User Manual for Vfold-Pipeline

(version 2.0)

Shi-Jie Chen Research Group

The Chen research group investigates the physical mechanism of RNA folding and develops predictive models for RNA structures and functions.

Shi-Jie Chen
Principle Investigator
Curators' Distinguished Professor
Email: chenshi@missouri.edu
Group website: https://vfold.missouri.edu

Mizzou
University of Missouri

Department of Physics and Astronomy
Department of Biochemistry
MU Institute for Data Science and Informatics

Columbia, MO 65211, USA

Contents

1 Introduction

The Vfold-Pipeline package is used for RNA 3D structure prediction from sequence. The workflow of the Vfold-Pipeline web server is illustrated in Fig. 1. It first predicts 2D structures using the Vfold2D [1, 2, 3, 4] module and then predicts 3D structures based on the predicted 2D structures using the Vfold3D[5]/VfoldLA[6] module. The Vfold2D module predicts 2D structures based on free energy. Specifically, for a given RNA sequence, the Vfold2D module gives an optimal 2D structure and an ensemble of suboptimal 2D structures ranked by free energies. Before 2D structure prediction, the pipeline automatically performs a sequence search in the Rfam database [7]. If any constraints are found, they will be used for 2D structure prediction. Moreover, the constraints from SHAPE experiments can be used in our Vfold2D model. Given 2D structures, the Vfold3D and VfoldLA modules predict 3D structures based on the assembly of the motif and loop templates extracted from the known 3D structures, respectively. This package first launches the Vfold3D method, and if it fails in giving prediction, then the VfoldLA method is used. This Vfold-Pipeline package combines the two 2D and 3D structure prediction modules into a pipeline and optimizes the calculation, making it convenient for users to predict RNA 3D structures from sequences.

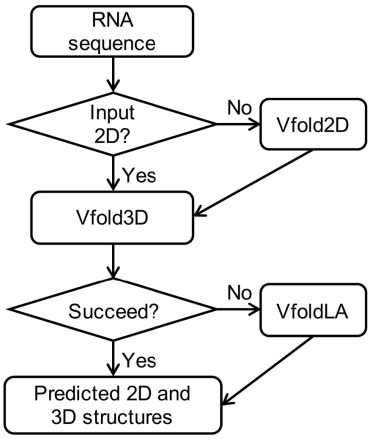

Figure 1: The workflow of the RNA 2D and 3D structure prediction procedures implemented in the Vfold-Pipeline package.

2 User Guide

2.1 Overview of the Vfold-Pipeline Package

The Vfold-Pipeline package consists of two modules, namely the Vfold2D and the Vfold3D/VfoldLA modules.

(1) There are two separate tools in the Vfold2D module:

vfold2D_cl.out is the classic version of the software for the prediction of RNA secondary structure without pseudoknots. The thermodynamic parameters for base stacks, including mismatched base stacks, are from the Turner parameters (04 version) or from the MFOLD, and the loop/junction entropies are from the Vfold2D model. The SHAPE constraints can also be applied to the prediction. It is written in C++ language stored in the directory "Vfold2D/bin/" after installation.

vfold2D_pk.out predicts RNA secondary structure including H-type pseudoknots. The thermodynamic parameters for base stacks are from the Turner parameters (04 version) or from the MFOLD, and the loop/junction entropies are from the Vfold2D model. The SHAPE constraints can also be applied to the prediction. It is written in C++ language stored in the directory "Vfold2D/bin/" after installation.

(2) There are two tools in the Vfold3D/VfoldLA module:

vfold3DLA.out is used to predict 3D structures by the Vfold3D or VfoldLA models. Given the 2D structure, it first uses the Vfold3D model to assemble the motif templates to form a complete structure. If the Vfold3D model fails in predicting the 3D structures due to the lack of the appropriate motif templates, the VfoldLA model will be used to predict the 3D structures by assembling the loop templates. The VfoldLA model also performs short-time coarse-grained molecular dynamics simulations to refine the assembled structures. The main difference between the Vfold3D and the VfoldLA models is that the VfoldLA model treats a motif (such as junction loops) as several separate loops while the Vfold3D model treats it as a whole. Therefore, the Vfold3D model can give more accurate 3D structures if it can manage to give predictions. It is written in C++ language stored in the directory "Vfold3DLA/bin/" after installation.

(3) Four Python scripts are used to connect the Vfold2D and Vfold3D/VfoldLA modules into a pipeline:

run_VfoldPipeline.py is the main script to run the pipeline. Users can specify parameters in an input file, such as the consideration of pseudoknotted 2D structures, the maximum number of 2D structures used for 3D structure prediction, the maximum number of predicted 3D structures and so on. It is written in Python language stored in the directory "bin/".

rfam_sequence_search.py as an Rfam sequence search module, is used to extract 2D constraints from the Rfam database [7], a collection of RNA families. The found constraints will be used in the 2D structure prediction. It is written in Python language stored in the directory "bin/".

prediction_2D.py as a 2D structure prediction module, is used to predict 2D structures for PK and non-PK structures by running vfold2D_cl.out and vfold2D_pk.out, respectively. It is written in Python language stored in the directory "bin/".

prediction_3D.py as a 3D structure prediction module, is used to predict 3D structures by running vfold3DLA.out. It is written in Python language stored in the directory "bin/".

2.2 Installation

(1) Machine requirements

(a) Linux operating system

(b) $\geq$ 10 threads

(c) $\geq$ 5 G RAM

(d) GCC (version $\geq$ 4.8) supporting c++ 11 standard

(2) Prerequisites: the following software should be installed on the machine.

MPICH (version $\geq$ 3.2) via http://www.mpich.org/downloads/. If you are using the linux distribution Ubuntu, use the command `sudo apt-get install mpich`. MPICH platform is used for parallel molecular dynamics simulations in the VfoldLA model.

QRNAS It is used to minimize the energy of the predicted RNA 3D structures [8]. Download it from https://github.com/sunandanmukherjee/QRNAS. Unpack the downloaded file and rename the directory as "QRNAS" if not.

LAMMPS (version 3Mar2020) It is used for MD simulations. You can download it from https://www.lammps.org/download.html and choose the version corresponding to **3March2020**. If this version is not available on the web page, please download the correct version from: https://download.lammps.org/tars/index.html. After downloading, unpack the tarball file using the command `tar -xzvf lammps-3Mar2020.tar.gz`. After unpacking, you get a directory named "lammps-3Mar20", rename it if not.

Python (version $\geq$ 3.6) with the following packages installed:

(a) **requests**: can be installed by "pip install requests"

(b) **json**: can be installed by "pip install json"

(c) **time**: can be installed by "pip install time"

The above three Python packages are used to extract 2D constraints from the Rfam database online. So the user's machine should connect to the Internet in order to perform a sequence search in the Rfam database, otherwise our pipeline will not perform the sequence search but can still predict the 2D structures by the Vfold2D model.

(3) Installation of the Vfold-Pipeline package

(a) `tar -jxvf VfoldPipeline_standalone.tar.bz2`

(b) move the `QRNAS` directory to `VfoldPipeline_standalone/Vfold3DLA/external_software/`

(c) move the `lammps-3Mar20` directory to `VfoldPipeline_standalone/Vfold3DLA/external_software/`

(d) `cd VfoldPipeline_standalone/`

(e) `bash ./Install.sh`

(f) After installation of the above programs, please **set the environment variable** `VfoldPipeline` to the directory where the VfoldPipeline package is installed by adding:
export VfoldPipeline=/Path/To/VfoldPipeline/Package
to ~/.bashrc. The reminder of this step will occur in the final screen output after installation.

2.3 Usage of the Vfold-Pipeline package

The Vfold-Pipeline package provides a Python interface to implement the 2D and 3D structure prediction pipeline.

The main script is `run_VfoldPipeline.py` which is run by Python of version $\geq$ 3.6 and is stored in "Vfold-Pipeline_standalone/bin/".

```
usage: python run_VfoldPipeline.py input_file.txt
```

"input_file.txt" is a text file storing all the information for the new job and it has a strict format shown as follows. The comments after the symol "#" can be deleted. We have provided three input_file.txt examples in the directories "example-1/2/3". Users can modify them according to your own jobs.

```
rna_name: 1e95           # The name of the RNA
seq: AUGC                # Single-stranded RNA sequence in "AUCGaucg"
2D:                      # The 2D structure users provide in dot-bracket
                           format; leave blank if users want to perform 2D
                           structure prediction by Vfold2D
if_PK: Yes               # "No" means not considering pseudoknots in the 2D
                            structure prediction by Vfold2D; "Yes" means
                            considering
max_num_2D: 2            # The maximum number of 2D structures to be used
                           for 3D structure prediction
temperature: 37.0        # The temperature used in 2D structure prediction
                           by Vfold2D
excluded_pdblist: 1e95   # The PDB list excluded from the template database
                           for 3D structure prediction; leave blank if none
                           is excluded
shape_file: shape.txt    # The SHAPE file used for 2D structure prediction
                           by Vfold2D; leave blank if no SHAPE file; In the
                           SHAPE file, the first and second columns
                           correspond to the nucleotide index and the SHAPE
                           reactivity value, respectively. The nucleotides
                           missing SHAPE reactivity values can be omitted.
rmsd_cutoff: 5.0         # The RMSD cutoff used in VfoldLA for 3D structure
                           prediction
cluster_num: 10          # The maximum number of predicted 3D structures by
                           VfoldLA
output_directory: ./     # The output directory
```

2.4 Examples

(A) Example 1: structure prediction for the tRNA 1ffy.

According to the parameters in the “input_file.txt”, the Vfold-Pipeline package first predicts its non-PK 2D structures and then predicts the 3D structures using only 1 predicted non-PK 2D structure. The PDB files 1ffy, 1qu2, 1qu3 and 5o2r are excluded from the template databases for 3D structure prediction by Vfold3D and VfoldLA.

This example can be run via the command “bash run.sh” or “python ../bin/run_VfoldPipeline.py input_file.txt”.

The final predicted 2D and 3D structures are stored in the ”1ffy/final_results/” folder. We have provided the reference results in the folder “1ffy-Reference-Results/”.

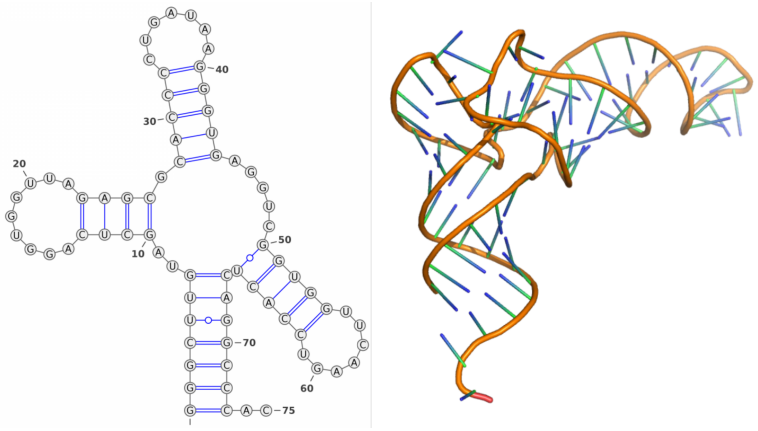

Figure 2: The predicted 2D and 3D structures in Example 1 for the tRNA 1ffy.

(B) Example 2: structure prediction for the pseudoknotted RNA 1e95 with SHAPE constraints.

According to the parameters in the “input_file.txt”, the Vfold-Pipeline package first predicts its non-PK and PK 2D structures with the SHAPE constraints in “example_shape.txt” and then predicts the 3D structures using up to 2 predicted non-PK and PK 2D structures.

This example can be run via the command “bash run.sh” or “python ../bin/run_VfoldPipeline.py input_file.txt”.

The final predicted 3D structures are stored in the “1e95/final_results/” folder. We have provided the reference results in the folder “1e95-Reference-Results/”.

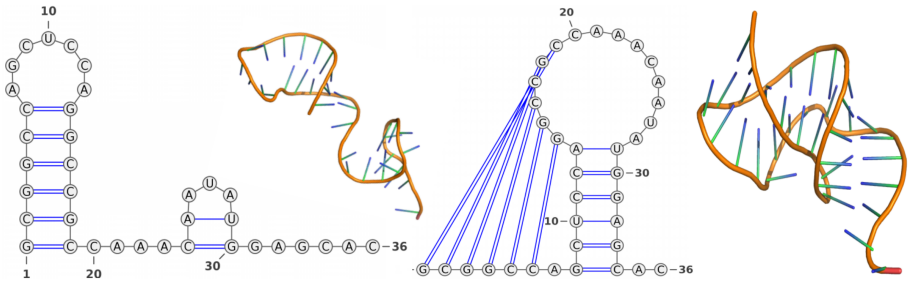

Figure 3: The predicted non-PK and PK 2D structures and their corresponding predicted 3D structures in Example 2 for the RNA 1e95.

(C) Example 3: structure prediction for the pseudoknotted RNA 1e95 given the 2D structure.

According to the parameters in the “input_file.txt”, the Vfold-Pipeline package uses the given 2D structure to predict the 3D structures.

This example can be run via the command “bash run.sh” or “python ../bin/run_VfoldPipeline.py input_file.txt”.

The final predicted 3D structures are stored in the “1e95/final_results/” folder. We have provided the reference results in the folder “1e95-Reference-Results/”.

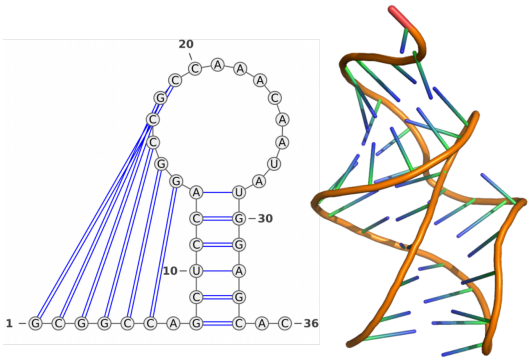

Figure 4: The given 2D structure and predicted 3D structures in Example 3 for the RNA 1e95.

3 References

[1] Zhao CH, Xu XJ and Chen S-J. (2017) *Predicting RNA structure with Vfold*. Methods in Molecular Biology, 1654: 3-15.

[2] Xu XJ and Chen S-J. (2016) *A method to predict the structure and stability of RNA/RNA complexes*. Methods Mol Biol. 1490:63-72.

[3] Xu XJ and Chen S-J. (2015) *Modeling the structure of RNA scaffold*. Methods Mol Biol. 1316: 1-11.

[4] Xu XJ, Zhao PN and Chen, S-J. (2014) *Vfold: a web server for RNA structure and folding thermodynamics prediction*. PLoS ONE

[5] Cao S and Chen S-J. (2011) *Physics-based de novo prediction of RNA 3D structures*. J. Phys. Chem. B., 115: 4216-4226.

[6] Xu XJ and Chen S-J. (2017) *Hierarchical assembly of RNA three-dimensional structures based on loop templates*. J. Phys. Chem. B., 122(21): 5327-5335.

[7] Kalvari, Ioanna, et al. (2021) *Rfam 14: expanded coverage of metagenomic, viral and microRNA families.*. Nucleic Acids Research, 49: D192-D200.

[8] Stasiewicz, Juliusz, et al. (2019) *QRNAS: software tool for refinement of nucleic acid structures.*. BMC structural biology, 19: 1-11.