ISSN: 0256-307X

中国物理快报
Chinese Physics Letters

Volume 27 Number 11 November 2010

A Series Journal of the Chinese Physical Society
Distributed by IOP Publishing

Online: http://www.iop.org/journals/cpl
http://cpl.iphy.ac.cn

CHINESE PHYSICAL SOCIETY

CHIN. PHYS. LETT. Vol. 27, No. 11 (2010) 118702

An RNA Base Discrete State Model toward Tertiary Structure Prediction *

ZHANG Jian(张建)**, ZHANG Yu-Jie(张玉洁), WANG Wei(王炜)**

National Laboratory of Solid State Microstructure, Department of Physics, Nanjing University, Nanjing 210093

(Received 27 May 2010)

We report a new ribonucleic acid (RNA) base discrete state model, which was first developed in our lab and designed to provide an efficient and accurate way of representing RNA structures toward RNA three-dimensional structure predictions. Since RNA free energy is largely determined by base pairs and base stackings instead of backbone trajectories, we directly model the RNA base configurations with respect to its previous one along the sequence. This is in sharp contrast with all previous works where the backbone trace was represented. To test how faithfully the discrete model can reproduce the chain trace in continuous space, we randomly select partial chains from the native structure of 23S ribosome RNA and re-grow them. The rms distance of the re-grown structures from the native ones is ~ 1.7 *Å for an optimized 16-state discrete model and gradually increases to* ~ 3.3 *Å for long chains of length 50. The efficiency is also good, e.g. the program will finish within several tens of second for long loops of length 50. Our model may facilitate the RNA three-dimensional structure predictions in the near future when combined with appropriate free energy evaluation methods.*

PACS: 87.14.Gk, 87.15.Bg, 87.64.Aa DOI: 10.1088/0256-307X/27/11/118702

RNA is more than a messenger between genome and protein. The biological functions of RNA range from carrying genetic information, participating in protein synthesis, catalyzing biochemical reactions, and regulating gene expressions, to acting as a structural molecule in cellular organelles.[1] For understanding how RNA molecules perform these tasks, knowledge of three-dimensional structures of RNA is often required. Although the most reliable sources of RNA structural information are from experiments, including x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryoelectron microscopy, experimental measurements of RNAs are technically challenging and costly in both time and effort. Computational prediction of RNA structures, therefore, can provide an alternative source of structural information.

Prediction of RNA secondary structures based on comparative sequence analysis or free energy minimization has been very successfull.[2–8] The predicted secondary structures provide useful insights into the RNA functions. However, to fully appreciate the RNA functions and dynamics, and to design new RNA relevant drugs, people eventually need their three-dimensional structures.

For RNA three-dimensional structure predictions, model building using homology information is a preferred strategy. However, due to the technical difficulties associated with RNA experiments, the number of experimentally determined RNA structures deposited in protein database (PDB) is few, with less than 800 pure RNA structures and about 2600 protein-nucleic acid (including DNA and RNA) complexes as of the year 2010. This limits the usage of the homology model building strategy. Another promising method is by minimizing the free energy (MFE) of RNA tertiary structures, based on the physical hypothesis that the native structures of RNAs have the lowest free energies. The MFE method basically includes two components, a structure sampling method and a free energy estimation strategy, as we often see in protein folding studies.[9–13]

To facilitate structure sampling, a structure representation model is needed. Previously, most RNA structure models were designed to represent the RNA backbones.[14–17] However, due to the additional degrees of freedom associated with sugar rings and bases (including purines or pyrimidines), the RNA backbones built by these models cannot uniquely determine the base-pair and base-stacking patterns, which however determine the free energy of the RNA to a large extent. Therefore such backbone-based models need an additional side-chain-rotamer-like library or rely on all-atom simulations to finally pin down the configurations of RNA bases in order to accurately calculate the corresponding free energies.

In this Letter, we report a new RNA base discrete state model, which is developed to model the RNA base structures and base-pair patterns directly. This model is designed to reflect the fact that the free energy of a given structure is largely determined by the base-pair and base-stacking patterns. Instead of giving the positions of backbone atoms, our model provides the positions of RNA bases (purines or pyrimidines), according to a pre-built database of spatial distances and orientations between successive nucleotides. The structures built by our model can thus be used directly to calculate the free energy contributed by base-pairs and base-stackings when combined with appropriate free energy estimation methods.

Our RNA base model is basically a discrete model that describes the relative spatial orientations of the next nucleotide with respect to its previous one along

*Supported by the National Natural Science Foundation of China (10504012, 10834002 and 10974088) and Jiangsu Province (No BK2009008).

**Email: jzhang@nju.edu.cn,wangwei@nju.edu.cn

the sequence; this information is extracted from an RNA structural database and will be used to build RNA chains. To extract such information, we first find an RNA database, which was built by Murray *et al.*[18] from Duke University and named RNA05. The RNA05 database contains 171 coordinate files that were selected from the Nucleic Acid Database as of February 2005. To characterize the spatial relationship between successive nucleotides, we chop the RNA chains into successive nucleotide pairs. For example, if the nucleotides in a chain is indexed from 1 to N along the sequence from $5'$ to $3'$, we then have totally $N-1$ successive pairs, $1 \to 2$, $2 \to 3$, $\cdots$, and $N-1 \to N$. The arrow's direction is along that of the sequence. We then partition these pairs into four classes, namely, purine→purine, purine→pyrimidine, pyrimidine→purine, pyrimidine→pyrimidine. For each nucleotide pair, we setup a local coordinate system on the first nucleotide, with the origin at the centroid of the base heavy atoms, the x axis passing through the N1 atom (purines) or N3 atom(pyrimidines), and the z axis perpendicular to the base plane.[17] Within this local coordinate system we calculate the coordinates of the base heavy atoms of the second nucleotide. We then cluster these local configurations with a K-mean clustering method. The K (K can be 4, 8, 16, 32 or even larger numbers in principle) centers of these clusters thus characterize the spatial position of a nucleotide with respect to its previous one along the sequence. The position vectors of these center nucleotides and their spatial orientations with respect to the first one are recorded in a library and will be used to grow RNA chains. We call this method of clustering and discretizing RNA successive bases the RNA base discrete K-state model, where K is the number of centers preset in the K-mean clustering procedure.

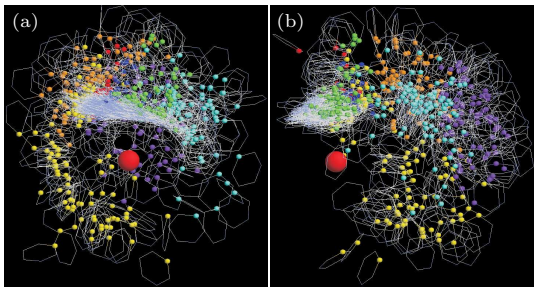

Fig. 1. The spatial positions of the second nucleotide bases with respect to their previous ones along the sequence in all 1855 pyrimidine-pyrimidine pairs. (a) Front view, (b) side view. The N1 atoms in the bases are represented by small spheres. The different colors denote different clusters the nucleotides belong to. The K value is set to be 8 when calculating this figure. The large red sphere is at the origin of the local coordinate system, determined by the first nucleotide.

From the RNA05 database, we obtained 2713 purine→purine pairs, 2133 purine→pyrimidine pairs, 2097 pyrimidine→purine pairs, and 1855 pyrimidine→pyrimidine pairs. As an example, we plot the K-mean clustering results for the pyrimidine→pyrimidine pairs in Fig. 1, which shows the spatial positions of the second pyrimidine with respect to the first one and the clusters they belong to. In Fig. 1, the spheres represent the N1 atoms in the pyrimidines and different colors denote different clusters the nucleotides belong to. Here the largest cluster is marked by red small spheres; their number is few for the first sight because most of them overlap with each other in space. This cluster corresponds to the A-form conformation of the RNA chains.

We tested how well our discrete model can reproduce the trace of RNA chains in continuous space by mapping the original RNA conformation into the discrete space associated with the RNA discrete state model and check their similarity. Here by "mapping" we mean that we try to find the most similar structure to the original one among all possible structures in the discrete space.

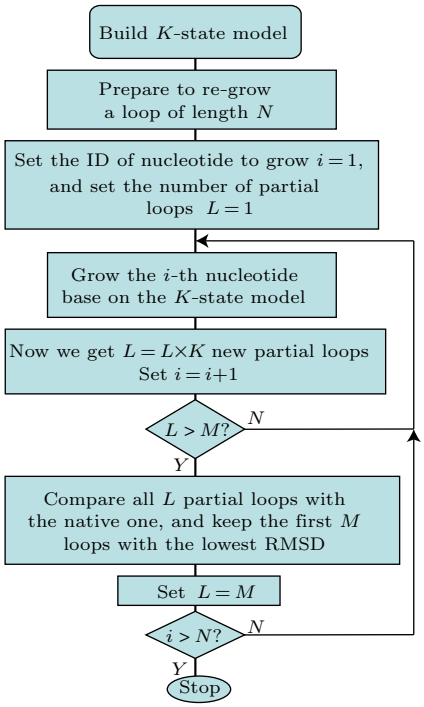

Fig. 2. The flowchart showing how we grow an RNA chain of length N with a K-state discrete model. We take the parameters as follows: i, the sequence identify of the nucleotide; L, the current number of partial chains; M, the maximum number of partial chains that we keep.

We took the 23S ribosome RNA (PDB ID: 1S72)[19] and a heuristic "build-up" algorithm[7,20] to do this test. The testing procedure is shown by the flowchart in Fig. 2. In detail, we randomly select a fragment of length N from the ribosome RNA, then we re-grow this RNA chain in single nucleotide increments starting from the $5'$ to the $3'$ terminus. Suppose that at the ith step we already have L partial chains, then we add a new nucleotide to each of these partial chains with all the K possible positions tried. This gives $L \times K$ partial chains, of which only the M structures with the lowest rms distance[21] from the partial original structure are retained. This process

is repeated until the last nucleotide in the fragment is added. The chain that best represents the original RNA chain is chosen such that it has the lowest rms distance with respect to the original among the final M structures. Note that the above-described procedure is not a blind prediction, since we need the native structure to guide the growth. This procedure is only to test whether the discrete model can reproduce the trace of original continuous chains faithfully, a prerequisite ability for being used to model RNA chains and to do prediction.

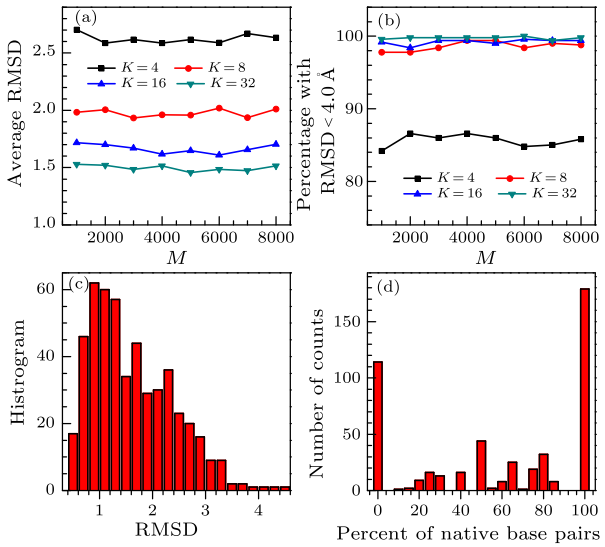

Fig. 3. The performance of the RNA base discrete state model as functions of K and M, which is the number of configurations we keep each time when we add a new nucleotide to the partial fragment. Large values of K or M mean higher accuracy whereas lower running efficiency. (a) The average rms distance on 500 times of calculation. (b) The percentage of rms distances that are smaller than 4 Å in the 500 times of calculations. (c) The distribution of rms distances with $K = 16$, $M = 1000$. (d) The distribution of percentage of native base pairs with $K = 16$, $M = 1000$.

We performed the test for different combinations of K and M values; the results are shown in Fig. 3. For each combination, the calculation was performed 500 times for different randomly chosen fragments of length 8 taken from the 23S ribosome RNA (PDB ID: 1S72). It can be seen that with the increase of K from 4 to 32, the average rms distance of the re-grown structures with respect to the original one is decreased, from ~ 2.6 Å to ~ 1.5 Å. This trend is expected since with the increase of K, the model becomes more flexible and can accommodate more structural changes. Figure 3(b) also shows this trend, with the increase of K, the percent of re-grown chains with the rms distance smaller than 4 Å is increased from $\sim$ 85% for $K = 4$ to $\sim$ 100% for $K = 32$. According to Fig. 3, it can be concluded that $K = 16$ is a good choice for the discrete model, as a compromise between accuracy and efficiency. We can also see that the re-growing algorithm is not sensitive to the choice of M, which is a good aspect and allows us to use a smaller M for the sake of running efficiency.

Besides the comparison of rms distances, we also tested how well the discrete model can reproduce the native base pairs. To determine if two bases form a base pair, we define $\boldsymbol{O}_i$ as the vector of the center of base ring of nucleotide i and $\boldsymbol{n}_i$ as the normal vector of the plane of the same ring. Note that when calculating these parameters, only the N1, C2, N3, C4, C5 and C6 atoms of the nucleotides are used, no matter whether they are A, U, G, or C. To judge if two bases of nucleotides i and j form a pair, we define θ as the angle between their normal vectors $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$, and $\boldsymbol{d} = \boldsymbol{O}_i - \boldsymbol{O}_j$, $d_z = \boldsymbol{d} \cdot \boldsymbol{n}_i$, and $d_{xy} = |\boldsymbol{d} - d_z \boldsymbol{n}_i|$. If the following three criteria are met, i.e., $\theta < 20°$, $d_z < 1.1$ Å, and $d_{xy} < 6.1$ Å, the two bases are deemed forming a base pair. The base pairs determined from the native structure (PDB 1S72) are called native base pairs. We tested the performance of the discrete model by calculating the percent of correctly reproduced base pairs in the regrown chains over the native ones. Note that to partially account for the less flexibility of the discrete model, a factor of 1.2 was multiplied on the above criteria when judging the base pairs in the regrown chains, in contrast to the factor of 1.0 when calculating native pairs from the PDB structure. The testing results for a chain of length 8 with parameters of $K = 16$ and $M = 1000$ are shown in Figs. 3(c) and 3(d), where the distribution of rms distances as well as the percentage of native base pairs in the regrown chains are given, respectively. It can be seen that for most regrown chains (95%), the rms distance is smaller than 3.0 Å. However, the percent of reproduced base pairs is relatively low - about 50% of regrown chains have correctly reproduced $> 80\%$ native base pairs. This problem is attributed to the discrete nature of the model and may be partly solved by further refining the obtained structure using molecular dynamics. The comparison of the results of rms distances and base pairs also indicates that the native pair formation is a far more stringent criterion for judging the quality of structures.

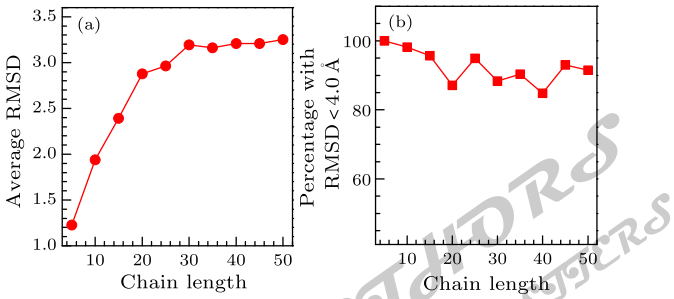

Fig. 4. The performance of the RNA base discrete model as a function of chain length. (a) The averaged rms distance on 500 calculations. (b) The percentage of rms distances that has a value smaller than 4 Å among 500 calculations.

We also studied the performance of the model as a function of chain length; the results are shown in Fig. 4. It can be seen that as the chain length is in-

creased from a few nucleotides to 50, the performance does deteriorate but not that much. The average rms distance is increased from 1.2 Å to ∼ 3.2 Å, and the percentage of rms distances that is smaller than 4 Å is decreased slightly, from 100% to ∼ 90%. However, although the performance becomes slightly worse, it is still acceptable since most of the RMSDs are still smaller than 4 Å.

To visualize the performance of the discrete state model, we compare the re-grown structure with the original one and plot them in Fig. 5. Figure 5(a) shows the best case when $K = 16$ and the chain length is 8; the RMSD between two structures is 0.5 Å. In contrast, Fig. 5(b) shows the worst case, corresponding to the rms distance of 4.4 Å. It can be seen that if the original chain resembles an A-form structure, the discrete-state model faithfully reproduces the original structure, with the re-grown bases locating closely to the corresponding original one. However, if the original structure has sharp turns (the case in Fig. 5(b) has three of them, a tough test), the mapping to discrete space is not very successful. However, it is still acceptable with a RMSD of 4.4 Å.

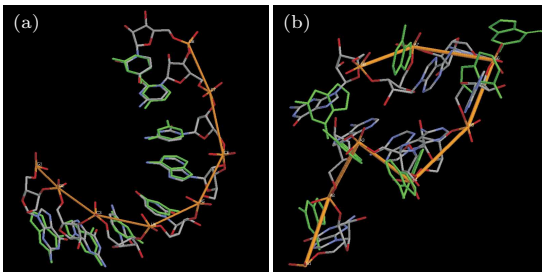

Fig. 5. The superposition of the original RNA structure and the mapped one, which is colored green. The K is set to be 16 and the length is 8. (a) The best case with the rms distance between two structures equal to 0.5 Å. (b) The worst case with the rms distance of 4.4 Å.

Figure 6 shows the testing results for a much longer RNA chain, with the chain length increased to 50-nucleotides. Figure 6(a) corresponds to the best case among 500 calculations (rms distance 3.0Å), and Fig. 6(b) the worst case, with the RMSD = 5.7 Å. Taking consideration of the long chain length, the rms distance of 3.0 Å in the first case is really a promising result. In the latter case, the deviation of the mapped structure from the native one mainly locates at the middle part, where the RNA bases are not paired with other bases and thus highly flexible, according to the native structure. Presumably, the RNA chain in this region undergoes large fluctuations and continuous changing of conformations, and may not be well represented by any single structure.

The efficiency of our model is good in terms of the speed of chain regrowth. For a chain of length 8 with $K = 16$ and $M = 4000$, the regrowing procedure needs only 3 s. For a much longer chain of 50 with the same other parameters, the calculation will finish within 50 s. This efficiency is paramount for high-throughput structural sampling in RNA tertiary structure predictions. Furthermore, our model directly gives the positions of RNA bases, allowing the direct calculation of the free energy of RNA chains without resorting to other procedures such as all-atom structural refinement of the configuration of bases. This feature further increases the efficiency of the model.

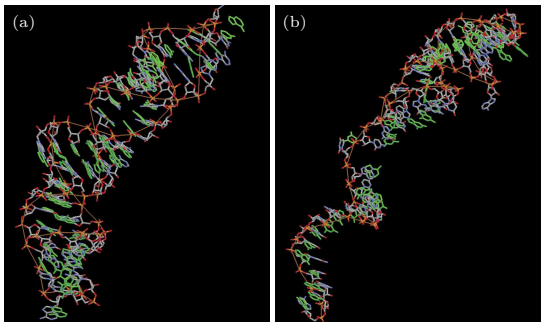

Fig. 6. The superposition of the original RNA structure and the mapped one, which is colored green. The K is set to 16 and the length is 50. (a) The best case with the rms distance between two structures 3.0 Å. (b) The worst case with the rms distance of 5.7 Å.

In conclusion, we have developed an RNA base discrete state model, which is totally different from the previous ones. By testing the similarity between the original structure and the mapped one in the discrete space, we show that our discrete model can reproduce the continuous RNA trajectory accurately. The most important feature of the model is its consistence with the fact that the RNA free energy is largely determined by base pairs and base stackings, allowing the direct calculation the free energies of the regrown structures, and no need resorting to rotamer-library based or all-atom based refinement of the configurations of bases. This is a promising feature and may facilitate the RNA three-dimensional structure predictions, which is ongoing in our lab.

References

[1] Gesteland R F et al 2006 *The RNA World* 3rd edn (New York: Cold Spring Harbor Press) Chap 11
[2] Mathews D H et al 1999 *J. Mol. Biol.* **288** 911
[3] Zuker M 2003 *Nucleic Acids Res.* **31** 3406
[4] Hofacker I L et al 1994 *Monatshefte Chemie* **125** 167
[5] Doshi K J et al 2004 *BMC Bioinformatics* **5** 105
[6] Gardner P P et al 2004 *BMC Bioinformatics* **5** 140
[7] Zhang J et al 2008 *J. Chem. Phys.* **128** 125107
[8] Zhang J et al 2009 *RNA* **15** 2248
[9] Bradley P et al 2005 *Science* **309** 1868
[10] Zhang C et al 2004 *Protein Science* **13** 400
[11] Zhang J, Qin M and Wang W 2006 *Proteins* **62** 672
[12] Zuo G H et al 2005 *Chin. Phys. Lett.* **22** 1809
[13] Xu W X et al 2005 *Chin. Phys. Lett.* **22** 258
[14] Chen S J 2008 *Annu. Rev. Biophys.* **37** 197
[15] Parisien M and Major F 2008 *Nature* **452** 51
[16] Yingling Y G and Shapiro B A 2006 *J. Mol. Graphics Modelling* **25** 261
[17] Das R and Baker D 2007 *Proc. Natl. Acad. Sci. USA* **104** 14664
[18] Murray L J W, Arendall W B III et al 2003 *Proc. Natl. Acad. Sci.* USA **100** 13904
[19] Klein D J et al 2004 *J. Mol. Biol.* **340** 141
[20] Zhang J F, Chen R and Liang J 2004 *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 26 2976 (San Francisco 1–4 September 2004)
[21] Kearsley S K 1989 *Acta Cryst.* A **45** 208

Chinese Physics Letters

Volume 27 Number 11 2010

(Continued on inside back cover)

FUNDAMENTAL AREAS OF PHENOMENOLOGY(INCLUDING APPLICATIONS)

CROSS-DISCIPLINARY PHYSICS AND RELATED AREAS OF SCIENCE AND TECHNOLOGY

GEOPHYSICS, ASTRONOMY, AND ASTROPHYSICS