, Huang X, Sim AY, Levitt M (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. RNA 17: 1066–1075.
53. Sim AY, Schwander O, Levitt M, Bernauer J (2012) Evaluating mixture models for building RNA knowledge-based potentials. J Bioinform Comput Biol 10: 1241010.
54. Guilhot-Gaudeffroy A, Azé J, Bernauer J, Froidevaux C. Apprentissage de fonctions de tri pour la prédiction d'interactions protéine-ARN; 2014; Rennes. Revue des Nouvelles Technologies de l'Information, RNTI-E-26. pp. 479–484.
55. Fonseca R, Pachov D, Bernauer J, van den Bedem H (2014) Characterizing RNA ensembles from NMR data with kinematic models. Nucleic Acids Res: in press.
56. Philips A, Milanowska K, Lach G, Boniecki M, Rother K, et al. (2012) MetalionRNA: computational predictor of metal-binding sites in RNA structures. Bioinformatics 28: 198–205.

A COLLABORATIVE FILTERING APPROACH FOR PROTEIN-PROTEIN DOCKING SCORING FUNCTIONS

Thomas BOURQUARD, Julie BERNAUER, Jérôme AZÉ, Anne POUPON. In *PLoS One*, 6(4):e18541, 2011.

OPEN ACCESS Freely available online

PLoS one

A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions

Thomas Bourquard[1,2,3], Julie Bernauer[2], Jérôme Azé[1], Anne Poupon[4,5,6]*

1 Bioinformatics Group, INRIA AMIB, Laboratoire de Recherche en Informatique, Université Paris-Sud, Orsay, France, **2** Bioinformatics Group, INRIA AMIB, Laboratoire d'Informatique (LIX), Ecole Polytechnique, Palaiseau, France, **3** INRIA Nancy Grand Est, LORIA, Vandoeuvre-lès-Nancy, France, **4** BIOS Group, INRA, UMR85, Unité Physiologie de la Reproduction et des Comportements, Nouzilly, France, **5** CNRS, UMR6175, Nouzilly, France, **6** Université François Rabelais, Tours, France

Abstract

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate conformations mimicking the complex existing *in vivo* between two proteins, and a scoring function is used to rank them in order to extract a native-like one. We have already shown that using Voronoi constructions and a well chosen set of parameters, an accurate scoring function could be designed and optimized. However to be able to perform large-scale *in silico* exploration of the interactome, a near-native solution has to be found in the ten best-ranked solutions. This cannot yet be guaranteed by any of the existing scoring functions. In this work, we introduce a new procedure for conformation ranking. We previously developed a set of scoring functions where learning was performed using a genetic algorithm. These functions were used to assign a rank to each possible conformation. We now have a refined rank using different classifiers (decision trees, rules and support vector machines) in a collaborative filtering scheme. The scoring function newly obtained is evaluated using 10 fold cross-validation, and compared to the functions obtained using either genetic algorithms or collaborative filtering taken separately. This new approach was successfully applied to the CAPRI scoring ensembles. We show that for 10 targets out of 12, we are able to find a near-native conformation in the 10 best ranked solutions. Moreover, for 6 of them, the near-native conformation selected is of high accuracy. Finally, we show that this function dramatically enriches the 100 best-ranking conformations in near-native structures.

Citation: Bourquard T, Bernauer J, Azé J, Poupon A (2011) A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions. PLoS ONE 6(4): e18541. doi:10.1371/journal.pone.0018541

Editor: Iddo Friedberg, Miami University, United States of America

Received December 6, 2010; **Accepted** March 3, 2011; **Published** April 22, 2011

Copyright: © 2011 Bourquard et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Anne.Poupon@tours.inra.fr

Introduction

Most proteins fulfill their functions through the interaction with other proteins [1]. The interactome appears increasingly complex as the experimental means used for its exploration gain in precision [2]. Although structural genomics specially addressing the question of 3D structure determination of protein-protein complexes have led to great progress, the low stability of most complexes precludes high-resolution structure determination by either crystallography or NMR. 3D structure of complexes are thus poorly represented in the Protein Data Bank (PDB) [3]. The fast and accurate prediction of the assembly from the structures of the individual partners, called protein-protein docking, is therefore of great value [4]. However, available docking procedures technically suitable for large-scale exploration of the proteome have shown their limits [5,6]. Indeed, amongst the easily available methods for such exploration, a near-native solution is found in the 10 best-ranked solutions (top 10) only for 34% of the studied complexes. For biologists, exploring 10 different conformations for experimental validation is already a huge effort. Making this exploration knowing that the prediction will be confirmed only in one case out of three is completely unacceptable. Consequently, large-scale protein-protein docking will be useful for biologists only if a near-native solution can be found in the top 10 in almost all cases (ideally in the top 5 or even the top 3).

A docking procedure consists in two tasks, generally consecutive and largely independent. The first one, called exploration, consists in building a large number of candidates by sampling the different possible orientations of one partner relatively to the other. The second task consists in ranking the candidates using a scoring function in order to extract near-native conformations. To be accurate, scoring functions have to take into account both the geometric complementarity and the physico-chemistry of amino acids in interaction, since they both contribute to the stability of the assembly [7,8].

Modeling multi-component assemblies often involves computationally expensive techniques, and exploring all the solutions is often not feasible. Consequently, we previously introduced a coarse-grained model for protein structure based on the Voronoi tessellation. This model allowed the set up of a method for discriminating between biological and crystallographic dimers [9], and the design of an optimized scoring function for protein-protein docking [10,11]. These results show that this representation retains the main properties of proteins and proteins assemblies 3D structures, making it a precious tool for building fast and accurate scoring methods. We have also explored the possibility to use a power diagram or Laguerre tessellation model, which gives a more realistic representation of the structure. However we have shown that this model does not give better results and increases algorithmic complexity [12].

OPEN ACCESS Freely available online PLoS one

A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions

Thomas Bourquard[1,2,3], Julie Bernauer[2], Jérôme Azé[1], Anne Poupon[4,5,6]*

1 Bioinformatics Group, INRIA AMIB, Laboratoire de Recherche en Informatique, Université Paris-Sud, Orsay, France, **2** Bioinformatics Group, INRIA AMIB, Laboratoire d'Informatique (LIX), École Polytechnique, Palaiseau, France, **3** INRIA Nancy Grand Est, LORIA, Vandoeuvre-lès-Nancy, France, **4** BIOS Group, INRA, UMR85, Unité Physiologie de la Reproduction et des Comportements, Nouzilly, France, **5** CNRS, UMR6175, Nouzilly, France, **6** Université François Rabelais, Tours, France

Abstract

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate conformations mimicking the complex existing *in vivo* between two proteins, and a scoring function is used to rank them in order to extract a native-like one. We have already shown that using Voronoi constructions and a well chosen set of parameters, an accurate scoring function could be designed and optimized. However to be able to perform large-scale *in silico* exploration of the interactome, a near-native solution has to be found in the ten best-ranked solutions. This cannot yet be guaranteed by any of the existing scoring functions. In this work, we introduce a new procedure for conformation ranking. We previously developed a set of scoring functions where learning was performed using a genetic algorithm. These functions were used to assign a rank to each possible conformation. We now have a refined rank using different classifiers (decision trees, rules and support vector machines) in a collaborative filtering scheme. The scoring function newly obtained is evaluated using 10 fold cross-validation, and compared to the functions obtained using either genetic algorithms or collaborative filtering taken separately. This new approach was successfully applied to the CAPRI scoring ensembles. We show that for 10 targets out of 12, we are able to find a near-native conformation in the 10 best ranked solutions. Moreover, for 6 of them, the near-native conformation selected is of high accuracy. Finally, we show that this function dramatically enriches the 100 best-ranking conformations in near-native structures.

Citation: Bourquard T, Bernauer J, Azé J, Poupon A (2011) A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions. PLoS ONE 6(4): e18541. doi:10.1371/journal.pone.0018541

Editor: Iddo Friedberg, Miami University, United States of America

Received December 6, 2010; **Accepted** March 3, 2011; **Published** April 22, 2011

Copyright: © 2011 Bourquard et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Anne.Poupon@tours.inra.fr

Introduction

Most proteins fulfill their functions through the interaction with other proteins [1]. The interactome appears increasingly complex as the experimental means used for its exploration gain in precision [2]. Although structural genomics specially addressing the question of 3D structure determination of protein-protein complexes have led to great progress, the low stability of most complexes precludes high-resolution structure determination by either crystallography or NMR. 3D structure of complexes are thus poorly represented in the Protein Data Bank (PDB) [3]. The fast and accurate prediction of the assembly from the structures of the individual partners, called protein-protein docking, is therefore of great value [4]. However, available docking procedures technically suitable for large-scale exploration of the proteome have shown their limits [5,6]. Indeed, amongst the easily available methods for such exploration, a near-native solution is found in the 10 best-ranked solutions (top 10) only for 34% of the studied complexes. For biologists, exploring 10 different conformations for experimental validation is already a huge effort. Making this exploration knowing that the prediction will be confirmed only in one case out of three is completely unacceptable. Consequently, large-scale protein-protein docking will be useful for biologists only if a near-native solution can be found in the top 10 in almost all cases (ideally in the top 5 or even the top 3).

A docking procedure consists in two tasks, generally consecutive and largely independent. The first one, called exploration, consists in building a large number of candidates by sampling the different possible orientations of one partner relatively to the other. The second task consists in ranking the candidates using a scoring function in order to extract near-native conformations. To be accurate, scoring functions have to take into account both the geometric complementarity and the physico-chemistry of amino acids in interaction, since they both contribute to the stability of the assembly [7,8].

Modeling multi-component assemblies often involves computationally expensive techniques, and exploring all the solutions is often not feasible. Consequently, we previously introduced a coarse-grained model for protein structure based on the Voronoi tessellation. This model allowed the set up of a method for discriminating between biological and crystallographic dimers [9], and the design of an optimized scoring function for protein-protein docking [10,11]. These results show that this representation retains the main properties of proteins and proteins assemblies 3D structures, making it a precious tool for building fast and accurate scoring methods. We have also explored the possibility to use a power diagram or Laguerre tessellation model, which gives a more realistic representation of the structure. However we have shown that this model does not give better results and increases algorithmic complexity [12].

In this study, using the Voronoi representation of protein structure, and an in-lab conformation generation algorithm, we show different ways of optimizing the scoring method based on probabilistic multi-classifiers adaptation and genetic algorithm.

Methods

Structure Representation and Conformation Generation

Like in our previous work [9–12], a coarse-grain model is used to represent the protein structure. We define a single node for each residue (the geometric center of side chain, including C_α), the Delaunay triangulation (dual of the Voronoi diagram) of each partner is then computed using CGAL [13] and the Voronoi tessellation is built. The generation of candidate conformations is performed as follows. For each node, a pseudo-normal vector is built by summing the vectors linking this node to its neighbors. In non-convex regions, this vector might point towards the interior of the protein. In this case the opposite vector is taken. Depending on the amino acid type, the length of this vector is made equal to the radius of a sphere whose volume is equal to the average volume occupied by this type of amino acid. This mean volume is taken from Pontius *et al.* [14]. For each possible pair of vectors (one in each partner), one of the vectors is translated so as to bring its extremity on the extremity of the first vector (step 1 on Figure 1). The second partner is then rotated so as to oppose the two vectors (step 2 on Figure 1). The second partner is then rotated around this new axis (step 3 on Figure 1), and a conformation of the complex is build every 5° rotation.

Although not all degrees of freedom are considered (the two normal vectors are always aligned in our method, but we could have considered varying the angle between them), we obtain a near-native conformation for all the complexes in the learning set.

Learning set

Our positive examples set is composed of native structures. We complemented our previous set [12] with the reference set from [15]. This set contains 211 bound-unbound and unbound-unbound complexes (complexes for which the 3D structure of at least one partner is known). SCOP [16] was used to remove redundancy (for two complexes AB and CD, if A and C belong to the same SCOP family, and B and D also belong to the same family, the complex is eliminated).

Negative examples (decoys, or non-native conformations) were generated by applying the previously described generation method to each complex of our native structures set. Only conformations having a minimal interface area of 400 Å^2 and a root mean square deviation (RMSD) relative to the native conformation higher than 10 Å were retained. Within this ensemble, 15 non-native conformations were chosen for each native conformations, resulting in 2980 negative examples in the learning set. As observed in our previous studies, missing values are a serious issue for scoring function optimization. All the non-native conformations presenting a too high number of missing values were removed. This number was taken to be twice the number of missing values in the corresponding native structure. 20 such non-native conformations for each native structure were randomly chosen from the initial decoys set.

Training Parameters

The coarse-grained Voronoi tessellation allows simple description of the protein-protein interface. 96 training attributes [12] based on properties of residues and pairs present at the interface have been used. For pair parameters, residues were binned in six categories: hydrophobic (ILVM), aromatics (FYW), small (AGSTCP), polar (NQ), positive (HKR) and negative (DE). These categories are also used to compute the 12 last parameters. Retained parameters are:

- c_1: The Voronoi interface area.
- c_2: The total number of interface residues.
- c_3 to c_{22}: The fraction of each type of interface residues.
- c_{23} to c_{42}: The mean volume of Voronoi cells for the interface residues.
- c_{43} to c_{63}: The fraction of pairs of interface residues.
- c_{64} to c_{84}: The mean node-node distance in pairs of interface residues.

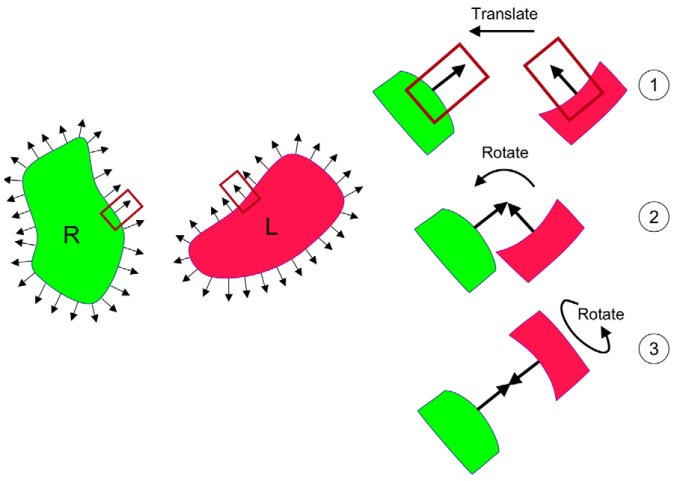


Figure 1. Conformation generation method.
doi:10.1371/journal.pone.0018541.g001

- c_{85} to c_{90}: The fraction of interface residues for each category.
- c_{91} to c_{96}: The mean volume of Voronoi cells for the interface residues for each category.

All parameters were computed on the complete interface, defined as all the residues having at least one neighbor belonging to the second partner, including residues in contact with solvent molecules.

Genetic algorithm

Using previously defined training attributes, genetic algorithms are used to find family of functions that optimize the ROC (Receiver Operating Characteristics) criterion. We used a $\lambda+\mu$ scheme, with $\lambda=10$ parents and $\mu=80$ children, and a maximum of 500 generations. We used a classical cross-over and auto-adaptative mutations. The ROC criterion is commonly used to evaluate the performance of learning procedures by measuring the area under the ROC curve (AUC). The ROC curve is obtained by plotting the proportion of true positives against false positives.

The scoring functions used in this work have the form:

$$S(conf)=\sum_i w_i|x_i(conf)-c_i| \quad (1)$$

where x_i is the value of parameter i and w_i and c_i are the weights and centering values respectively for parameter i. w_i and c_i are optimized by the learning procedure. Learning was performed in a 10-fold cross-validation setting. Ten groups of models were randomly chosen, each excluding 10% of the training set. Learning was repeated n times for each training subset. Consequently, each conformation is evaluated using n different scoring functions, and for final ranking, the sum of the ranks obtained by each function is used.

As described in the Results section, the number of functions n used in the computation of the final rank might have an impact on the quality of the global ranking.

Collaborative filtering methods

Several previous studies have shown the strength of Collaborative Filtering (CF) techniques in Information Retrieval problems [17] to increase the accuracy of the prediction rate. In a common CF recommender system, there is a list of m users, $U_1,U_2,\ldots,U_m$ and a list of p items, $I_1,I_2,\ldots,I_p$ and each user gives a mark to each object. This mark can also be inferred from the user's behaviour. The final mark of each object is then defined by the ensemble of marks received from each user.

In the present work, a classifier is a user, and conformations are the items. Each classifier (user) assigns to each item (conformation) a binary label (or mark): *native* (+) or *non native* (−).

12 classifiers have been trained on the learning set (see "Results and Discussion"), deriving from four different methods: decision trees, rules, logistic regression and SVM (Support Vector Machine). Most optimizations were done using Weka [18]. The *SVMlight* [19] software was used for SVM computations.

In a first approach, we have used a default voting system: the conformations are ranked according to the number of + marks they have received. Since we have 12 classifiers, this determines 13 different categories: 12+, 11+, …, 0+.

Because 13 categories is far from enough to efficiently ranks a very large number of conformations, we have also used a second approach using an amplification average voting system. In this system, the votes of each classifier are weighted by the precision. Consequently, the + vote of each classifier is different from the + vote of a different classifier. This results in 2^{12} categories. The categories are ordered according to:

$$S_{CF}=\frac{\exp^{S-}}{\exp^{S+}} \quad (2)$$

Where S_+ (respectively S_-) is the sum of the precisions of the classifiers that have voted + (respectively −) for conformations of this category. This score is assigned to each conformation of the considered category.

$$S_+=\sum_{i=1}^{n}11(vote_i=+)\times pr_i \quad S_-=\sum_{i=1}^{n}11(vote_i=-)\times pr_i \quad (3)$$

Where $vote_i$ represents the vote of the i^{th} classifier and pr_i represents its precision. In a unweighted approach, pr_i is set to 1 for all the classifiers.

Finally, the CF and GA methods have been coupled. For each conformation evaluated with at least one positive vote ($S_+>0$), the score $S_{CF-GA}(C)$ of a given conformation C is the product of the rank obtained by C in the GA, and $S_{CF}(C)$. For conformations receiving only negative votes, the score $S_{CF-GA}(C)$ is set to be maximal. The evaluated conformations are then re-ranked according to this score (in decreasing order). It should be noted that scores (and consequently ranks) obtained through this method are not necessarily unique. To measure the number of possible ranks for each method, taking into account the number of examples to classify, we will use the granularity as defined in equation 4.

$$granularity(S)=\frac{\text{number of ranks}}{\text{number of examples}} \quad (4)$$

Where S is a set of evaluated conformations.

Evaluation of learning accuracy

The most commonly used criterion for evaluating the efficiency of a learning procedure is the Area Under the ROC curve (ROC AUC). The ROC curve is obtained by plotting the proportion of true positives against the proportion of false positives. A perfect learning should give an AUC of 1 (all the true positives are found before any of the negatives), whereas a random function has an AUC of 0.5 (each prediction has probabilities of 0.5 to be correct or incorrect).

To measure the performances of the different scoring functions we use precision, recall and accuracy using TP,FP,TN and FN as in the confusion matrix (see Table 1). We will also use false negative rate (FNR) and true negative rate (TNR).

Table 1. Confusion matrix.

		solution	
		+	-
prediction	+	TP	FP
	-	FN	TN

TP: true positives, FP: false positives, FN: false negatives, TN: true negatives.
doi:10.1371/journal.pone.0018541.t001

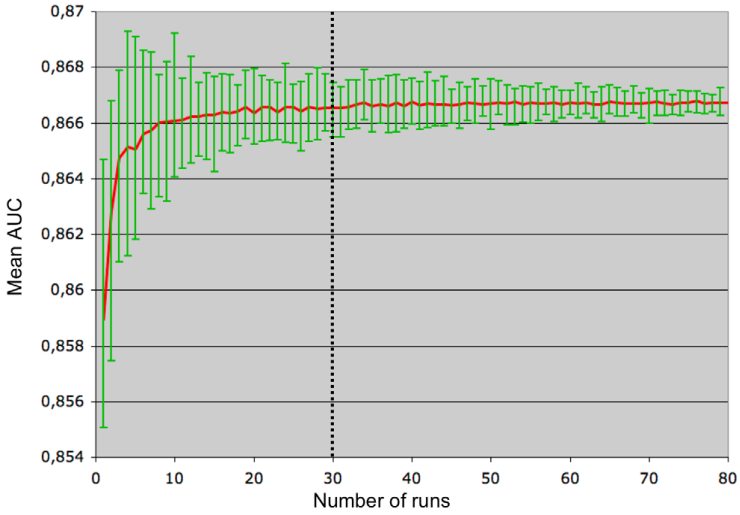


Figure 2. Genetic Algorithm performance as a function of the number of runs. For each number of runs *n*, the measure of the AUC has been repeated 50 times using a 10-fold cross-validation protocol. Average, minimum and maximum values are plotted.
doi:10.1371/journal.pone.0018541.g002

These values can be computed as:

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} \quad Accuracy = \frac{TP+TN}{total}$$

$$FNR = \frac{FN}{TN+FN} \quad TNR = \frac{TN}{TN+FN}$$

CAPRI Experiments

To evaluate the accuracy of our CF-GA scoring procedure, we developed two protocols based on targets 22 to 40 of the CAPRI (Critical Assessment of PRedicted Interaction) experiment. CAPRI is a blind prediction experiment designed to test docking and scoring procedures [20,21]. In the scoring experiment, a large set of models submitted by the docking predictors is made available to the community to test scoring functions independently of conformation generation algorithms.

Four targets where eliminated for different reasons:

- The structure of target 31 has not yet been released, making it impossible to evaluate the obtained results.
- The native 3D structure of target 30 is still a vexed question [22].
- Targets 33 and 34 are protein-RNA complexes and our scoring method is not adapted to this problem yet.

For each target, the scoring ensemble was evaluated using GA, CF and CF-GA methods.

For reasons exposed in "Results and Discussion", candidate conformations where evaluated according to two different sets of criteria.

In the *fnat* evaluation, we use only the *fnat* criterion, which is the fraction of native contacts (the fraction of contacts between the two partners in the evaluated conformation that do exist in the native structure). Four quality classes can be defined:

- High ($fnat \geq 0.5$),
- Medium ($0.3 \leq fnat < 0.5$),
- Acceptable ($0.1 \leq fnat < 0.3$),
- Incorrect ($fnat < 0.1$)

CAPRI evaluation [21,23] also uses two other criteria: the I_{RMSD} (*RMSD* between prediction and native structure computed

Table 2. Precision, recall and accuracy of the retained classifiers.

Classifier	Precision	Recall	Accuracy
SVM-RBF	1.000	0.606	0.975
PART-M2	0.777	0.737	0.970
J48-M2	0.704	0.697	0.963
JRIP-N10	0.665	0.520	0.954
JRIP-N2	0.65	0.591	0.955
PART-M10	0.645	0.561	0.953
PART-M5	0.642	0.626	0.955
SVM-Q2	0.64	0.727	0.958
J48-M5	0.630	0.586	0.953
JRIP-N5	0.615	0.566	0.951
Logistic	0.607	0.414	0.947
J48-M10	0.564	0.465	0.944

Classifiers have been trained on the same learning set as the genetic algorithm, in 10-fold cross-validation.
doi:10.1371/journal.pone.0018541.t002

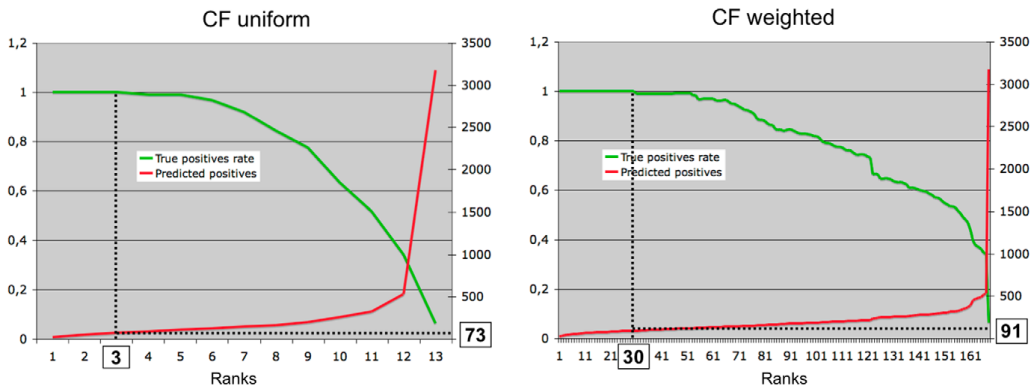


Figure 3. True positive rate for uniform and weighted collaborative filtering. The true positive rate (green) and the total number of positives are plotted for uniform (left) and weighted (right) collaborative filtering, as a function of the category. The vertical and horizontal dotted lines give the category, and the corresponding number of conformations predicted as positives, above which the true positive rate decreases under 1.
doi:10.1371/journal.pone.0018541.g003

only on interface atoms) and L_{RMSD} (*RMSD* computed on all the atoms of the smallest protein, the largest protein of prediction and native structure being superimposed). Again four quality classes are defined:

- High: (*fnat* $\geq$0.5) and ($I_{RMSD} \leq$1 or $L_{RMSD} \leq$1)
- Medium: [(0.3$\leq$ *fnat* $<$0.5) and ($I_{RMSD} \leq$2.0 or $L_{RMSD} \leq$5.0)]**or** [(*fnat* $>$0.5 and $I_{RMSD}>$1.0 or $I_{RMSD}>$1.0)]
- Acceptable: [(0.1$\leq$ *fnat* $<$0.3) and ($I_{RMSD} \leq$4.0 or $L_{RMSD} \leq$10.0)] **or** [*fnat* $>$0.3 and ($L_{RMSD} >$5.0 or $I_{RMSD} >$2.0)]
- Incorrect.

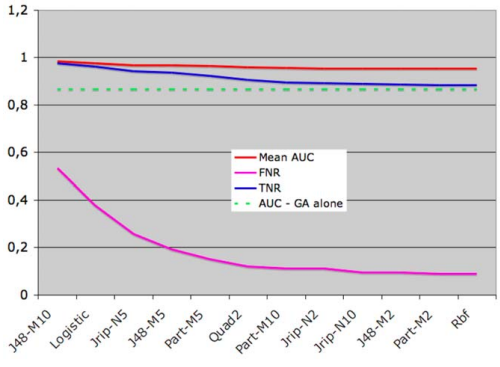


Figure 4. Evolution of AUC, true negative rate (TNR) and false negative rate (FNR) in CF-GA using increasing number of classifiers. Classifiers were added to the collaborative filter, using averaged voting, in increasing precision order. For example, abscissa "JRIP-N5" corresponds to the CF-GA method using J48-M10, Logistic and JRIP-N5 classifiers. Green and red curves correspond to AUC of GA method (which is constant since it doesn't use the classifiers, shown for comparison) and CF-GA method respectively. TNR: true negative rate; FNR: false negative rate.
doi:10.1371/journal.pone.0018541.g004

Results and Discussion

In our previous work, we have used different flavors of genetic algorithm (GA) optimization to obtain scoring functions for protein-protein docking. Since we have reached the limits of the precision that can be obtained with GA alone, we combined the GA-based scoring function with scoring functions built using four other learning algorithms:

- Logistic regression (LR) [24];
- Support Vector Machines [25], using either radial-based function (RBF), linear kernel (LK), polynomial kernel (PK) or 2 and 4 quadratic kernels (QK2 and QK4);
- Decision trees, using the C4.5 learner [26] and, J48, its implementation in Weka [18], using 2, 5 and 10 as minimum numbers of examples required to build a leaf (classifiers J48-M2, J48-M5 and J48-M10 respectively);
- Two-rules learners, using two different implementations (JRIP [27] and PART [28]), using again 2, 5 and 10 as minimum numbers of examples required to build a rule (classifiers JRIP-M2, JRIP-M5, JRIP-M10, PART-M2, PART-M5 and PART-M10).

Here we show how these 15 classifiers can be combined, in a collaborative scheme and with the genetic algorithm procedure.

Predictions obtained with the genetic algorithm procedure

The sensitivity (ability to discriminate true positives from false positives, also called recall) of the genetic algorithm (GA) has been evaluated using the ROC criterion. Since GA is a heuristic, optimization must be repeated. The number of repetitions necessary for obtaining a reliable result largely depends on the specificity of the problem. To determine the number of repetitions needed in our case, we have plotted the area under the ROC curve (AUC) as a function of the number of runs. For each value of the number of runs n, the experiment has been repeated 50 times in 10-fold cross-validation. This allows to compute, for each n, the mean value and the variance of the AUC. As can be seen on

Figure 2, the AUC reaches a plateau (0.866, the difference with AUC with 1 repetition is significant) when the number of runs is higher than 30, and the variance is then less than 10^{-9}.

Based on this result, GA runs will be repeated 30 times in the following.

Classifiers

The precision, recall and accuracy have been computed for each of the chosen classifiers. Three of them (LK, PK and QK4) have precision lower than 0.5, meaning that their predictions are even worse than random. Consequently these three classifiers were discarded. The values obtained for the remaining 12 classifiers are given in Table 2. The results obtained show that the different classifiers have very good accuracies. This result is largely due to the fact that the number of positive examples is about ten times lower than the number of negative examples. Consequently, a classifier which predicts all candidates as negative would have an accuracy of 0.9, but a precision of 0 and a recall of 0 for the positive examples. SVM-RBF has a precision of 1, showing that this classifier does not give any false positives, however, the recall is only 0.606, which means that it misses 40% of the positives. Apart from SVM-RBF, all classifiers have relatively low precision and recall.

The different classifiers have first been combined using an uniform collaborative filtering scheme. In this configuration, each classifier votes for each conformation. Its vote can be *positive* or *negative*. Consequently, a given conformation can receive from 12 to 0 positive votes. Thus, 13 different groups are created, which can be ordered by decreasing numbers of positive votes. When applied to the learning set in 10-fold cross-validation, the three best categories (13, 12, and 11 positive votes) contain only native conformations (Figure 3). This means that the 73 best ranked conformations are true positives.

However, when considering thousands of conformations, 13 categories are not sufficient for efficiently ranking, since many non-equivalent conformations have the same rank (granularity 0.05). To address this problem, we have used an averaged voting protocol (weighted collaborative filtering). Each classifier still votes "positive" or "negative" for each conformation, but the vote is weighted by the precision of the classifier. Since the 12 precisions are all different, the votes of the different classifiers are not equivalent anymore, which results in $2^{12} = 4096$ different categories. Consequently, conformations can be classified in 4096 categories, which can be ranked as a function of their positive score ($score+$, see Methods). Again, the best categories contain

Table 3. Evaluation of the CF-GA method.

Target	With RMSD filtering									No RMSD filtering					
	GA			CF			CF-GA			CF – GA			CF then GA		
	best	N	R	best	N	R	best	N	R	best	N	R	best	N	R
fnat criterion only															
T22	***	3	4	***	2	3	***	2	4	***	2	4	***	2	4
T23	**	10	1	***	9	1	***	9	1	***	10	1	***	10	1
T25	*	1	5	***	2	2	***	2	2	***	4	2	***	4	2
T26	*	2	6	**	5	1	**	5	1	**	6	1	**	7	1
T27	***	3	6	***	5	1	***	4	1	***	5	1	***	6	1
T29	**	2	5	**	4	1	**	2	2	***	6	2	***	6	2
T32	*	2	3	***	2	3	***	3	10	*	2	6	-	0	16
T35	-	0	-	*	1	2	*	1	1	*	1	1	-	0	-
T37	*	1	1	*	1	5	*	1	3	*	1	3	-	0	16
T39	-	0	107	-	0	48	-	0	99	-	0	205	-	0	-
T40A	***	2	2	***	2	1	***	3	1	***	4	1	***	5	1
T40B	-	0	13	-	0	13	-	0	16	-	0	48	-	0	40
All *CAPRI* criteria															
T25	-	0	13	*	1	5	*	1	5	*	2	6	*	2	6
T29	*	1	8	*	2	1	*	2	2	**	6	2	**	6	1
T32	-	0	36	**	1	3	**	1	10	-	0	18	-	0	16
T35	-	0	NA	-	0	NA	-	0	NA	-	0	167	-	0	-
T37	-	0	14	-	0	13	-	0	17	-	0	18	-	0	18
T39	-	0	NA	-	0	NA	-	0	NA	-	0	652	-	0	-
T40A	**	1	3	*	1	1	*	1	1	***	5	1	**	4	1
T40B	-	0	28	-	0	13	-	0	16	-	0	48	-	0	33

Best quality conformation found in the top 10 ranked solutions from target 22 to target 40 for genetic algorithm (GA), collaborative filtering (CF) and combination of the previous two (CF-GA) methods, with RMSD filtering. Same results are given for the CF-GA method without RMSD filtering, and for the *CF then GA* method. N: Numbers of acceptable or better solutions in the top 10; R: rank of the first acceptable or better solution for each target. Numbers of high quality (***), medium quality (**), acceptable (*) and incorrect conformations in each ensemble and for each method when using RMSD filtering are given in Table 4.

doi:10.1371/journal.pone.0018541.t003

Table 4. Total number of conformations is each category before and after RMSD filtering, using *fnat* criterion.

	GA				CF				CF-GA				Without RMSD filtering			
Target	***	**	*	I	***	**	*	I	***	**	*	I	***	**	*	I
fnat criterion only																
T22	12	14	6	40	12	13	7	40	12	13	8	36	32	29	98	113
T23	4	10	23	12	4	10	23	12	2	10	24	13	24	36	189	37
T25	1	0	2	42	1	0	2	38	1	0	2	2	13	2	13	88
T26	12	4	21	165	10	8	21	166	9	7	20	167	537	33	106	641
T27	29	39	39	186	34	36	40	183	32	37	40	192	399	131	106	654
T29	0	3	0	67	0	2	12	60	1	1	16	58	62	78	59	163
T32	1	0	8	171	1	0	7	172	0	1	9	172	1	11	184	376
T35	0	0	3	168	0	0	2	157	0	0	1	159	0	0	8	491
T37	2	2	23	339	1	3	24	337	4	0	21	347	45	34	119	1497
T39	1	1	5	325	0	1	6	324	0	1	5	321	4	1	20	1275
T40A	1	0	7	247	1	0	7	244	1	0	5	248	366	36	119	1439
T40B	2	1	2	249	2	0	1	249	2	1	0	186	165	22	72	1701
All *CAPRI* criteria																
T25	0	0	1	44	0	0	1	40	0	1	0	44	0	6	14	96
T29	0	0	3	67	0	0	2	72	0	1	2	73	1	76	66	219
T32	0	1	0	179	0	1	0	180	0	0	1	181	0	3	12	557
T35	0	0	0	161	0	0	0	159	0	0	0	160	0	0	3	496
T37	0	3	3	360	0	2	3	360	0	3	2	367	11	46	42	1596
T39	0	1	0	371	0	0	0	331	0	0	0	327	0	3	1	1296
T40A	0	0	2	252	0	0	2	250	0	1	0	253	90	151	150	1569
T40B	2	0	0	252	2	0	0	250	1	1	0	187	102	54	30	1774

Numbers of high quality (***), medium quality (**), acceptable (*) and incorrect (I) conformations in the CAPRI scoring ensembles for each target using *fnat* criterion only or all CAPRI criteria, with and without RMSD filtering.

doi:10.1371/journal.pone.0018541.t004

only true positives (see Figure 3). The results are even better than those obtained with uniform CF, since the first non-native conformation belongs to category 31, which means that the 91 best ranked conformations are natives.

However, when considering millions of conformations, 4096 categories are still not sufficient (granularity 0.15). For example, when using the weighted-CF method on the learning set, the best category (only positive votes) contains 24 conformations. Consequently, this method cannot be used for ranking large data sets.

Combination of collaborative filtering and genetic algorithm

Since CF efficiently eliminates non-native conformations, we have used CF to weight the GA score (see Methods). This is what we call the collaborative filtering - genetic algorithm (CF-GA) method. The averaged voting configuration was used, and the CF-GA score is obtained by multiplying the GA score by the ratio of the exponential of positive and negative CF scores. Consequently, the score of conformations classified as negatives by a majority of classifiers have very low CF-GA scores. Figure 4 shows the evolutions of AUC, true negative rate (TNR) and false negative rate (FNR) as we add more classifiers in the CF (in increasing precision order).

Another way of combining the two methods is to: first classify the candidate conformations using the CF, retain only the candidates of the best classes, then use the GA to rank them. To evaluate this approach, we retained all the candidate conformations which rank was lower than N ($N = (10,20,...,100)$ have been tested). These were then ranked using GA. The best results have been obtained with $N = 100$, but this method proved less efficient than the CF-GA (see *CF then GA* in Table 3).

Using the 12 classifiers, the AUC is 0.98, but more importantly, the FNR is only 0.09, meaning that more than 90% of the conformations classified as natives are indeed natives. Unlike

Table 5. Enrichment in acceptable or better solutions.

	CF-GA				**CF**				**GA**			
Target	*fnat*		*capri*		*fnat*		*capri*		*fnat*		*capri*	
	E⋆	**E⋆⋆**	**E⋆**	**E⋆⋆**	**E⋆**	**E⋆⋆**	**E⋆**	**E⋆⋆**	**E⋆**	**E⋆⋆**	**E⋆**	**E⋆⋆**
20% best ranked conformations												
T22	*0.6*	0.74	-	-	*0.45*	*0.58*	-	-	*0.56*	0.74	-	-
T23	1.19	1.75	-	-	1.19	1.75	-	-	1.23	1.63	-	-
T25	3.33	5	3	NA	3.04	4.56	4.56	NA	1.67	*0*	*0*	*0*
T26	1.5	1.26	-	-	1.58	1.71	-	-	1.41	0.95	-	-
T27	1.1	1.22	-	-	1.07	1.12	-	-	1.1	1.1	-	-
T29	6.67	5	5	NA	1.89	2.64	5.29	NA	1.21	2.71	1.81	*0*
T32	2.78	5	5	5	2.5	5	5.03	5.03	2.53	5.06	5.06	NA
T35	1.78	NA	NA	NA	2.48	NA	NA	NA	*0*	NA	NA	NA
T37	1.86	2.51	3.34	3.34	1.25	2.5	3	2.5	2.04	3.82	4.08	3.4
T39	*0*	*0*	*0*	*0*	0.72	*0*	NA	NA	*0*	*0*	NA	NA
T40A	3.75	5	4.98	NA	3.71	4.94	4.94	NA	3.32	4.98	4.98	4.98
T40B	1.99	3.32	4.98	4.98	3.29	4.94	4.94	4.94	3.71	3.71	3.71	1.85
Average	**2.21**	**2.8**	**4.04**	**3.33**	**1.93**	**2.7**	**4.62**	**4.16**	**1.56**	**2.24**	**3.27**	**2.05**
20% worst ranked conformations												
T22	1.65	1.85	-	-	1.2	1.34	-	-	1.53	1.66	-	-
T23	*0.66*	0.7	-	-	*0.66*	0.7	-	-	1.09	1.23	-	-
T25	*0*	*0*	*0*	NA	1.52	*0*	*0*	NA	*0*	*0*	*0*	*0*
T26	0.82	0.63	-	-	0.92	0.57	-	-	1.13	1.9	-	-
T27	0.91	0.79	-	-	0.93	0.91	-	-	0.92	0.73	-	-
T29	3.33	*0*	*0*	NA	*0.38*	*0*	*0*	NA	1.51	*0*	*0*	*0*
T32	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	NA
T35	1.78	NA	*0*	NA	*0*	NA	*0*	NA	*0*	NA	*0*	NA
T37	0.56	*0*	*0*	*0*	0.54	*0*	*0*	*0*	*0*	*0*	*0*	*0*
T39	1.44	*0*	*0*	*0*	2.15	*0*	*0*	NA	*0*	*0*	*0*	NA
T40A	*0*	*0*	*0*	NA	*0*	*0*	*0*	NA	*0*	*0*	*0*	*0*
T40B	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Average	**0.93**	**0.36**	**0**	**0**	**0.69**	**0.32**	**0**	**0**	**0.51**	**0.5**	**0**	**0**

The enrichment in acceptable or better conformations (E⋆) is computed as the proportion of such conformations in the 20% best ranked conformations (respectively worst ranked conformations) divided by the proportion of such conformations is the complete set. Same computation for medium quality or better conformations (E⋆⋆). These enrichments are computed using either *fnat* or CAPRI criteria (*capri*), and for the three methods (GA: genetic algorithm, CF: collaborative filtering, CF-GA: hybrid method). Values in italic are not statistically significant.

doi:10.1371/journal.pone.0018541.t005

Table 6. Best conformation present in the top 10 for different scoring groups.

Groups	T22	T23	T25	T26	T27	T29	T32	T35	T37	T39	T40A	T40B
C Wang	0	0	**	0	*	0	0	*	**	0	***	**
A.M.J.J Bonvin	*	0	*	-	*	**	0	0	*	0	***	**
H. Wolfson	-	-	**	**	*	0	0	0	*	0	*	0
P. A. Bates	-	-	-	-	*	**	0	0	***	0	***	0
Z. Weng	-	-	-	**	*	**	0	0	***	0	***	0
J. F.-Recio	-	-	**	-	*	***	0	0	0	0	0	0
X. Zou	-	-	-	-	*	-	0	0	***	0	***	***
T. Haliloglu	-	-	-	-	-	-	-	-	**	0	***	**
C. J. Camacho	-	-	-	-	**	**	-	-	-	-	***	***
M. Takeda-Shitaka	-	-	-	0	0	0	0	0	-	-	***	**
I. Vakser	-	-	-	-	-	-	**	0	0	0	-	-

Table 6. Cont.

Groups	T22	T23	T25	T26	T27	T29	T32	T35	T37	T39	T40A	T40B
CF-GA Method	⋆⋆⋆[a]	⋆⋆⋆[a]		⋆⋆[a]	⋆⋆⋆[a]			0	0	0		0
			⋆			⋆⋆	⋆⋆⋆				⋆	

0: no acceptable or better solution found, -: group has not participated,
[a]: *fnat* evaluation.
doi:10.1371/journal.pone.0018541.t006

collaborative filtering (CF), the GA method gives unique ranks for all conformations (granularity 1). Using the CF-GA method, the global granularity is lower, mostly because conformations classified as non-natives by a majority of classifiers have very few different, but very high, ranks. However, the scores obtained by the 100 best ranked conformations are almost always unique (granularity 0.995), which allows an efficient sorting of the best conformations.

Finally, our tests have shown that similar conformations have a tendency to have very close ranks. To obtain as much diversity as possible in the best ranked solutions, we removed this redundancy using the RMSD between the conformations. A conformation is kept only if its RMSD with better ranked conformations is higher than 5 Å.

Analysis of the most informative parameters in CF and GA allows to better understand the complementarity of the two methods. Indeed, whereas in GA the most informative parameters measure properties of individual residues, CF relies mostly on parameters relative to contacts at the interface. Interestingly, the distance between small amino acids (AGSTCP) appears as the most discriminating parameters in 9 of the 10 analysed filters (the two SVM filters have been excluded). 5 other distances appear in the 10 most discriminating parameters for CF: Hydrophobic-Small, Polar-Positive, Hydrophobic-Negative, Negative-Negative and Polar-Small. The remaining 4 parameters are frequencies of pairs: Hydrophobic-Negative, Polar-Positive, Hydrophobic-Hydrophobic and Polar-Negative. Among the 10 most informative parameters in GA, 7 are relative to individual residues: volumes of R, E, K, P and I; and frequencies of K and 2. The surface of the interface appears in 4th position, and only 2 parameters are relative to contacts at the interface: frequency of Hydrophobic-Polar pairs and distances between Hydrophobic amino acids in contact.

Ranking of CAPRI ensembles

The CF-GA ranking was applied to CAPRI targets, which were excluded from the learning set. Since no acceptable or better solutions was present in the scoring ensembles for targets 24, 36 and 38, these targets were removed of the analysis.

In a first evaluation, we have used only the *fnat* (fraction of native contacts, see Methods) as a quality measure for all structures in the different scoring ensembles. As explained in the Methods section, CAPRI evaluators do consider the *fnat* criterion, but also I_{RMSD} and L_{RMSD} which are different and complementary measures of the distance between the proposed conformation and the native structure. We were unable to reproduce faithfully these measures since they require manual modifications of both the proposed conformation and the native structure (see Methods). Only for targets T25, T29, T32, T35, T37, T39 and T40 were these measures available from the CAPRI website. Consequently, although the *fnat* indicator is less stringent than the criteria used by CAPRI evaluators, all targets have been analysed using solely the *fnat* criterion. In parallel, for those targets for which they are available, an evaluation using all CAPRI criteria has been conducted.

We first evaluated the ability of our scoring method to find the native structure within the scoring ensemble. For each target, the native structure was introduced in the scoring ensemble. We were able to rank the native solution in the top 10 for 5 out of 12 targets, and in the top 100 for 9 out of 12 targets.

Our next test was to rank the conformations in the CAPRI ensembles, and count the number of acceptable or better solutions in the top 10. Table 3 shows the results obtained using GA, CF and CF-GA. Numbers of high quality (⋆⋆⋆), medium quality (⋆⋆), acceptable (⋆) and incorrect conformations in each ensemble and for each method when using RMSD Filtering are given in Table 4.

As can be seen in Table 3, CF-GA is able to rank at least one acceptable or better solution in the top 10 for 10 out of 12 targets. The rank of the first acceptable or better solution is even lower than 4 for 9 targets, and medium quality or better for 8 targets (it should be noted however that for target 35 only acceptable or incorrect conformations were present in the ensemble). When considering all of the CAPRI criteria, CF-GA ranks acceptable or better solutions in the top 10 for 4 out of 8 targets. Interestingly, there seems to be no correlation between our ability to rank the native solution in the top 10 and our ability to ranked an acceptable or better solution in the top 10. Indeed, for targets T22, T26, T27, T29 and T40_A, the native structure is not ranked in the top 10 (even not in the top 100 for T27 and T29), but acceptable or better conformations are found.

CF and CF-GA give very similar results. The best quality conformations and numbers of acceptable or better solutions found in the top 10 are equivalent. However, the average rank of the first acceptable conformation is lower for CF than for CF-GA (3 *vs.* 3.81; target 39 was excluded from this computation since we considered that the ranks obtained were too high to be significant).

When not using RMSD filtering, the use of the *fnat* criterion doesn't affect CF-GA global performance. However, using all CAPRI criteria, CF-GA ranks an acceptable or better conformation in the top 10 for only 3 targets out of 8. For target 32, the high quality solution that is found at rank 10 with RMSD filtering, appears at rank 18 without RMSD filtering. More generally, results in Table 3 also show that using RMSD filtering decreases the mean rank of the first acceptable or better solution (3.81 *vs.* 6.36, excluding target 39), but also decreases the mean number of acceptable or better solutions in the top 10 (2.67 *vs.* 3.42, including target 39).

To further evaluate these methods, the enrichment in acceptable or better solutions in the 20% best ranked and 20% worst ranked conformations were computed. Results (Table 5) clearly show that the top 20% is largely enriched in acceptable or

better solutions, and even more in medium or better solutions when considering the *fnat* criterion. The comparison between these two categories is more difficult when using all of the CAPRI criteria, since in most cases the computation cannot be made. It can also be seen that CF-GA is better at enriching the top 20% in acceptable or better solutions. It should also be noted that for the three methods, using CAPRI criteria, no acceptable or better solution is ranked in the worst 20%.

We have compared these results with the ones obtained by other scoring groups on the 12 targets. As can be seen from Table 6, two of the targets for which we do not find an acceptable or better solution in the top 10 (T35 with all CAPRI criteria, and T39 with either quality measures) were difficult targets, and only one group obtained an acceptable solution for T35, none for T39. It should also be noted that target 35 is not a biological complex, but the assembly of two different modules belonging to the same protein chain.

Target 37 was found by most scorers. Our failure for this target is probably related to the fact that this complex is made of three protein chains (A, C and D), and the docking was conducted using only two of these chains. The resulting candidate interfaces, since they represent only a portion of the native interface, are two small to be favourably ranked by our method. Target 40 is also a trimer (chains A, B and C), but this time with two distinct interfaces (CA: target 40A, and CB: target 40B). The GA-CF method successfully finds the CA interface, but fails to favourably rank a good conformation for interface CB. The CA interface is significantly larger than CB (1009.5 $Å^2$ *vs.* 731.3 $Å^2$). Here again, the size of this second interface is two small for our method, especially since much larger interfaces (corresponding to the CA interface) are found in the proposed conformations.

For targets 22, 23, 26 and 27, the CAPRI criteria for all proposed conformations are not available. We have compared the categories given to the different conformations by the two criteria sets. Results shown Table 7 show that 99.4% of the conformations evaluated as high quality using the *fnat* criterion are evaluated as at least acceptable using all criteria (76.8% are even evaluated as medium or better), and 84.7% of the conformations evaluated as medium using the *fnat* criterion are evaluated as acceptable or better using CAPRI criteria. Consequently, the solutions found in the top 10 for targets 22, 23, 26 and 27 would very likely be considered as acceptable or better using CAPRI criteria. The conformations retained for targets 22, 23, 26 and 27 have *fnat* values of 0.95, 0.61, 0.45 and 1 respectively. Upon visual inspection (see Figure 5), and global RMSD computation, we estimated that their CAPRI status would be high, medium, acceptable and high respectively.

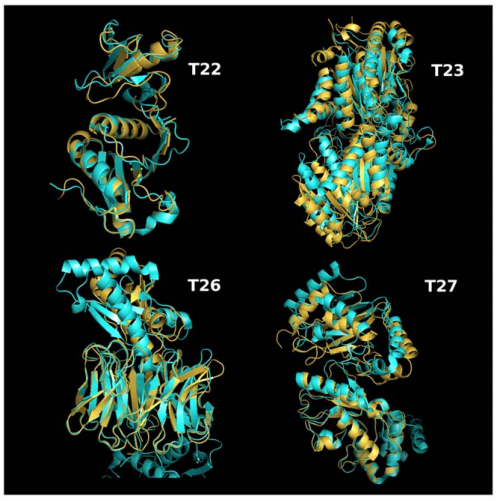


Figure 5. Conformations retained for targets 22, 23, 26 and 27. Native structure in orange, prediction in blue.
doi:10.1371/journal.pone.0018541.g005

Apart from the results obtained by our scoring function, this study shows that the *fnat* criterion, although and because it is less stringent than the CAPRI criteria, allows a better estimation of the performances of prediction methods. Indeed, predictions that correctly identify the interface area on both protein would be considered *incorrect* using the CAPRI criteria, but *acceptable* using the *fnat* criterion. For predictions having correct contacts, classified as *high* with the *fnat*, the CAPRI criteria often classifies them as *medium* or even *low*, mostly because of errors in global relative orientations of the two partners. Consequently, the *incorrect* class with the CAPRI criteria doesn't distinguish between these predictions, which have a very high biological utility, and predictions having few native contacts, which are *biologically* wrong. Thus it appears that, from the biologist's point of view, the *fnat* criterion is certainly more useful.

Globally the CF-GA method performs very well, ranking acceptable or better solutions in the top 10 for 8 out of 12 targets. The comparison with other methods is very difficult, since the other methods are evolving and the different groups have not participated to the same rounds. However, it can be seen that the performances of CF-GA compare favorably with current well-performing techniques.

Conclusion

We have shown that the use of a collaborative filtering strategy combined to a learning procedure leads to an efficient method. Using this technique, we are able to rank at least one acceptable or better solution in top 10 for 10 out of 12 CAPRI targets using solely the *fnat* criterion, and 4 out of 8 when using all CAPRI criteria, in cases where scoring ensembles contain acceptable or better solutions. We have also shown that the set of 20% best ranked conformations is largely enriched in medium or better conformations, whereas the set of 20% worst ranked solutions contains very few good models.

Table 7. Comparison between *Capri* and *fnat* evaluations.

	Capri				**Total**
fnat	⋆⋆⋆	⋆⋆	⋆	**Incorrect**	
⋆⋆⋆	204	298	148	4	654
⋆⋆		21	106	23	150
⋆			44	547	591
Incorrect				7069	7069
Total	204	319	298	7643	8464

For all the conformations in the CAPRI scoring ensembles, the classifications as high-quality, medium-quality, acceptable or incorrect conformation using only *fnat*, or complete CAPRI are compared. For example, there are 298 conformations classified as medium-quality using CAPRI criteria and high-quality by *fnat* criterion.
doi:10.1371/journal.pone.0018541.t007

The use of RMSD-filtering allows to increase the diversity of the conformations present in the top 10, which decreases the mean rank of the first acceptable or better conformation, but also decreases the number of acceptable or better conformations in the top 10. This is an advantage in an exploration perspective, since the proposed conformations are very different from each other. But this is also a disadvantage in an optimization or refinement perspective, since, for example, a very favourably ranked medium quality conformation can eliminate a high quality conformation having a slightly higher rank.

Finally, we have seen that our method fails on trimers. In the case of target 40 this is largely due to the fact that our method searches the best interface, and is not trained to look for multiple interfaces. Finding these interfaces would probably require training the method specifically on complexes with more than two chains.

Author Contributions

Conceived and designed the experiments: TB JB JA AP. Performed the experiments: TB JB JA AP. Analyzed the data: TB JB JA AP. Contributed reagents/materials/analysis tools: TB JB JA AP. Wrote the paper: TB JB JA AP.

References

1. Wodak SJ, Janin J (2002) Structural basis of macromolecular recognition. Adv Protein Chem 61: 9–73.
2. Sanderson CM (2009) The cartographers toolbox: building bigger and better human protein interaction networks. Brief Funct Genomic Proteomic 8: 1–11.
3. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The Protein Data Bank. Nucleic Acids Research 28: 235–242.
4. Ritchie DW (2008) Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 9: 1–15.
5. Mosca R, Pons C, Fernández-Recio J, Aloy P (2009) Pushing structural information into the yeast interactome by high-throughput protein docking experiments. PLoS Comput Biol 5: e1000490.
6. Kastritis PL, Bonvin AM (2010) Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 9: 2216–25.
7. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47: 409–43.
8. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ (2008) Principles of exible protein-protein docking. Proteins 73: 271–89.
9. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A (2008) DiMoVo: a Voronoi tessellationbased method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics 24: 652–8.
10. Bernauer J, Poupon A, Azé J, Janin J (2005) A docking analysis of the statistical physics of protein-protein recognition. Phys Biol 2: S17–23.
11. Bernauer J, Azé J, Janin J, Poupon A (2007) A new protein-protein docking scoring function based on interface residue properties. Bioinformatics 23: 555–62.
12. Bourquard T, Bernauer J, Azé J, Poupon A (2009) Comparing Voronoi and Laguerre tessellations in the protein-protein docking context;. In: Sixth International Symposium on Voronoi Diagrams (ISVD). pp 225–232.
13. Boissonnat JD, Devillers O, Pion S, Teillaud M, Yvinec M (2002) Triangulations in CGAL. Comput Geom Theory Appl 22: 5–19.
14. Pontius J, Richelle J, Wodak S (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264: 121–36.
15. Chen R, Mintseris J, Janin J, Weng Z (2003) A protein-protein docking benchmark. Proteins 52: 88–91.
16. Murzin A, Brenner S, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–40.
17. Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. Advances in Artificial Intelligence.
18. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The weka data mining software: An update. SIGKDD Explorations 11: 10–18.
19. Joachims T (1999) Making large-scale support vector machine learning practical. CambridgeMA, , USA: MIT Press. pp 169–184.
20. Janin J, Henrick K, Moult J, Eyck L, Sternberg M, et al. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52: 2–9.
21. Lensink MF, Mendez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 69: 704–18.
22. Tong Y, Chugha P, Hota PK, Alviani RS, Li M, et al. (2007) Binding of Rac1, Rnd1, and RhoD to a novel Rho GTPase interaction motif destabilizes dimerization of the plexin-B1 effector domain. J Biol Chem 282: 37215–24.
23. Mendez R, Leplae R, Lensink MF, Wodak SJ (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. Proteins 60: 150–69.
24. le Cessie S, van Houwelingen JC (1992) Ridge estimators in logistic regression. Applied Statistics 41: 191–201.
25. Schölkopf B, Burges CJ (1998) Advances in kernel methods - support vector learning .
26. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann.
27. Cohen WW (1995) Fast effective rule induction. In: In: Twelfth International Conference on Machine Learning. pp 115–123.
28. Eibe Frank IHW (1998) Generating accurate rule sets without global optimization. In: In: Fifteenth International Conference on Machine Learning. pp 144–151.

COARSE-GRAINED AND ALL-ATOM KNOWLEDGE-BASED POTENTIALS FOR RNA

Julie BERNAUER, Xuhui HUANG, Adelene Y.L. SIM, Michael LEVITT. In *RNA*, 17(6):1066-75, 2011.

Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation

JULIE BERNAUER,[1,4] XUHUI HUANG,[2,4] ADELENE Y.L. SIM,[3] and MICHAEL LEVITT[4]
[1]INRIA AMIB Bioinformatique, Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau, France
[2]Department of Chemistry, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[3]Department of Applied Physics, Stanford University, Stanford, California 94305-4090, USA
[4]Department of Structural Biology, Stanford University, Stanford, California 94305-5126, USA

ABSTRACT

RNA molecules play integral roles in gene regulation, and understanding their structures gives us important insights into their biological functions. Despite recent developments in template-based and parameterized energy functions, the structure of RNA—in particular the nonhelical regions—is still difficult to predict. Knowledge-based potentials have proven efficient in protein structure prediction. In this work, we describe two differentiable knowledge-based potentials derived from a curated data set of RNA structures, with all-atom or coarse-grained representation, respectively. We focus on one aspect of the prediction problem: the identification of native-like RNA conformations from a set of near-native models. Using a variety of near-native RNA models generated from three independent methods, we show that our potential is able to distinguish the native structure and identify native-like conformations, even at the coarse-grained level. The all-atom version of our knowledge-based potential performs better and appears to be more effective at discriminating near-native RNA conformations than one of the most highly regarded parameterized potential. The fully differentiable form of our potentials will additionally likely be useful for structure refinement and/or molecular dynamics simulations.

Keywords: RNA structure; knowledge-based potential; scoring

INTRODUCTION

RNA molecules are responsible for a wide range of biological processes occurring in the cell. To function, RNAs adopt detailed three-dimensional (3-D) folds (Gesteland et al. 2006). Understanding these structural intricacies gives insights to molecular evolution and structure-function relationships. Recently it was shown that, with high-resolution 3-D motifs, it is possible to design optimal sequences that improve RNA function (Das et al. 2010). This highlights the need for accurate RNA structure prediction and evaluation tools.

It had been hoped (Tinoco and Bustamante 1999) that the four nucleotides alphabet of RNA would make RNA structure prediction a more tractable problem than for proteins, since the latter have wider structural diversity arising from their 20 natural amino acids library. However, predicting the fold of RNA molecules, especially larger systems, is still a daunting task. Fortunately, RNA structure prediction is simplified by the hierarchical folding process of most RNAs (Brion and Westhof 1997; Batey et al. 1999; Tinoco and Bustamante 1999). An extended RNA first forms stable secondary structure defined by base-pairing, then packs into a globular 3-D form.

Given the efficient techniques developed for secondary structure prediction (Zuker 2003; Mathews 2006; Reeder et al. 2006; Shapiro et al. 2007; Hofacker 2009), the major remaining difficulty is determining the detailed local structure of bases and how they affect the RNA's global 3-D structure. Typical base interactions are base-pairing (canonical and noncanonical) and base-stacking. Even tertiary interactions—which usually contribute strongly to an RNA molecule's overall 3-D fold—like the tetraloop-tetraloop receptor (a well-defined base-pairing interaction between two distant small motifs) can be reduced to such local base interactions. Extensive work has been done to classify these

Reprint requests to: Julie Bernauer, INRIA AMIB Bioinformatique, Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau, France; e-mail: julie.bernauer@inria.fr.
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2543711.

 Published by Cold Spring Harbor Laboratory Press. Copyright © 2011 RNA Society.

Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation

JULIE BERNAUER,[1,4] XUHUI HUANG,[2,4] ADELENE Y.L. SIM,[3] and MICHAEL LEVITT[4]
[1]INRIA AMIB Bioinformatique, Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau, France
[2]Department of Chemistry, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[3]Department of Applied Physics, Stanford University, Stanford, California 94305-4090, USA
[4]Department of Structural Biology, Stanford University, Stanford, California 94305-5126, USA

ABSTRACT

RNA molecules play integral roles in gene regulation, and understanding their structures gives us important insights into their biological functions. Despite recent developments in template-based and parameterized energy functions, the structure of RNA—in particular the nonhelical regions—is still difficult to predict. Knowledge-based potentials have proven efficient in protein structure prediction. In this work, we describe two differentiable knowledge-based potentials derived from a curated data set of RNA structures, with all-atom or coarse-grained representation, respectively. We focus on one aspect of the prediction problem: the identification of native-like RNA conformations from a set of near-native models. Using a variety of near-native RNA models generated from three independent methods, we show that our potential is able to distinguish the native structure and identify native-like conformations, even at the coarse-grained level. The all-atom version of our knowledge-based potential performs better and appears to be more effective at discriminating near-native RNA conformations than one of the most highly regarded parameterized potential. The fully differentiable form of our potentials will additionally likely be useful for structure refinement and/or molecular dynamics simulations.

Keywords: RNA structure; knowledge-based potential; scoring

INTRODUCTION

RNA molecules are responsible for a wide range of biological processes occurring in the cell. To function, RNAs adopt detailed three-dimensional (3-D) folds (Gesteland et al. 2006). Understanding these structural intricacies gives insights to molecular evolution and structure-function relationships. Recently it was shown that, with high-resolution 3-D motifs, it is possible to design optimal sequences that improve RNA function (Das et al. 2010). This highlights the need for accurate RNA structure prediction and evaluation tools.

It had been hoped (Tinoco and Bustamante 1999) that the four nucleotides alphabet of RNA would make RNA structure prediction a more tractable problem than for proteins, since the latter have wider structural diversity arising from their 20 natural amino acids library. However, predicting the fold of RNA molecules, especially larger systems, is still a daunting task. Fortunately, RNA structure prediction is simplified by the hierarchical folding process of most RNAs (Brion and Westhof 1997; Batey et al. 1999; Tinoco and Bustamante 1999). An extended RNA first forms stable secondary structure defined by base-pairing, then packs into a globular 3-D form.

Given the efficient techniques developed for secondary structure prediction (Zuker 2003; Mathews 2006; Reeder et al. 2006; Shapiro et al. 2007; Hofacker 2009), the major remaining difficulty is determining the detailed local structure of bases and how they affect the RNA's global 3-D structure. Typical base interactions are base-pairing (canonical and noncanonical) and base-stacking. Even tertiary interactions—which usually contribute strongly to an RNA molecule's overall 3-D fold—like the tetraloop-tetraloop receptor (a well-defined base-pairing interaction between two distant small motifs) can be reduced to such local base interactions. Extensive work has been done to classify these

Reprint requests to: Julie Bernauer, INRIA AMIB Bioinformatique, Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau, France; e-mail: julie.bernauer@inria.fr.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2543711.

 Published by Cold Spring Harbor Laboratory Press. Copyright © 2011 RNA Society.

base interactions (Murray et al. 2003; Sykes and Levitt 2005; Das and Baker 2007; Frellsen et al. 2009). Recent advances in RNA structure prediction techniques make use of base-pairing and stacking preferences either in the form of an energy function (Dima et al. 2005; Jonikas et al. 2009; Flores and Altman 2010) or through the use of fragment libraries taken from known RNA structures (Das and Baker 2007; Parisien and Major 2008).

Despite our understanding and classification of base interactions, for a given RNA, there are still many possible conformations consistent with reasonable secondary structures. Therefore the selection of good native-like models from an ensemble of conformations (also known as decoys) is a vital, yet very challenging task. Das and colleagues showed that a low resolution energy function was insufficient to discriminate good models (Das and Baker 2007)—defined by low root-mean-squared deviation (RMSD) to the native state—and that, with the addition of some higher resolution terms, its discriminatory power increased significantly (Das et al. 2010). This energy function—made available in the Rosetta package—tackles small RNA motifs more effectively than physics-based energy functions. However, the Rosetta RNA energy function is based on careful parameterization of weights for the various energetic components arising from preferred RNA base orientations and interactions, and therefore it is unclear how its efficacy scales with RNA size.

Similar problems in the protein folding world have led to the development of knowledge-based (KB) potentials. For instance, the potential of mean force (PMF) was generated from distance distributions between protein atoms, and was shown to be effective in screening and refining protein decoys (Samudrala and Moult 1998; Zhou and Zhou 2002; Zhang et al. 2004; Summa and Levitt 2007). To derive such a potential, a training set of high-resolution, nonredundant structures is required. The smaller number of high-resolution RNA structures available has thus far stalled the development of such distance-based KB potentials specifically designed for RNA.

In this study, we derive differentiable all-atom and coarse-grained KB potentials for RNA structures, using careful statistical treatment to handle low count regions. Unlike Rosetta, or other existing RNA potentials, our KB potentials implicitly incorporate all base interactions into distance-based potentials, eliminating the need for accurate weighting of energetic components. Our results show that our all-atom potential is effective in scoring RNA decoys for the selection of good native-like models in RNA systems of different sizes. When the native structure is derived by NMR, some of the near-native decoy structures scored with the all-atom potential have an energy that is below that of the NMR-determined native state: These structures may be closer to the true native state and thus constitute refined native states. The fully differential forms of our potentials facilitate their use in molecular dynamics (MD) and structure refinement.

RESULTS

Selection of representative RNA data set

The generation of an effective KB potential requires the careful selection of representative RNA structures. This data set of RNA structures should be high-resolution (to capture the intricate base–base interactions), nonredundant (to ensure that no particular RNA structure dominates), and sufficiently large (to provide good statistics). These conflicting criteria are difficult to meet and are not satisfied by the existing structure sets available in the literature such as RNAbase (Murthy and Rose 2003) or NDB (Murray et al. 2003).

We developed a protocol that combines automated and manual data curations designed to facilitate the extraction of high-quality, representative RNA structures (Supplemental Fig. 1; Materials and Methods). The process selects RNA-only structures that are solved by X-ray crystallography up to a resolution of at least 3.5 Å, in the absence of bound ligands or proteins. Structures that have identical sequences are filtered out to prevent redundancy in the data set.

The complete extraction procedure applied to the PDB (Berman et al. 2007) led to 77 selected RNA structures (total 7251 nucleotides [nt]). Fifty-four molecules in our data set also belong to the Stombaugh et al. (2009) data set, which contains 304 structures. Our data set is much smaller due to the stringent criteria used. The finalized data set used in the generation of the continuous RNA potential is summarized in Supplemental Table 1. Other than being useful for this work, the data set is generic enough to be used for other learning purposes and we have therefore made it publicly available (http://csb.stanford.edu/rna).

Generation of RNA potential

There are several ways to extract information from our structural data set. Some common methods to do this use known RNA base–base interactions like base-pairing and base-stacking, and generate independent potentials that are specific to these interactions (Sharma et al. 2008; Jonikas et al. 2009; Das et al. 2010). This approach, however, requires careful parameterization of the different energetic components. Alternatively, the BARNACLE model (Frellsen et al. 2009) uses dihedral angles from RNA rotamers (Murray et al. 2003) to train an angle-based RNA potential. While this appears to work well for sampling small RNA systems constrained by secondary structure information, it seems less likely that such a potential will capture tertiary interactions between distant motifs. Instead, we make use of distributions of inter-atomic distances, which allows us, in principle, to incorporate information from a wide array of interaction types.

We generated two RNA KB potentials: a coarse-grained five-point (P, C4′ backbone atoms, and C2, C4, and C6

base-planar atoms per base) version, and an all-atom version. The former is likely to be more effective for fast, efficient sampling due to the simplified representation of each base. This same five-atoms description was shown to be sufficient in describing base orientations (Sykes and Levitt 2005). The all-atom potential, on the other hand, may be useful for high-resolution RNA structure refinement, as a result of its inherent amount of structural detail, as it is for proteins (Chopra et al. 2010). Due to their nonoverlapping utility, both potentials were developed and tested here.

The distance computation led to ~1 million distances <16 Å for the five-atoms per nucleotide model. Among them, 64% are due to the ribosomal RNA family (51% being due to the only complete ribosome structure included in our data set).

To obtain a potential from these distance measurements, we built a PMF as described previously for proteins (Samudrala and Moult 1998; Lu and Skolnick 2001). The potential between two atoms i and j at distance d_{ij} apart can be written as an energy function (Samudrala and Moult 1998) expressed as

$$E = -kT \sum_{ij} \ln\left(\frac{P_{obs}(d_{ij})}{P_{ref}(d_{ij})}\right)$$

where T is the temperature (taken to be 300 K) and k the Boltzmann constant. $P_{obs}(d_{ij})$ and $P_{ref}(d_{ij})$ represent the observed and reference probabilities, respectively, for the atoms i and j to be separated by distance d_{ij}.

Unlike previous work, in this study, $P_{obs}(d_{ij})$ and $P_{ref}(d_{ij})$ are not computed by binning distances, which could significantly affect the results. Instead, these are probability distributions obtained from statistical analysis (see Materials and Methods): We used a Dirichlet Process Mixture Model to obtain the analytical form of the potential as a sum of Gaussian functions. Another feature of this potential is that it is fully differentiable, making it suitable for energy minimization or MD. To our knowledge, this is the first RNA KB potential that can be directly applied to continuum MD, though a KB potential for discrete MD has been designed (Sharma et al. 2008).

In developing KB potentials, the choice of the reference state is key. Some options include an ideal gas reference state (Zhou and Zhou 2002) or a quasi-chemical approximation (Lu and Skolnick 2001), which originates from "uniform density" reference state defined by Sippl (1990). This study used the latter with a composition-independent scale, i.e., the observed distances from all possible pairs are combined together to represent the reference state.

Assessment of potentials by decoy scoring

To assess the quality of our KB potentials we used them to score a variety of RNA decoys, and observed their abilities to distinguish good, near-native models. Scoring is a quick and simple way to evaluate the quality of a potential, compared to more computationally intensive methods like refinement and sampling. As a comparison, we scored the same decoys using the latest high-resolution scoring function from Rosetta (Das et al. 2010).

One set of decoys was generated by position-restrained molecular dynamics and replica-exchange molecular dynamics (REMD) simulations (see Materials and Methods), methods that cover a wide near-native RMSD range (from 0.1 to ~12 Å). Five different RNA structures were used, and scores evaluated using the KB potentials generated from the full data set (Fig. 1). The cropped (using a data set where the five structures were all removed) and full versions of the KB potentials yield similar results (see Supplemental Fig. 2). In all five cases, the all-atom and coarse-grained KB potentials and Rosetta were very effective in identifying near-native decoys, as indicated by the strong scoring funnel toward the native structure.

The assessment of potentials using a single method for decoy generation may be insufficient to determine their

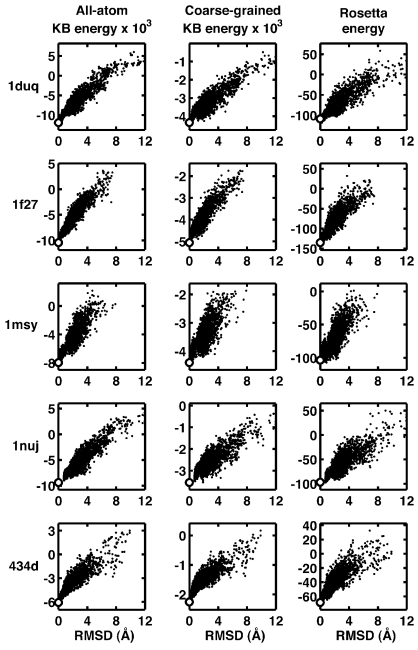


FIGURE 1. Energy as a function of RMSD for decoys generated using position-restrained dynamics together with replica-exchange molecular dynamics for five different systems (rows). All-atom KB, coarse-grained KB, and Rosetta energies are shown in the *left*, *middle*, and *right* columns, respectively. In each case, a funnel shape toward the native structure (white circle) is seen, characteristic of a scoring function that is effective at distinguishing near-native structures from less native-like structures.

limitations (Handl et al. 2009). We therefore generated a second set of decoys using normal modes (NM). These decoys cover a narrower range of RMSD but present different geometrical distortions from the prior physics-based force-field methods. The all-atom, coarse-grained and Rosetta potentials show similar efficacy (Fig. 2; Supplemental Fig. 3): The funnel shape characteristic of good potentials is less pronounced, suggesting weaker ability of all three potentials to differentiate such decoys.

Last, we tested the potentials' capabilities to score diverse RNA structures assembled by RNA-like fragments which had no native base-pairing enforced (see Supplemental Figs. 4–7). Not surprisingly, due to the reduced constraints, all three potentials were less effective in scoring these decoys. In general, our all-atom KB potential (full version) still appears to quantitatively do better than Rosetta (see Table 1). Das and colleagues showed that the combination of refinement and scoring improved the discriminatory power of the Rosetta potential (Das et al. 2010), suggesting that, with atomic refinement, our all-atom KB potential could possibly perform well too.

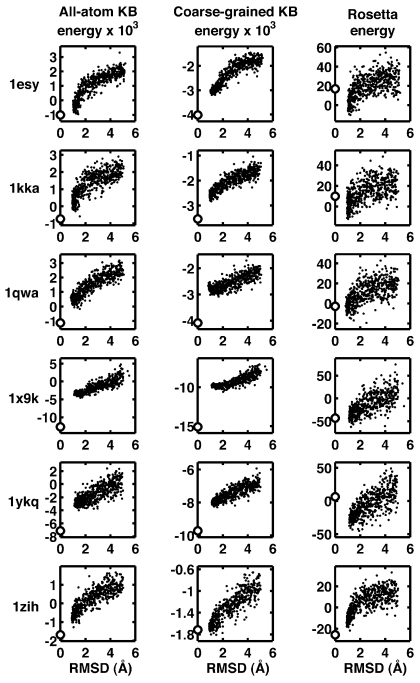


FIGURE 2. Energy as a function of RMSD for decoys generated by normal modes for six RNA structures (more in Supplemental Fig. 3). Scoring using our two KB potentials (all-atom on *left*, coarse-grained in *middle*) and Rosetta (*right*) are shown. Native scores are represented as white circles. A funnel toward low RMSD is seen in most cases. However, in several instances, some decoys score better than the native structure, a behavior that is more pronounced for the Rosetta scoring function.

Evaluation metrics

For a quantitative comparison between all three potentials, we counted the number of decoys that scored lower than the native structure (Table 1). This gives us an indication of the number of structures that will be erroneously selected ahead of the native structure due to limitations in the potentials. Alternatively, we also define an Enrichment Score (*ES*), a useful metric based on identifying the top 10% scoring ($E_{top10\%}$) and best 10% RMSD values ($R_{top10\%}$), then evaluating their degree of overlap (this choice percentage is somewhat arbitrary). The Enrichment Score (Tsai et al. 2003) is defined as

$$ES = \frac{\left|E_{top10\%} \cap R_{top10\%}\right|}{0.1 \times 0.1 \times N_{decoys}}$$

where $|E_{top10\%} \cap R_{top10\%}|$ is the number of structures in the intersection of $E_{top10\%}$ and $R_{top10\%}$. $E_{top10\%}$ corresponds to the set of structures with energies in the best 10% of the energy range. $R_{top10\%}$ corresponds to the set of structures having RMSD in the lowest 10% of the RMSD range.

For a perfectly linear scoring function for which $E_i = c \times R_i$ for each structure i and c is a constant, this would give

$$ES = \frac{\left|E_{top10\%} \cap R_{top10\%}\right|}{0.1 \times 0.1 \times N_{decoys}} = \frac{0.1 \times N_{decoys}}{0.1 \times 0.1 \times N_{decoys}} = 10$$

In a random scoring case, we would have

$$ES = \frac{\left|E_{top10\%} \cap R_{top10\%}\right|}{0.1 \times 0.1 \times N_{decoys}} = \frac{0.1 \times 0.1 \times N_{decoys}}{0.1 \times 0.1 \times N_{decoys}} = 1$$

Hence, we have

$$ES = \begin{cases} 10, \text{ perfect scoring} \\ 1, \text{ perfectly random} \\ < 1, \text{ bad scoring} \end{cases}$$

What constitutes a good scoring function is not obvious, though it clearly should have an *ES* between 1 and 10, the closer to 10 the better. For MD and NM, where RNA decoys have secondary structures similar to their respective native states, our all-atom KB potential appears to generally do best (Table 1).

DISCUSSION

Captured structural features

Common RNA base interactions typically explicitly represented in RNA potentials or force-fields are base-pairing

TABLE 1. Quantitative comparisons of the decoy-screening capabilities of our KB potentials (all-atom and coarse-grained) with the Rosetta RNA potential

				Number of Structures below Native Energy			Enrichment Score		
Decoy generation method	RNA	Chain length	Experimental resolution (Å)	All-atom KB	Coarse-grained KB	Rosetta	All-atom KB	Coarse-grained KB	Rosetta
(A) Position-restrained dynamics and REMD	1duq	26	2.1	0	1	190	7.6	7.3	7.1
	1f27	30	1.3	0	0	0	7.9	7.2	6.2
	1msy	27	1.41	0	0	8	5.7	4.9	3.6
	1nuj	24	1.8	0	0	3	7.3	7.0	6.9
	434d	14	1.16	0	1	0	7.7	7.9	6.8
(B) Normal modes	1duq	26	2.1	0	0	0	7.0	4.4	3.8
	1esy	19	NMR	0	0	203	5.4	5.6	5.6
	1f27	30	1.3	0	0	0	5.8	4.8	2.6
	1i9v	76	2.6	0	0	0	2.6	4.4	3.0
	1kka	17	NMR	0	0	178	4.6	5.2	4.6
	1msy	27	1.41	0	0	0	5.6	4.0	4.6
	1nuj	24	1.8	0	0	0	7.4	3.8	2.4
	1qwa	21	NMR	0	0	65	3.2	1.0	3.8
	1x9k	62	3.17	0	0	32	1.6	1.0	3.0
	1xjr	46	2.7	0	0	0	5.4	5.4	2.2
	1ykq	49	1.9	0	0	350	3.4	3.8	2.8
	1zih	12	NMR	0	11	0	5.4	2.8	6.6
	28sp	28	NMR	0	0	0	4.0	1.6	1.8
	2f88	34	NMR	0	0	0	5.4	2.6	4.4
	434d	14	1.16	0	0	0	7.4	0.6	5.2
	1a4d	41	NMR	17	497	0	3.8	2.2	0.8
	1csl	28	1.6	0	0	0	1.5	0.2	1.3
	1dqf	19	2.2	0	0	0	1.8	1.0	1.0
	1esy	19	NMR	305	505	65	3.7	1.8	1.2
	1i9x	26	2.18	0	0	0	1.3	0.8	1.5
	1j6s	24	1.4	0	468	0	1.4	1.0	0.6
	1kd5	22	1.58	0	0	0	0.3	1.0	0.2
	1kka	17	NMR	467	495	69	1.2	0.4	0.6
	1l2x	27	1.25	0	0	0	3.2	1.8	1.8
	1mhk	32	2.5	0	4	0	1.2	0.6	1.0
	1q9a	27	1.04	0	497	0	0.5	0.5	0.8
	1qwa	21	NMR	187	505	26	1.2	0.8	1.0
	1xjr	46	2.7	0	0	0	2.0	1.0	1.2
	1zih	12	NMR	36	504	0	5.0	5.7	2.0
	255d	24	2.0	0	0	0	0.7	0.7	1.3
	283d	24	2.3	0	0	0	0.8	0.8	0.7
	28sp	28	NMR	299	504	0	1.5	1.3	1.7
	2a43	26	1.34	0	0	0	2.0	2.0	0.6
	2f88	34	NMR	0	498	0	1.3	1.2	1.3
AVERAGE VALUES	(A)		5	0	0	40	7.2	6.9	6.1
	(B)		15	0	1	55	4.9	3.4	3.8
	(C)		20	66	224	8	1.8	1.3	1.1
	X-Ray		27	0	36	22	7.8	6.0	5.7
	NMR		13	101	271	47	3.5	2.5	2.7
	All		40	33	112	30	3.7	2.8	2.7

Overall, the all-atom KB potential is a more discriminating scoring function than Rosetta for all three decoys sets as well as for X-ray and NMR structures. This is seen in the Enrichment Scores (*ES*), where the all-atom KB potential has a higher *ES* than that for Rosetta in 29 cases whereas Rosetta is better in only nine of the 40 cases. The average values of *ES* for each decoy set (A, B, and C) show that set A, which is derived by restrained molecular dynamics, is easiest to discriminate, whereas set C, which is derived by FARNA, is the hardest to discriminate. Decoys derived from structures determined by X ray are much easier to discriminate than those derived by NMR. Similar trends are seen in the number of structures below native energy: Our all-atom KB potential finds no such structures for decoys whose native structure is determined by X-ray crystallography. Overall, there is no significant difference between our coarse-grained KB potential and Rosetta (except for RNA structures solved by NMR). The significant number of decoys with scores below that of NMR-derived native structures for both of our KB potentials suggests that these potentials might be useful for near-native decoy refinement. This could also be an artifact of our KB potentials being derived from X-ray structures. The largest ES for each decoy are shaded in light green, while the largest number of structures below native energy are shaded in pink.

and base-stacking. While we did not include these terms explicitly in our KB potentials, the distance-dependent potentials we developed should inherently include this information. Figure 3 shows that our potentials and their first derivatives are smooth and important base-interaction features are also captured as troughs in the potentials. For instance, base-stacking is more pronounced between purines, but least between pyrimidines (Saenger 1984). The potential between C4 atoms in guanines (purines; Fig. 3A) shows a well at ~4.4 Å, which corresponds to the distance between base-stacked C4 atoms. On the contrary, this basin is absent for a similar potential between uracils (pyrimidines; Fig. 3B), consistent with the weak base-stacking interaction. The base-pairing interaction between guanine and cytosine (and adenine and uracil; not shown) is also captured (Fig. 3C).

The full energy landscape of an RNA is hyper-dimensional and cannot be adequately visualized from these potential plots, which are low-dimensional projections of the full energy surface. Nonetheless, the success of our potentials in scoring RNA decoys suggests that our KB representation of the RNA landscape is reasonable. Most native conformations can be accurately identified by our KB potentials even in its coarse-grained form (Figs. 1, 2; Supplemental Figs. 3–7). There are, however, structures that score close to, or lower than the native. To have a sense of which structural features are well captured by our potentials, we superimposed the best-scored decoys to the native state, and observed their structural differences.

Unsurprisingly, due to their dominance in KB statistics, helical topologies are well preserved and captured by our KB potential scoring (see Fig. 4). This also appears to be the case for Rosetta scoring. However, Rosetta is less effective in scoring the correct loop structure compared to our all-atom KB potential. The KB potential, unlike Rosetta, does not contain explicit base-pairing and base-stacking terms and hence does not necessarily favor a helix-like stacking for loops. This might be why our all-atom KB potential outperforms Rosetta in scoring the GUAA tetraloop (Fig. 4). Success in modeling such small motifs by Rosetta (Das et al. 2010) suggests that all-atom refinement of the models could

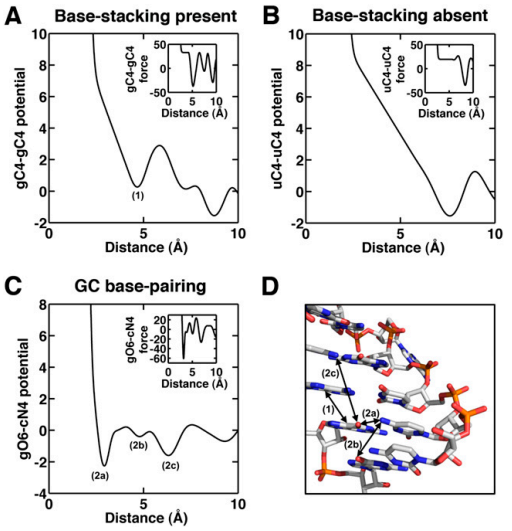


FIGURE 3. Structural features captured by the KB potential. The plots (*A–C*) show the potentials for specific atom pairs. In each plot, the corresponding force is shown in the *inset*. (*A*) gC4-gC4 potential showing a base-stacking well ~4.4 Å labeled (1). (*B*) uC4-uC4 potential showing no base-stacking well. (*C*) gO6-cN4 potential showing a deep base-pairing well (2a) and various structural wells (2b and 2c). (*D*) Distances represented in the different wells shown on the Rev binding element of HIV-1 structure (PDB id: 1duq).

improve scoring, but such analysis is beyond the scope of this current work.

From Table 1, the experimental method used to solve the native structure has an impact on the data: It seems to be more difficult to obtain good scoring for RNA structures solved by NMR. When the native structure is derived by NMR, some of the near-native decoy structures scored with the all-atom potential have energies below that of the NMR-determined native state. While this could likely be attributed to the use of X-ray structures in generating the KB and Rosetta potentials, the behavior could also be partly due to a single NMR reference structure not fully representing the true native state. NMR structures are usually more varied and their accuracies are hard to evaluate, in contrast to X-ray structures where resolution and R_{free} factor provide good insights into the quality of structures.

The quality of scoring also depends on the nature of the decoy set. For example, for structures 1q9a and 28sp, FARNA failed to model bulge regions present in the native RNA, so all FARNA decoys used lacked such bulges. Hence scoring results were bad for both KB potentials and Rosetta (Table 1).

In general, the coarse-grained KB potential is less effective at screening decoys, likely because high-resolution information is omitted from its representation. This could explain the reduced ability of the coarse-grained KB potential to discriminate good decoys of small RNAs (e.g., 1zih, 12 bases; 434d, 14 bases).

Fully differentiable potentials for refinement and modeling

As mentioned previously, our KB potentials are fully differentiable and could be effective for refinement of near-native RNA decoys. The scoring results on the different types of RNA decoys (Figs. 1, 2; Supplemental Figs. 3–7) indicate that our potentials might be promising for refinement, since they show strong funnels toward the native state. However, being able to refine a structure well also depends on the energy landscape close to the native structure (Chopra et al. 2008)—we cannot visualize this by the simple scoring scheme we have adopted here.

We can also make use of our KB potentials to run MD simulations on different RNA systems. However, it is unclear whether these potentials can effectively model unfolded or intermediate RNA states. Modeling such extended conformations may require long-range interactions, but such distances are lacking in X-ray structures of globular native RNA. To better address this problem, and possibly improve the geometry of base-interactions, we envision having to explicitly include base-pairing interactions or other orientation-dependent interactions like those used in recent studies (Dima et al. 2005; Stombaugh et al. 2009; Zirbel et al. 2009). In future work, we will look at structural refinement of

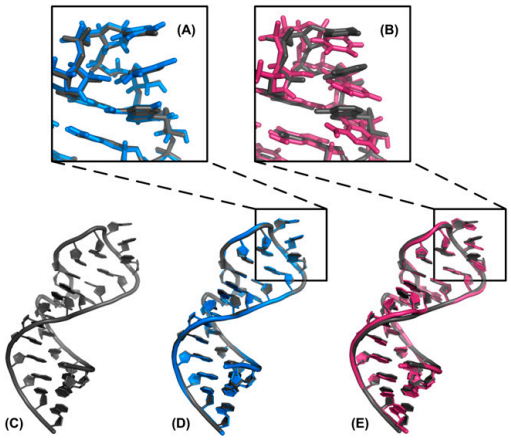


FIGURE 4. Comparison of the best scoring decoys for the GUAA tetraloop (PDB id: 1msy). The native structure is shown in *C* and the superimposed decoys selected by the all-atom KB potential and Rosetta are illustrated in *D* and *E*, respectively. In both *D* and *E* the native structure is also shown in gray. The close-up views of the tetraloop for both best scored decoys are shown in panels *A* and *B*, respectively. The Rosetta scoring function incorrectly selects a structure with stronger base planar stacking than found in the native structure.

near-native decoys to investigate the quality of our potentials close to the native energy basins, and then evaluate the need for additional terms in our KB potentials.

Simplified treatment of solvent and electrostatics

A major advantage of using KB potentials is the implicit treatment of electrostatics and solvent, through the use of pairwise atomic distances in experimentally determined structures. This removes the need to include solvents and ions in any sampling or scoring procedure, reducing the size of the problem, thus allowing the handling of large RNA systems. Since distance statistics were taken from crystal structures grown under a diverse range of ionic conditions (albeit most crystals were grown in the presence of divalent ions), our KB potentials cannot be directly related to specific ionic conditions. Rather, our potentials are likely applicable to the broad range of ionic conditions under which most RNAs fold to their native form. Arguably our KB treatment of electrostatics and hydration is crude and unphysical, since we intentionally did not take into account differences in ionic conditions of the different crystallized RNAs, and also did not differentiate between diffuse ions from partially or fully dehydrated ones. However, the significant reduction in computational complexity definitely improves sampling efficiency. We can make use of the KB potential to seed different structures for more intricate explicit solvent and ions MD simulations.

CONCLUSION

We built fully differentiable KB potentials from a carefully curated data set of high-resolution RNA structures and used decoys to assess their qualities. While such an evaluation scheme has its limitations (Handl et al. 2009), it is a fast and easy method to determine the quality of potentials. We minimized any bias by scoring decoys generated from three different approaches. Even in the absence of a priori information, our RNA potentials—in particular that with all-atom representation—lead to effective discrimination of RNA decoys, comparable to, and in some cases bettering, existing parameterized or template-based techniques.

MATERIALS AND METHODS

RNA data set and distance collection

We built our RNA data set by selecting RNA structures that fulfill the following specific requirements:

1. Each structure has been solved by X-ray crystallography to a resolution >3.5 Å.
2. The solved RNA structure should not be bound to proteins or ligands.
3. Less than 5% of the nucleotides in the RNA are modified or missing.
4. The data set does not contain two structures with sequence identity >80%.
5. The structure should be representative of the biologically active molecule (symmetric molecules are built if needed).

The RNA selection process consists of automated and manual portions. The PDB (2007 annual release) was scanned for suitability by using an in-house extension of the BioPython PDB module (Hamelryck and Manderick 2003) for nucleic acids. The lengths and sequences of RNA structures that meet criteria 1 and 2 (see above) were extracted and analyzed using the same module.

To account for identical RNAs, these sequences obtained were aligned using the program Blast (Altschul et al. 1990) and hierarchical clustering based on sequence identity was performed using the statistical program R (R Development Core Team 2008). These clusters of sequences were then manually evaluated. For each cluster, the structure corresponding to the longest sequence was retained. The structural details were manually curated and biological functions extracted from the relevant literature. When the biologically relevant molecule was not found in the asymmetric unit, symmetric chains were built using PyMOL (http://www.pymol.org/) (DeLano 2002) and added to the structure file.

Once selected, structures were labeled using a family tag (Ribosomal RNAs, Ribozymes, Transfer RNAs, Viral RNAs, SRP RNAs and miscellaneous; see Supplemental Table 1). This data set is available at http://csb.stanford.edu/rna.

Statistical analyses and functional forms

Computing $P_{obs}(d)$ and $P_{ref}(d)$ as shown in Equation 1 from distance measurements is essentially a density estimation problem. The probabilities are inferred from the distances $\{d_1,\ldots,d_n\}$, which are assumed to be exchangeable observations of P_{obs} and P_{ref}. There are many alternative ways for performing density estimation in univariate sets. In previous studies, fixed binning and spline fitting were mainly used. This strategy can induce a lot of artifacts due to low count and noise and the resultant probability density often may not be a good representation of the data. Thus we decided to rely on classical statistical techniques. In this study, we used a Dirichlet process mixture model, which leads to analytically differentiable potential functions. Density estimation was performed using the implementation of Dirichlet process mixture models in the Flexible Bayesian Modeling package written by R.M. Neal. This software defines a hierarchical structure for the prior of the parameters $\phi = \{\mu,\sigma^2\}$. The reader should refer to Neal (1998) for further details.

Normal mixture models are also widely used for density estimation. The density function is assumed to be a mixture of a number of Gaussian components weighted by factors $\omega = \{\omega_1,\ldots,\omega_n\}$. The density function has the form

$$P(d) = \sum_{j=1}^{N} \omega_j \mathrm{N}(\mu_j, \sigma_j^2).$$

Given a fixed number of components N, it is easy to find the function $P(d)$ that maximizes the likelihood of the data set. However, determining the optimal number of components in a statistically meaningful way is a difficult problem to which much research has been devoted (McLachlan and Peel 2000).

An alternative that has been investigated more recently is to extend the finite mixture model to an infinite mixture of components.

One can then use a purely Bayesian approach to infer the parameters of the model, with a clever prior for the mixing proportions of the components. A good choice for this prior is a Dirichlet process, which results in what is known as a Dirichlet process mixture model. These models can have strong advantages over their finite counterparts (Rasmussen 2000):

- In many applications it may be more appropriate not to limit the number of components.
- The number of represented classes is automatically determined.
- The use of reversible jump Markov Chain Monte Carlo (MCMC) effectively avoids local minima that plague mixtures trained by optimization methods.
- It is simpler to work at the infinite limit than to work with finite mixtures of unknown size.

To overcome signal instabilities generated by the estimations at small distances where the number of counts is small for both $P_{obs}(d)$ and $P_{ref}(d)$, the first part of the potential is assumed to be linear up to the first descending inflection point. The linear approximation proved to be sufficient to obtain reasonable looking potential shapes for the coarse-grained and all-atom potentials and shows better results than sigmoid estimates (data not shown). To ensure a smooth truncation at the distance cutoff (taken to be 14 Å), the whole signal was multiplied by a negative sigmoid function centered on the cutoff distance. Both of those assumptions lead to continuously differentiable energy and force functions, suitable for MD simulations.

Generation of decoy structures

In this study, we proposed a method to generate RNA decoy sets with RMSD ranging continuously from 0 Å to >10 Å. Our method is based on using MD simulations for sampling (Huang et al. 1996). Typical RNA MD simulations in explicit solvent will generate configurations that have RMSD values a few angstroms (typically 2 Å) away from the crystal structure. In order to generate near-native decoy structures (i.e., with RMSD <2 Å), we applied a position restraint potential on each heavy atom of the RNA molecule to constrain the motions of RNA. On the other hand, in order to generate decoy structures that are far from the native structure, we applied REMD, an enhanced sampling algorithm, to sample the configuration space far from the native structure.

MD simulations are often trapped in local free energy minima when sampling a rugged free energy landscape for biomolecular folding. REMD was developed to overcome this problem by inducing a random walk in temperature space, such that broad sampling is achieved at high temperature to avoid kinetic traps at low temperature (Hansmann and Okamoto 1999; Sugita and Okamoto 1999). In REMD, multiple simulations are run, each at a different temperature. A random walk in temperature space is achieved by periodically attempting to swap the conformations at two neighboring temperatures. The probability of accepting a swap is

$$P(i \rightarrow j) = \min\left(1, e^{(\beta_j - \beta_i)(U_i - U_j)}\right)$$

where $P(i \rightarrow j)$ is the probability of transitioning from temperature $T(i)$ to temperature $T(j)$, β_i is $1/(kT_i)$, with k the Boltzmann constant, and U_i is the potential energy of the conformation at $T(i)$. Thus, the detailed balance condition is satisfied.

Our simulations used the AMBER 03 potential for nucleic acids (Chen and Pappu 2007). The GROMACS molecular dynamics simulation package (Hess et al. 2008) was used due to its speed. The RNA molecule was solvated in a water box with any solute atom at least 10 Å away from the wall of the box. Sodium cations (Na^+) were added to neutralize the system. The simulation system was minimized using a steepest descent algorithm, followed by a 100 psec MD simulation applying a position restraint potential to the RNA heavy atoms. All simulations were run with constant NVT by coupling to a Nose-Hoover thermostat (Hoover 1985) with a coupling constant of 0.02 $psec^{-1}$. A cutoff of 10 Å was used for nonbonded interactions. Long-range electrostatic interactions were treated with the Particle-Mesh Ewald (PME) method (Darden et al. 1995). Nonbonded pair lists were updated every 10 steps with an integration step size of 2 fsec in all simulations. All bonds were constrained using the LINCS algorithm (Hess et al. 1997).

Five representative RNA systems were chosen from our initial RNA data set to generate the decoy structures. For each RNA system, 20 1-nsec position restraint simulations were performed with each heavy atom constrained to its initial position by a harmonic potential,

$$E = k(r - r_0)^2$$

where k, the force constant, equals 0, 10, 20, ... 90, 100, 200, 300, ... 900 respectively in each of the 20 simulations. In addition, 1-nsec REMD simulations are also performed for each RNA system. The temperature list was roughly exponentially distributed, with 50 temperatures ranging from 285 to 592K.

Normal-mode decoys were generated using our normal-mode perturbation method (Summa and Levitt 2007). Quasi elastic modes of each native structure are computed using just the single-bond torsion angles as degrees of freedom. The potential energy and kinetic energy matrices, V and T, were generated by numerical differentiation (Levitt et al. 1985) using the Tirion-like (Tirion 1996) energy function:

$$U_{ij} = 90 \cdot \left(r^2 - R^2\right)^2 \Big/ \left\{R^4 \cdot \left[aR^4 + (1-a)r^4\right]\right\}$$

where r is the separation of atoms i and j, R is the constant separation of the same atoms in the native structure, and the constant a is set to 0.2. Using this function, the energy and its first derivative are zero at the native state ($r = R$) and the second derivative is always positive and decreases as R^{-6}. Eigenvectors derived in torsion-angle space involve combinations of torsion angles that do not move atoms along straight lines in Cartesian coordinates. In the past (Summa and Levitt 2007), we used the shifts of atomic positions caused by a very small shift along a torsional mode denoted as v_{ij} for the ith Cartesian coordinate of the jth mode. These shift vectors are not necessarily orthogonal in Cartesian coordinates $\left(\sum v_{ik} v_{ij} \neq 0\right)$ so that adding components from such vectors can fail to span the subspace of K modes properly. We dealt with this problem by using the actual torsion angle changes associated with each normal mode. The angle changes for the 20 lowest modes were added together with random amplitudes and then used to perturb the native structure in torsion angle space. This gave a structure that still had stereochemically correct bond lengths and angles but could have bad contacts. The RMSD of this structure was recorded, as was the number of bad contacts. The procedure was then repeated 50,000 times using

different random amplitudes whose magnitude slowly increased so as to ensure that we generated decoys with a uniform range of RMSD values up to some specified maximum value. The RMSD values of the structures were used to count how many structures fell in RMSD bins 0.1 Å wide. A required number of structures in each bin was specified (10 here) and the maximum RMSD value was set to 5 Å so that we aim to have 50 bins each containing 10 decoys. The 50,000 tries generated 100-fold more decoys and we chose the smallest RMSD with the smallest number of bad contacts in each bin. This gave ~500 decoys that were then refined by Encad energy minimization to ensure that none of the decoys would be easy to discriminate due to bad contacts.

The FARNA decoys used in this study were obtained from http://www.stanford.edu/~rhiju/data.html, and described in detail in the corresponding FARNA article (Das and Baker 2007).

Scoring with Rosetta RNA

Scoring of RNA decoys using the Rosetta scoring function was conducted with the Rosetta 3.0 package (http://www.rosettacommons.org). The addition of hydrogen atoms to native structures often introduces steric clashes. Therefore, for consistency, all hydrogen atoms were removed (decoys and native). In most cases, the terminal 5′-phosphate was missing, and was inserted based on ideal RNA base geometry. To relieve strain and steric clashes from the addition of the phosphate, only the corresponding bases were allowed to move in a simple implicit solvent minimization procedure (AMBER 99 force-field [Wang et al. 2000]; Generalized Born electrostatics [Tsui and Case 2000] with inverse Debye-Huckel length of 0.19 $Å^{-1}$; maximum of 500 steps implemented in Nucleic Acid Builder [Macke and Case 1997]). Such a short and constrained minimization procedure adequately removes steric clashes introduced by the terminal phosphate, while appropriately maintaining the RNA fold. Atomic movements introduced were minimal in all cases, with small RMSD changes. These same structures were also used in our KB potential scoring.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work is part of the "GNAPI Associate Team." We thank the INRIA Équipe Associée program for financial support. This work was supported by a National Institutes of Health award (GM041455) to M.L., a Human Frontiers in Science Program grant to M.L., and a Hong Kong University Grants Council award (RPC10SC03) and RGCHKUST6/CRF/10 to X.H. A.Y.L.S. acknowledges support from the Agency for Science, Technology and Research (A*STAR), Singapore. The authors acknowledge support from NSF award CNS-0619926 for computer resources.

Received November 15, 2010; accepted March 1, 2011.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Batey RT, Rambo RP, Doudna JA. 1999. Tertiary motifs in RNA structure and folding. *Angew Chem Int Ed Engl* **38:** 2326–2343.

Berman H, Henrick K, Nakamura H, Markley JL. 2007. The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35:** D301–D303.

Brion P, Westhof E. 1997. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct* **26:** 113–137.

Chen AA, Pappu RV. 2007. Parameters of monovalent ions in the AMBER-99 forcefield: Assessment of inaccuracies and proposed improvements. *J Phys Chem B* **111:** 11884–11887.

Chopra G, Summa CM, Levitt M. 2008. Solvent dramatically affects protein structure refinement. *Proc Natl Acad Sci* **105:** 20239–20244.

Chopra G, Kalisman N, Levitt M. 2010. Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins* **78:** 2668–2678.

Darden T, York D, Pedersen L. 1995. A smooth particle mesh Ewald potential. *J Chem Phys* **103:** 3014–3021.

Das R, Baker D. 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci* **104:** 14664–14669.

Das R, Karanicolas J, Baker D. 2010. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* **7:** 291–294.

DeLano WL. 2002. *The PyMOL user's manual*. DeLano Scientific, San Carlos, CA.

Dima RI, Hyeon C, Thirumalai D. 2005. Extracting stacking interaction parameters for RNA from the data set of native structures. *J Mol Biol* **347:** 53–69.

Flores SC, Altman RB. 2010. Turning limited experimental information into 3D models of RNA. *RNA* **16:** 1769–1778.

Frellsen J, Moltke I, Thiim M, Mardia KV, Ferkinghoff-Borg J, Hamelryck T. 2009. A probabilistic model of RNA conformational space. *PLoS Comput Biol* **5:** e1000406. doi: 10.1371/journal.pcbi.1000406.

Gesteland RF, Cech T, Atkins JF. 2006. *The RNA world: The nature of modern RNA suggests a prebiotic RNA*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Hamelryck T, Manderick B. 2003. PDB file parser and structure class implemented in Python. *Bioinformatics* **19:** 2308–2310.

Handl J, Knowles J, Lovell SC. 2009. Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics* **25:** 1271–1279.

Hansmann UH, Okamoto Y. 1999. New Monte Carlo algorithms for protein folding. *Curr Opin Struct Biol* **9:** 177–183.

Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. 1997. LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* **18:** 1463–1472.

Hess B, Kutzner C, Van der Spoel D, Lindahl E. 2008. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* **4:** 435–447.

Hofacker IL. 2009. RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* **Chapter 12:** Unit12 12. doi: 10.1002/0471250953.bi1202s26.

Hoover W. 1985. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* **31:** 1695–1697.

Huang ES, Subbiah S, Tsai J, Levitt M. 1996. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J Mol Biol* **257:** 716–725.

Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15:** 189–199.

Levitt M, Sander C, Stern PS. 1985. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol* **181:** 423–447.

Lu H, Skolnick J. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44:** 223–232.

Macke TJ, Case DA. 1997. Modeling unusual nucleic acid structures. In *Molecular modeling of nucleic acids* (ed. NB Leontis, J SantaLucia Jr), Vol. 682, pp. 379–393. American Chemical Society.
Mathews DH. 2006. Revolutions in RNA secondary structure prediction. *J Mol Biol* **359:** 526–532.
McLachlan G, Peel D. 2000. *Finite mixture models*. Wiley, New York.
Murray LJ, Arendall WB III, Richardson DC, Richardson JS. 2003. RNA backbone is rotameric. *Proc Natl Acad Sci* **100:** 13904–13909.
Murthy VL, Rose GD. 2003. RNABase: An annotated database of RNA structures. *Nucleic Acids Res* **31:** 502–504.
Neal RM. 1998. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat* **9:** 249–265.
Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452:** 51–55.
R Development Core Team 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rasmussen CE. 2000. The infinite Gaussian mixture model. In *Advances in neural information processing systems* (ed. SA Sollaet al.), Vol. 12, pp. 554–560. MIT Press, Boston.
Reeder J, Hochsmann M, Rehmsmeier M, Voss B, Giegerich R. 2006. Beyond Mfold: Recent advances in RNA bioinformatics. *J Biotechnol* **124:** 41–55.
Saenger W. 1984. *Principles of nucleic acid structure*. Springer-Verlag, New York.
Samudrala R, Moult J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* **275:** 895–916.
Shapiro BA, Yingling YG, Kasprzak W, Bindewald E. 2007. Bridging the gap in RNA structure prediction. *Curr Opin Struct Biol* **17:** 157–165.
Sharma S, Ding F, Dokholyan NV. 2008. iFoldRNA: Three-dimensional RNA structure prediction and folding. *Bioinformatics* **24:** 1951–1952.
Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213:** 859–883.
Stombaugh J, Zirbel CL, Westhof E, Leontis NB. 2009. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* **37:** 2294–2312.
Sugita Y, Okamoto Y. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **314:** 141–151.
Summa CM, Levitt M. 2007. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci* **104:** 3177–3182.
Sykes MT, Levitt M. 2005. Describing RNA structure by libraries of clustered nucleotide doublets. *J Mol Biol* **351:** 26–38.
Tinoco I Jr, Bustamante C. 1999. How RNA folds. *J Mol Biol* **293:** 271–281.
Tirion MM. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* **77:** 1905–1908.
Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. 2003. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **53:** 76–87.
Tsui V, Case DA. 2000. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **56:** 275–291.
Wang J, Cieplak P, Kollman PA. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?. *J Comput Chem* **21:** 1049–1074.
Zhang C, Liu S, Zhou H, Zhou Y. 2004. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci* **13:** 400–411.
Zhou H, Zhou Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11:** 2714–2726.
Zirbel CL, Sponer JE, Sponer J, Stombaugh J, Leontis NB. 2009. Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res* **37:** 4898–4918.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.

CHARACTERIZING RNA ENSEMBLES FROM NMR DATA WITH KINEMATIC MODELS

Rasmus FONSECA, Dimitar V. PACHOV, Julie BERNAUER, Henry VAN DEN BEDEM. In *Nucleic Acids Research*, 42(15):9562-72, 2014.

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku707

Characterizing RNA ensembles from NMR data with kinematic models

Rasmus Fonseca[1,2,3], Dimitar V. Pachov[4], Julie Bernauer[1,2] and Henry van den Bedem[5,*]

[1]AMIB Project, INRIA Saclay-Île de France, 1 rue Honoré d'Estienne d'Orves, Bâtiment Alan Turing, Campus de l'École Polytechnique, 91120 Palaiseau, France, [2]Laboratoire d'Informatique de l'École Polytechnique (LIX), CNRS UMR 7161, École Polytechnique, 91128 Palaiseau, France, [3]Department of Computer Science, University of Copenhagen, Nørre Campus, Universitetsparken 5, DK-2100 Copenhagen, Denmark, [4]Department of Chemistry, Stanford University, 333 Campus Dr., Stanford, CA 94305, USA and [5]Joint Center for Structural Genomics, Stanford Synchrotron Radiation Lightsource, Stanford University, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

Received May 18, 2014; Revised July 18, 2014; Accepted July 21, 2014

ABSTRACT

Functional mechanisms of biomolecules often manifest themselves precisely in transient conformational substates. Researchers have long sought to structurally characterize dynamic processes in noncoding RNA, combining experimental data with computer algorithms. However, adequate exploration of conformational space for these highly dynamic molecules, starting from static crystal structures, remains challenging. Here, we report a new conformational sampling procedure, KGSrna, which can efficiently probe the native ensemble of RNA molecules in solution. We found that KGSrna ensembles accurately represent the conformational landscapes of 3D RNA encoded by NMR proton chemical shifts. KGSrna resolves motionally averaged NMR data into structural contributions; when coupled with residual dipolar coupling data, a KGSrna ensemble revealed a previously uncharacterized transient excited state of the HIV-1 trans-activation response element stem–loop. Ensemble-based interpretations of averaged data can aid in formulating and testing dynamic, motion-based hypotheses of functional mechanisms in RNAs with broad implications for RNA engineering and therapeutic intervention.

INTRODUCTION

Non-coding ribonucleic acids (ncRNAs) mediate important cellular processes. Transfer RNA and ribosomal RNA are essential functional components in protein synthesis (1). Short interfering RNAs (siRNAs) and microRNAs (miRNAs) are the effector molecules in RNA interference, the process of silencing expression of specific genes in cells, and hold great promise as therapeutics (2,3). Riboswitches regulate gene expression by adopting alternative, 3D conformations in response to binding events (4). In RNA nanomedicine, these and other functional RNAs are fused into nanoparticles for targeted intracellular delivery, silencing cancer and infectious disease-specific genes (5).

RNA molecules are highly dynamic, sampling a wide range of conformational rearrangements to interact with binding partners and perform their function (6,7). The native ensemble of biomolecules, i.e. the set conformational states the molecule adopts *in vivo*, cannot be observed directly. Solution-state nuclear magnetic resonance (NMR) spectroscopy can probe the RNA conformational landscape at timescales ranging from picosecond to seconds or longer, often providing detailed evidence of dynamically interchanging, sparsely populated substates (8,9). Structurally characterizing conformational substates would offer tremendous potential for uncovering functional mechanisms (10), particularly for riboswitches (11), or predicting molecular interactions of RNA sub-units, such as in nanostructures (12). However, resolving motionally averaged NMR measurements into constituent, structural contributions that represent key features of the data remains extremely challenging (13).

The value of analyzing NMR spectroscopy data guided by a conformational ensemble has long been recognized (14,15). Conformational diversity for RNA ensemble analyses is often provided by sophisticated molecular dynamics simulations (16,17). Long trajectories with specialized force fields on dedicated supercomputers are required to adequately sample conformational space, limiting ensemble analyses to modestly-sized RNA molecules (18). Here, we present an efficient conformational sampling procedure, Kino-geometric sampling for RNA (KGSrna), which can report on ensembles of RNA molecular conformations orders of magnitude faster than MD simulations. KGSrna represents an RNA molecule with rotatable, single bonds as degrees-of-freedom and groups of atoms as rigid bod-

*To whom correspondence should be addressed. Tel: +1 650 776 5365; Fax: +1 650 926 3292; Email: vdbedem@stanford.edu

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku707

Characterizing RNA ensembles from NMR data with kinematic models

Rasmus Fonseca[1,2,3], Dimitar V. Pachov[4], Julie Bernauer[1,2] and Henry van den Bedem[5,*]

[1]AMIB Project, INRIA Saclay-Île de France, 1 rue Honoré d'Estienne d'Orves, Bâtiment Alan Turing, Campus de l'École Polytechnique, 91120 Palaiseau, France, [2]Laboratoire d'Informatique de l'École Polytechnique (LIX), CNRS UMR 7161, École Polytechnique, 91128 Palaiseau, France, [3]Department of Computer Science, University of Copenhagen, Nørre Campus, Universitetsparken 5, DK-2100 Copenhagen, Denmark, [4]Department of Chemistry, Stanford University, 333 Campus Dr., Stanford, CA 94305, USA and [5]Joint Center for Structural Genomics, Stanford Synchrotron Radiation Lightsource, Stanford University, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

Received May 18, 2014; Revised July 18, 2014; Accepted July 21, 2014

ABSTRACT

Functional mechanisms of biomolecules often manifest themselves precisely in transient conformational substates. Researchers have long sought to structurally characterize dynamic processes in non-coding RNA, combining experimental data with computer algorithms. However, adequate exploration of conformational space for these highly dynamic molecules, starting from static crystal structures, remains challenging. Here, we report a new conformational sampling procedure, KGSrna, which can efficiently probe the native ensemble of RNA molecules in solution. We found that KGSrna ensembles accurately represent the conformational landscapes of 3D RNA encoded by NMR proton chemical shifts. KGSrna resolves motionally averaged NMR data into structural contributions; when coupled with residual dipolar coupling data, a KGSrna ensemble revealed a previously uncharacterized transient excited state of the HIV-1 trans-activation response element stem–loop. Ensemble-based interpretations of averaged data can aid in formulating and testing dynamic, motion-based hypotheses of functional mechanisms in RNAs with broad implications for RNA engineering and therapeutic intervention.

INTRODUCTION

Non-coding ribonucleic acids (ncRNAs) mediate important cellular processes. Transfer RNA and ribosomal RNA are essential functional components in protein synthesis (1). Short interfering RNAs (siRNAs) and microRNAs (miRNAs) are the effector molecules in RNA interference, the process of silencing expression of specific genes in cells, and hold great promise as therapeutics (2,3). Riboswitches regulate gene expression by adopting alternative, 3D conformations in response to binding events (4). In RNA nanomedicine, these and other functional RNAs are fused into nanoparticles for targeted intracellular delivery, silencing cancer and infectious disease-specific genes (5).

RNA molecules are highly dynamic, sampling a wide range of conformational rearrangements to interact with binding partners and perform their function (6,7). The native ensemble of biomolecules, i.e. the set conformational states the molecule adopts *in vivo*, cannot be observed directly. Solution-state nuclear magnetic resonance (NMR) spectroscopy can probe the RNA conformational landscape at timescales ranging from picosecond to seconds or longer, often providing detailed evidence of dynamically interchanging, sparsely populated substates (8,9). Structurally characterizing conformational substates would offer tremendous potential for uncovering functional mechanisms (10), particularly for riboswitches (11), or predicting molecular interactions of RNA sub-units, such as in nanostructures (12). However, resolving motionally averaged NMR measurements into constituent, structural contributions that represent key features of the data remains extremely challenging (13).

The value of analyzing NMR spectroscopy data guided by a conformational ensemble has long been recognized (14,15). Conformational diversity for RNA ensemble analyses is often provided by sophisticated molecular dynamics simulations (16,17). Long trajectories with specialized force fields on dedicated supercomputers are required to adequately sample conformational space, limiting ensemble analyses to modestly-sized RNA molecules (18). Here, we present an efficient conformational sampling procedure, Kino-geometric sampling for RNA (KGSrna), which can report on ensembles of RNA molecular conformations orders of magnitude faster than MD simulations. KGSrna represents an RNA molecule with rotatable, single bonds as degrees-of-freedom and groups of atoms as rigid bod-

*To whom correspondence should be addressed. Tel: +1 650 776 5365; Fax: +1 650 926 3292; Email: vdbedem@stanford.edu

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

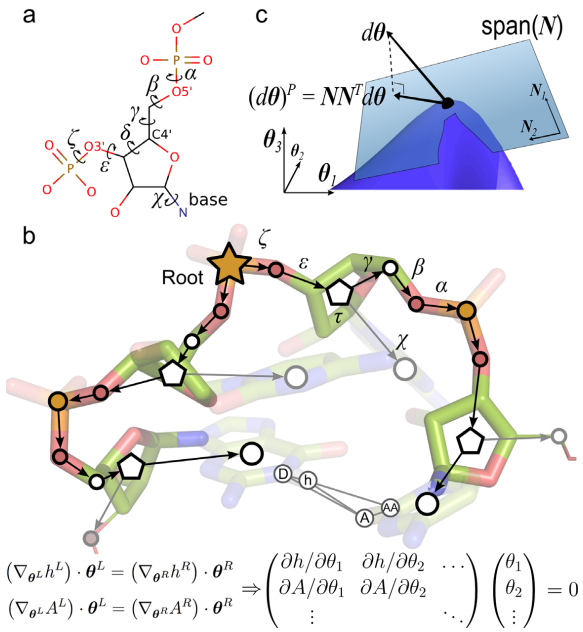


Figure 1. Kinematic representation of RNA. (**a**) A single nucleotide of RNA with its torsional degrees-of-freedom. (**b**) Edges in the directed spanning tree encode n torsional degrees-of-freedom $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and vertices (circles) encode rigid bodies. Pentagons represent riboses, which have an additional internal degree-of-freedom governing their conformation (puckering). The hydrogen bond h-A closes a kinematic cycle, and is one of m distance constraints. As the position of the hydrogen atom h changes through perturbation of dihedral angles in the left branch of the tree, the new position of h should be matched by appropriate changes in the right branch, i.e. $(\nabla_{\theta^L} h^L) \cdot \boldsymbol{\theta}^L = (\nabla_{\theta^R} h^R) \cdot \boldsymbol{\theta}^R$. Similarly, a change in position of heavy atom A from the right tree should be matched by changes in the left tree. These instantaneous distance constraints define the $6m \times n$ Jacobian matrix $\boldsymbol{J}$. (**c**) A schematic representation of the subspace of conformational space defined by the closure constraints. The subspace (blue surface) is highly nonlinear, but can be locally approximated by its tangent space, the null-space of $\boldsymbol{J}$ (translucent blue plane).

ies (Figure 1a). In this representation, non-covalent bonds form distance constraints, which create nested, closed rings (Figure 1b). Torsional degrees-of-freedom in a closed ring demand carefully coordinated changes to avoid breaking the non-covalent bond, which greatly reduces the conformational flexibility (19–22). The reduced flexibility from a network of nested, closed rings consequently deforms the biomolecule along preferred directions on the conformational landscape. In contrast to techniques based on explicit constraint counting (19,22), our new procedure projects degrees-of-freedom onto a lower-dimensional subspace of conformation space, in which the geometries of the non-covalent bonds are maintained exactly under conformational perturbation.

The dimensionality reduction additionally enables efficient exploration of conformational space and reduces the risk of overfitting sparse experimental data. Kino-geometric sampling of 3D RNA models can recover the conformational landscape encoded by proton chemical shifts in solution. Combined with NMR residual dipolar coupling (RDC) measurements, our procedure can automatically determine the size and weights of a parsimonious conformational ensemble that, provably, best agrees with the data. Our results can guide interpretation of proton chemical shift (23) or RDC data (16,17), and complement insights obtained from single, averaged models (24–26), ensembles resulting from experimentally guided modeling procedures (27–29), normal mode analysis (30), Monte Carlo simulations (31) or *de novo* tertiary structure prediction (32–34).

MATERIALS AND METHODS

A kinematic model for non-coding RNA

We encode a polynucleotide as a rooted, directed spanning tree, i.e. an acyclic graph that connects all vertices such that each one, except the root, has only one incoming, directed edge. Each vertex represents a rigid group of atoms (rigid body), and each edge represents a rotational (torsional) degree-of-freedom for the backbone (α, β, γ, δ, ε and ζ) or the N-glycosidic bond of the nucleoside (χ) (Figure 1a). We use the linear, branched structure of the polynucleotide to identify rigid bodies based only on knowledge of bond-flexibility. Initially each atom is a rigid body. Atoms that are (partially) double bonded have their rigid bodies merged and atoms with only one covalent neighbor are merged with their neighbor. Thus, rigid bodies are the largest conformational sub-units not containing internal rotational degrees-of-freedom.

Perturbing the torsional angle δ generally breaks the geometry of riboses. To efficiently sample ribose conformations we introduced a differentiable coordinate transformation τ for the angle δ, which maintains ideal geometry of the ribose when δ is perturbed (Supplementary Figure S1).

A vector $\boldsymbol{\theta} \in \mathbb{S}^n$, with $\mathbb{S}$ the unit circle, completely specifies a conformation for a molecule with n rotational degrees-of-freedom. Each hydrogen bond defines a closed ring or *kinematic cycle*. In this study, we consider hydrogen bonds between Watson–Crick (WC) pairs only. For any hydrogen bond, rotation along the h-A axis is allowed but no other distortion of the geometry is permitted (Figure 1b). Interatomic forces are implicitly encoded by rigid links between adjacent atoms and a long-range hard-sphere interaction potential based on van der Waals radii.

Conformational perturbation with constraints

Randomly perturbing a conformation would break the hydrogen bonds. We developed two complementary conformational sampling mechanisms that, in linear approximation, maintain distance constraints exactly. A *null-space* perturbation sensitively samples local neighborhoods of conformation space and a *rebuild* perturbation can rapidly explore more distant areas (20).

Null-space perturbation. A null-space perturbation of a conformation $\boldsymbol{\theta}$ projects an n-dimensional trial-vector onto the null-space of the Jacobian matrix $\boldsymbol{J}$. This $6m \times n$ matrix is defined by the instantaneous, or velocity, kinematic relation $d\boldsymbol{x} = \boldsymbol{J} d\boldsymbol{\theta}$, where $\boldsymbol{x}$ are the m 6D coordinates describing the end-effectors, i.e. the position and orientation

of the donor and acceptor atoms that define m closure constraints (20,22,35) (Figure 1b). The matrix $\boldsymbol{J}$ can be obtained as follows. If $h^{\mathrm{L,R}} : \boldsymbol{\theta}^{\mathrm{L,R}} \mapsto (h_x, h_y, h_z, h_\alpha, h_\beta, h_\gamma)$ is the map relating the degrees-of-freedom to the position and orientation of the hydrogen donor atom going around the left (L) or right (R) side of the cycle, then the constraint equations are $\left(\nabla_{\boldsymbol{\theta}^{\mathrm{L}}} h^{\mathrm{L}}\right) \cdot \boldsymbol{\theta}^{\mathrm{L}} = \left(\nabla_{\boldsymbol{\theta}^{\mathrm{R}}} h^{\mathrm{R}}\right) \cdot \boldsymbol{\theta}^{\mathrm{R}}$, i.e. infinitesimal displacements for h resulting from conformational changes from the left side of the cycle should match those from the right. Similar expressions hold for the acceptor atom A. This leads to six constraints per cycle. Thus, the entries of $\boldsymbol{J}$ contain the derivative of all hydrogen bond end-atom positions with respect to each degree-of-freedom (Figure 1b). The null-space of $\boldsymbol{J}$, i.e. the subspace spanned by vectors $d\boldsymbol{\theta}$ for which $\boldsymbol{J}d\boldsymbol{\theta} = \boldsymbol{0}$, and which leave the end-atom positions and orientations invariant, is generally $n-5m$-dimensional (Figure 1c). Vectors $d\boldsymbol{\theta}$ in this lower-dimensional space are *redundant* degrees-of-freedom that can be perturbed while maintaining distance constraints. Applying sufficiently small vectors from the null-space to a conformation will ensure that hydrogen-bond geometry is maintained. The right-singular vectors of the singular value decomposition $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ form a basis, $\boldsymbol{N}$, of the null-space of the Jacobian. A null-space perturbation projects a random trial-vector $\Delta\boldsymbol{\theta}$ onto the null-space, and adds it to the selected seed conformation: $\boldsymbol{\theta}_{\mathrm{new}} = \boldsymbol{\theta}_{\mathrm{seed}} + \boldsymbol{N}\boldsymbol{N}^{\mathrm{T}}\Delta\boldsymbol{\theta}$. In contrast to techniques based on Laman constraint counting (19,22), our null-space method does not rely on the molecular conjecture (36) to identify exactly all rigid and flexible substructures in the molecule.

Rebuild perturbation. The sampling step size of a null-space perturbation is limited by the linearized forward kinematics. A rebuild perturbation allows for a larger step size, and accommodates sampling of preferred ribose conformations. A rebuild perturbation randomly selects a backbone segment of up to two nucleotides not constrained by base pairing or stacking interactions, and breaks the O3′–P bond at the 3′-end.

A new τ-angle is sampled for each ribose in the segment according to a bimodal probability distribution

$$P(\tau) = 0.6N_\tau(-154^\circ, 11.5^\circ) + 0.5N_\tau(44.7^\circ, 17.2^\circ)$$

where N_τ denotes the normal distribution, with peaks at the C3′-endo and C2′-endo conformations (Supplementary Figure S1). Glycosidic angles, χ, in the segment are resampled to a random value. After resampling, all backbone torsions of the segment, except those in riboses, starting at the P–O5′ bond at the 5′-end, are used to reclose the O3′–P bond. Reclosing of the segment is performed by iteratively applying the Moore–Penrose inverse of the Jacobian matrix, which, in linear approximation, minimizes the distance to reclose the O3′–P bond as a function of the backbone torsions.

Sampling procedure

An overview of the sampling procedure is shown in Figure 2. KGSrna takes as input an initial conformation $\boldsymbol{\theta}_{\mathrm{init}}$, an exploration radius r_{init} and a set of canonical WC pairs to identify hydrogen bonds A(N3)–U(H3) and G(H1)–C(N3) as distance constraints. WC pairs are obtained from the RNAView program (37). Next, it grows a pool of conformations by repeatedly perturbing either $\boldsymbol{\theta}_{\mathrm{init}}$ or a previously generated seed conformation, $\boldsymbol{\theta}_{\mathrm{seed}}$, in the pool that is within r_{init} C4′ root mean square deviation (RMSD) of $\boldsymbol{\theta}_{\mathrm{init}}$. The seed conformation is selected by first generating a completely random conformation $\boldsymbol{\theta}_{\mathrm{random}}$. Next, the conformation closest to $\boldsymbol{\theta}_{\mathrm{random}}$ from all previously generated conformations that are within a spherical shell of random radius from $\boldsymbol{\theta}_{\mathrm{init}}$ and width $r_{\mathrm{init}}/100$ is selected as $\boldsymbol{\theta}_{\mathrm{seed}}$, and then $\boldsymbol{\theta}_{\mathrm{random}}$ is discarded. This guarantees that samples in sparsely populated regions within the exploration sphere are more likely to be chosen as seeds and that the sample population will distribute widely. A rebuild perturbation of two free nucleotides or a null-space perturbation is then performed at a 10/90 rate. To characterize the apical loop of HIV-1 TAR, see below, the C2′-endo peak was up-shifted by 60° to oversample non-helical ribose conformations. A null-space perturbation can start from a seed generated by a rebuild perturbation or vice versa, allowing detailed exploration of remote parts of conformation space. The trial-vector is scaled down to ensure no torsional change exceeds 0.1 radians = 5.7°. If no clashes between atoms were introduced in generating a new sample, $\boldsymbol{\theta}_{\mathrm{new}}$, it is accepted in the conformation pool. An efficient grid-indexing method is used for clash detection by overlapping van der Waals radii (38). The van der Waals radii were scaled by a factor 0.5.

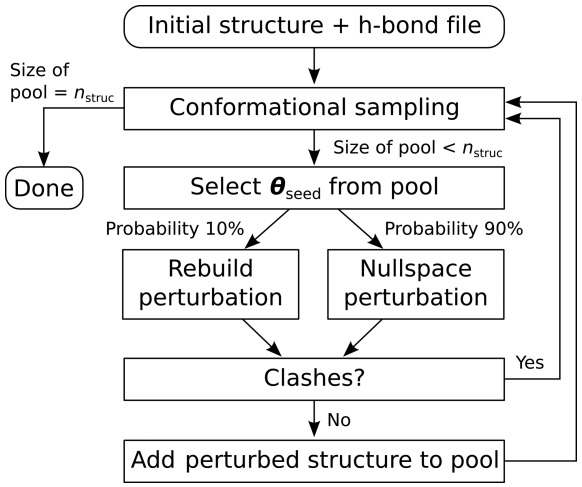


Figure 2. A flowchart of the KGSrna sampling algorithm. KGSrna takes as inputs an initial conformation and a file of hydrogen bonds A(N3)–U(H3) and G(H1)–C(N3) as distance constraints. Next, a pool of conformations is initialized with the input structure and then grown by repeatedly perturbing a randomly selected conformation from the pool with a rebuild or null-space perturbation at a 10/90 rate. If no clashes between atoms were introduced in the perturbed conformation, it is added to the pool. The procedure is repeated until a desired number n_{struct} of conformations is obtained.

Benchmark set

A benchmark set of 60 ncRNAs was compiled from the Biological Magnetic Resonance Bank (BMRB) (39) by selecting all single-chain RNA molecules that contain more

than 15 nucleotides, were solved with NMR spectroscopy, and have measured chemical shift data available (Supplementary Table S1). Some molecules were removed to ensure that the edit-distance between the sequences of any pair was at least 5. To ensure uniformity in the benchmark set, the HPO_4 group was removed at the 5′-end of the molecules 2LUB, 2LHP, 2AU4 and 2L94.

Back calculating NMR properties from ensembles

Observed chemical shifts were obtained from the BMRB. Chemical shifts were back calculated using the software NUCHEMICS (40). A flowchart of the procedure used to obtain back calculated chemical shifts is shown in Supplementary Figure S1(a). RDC data were back calculated using the program PALES (41), with command line option -inD loopshort.tab -bestFit -pdb $file -H, where loopshort.tab contained measured RDC data for nucleotides 30–35 published by Dethoff *et al.* (42).

Symmetric Kullback–Leibler divergence

The symmetrized KL divergence $D_{\mathrm{KL}}(P\|Q)$ represents the difference between two discrete probability distributions P and Q (43). $D_{\mathrm{KL}}(P\|Q)$ is often interpreted as the average number of bits of information lost when approximating one distribution by the other. It is defined as:

$$D_{\mathrm{KL}}(P\|Q) = \frac{D'_{\mathrm{KL}}(P\|Q) + D'_{\mathrm{KL}}(Q\|P)}{2}$$

where $D'_{\mathrm{KL}}(P\|Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i)$.

To obtain the discrete probability distributions for a set of measured or predicted CS values, we created histograms with 30 bins in the range between the predicted mean ± four standard deviations. Small pseudocounts were added to empty bins to avoid a vanishing denominator in the divergence calculation, after which the distribution was renormalized.

Fitting $^1D_{\mathrm{CH}}$ residual dihedral coupling data

We calculated 20 000 Monte Carlo samples each starting from models 1 to 10 in the NMR bundle of wild type HIV-1 TAR with pdb id 1ANR. The sampling included the full, 29 nucleotide long models. Sampling was biased toward generically pairing the C30 and A35 bases and the U31 and G34 bases using a Metropolis criterion (Supplementary Figure S1(b)). A newly sampled structure was accepted into the pool of samples with a probability $\min(1, e^{-\Delta d})$, where Δd is the change from the seed to the new structure, i.e. $d = \min_{p\in P(30,35)} |2A - |p|| + \min_{p\in P(31,34)} |2A - |p||$ and P(A, B) is the set of all vectors between any charged hydrogen in base A and any hydrogen acceptor in base B or vice versa. The seed selection was also modified. Instead of selecting a seed from the pool based on RMSD to the starting model it was based on Δd from the starting model. For each of the 10 sets of 20 000 samples, we calculated RDCs with the software PALES. For each set, we fitted an optimal ensemble with a new algorithm we developed, called rdcFit, using the following quadratic program:

$$\min_{w} \left\| {}^1\boldsymbol{D}^{\mathrm{o}}_{\mathrm{CH}} - \sum_i w_i \, {}^1\boldsymbol{D}^{\mathrm{c}}_{\mathrm{CH},i} \right\|^2$$
$$\text{s.t. } w_i \geq 0 \text{ for all } i$$
$$0 \leq \sum_i w_i \leq 1,$$

where ${}^1\boldsymbol{D}^{\mathrm{o}}_{\mathrm{CH}}$ is a vector of observed RDCs and ${}^1\boldsymbol{D}^{\mathrm{c}}_{\mathrm{CH},i}$ a set of vectors of back-calculated RDCs for nucleotides 30–35. The vector $\boldsymbol{w}^{\mathrm{T}}$ is the fitted variable that simultaneously determines the optimal size of the ensemble and the relative weights of its members, under the constraint that the weights sum to unity. Unlike stochastic or heuristic optimization procedures, a constrained quadratic fit deterministically identifies the global optimum of the fitted parameters, i.e. both the size and weights of the ensemble.

To further optimize a transitional, ES-like state identified by the fitting procedure, 20 000 additional KGSrna Monte Carlo samples were calculated starting from the ES-like state. Our Metropolis criterion was restricted to hydrogen bonds C30(N4)–A35(N1), C30(N3)–A35(N6), U31(O2)–G34(N1) and U31(N3)–G34(O6) in this step to improve the ES-like state.

Molecular dynamics simulations

All molecular dynamics simulations were performed with GROMACS 4.6.1 and the CHARMM 27 all-atom force field. The KGSrna structure was solvated in an octahedral unit cell with TIP3 water molecules and electrostatically neutralized by 28 Na ions (concentration 0.05 M and no ions within 6 Å of any RNA atom). The resulting system contained 13 054 water molecules and 40 120 atoms. For each of the 15 runs, the simulation system was minimized using a steepest descent algorithm, followed by a 150 ps MD equilibration applying a position restraint potential to the RNA heavy atoms. All simulations were run for 100 ns with constant NPT at a temperature of 300 K by coupling to a Nose–Hoover thermostat with a coupling constant of 0.6 ps and a Parrinello–Rahman barostat at a reference pressure of 1.0 bar. The van der Waals cutoff was set to 10 Å with a switching distance of 9 Å and the short-range electrostatics was set to 12 Å. Long-range electrostatic interactions were treated with the Particle-Mesh Ewald (PME) method. Non-bonded pair lists were updated every 10 steps with an integration step size of 2 fs in all simulations. All bonds were constrained using the LINCS algorithm.

Availability

The KGSrna software is available at http://smb.slac.stanford.edu/~vdbedem.

RESULTS

Accessing the native ensemble

Efficient exploration of the native ensemble requires broad and uniform sampling. Sampled conformations need to diffuse away quickly from an initial structure, while simultaneously at least one member of the native ensemble should

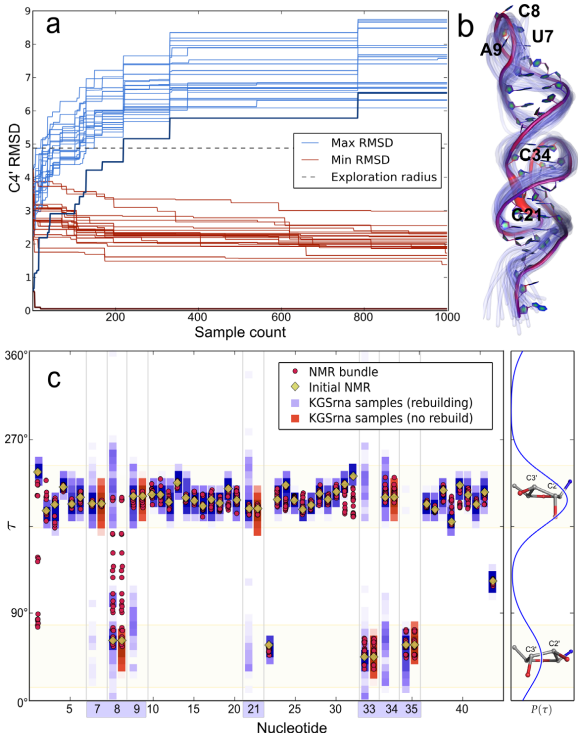


Figure 3. Sampling properties of 1000 KGSrna samples illustrated with the TYMV pseudoknot. Sampling was started from model one of the NMR bundle with pdb id 1A60, and the sampling-radius was set to 4.6 Å. (**a**) The evolution of the minimum (red curves) and maximum (blue curves) C4′ RMSD to each structure in the NMR bundle. Bold curves correspond to the starting structure. (**b**) The backbone of the initial structure indicating varying degrees of ridigity for torsional degrees-of-freedom. A thicker and more red-shifted backbone indicates higher variances for those degrees-of-freedom. The backbones of 25 randomly chosen KGSrna samples are shown in translucent blue to reflect flexibility. (**c**) Distributions of the τ-angles in the NMR-bundle and the KGSrna samples. Ribose conformations of the 1000 samples are displayed vertically as normalized color-coded histograms with a bin-width of 1.8°. Rebuild perturbations recover the full range of τ-angles in the NMR bundle for free nucleotides (highlighted on the x-axis), as shown by residue 8. The distribution from which τ-angles are sampled is shown on the right. The large peak corresponds to C3′-endo conformations and the smaller one to C2′-endo conformations.

be found close to any sampled conformation. We first validated these characteristics for KGSrna on a benchmark set of 60 RNA molecules with an average length of 30 nucleotides (nt) determined by NMR spectroscopy from the BMRB (Supplementary Table S1). We view the NMR bundle as structural representatives of a native ensemble, i.e. a 'synthetic' ensemble. For each RNA molecule, we created a set of 1000 samples starting from the first model of the NMR bundle. The exploration radius was fixed at the largest pairwise RMSD in each NMR bundle. Creation of 1000 samples took on average 372 s. Figure 3a shows the evolution of the C4′ RMSD between 1000 KGSrna samples and the NMR bundle of the 44 nt pseudoknotted acceptor arm of the transfer RNA-like structure of turnip yellow mosaic virus (TYMV). The procedure quickly expands its sampling neighborhood from the starting model to exceed its preset exploration radius of 4.9 Å (Figure 3a bold blue line). Within ∼300 sampling steps, the distance to the starting model reaches a limiting distance of ∼1.5 Å beyond the exploration radius, a trend that was consistent across our benchmark set (Supplementary Table S1). The maximum RMSD to each member of the NMR bundle of the sample set, represented by the blue lines, ranges from 6.1 to 8.7 Å. These trends indicate that samples diffuse quickly and uniformly through the synthetic ensemble, away from the starting model and consistently equidistant to all members of the NMR bundle.

Each member of the NMR bundle is closely approximated by a KGSrna sampled conformation as shown by the minimum RMSD (red lines). For the TYMV pseudoknot, the minimum RMSD also converges rapidly to a limiting value of ∼2 Å. This trend is also consistent across our benchmark set, with an average minimum RMSD of 1.2 Å (Supplementary Table S1). These metrics compare favorably to RNA 3D structure prediction algorithms (44), which suggests that our procedure can be used as an efficient conformational search procedure to further refine *ab initio* structures.

Regions of the molecule that are highly constrained by hydrogen bonds are difficult to deform, which is intrinsic in our kinematic representation by sharply reduced trial deviations after projecting into the null-space. In Figure 3b we color-coded the backbone of the TYMV pseudoknot to reflect the degree of rigidity for each backbone degree-of-freedom. The acceptor arm of TYMV contains two loops. Loop I (C21–U24), which spans the major groove, is clearly identifiable with highly flexibly degrees-of-freedom and loop II (U33–A35), which spans the minor groove, also stands out. The T loop, at the 3′-end, is somewhat less flexible. The variance in atom positions is illustrated by 25 randomly chosen KGSrna samples, shown in translucent blue. While conformational heterogeneity appears to be concentrated around the 3′- and 5′-end of the molecule, it originates primarily from the three least constrained backbone regions, loops I and II, and the T loop.

We then examined if rebuilding segments while sampling preferred ribose conformations would accurately represent ribose conformations of the NMR bundles. While C3′-endo to C2′-endo conformational transitions are rare in double strand regions, they are expected to occur more frequently in loop regions (45). In the benchmark set only 4.3% of all nucleotides occur with both C3′-endo and C2′-endo ribose conformations in the NMR bundle, but among unconstrained loop residues there are 35%. Ribose conformations can have important long-range structural effects, changing helical conformations and playing critical roles in binding events (46).

Unsurprisingly, rebuilding of unconstrained segments results in broader sampling of the ribose conformation. Starting from the first model of the NMR bundles, only nine out of 196 ribose conformations (4.8%) with both C3′-endo and C2′-endo are fully recovered using null-space perturbations. In contrast, rebuild perturbations recover all but four ribose conformations (98%). These four are all in less common conformations such as O4′-endo or C1′-endo. Figure 3c shows the range of τ-angles sampled by each residue

from the TYMV pseudoknot obtained from null-space perturbations only (magenta squares) and with rebuild perturbations enabled (blue squares). Without rebuild perturbations, sampled τ-angles remain close to their initial values obtained from the first model of the NMR conformational ensemble (yellow diamonds). For nucleotide C8 for example, null-space perturbations are unable to recover the full range of τ-angles in the NMR bundle (red circles), but rebuild perturbations do.

KGSrna recovers proton chemical shifts

Chemical shifts are time-averaged measurements on conformational ensembles at sub-millisecond timescales (23). Non-exchangeable ^{1}H chemical shifts (CS) predicted directly from RNA 3D structural models are generally in excellent agreement with those reported from experiments in the BMRB. Experimental ^{1}H CS are widely available, are sensitive to conformational changes and have aided in structurally characterizing conformational substates (47). Researchers have combined measured ^{1}H CS for proteins with structure prediction algorithms that use a database of structural fragments to determine atomically detailed *de novo* conformations (48). Das *et al.* recently established that proton chemical shifts can aid structure prediction algorithms in distinguishing decoys from a native state in RNAs (26).

Here, we regard measured ^{1}H CS on a benchmark set of 3D RNA structures as time- and ensemble-averaged distributions over the conformational landscape. We examined the ability of KGSrna to sample native dynamical ensembles that recover sugar (H1′) and nucleobase (H2, H5, H6 and H8) CS distributions for unconstrained (non-helical) and WC paired (helical) regions. We used the well-calibrated program NUCHEMICS (40) to predict ^{1}H CS from our 3D RNA structures.

KGSrna enables broad sampling to identify sparsely populated substates, while maintaining conformational distributions similar to those measured. Figure 4a shows the distribution of measured and predicted ^{1}H CS for helical (top row) and non-helical (bottom row) regions for each proton type over the whole benchmark set. Figure 4b shows the location of the probes. For helix backbone and base protons, the medians of the distributions are virtually identical. This suggests that, on average, our kinematic representation of RNA results in an unbiased exploration of the conformational landscape encoded in the measured proton chemical shifts.

For helical and non-helical regions, aggregate and individual (Supplementary Figure S1) sampling distributions of ^{1}H CS obtained with KGSrna are visually similar to the distributions obtained from experimental measurements. To further compare similarities between the measured chemical shift distributions P^{M} and the predicted distribution P^{KGSrna}, we calculated the symmetrized Kullback–Leibler (KL) divergences $D_{\mathrm{KL}}(P^{\mathrm{M}} \| P^{\mathrm{KGSrna}})$ of P^{KGSrna} from P^{M} and compared those to the KL divergences $D_{\mathrm{KL}}(P^{\mathrm{M}} \| P^{\mathrm{init}})$ and $D_{\mathrm{KL}}(P^{\mathrm{M}} \| P^{\mathrm{NMR}})$ for our benchmark set (Figure 4c; Supplementary Table S1; Materials and Methods). The distributions P^{init} and P^{NMR} are the predicted distributions calculated from the first model of the NMR bundle only and the full NMR bundle. Both P^{KGSrna} and P^{NMR} deviate from P^{M}, in part owing to weighted motional averaging of the measured shifts. Similarities between the KGSrna predicted distributions and the measured distributions exceeded those of the predicted distributions from the first model. The distribution of chemical shifts contributed by helical regions is expected to diverge less widely than that contributed from non-helical regions, owing to restrained conformational diversity. However, KGSrna predicted chemical shifts for both regions to a similar, accurate degree compared to a single starting model or the full NMR bundle. In 58 out of 60 of cases, KGSrna improved agreement with the distribution of measured ^{1}H CS in non-helical regions (58 out of 60 for helical regions too) compared to the distribution calculated from the first model. The average KL divergence reduction was 37% (42% for helical regions (Supplementary Table S1, Figure 4c). This suggests that KGSrna is able to diverge from the starting model, and explores beyond a local neighborhood of conformational space. In addition, in 70% of cases KGSrna improved agreement with the distribution of measured ^{1}H CS in non-helical regions (58% for helical regions) compared to the distribution calculated from the full NMR bundle. Predictions for non-helical regions were improved by our rebuilding procedure, conceivably resolving structural disorder inadequately represented by the NMR bundle (40). The similarities between predicted and measured distributions suggest that a simple kinematic model with constraints samples the conformational landscape according to the same distribution as RNA in solution.

We then asked how accurately just a single KGSrna sample could recover measured chemical shifts. The error between measured and predicted CS is attributable to measurement errors and systematic errors in prediction. Additionally, measured chemical shifts are a weighted motional average. We therefore regarded the NMR 3D conformer that best agrees with measured chemical shifts as a benchmark of predictive value.

We calculated the RMSD ($\mathrm{RMSD}_{\mathrm{CS}}$) between the measured and predicted chemical shifts for all proton types for each 3D model in the NMR bundles and in the KGSrna sample sets (Supplementary Table S1). The minimum $\mathrm{RMSD}_{\mathrm{CS}}$(M, KGSrna) ranges from 0.17 to 0.54 ppm (mean 0.30 ppm) and the minimum $\mathrm{RMSD}_{\mathrm{CS}}$(M, NMR) from 0.16 to 0.53 ppm (mean 0.30 ppm). A recent study observed a mean minimum weighted $\mathrm{RMSD}_{\mathrm{CS}}$ of 0.23 ppm (ranging from 0.16 to 0.35 ppm) for an ensemble of 8000 conformers obtained from molecular dynamics simulations for four RNAs, but the proton chemical shifts were weighted to favor those that better agreed with measured values (23). In 80% of cases in our benchmark set, the $\mathrm{RMSD}_{\mathrm{CS}}$ of the best KGSrna conformer is lower than that of the best conformer identified from the NMR bundle (Figure 4d). The average improvement over the starting model is 18% (p-value < 0.01), and in some cases exceeds 40%. As proton chemical shifts can discriminate a native state, this result suggests that a simple kinematic representation yields a powerful conformational search algorithm.

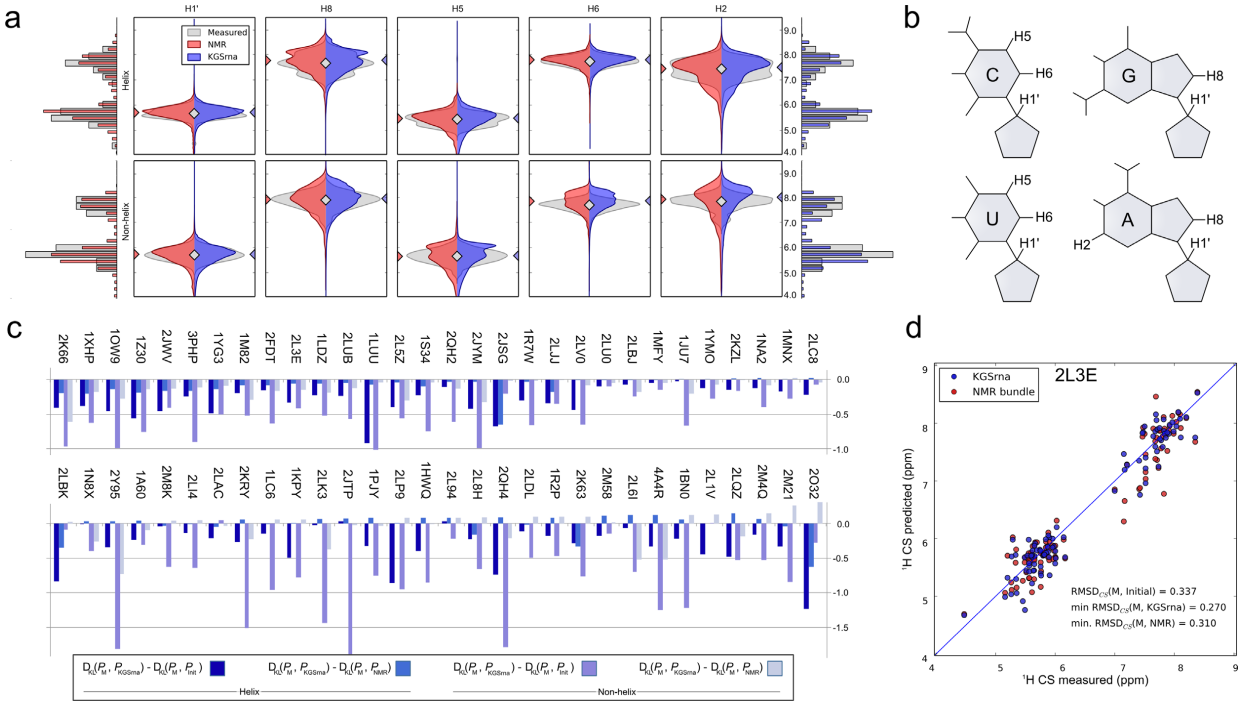


Figure 4. Agreement between measured ^{1}H chemical shifts and those back calculated from KGSrna and NMR 3D structures. (**a**) Depicted chemical shift values are aggregated by proton type in helical (top) and non-helical regions (bottom). The discrete distributions were smoothed with a Gaussian kernel density estimator (bandwidth $n^{-0.2}$ where n is number of data points) for easier visualization. Measured values were taken from the BMRB and KGSrna samples and NMR bundle values were back calculated using NUCHEMICS. Marginal distributions are shown as histograms with bin-widths of 0.275 ppm. (**b**) The symmetric Kullback–Leibler divergence indicates the degree of similarity of two distributions and is calculated for the marginal distributions of measured-to-KGSrna, measured-to-initial and measured-to-NMR. The differences between measured-to-KGSrna, measured-to-initial and measured-to-NMR are shown in the bar plot. A negative value indicates better agreement of the ensemble of KGSrna 3D structures with measured values than its comparison 3D structures. (**c**) Predicted ^{1}H chemical shifts calculated from the 3D structures from the KGSrna ensemble and the NMR bundle compared to the measured values of the 32nt P2a-J2a/b-P2b (helix-bulge-helix) of human telomerase RNA (pdb id 2L3E). The data points are expected to lie along a 45° line if measured ^{1}H CS are accurately predicted.

KGSrna reveals a hairpin loop excited state from $^1D_{CH}$ HIV-1 TAR data

The 5′-end of the human immunodeficiency virus type-1 (HIV-1) transcript contains a 59-nucleotide trans-activation response element (TAR) stem–loop (49). In the ground state, HIV-1 TAR binds human cyclin T1 and viral trans-activator protein Tat that activate and enhance transcription of the HIV-1 genome (42,50–52). The HIV-1 TAR apical hairpin loop plays a key role in binding Tat. Available structures for the HIV-1 TAR apical loop exhibit significant conformational differences, which indicate that the loop is highly flexible. However, a full atomic characterization of the structure and dynamics of the HIV-1 TAR hairpin loop remains elusive. Al-Hashimi *et al.* recently proposed a two-state model (ground and excited state, GS and ES) of the apical HIV-1 TAR hairpin loop from NMR $R_{1\rho}$ relaxation dispersion measurements and mutagenesis (53). Their study suggested formation of a $U_{31}G_{32}G_{33}G_{34}$ tetraloop in the ES, with a non-canonical closing base-pair C30–A35.

RDCs report the amplitude of motions that reorient C-H and N-H bond vectors on the sub-millisecond time-scale. Experimentally observed RDCs are a weighted average of all conformational substates. 'Sample-and-select' strategies, which rely on generating a large number of samples from which a subset is selected that best explains experimental data, have previously led to insights into conformational dynamics and functional mechanisms in X-ray crystallography and NMR data (54–56).

KGSrna was designed to explore correlated conformational variability resulting from nested, closed rings. The degrees-of-freedom of the HIV-1 TAR hairpin apical loop participate in the nested, closed rings formed by the canonical base-pairs in the stem. To test if KGSrna can structurally characterize conformational substates of the loop guided by RDC data, we calculated 20 000 samples each starting from the full, 29 nucleotide models 1 to 10 in the NMR bundle with pdb id 1ANR of free HIV-1 TAR. To enable structural characterization of the dynamics leading to the ES, we biased our sampling toward broad, non-specific conformational pairing of C30–A35 and U31–G34. A Metropolis criterion skewed the sample set to include favorable interactions of any charged hydrogen in base A with any hydrogen acceptor in base B (Materials and Methods). For each of the 200 000 samples, we back-calculated RDCs with the program PALES (41). PALES accurately calculates the overall alignment of the RNA molecule (18). From each batch of

20 000, we then determined a weighted ensemble that optimally explained the experimentally observed RDCs using a new constrained quadratic fit algorithm (rdcFit) that we adapted from an application we previously developed for X-ray crystallography applications (qFit) (54,55).

This procedure identified a 10-member, weighted ensemble from the sample set starting from model seven in the NMR bundle that agrees extremely well with experimentally observed RDC values (Figure 5a). The coefficient of determination between observed $^1\boldsymbol{D}_{\mathrm{CH}}$ values and those predicted from the weighted ensemble equals 0.98. The predicted values of the ensemble accurately reflect the mobility of riboses and nucleobases, with $^1\boldsymbol{D}_{\mathrm{CH}}$ small in magnitude indicating elevated mobility (Figure 5b). The RMSD between observed and predicted $^1\boldsymbol{D}_{\mathrm{CH}}$ values is 1.55 Hz, below the experimental error of 2–4 Hz (16,18).

Our ensemble characterizes disparities in mobility between nucleotides in exquisite atomic detail, consistent with the RDC data (Figure 5a inset). In our ensemble nucleobases U31, G32 and A35 are most mobile, with motions indicating looping in and out. Small magnitudes of U31 $^1\boldsymbol{D}_{\mathrm{C6H6}}$, G32 $^1\boldsymbol{D}_{\mathrm{C8H8}}$, and A35 $^1\boldsymbol{D}_{\mathrm{C8H8}}$ experimental values support this interpretation (Figure 5b). G34 exhibits a more limited range of motion in the ensemble, consistent with larger values of G34 $^1\boldsymbol{D}_{\mathrm{C8H8}}$. G32 is more mobile than G33, and is looped out for all members of the ensemble. Stacking interactions reported in MD simulations between U31 and G32 (42) or G32 and G33 (49) are not represented in our ensemble. A previously reported and experimentally confirmed base triple (A22–U40)U31 (47) was not observed in our ensemble.

In the conformation of our ensemble most closely exhibiting features attributed to the GS, we confirmed the formation of a stabilizing cross-loop WC bp C30·G34 (49) (Figure 5c, left). Nucleobases A35 and U31 are looped out in this conformation, while G33 is in front of the loop, possibly interacting with C30 and G34. A second conformation of our ensemble exhibits features closely associated with a transition to the ES, suggesting that C30 and A35 are poised to adopt a reverse wobble pair, with hydrogen bonds C30(N4)–A35(N1) and C30(N3)–A35(N6). G34 is positioned to adopt a GU wobble pair with U31 through hydrogen bonds U31(O2)–G34(N1) and U31(N3)–G34(O6) (57) (Figure 5c, right).

The G34 glycosidic angle in our ensemble is (high) anti, ranging from −110.2° to −83.7°. In the GS of our ensemble, G34 adopts an anti base (−93.2°), while it is high anti (−83.7°) in the transition to ES. In the formation of UUCG tetraloops, it is common for the guanine to loop out to accommodate a transition from anti to syn (58). Experimental $^1\boldsymbol{D}_{\mathrm{C8H8}}$ data do not appear to support a similar large amplitude motion of the G34 base when adjusting from anti to syn. Instead, our ensemble suggests that G34 gently readjusts to accommodate A35 looping in.

To confirm this intermediate state toward the ES, we generated an additional 20 000 samples starting from this conformation, instructing KGSrna to further optimize the CA reverse wobble and the GU wobble pairs with the Metropolis criterion (Materials and Methods). In the model with most ideal hydrogen-bond geometry between these bases, we observe ribose conformations suggesting that C30 and A35 are adopting a C3′-endo conformation, continuing the A-form helical stem from bp C29·G36. To examine if the ES is kinetically accessible from this intermediate state, we started 15 independent, 100 ns molecular dynamics (MD) simulations (see 'Materials and Methods' section).

Consistent with the transient character of the ES, the CA reverse wobble pair and GU wobble pair were maintained for 20–65 ns in duration for 4 out of 15 or 27% of the MD simulations. In the remaining simulations, pairings did not occur or were short-lived. Figure 4d shows the evolution of the distances of the hydrogen bonds between the CA pair and the GU pair for the MD simulation that maintained pairing for nearly 65 ns. The GU pair interacts more weakly than the CA pair, with U31 staggered toward the apex of the loop (Supplementary Figure S2). G34 forms additional hydrogen bonds G34(N2)–U31(O4′) and G34(N1)–C30(O2′) that stabilize the ES (Supplementary Figure S2). In the independent MD simulations, looping out of U31 generally disrupts its pairing with G34. Subsequently, the reverse wobble CA pair is disrupted. The helical conformation is extended in the ES; the riboses of G34 and A35 largely adopt C3′-endo conformations in the simulation. The riboses of C30 and U31 adopt a C2′-endo conformation for the duration of U31–G34 pairing, after which C30 adopts a C3′-endo conformation (Supplementary Figure S3). G32 and G33 intermittently stack during the simulation (Figure 5d inset). To our knowledge, this is the first time sustained and simultaneous pairing of C30–A35 and U31–G34 observed in MD simulations of HIV-1 TAR.

DISCUSSION

Molecular dynamics simulations can often provide new and highly detailed insight into specific, atomic interactions and functional mechanisms of biomolecules. By contrast, recent advances suggest that random sampling algorithms, coupled with knowledge-based potentials (33) and/or sparse experimental data (26), are better suited to provide broad exploration of the conformational landscape. Our analysis demonstrates that conformational ensembles of non-coding RNAs in solution can be accessed from efficiently sampling coordinated changes in rotational degrees-of-freedom that preserve the hydrogen bonding network. Compared to exploring the conformational landscape with molecular dynamics simulations, our highly simplified structural representation obtains similar agreement between measured and predicted chemical shifts from fewer samples and unweighted RMSD. Coordinated changes enforced by the kinematic representation deform the molecule along preferred directions on the conformational landscape, overlapping with those avoiding hydrogen bond dissociation. These intrinsic constraints on deformation enable our procedure to efficiently probe the conformational diversity resulting from equilibrium fluctuations of the ensemble, suggesting that a kinematic representation is capable to encode the dominant forces within a folded polynucleotide corresponding to sub-millisecond, RDC time scales.

Our procedure directly encodes rigidity of RNA molecules. By analyzing how flexibility propagates through amino acids in room temperature X-ray diffraction data we previously established that 3D networks related to func-

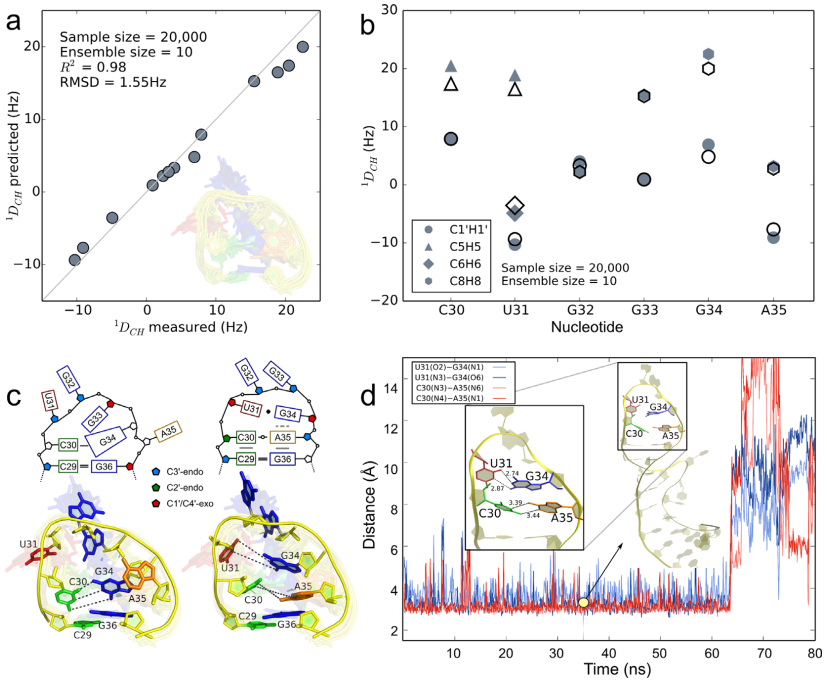


Figure 5. Structural characterization of conformational substates of the apical Tat binding loop of HIV-1 TAR. (**a**) Paired measured RDCs for apical loop nucleotides and those predicted from a 10-member weighted ensemble (inset) obtained from fitting 20 000 KGSrna samples to measured RDCs with a quadratic program. The data points are expected to lie along a 45° line if measured RDCs are accurately predicted. The coefficient of determination for the predicted RDCs equals 0.98. (**b**) Observed (solid symbols) and predicted (open symbols) RDCs for apical loop nucleotides. Smaller magnitudes for RDCs generally indicate more angular mobility in the bond vectors. (**c**) Schematic of the GS (top panel, left) and the ES (top panel, right) corresponding to the 3D structures closest to the GS and the ES in the 10-member ensemble. The bottom left panel shows the GS highlighted in the ensemble, with the other members translucent in the background. The bottom right panel shows the ES identified from biasing the sampling toward pairing C30–A35 and U31–G34. (**d**) Time evolution of the hydrogen bond distances between reverse wobble pair C30–A35 (blue colors) and GU wobble pair U31–G34 (red colors) in the ES of HIV-1 TAR for 80 ns of the molecular dynamics trajectory. The distances shown are between heavy donor and acceptor atoms, sampled every 100 ps. Along the trajectory, the apical loop maintains a helical structure (inset at 35 ns) until, at 65 ns, pairing of U31–G34 and subsequently C30–A35 is disrupted.

tional mechanisms partition protein molecules (55). These networks can provide important mechanistic insights into binding events and the role of allostery in activation. While a purely kinematic model does not suffice to determine strain, i.e. the deformation of a biomolecule due to stress, kinematics can elucidate long-range effects of locking or unlocking degrees-of-freedom through mutations and altered non-covalent bonds.

Combined with experimental data, KGSrna enables structural biologists to quickly formulate and test hypothesis about conformational dynamics, and offers tremendous potential for uncovering functional mechanisms. Our integrative analysis of HIV-1 TAR, linking structural experimental data and relaxation dispersion data with advanced computational algorithms, enabled us to identify an intermediate state that relaxes to the ES. However, the detailed structure of ground and evanescent excited states and their precise transitional mechanisms remain unresolved. While other researchers have posited a U31–G34 reverse wobble base-pair based on analogy with UUCG tetraloops and downfield shifted chemical shifts, we find that the anti-G34 base suggests a staggered U31–G34 wobble pair in the excited state. A mechanism analogous to the formation of UNCG tetraloops proposed for the HIV-1 TAR hairpin ES is not supported by our structural analysis of RDC data (53).

The set of feasible conformations for the apical loop of HIV-1 TAR is huge. While a convex quadratic fit of predicted to experimentally observed RDCs provably determines the global optimum, the quality of that optimum is limited by conformational sampling. The fitted ensemble approximates true substates, which, averaged, constitute the NMR measurements. While it is tempting to associate fractional contributions with population lifetimes, we should expect those to compensate for any conformational inaccuracies. A more direct correspondence between fractional contributions and population lifetimes will require bounding conformational space and/or additional data (54,55).

Our results suggest that diffusive motions of RNA are restricted to a lower-dimensional subspace of conformation space. As secondary structure prediction algorithms for RNA have matured greatly, this insight can have important implications for the efficiency and accuracy of search methods in 3D structure prediction and biomolecular docking applications. These methods rely on coarse-grained representations, which are often derived heuristically (31). By

contrast, our null-space procedure naturally reduces the dimensionality of the system. It automatically partitions the molecule into 'free' and 'cycle' degrees-of-freedom, of which only the latter require coordinated changes.

NMR relaxation dispersion experiments can provide highly detailed insight into transient, sparsely populated substates, but high energetic barriers frequently prevent access through molecular dynamics simulations. Our method provides a widely applicable, new avenue to uncover RNA conformational diversity from a variety of data sources. Combined with mutagenesis, our new approach can be used to relate motion to function with implications for RNA engineering and drug design.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank Sybren Wijmenga for providing a copy of the software NUCHEMICS. We are grateful to Kurt Wüthrich, James Fraser and R. Bryn Fenwick for stimulating discussions and careful reading of the manuscript. J.B. acknowledges access to the HPC resources of TGCC (Très Grand Centre de calcul du CEA) under the allocation t2013077065 made by GENCI (Grand Equipement National de Calcul Intensif). This work is part of the ITSNAP Associate Team. We thank the Inria Équipe Associée program for financial support.

FUNDING

U.S. National Institute of General Medical Sciences Protein Structure Initiative [U54GM094586]; SLAC National Accelerator Laboratory LDRD (Laboratory Directed Research and Development) [SLAC-LDRD-0014-13-2 to H.v.d.B.]. Funding for open access charge: SLAC National Accelerator Laboratory LDRD (Laboratory Directed Research and Development) [SLAC-LDRD-0014-13-2].
Conflict of interest statement. None declared.

REFERENCES

1. Nissen,P., Hansen,J., Ban,N., Moore,P.B. and Steitz,T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
2. Dorsett,Y. and Tuschl,T. (2004) siRNAs: applications in functional genomics and potential as therapeutics. *Nat. Rev. Drug. Discov.*, **3**, 318–329.
3. Cooper,T.A., Wan,L. and Dreyfuss,G. (2009) RNA and disease. *Cell*, **136**, 777–793.
4. Tucker,B.J. and Breaker,R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.
5. Zhou,J., Shu,Y., Guo,P., Smith,D.D. and Rossi,J.J. (2011) Dual functional RNA nanoparticles containing phi29 motor pRNA and anti-gp120 aptamer for cell-type specific delivery and HIV-1 inhibition. *Methods*, **54**, 284–294.
6. Leulliot,N. and Varani,G. (2001) Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry*, **40**, 7947–7956.
7. Kim,H., Abeysirigunawarden,S.C., Chen,K., Mayerle,M., Ragunathan,K., Luthey-Schulten,Z., Ha,T. and Woodson,S.A. (2014) Protein-guided RNA dynamics during early ribosome assembly. *Nature*, **506**, 334–338.
8. Bothe,J.R., Nikolova,E.N., Eichhorn,C.D., Chugh,J., Hansen,A.L. and Al-Hashimi,H.M. (2011) Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. *Nat. Methods*, **8**, 919–931.
9. Buck,J., Ferner,J.A.N., Wacker,A., Urtig,B.F. and Schwalbe,H. (2011) Mapping the landscape of RNA dynamics with NMR spectroscopy. *Acc. Chem. Res.*, **44**, 1292–1301.
10. Zhang,Q., Stelzer,A.C., Fisher,C.K. and Al-Hashimi,H.M. (2007) Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, **450**, 1263–1267.
11. Lipfert,J., Das,R., Chu,V.B., Kudaravalli,M., Boyd,N., Herschlag,D. and Doniach,S. (2007) Structural transitions and thermodynamics of a glycine-dependent riboswitch from Vibrio cholerae. *J. Mol. Biol.*, **365**, 1393–1406.
12. Guo,P. (2010) The emerging field of RNA nanotechnology. *Nat. Nanotechnol.*, **5**, 833–842.
13. Shi,X., Beauchamp,K.A., Harbury,P.B. and Herschlag,D. (2014) From a structural average to the conformational ensemble of a DNA bulge. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E1473–E1480.
14. Emani,P.S., Bardaro,M.F., Huang,W., Aragon,S., Varani,G. and Drobny,G.P. (2014) Elucidating molecular motion through structural and dynamic filters of energy-minimized conformer ensembles. *J. Phys. Chem. B*, **118**, 1726–1742.
15. Bouvignies,G., Vallurupalli,P., Hansen,D.F., Correia,B.E., Lange,O., Bah,A., Vernon,R.M., Dahlquist,F.W., Baker,D. and Kay,L.E. (2011) Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature*, **477**, 111–114.
16. Frank,A.T., Stelzer,A.C., Al-Hashimi,H.M. and Andricioaei,I. (2009) Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res.*, **37**, 3670–3679.
17. Borkar,A.N., De Simone,A., Montalvao,R.W. and Vendruscolo,M. (2013) A method of determining RNA conformational ensembles using structure-based calculations of residual dipolar couplings. J. Chem. Phys., **138**, 215103.
18. Salmon,L., Bascom,G., Andricioaei,I. and Al-Hashimi,H.M. (2013) A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed. *J. Am. Chem. Soc.*, **135**, 5457–5466.
19. Jacobs,D.J., Rader,A.J., Kuhn,L.A. and Thorpe,M.F. (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150–165.
20. van den Bedem,H., Lotan,I., Latombe,J.C. and Deacon,A.M. (2005) Real-space protein-model completion: an inverse-kinematics approach. *Acta Cryst.*, **D61**, 2–13.
21. Yao,P., Dhanik,A., Marz,N., Propper,R., Kou,C., Liu,G., van den Bedem,H., Latombe,J.-C., Halperin-Landsberg,I. and Altman,R.B. (2008) Efficient algorithms to explore conformation spaces of flexible protein loops. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 534–545.
22. Yao,P., Zhang,L. and Latombe,J.-C. (2012) Sampling-based exploration of folded state of a protein under kinematic and geometric constraints. *Proteins Struct. Funct. Bioinform.*, **80**, 2–43.
23. Frank,A.T., Horowitz,S., Andricioaei,I. and Al-Hashimi,H.M. (2013) Utility of 1H NMR chemical shifts in determining RNA structure and dynamics. *J. Phys. Chem. B*, **117**, 2045–2052.
24. Flores,S.C. (2013) Fast fitting to low resolution density maps: elucidating large-scale motions of the ribosome. *Nucleic Acids Res.*, **42**, e9.
25. van der Werf,R.M., Tessari,M. and Wijmenga,S.S. (2013) Nucleic acid helix structure determination from NMR proton chemical shifts. J. Biomol. NMR, **56**, 95–112.
26. Sripakdeevong,P., Cevec,M., Chang,A.T., Erat,M.C., Ziegeler,M., Schwalbe,H., Sigel,R.K.O., Turner,D.H. and Das,R. (2014) Consistent structure determination of noncanonical RNA motifs from 1H NMR chemical shift data alone. *Nat. Methods*, **11**, 8–13.
27. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
28. Ding,F., Lavender,C.A., Weeks,K.M. and Dokholyan,N.V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat. Methods*, **9**, 603–608.

29. Parisien,M. and Major,F. (2012) Determining RNA three-dimensional structures using low-resolution data. *J. Struct. Biol.*, **179**, 252–260.
30. Zacharias,M. and Sklenar,H. (2000) Conformational deformability of RNA: a harmonic mode analysis. *Biophys. J.*, **78**, 2528–2542.
31. Sim,A.Y.L., Levitt,M. and Minary,P. (2012) Modeling and design by hierarchical natural moves. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 2890–2895.
32. Shapiro,B.A., Yingling,Y.G., Kasprzak,W. and Bindewald,E. (2007) Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.*, **17**, 157–165.
33. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
34. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.
35. Burdick,J.W. (1989) On the inverse kinematics of redundant manipulators: characterization of the self-motion manifolds. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE Comput. Soc. Press, pp. 264–270.
36. Katoh,N. and Tanigawa,S. (2009) A proof of the molecular conjecture. In *Proceedings of the 25th Annual Symposium on Computational Geometry*. ACM Press, Aarhus, Denmark, pp. 296–305.
37. Yang,H., Jossinet,F. and Leontis,N. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
38. Halperin,D. and Overmars,M.H. (1998) Spheres, molecules and hidden surface removal. *Comp. Geom.-Theor. Appl.*, **11**, 83–102.
39. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
40. Cromsigt,J.A., Hilbers,C.W. and Wijmenga,S.S. (2001) Prediction of proton chemical shifts in RNA – their use in structure refinement and validation. *J. Biomol. NMR*, **21**, 11–29.
41. Zweckstetter,M. (2008) NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Prot.*, **3**, 679–690.
42. Dethoff,E.A., Hansen,A.L., Musselman,C., Watt,E.D., Andricioaei,I. and Al-Hashimi,H.M. (2008) Characterizing complex dynamics in the transactivation response element apical loop and motional correlations with the bulge by NMR, molecular dynamics, and mutagenesis. *Biophys. J.*, **95**, 3906–3915.
43. Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.
44. Laing,C. and Schlick,T. (2011) Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struct. Biol.*, **21**, 306–318.
45. Levitt,M. and Warshel,A. (1978) Extreme conformational flexibility of the furanose ring in DNA and RNA. *J. Am. Chem. Soc.*, **100**, 2607–2613.
46. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
47. Huthoff,H., Girard,F., Wijmenga,S.S. and Berkhout,B. (2004) Evidence for a base triple in the free HIV-1 TAR RNA. *RNA*, **10**, 412–423.
48. Shen,Y., Lange,O., Delaglio,F., Rossi,P., Aramini,J.M., Liu,G., Eletsky,A., Wu,Y., Singarapu,K.K., Lemak,A. *et al.* (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 4685–4690.
49. Kulinski,T., Olejniczak,M., Huthoff,H., Bielecki,L., Pachulska-Wieczorek,K., Das,A.T., Berkhout,B. and Adamiak,R.W. (2003) The apical loop of the HIV-1 TAR RNA hairpin is stabilized by a cross-loop base pair. *J. Biol. Chem.*, **278**, 38892–38901.
50. Aboul-ela,F., Karn,J. and Varani,G. (1996) Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic Acids Res.*, **24**, 3974–3981.
51. Tahirov,T.H., Babayeva,N.D., Varzavand,K., Cooper,J.J., Sedore,S.C. and Price,D.H. (2010) Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature*, **465**, 747–751.
52. Lu,H., Li,Z., Xue,Y. and Zhou,Q. (2013) Viral-host interactions that control HIV-1 transcriptional elongation. *Chem. Rev.*, **113**, 8567–8582.
53. Dethoff,E.Q., Petzold,K., Chugh,J., Casiano-Negroni,A. and Al-Hashimi,H.M. (2012) Visualizing transient low-populated structures of RNA. *Nature*, **491**, 724–728.
54. van den Bedem,H., Dhanik,A., Latombe,J.-C. and Deacon,A.M. (2009) Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. *Acta Cryst.*, **D65**, 1107–1117.
55. van den Bedem,H., Bhabha,G., Yang,K., Wright,P.E. and Fraser,J.S. (2013) Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat. Methods*, **10**, 896–902.
56. Fenwick,R.B., van den Bedem,H., Fraser,J.S. and Wright,P.E. (2014) Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E445–E454.
57. Varani,G. and McClain,W.H. (2000) The G·U wobble base pair. *EMBO Rep.*, **1**, 18–23.
58. Chen,A.A. and García,A.E. (2013) High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 16820–16825.

APPENDIX II

RÉSUMÉ EN FRANÇAIS

RÉSUMÉ

Ce travail s'intéresse au développement d'algorithmes, modèles et autres approches computationnelles pour la biologie structurale et ses applications : la thérapeutique et les nanotechnologies. Il traite en détails de trois sujets principaux : (i) les modèles gros-grain et l'apprentissage supervisé pour la prédiction d'interactions 3D entre biomolécules (amarrage), (ii) la mise en place et l'utilisation de potentiels statistiques pour l'évaluation de structures, (iii) le développement de méthodes inspirées de la cinématique pour l'étude de la dynamique des biomolécules. Ces problèmes sont abordés pour les protéines et les acides ribo-nucléiques (ARN) et sont la clé de nombreux problèmes en bioinformatique structurale. Complétée par des approches multi-échelles en développement pour l'échantillonnage de structures moléculaires, telles que l'échantillonnage par théorie des jeux ou l'apprentissage non-supervisé pour la classification de conformations, cette approche intégrée pourra permettre dans un avenir proche le developpement de méthodes efficaces pour l'analyse et le design de machines moléculaires.

A Modèles gros-grain et apprentissage supervisé pour l'amarrage

1 Les modèles de Voronoï gros-grain : une bonne représentation des complexes protéiques ?

Les protocoles d'amarrages comportent deux étapes principales successives : d'abord, une grande quantité de conformations putatives est générée (exploration), puis une fonction de score est utilisée pour les classer (*scoring*). Cette fonction de score doit prendre en compte à la fois la complémentarité géométrique et les propriétés physico-chimiques des molécules en interaction. C'est cette deuxième étape que j'ai principalement étudiée, à travers le développement de fonctions de scores rapides et fiables.

Pour obtenir de bons descripteurs des propriétés des protéines, nous construisons tout d'abord le diagramme de Voronoï de la structure 3D des protéines de manière efficace. Le diagramme de Voronoï, tout comme le diagramme de Laguerre, se sont avérés être de bons modèles de la structure 3D des protéines [BAJP05, BBAP09, Pou04]. En particulier, cette formalisation permet une bonne description des propriétés d'empilement des résidus à l'interface entre deux protéines. Ainsi, il est possible d'obtenir un ensemble de descripteurs à partir de mesures sur des complexes de structure connue et sur des leurres [BAJP05, BAJP07]. Il est certainement possible d'étendre cette méthode à d'autres types d'interactions et à d'autres types de complexes, en particulier les complexes protéine-ARN que nous avons également étudiés [GG14] (voir Figure R1, mais cette piste reste à explorer.

Pour permettre une construction simple et optimale des modèles de Voronoï et autres constructions géométriques de façon générique, nous avons développé une bibliothèque C++ à en-têtes seules. Cette

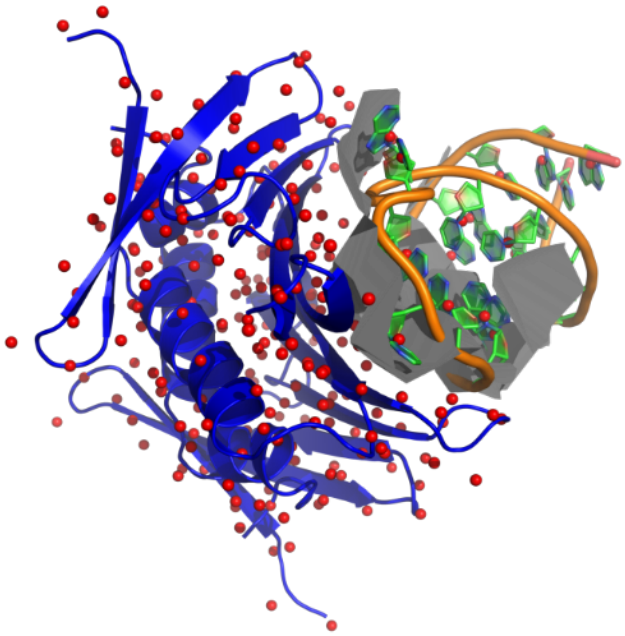


FIGURE R1 : Représentation gros-grain et interface de Voronoï pour le complexe entre une protéine d'enveloppe du phage PP7 et une épingle d'ARN (code PDB 2QUX) Le modèle de Voronoï capture les propriétés de l'interaction comme l'empilement des acides aminés et des nucléotides.

bibliothèque logicielle, appelée ESBTL (*Easy Structural Biology Template Library*) est utilisée par la suite CGAL (*Computational Geometry Algorithms Library*) pour la prise en charge de fichiers PDB [LCB10].

2 Prédiction fiable des complexes protéine-protéine et protéine-ARN par apprentissage supervisé

Les descripteurs issus de l'analyse des diagrammes de Voronoï sont utilisés en entrée d'une procédure d'apprentissage supervisé. Plusieurs approches ont été évaluées for le docking protéine-protéine : la régression logistique, les machines à vecteurs de support (SVM), et un algorithme génétique. Pour ce dernier, nous avons utilisé ROGER (*ROc based Genetic learnER*), algorithme qui maximise l'aire sous la courbe de ROC (*Receiver Operating Characteristic*). Les conformations putatives sont générées par deux algorithmes de docking differents puis classées. Dans la plupart des cas, le rang des solutions presque-natives est largement amélioré [BAJP07, BPAJ05].

Le succès de cette stratégie pour le problème de l'amarrage a également permis de répondre à une question biologique qui se pose souvent lors de la résolution expérimentale d'une structure. La connaissance de l'état d'oligomérisation d'une protéine est souvent essentielle à la compréhension de sa fonction et des mécanismes biologiques. Dans un cristal, chaque monomère de protéine est en contact avec beaucoup d'autres monomères dans le réseau, formant de nombreuses interfaces parmi lesquelles peu sont biologiquement intéressantes. Pouvoir différencier ces dimères "cristallins" des "dimères biologiques" est difficile. Le modèle de Voronoï permet toutefois de mettre en lumière des différences significatives

entre ces dimères, et nous a permis de mettre en place une méthode performante de discrimination entre interfaces spécifiques et non-spécifiques [BBR+08].

La construction de Voronoï, combinée à un choix astucieux de descripteurs, nous a permis de mettre en place et d'optimiser plusieurs fonctions de score. Pour pouvoir réaliser des expériences *in silico* à grande échelle, par exemple explorer l'interactome, il faut qu'une solution presque-native soit classée dans les dix premiers. Nous avons donc introduit une procédure de classement tirant parti des précédentes fonctions de score. Les rangs obtenus à l'aide de différents classifieurs (arbres de décision, règles et machine à vecteurs de support) sont combinés par filtrage collaboratif. L'évaluation de cette approche sur les ensembles du challenge international CAPRI (*Critical Assessment of PRediction of Interactions*) montre que la fonction ainsi obtenue permet d'enrichir considérablement les 100 premières structures en conformations presque-natives [BBAP11].

Les complexes protéine-ARN sont impliqués dans de nombreuses fonctions essentielles pour la cellule. Cependant la résolution expérimentale de leur structure est encore plus difficile que pour les protéines. Il est aussi souvent très difficile d'adapter les approches d'amarrage développées pour les complexes protéine-protéine : l'ARN est très flexible et peu de données expérimentales sont disponibles. Nous avons étendu le protocole de référence RosettaDock aux complexes protéine-ARN à l'échelle gros-grain et l'échelle atomique. À l'aide d'un algorithme génétique, après avoir extrait et nettoyé un jeu de données de référence, nous montrons que notre score RosettaDock est non seulement capable de discriminer entre complexes biologiques et leurres, mais peut aussi classer correctement les conformations proches de la solution. Notre approche s'est avérée efficace et robuste pour générer et identifier les structures biologiques sur les deux jeux d'essai publics à la fois sur les partenaires liés et non-liés. En plus d'avoir des performances comparables aux approches connues, voire meilleures, c'est la seule méthode qui permet une recherche multi-échelle optimisée permettant à terme un protocole entièrement flexible [GGFAB14].

Bien que les méthodes présentées ici permettent d'obtenir de bons résultats pour le docking rigide ou semi-flexible, la génération de conformations pour les structures très flexibles reste une question difficile qu'il faut envisager à l'aide de nouveaux algorithmes multi-échelle. L'amarrage de ces conformations va aussi certainement demander de disposer de meilleures méthodes de partitionnement de données.

B De la biophysique aux données : les potentiels statistiques pour l'ARN

Les molécules biologiques peuvent adopter différentes conformations. Pouvoir identifier les conformations correspondant à des fonction biologiques, par exemple pour prédire les interactions, est un problème difficile. C'est le cas tout particulièrement pour les ARN qui sont très flexibles. Les potentiels statistiques se sont avérés efficaces pour ce type de problème pour les protéines, par exemple pour la prédiction de structure. Pour l'ARN, nous avons développé des potentiels statistiques dérivables à partir d'un jeu de données de structures d'ARN que nous avons extrait et nettoyé. Ces potentiels ont été construits pour deux représentations : gros-grain et tout atome.

Pour pouvoir obtenir ces potentiels nous avons construits différents modèles de mélange. Les modèles obtenus par processus de Dirichlet [BHSL11] se sont révélés plus performants que ceux obtenus par estimation de densité ou espérance-maximisation [SSLB12].

Pour l'évaluation, nous nous sommes focalisés sur un aspect du problème de prédiction : l'identification d'une conformation native dans un jeu de conformations presque natives. Sur une grande quantité de données, obtenues à partir de trois méthodes différentes et indépendantes (données expérimentales et synthétiques), nous montrons, à l'aide de critères d'évaluation classiques et spécialement définis pour

cet usage, que nos potentiels permettent de distinguer la structure native mais aussi d'identifier les structures presque natives proches, même au niveau gros-grain. Les potentiels donnent de bons résultats et présentent de meilleures performances en discrimination de structures presque natives que l'une des meilleures méthodes disponibles. La fonction de potentiel obtenue est aussi très lisse par construction et permet la minimisation par descente de gradient, technique très largement utilisée dans les méthodes biophysiques classiques.

L'extension de ces potentiels à différents types de molécules et aux interactions pourrait être une option intéressante. La faible quantité de données expérimentales reste toutefois un problème majeur. Des développements ciblés (par exemple pour la reconnaissance par les anticorps) combinés à des modèles plus fins utilisant des méthodes de partitionnement pourrait être une piste d'exploration intéressante.

C Un modèle inspiré de la robotique : la cinématique inverse

L'analyse des données de RMN (Résonance Magnétique Nucléaire) fait appel à une grande diversité de conformations. Ces conformations sont en général obtenues par des simulations de dynamique moléculaire. De longues trajectoires sont obtenues grâce à des champs de force spécialisés sur des supercalculateurs (parfois dédiés) pour obtenir un échantillonnage de l'espace conformationnel. Ces analyses sont donc réservées à des molécules de taille modeste.

Pour s'affranchir de ces limitations, nous avons développé pour l'ARN une technique d'échantillonnage conformationnel performante : KGSrna (*Kino-Geometric Sampling*). Cette méthode permet d'obtenir des échantillons avec une rapidité supérieure de plusieurs ordres de grandeur par rapport à la modélisation moléculaire.

Dans ce modèle, une molécule d'ARN est représentée par des articulations, les liaisons simples, qui sont les degrés de liberté (torsion), et par des groupes d'atomes qui forment des corps rigides. Dans cette représentation, les liaisons non-covalentes forment des contraintes de distances qui créent des cycles imbriqués sur un arbre couvrant enraciné. Les degrés de liberté torsionnels à l'intérieur d'un cycle demandent d'effectuer des changements coordonnés pour éviter de rompre les liaisons non-covalentes, ce qui réduit de façon drastique la flexibilité conformationnelle.

Cette flexibilité, bien que réduite par le réseau de cycles, permet de déformer la molécule le long de directions privilégiées dans l'espace des conformations. Cette nouvelle procédure projette les degrés de liberté sur un sous-espace de dimension inférieure à celle de l'espace des conformations dans lequel la géométrie des liaisons non-covalentes est maintenue par pertubation. La réduction de la dimensionalité permet de plus l'exploration efficace de l'espace des conformations et limite le risque de surajustement (*overfitting*) des données. L'échantillonage de structures 3D d'ARN avec KGS permet de retrouver l'espace des conformations décrit par les déplacements chimiques en solution et aide considérablement l'interprétation des données [FPBv14].

La performance de cette approche, ainsi que sa nature intrinsèquement parallèle, en ferait une approche de choix pour traiter des complexes macromoléculaires sur des architectures parallèles.

D Perspectives

D'un point de vue biologique, disposer de techniques efficaces pour la modélisation structurale des protéines et des ARN nous permettrait d'étudier des problèmes importants par leurs aspects thérapeutiques.

Ces problèmes sont actuellement hors de portée des techniques *in vitro* mais aussi *in silico*. Par exemple, il serait possible de modéliser rapidement des assemblages aussi complexes que le ribosome, les capsides virales ou les complexes de réparation de l'ADN impliqués dans les cancers.

De nombreuses collaborations avec des expérimentalistes pourraient voir le jour suite au développement de ces méthodes, à la fois robustes et bien définies informatiquement et algorithmiquement. Ces collaborations permettraient de nombreuses avancées pour l'étude et le design de biomolécules ciblés pour des applications thérapeutiques (telles que le cancer ou les maladies virales) ou encore nanotechnologiques [Guo10]. Ceci implique des outils informatiques et un développement algorithmique à chaque étape du développement du projet. Nous disposons pour l'instant de modules pour la prédiction et l'affinement atomique de complexes protéine-protéine et protéine-ARN et travaillons à étendre ces modèles à plus grande échelle, en utilisant les modèles de cinématique inverse ou de théorie des jeux pour les gros assemblages.

Malgré les progrès des techniques expérimentales au cours de la dernière décennie, obtenir une description précise d'un phénomème biologique à l'échelle atomique reste difficile. L'augmentation de la quantité de données disponibles permettra sans aucun doute l'augmentation du développement de modèles construits à partir de ces données, pour lesquels des algorithmes efficaces seront développés. Des prédictions plus précises, hors de portée des modèles biophysiques actuels, pourront être ainsi proposées.