Objective: Create a model that predicts RNA structure from sequence.

Context/Background:

The higher order structures of biomacromolecules are very challenging to predict but critical for functional understanding and manipulation. RNA in particular is very challenging to work with *in vitro*; the factors driving the structure and function of RNA molecules remain poorly understood.

While it is easy to think of it as a less-stable version of DNA, its structure is impacted much less by Watson-Crick base pairing and instead is heavily dictated by base packing and non-Watson-Crick base pairing (Vincens and Kieft 2022 https://doi.org/10.1073/pnas.2112677119). Furthermore, it is thought to be highly dynamic, and predicting a single "native state" is thus significantly more inaccurate and limiting than in protein structure prediction. The leading theory is that we are living in an RNA-first world; these molecules are critically enmeshed in the most conserved processes in nature. We are continually fighting off RNA viruses, and increasing numbers of RNA therapies are being continuously developed. Effective understanding of all of these problems requires an understanding of RNA structure.

"Reactivity" in our dataset refers to chemical mapping reactivities, a widely used readout of RNA secondary structure that is both high throughput and able to resolve down to single nucleotides. For each sequence, we can have data on reactivity measured with either dimethyl sulfate (DMS) or 2-aminopyridine-3-carboxylic acid imidazolide (2A3) chemical modifiers. DMS reacts with the base-pairing faces of unpaired adenosine and cytidine nucleotides; while 2A3 acylates the 2'-hydroxyl on the ribose, thereby serving as a probe of backbone flexibility.

Problem Scope

Develop a model that will take an RNA sequence and predict the reactivity measure for each nucleotide, serving as a proxy for macromolecular structure.

Data Source(s)

train_data_QUICK_START.csv from (closed) Stanford Ribonanza RNA Folding Kaggle competition

https://www.kaggle.com/competitions/stanford-ribonanza-rna-folding/data?select=train_data_QUICK_START.csv

Competition Citation: Rhiju Das, Shujun He, Rui Huang, Jill Townley, Rachael Kretsch, Thomas Karagianes, John Nicol, Grace Nye, Christian Choe, Jonathan Romano, Maggie Demkin, Walter

Reade, and Eterna players . (2023). Stanford Ribonanza RNA Folding. Kaggle. https://kaggle.com/competitions/stanford-ribonanza-rna-folding

Other Informational Sources

The Kaggle competition page has extensive literature links, discussion boards, and proposed posted solutions

Existing RNA structure prediction algorithms have associated publications and documentations that may also prove helpful

A general literature search (via PubMed or equivalent) can also provide insights

Potential Limitations or Constraints

Existing RNA structural data on which existing algorithms are based was not collected in an unbiased manner for the purposes of generating a dataset for algorithm training purposes, therefore there may be unknown factors impacting either the sequences selected or the reactivity readouts.

Additionally, experimental artifacts and limitations impact the reactivity readouts (e.g. first and last positions not being able to provide a readout).

Finally, while the dataset is relatively large, if you consider the fact that we are looking at energetics between individual atoms and groups of atoms it becomes clear that the problem in question is incredibly complicated.