Estimating the fraction of non-coding RNAs in mammalian transcriptomes

Yurong Xin, Giulio Quarta, Hin Hark Gan, and Tamar Schlick
Department of Chemistry and Courant Institute of Mathematical Sciences
New York University, New York, NY, 10012

Introduction

- Recent studies of mammalian transcriptomes have identified numerous RNA transcripts that do not code for proteins; their identity, however, is largely unknown.
- Do these transcripts correspond to genuine non-coding RNAs (ncRNAs) with biological functions, or to other RNAs that may be non-functional transcripts?
- Determining the fraction of genuine non-coding RNAs in putative ncRNAs will advance our understanding of the composition of mammalian transcriptomes and the general importance of ncRNAs for cellular function.

Methods

- Six putative ncRNA datasets are analyzed: (1) FANTOM3 ncRNAs, (2) FANTOM3 stringent ncRNAs, RNAz predictions (3) set1.P0.5 (P>0.5), (4) set1.P0.9 (P>0.9), (5) set2.P0.5 (P>0.5), and (6) EvoFold predictions.
- The relative z-score is derived from the z-score of the monkey test; it assesses sequence randomness.

1. The relative z-score classifies different nucleotide sequence groups into three clusters: genome/intergene/intron, mRNA/ncRNA, and repeat.

2. None of the six putative ncRNA datasets have a relative z-score close to the known ncRNA class.

3. The fraction model is developed using the statistical feature of the relative z-score. Our analysis suggests a wide range of ncRNA fractions (5–50%) in the six putative ncRNA datasets.

4. Thermodynamic analyses for short FANTOM3 RNAs (23% passing rate) agree with the prediction (18%) by our model.

Conclusions

- A first-level approximation model is proposed to estimate the fraction of genuine ncRNAs in unknown RNA transcripts.
- The analyses suggest that fewer genuine ncRNAs may exist in the experimental or computational datasets.

Acknowledgements

The work was supported by Human Frontier Science Program (HFSP) and by a joint NSF/NIGMS initiative in Mathematical Biology (DMS-0201160). We thank Stefan Washietl and Jakob Skou Pedersen for kindly providing us sequences predicted by the RNAz and EvoFold programs, respectively.