Biophysical *Journal*
New and Notable

Graph, pseudoknot, and SARS-CoV-2 genomic RNA: A biophysical synthesis

Shi-Jie Chen[1,2,3,*]
[1]Department of Physics; [2]MU Institute for Data Science and Informatics; and [3]Department of Biochemistry, University of Missouri, Columbia, Missouri

The COVID-19 pandemic poses an urgent challenge to the scientific community. Current worldwide efforts to develop effective vaccines and therapeutic drugs targeting protein or the RNA genome have led to unprecedented demand for accurate modeling of the structure and intermolecular interactions for the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus. One of the outstanding problems is the structural modeling for the viral RNA genome. At the genomic level, the SAR-CoV-2 virus is a positive-sense RNA virus encoded in a 30,000-nucleotide RNA sequence. Decades of coronavirus research have suggested highly conserved structural domains that play crucial roles in viral replication and infection. These structural domains offer a plethora of targets for drug design. One such major drug target is the frameshift stimulation element (FSE).

During translation of the viral RNA, the downstream RNA structure mechanically blocks ribosome movement, resulting in a translational pause, which can accommodate a subsequent shift in the open reading frame and production of multiple proteins, such as gag and pol. Aberrant frameshift efficiencies lead to abnormal ratios of protein production, damaging viral assembly and replication. Therefore, altering or inhibiting the frameshift efficiency through drug binding to the RNA is a highly promising strategy for inhibiting SARS-CoV-2 activity.

At the center of the frameshift machinery is the FSE RNA. Computational and experimental studies suggest that the native (functional) structure of FSE is an H-type pseudoknot. With the pseudoknot structure, the frameshift machinery is a three-component system: 1) a 7-nucleotide single-stranded slippery region (nucleotides 13405–13411); 2) a downstream pseudoknot (13418–13488) composed of helix stems S1 (13418–13427/13438–13447) and S2 (13431–13437/13478–13485), which are cross-linked by loops L1 (13428–13430) and L2 (13448–13477), and a third stem S3 (13448–13456/13466-13475) inside loop L2; and 3) a 6-nucleotide spacer (13412–13417) between the slippery sequence and the downstream pseudoknot.

Inspired by the therapeutic importance for understanding the sequence-structure relationship for FSE, as reported in this issue of *Biophysical Journal*, Schlick et al. (1) recently performed a systematic computational study to search for the structurally critical nucleotides that may serve as drug targets. Specifically, Schlick et al. exhaustively scanned the mutations to identify nucleotides whose mutations would destroy the native structure. The study led to several surprising findings, including the discovery that mutating only two to three critical nucleotides would be sufficient to cause dramatic structural changes and hence possible disruption of frameshifting.

Conformational sampling is one of the major challenges for computational study of sequence-structure relationships. Incomplete or poor-quality sampling is often the culprit for inaccurate predictions of RNA folding. Schlick et al. developed and employed a highly innovative graph-theoretic approach, RNA-As-Graphs (RAG), to tackle the sampling problem (2). The key strategy of RAG is to transform an RNA two-dimensional (2D) structure into a tree (or dual) graph in which loops (helices) and helices (loops) are represented as vertices and edges, respectively. For example, in terms of a dual graph, a two-stem H-type pseudoknot can be represented as two vertices (two helix stems) connected by two edges (two cross-linked loops).

Because mapping from a structure to a graph effectively silences information about helix and loop lengths and retains only the "topology" of the network of helix-loop connectivity, RAG leads to a drastic reduction in the (graphical) conformational space. Another notable feature that distinguishes the RAG model from other coarse-grained RNA folding models

Submitted January 24, 2020, and accepted for publication January 28, 2021.

*Correspondence: chenshi@missouri.edu

Editor: Susan Schroeder.
https://doi.org/10.1016/j.bpj.2021.01.030

is that the model provides a platform for direct application of various powerful and rigorous graph theory tools. For example, the graph partitioning algorithm gives a rigorous method to modularize a structure. Moreover, the exact enumeration of all the possible graphs for a given number of vertices makes it possible to exhaustively sample all possible RNA motifs and structures, including new folds not yet discovered in experimentally determined structures.

In the study of sequence-structure relationships, exploring the sequence space with high computational efficiency is another major challenge. To tackle this challenge, Schlick et al. developed the RAG-IF approach to select sequences that fold into a given target structure (the inverse folding problem) (3). Realizing that RNA folding is intrinsically a three-dimensional (3D) problem, Schlick et al. integrated 2D and 3D folding algorithms in RAG-IF. Starting from a given target structure, RAG-IF parses the graph of the whole structure into subgraphs. For each subgraph, the algorithm employs a genetic algorithm to perform systematic mutations. For a given mutant sequence, RAG-IF predicts the 3D folds for each subgraph. Assembly of the subgraph 3D structures generates an ensemble of whole 3D structures, which are then scored and ranked by a knowledge-based statistical potential. Based on the 3D structure scoring, RAG-IF selects the top 200 unique sequences and submits the sequences for 2D structure prediction. Finally, the predictor chooses sequences predicted to fold into the target 2D structure.

The RAG and RAG-IF methods provide a highly efficient and effective search tool for critical nucleotides and mutations. Specifically, Schlick et al. generate a set of possible alternative (non-native) folds by graphically transforming the native fold. Treating these alternative graphs (structures) as target structures, RAG/RAG-IF finds the corresponding sequences (mutations) for each target. One of the remarkable features of RAG/RAG-IF is the ability to generate a great variety of different RNA topologies, including those containing non-native helix stems. By creating native-like and non-native helix stems, the RAG/RAG-IF approach offers a mathematically elegant and computationally efficient tool for investigating the sequence-structure relationship for large structural arrangements. Large conformational changes are essential for many RNA functions—for example, RNA catalysis in the different steps of the spliceosome cycle.

The above approach leads to a number of minimal mutants that destroy the pseudoknot structure and/or helix stem S2. For example, mutants [13441A-G, 13443A-C] in S1 and [13483G-C, 13485U-C] in S2 cause switches from the native pseudoknot to a three-way junction and a three-stem structure with an internal loop, respectively. Although computation was mostly focused on stem 2, the same approach can be used to find critical mutations that destroy stems 1 and 3. The successful applications of the graph-theoretic approaches demonstrate the advantage of coarse-grained modeling for RNA folding (4) and the power of rigorous mathematical tools, such as graph theory, in biophysical modeling. As shown below, these graph-theoretic approaches may play a unique role in tackling further challenges in the biophysics of SARS-CoV-2 FSE.

There are two major challenges in the biophysical modeling of the FSE RNA structures. First, the FSE RNA may form multiple alternative low-energy structures at both the 2D and the 3D structure levels (5,6). For example, at the 2D structure level, nucleotide 13448G can switch between base pairing with 13417U and with 13475U, causing two different 2D structures. At the 3D structure level, computer simulation indicates the formation of different 3D folds with the 5′-end spacer sequence threading (or not threading) through the junction region between stems S1 and S3. In addition, the formation of the different 3D structures can be further complicated by metal ion effects (7). Second, the folding of FSE may be influenced by potential long-range interactions between FSE and other nonlocal regions in the SARS-CoV-2 genome. For example, including nucleotides upstream from the slippery site can lead to the formation of alternative helices (8). Including 50 nucleotides upstream from the slippery site, VfoldPK (9), a free energy-based model for the folding of pseudoknotted structures, predicts the formation of two low-energy structures with different sets of basepairs for helix S1 and a new helix formed with long-range base pairing between the slippery-spacer region and the 50-nucleotide upstream sequence. Consideration of such a full structure may be necessary to design drug binding to the FSE target. However, in the process of viral translation, the sliding ribosome may disrupt upstream structures. Therefore, predicting the influence of the disruption of the long-range interactions in the upstream structure on the possible structural rearrangements in the downstream FSE is important. The RAG model, with its unique graph-based structure sampling algorithm, offers a highly promising tool for predicting alternative folds, even for larger systems.

The fact that RAG/RAG-IF models rely on RNA 2D structure prediction highlights the need for an accurate 2D structure prediction program. Sequence alignment can often provide reliable information about conserved nucleotides and basepairs. However, sequence alignment often cannot give all the basepairs for a structure. Therefore, we need physical models to predict the (remaining) basepairs. Although the RAG algorithm offers an excellent solution to the conformational sampling problem, calculating the free energy for structures, especially for those containing convoluted pseudoknots and loop-helix base triple interactions (10), requires an accurate physical model.

Finally, frameshifting occurs stochastically as a result of the competition between the unfolding of the

downstream RNA structure and the disruption of cognate codon-anticodon interactions in the slippery region. The competition is further complicated by the buildup of the elastic force in the intervening spacer. A quantitative prediction of the frameshift efficiency requires the consideration of the whole three-component system. Therefore, handling the coupled fluctuations of the three components is a key issue in modeling the system. The problem complexity is further compounded if systematic mutations and inverse folding problems are considered. By efficiently and exhaustively enumerating the graphs (structures), the RAG/RAG-IF model may offer a highly attractive tool for generating a statistical ensemble of fluctuating states for the three-component FSE system, both thermodynamically and kinetically.

ACKNOWLEDGMENTS

The author thanks Yangwei Jiang, Jun Li, Sicheng Zhang, and Yuanzhe Zhou for many useful discussions and Travis Hurst for critical reading of the manuscript.

This work was supported by the National Institutes of Health under Grants R01-GM117059 and R35-GM134919 to S.-J.C.

REFERENCES

1. Schlick, T., Q. Zhu, …, S. Yan. 2021. Structure-altering mutations of the SARS-CoV-2 frame shifting RNA element. *Biophys. J.* 120:1040–1053.
2. Fera, D., N. Kim, …, T. Schlick. 2004. RAG: RNA-As-Graphs web resource. *BMC Bioinformatics.* 5:88.
3. Jain, S., Y. Tao, and T. Schlick. 2020. Inverse folding with RNA-As-Graphs produces a large pool of candidate sequences with target topologies. *J. Struct. Biol.* 209:107438.
4. Šulc, P. 2020. The multiscale future of RNA modeling. *Biophys. J.* 119:1270–1272.
5. Zhang, K., I. N. Zheludev, …, R. Das. 2020. Cryo-electron microscopy and exploratory antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *bioRxiv* https://doi.org/10.1101/2020.07.18.209270.
6. Schroeder, S. J. 2020. Perspectives on viral RNA genomes and the RNA folding problem. *Viruses.* 12:1126.
7. Omar, S. I., M. Zhao, …, M. T. Woodside. 2020. Modeling the structure of the frameshift stimulatory pseudoknot in SARS-CoV-2 reveals multiple possible conformers. *bioRxiv* https://doi.org/10.1101/2020.06.08.141150.
8. Tammy, C. T., M. F. Lan, …, S. Rouskin. 2020. Structure of the full SARS-CoV-2 RNA genome in infected cells. *bioRxiv* https://doi.org/10.1101/2020.06.29.178343.
9. Xu, X., P. Zhao, and S. J. Chen. 2014. Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One.* 9:e107504.
10. Cao, S., D. P. Giedroc, and S. J. Chen. 2010. Predicting loop-helix tertiary structural contacts in RNA pseudoknots. *RNA.* 16:538–552.