January 7, 2019

Dear Editors of *Nature Communications*,

Enclosed please find our manuscript entitled “LinearPartition: Linear-Time Approximation of RNA Folding Partition Function and Base Pairing Probabilities”, to be considered for publication as an article in *Nature Communications.* We present a *linear-time* algorithm to approximate the partition function for RNA folding, while other algorithms’ runtimes scale (at least) cubically with the sequence length.

Macromolecular structure prediction is an important problem in biology and it has wide applications such as drug design. For many such sequences including proteins and RNAs, multiple structures exist at equilibrium with varying probabilities. This can be captured by the notion of “partition function” from thermodynamics which is the normalization factor that defines a distribution over all possible structures in the ensemble. In RNA folding in particular, the partition function is a central concept that has many other applications, such as deriving the base pairing probabilities which lead to more accurate structure predictions, stochastic sampling from the distribution, accessibility estimation, and determining confidence in structure prediction.

McCaskill (1990) pioneered the standard algorithm in use today for computing the partition function and base pairing probabilities for a given RNA, whose runtime scales *cubically* with the sequence length. In the 30 years since then, it has been widely adopted (cited 1,427 times on Google Scholar) and is implemented in numerous software packages. However, its cubic-factor slowness prevents it from being applied to longer sequences such as long non-coding RNAs and mRNAs. Our work, though approximate, is the first major speedup for this long-standing problem in 30 years, and our linear-time algorithm is orders of magnitude faster than the standard cubic-time one (for example, 2,771x faster on a sequence of length 32,753 from the RNAcentral dataset: 2.5 days vs. 1.3 minutes).

Remarkably, even though our algorithm approximates the partition function, this approximation nevertheless yields more accurate distributions of the ensemble, especially on longer RNA families. The quality of distribution is measured by correlation with the ground-truth structures as well as downstream prediction accuracies using derived base-pairing probabilities. Therefore, we think this work is a major advancement in the field of RNA computational biology. Looking forward, our work will have impact by providing fast calculations for full length sequences, such as mRNAs.

We assure that this work is original and has not been published nor is it under consideration for publication elsewhere. The manuscript has been uploaded to the arXiv preprint server: https://arxiv.org/abs/1912.13190. We have no conflicts of interest to disclose.

A small part of this work uses another work from our group which is under review by another journal (and available at https://arxiv.org/abs/1912.12796). For your convenience we uploaded that paper as a “related manuscript”. Please address all correspondence concerning this manuscript to me at liang.huang.sh@gmail.com.

Thank you for consideration of our manuscript.

Liang Huang, *Ph.D.*
Distinguished Scientist, Baidu Research USA
Assistant Professor, Oregon State University