

Statistical potentials for 3D structure evaluation: From proteins to RNAs

Ya-Lan Tan(谭雅岚), Chen-Jie Feng(封晨洁), Xunxun Wang(王勋勋), Wenbing Zhang(张文炳), and Zhi-Jie Tan(谭志杰)

Citation: Chin. Phys. B, 2021, 30 (**2**): 028705. DOI: 10.1088/1674-1056/abc0d6
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Retrieval of multiple scattering contrast from x-ray analyzer-based imaging

Heng Chen(陈恒), Bo Liu(刘波), Li-Ming Zhao(赵立明), Kun Ren(任坤), and Zhi-Li Wang(王志立)
Chin. Phys. B, 2021, 30 (**1**): 018701. DOI: 10.1088/1674-1056/abbbed

Dynamics and coherence resonance in a thermosensitive neuron driven by photocurrent

Ying Xu(徐莹), Minghua Liu(刘明华), Zhigang Zhu(朱志刚), Jun Ma(马军)
Chin. Phys. B, 2020, 29 (**9**): 098704. DOI: 10.1088/1674-1056/ab9dee

Entrainment mechanism of the cyanobacterial circadian clock induced by oxidized quinone

Ying Li(李莹), Guang-Kun Zhang(张广鹃), Zi-Gen Song(宋自根)
Chin. Phys. B, 2020, 29 (**9**): 098703. DOI: 10.1088/1674-1056/aba615

A 3D biophysical model for cancer spheroid cell-enhanced invasion in collagen-oriented fiber microenvironment

Miaomiao Hai(海苗苗), Yanping Liu(刘艳平), Ling Xiong(熊玲), Guoqiang Li(李国强), Gao Wang(王高), Hongfei Zhang(张鸿飞), Jianwei Shuai(帅建伟), Guo Chen(陈果), Liyu Liu(刘雳宇)
Chin. Phys. B, 2020, 29 (**9**): 098702. DOI: 10.1088/1674-1056/ab9c02

Improving RNA secondary structure prediction using direct coupling analysis

Xiaoling He(何小玲), Jun Wang(王军), Jian Wang(王剑), Yi Xiao(肖奕)
Chin. Phys. B, 2020, 29 (**7**): 078702. DOI: 10.1088/1674-1056/ab889d

---

TOPICAL REVIEW — Modeling and simulations for the structures and functions of proteins and nucleic acids

Statistical potentials for 3D structure evaluation: From proteins to RNAs*

Ya-Lan Tan(谭雅岚), Chen-Jie Feng(封晨洁), Xunxun Wang(王勋勋), Wenbing Zhang(张文炳)†, and Zhi-Jie Tan(谭志杰)‡

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

(Received 27 June 2020; revised manuscript received 2 September 2020; accepted manuscript online 14 October 2020)

Structure evaluation is critical to *in silico* 3-dimensional structure predictions for biomacromolecules such as proteins and RNAs. For proteins, structure evaluation has been paid attention over three decades along with protein folding problem, and statistical potentials have been shown to be effective and efficient in protein structure prediction and evaluation. In recent two decades, RNA folding problem has attracted much attention and several statistical potentials have been developed for RNA structure evaluation, partially with the aid of the progress in protein structure prediction. In this review, we will firstly give a brief overview on the existing statistical potentials for protein structure evaluation. Afterwards, we will introduce the recently developed statistical potentials for RNA structure evaluation. Finally, we will emphasize the perspective on developing new statistical potentials for RNAs in the near future.

Keywords: statistical potential, 3-dimensional structure evaluation, RNA, protein

PACS: 87.10.Vg, 87.15.bg, 87.14.gn, 87.14.E– **DOI:** 10.1088/1674-1056/abc0d6

1. Introduction

It is widely known that RNAs can pass the genetic information from DNAs to control the synthesis of proteins and can also serve as the storage of genetic information.[1] Beyond the sequence-level functions, a variety of non-coding RNAs have been discovered in recent three decades, which have extensively biological functions such as catalysis and genetic regulation.[2,3] The functions of noncoding RNAs are generally coupled to their structures or proper structure change. For example, ribozymes can only perform the catalysis of splicing or cleavage when they fold into their native 3-dimensional (3D) structures,[4] and riboswitches can only regulate the genetic expression at translation or transcription level through proper change of their 3D structures upon metabolite binding.[5] Therefore, understanding RNA structures, especially RNA 3D structures, would be critical to understanding their biological functions. Experimental methods have been widely employed to derive RNA 3D structures, such as x-ray crystallography, NMR spectroscopy, and cryo-electron.[6,7] However, until now, RNA 3D structures deposited in the protein data bank (PDB) database[8] are still limited, partially because the experimental methods are generally time-consuming and of high cost. In parallel, various computational models have been developed to predict RNA 3D structures in silico recently,[9–12] including knowledge-based and physics-based models.[13–36] Generally, a predictive model would predict an ensemble of structure candidates, and thus, a reliable scoring function becomes highly required to evaluate the quality of the candidate structures. Furthermore, a reliable scoring function can also be employed to guide RNA structure prediction and to optimize predicted structures.[37,38]

For proteins, statistical potentials, also called knowledge-based potentials or mean force potentials, have been proven to be an effective and efficient scoring function for structure evaluation in protein and protein-complex 3D structure predictions, along with the protein folding problem from three decades ago.[39–48] In the recent decade, several statistical potentials have been developed for RNA structure evaluation or prediction,[49–56] most of which were built in similar ways to those for proteins due to the prior progress made for proteins and some similar polymer features between proteins and RNAs such as backbone and side chain. Nevertheless, there is inherent difference between proteins and RNAs, such as residue types and driving forces for structure folding. Therefore, to build high-performance statistical potentials for RNAs may still require the involvement of the structure characteristics of RNAs, in addition to utilizing similar modeling experience for proteins.

In this review, we will firstly introduce the derivation of general formulas for statistical potentials based on fundamental principles. Afterwards, we will give a brief overview on typical traditional statistical potentials for protein 3D structure

*Project supported by the National Natural Science Foundation of China (Grant Nos. 11774272 and 11575128). The numerical calculations related to our work in this review were performed on the supercomputing system in the Supercomputing Center of Wuhan University.

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

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

© 2021 Chinese Physical Society and IOP Publishing Ltd

prediction and evaluation due to the important reference role for RNAs. After that, we will introduce the recent advances in the statistical potentials for RNA 3D structure evaluation. Finally, we will emphasize a perspective on the further development of new statistical potentials with higher performance for RNA 3D structure evaluation.

2. Fundamental principles for statistical potentials

There are two fundamental principles for deriving and utilizing a statistical potential for proteins or RNAs: (i) the probability of a conformation of a biomacromolecule in a conformation ensemble obeys the Boltzmann's law which connects the effective potential energy of the system to its probability;[39,57,58] (ii) the native structure of a biomacromolecule corresponds to the conformation with the lowest free energy, which was first demonstrated by Anfinsen in 1973[59] and connects the free energy minimum of a biomacromolecule to its native structure.[58]

In addition, a native folded structure is actually driven and stabilized by various complex intramolecular and intermolecular interactions, while there would not exist a perfect potential energy function to exactly describe all these interactions for a biomacromolecule system.[58] Practically, a biomacromolecule system can be considered as an equilibrium ensemble of different atom (atom block) pair types with different types of characterizing geometrical parameters such as inter-atom contact, inter-atom distance, inter-atom angle, inter-block orientation, and so on. If a geometrical parameter is chosen, the probability of describing different values of the geometrical parameter can be utilized to build a statistical potential as a function of the geometrical parameter for different atom (or atom block) pair types.

2.1. Derivation of statistical potential

In principle, any kind of geometrical parameters, such as distances or angles between atoms which can be utilized to distinguish a native conformation from decoy ones, can be adopted to derive a statistical potential.[48] According to the inverse of Boltzmann's law, the potential energy for a particular atom pair type in the native state ensemble of biomacromolecules can be expressed by[40]

$$E^{\rm obs}(s) = -k_{\rm B}T\ln[P^{\rm obs}(s)] - k_{\rm B}T\ln[Z^{\rm obs}], \tag{1}$$

where $k_{\rm B}$ and T are the Boltzmann constant and the temperature in Kelvin, respectively. $E^{\rm obs}(s)$ and $P^{\rm obs}(s)$ are the energy and observed probability of a particular atom pair type for a geometrical parameter s in the native structure ensemble. $Z^{\rm obs}$ is the summation of Boltzmann factor over geometrical parameter s in the native state ensemble

$$Z^{\rm obs} = \sum_s \exp\left[-E^{\rm obs}(s)/k_{\rm B}T\right]. \tag{2}$$

However, $E^{\rm obs}(s)$ includes not only the interactions we focus on but also other interactions from surrounding medium.[60] Therefore, to obtain an effective statistical potential $\Delta E(s)$, a reference state needs to be involved. Similarly, according to Eq. (1), the inverse of Boltzmann's law for a reference state can be written as[40]

$$E^{\rm ref}(s) = -k_{\rm B}T\ln[P^{\rm ref}(s)] - k_{\rm B}T\ln[Z^{\rm ref}], \tag{3}$$

where $E^{\rm ref}(s)$ and $P^{\rm ref}(s)$ are the potential energy and the probability of the geometrical parameter s in a reference state, and $Z^{\rm ref}$ is the summation of Boltzmann factor over geometrical parameter s in the reference state ensemble

$$Z^{\rm ref} = \sum_s \exp\left[-E^{\rm ref}(s)/k_{\rm B}T\right]. \tag{4}$$

Therefore, a net statistical potential is obtained as a function of geometrical parameter s[40]

$$\Delta E(s) = E^{\rm obs}(s) - E^{\rm ref}(s), \tag{5}$$

and from Eqs. (1) and (3), $\Delta E(s)$ can be computed from the probability $P^{\rm obs}(s)$ and $P^{\rm ref}(s)$ by

$$\Delta E(s) = -k_{\rm B}T\ln\left[\frac{P^{\rm obs}(s)}{P^{\rm ref}(s)}\right] - k_{\rm B}T\ln\left[\frac{Z^{\rm obs}}{Z^{\rm ref}}\right]. \tag{6}$$

In addition, at a fixed temperature, $Z^{\rm obs}/Z^{\rm ref}$ is a constant and independent of the geometrical parameter s for a given amino/nucleic acid sequence,[40] and consequently $-k_{\rm B}T\ln[Z^{\rm obs}/Z^{\rm ref}]$ can be treated as zero point of the potential. Therefore, a general expression of a statistical potential can be obtained as[40]

$$\Delta E(s) = -k_{\rm B}T\ln\left[\frac{P^{\rm obs}(s)}{P^{\rm ref}(s)}\right]. \tag{7}$$

As shown in Eq. (7), the geometrical parameters s, which may involve two or more atoms (or atom group) and the reference states, are crucial for building statistical potentials, and the core difference between various statistical potentials is attributed to their choice.

2.2. Utilization of statistical potential

After deriving a statistical potential from Eqs. (1)–(7), the total energy $\Delta E(S,C)$ for a conformation C of a given sequence S relatively to the reference state can be given by[40]

$$\Delta E(S,C) = \sum \Delta E(s), \tag{8}$$

where the summation is over all atom (atom block) pairs with the additive assumption for statistical potentials.[40]

3. Statistical potentials for protein 3D structures

Since the pioneering work of Tanaka and Scheraga in 1976,[39] various strategies have been proposed to develop effective statistical potentials for proteins at atom or residue level.[40,42,43,61–64] These strategies can be roughly classified into the following categories: (i) developing various reference states or circumvent reference states;[43,65–70] (ii) classifying short-, medium-, and long-range interactions;[39,57,72–74] (iii) considering different geometrical parameters such as multi-body[63,75–79] and orientation-dependent interactions.[80–86] The geometrical parameters and major features of existing approaches employed in developing statistical potentials for proteins have been summarized in Table 1.

Table 1. The mainly existing approaches in developing statistical potentials for proteins.

Approaches	Geometrical parameters	Main features	References
Proposing reference states	Distance between atom/residue pairs	Simple and efficient; dependent on proposed reference state	[43,65–69]
Circumventing reference states	Distance between atom/residue pairs	Free of reference states; dependent on the quality of training decoy sets	[47,70,71]
Involving short-, medium-, and long-range interactions	Distance between atom/residue pairs; torsional angles of backbone	Differentiating short-, medium-, and long-range interactions; dependent on the selection of interaction boundary	[91,92,94]
Extracting multi-body interactions	Contact between atom/residue triplets or quadruplets	Capturing information of multi-body interactions; requiring abundant training set	[63,75,77–79,95–97,99–101]
Involving orientation-dependent interaction	Orientation between atom/residue block pairs	Capturing relative position between considered atom/residue block pairs; dependent on block orientation definition	[80–86]

3.1. Existing widely used reference states

As shown in the above derivation of statistical potentials, a reference state is necessarily required to build a statistical potential. An ideal reference state is a conformation ensemble composed of a non-redundant and complete spectrum of conformations in phase space and without non-bonded interactions between atoms.[48] Nevertheless, to obtain an ideal reference state might be a mission impossible. Thus, various approximations have been proposed to simulate the reference state to derive effective statistical potentials.[43,65–69] Generally, distance r between two atoms is treated as the geometrical parameter s to build statistical potentials, and correspondingly, $P^{\rm obs}(s)$ and $P^{\rm ref}(s)$ can be rewritten as $P_{ij}^{\rm obs}(r)$ and $P_{ij}^{\rm ref}(r)$ which represent the probability of the distance between atom types i and j within distance interval $[r, r+{\rm d}r]$ for the native and reference states, respectively. Here, the observed probability from the native structure ensemble can be given by[48]

$$P_{ij}^{\rm obs}(r) = \frac{N_{ij}^{\rm obs}(r)}{N_{ij}^{\rm obs}}, \tag{9}$$

where $N_{ij}^{\rm obs}(r)$ is the count of the distance between atom types i and j within distance interval $[r, r+{\rm d}r]$, and $N_{ij}^{\rm obs}$ is the summation of $N_{ij}^{\rm obs}(r)$ over r.

The existing reference states can be roughly classified into two types: based on experimental structures in PDB database[43,65,68] and based on statistical physics models.[66,67,69] The former type includes averaging reference state, quasi-chemical approximation reference state, and atom-shuffled reference state, and the latter type includes finite-ideal-gas reference state, spherical-non-interacting reference state, and random-walk-chain reference state.

3.1.1. On the basis of experimental structure database

Averaging reference state Samudrala and Moult proposed the RAPDF potential in 1988 by using the averaging reference state,[43] in which the distribution averaged over all atom pair types in experimental structures was used to approximately represent the distribution of different atom pair types in the reference state. Here, the probability $P_{ij}^{\rm ref}(r)$ is expressed as[43]

$$P_{ij}^{\rm ref}(r) = \frac{N_{ij}^{\rm ref}(r)}{N_{ij}^{\rm ref}} = \frac{\sum_{ij} N_{ij}^{\rm obs}(r)}{\sum_{ij}\sum_{r} N_{ij}^{\rm obs}(r)} = \frac{N^{\rm obs}(r)}{N^{\rm obs}}, \tag{10}$$

where $N^{\rm obs}(r)$ is the number of atom pairs within distance interval of $[r, r+{\rm d}r]$ from experimental structures in database, and $N^{\rm obs}$ is the summation of $N^{\rm obs}(r)$ over all distance intervals.

Quasi-chemical approximation reference state Lu and Skolnick assumed that the number of a certain atom pair type counted from the ideal reference state should be proportional to the molar fraction of the corresponding one from the experimental structures in database, and employed the quasi-chemical approximation reference state to drive the statistical potential of KBP.[65] Here, $P_{ij}^{\rm ref}(r)$ is given by[65]

$$P_{ij}^{\rm ref}(r) = \frac{N_{ij}^{\rm ref}(r)}{N_{ij}^{\rm ref}} = \frac{x_i x_j N^{\rm obs}(r)}{N_{ij}^{\rm obs}}, \tag{11}$$

where x_i is the mole fraction of atom type i in experimental structures, $N_{ij}^{\rm obs}$ is the counts of the pair of atom types i and j over all distance intervals, and $N^{\rm obs}(r)$ is the total count of all atom pairs within distance interval $[r, r+{\rm d}r]$.

Atom-shuffled reference state Shuffled reference states, including residue-shuffled, sequence-shuffled, and atom-shuffled modes, were proposed by Rykunov and Fiser.[68] Here, the atom-shuffled reference state is introduced, as the representative of shuffled reference states. In the reference state, all structures are obtained through the shuffled operation that the identities of all atoms in experimental structures are randomly shuffled while the coordinates of the atoms are kept fixed. Thus, $P_{ij}^{\mathrm{ref}}(r)$ can be written as[68]

$$P_{ij}^{\mathrm{ref}}(r)=\frac{N_{ij}^{\mathrm{shuffled}}(r)}{N_{ij}^{\mathrm{shuffled}}}. \tag{12}$$

Here, $N_{ij}^{\mathrm{shuffled}}(r)$ and $N_{ij}^{\mathrm{shuffled}}$ are both counted over the shuffled structures from experimental ones.

From the above, in the experimental structure-based reference states, the information of experimental structures is used to simulate the reference states through summing up atom types, involving molar fraction of atom types, and shuffling atom/residue types, and consequently, these reference states can be easily modeled given the native structure database. However, this type of reference states naturally excludes a huge number of nonnative conformations in phase space, and would differ significantly from the ideal reference state in conformation space.

3.1.2. On the basis of statistical physics models

Finite-ideal-gas reference state By applying the ideal-gas-like pair distribution function to model that of the reference state for proteins, Zhou and Zhou developed the statistical potential of Dfire.[66] The pair distribution function for proteins in the ideal gas state is given by[88]

$$g_{ij}(r)=\frac{N_{ij}^{\mathrm{obs}}(r)/4\pi r^2\Delta r}{N_iN_j/V}, \tag{13}$$

where $N_{ij}^{\mathrm{obs}}(r)$ is the number of the pair of atom types i and j within distance interval $[r,r+\mathrm{d}r]$, N_i and N_j are the total number of atom types i and j over all the distance bins, respectively. V is the volume of a protein. Since the potential energy between atom types i and j, $\Delta E_{ij}(r)$, equals to $-k_{\mathrm{B}}T\ln g_{ij}(r)$[88] and $\Delta E_{ij}(r)=0$ for the distance between these two considered atoms larger than a cutoff distance r_{c}, $P_{ij}^{\mathrm{ref}}(r)$ can be derived as[66]

$$P_{ij}^{\mathrm{ref}}(r)=\frac{N_{ij}^{\mathrm{obs}}(r_{\mathrm{c}})}{N_{ij}^{\mathrm{obs}}}\left(\frac{r}{r_{\mathrm{c}}}\right)^{\alpha}=P_{ij}^{\mathrm{obs}}(r_{\mathrm{c}})\left(\frac{r}{r_{\mathrm{c}}}\right)^{\alpha}. \tag{14}$$

Here, a dimension parameter α is involved since macromolecule systems are actually not ideal gas even at high temperature.

Spherical-non-interacting reference state Shen and Sali proposed the spherical-non-interaction reference state for proteins,[67] in which a native structure is replaced by a sample sphere with the same radius of gyration as the native structure. Since this reference state takes the native structures of different sizes into account, $P_{ij}^{\mathrm{ref},p}(r)$, calculated by a single native structure p, can be expressed by[67]

$$\begin{aligned}P_{ij}^{\mathrm{ref},p}(r)&=P_{ij}^{\mathrm{ref},p}(r,a)\\&=\begin{cases}\dfrac{3r^2(r-2a)^2(r+4a)}{16a^6}, & r_{\mathrm{c}}>2a;\\ \dfrac{6r^2(r-2a)^2(r+4a)}{r_{\mathrm{c}}^3(r_{\mathrm{c}}^3-18a^2r_{\mathrm{c}}+32a^3)}, & r_{\mathrm{c}}\le 2a,\end{cases}\end{aligned} \tag{15}$$

where r_{c} is the cutoff distance, and a is the radius of the sphere which has the same radius of gyration R_{g} as the experimental protein structure. Through taking a weighted average of $P_{ij}^{\mathrm{obs},p}(r)/P_{ij}^{\mathrm{ref},p}(r)$ over all proteins, the final statistical potential can be calculated by[67]

$$\Delta E_{ij}(r)=-k_{\mathrm{B}}T\ln\left[\sum\nolimits_p w_p\frac{P_{ij}^{\mathrm{obs},p}(r)}{P_{ij}^{\mathrm{ref},p}(r)}\right], \tag{16}$$

where the weight w_p of a sampled protein p is defined by the ratio between the number of all atom pairs in the sampled structure and the number of atom pairs in all samples, and $P_{ij}^{\mathrm{obs},p}(r)$ is the probability counted from the sample native structure.

Random-walk-chain reference state To simulate the intrinsic connectivity of protein chains, Zhang and Zhang proposed the random-walk-chain reference state.[69] According to the polymer theory of freely-joined chain model, $P_{ij}^{\mathrm{ref},p}(r)$, calculated for one single native chain, can be expressed as[69]

$$\begin{aligned}P_{i,j}^{\mathrm{obs},p}(r)&=P^{\mathrm{ref},p}(r)=\int P^{\mathrm{ref},p}(r,n)\,\mathrm{d}n\\&=\sum\nolimits_{n=1}^{N}4\pi r^2\left(\frac{3}{2\pi nl^2}\right)^{3/2}\exp\left(-\frac{3r^2}{2nl^2}\right)\Delta r,\end{aligned} \tag{17}$$

where N is the number of residues in an experimental structure p. l is Kuhn length, and l is an adjustable parameter to scale the free-joint chain to a realistic protein chain. $P_{ij}^{\mathrm{ref}}(r)$ is equal to the summation of $P_{ij}^{\mathrm{ref},p}(r)$ over all experimental structures. Finally, it is assumed that the potential energy between atoms of types i and j $\Delta E_{ij}(r)=0$ for inter-atom distance larger than a cutoff distance r_{c}, and $P_{ij}^{\mathrm{ref}}(r)$ can be derived as[69]

$$\begin{aligned}P_{ij}^{\mathrm{ref}}(r)&=\sum_p P_{i,j}^{\mathrm{obs},p}(r_{\mathrm{c}})\left(\frac{r}{r_{\mathrm{c}}}\right)^2\\&\quad\times\frac{\sum_{n=1}^{N}\exp\left(-3r^2/2nl^2\right)/n^{3/2}}{\sum_{n=1}^{N}\exp\left(-3r_{\mathrm{c}}^2/2nl^2\right)n^{3/2}}.\end{aligned} \tag{18}$$

Thus, the statistical potential based on the random-walk-chain reference state can be obtained according to Eqs. (7) and (18).

In this type of reference states, different from experimental structure-based reference states, various statistical physics models such as ideal gas-like model, spherical cluster model, and random-walk-chain are proposed to simulate the reference

state instead of the experimental 3D structures. Thus, for protein systems, the physical-model-based reference states can better capture the huge conformation space than the experimental structure-based reference states and could have better performance. Furthermore, the extra parameters are involved in the reference states, e.g., dimension parameter α and Kuhn length l, which can better match the spatial scale of physical models to that of a realistic protein system. However, these physical-model-based reference states ignore the structure details of proteins which may differ visibly from an ideal-gas or random-walk chain system.

3.2. Circumvention of reference states

Since it is difficult to simulate the ideal reference state, to circumvent the problem, Huang and Zou developed an iterative strategy to build their statistical potentials for protein-ligand binding, protein–protein and intra-protein interactions.[47,70,71] In the strategy, the iterative starts from a set of initial potentials[47]

$$\Delta E_{ij}^{0} = -k_{\mathrm{B}}T \ln \frac{\rho_{ij}^{\mathrm{obs}}(r)}{\rho_{ij}^{\mathrm{bulk}}}, \tag{19}$$

where $\rho_{ij}^{\mathrm{obs}}(r)$ is the observed number density of pair of atom types i and j in the distance interval $[r-\Delta r/2, r+\Delta r/2]$ in the native structures database, and $\rho_{\mathrm{bulk}}^{\mathrm{obs}}$ is the average number density of pair of atom types i and j in a reference sphere of radius $R_{\max}$. The initial potential $\Delta E_{ij}^{0}(r)$ is not far away from the true potential $\Delta E(r)$, which can help the iterative process to avoid being trapped at local energy minima. The difference between the predicted pair distribution $g_{ij}(r)$ and the observed pair distribution $g_{ij}^{\mathrm{obs}}(r)$ was narrowed step by step, and until the potential $\Delta E_{ij}^{(n)}(r)$ could distinguish the native structures from decoys, the iterative process was terminated. Here, the iterative equation is given by[47]

$$\Delta E_{ij}^{(n+1)}(r) = \Delta E_{ij}^{(n)}(r) + \frac{1}{2}k_{\mathrm{B}}T\left[g_{ij}(r) - g_{ij}^{\mathrm{obs}}(r)\right]. \tag{20}$$

Through the iteration process based on a statistical mechanics modeling, this method can circumvent the problem of reference states. Essentially, the observed pair distribution $g_{ij}^{\mathrm{obs}}(r)$ serves as a criterion for the performance of the statistical potential and the iteration process can yield statistical potentials with high performance,[47,70,71] although this method may strongly require good training sets containing both native and decoy structures for the iteration process.

3.3. Classification of short-, medium-, and long-range interactions

Tanaka and Scheraga have classified the inter-residue interactions into short-, medium-, and long-ranged ones for proteins,[39] and the criteria for interaction ranges of proteins are based on the separations between residues along protein sequences. Afterwards, the preference of amino acids to form short-, medium-, and long-range contacts has been analyzed for different types of proteins,[72–74,89] and several potentials have also been proposed based on distinguishing interaction ranges.[57,72,91–94] The classification of interaction ranges may be very necessary since such treatment can involve more accurate information extracted from native structure database in different interaction ranges.

3.3.1. Backbone torsion-based potentials

To measure the influence of amino acids on the backbone conformation of neighboring residues along protein sequences, backbone torsion-dependent potentials were proposed by Rooman *et al.*[91] and Kocher *et al.*[92] A succession of dihedral angles ϕ, ψ, and ω of each residue represent the backbone of proteins, and these dihedral angles can be clustered into 7 domains.[91] Based on Eqs. (7) and (8), the total torsional energy for a protein conformation can be expressed as the summation of residue-to-torsion potentials along the protein conformation[92]

$$\Delta E(S,C) = -k_{\mathrm{B}}T \sum_{i,j,k=1}^{N} \frac{1}{\xi_k} \ln \frac{P_{i-k,j-k}^{a_i,a_j}(t_k)}{P_{i-k,j-k}(t_k)}, \tag{21}$$

where $P_{i-k,j-k}^{a_i,a_j}(t_k)$ is the observed probability that pairs of amino acids a at positions i and j are associated with the torsion domain t at position k. $P_{i-k,j-k}(t_k)$ corresponds to the sequence nonspecific probability of t_k. N is the number of residues for a protein chain, and ξ_k is a normalization factor. Furthermore, the torsion-dependent potentials were divided into a short-range one and a medium-ranged one, in which only residues and torsion domain in the intervals $[k-1, k+1]$ and $[k-8, k+8]$ were taken into account, respectively. Except for the above described residue-to-torsion potential, a torsion-to-residue potential was proposed, and can be obtained from the probability $P_{i-k,j-k}^{t_i,t_j}(a_k)$ in which pairs of torsion domains t at positions i and j are associated with the amino acid a at position k. The torsion-dependent potentials can well describe the interactions in protein surface and can well predict the stability of surface mutations.[93]

3.3.2. Distance-dependent potentials

Similar with the backbone torsion-based potentials, distance-dependent interactions have also been divided into two contributions: local (short- and medium-ranged) interactions and nonlocal (long-ranged) interactions.[92,94] The local/nonlocal interactions can be derived from $P_{|i-j|}^{ai,aj}(r)$, which represents the propensities of two residues a_i and a_j with a spatial distance r and within residue separation $|i-j|$ along sequences. In the work of Kocher *et al.*,[92] the potentials between residue pairs separated by 1–7 sequence position were

considered as medium-ranged potentials and those between residue pairs separated by more than 8 residues were treated as nonlocal (long-ranged) ones. In the work of Melo *et al.*, the residue separation of 9 was regarded as the boundary of local and nonlocal interactions.[94]

3.4. Extraction of multi-body interactions

Beyond widely used pairwise two-body potentials, multi-body potentials, especially for three- and four-body potentials, have been proposed for capturing higher-order interactions.[63,75–79,95–97] In addition, several geometrical techniques, such as Voronoi diagram, Delaunay triangulation,[98] and alpha shape[76,96] were employed to define the neighboring bodies and to identify the multi-body interactions.

3.4.1. Three-body potentials

Three-body contacts have been introduced in several studies where physical models were developed with explicitly incorporating a three-body interaction.[95,96,99–101] Based on Delaunay triangulation, Li and Liang used the alpha shape to define three-body interactions for protein structures.[96] Here, we briefly introduce this approach as an example of this class of statistical potential. In the model of Li and Liang, volume overlap was used to identify three-body contacts, and such volume overlap occurs if three atoms from nonbonded residues share a Voronoi vertex.[96] Furthermore, Li and Liang introduced a nonadditive coefficient by measuring the deviation of three-body interactions from three independent pairwise interactions.[96] The nonadditive coefficient v of three-body interactions was defined as[96]

$$v(i,j,k) \equiv \frac{P(i,j,k)}{P'(i,j,k)}, \tag{22}$$

where $P(i,j,k)$ is the three-body propensity for residues of types i, j, and k in native structure database, and $P'(i,j,k)$ is the propensity if the three-body interaction is the simple consequence of three independent pairwise contact interactions. The expressions of $P(i,j,k)$ and $P'(i,j,k)$ are written as follows:[96]

$$P(i,j,k) \equiv \frac{q(i,j,k)}{p(i,j,k)}, \tag{23}$$

$$\begin{aligned} P'(i,j,k) &= P(i,j)\cdot P(j,k)\cdot P(i,k) \\ &= \frac{q(i,j)}{p(i,j)}\cdot\frac{q(j,k)}{p(j,k)}\cdot\frac{q(i,k)}{p(i,k)}. \end{aligned} \tag{24}$$

Here, $P(i,j)$ is the pairwise contact propensity for residue types i and j, and $P(i,j)$ equals to the odds ratio of the observed probability $q(i,j)$ for two-body contact of residues i and j to the excepting probability $p(i,j)$ for two-body contact of residues i and j. The observed probability $q(i,j,k)$ and expecting $p(i,j,k)$ are for three-body contact, respectively. The details of calculating $p(i,j)$, $q(i,j)$, $q(i,j,k)$, and $q(i,j,k)$ can be found in Ref. [96]. Finally, the potential energy $\Delta E(M)$ for a molecule M modeled as a set of fused hard spheres can be expressed as[96]

$$\Delta E(M) = -\sum_{\sigma_{ij}\in\kappa} \ln P(i,j) - \sum_{\sigma_{ijk}\in\kappa} \ln v(i,j,k), \tag{25}$$

where $\sigma_{ij} \in \kappa$ represents pairwise alpha contacts, and $\sigma_{ijk} \in \kappa$ represents three-body alpha contacts.

3.4.2. Four-body potentials

To capture more complicated interactions, several four-body potentials have been developed,[63,75,77–79,95,97] and Delaunay tessellation algorithms were widely used for finding the nearest neighbors in native structures. Krishnamoorthy and Tropsha have verified that their four-body potentials derived with Delaunay tessellation performed better than pairwise two-body potentials in distinguishing correct sequences/structures and generating Z-scores.[77] Later, considering more detailed interactions between backbones and side chains and including some of the sequential information of proteins, Feng *et al.* have proposed a new four-body statistical potential based on Delaunay tessellation.[78] Instead of using Delaunay tessellation, Gniewek *et al.* have used a simple geometric construction to develop a four-body contact potential, and they found that the performance of their optimized coarse-grained contact potentials was comparable to the performance of DFIRE for all proteins larger than 80 amino acids.[79]

As shown above, due to the huge computational complexity and limited structure data, most existing multi-body potentials are based on local residue contact rather than distance and with the growth of protein structures deposited in database, multi-body distance-dependent statistical potentials can be developed for capturing higher order interactions in protein structures.

3.5. Involvement of orientation-dependent interaction

Atoms in proteins generally carry charges and thereby there can exist orientation-dependent correlations between polar atom blocks due to dipole–dipole-like interaction and associated polarization effect.[85] Consequently, orientation-dependent correlation between atom blocks has been considered as a complementary contribution to above described statistical potentials. Recently, Wang and Huang have shown that the involvement of orientation contribution can visibly improve the performance of the statistical potential.[86] In geometry, the relative position and orientation between two 3D objects need to be determined by six degrees of freedom, including three translations and three rotations, or one distance and five angular variables.[83] Hence, the definition and calculation approaches for the orientation between two blocks are more diverse than those for the distance of two blocks, and a

more adequate native structure database is needed for developing orientation-dependent statistical potentials.

3.5.1. Definition of relative orientation

Ideally, for proteins, an amino acids residue can be represented by several blocks, and each block is centered by each of heavy atoms. As shown in Fig. 1(a), the relative orientation of a block pair is described by five angles: θ_i, φ_i, θ_j, φ_j, and ω_{ij}. The first four angles are the polar coordinates of the $\boldsymbol{r}_{ij}$ vector in the local 3D reference frame of each amino acid and ω_{ij} describes the relative rotation of $\boldsymbol{V}_z$ vectors of two amino acids along the $\boldsymbol{r}_{ij}$ axis. The local 3D reference frame is defined as

$$\begin{aligned} \boldsymbol{V}_z &= \frac{(\boldsymbol{r}_{12}+\boldsymbol{r}_{13})}{|\boldsymbol{r}_{12}+\boldsymbol{r}_{13}|}, \\ \boldsymbol{V}_y &= \frac{\boldsymbol{V}_z \times \boldsymbol{r}_{13}}{|\boldsymbol{V}_z \times \boldsymbol{r}_{13}|}, \\ \boldsymbol{V}_x &= \boldsymbol{V}_y \times \boldsymbol{V}_z, \end{aligned} \tag{26}$$

where $\boldsymbol{r}_{12} = \boldsymbol{r}(i_1) - \boldsymbol{r}(i)$ and $\boldsymbol{r}_{13} = \boldsymbol{r}(i_2) - \boldsymbol{r}(i)$ are the relative vectors from atom i to atoms i_1 and i_2, respectively; $\boldsymbol{V}_z$ is the normal vector to the plane within which $\boldsymbol{V}_x$ and $\boldsymbol{V}_z$ lie.

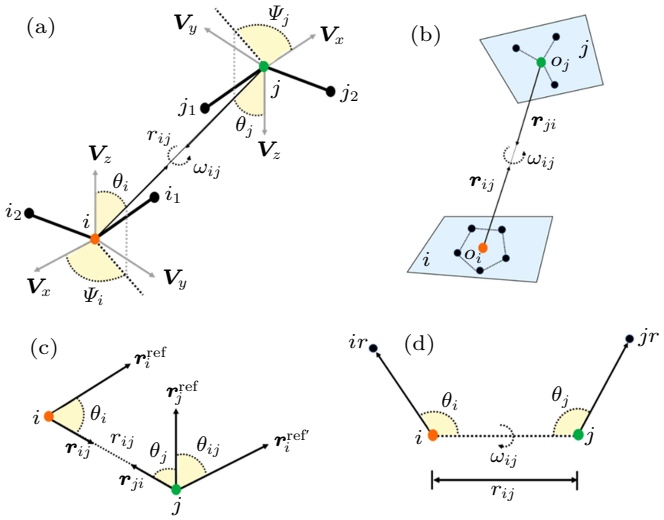


Fig. 1. (a) The complete definition of relative position and orientation between two blocks centered by atoms i and j, respectively, in which one scalar distance r_{ij} and five scalar angles $(\theta_i, \Psi_i, \theta_j, \Psi_j, \omega_{ij})$ parameters are used. (b) The definition of relative position and orientation between two blocks i and j for OPUS-PSP, in which two direction vectors $\boldsymbol{r}_{ij}$ and $\boldsymbol{r}_{ji}$ and one scalar inter-rotation angle ω_{ij} parameters are used. (c) The definition of relative position and orientation between two polar atoms i and j for dDFIRE, in which one scalar distance r_{ij} and three scalar angles $(\theta_i, \theta_j, \theta_{ij})$ parameters are used. (d) The definition of relative position and orientation between two atoms i and j for ORDER_AVE, in which one scalar distance r_{ij} and three scalar angles $(\theta_i, \theta_i, \omega_{ij})$ parameters are used.

Afterwards, by taking the relative position and orientation as geometrical parameters, equation (7) can be rewritten as

$$\Delta E_{ij} = -k_{\mathrm{B}}T \ln \left[\frac{P_{ij}^{\mathrm{obs}}(r_{ij}, \theta_i, \varphi_i, \theta_j, \varphi_j, \omega_{ij})}{P_{ij}^{\mathrm{ref}}(r_{ij}, \theta_i, \varphi_i, \theta_j, \varphi_j, \omega_{ij})} \right], \tag{27}$$

where P_{ij}^{obs} and P_{ij}^{ref} are the joint probability at the given relative distance and orientation of observing two atoms of types i and j, for native and reference states, respectively. However, this six-dimensional joint probability distribution might bring the problem that the data of experimental structures are not sufficient to accurately describe the 6D joint distribution of a specific pair of block types.

3.5.2. Recent developed orientation-dependent potentials

A perfect definition of relative orientation for proteins is so detailed that building an orientation-dependent potential would experience the challenge of sparse experimental data and long computational quantity. The existing strategies used to overcome the problem mainly include: (i) reducing representations of proteins; (ii) reducing variables of defining the orientation; and (iii) reducing dimensions of the joint probability distribution. In this subsection, several recent orientation-dependent statistical potentials are briefly introduced as follow.

OPUS-PSP OPUS-PSP potential, proposed by Lu *et al.*, defines orientations for blocks of atoms bonded rigidly within the same residue.[80] In OPUS-PSP potential, 20 amino acid residues are decomposed into 19 rigid-body block types, in which all heavy atoms are assumed to be in the same plane, and each residue contains several blocks. As shown in Fig. 1(b), the relative orientation of blocks i and j respectively with the origins o_i and o_j is determined by two relative direction vectors $\boldsymbol{r}_{ij}$ and $\boldsymbol{r}_{ji}$, and an inter-rotation angle ω_{ij}. For the differences among these 19 rigid-body blocks, the definition of local 3D reference frame for each block type and inter-rotation angle is specified, which can be found in Lu *et al.*[80] Finally, the orientation-dependent energy function can be expressed by[80]

$$E(\Omega_{ij}, i, j) = -k_{\mathrm{B}}T \ln \frac{P_{ij}^{\mathrm{obs}}(\Omega_{ij}, i, j)}{P_{ij}^{\mathrm{ref}}(\Omega_{ij}, i, j)}, \tag{28}$$

where P_{ij}^{obs} and P_{ij}^{ref} are the probabilities at the given orientation state for two block types i and j in contact for native and reference states, respectively, and the "block contact pairs" are considered by averaging number of heavy atoms pairs in contact (when the distance between two atoms is less than 5 Å, they are defined as in contact). $\Omega_{ij} = (r_{ij}, \boldsymbol{r}_{ji}, \omega_{ij})$ defines the relative orientation of block types i and j. In OPUS-PSP, the quasi-chemical approximation was used as the reference state.

dDFIRE Yang and Zhou introduced the orientation dependence of polar atoms into the DFIRE potential and derived the dDFIRE potential, and some improvements were observed over DFIRE in modeling the refolding of protein terminal regions with secondary structure.[85] In dDFIRE potential, a simplified definition involving four variables for an atom pair was used instead of the complete definition of orientation involving six variables. As shown in Fig. 1(c), a reference vector $\boldsymbol{r}_i^{\mathrm{ref}}$ ($\boldsymbol{r}_j^{\mathrm{ref}}$) of a given polar atom i (j) is produced by summing

the bond vectors from the polar atom to other chemically connected heavy atoms. Then, one distance r_{ij} and three angles θ_i, θ_j, θ_{ij} can be generated to define the relative orientation between polar atoms i and j, where r_{ij} is the distance between two atoms, θ_{ij} is the angle between $\boldsymbol{r}_i^{\rm ref}$ and $\boldsymbol{r}_j^{\rm ref}$, θ_i is the angle between $\boldsymbol{r}_i^{\rm ref}$ and the distance vector $\boldsymbol{r}_{ij}$, and θ_j is the angle between $\boldsymbol{r}_j^{\rm ref}$ and the distance vector $\boldsymbol{r}_{ij}$. Furthermore, to reduce the amount of required experimental structure data, it is assumed that these three angles θ_i, θ_j, and θ_{ij} are independent of each other. Thus, the orientation-dependent potential can be written as[85]

$$\begin{aligned} & E_{ij}^{\rm dDFIRE}(r_{ij},\theta_i,\theta_j,\theta_{ij}) \\ = \; & E_{ij}^{\rm DFIRE}(r_{ij}) + E_{ij}(\theta_i|r_{ij}) \\ & + E_{ij}(\theta_i|r_{ij}) + + E_{ij}(\theta_{ij}|r_{ij}) \\ = \; & -k_{\rm B}T\left(\ln\frac{P_{ij}^{\rm obs}(r_{ij})}{P_{ij}^{\rm ref}(r_{ij})} + \ln\frac{P_{ij}^{\rm obs}(\theta_i|r_{ij})}{P_{ij}^{\rm ref}(\theta_i|r_{ij})}\right. \\ & \left. + \ln\frac{P_{ij}^{\rm obs}(\theta_j|r_{ij})}{P_{ij}^{\rm ref}(\theta_j|r_{ij})} + \ln\frac{P_{ij}^{\rm obs}(\theta_{ij}|r_{ij})}{P_{ij}^{\rm ref}(\theta_{ij}|r_{ij})}\right). \end{aligned}$$

In dDFIRE, the finite-ideal-gas reference state was used as the reference state.

ORDER_AVE ORDER_AVE is a four-dimensional joint probability distribution with five atoms per residue.[102] As shown in Fig. 1(d), for atom pair i and j, two heavy atoms ir and jr that directly connect to i and j, were selected, and the orientations of atoms i and j can be described by four variables, r_{ij}, α, β, and γ. Here, r_{ij} is the distance between atoms of types i and j; α and β are the scalar angles of the atom groups ir–i–j and i–j–jr, respectively; and γ is the dihedral angle of atom groups ir–i–j–jr. Finally, the orientation-dependent energy ΔE_{ij} for the atom pair i and j can be expressed as[102]

$$\Delta E_{ij}(r_{ij},\alpha,\beta,\gamma) = -k_{\rm B}T\ln\frac{P_{ij}^{\rm obs}(r_{ij},\alpha,\beta,\gamma)}{P_{ij}^{\rm ref}(r_{ij},\alpha,\beta,\gamma)}, \tag{29}$$

where $P_{ij}^{\rm ref}$ was calculated based on the averaging reference state.

KORP KORP, a side-chain independent potential, was defined by a 6D joint probability that only depends on the relative orientation and position of three backbone atoms.[83] As shown in Fig. 1(a), six parameters including a distance r_{ij} and five angles $(\theta_i,\varphi_i,\theta_j,\varphi_j,\omega_{ij})$ are used. However, it is quite challenging to obtain a detailed 6D joint probability based on the relatively small number of experimental structures. Thus, a constant z was added to prevent infinite values for very small probabilities and to improve the numerical stability for low-count statistics, and the orientation-dependent potential can be given by[83]

$$\Delta E_{ij} = -k_{\rm B}T\ln\frac{P_{ij}^{\rm obs}(r_{ij},\theta_i,\varphi_i,\theta_j,\varphi_j,\omega_{ij})+z}{P_{ij}^{\rm ref}(r_{ij},\theta_i,\varphi_i,\theta_j,\varphi_j,\omega_{ij})+z}, \tag{30}$$

where $P_{ij}^{\rm ref}$ was calculated based on the averaging reference state, which simply averages over the 20 different amino acid types to represent the reference probability regardless of residue types.

The above described involvements of orientation-dependent potentials can provide a very valuable contribution to a statistical potential,[80,83,85,102] while due to sparse experimental data and long computational cost, the quality of an orientation-dependent potential may strongly rely on the definition of relative orientation for atom groups in proteins.

4. Statistical potentials for RNA 3D structures

In the recent decade, as the number of experimental RNA 3D structures deposited in PDB database[8] increases, several statistical potentials have been derived based on known structures for RNA structure prediction and assessment,[37,49–56] partially with the aid of the experience for proteins. However, the existing statistical potentials for RNAs are significantly less than those for proteins and still do not reach a satisfactory performance. In the following, we will give an overview on the recently developed statistical potentials for RNA 3D structure evaluation and will emphasize the possible development on statistical potentials for RNAs in the near future.

4.1. Existing statistical potentials for RNAs

To our knowledge, the existing statistical potentials for RNA 3D structure evaluation[50–56] were developed (partially) with the aid of the experience or those in protein structure prediction and assessment. The existing statistical potentials for RNAs have been organized in Table 2, and will be introduced explicitly in the following.

4.1.1. Distance- and torsion angle-dependent statistical potentials

RASP Based on the averaging reference state, Capriotti *et al.* have built RASP at both coarse-grained and all-atom (23 clustered atom types) levels from a non-redundant training set of 85 RNA structures, and distance between atom pairs was considered as the geometrical parameter.[51] Capriotti *et al.* have shown that RASP has a better performance than NAST[37] which is a nucleotide-level coarse-grained statistical potential composed of bond, angle, dihedral, and non-bond terms.

KB potential Using the quasi-chemical approximation reference state, Bernauer *et al.* have derived distance-dependent and fully differentiable statistical potentials at both coarse-grained and all-atom (85 atom types) levels from a training set of 77 native RNAs.[50] Owing to the implementation of Dirichlet process mixture models, the statistical potentials are fully differentiable, which makes them applicable for molecular dynamics simulations.

3dRNAscore Different from RASP and KB potentials, 3dRNAscore is composed of the distance- and torsion angle-dependent potentials based on the averaging reference state.[52] In 3dRNAscore, 85 atom types and 7 torsion angle types were defined, and a set of 317 RNA native structures was used for training the potential.[52] In addition, to appropriately consider the contributions of the two kinds of energy terms, a weight factor w has been optimized by four typical RNA decoys.[52] The test of 3dRNAscore has shown that, the dihedral-dependent potential alone performs visibly worse than the distance-dependent potential averagely, while the combined potential with dihedral-dependent and distance-dependent contributions generally performs better than the distance-dependent potential.[52] Thereby, the dihedral-dependent potential has been verified useful to improve the accuracy of a distance-dependent potential.

According to the results reported by Wang *et al.*,[52] overall, the performance of RASP is worse than that of KB potential and 3dRNAscore, and 3dRNAscore has the best performance among these three statistical potentials.[52] The mainly reasons might be that the KB potential of all-atom and 3dRNAscore are based on 85 atom types, while only 23 clustered atom types are involved in the all-atom version of RASP. In addition, the best performance of 3dRNAscore might be attributed to the explicit emphasis on the local torsional structure feature of RNA backbones which can be important for RNA 3D structures.

4.1.2. Comparisons of different reference states for RNA 3D structure evaluation

For RNAs, only the averaging and quasi-chemical approximation reference states were used to build statistical potentials until our very recent study.[55] To understand the existing widely used six reference states for RNAs, we performed a comprehensive examination on these six reference states, including the averaging, quasi-chemical approximation, atom-shuffled, finite-ideal-gas, spherical-non-interacting, and random-walk-chain reference states; see also Subsection 3.1. In the work,[55] using the same training set and parameters, e.g., bin width and distance cutoff, we constructed six all-atom (85 atom types) distance-dependent statistical potentials corresponding to the six reference states, and examined these potentials against three RNA test sets; see Table 2. Our extensive examinations showed that, overall, for identifying native structures and ranking decoy structures, the finite-ideal-gas[66] and random-walk-chain[69] reference states were slightly better than others, while for identifying near-native structures, there was slight difference among the six reference states. Moreover, compared among RASP, KB potential, and 3dRNAscore, 3dRNAscore has the similar performance to the top ones of these six statistical potentials, while RASP and KB potential have visibly lower performance than the others.[55]

Table 2. The existing statistical potentials for RNA structure evaluation.

Reference states	Statistical potentials	Geometrical parameters	Atom types	Capacity of training set	References
Averaging	RASP-ALL	Distance between atom pairs	23	85	[51]
	Avg-REF	Distance between atom pairs	85	109	[55]
	3dRNAscore	Distance between atom pairs and torsional angles of backbone	85	317	[52]
Quasi-chemical approximation	All-atom KB potential	Distance between atom pairs	85	77	[50]
	QChA-REF	Distance between atom pairs	85	109	[55]
Atom-shuffled	Ash-REF	Distance between atom pairs	85	109	[55]
Finite-ideal-gas	FIG-REF	Distance between atom pairs	85	109	[55]
Spherical-non-interacting	SNI-REF	Distance between atom pairs	85	109	[55]
Random-walk-chain	RWC-REF	Distance between atom pairs	85	109	[55]
Multinomial reference distribution	RAMP	Atomic quadruplet interaction	4	85	[56]
Free of reference states	RNA3DCNN	Surrounding environment of each nucleotide (atomic occupation number, mass, and charge within the grid box)	85	414	[53]

4.1.3. Four-body contact potential

Masso has developed a four-body contact potential of RAMP, which is the first multibody statistical potential for RNAs.[56] In RAMP, atomic four-body nearest-neighbors were generated by the Delaunary tessellation[98] for each structure in a training set of 85 native RNAs, and each RNA was represented by four atom types of C, N, O, and P.[56] Thus,

35 distinct quadruplet types can be produced by 4-letter atomic alphabet. With atomic quadruplet contacts as geometric parameter and the inversed Boltzmann's law, equation (7) can be rewritten as[56]

$$\Delta E_{ijkl} = -k_{\mathrm{B}}T \ln \frac{P^{\mathrm{obs}}_{ijkl}}{P^{\mathrm{ref}}_{ijkl}}, \tag{31}$$

where P^{obs}_{ijkl} is the observed probability of atomic quadruplet types (i,j,k,l) within an edge-length cutoff (i.e., 8 Å or 10 Å) in the native state, and P^{ref}_{ijkl} is the expected probability of atomic quadruplet types (i,j,k,l) within an edge-length cutoff for a reference state. With a multinomial reference distribution, P^{ref}_{ijkl} can be computed by[56]

$$P^{\mathrm{ref}}_{ijkl} = \frac{4!}{\prod_{n=1}^{4}(t_n!)} \prod_{n=1}^{4} a_n^{t_n}. \tag{32}$$

Here, t_n is the number of atom type n occurred in atomic quadruplet type (i,j,k,l), and $\sum_{n=1}^{4}(t_n) = 4$. a_n is the relative frequency that atom of type n occurs in native RNA structures database, and $\sum_{n=1}^{4}(a_n) = 1$. However, overall, RAMP has a worse performance than RASP-ALL[97] in identifying native structures.[56] The main reason might be that RAMP is a coarse-grained-level contact potential with only 4 clustered atom types, while RASP-ALL is a distance-dependent potential with 23 clustered atom types.

4.1.4. Utilization of 3D convolutional neural network

Completely diverged from the above-described traditional statistical potentials, Li *et al*. have employed a 3D convolutional neural network to develop two 3DCNN-based scoring models, named RNA3DCNN_MD and RNA3DCNN_MDMC, for assessing near-native RNA decoys and RNA decoys with large root-mean-square-deviation (RMSD) fluctuation, respectively.[53] Two training sets were generated by molecular dynamics (MD) simulations and Monte Carlo (MC) structure prediction for each of 414 RNAs. RNA3DCNN_MD was trained by the first set, while RNA3DCNN_MDMC was trained by the training set combined the two training sets. The test showed that, for decoys with RMSD less than 1.0 Å, the performance of RNA3DCNN was similar to or worse than that of traditional statistical potentials, while for decoys generated by RNA structure predictive models, RNA3DCNN performed obviously superior to other existing statistical potentials in identifying native structures.[53]

4.2. Limitation of current statistical potentials and perspective for RNAs

According to Eq. (1), a set of pairwise potentials can be extracted from the distance distribution of atom pairs by the inverse of Boltzmann's law. However, an individual pairwise potential involves not only the interaction between two atoms but also the influence of the surrounding medium on the interaction.[60] Therefore, reference states and other approaches have been developed aiming to eliminate redundant information. Based on the experiences in protein structure prediction and assessment from which the existing RNA statistical potentials benefit, we will make a brief discussion for developing new statistical potentials for RNA structure evaluation.

4.2.1. On reference states

The existing six reference states can be classified into two types: based on experimental RNA 3D structures[43,65,68] and based on physical modeling.[66,67,69] The physical-model-based reference states are built by various physical models, and they can better describe the conformation space of RNA decoys than the 3D-structure-based reference states. However, the physical-model-based reference states are still deviate significantly from the ideal one due to the ignorance of detailed structure features, and our recent study shows that the performance of statistical potentials based on physical-model-based reference states is still not at a high level for realistic decoys from RNA structure prediction models.[55] For example, the ideal-gas reference state completely neglects the connectivity of RNA chains, and a random-walk reference state ignores the intrinsic rigidity of RNA chains. Moreover, both of the reference states neglect side chains and atom types. Therefore, it is still necessary to propose more realistic reference states for RNAs or circumvent the problem of reference states.

One possible way to simulate more realistic RNA chains is to involve the chain rigidity of RNAs, e.g., an RNA chain is modeled as a semiflexible polymer rather than a completely random polymer.[103,104] Another possible way is to develop a combined reference state composed of the physical-model-based and structure-based reference states which may complement each other. A third possible way is to develop an ensemble-based reference state of ideal-gas-like states of different dimension parameters or of random-walk-chain states of difference Kuhn lengths, since backbone atoms and side chain ones may behave very differently. Due to the difficulty in modeling an ideal reference state, an alternative way is to circumvent the problem. Huang and Zou have proposed an iterative method to develop scoring functions for protein–ligand and protein–protein docking.[61,62] This method provides a possible way to circumvent reference states for modeling statistical potentials for RNAs, although the performance may rely strongly on the high-quality decoy training set which should be generated by various structure prediction models.

4.2.2. On short-, medium-, and long-range interactions

Tanaka and Scheraga[39] have classified the inter-residue interactions into short-, medium-, and long-ranged ones for proteins, and afterwards, the preference of amino acids to

form medium- and long-range contacts between has been analyzed for proteins.[72–74,89] Until now, most statistical potentials for RNAs do not distinguish the contributions of short-, medium-, and long-ranged interactions. However, these different residue-separation-ranged interactions should play different roles in RNA folding process or in stabilizing folded RNA structures. For example, the short-, medium-, and long-ranged potentials may correspond to the interactions within loops, those stabilizing secondary segments, and those stabilizing tertiary contacts for RNAs. Although the local interactions were split from non-local interactions in RASP by Capriotti *et al.*,[51] the local potential was built at each nucleotide separation which is too subtle to avoid sparse data problem and might not be able to represent specific topological features. Therefore, the classification of short-, medium-, and long-ranged interactions is in accordance with the folding and structure features of RNAs and may be valuable for developing statistical potentials of higher performance for RNA structure evaluation.

4.2.3. On geometrical parameters

Compared with pairwise two-body potential, multi-body potential could capture more complex 3D interactions.[78,96] Up to now, RAMP is the only RNA four-body statistical potential, but its overall performance is not better than that of RASP-ALL,[56] which may be attributed to that RAMP only involves four atom types to represent RNA structures and RAMP is a (spatially) local contact potential ignoring spatially nonlocal interactions. Therefore, for building a multi-body potential, a proper geometrical parameter may be very important which should not only describe the key geometrical relation between multi-body atoms, but also be conveniently used to build a multi-body potential based on limited native structure database. Another key problem in modeling multi-body potential may be the coarse-graining of RNA atoms which should not only retain the major structure feature of RNAs, but also effectively decrease the complexity of multi-body potentials and the demand for structure data in training set.

On the other hand, most existing RNA statistical potentials are distance-dependent, expect for 3dRNAscore which is composed of the distance- and the local torsional angle-dependent terms.[52] The better performance of 3dRNAscore than prior traditional statistical potentials is mainly attributed to the explicit emphasis on the local torsional structure feature of RNA backbone. Hence, other geometrical parameters can be involved into distance-dependent statistical potentials to improve the performance, such as orientation, which has been highlighted important for proteins;[86] see Subsection 3.5. Another geometrical parameter can be the nonlocal torsional angle beyond one nucleotide, which may also improve the description for RNA backbone and be helpful for a distance-dependent potential. Owing to current limited experimental RNAs, it is very important to define the orientation variables, the importance of which has been discussed above. Based on current limited experimental RNAs, it is very important to define the orientation variables and deal with the dependence among the variables. Compared to 20 amino acids with 167 atoms for proteins, RNAs has 4 nucleotides with 85 atom types, which may make it easier to build high-performance orientation-dependent potentials for RNAs.

4.2.4. On unique feature of RNAs

From the above, the existing statistical potentials for RNAs were developed mainly according to those for proteins. There are some important similar structure features between RNAs and proteins. For examples, first, RNAs and proteins are both polymers with backbones and side chains. Second, there are distinctively ordered secondary segments in RNA and protein structures, such as helices and loops for RNAs, and α-helices, β-sheets, and loops for proteins; third, the tertiary structures of RNA and proteins are both very compact and approximately sphere-like. Such similarity between RNA and proteins structures has enabled the great benefit of the existing statistical potentials for RNAs from those for proteins.

However, there are also several distinctive differences between RNAs and proteins. The major forces for stabilizing structures are very different for RNAs and proteins. RNAs are generally stabilized by strong secondary interactions of base-pairing/stacking and moderate tertiary interactions, while proteins are generally stabilized by hydrophilic and hydrophobic interactions. Consequently, the secondary and tertiary interactions for RNAs may be roughly classified while those for proteins are mutually coupled. The separate involvement of secondary and tertiary contributions in a statistical potential may improve its performance. Furthermore, base-stacking can promote the helical orientation of RNAs and consequently RNA structures are apparently less spherical than protein structures,[105,106] which may suggest that angular or orientation-dependent contribution is more important for a statistical potential for RNAs.

Very importantly, RNAs are strongly charged polyanion,[106–109] which is distinctively different from proteins generally with certain positively and negatively charged domains. Generally, for RNAs, metal ions are essentially required to neutralize the negative charges on RNA backbones, and consequently would favor the folding of compact native structures. Furthermore, cations with different valences and concentrations could have very different influences on the deformation of folded RNA structures.[108,109,111–113] For example, due to polyanionic nature, RNA structures can become less compact at lower salt and an RNA can have visibly more compact structure at multivalent salt than at monovalent

salt.[32,33,35] Therefore, the involvement of ion-electrostatic interaction in a statistical potential may be helpful for improving its performance for RNA 3D structure evaluation, especially for the NMR structures.[55] Finally, machine and deep learning methods which could dig critical and uneasy to be detected RNA feature information are also able to be applied to build statistical potentials.[53]

5. Conclusions

Since various computational RNA 3D structure prediction models have appeared in recent years, scoring functions are essentially needed for evaluating the predicted structure candidates. Because statistical potentials for protein 3D structure prediction and evaluation have been developed from a long time ago, there are lots of experience which can be very helpful for RNAs. In this review, we firstly introduced the fundamental principles of deriving the formulas of statistical potentials. Afterwards, we gave a brief view on the development of statistical potentials for proteins, and we emphasize the choice of geometrical parameters and reference states which are crucial for building different statistical potentials. Furthermore, we introduced the recent advances in statistical potentials for RNAs. Finally, we make a detailed discussion on possible further improvements of statistical potentials for RNAs, which may be helpful for developing high-performance statistical potentials for RNA 3D structure evaluation.

Acknowledgements

We are grateful to Professors Shi-Jie Chen (University of Missouri), Xiaoqin Zou (University of Missouri), Jian Zhang (Nanjing University), Yi Xiao (Huazhong University of Science and Technology), and Sheng-You Huang (Huazhong University of Science and Technology) for valuable discussions.

References

[1] Gesteland R F, Cech T R and Atkins J F 2006 *Q. Rev. Biol.* **81** 272
[2] Dethoff E A, Chugh J, Mustoe A M and Alhashimi H M 2012 *Nature* **482** 322
[3] Guttman M and Rinn J L 2012 *Nature* **482** 339
[4] Doherty E A and Doudna J A 2001 *Annu. Rev. Biophys. Biomol. Struct.* **30** 457
[5] Edwards T E, Klein D J and Ferré-D'Amaré A R 2007 *Curr. Opin. Struct. Biol.* **17** 273
[6] Aviv T, Amborski A N, Zhao X S, Kwan J J, Johnson P E, Sicheri F and Donaldson L W 2006 *J. Mol. Biol.* **356** 274
[7] Baird N J, Ludtke S J, Khant H, Chiu W, Pan T and Sosnick T R 2010 *J. Am. Chem. Soc.* **132** 16352
[8] Rose P W, Prlić A, Altunkaya A, Bi C, Bradley A R, Christie C H, Costanzo L D, Duarte J M, Dutta S, Feng Z, Green R K, Goodsell D S, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose A S, Shao C, Tao Y P, Valasatava Y, Voigt M,Westbrook J D,Woo J, Yang H, Young J Y, Zardecki C, Berman H M and Burley S K 2017 *Nucleic Acids Res.* **45** D271
[9] Miao Z, Adamiak R W, Antczak M, *et al.* 2017 *RNA* **23** 655
[10] Sun L Z, Zhang D and Chen S J 2017 *Annu. Rev. Biophys.* **46** 227
[11] Schlick T and Pyle A M 2017 *Biophys. J.* **113** 225
[12] Shi Y Z, Wu Y Y, Wang F H and Tan Z J 2014 *Chin. Phys. B* **23** 078701
[13] Ding F, Sharma S, Chalasani P, Demidov V V, Broude N E and Dokholyan N V 2008 *RNA* **14** 1164
[14] Parisien M and Major F 2008 *Nature* **452** 51
[15] Jossinet F, Ludwig T E and Westhof E 2010 *Bioinformatics* **26** 2057
[16] Das R, Karanicolas J and Baker D 2010 *Nat. Methods* **7** 291
[17] Cao S and Chen S 2011 *J. Phys. Chem. B* **115** 4216
[18] Rother M, Rother K, Puton T and Bujnicki J M 2011 *Nucleic Acids Res.* **39** 4007
[19] Zhao Y, Huang Y, Gong Z, Wang Y, Man J and Xiao Y 2012 *Sci. Rep.* **2** 734
[20] 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
[21] Zhang J, Bian Y, Lin H and Wang W 2012 *Phys. Rev. E* **85** 21909
[22] Xia Z, Bell D R, Shi Y and Ren P Y 2013 *J. Phys. Chem. B* **117** 3135
[23] Kim N, Laing C, Elmetwaly S, Jung S, Curuksu J, Schlick T 2014 *Proc. Natl. Acad. Sci. USA* **111** 4079
[24] Xu X, Zhao P and Chen S 2014 *PloS one* **9** e107504
[25] Shi Y Z, Wang F H, Wu Y Y and Tan Z J 2014 *J. Chem. Phys.* **141** 105102
[26] Shi Y Z, Jin L, Wang F H, Zhu X L and Tan Z J 2015 *Biophys. J.* **109** 2654
[27] Boniecki M J, Lach G, Dawson W K, Tomala K, Lukasz P, Soltysinski T, Rother K M and Bujnicki J M 2016 *Nucleic Acids Res.* **44** e63
[28] Li J, Zhang J, Wang J, Li W and Wang W 2016 *PLoS Comput. Biol.* **12** e1005032
[29] Jain S and Schlick T 2017 *J. Mol. Biol.* **429** 3587
[30] Bell D R, Cheng S Y, Salazar H and Ren P 2017 *Sci. Rep.* **7** 45812
[31] Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y and Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
[32] Jin L, Shi Y Z, Feng C J, Tan Y L and Tan Z J 2018 *Biophys. J.* **115** 1403
[33] Shi Y Z, Jin L, Feng C J, Tan Y L and Tan Z J 2018 *PLoS Comp. Biol.* **14** e1006222
[34] Zhang B, Qiu H, Jiang J, Liu J and Shi Y Z 2019 *J. Chem. Phys.* **151** 165101
[35] Jin L, Tan Y L, Wu Y, Wang X, Shi Y Z and Tan Z J 2019 *RNA* **25** 1532
[36] Zhang J, Dundas J, Lin M, Chen R, Wang W and Liang J 2009 *RNA* **12** 2248
[37] Jonikas M A, Radmer R J, Laederach A, Das R, Pearlman S, Herschlag D and Altman R B 2009 *RNA* **15** 189
[38] Zhang D and Chen S J 2018 *J. Chem. Theory Comput.* **14** 2230
[39] Tanaka S and Scheraga H A 1976 *Macromolecules* **9** 945
[40] Sippl M J 1990 *J. Mol. Biol.* **213** 859
[41] Thomas P D and Dill K A 1996 *J. Mol. Biol.* **257** 457
[42] Melo F and Feytmans E 1997 *J. Mol. Biol.* **267** 207
[43] Samudrala R, Moult J 1998 *J. Mol. Biol.* **275** 895
[44] Skolnick J, Kolinski A and Ortiz A 2000 *Proteins* **38** 3
[45] Buchete N V, Straub J E and Thirumalai D 2004 *Curr. Opin. Struct. Biol.* **14** 225
[46] Chi Z, Song L, Zhu Q and Zhou Y Q 2005 *J. Med. Chem.* **48** 2325
[47] Huang S Y and Zou X Q 2011 *Proteins* **79** 2648
[48] Deng H, Jia Y, Wei Y and Zhang Y 2012 *Proteins: Struct. Funct. Genet.* **80** 2311
[49] Dima R I, Hyeon C and Thirumalai D 2005 *J. Mol. Biol.* **347** 53
[50] Bernauer J, Huang X, Sim A Y L and Levitt M 2011 *RNA* **17** 1066
[51] Capriotti E, Norambuena T, Marti-Renom M A and Melo F 2011 *Bioinformatics* **27** 1086
[52] Wang J, Zhao Y, Zhu C and Xiao Y 2015 *Nucleic Acids Res.* **43** e63
[53] Li J, Zhu W, Wang J, L W, Gong S, Zhang J and Wang W 2018 *PLoS Comput. Biol.* **14** e1006514
[54] Yang Y, Gu Q, Zhang B, Shi Y Z, Shao Z G 2018 *Chin. Phys. B* **27** 38701
[55] Tan Y L, Feng C J, Jin L, Shi Y Z and Tan Z J 2019 *RNA* **25** 793
[56] Masso M J 2018 *Theor Biol.* **453** 58
[57] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534
[58] Sippl M J 1993 *J. Comput. Aid Mol. Des.* **7** 473
[59] Anfinsen C B. 1973 *Science* **181** 223
[60] Sippl M J 1995 *Curr. Opin. Struct. Biol.* **5** 229
[61] Nishikawa K and Matsuo Y O 1993 *Protein Eng.* **6** 811
[62] Kocher J P A, Rooman M J and Wodak S J 1994 *J. Mol. Biol.* **235** 1598
[63] Singh R K, Tropsha A and Vaisman I I 1996 *J. Comp. Biol.* **3** 213
[64] Miyazawa S and Jernigan R L 2005 *J. Chem. Phys.* **122** 24901

[65] Lu H and Skolnick J 2001 *Proteins: Struct. Funct. Bioinform.* **44** 223
[66] Zhou H and Zhou Y 2002 *Protein Sci.* **11** 2714
[67] Shen M Y and Sali A 2006 *Protein Sci.* **15** 2507
[68] Rykunov D and Fiser A 2010 *BMC Bioinformatics* **11** 128
[69] Zhang J and Zhang Y 2010 *PloS One* **5** e15386
[70] Huang S Y and Zou X Q 2008 *Proteins: Struct. Funct. Bioinform.* **72** 557
[71] Huang S Y and Zou X Q 2006 *J. Comput. Chem.* **27** 1866
[72] Gromiha M M and Selvaraj S 2004 *Prog. Biophys. Mol. Biol.* **86** 235
[73] Gromiha M M and Selvaraj S. 1999 *Biophys. Chem.* **77** 49
[74] Gromiha M M and Selvaraj S. 1997 *J. Biol. Phys.* **23** 151
[75] Carter C W, Lefebvre B C, Cammer S A, Tropsha A, Edgell M H 2001 *J. Mol. Biol.* **311** 625
[76] Li X, Hu C, Liang J 2003 *Proteins: Struct. Funct. Bioinform.* **53** 792
[77] Krishnamoorthy B and Tropsha A 2003 *Bioinformatics* **19** 1540
[78] Feng Y, Kloczkowski A and Jernigan R L 2007 *Proteins: Struct. Funct. Bioinform.* **68** 57
[79] Gniewek P, Leelananda S P, Kolinski A, Jernigan R L and Kloczkowski A 2011 *Proteins: Struct. Funct. Bioinform.* **79** 1923
[80] Lu M, Dousis A D and Ma J 2008 *J. Mol. Biol.* **376** 288
[81] Yu Z, Yao Y, Deng H and Yi M 2019 *BMC Bioinformatics* **20** 211
[82] Zhou H and Skolnick J 2011 *Biophys. J.* **101** 2043
[83] López-Blanco J R and Chacón P 2019 *Bioinformatics* **35** 3013
[84] Kortemme T, Morozov A V and Baker D 2003 *J. Mol. Biol.* **326** 1239
[85] Yang Y and Zhou Y 2008 *Proteins: Struct. Funct. Bioinform.* **72** 793
[86] Wang X and Huang S Y 2019 *J. Chem. Inf. Model.* **59** 3080
[87] Buchete N V, Straub J E and Thirumalai D 2004 *Protein Sci.* **13** 862
[88] Friedman H L 1985 *A Course in Statistical Mechanics* (Prentice-Hall, Englewood Cliffs)
[89] Kumarevel T S, Gromiha M M, Selvaraj S, Gayatri K, Kumar P K R 2002 *Biophys. Chem.* **99** 189
[90] Miyazawa S and Jernigan R L 1999 *Proteins* **36** 347
[91] Rooman M J, Kocher J P A and Wodak S J 1992 *Biochemistry* **31** 10226
[92] Kocher J P, Rooman M J and Wodak S J 1994 *J. Mol. Biol.* **235** 1598
[93] Gilis D and Rooman M 1997 *J. Mol. Biol.* **272** 276
[94] Melo F, Sanchez R and Sali A 2002 *Protein Sci.* **11** 430
[95] Munson P J and Singh R K 1997 *Protein Sci.* **6** 1467
[96] Li X and Liang J 2005 *Proteins: Struct. Funct. Bioinform.* **60** 46
[97] Masso M 2017 *Biomed. Res. Int.* 5760612
[98] Barber C B, Dobkin D P and Huhdanpaa H 1996 *Acm. T. Math. Software* **22** 469
[99] Eastwood M P and Wolynes P G 2001 *J. Chem. Phys.* **114** 4702
[100] Rossi A, Micheletti C, Seno F and Maritan A 2001 *Biophys. J.* **80** 480
[101] Godzik A and Skolnick J 1992 *Proc. Natl. Acad. Sci. USA* **89** 12098
[102] Liu Y, Zeng J and Gong H 2014 *Proteins* **82** 2383
[103] Wu Y Y, Bao L, Zhang X and Tan Z J 2015 *J. Chem. Phys.* **142** 125103
[104] Zhang X, Bao L, Wu Y Y, Zhu X L and Tan Z J 2017 *J. Chem. Phys.* **147** 054901
[105] Hyeon C, Dima R I and Thirumalai D 2006 *J. Chem. Phys.* **125** 194905
[106] Xiong G, Xi K, Zhang X and Tan Z J 2018 *Chin. Phys. B* **27** 018203
[107] Chen S J 2008 *Annu. Rev. Biophys.* **37** 197
[108] Tan Z J and Chen S J 2010 *Biophys. J.* 2010 **99** 1565
[109] Tan Z J and Chen S J 2011 *Biophys. J.* **101** 176
[110] Lipfert J, Doniach S, Das R and Herschlag D 2014 *Annu. Rev. Biochem.* **83** 813
[111] Wu Y Y, Zhang Z L, Zhang J S, Zhu X L and Tan Z J 2015 *Nucleic Acids Res.* **43** 6156
[112] Xi K, Wang F H, Xiong G, Zhang Z L and Tan Z J 2018 *Biophys. J.* **114** 1776
[113] Bao L, Zhang X, Jin L and Tan Z J 2016 *Chin. Phys. B* **25** 18703