TOPICAL REVIEW - STATISTICAL PHYSICS AND COMPLEX SYSTEMS

RNA structure prediction: Progress and perspective

TOPICAL REVIEW — Statistical physics and complex systems

RNA structure prediction: Progress and perspective*

Shi Ya-Zhou(时亚洲), Wu Yuan-Yan(吴园燕), Wang Feng-Hua(王凤华), and Tan Zhi-Jie(谭志杰)†

Department of Physics and Key Laboratory of Artificial Micro- and Nano-structures of Ministry of Education, School of Physics and Technology, Wuhan University, Wuhan 430072, *China*

(Received 17 February 2014; revised manuscript received 15 April 2014; published online 23 May 2014)

Many recent exciting discoveries have revealed the versatility of RNAs and their importance in a variety of cellular functions which are strongly coupled to RNA structures. To understand the functions of RNAs, some structure prediction models have been developed in recent years. In this review, the progress in computational models for RNA structure prediction is introduced and the distinguishing features of many outstanding algorithms are discussed, emphasizing three-dimensional (3D) structure prediction. A promising coarse-grained model for predicting RNA 3D structure, stability and salt effect is also introduced briefly. Finally, we discuss the major challenges in the RNA 3D structure modeling.

Keywords: RNA structure prediction, secondary structure, three-dimensional (3D) structure, coarse-grained model

PACS: 87.14.gn, 87.15.bd, 87.15.bg, 87.15.Cc **DOI:** 10.1088/1674-1056/23/7/078701

1. Introduction

It is well known that the primary role of RNAs is to convert the genetic information stored in DNA into proteins. Rapid methods for sequencing DNA have revealed that over 75% of human genomic DNA is transcribed into RNA, even though only about 5% of the DNA codes are translated into proteins.[1] Within the past few years, many additional functions of non-coding RNAs, i.e., functional but not in the sense of being translated into protein, have been discovered, which indicates that RNAs play a more important role in the life process than that stipulated by the traditional consensus in molecular biology.[1,2] Besides the familiar non-coding RNAs such as the ribosomal RNAs and transfer RNAs, other non-coding RNAs have been discovered, including ribozymes, small-interface RNAs, and riboswitches, which are known, respectively, to catalyze mRNA splicing, control protein synthesis, and regulate transcription and translation.[3–7]

To become functional, RNAs generally adopt compact tertiary structures. Understanding these structural features not only gives insight into to the structure–function relationship but also aids drug design.[8] Until now, the experimental methods for measuring RNA three-dimensional (3D) structure include X-ray crystallography and NMR spectroscopy, both of which have been successfully employed to obtain high-resolution RNA 3D structures. However, due to the flexible nature of RNA structures and the high cost of the experimental methods, the number of RNA structures in the database is much less than that of RNA sequences. Therefore, computational modeling has become very important to obtain RNA structures.[9–12]

The RNA folding process is generally hierarchical, i.e., helical elements in a secondary structure are formed first, and then the compact tertiary structures are folded with the associations of RNA motifs, the formation of tertiary contacts, and possible minor rearrangements of secondary segments.[8,13,14] Due to the strong base-pairing and base-stacking interactions, RNA secondary structures are usually relatively stable. Therefore, RNA folding can be generally divided into two steps: RNA secondary structure folding and RNA tertiary structure folding. In the literature, RNA structure modeling is generally classified into two levels: two-dimensional (2D) secondary structure modeling and 3D structure modeling.

In this review, we will focus on the recent advances in RNA structure prediction. The main text is organized as follows. In Section 2, we will give a brief overview of the predictive algorithms for prediction of 2D RNA secondary structures. In Section 3, we will give a detailed introduction of the recent progress in modeling RNA 3D structure, including the knowledge-based models and physics-based models. At the end of Section 3, we will briefly introduce a predictive coarse-grained (CG) model, newly developed for predicting RNA 3D structure, stability and salt effects. Finally, we will discuss the challenges in modeling RNA 3D structure.

2. RNA secondary structure prediction

The RNA folding process is generally hierarchical, which means that local interactions occur first and are energetically stronger than tertiary interactions.[8,13,14] Therefore, the RNA

*Project supported by the National Natural Science Foundation of China (Grant Nos. 11074191, 11175132, and 11374234), the National Basic Research Program of China (Grant No. 2011CB933600), and the Program for New Century Excellent Talents of China (Grant No. NCET 08-0408).

†Corresponding author. E-mail: zjtan@whu.edu.cn

secondary structure provides a scaffold to its native 3D structure, and it can be predicted without the knowledge of tertiary interactions. Over the past three decades, many computational models have been developed for predicting RNA secondary structure, and these have been reviewed elsewhere.[13,15–17] The major models are listed in Table 1 along with the corresponding websites and references. These computational models can be roughly classified into two categories: comparative sequence-based methods and free energy-based methods. In the following, we will give a brief introduction to the models.

Table 1. The major algorithms for RNA secondary structure prediction.

Algorithms	Ref.	Features	Web page
RNAalifold	[18]	Compute minimum energy structures from aligned sequences using covariation analysis.	http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi
Pfold/PPfold	[19], [20]	Predict RNA secondary structures using phylogeny and auxiliary data.	http://daimi.au.dk/∼compbio/pfold/
MARNA	[21]	An approach for multiple alignments of RNAs with sequence/structure comparisons.	http://rna.informatik.uni-freiburg.de/MARNA/Input.jsp
Mfold	[22]	Search minimization free energy structures by dynamic programming algorithm.	http://mfold.rna.albany.edu/?q=mfold
RNAStructure	[23]	Predict RNA secondary structures as well as base pairs probabilities.	http://rna.urmc.rochester.edu/RNAstructureWeb/
RNAfold	[24]	Predict minimum free energy structures and base pair probabilities from single sequences.	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
Contrafold	[25]	Predict structures using discriminative training and feature-rich scoring.	http://contra.stanford.edu/contrafold/
Sfold	[26]	Predict structures by a statistical sampling based on Boltzmann-weighted ensemble.	http://sfold.wadsworth.org/cgi-bin/index.pl
RNAShapes	[27]	Select suboptimal shapes using a reduced representation of RNA structures.	http://bibiserv.techfak.uni-bielefeld.de/rnashapes/
Kinwalker	[28]	A heuristic approach to kinetic RNA folding.	http://www.bioinf.uni-leipzig.de/Software/Kinwalker/
MPGAfold	[29]	Predict folding pathways and functional substructure with genetic algorithm.	http://www-lmmb.ncifcrf.gov/users/bshapiro//mpgaFold/mpgaFold.html
Pknots	[30]	Pseudoknot structure prediction using a dynamic programming algorithm.	http://selab.janelia.org/software.html
ILM	[31]	An iterated loop matching algorithm for pseudoknots structure prediction.	http://www.cs.wustl.edu/∼zhang/projects/rna/ilm/
Hotknot	[32]	A heuristic algorithm for predicting RNA secondary structures including pseudoknots.	http://www.cs.ubc.ca/labs/beta/Software/HotKnots
PknotsRG	[33]	Fold medium-sized pseudoknots based on thermodynamics model.	http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/

Note: This table summarizes the major algorithms in secondary structure prediction, and provides a URL for downloading or using online for each algorithm.

2.1. Comparative sequence-based methods

Assuming that RNAs with similar sequences in different species have similar structures, the secondary structures of RNAs can be predicted by comparative sequence analysis. This logic was first used to predict the secondary structure of tRNA correctly based on the expert knowledge of RNAs.[9,34]

Since the incipient manual comparative sequence analysis methods highly depend on the knowledge of users and are time-consuming, several automatic algorithms[18–21,35–37] that use multiple sequences were recently developed to predict RNA secondary structures: RNAalifold,[18] Pfold/PPfold,[19,20] TurboFold,[36] RNAforester,[37] and MARNA.[21] These algorithms have been benchmarked for prediction speed and accuracy with the same data set[38] and have been reviewed elsewhere.[16] Here, we only give a brief overview.

The RNAalifold algorithm predicts the consensus structure for a set of sequences from a given alignment with the use of sequence covariations and thermodynamic parameters.[18] Pfold couples a stochastic context-free grammar to a phylogenetic analysis with a high accuracy in predictions,[19] and its multithreaded version, PPfold can predict the structures for large RNA alignments at practical timescales with the phylogenetic calculations and the inside-outside algorithm.[20] An iterative method, TurboFold can predict secondary structures for multiple RNA sequences with the free energy model and the information from the comparative analysis between sequences.[36] RNAforester[37] and MARNA[21] are algorithms that first fold the RNA sequences with the free energy-based methods and then align the predicted structures with special tools.

2.2. Free energy-based methods

The most popular methods are to predict RNA secondary structures from a single sequence by calculating the free energy, which is based on the thermodynamic hypothesis that the conformation with the lowest free energy is the native structure.[39]

2.2.1. Free energy minimization

Based on an empirical nearest-neighbor model and thermodynamic parameters obtained from experiments,[40] Mfold[22] is a very important algorithm for searching the RNA secondary structure with minimum free energy from a single sequence, with the use of the dynamic programming algorithm. Mfold can predict secondary structures for RNAs with an accuracy of $\sim 70\%$.[22] Similarly, RNAstructure[23] and RNAfold[24] were developed for predicting the structures with minimum free energies as well as calculating the equilibrium partition functions and base pairing probabilities.

However, because of the assumption of free energy additivity and the absence of measured thermodynamic parameters of non-canonical base pairs and sequence-dependent loops, the free energy computed from the nearest-neighbor model for a conformation is not accurate.[41] Consequently, some advanced methods have introduced statistical data or abstract shapes of known structures to more accurately sample RNA secondary structures.[25–27] The Sfold is a typical statistical algorithm developed by Ding and Lawrence.[26] With a partition function calculation, Sfold can sample the optimal secondary structure according to Boltzmann probability with a stochastic dynamic programming algorithm.[26] RNAshapes uses the concept of abstract shapes to calculate cumulative probabilities of structures belonging to the identified shapes, and integrates the shapes well with dynamic programming algorithms to efficiently predict RNA secondary structures.[27]

2.2.2. Suboptimal structure prediction

Some RNA sequences may adopt secondary structures that are partially determined by folding kinetics or may have more than one conformation.[42] The simplest method is to explicitly generate all possible structures from the sequence, based on the thermodynamic parameters, and to further give possible folding pathways with the master equation method.[42] However, it is generally impossible to enumerate all possible structures for large RNAs due to the huge number of possible structures for a sequence.[15] To predict the multiple kinetically-stable structures for large RNAs would require approximations at different element levels.[43]

Recently, some automated procedures for efficiently generating a diverse set of suboptimal structures for large RNAs were proposed.[28,29] The Kinwalker introduced a heuristic approach for folding kinetics of large RNAs by constructing secondary structures through a stepwise combination of locally optimal substructures, which can be calculated by dynamic programming.[28] Another method employs MPGAfold to predict possible folding pathways and structures of functional intermediates with the use of a massively parallel genetic algorithm.[29]

2.2.3. Pseudoknot structure prediction

An RNA pseudoknot, which is minimally composed of two helical segments connected by single-stranded regions or loops, plays diverse fundamental roles in the control of viral replication, in the structural organization of complex RNAs and in the self-cleaving ribozyme catalysis.[44] Since there are crosslinkings of pseudoknot loops and thermodynamic parameters for pseudoknot loops have not been experimentally derived, to accurately predict secondary structures of RNA pseudoknots is still challenging, especially for the complex non-H-type pseudoknots.[17,30–33,44–47]

To address the problem, Rivas and Eddy introduced a few pseudoknot-specific parameters to fold optimal pseudoknotted RNAs (100–200 nt) with a dynamic programming algorithm.[30] The ILM model uses a combination of thermodynamic and covariance information to predict RNA secondary structures, including pseudoknots through an iterated loop matching algorithm.[31] As a heuristic algorithm, HotKnots predicts the secondary structures of pseudoknots by assembling the energetically favorable substructures determined from the free energy model.[32] Similarly, FlexStem constructs RNA secondary structures with pseudoknots by adding maximal stems according to the free energy model.[45] The Kinefold algorithm predicts the structures of pseudoknots with topological and geometrical constraints through a long-time-scale RNA folding simulation,[46] which follows the stochastic closing and opening of individual RNA helices. In addition, the thermodynamic parameters of a pseudoknots loop can also be derived from a lattice model.[44,47,48] Based on that, Vfold predicts the free energy landscapes of pseudoknots.[44,48]

3. RNA 3D structure prediction

Although the secondary structure provides the blueprint of an RNA molecule, the knowledge of RNA 3D structure is still indispensable for understanding its function comprehensively. For initial 3D structure modeling, the 3D structures of some common RNA molecules such as tRNAs,[9] group I introns,[10] and RNase P[11] have been successfully built by the RNA structure experts. In recent years, a variety of computational models have been developed for predicting the RNA 3D structures,[13,14,49–51] which have been tabulated in Table 2. These models can be roughly classified into two categories, based on whether they are knowledge-based or physics-based.

Table 2. The major algorithms for RNA 3D structure prediction.

Algorithms	Ref.	Classification	Model	Web page
MANIP	[52]	Graphics-based	All-atomic	http://www-ibmc.u-strasbg.fr/upr9002/westhof/index.html
ERNA-3D	[53]	Graphics-based	All-atomic	http://owww.molgen.mpg.de/~ag_ribo/ag_brimacombe/ERNA3D/ERNA-3D.html
RNA2D3D	[54]	Graphics-based	All-atomic	http://www.ccrnp.ncifcrf.gov/users/bshapiro/software.html
S2S/Assemble	[55], [56]	Graphics-based	All-atomic	http://bioinformatics.org/assemble/
ModeRNA	[57]	Homology-based	All-atomic	http://iimcb.genesilico.pl/moderna/
RNABuilder	[58]	Homology-based	All-atomic	https://simtk.org/home/rnatoolbox
3dRNA	[59]	Homology-based	All-atomic	http://biophy.hust.edu.cn/3dRNA/3dRNA.html
RNAComposer	[60]	Homology-based	All-atomic	http://rnacomposer.ibch.poznan.pl
MC-fold/MC-Sym	[61]	Physics-based	All-atomic	http://www.major.iric.ca/MajorLabEn/MC-Tools.html
FARNA/FARFAR	[62], [63]	Physics-based	All-atomic	http://rosie.rosettacommons.org/
RSIM	[64]	Physics-based	All-atomic	http://www.github.com/jpbida/rsim
BARNACLE	[65]	Physics-based	All-atomic	http://sourceforge.net/projects/barnacle-rna/
RNAnbnds	[66]	Physics-based	All-atomic	http://biophy.nju.edu.cn/index-en.htm
YUP	[67]	Physics-based	Coarse-grained: One-bead	http://www.harvey.gatech.edu/YammpWeb/default.html
NAST	[68]	Physics-based	Coarse-grained: One-bead	https://simtk.org/home/nast
iFold	[69], [70]	Physics-based	Coarse-grained: Three-bead	http://danger.med.unc.edu/tools.php
Vfold	[71], [72]	Physics-based	Coarse-grained: Three-bead	http://vfold.missouri.edu/chen-software02.html
Five-bead Model	[73], [74]	Physics-based	Coarse-grained: Five-bead	http://biomol.bme.utexas.edu/lab/
HiRE-RNA	[75], [76]	Physics-based	Coarse-grained: Six/seven-bead	http://www-lbt.ibpc.fr/LBT/index.php?page=lbt&hl=en

Note: This table summarizes the major methods in RNA 3D structure prediction, and provides a URL for downloading or using online for each method.

3.1. Knowledge-based modeling

With the increasing number of experimentally determined structures in the database, RNA 3D structures can be predicted by the assembly of known motifs or sequence alignment. Knowledge-based modeling mainly includes graphics-based modeling and homology-based modeling.

3.1.1. Graphics-based methods

Graphics-based modeling generally provides a graphical interface and allows users to construct RNA 3D structures by manipulating or assembling RNA segments.[52–56] The major graphics-based algorithms will be introduced as follows.

MANIP As an interactive tool, the MANIP allows the assembly of known 3D motifs into a complete RNA structure on the computer screen by users from the corresponding secondary structure predicted by comparative sequence analysis.[52] Although the MANIP is not an automatic procedure, it constitutes a quick and easy way to build the 3D structures of RNAs especially large-size RNAs, e.g., the RNase P RNA and the group I introns. In addition, multiple connection and base-pair tables that explicitly contain topological information of RNAs can be a reference for the precise modeling of interactions between RNAs.[52]

ERNA-3D To generate RNA 3D structures from sequences and secondary structures, the ERNA-3D provides a graphical interface for users to freely position the A-form helices and to directly pull the single inter-helical strands.[53] The main advantage of the algorithm is that there is almost no limitation to the size of RNAs. ERNA-3D has been used to successfully generate the 3D structures of mRNA, tRNAs, and rRNAs, including the 16 S rRNA, 23 S rRNA, and 5 S rRNA.[53]

RNA2D3D The RNA2D3D can rapidly predict rough 3D structures for large RNAs, e.g., ribozymes, viral kissing loops and a variety of RNA nanostructures, based on their secondary structures.[54] However, manual manipulation is required to generate satisfactory 3D structures through a graphical interface with special tools such as compacting, stem-stacking, segment-positioning, and energy-refinement.[54]

S2S/Assemble The S2S/Assemble algorithm is an interactive graphical algorithm in which users can easily display, manipulate and interconnect RNA data from sequence to structure, as well as analyzing and building RNA 3D architectures.[55,56] Since all interactions including base-pairing and base-stacking have to be annotated manually, it is difficult to perform a high-throughput structure prediction with this algorithm.

Although the graphics-based methods introduced above can be used to build 3D structures for large RNAs with hundreds of nucleotides rapidly and intuitively, because they are

manual tools, they need users to set up and refine the RNA structure models according to specific principles using the tools included in the software packages. Thus, in order to build plausible structures, it is necessary for users to have intimate knowledge of RNA structures.

3.1.2. Homology-based methods

Since the 3D structure of a macromolecule evolves much more slowly than its sequence, evolutionarily related macromolecules usually retain similar 3D structures despite their divergence on the sequence level. Based on this fact, the 3D structures of a macromolecule can be built through aligning the sequence of the target molecule to the structures of template molecules.[56] Homology-based modeling, also called template-based modeling or comparative modeling, has been rather successful in predicting 3D structure of proteins.[77,78] In analogy to the homology-based modeling for proteins, several algorithms[57–60,79] such as ModeRNA,[57] RNABuilder,[58] and 3dRNA[59,79] have been developed for building the RNA 3D structures.

ModeRNA The ModeRNA algorithm allows both simplistic structure prediction from a set of templates/alignments and user-controlled manipulations of structures, e.g., fragment assembly.[57] Compared with other modeling algorithms, the ModeRNA can recognize and model post-transcriptionally modified nucleosides. It should be noted that although ModeRNA is not a graphics-based tool, it still requires users to supply the alignment between the template RNAs and the target RNA, and to specify the base pairs between the inserted fragment and the rest of the RNA.

RNABuilder RNABuilder, now known as MMB, is another method for comparative modeling of RNA structures by using the distance and torsion angles from the aligned regions of the template as modeling restraints.[58] The distinguishing characteristic of RNABuilder is that it introduces a CG force field including base-pairing, base-stacking, tertiary interactions, and excluded-volume interactions, to bring the interacting bases into the base pairing geometry. The software package is written with an internal coordinate mechanics library, so the CG force field can be applied to the base rather than to individual atoms. Although all forces employed in RNABuilder are user-specified, it provides a new framework for predicting RNA 3D structures, combining experimental constraints and users' hypotheses. In addition, RNABuilder has been used successfully to build the entire 200-nt Azoarcus ribozyme structure with structural information from two template structures.[58]

3dRNA 3dRNA is a faster and more automated algorithm for building RNA 3D structures by assembling the A-form helices and various loops whose structures are extracted from known structures in a database.[59,79] For 300 tested RNAs including duplexes, hairpins, pseudoknots, and junctions, 3dRNA predicts reliable 3D structures based on their secondary structures. In addition, as a web server, 3dRNA can be used freely online, and with the sequence and secondary structure as input, one can obtain the predicted 3D structure quickly.[59]

As introduced above, the major advantage of homology-based modeling is that there is generally no limit on the size of RNAs to be modeled. However, the quality of the homology model is dependent on the quality of the sequence alignments and template structures. Although the number of known RNA structures stored in the PDB/NDB database is increasing rapidly, it still may not be easy to find proper template RNAs for a particular target RNA. Moreover, due to the high flexibility of RNAs, their structures generally change with the solution environment such as temperature, ion conditions,[80] and the existence of other ligands or macromolecules. In addition, the development of a good alignment for RNAs with complex structures usually requires laborious manual preparation based on the established expertise with regard to the most relevant RNA families. Thus, homology-based modeling is not always effective.

3.2. Physics-based modeling

Physics-based approaches are based on biophysical principles, which simulate the folding process in search of a conformation with minimal free energy. Since a full-atom structure modeling for an RNA generally involves many degrees of freedom and consequently huge computational complexity, several CG predictive models have also been proposed at different resolution levels with physical simplifications.

3.2.1. All-atomistic model

Until now, the all-atomistic molecular dynamics is popular for macromolecule modeling and can give a view of the actual motion of atoms with physics-based atomic force fields such as AMBER[81,82] and CHARMM.[83] However, due to the many degrees of freedom, it is still difficult to fold RNA 3D structures with these all-atomistic force fields, even with the latest computational facilities. Consequently, models with the all-atomistic fragments assembled have been developed based on the known fragments or secondary structures,[61–66,84–86] such as MC-fold/Mc-Sym pipeline,[61] FARNA/FARFAR,[62,63] RSIM,[64] and RNAnbds.[66,84]

MC-fold/MC-sym Since secondary structures provide enough structural constraints to automated 3D building, the MC-fold/MC-sym pipeline infers RNA secondary structure from sequence data and then assembles a series of 3D structures based on their secondary structures.[61] Unlike the thermodynamics approaches such as Mfold,[22] the MC-fold can predict RNA secondary structures including canonical and

non-canonical base pairs with the use of a knowledge-based scoring function based on a database of NCMs (nucleotide cycle motifs). The NCMs, which are circularly connected by covalent, pairing or stacking interactions, were built from an analysis of X-ray crystallographic structures. The fragment insertion simulation is performed by MC-sym with the 3D NCMs and the Las Vegas algorithm. MC-fold/MC-sym pipeline has been validated by building 3D structures of precursor microRNA and a new 3D structure of the human immunodeficiency virus (HIV1) *cis*-acting -1 frame-shifting segment.[61]

FARNA/FARFAR Das and Baker explored a fully automated and energy-based approach of FARNA to predict RNA 3D structure.[62] With the use of Monte Carlo algorithm and a simplified knowledge-based energy function that favors base pairing and stacking geometries, FARNA assembles trinucleotide fragments extracted from the ribosome crystal structure into an all-atomistic structure corresponding to the given sequence. The CG base-pairing potential used in FARNA is based on the statistical analysis of the bases in the ribosome and can account for not only Watson–Crick base pairs but also the interactions with the Hoogsteen and sugar edges. The FARFAR introduces a high-resolution refinement phase into FARNA to predict and design non-canonical small RNA structures with atomic accuracy.[63]

RSIM As an improved method of FARNA,[62] the RSIM predicts RNA 3D structures from secondary structure constraints using Monte Carlo algorithm with closed moves.[64] The important feature of RSIM is that it allows visualization of the predicted RNA conformational space as a graph, which provides insights into the possible RNA dynamics to further identify functional RNA conformations. However, the current implementation of RSIM is limited to non-pseudoknotted RNAs and requires manual intervention for complicated branched structures.[64]

RNAnbds The RNAnbds is designed to predict RNA 3D structures by fragment assembly based on the statistical configurations of bases according to their sequence/spatial neighbors in databases.[66,84] Combined with a statistical potential including base-pairing and base-stacking and a sequential Monte Carlo method, RNAnbds gives reliable prediction for short fragments (< 15 nt), especially loops with RMSDs < 4 Å between the predicted and the experimental fragments.

3.2.2. Coarse-grained model

Another way to reduce the computational cost is to reduce the number of particles by treating a group of functional atoms as a single bead.[87–89] The bead may represent only a few atoms or a large group of atoms depending on the resolution of the model (see Fig. 1). After the one-bead RNA model first developed by Malhotra and Harvey,[90] many CG models have been implemented to predict RNA 3D structures[67–76] or to model the interactions between RNAs and other molecules,[91,92] such as YUP,[67] NAST,[68] iFold,[69,70] Vfold,[71,72] etc. (see Table 2).

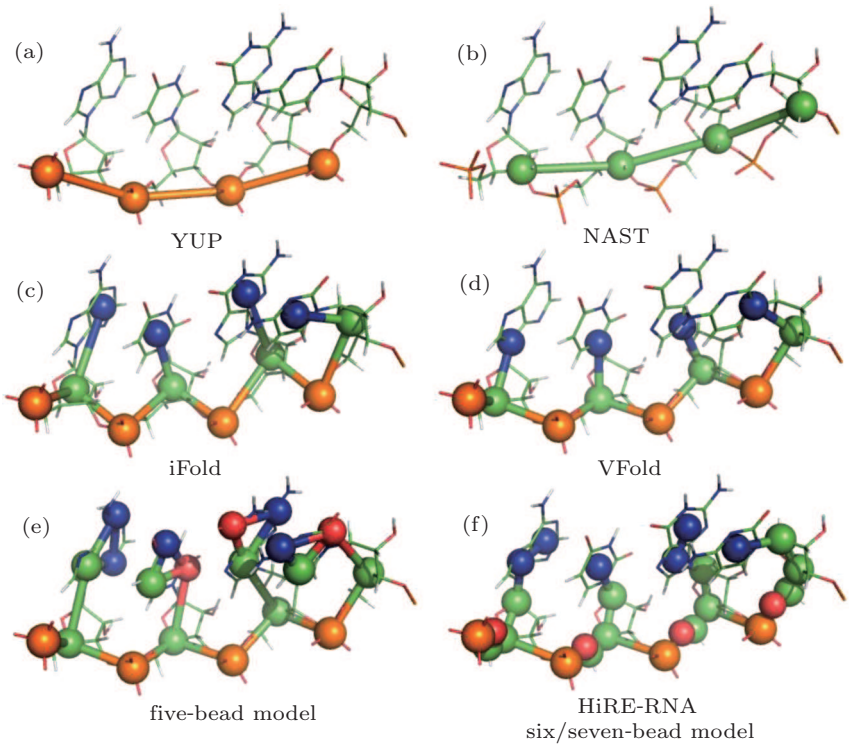

Fig. 1. (color online) Coarse-grained representations (ball-sticks) of different resolution models for one fragment of sequence: 5'-AUGC-3' superposed on an all-atom representation (thin sticks). (a), (b) One-bead model; (c), (d) three-bead model; (e), (f) high-resolution model.

YUP The YUP is a very flexible molecular mechanics algorithm for CG and multi-scaled modeling.[67] Based on the associated energy potentials and the methods such as Monte Carlo, molecular dynamics, and energy minimization, the algorithm has been used to model RNA structures as well as structures of protein and DNA. In YUP, to model large RNAs, one nucleotide is replaced by one pseudoatom placed at the center of a phosphorus atom, which effectively reduces the computational cost. Although YUP requires users to provide the secondary structure of RNAs and the information for the force field, it is an adaptive package for automatic CG modeling of RNAs.[67]

NAST Similar to YUP,[67] the NAST is another one-bead model, in which the C3' atom in a nucleotide is chosen to represent the entire nucleotide.[68] With an RNA-specific knowledge-based potential and a simple molecular dynamics algorithm, NAST can sample conformations that satisfy a given set of secondary structure and tertiary contact constraints. One advantage of NAST is its ability to incorporate experimental data such as ideal small-angle X-ray scattering data and experimental solvent accessibility data as the filters to rank the clusters of structurally similar conformations. NAST has been used to predict the 3D structures of the yeast phenylalanine tRNA (76 nt) and the P4–P6 domain of the Tetrahymena thermophile group I intron (158 nt) within 8 Å and 16 Å RMSDs from experimental structures, respectively.[68]

iFold iFold is a novel web-based algorithm developed by the Doknolyan's group, which can be used to predict RNA 3D structures from sequences.[69,70] To rapidly explore the possible conformational space of RNA molecules, the model employs a CG representation of three beads per nucleotide and the efficient discrete molecular dynamics simulations with stepwise potentials including base-pairing and base-stacking. The power of iFold has been demonstrated by predicting the 3D structures of 150 RNAs with diverse sequences and the majority of the predicted structures have < 4 Å deviations from experimental structures.[69]

Vfold Based on the experimental thermodynamic parameters for helices and the loop entropy from the random walks of virtual bonds in a diamond lattice, the Vfold provides the free energy landscape by enumerating all of the secondary structures including pseudoknots.[71,72] The corresponding 3D structures of the minimum and the local minima in the energy landscape can be built through the fragment assembly based on the secondary structures, and then refined by AMBER[81] energy minimization. The notable advantage of Vfold is the ability to generate RNA structures with cross-linked loops and helices such as pseudoknots and complexes.[72]

Five-bead model For higher resolution, a five-bead model was developed by Ren and co-workers.[73,74] In the model, each nucleotide is represented by five pseudoatoms, two of which are for the backbone and three are for the base to preferably capture base stacking and pairing. With molecular dynamics simulations and the simulated annealing algorithm, the model can be used to predict the 3D structures of small RNAs and to capture the 3D structures of large-size RNA by integrating the limited experimental data.[74] Since the model is designed to be compatible with existing all-atomistic modeling packages, the predicted CG structures can be directly mapped to all-atomistic structures and refined with an all-atomistic force field.[74]

HiRE-RNA HiRE-RNA is a generic high-resolution CG model with four beads for the RNA backbone, which enables a better description of protein–RNA binding interfaces.[75,76] The base pairing is modeled by hydrogen bonding interactions consisting of 2-, 3-, and 4-body terms, and the involved interactions include canonical Watson–Crick and wobble base pairs as well as the relatively rare A–C, A–G, and U–C pairs. The model has been built in molecular dynamics simulations to study the stability of RNA hairpins[75] and duplexes.[76] However, the parameters of the model may need further adjustment to match the predicted thermodynamic properties of RNAs to the experimental data.

3.2.3. A CG model for 3D structure, stability, and salt effect

As discussed above, most existing models focus on the 3D structure prediction. Since RNAs are highly charged polyanionic polymers, the RNA structures are sensitive to temperature and ion conditions.[93–96] Is there a model that can predict the RNA structures at a given temperature and ionic condition? To address this issue, a new three-bead CG model[97] has been developed very recently with three beads placed respectively on the existing P, C4' and N9 atoms for purine (or N1 for pyrimidine). With the Monte Carlo simulated annealing algorithm, the model could not only predict 3D structures, but also give reliable predictions on the stability and salt effect for small RNAs. The implicit-solvent force field used in the model contains the local geometric constraints of RNAs as well as the sequence-dependent base pairing and base stacking potentials and the electrostatic interactions between phosphate groups.

The model was first validated by predicting the 3D structures of 46 RNAs (≤ 45 nt) from their sequences, including pseudoknots and hairpins, with or without bulge/internal loops. For each sequence, from an initial random chain, the RNA is folded into a series of native-like structures with the decrease of temperature to a target temperature in the Monte Carlo annealing process. The overall mean RMSD and the overall minimum RMSD between the predicted structures of 46 tested RNAs and the corresponding experimental structures are 3.6 Å and 2.1 Å, respectively. Beyond that, the model

can also make reliable predictions on the stability of various RNA hairpins such as melting temperatures with the mean errors 1.0 °C from extensive experimental data at extensive Na^+ concentrations.[96] Meanwhile, it can provide the ensemble of probable 3D structures of RNA hairpins at different temperature/salt conditions. Therefore, the model may become a basis for a unified model for predicting 3D structures, as well as stability and salt effects, for large RNAs with complex structures.

4. Conclusion and perspective

Since understanding RNA structure, especially in 3D, is crucial to understanding the mysterious RNA world, various computational models have been developed in modeling RNA structures in recent years. RNA secondary structure can be relatively less difficult to predict from sequence alignment, thermodynamics-based dynamic programming algorithms or a combination of approaches. The accuracy of predicting secondary structures would be improved by having more structures described in a database or by refining the empirical thermodynamic parameters that are available. However, the art of predicting RNA 3D structures is just beginning to develop and faces many challenges.

First of all, predicting 3D structures for large-size RNAs is still a primary challenge, especially the determination of long-range tertiary interactions within RNAs.[98] Some knowledge-based methods can predict structures with no size limitation, but their requirements for descriptions of known structures and for hands-on interaction with expert user limit their wider application. For physics-based models, one possible way to address large-size RNAs is to reduce the sampled conformational space. Since RNA folding is generally hierarchical, the secondary structure information (e.g., base-pairing and base-stacking) obtained from various algorithms and experiments can be used to constrain the extent of sampling for 3D structure prediction. The architects of some models have taken the lead in assembling the 3D fragments into all-atomistic structures from secondary structures, and this strategy has proven to be effective.[99] Moreover, very recent studies show that the use of experimental data such as low-resolution SAXS data and multiplexed hydroxyl radical cleavage analysis can dramatically improve the efficiency and accuracy of structure prediction.[100–104] However, how to efficiently and accurately predict 3D structures for large RNAs from sequences is still a puzzle. In addition, "coarse-graining", i.e., treating group of functional atoms as a single bead, is another way to reduce the sample space. Generally, to obtain structures at higher resolution, all-atomic structures can be reconstructed from CG structures and optimized with all-atomistic force fields such as AMBER[81] and CHARMM.[83]

The second big challenge at the moment is that current studies of RNA functions suggest that the actual functional structures of RNAs might differ from the theoretical structures predicted on the basis of minimum free energy.[48,105–107] For instance, riboswitches usually perform their functions by changing structure rather than retaining a static native structure.[107] This indicates that the kinetic RNA structures that develop during the folding process can be important for RNA functions. Although a few methods such as iFold[69] and Vfold[72] can partly predict the 3D structures of the local minima in the energy landscape, it is still difficult for them to provide the folding kinetics of 3D structures because of the incomputable conformations. Nevertheless, it is plausible that one could predict the folding kinetics of a given RNA at secondary structure level using the existing algorithms such as Kinwalker[28] and MPGAfold,[29] and then transform the kinetically important structures to 3D representation by the fragment assembly or other physics-based methods to approximately obtain the kinetic RNA folding process at the 3D level, including folding pathways and rates, and kinetic structures.

Thirdly, RNA 3D structures are very sensitive to the solution environment, e.g., temperature, ions, ligands, and other macromolecules, due to the structures' high flexibility and the high density of negative charges on RNAs.[8,93,108–117] However, most of the current structure prediction models seldom consider any external conditions beyond the room/body temperature and high salt (namely, 1M NaCl). It is a challenge to predict RNA 3D structures at arbitrarily chosen temperature/ion conditions. Very recently, a new physics-based CG model has been developed to predict the RNA 3D structures at different salt/temperature conditions from sequences,[97] but the model is only applicable to small RNAs ($\leq$ 45 nt) and it is not suitable for solutions with multivalent ions such as Mg^{2+}, solutions that have been shown to play very important roles in RNA tertiary structures and functions.[93,117,118] The model can be possibly extended to involve explicit ions, while the conformation search cost would increase correspondingly.[95,119] Furthermore, it is known that cells contain up to 30% by volume of macromolecular species such as proteins, DNA, and RNA.[114] Recent experimental and theoretical studies have suggested that macromolecular crowding can greatly influence the structure stability and kinetic properties of RNAs.[116] Unfortunately, the existing prediction models seldom consider the influence of macromolecular crowding on RNA 3D structure. Additionally, some RNAs become functional only upon forming complexes with proteins or other molecules. For example, RNA–protein complexing plays an important role in protein synthesis, viral replication, cellular defense and developmental regulation.[120–123] To form complexes, RNAs usually adopt specific structures to provide binding sites for proteins and other molecules.[121–123] However, to model the complex structures of RNAs with other molecules is more challenging, due to the fewer available ex-

perimentally discovered structures[122] and the complicated interactions between RNAs and other molecules.

Nevertheless, the recent successes in RNA 3D structure modeling suggest that we can expect many exciting developments in RNA structure prediction in the coming decade.

Acknowledgments

We are grateful to Shi-Jie Chen, Wen-Bing Zhang, and Song Cao for their valuable discussion.

References

[1] Gesteland R F, Cech T R and Atkins J F 2006 *The RNA World* (3rd edn.) (New York: Cold Spring Harbor Press)
[2] Crick F 1970 *Nature* **227** 561
[3] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S R and Bateman A 2005 *Nucleic Acids Res.* **33** 121
[4] Doherty E A and Doudna J A 2001 *Annu. Rev. Biophys. Biomol. Struct.* **30** 457
[5] Hannon G J 2002 *Nature* **418** 244
[6] Chen J and Zhang W 2012 *J. Chem. Phys.* **137** 225102
[7] Edwards T E, Klein D J and Ferré-D'Amaré A R 2007 *Curr. Opin. Struct. Biol.* **17** 273
[8] Tinoco I and Bustamante C 1999 *J. Mol. Biol.* **293** 271
[9] Levitt M 1969 *Nature* **224** 759
[10] Michel F and Westhof E 1990 *J. Mol. Biol.* **216** 585
[11] Harris M E, Nolan J M, Malhotra A, Brown J W, Harvey S C and Pace N R 1994 *EMBO J.* **13** 3953
[12] Stagg S M, Mears J A and Harvey S C 2003 *J. Mol. Biol.* **328** 49
[13] Shapiro B A, Yingling Y G, Kasprzak W and Bindewald E 2007 *Curr. Opin. Struct. Biol.* **17** 157
[14] Laing C and Schlick T 2010 *J. Phys. Condens. Matter.* **22** 283101
[15] Mathews D H 2006 *J. Mol. Biol.* **359** 526
[16] Puton T, Kozlowski L P, Rother K M and Bujnicki J M 2013 *Nucleic Acids Res.* **41** 4307
[17] Mathews D H and Turner D H 2006 *Curr. Opin. Struct. Biol.* **16** 270
[18] Bernhart S H, Hofacker I L, Will S, Gruber A R and Stadler P F 2008 *Bioinformatics* **9** 474
[19] Knudsen B and Hein J 2003 *Nucleic Acids Res.* **31** 3423
[20] Sükösd Z, Knudsen B, Kjems J and Pedersen C N 2013 *Bioinformatics* **28** 2691
[21] Siebert S and Backofen R 2005 *Bioinformatics* **21** 3352
[22] Zuker M and Stiegler P 1981 *Nucleic Acids Res.* **9** 133
[23] Reuter J S and Mathews D H 2010 *Bioinformatics* **11** 129
[24] Hofacker I L and Stadler P F 2006 *Bioinformatics* **22** 1172
[25] Do C B, Woods D A and Batzoglou S 2006 *Boinformatics* **23** 90
[26] Ding Y and Lawrence C E 2003 *Nucleic Acids Res.* **31** 7280
[27] Giegerich R, Voss B and Rehmsmeier M 2004 *Nucleic Acids Res.* **32** 4843
[28] Geis M, Flamm C, Wolfinger M T, Tanzer A, Hofacker I L, Middendorf M, Mandl C, Stadler P F and Thurner C 2008 *J. Mol. Biol.* **379** 160
[29] Shapiro B A, Wu J C, Bengali D and Potts M J 2001 *Bioinformatics* **17** 137
[30] Rivas E and Eddy S R 1999 *J. Mol. Biol.* **285** 2053
[31] Ruan J, Stormo G D and Zhang W 2003 *Bioinformatics* **20** 58
[32] Ren J, Bastegari B, Condon A and Hoos H H 2005 *RNA* **11** 1494
[33] Reeder J and Giegerich R 2004 *Bioinformatics* **5** 104
[34] Madison J T, Everett G A and Kung H 1966 *Science* **153** 531
[35] Mathews D H and Turner D H 2002 *J. Mol. Biol.* **317** 191
[36] Harmanci A O Sharma G and Mathews D H 2011 *Bioinformatics* **12** 108
[37] Höchsmann M, Voss B and Giegerich R 2004 *IEEEACM Trans. Comput. Biol. Bioinform.* **1** 53
[38] Gardner P P and Giegerich R 2004 *Bioinformatics* **5** 140
[39] Tinoco I, Uhlenbeck O C and Levine M D 1971 *Nature* **230** 362
[40] Xia T, SantaLucia J, Burkand M E, Kierzek R, Schroeder S J, Jiao X, Cox C and Turner D H 1998 *Biochemistry* **37** 14719
[41] Zuker M 2000 *Curr. Opin. Struct. Biol.* **10** 303
[42] Zhang W and Chen S J 2002 *Proc. Natl. Acad. Sci. USA* **99** 1931
[43] Zhao P, Zhang W and Chen S J 2010 *Biophys. J.* **98** 1617
[44] Cao S and Chen S J 2006 *Nucleic Acids Res.* **34** 2634
[45] Chen X, He S M, Bu D, Zhang F, Wang Z, Chen R and Gao W 2008 *Bioinformatics* **24** 1994
[46] Xayaphoummine A, Bucher T and Isambert H 2005 *Nucleic Acids Res.* **33** 605
[47] Zhang J, Lin M, Chen R, Wang W and Liang J 2008 *J. Chem. Phys.* **128** 125107
[48] Chen S J 2008 *Annu. Rev. Biophys.* **37** 197
[49] Rother K, Rother M, Boniecki M, Puton T and Bujnicki J M 2011 *J. Mol. Model* **17** 2325
[50] Cruz J A, Blanchet M F, Boniecki M, Bujnicki J M, Chen S J, Cao S, Das R, Ding F, Dokholyan N V, Flores S C, Huang L, Lavender C A, Lisi V, Major F, Mikolajczak K, Patel D J, Philips A, Puton T, Santalucia J, Sijenyi F, Hermann T, Rother K, Rother M, Serganov A, Skorupski M, Soltysinski T, Sripakdeevong P, Tuszynska I, Weeks K M, Waldsich C, Wildauer M, Leontis N B and Westhof E 2012 *RNA* **18** 610
[51] Hajdin C E, Ding F, Dokholyan N V and Weeks K M 2010 *RNA* **16** 1340
[52] Massire C and Westhof E 1998 *J. Mol. Graph. Model.* **16** 197
[53] Zwieb C and Muller F 1997 *Nucleic Acids Symp. Ser.* **36** 69
[54] Martinez H M, Maizel J V and Shapiro B A 2008 *J. Biomol. Struct. Dyn.* **25** 669
[55] Jossinet F and Westhof E 2005 *Bioinformatics* **21** 3320
[56] Jossinet F, Ludwig T E and Westhof E 2010 *Bioinformatics* **26** 2057
[57] Rother M, Rother K, Puton T and Bujnicki J M 2011 *Nucleic Acids Res.* **39** 4007
[58] Flores S C and Altman R B 2010 *RNA* **16** 1769
[59] Zhao Y, Huang Y, Gong Z, Wang Y, Man J and Xiao Y 2012 *Sci. Rep.* **2** 734
[60] Popenda M, Szachniuk M, Antczak M, Purzycka K J, Lukasiak P, Bartol N, Blazewicz J and Adamiak R W 2012 *Nucleic Acids Res.* **40** e112
[61] Parisien M and Major F 2008 *Nature* **452** 51
[62] Das R and Baker D 2007 *Proc. Natl. Acad. Sci. USA* **104** 14664
[63] Das R, Karanicolas J and Baker D 2010 *Nat. Meth.* **7** 291
[64] Bida J P and Maher L J 2012 *RNA* **18** 385
[65] Frellsen J, Moltke I, Thiim M, Mardia K V and Ferkinghoff-Borg J 2009 *PloS Comput. Biol.* **5** e1000406
[66] Zhang J, Bian Y, Lin H and Wang W 2012 *Phys. Rev. E* **85** 021909
[67] Tan R K Z, Petrov A S and Harvey S C 2006 *J. Chem. Theory Comput.* **2** 529
[68] Jonikas M A, Radmer R J, Laederach A, Das R, Pearlman S, Herschlag D and Altman R B 2009 *RNA* **15** 189
[69] Ding F, Sharma S, Chalasani P, Demidov V V, Broude N E and Dokholyan N V 2008 *RNA* **14** 1164
[70] Sharma S, Ding F and Dokholyan N V 2008 *Bioinformatics* **24** 1951
[71] Cao S and Chen S J 2005 *RNA* **11** 1884
[72] Cao S and Chen S J 2011 *J. Phys. Chem. B* **115** 4216
[73] Xia Z, Gardner D P, Gutell R R and Ren P 2010 *J. Phys. Chem. B* **114** 13497
[74] Xia Z, Bell D R, Shi Y and Ren P 2013 *J. Phys. Chem. B* **117** 3135
[75] Pasquali S and Derreumaux P 2010 *J. Phys. Chem. B* **114** 11957
[76] Cragnolini T, Derreumaux P and Pasquali S 2013 *J. Phys. Chem. B* **117** 8047
[77] Zhang Y and Skolnick J 2013 *BMC Biol.* **11** 44
[78] Wu S and Zhang Y 2008 *Bioinformatics* **24** 924
[79] Zhao Y, Gong Z and Xiao Y 2011 *J. Biomol. Struct. Dyn.* **28** 815
[80] Tan Z J and Chen S J 2009 *Methods Enzymol.* **469** 465
[81] Case D A, Cheatham T E, Darden T, Gohlke H, Luo R, Merz K M, Onufriev A, Simmerling C, Wang B and Woods R J 2005 *J. Comput. Chem.* **26** 1668
[82] Gong Z, Zhao Y and Xiao Y 2010 *J. Biomol. Struct. Dyn.* **28** 431
[83] Brooks B R, Brooks C L, Mackerell A D, Nilsson L, Petrella R J, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner A R, Feig M, Fischer S, Gao J, Hodoscek M W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor R W, Post C B, Pu J Z, Schaefer M, Tidor B, Venable R M, Woodcock H L, Wu X, Yang W, York D M and Karplus M 2009 *J. Comput. Chem.* **30** 1545
[84] Zhang J, Zhang Y J and Wang W 2010 *Chin. Phys. Lett.* **27** 118702

[85] Zhang J, Dundas J, Lin M, Chen M, Wang W and Liang J 2009 *RNA* **15** 2248
[86] Laing C, Jung S, Kim N, Elmetwaly S, Zahran M and Schlick T 2013 *PLoS ONE* **8** e71947
[87] Paliy M, Melnik R and Shapiro B A 2010 *Phys. Biol.* **7** 036001
[88] de Pablo J J 2011 *Annu. Rev. Phys. Chem.* **62** 555
[89] Leonarski F, Trovato F, Tozzini V, Les A and Trylska J 2013 *J. Chem. Theory Comput.* **9** 4874
[90] Malhotra A and Harvey S C 1994 *J. Mol. Biol.* **240** 308
[91] Denesyuk N and Thirumalai D 2013 *J. Phys. Chem. B* **117** 4901
[92] Hori N and Takada S 2012 *J. Chem. Theory Comput.* **8** 3384
[93] Woodson S A 2005 *Curr. Opin. Chem. Biol.* **9** 104
[94] Tan Z J and Chen S J 2011 *Met. Ions Life Sci.* **9** 101
[95] Wang F H, Wu Y Y and Tan Z J 2013 *Biopolymers* **99** 370
[96] Tan Z J and Chen S J 2008 *Biophys. J.* **95** 738
[97] Shi *et al*. 2014 (in submission)
[98] Chauhan S and Woodson S A 2008 *J. Am. Chem. Soc.* **130** 1296
[99] Bailor M H, Mustoe A M, Brooks C L and Al-Hashimi H M 2011 *Curr. Opin. Struct. Biol.* **21** 296
[100] Wilkinson K A, Merino E J and Weeks K M 2006 *Nat. Protoc.* **1** 1610
[101] Parisien M and Major F 2012 *J. Struct. Biol.* **179** 252
[102] Das R, Kudaravalli M, Jonikas M, Laederach A, Fong R, Schwans J P, Baker D, Piccirilli J A, Altman R B and Herschlag D 2008 *Proc. Natl. Acad. Sci. USA* **105** 4144
[103] Seetin M J and Mathews D H 2011 *J. Comput. Chem.* **32** 2232
[104] Ding F, Lavender C A, Weeks K M and Dokholyan N V 2012 *Nat. Methods* **9** 603
[105] Pan J, Thirumalai D and Woodson S A 1997 *J. Mol. Biol.* **273** 7
[106] Hall K B 2012 *Methods Mol. Biol.* **875** 67
[107] Chen J, Gong S, Wang Y and Zhang W 2014 *J. Chem. Phys.* **140** 025102
[108] Zhang Y, Zhang J and Wang W 2011 *J. Am. Chem. Soc.* **133** 6882
[109] Draper D E 2013 *Biopolymers* **99** 1105
[110] Tan Z J and Chen S J 2005 *J. Chem. Phys.* **122** 44903
[111] Tan Z J and Chen S J 2006 *Biophys. J.* **90** 1175
[112] Tan Z J and Chen S J 2010 *Biophys. J.* **99** 1565
[113] Gong Z, Zhao Y, Chen C and Xiao Y 2011 *J. Biomol. Struct. Dyn.* **29** 403
[114] Ellis R J 2001 *Curr. Opin. Struct. Biol.* **11** 114
[115] Tan Z J and Chen S J 2012 *Biophys. J.* **103** 827
[116] Zhou H X 2008 *Arch. Biochem. Biophys.* **469** 76
[117] Draper D E 2008 *Biophys J.* **95** 5489
[118] Tan Z J and Chen S J 2011 *Biophys. J.* **101** 176
[119] Wu Y Y, Wang F H and Tan Z J 2013 *Phys. Lett. A* **377** 1911
[120] Hall K B 2002 *Curr. Opin. Struct. Biol.* **12** 283
[121] Wang Y, Chen X, Liu Z P, Huang Q, Wang Y, Xu D, Zhang X S, Chen R and Chen L 2013 *Mol. Biosyst.* **9** 133
[122] Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu H, Zhao Y amd Chen R 2006 *Nucleic Acids Res.* **34** D150
[123] Huang Y, Liu S, Guo D, Li L and Xiao Y 2013 *Sci. Rep.* **3** 1887