RNASurface: fast and accurate identification of motifs with high structural potential

Ruslan Soldatov[1], Svetlana Vinogradova[1,2], Andrey Mironov[1,2]

[1]*Institute for Information Transmission Problem, Moscow, Russia*

[2]*Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia*

solrust@mail.ru

RNA has an abundance of structural functions in cells. A lot of new classes of non-coding RNAs have been discovered during last decades. For example, the microRNA regulates gene expression through post-transcriptional repression . The riboswitches are cis-acting regulatory elements, which act as feedback regulator of metabolite abundance.

Biological functions of majority functional RNAs are crucially depend on a secondary structure, which is the scaffold of a tertiary structure. Prediction of RNA secondary structure can be done by minimum free energy (MFE) approach which based on dynamic programming. Measure of the RNA secondary structure significance partially reflects potential to perform cellular function: there were noticed that non-coding RNAs have less free energy than random sequences (Clote et al, 2005). Formally, structural potential of a sequence s determines as (Washietl et al, 2005)

$$Z = \frac{E(s)-\mu(s)}{\sigma(s)},$$

where E(s) is the minimum free energy, $\mu(s)$ and $\sigma(s)$ are the mean and the standard deviation of MFE of the set of shuffled sequences with preserved average dinucleotide content. Maintenance of dinucleotide content is important due to stacking interactions of base pairs.

For the given genome S of the size N, the surface of structural potential determines as matrix of Z-scores:

$$\{Z_{ij}, 1 \le i \le j \le N\}, \text{ where } Z_{ij} = Z(S_{ij})$$

Further, motif S_{ij} is locally-optimal if it is a local negative peak of the surface.

Here we present program RNASurface that allows fast reconstruction of the surface of structural potential with simultaneous identification of locally-optimal motifs.

An application to Bacillus subtilis demonstrates that this approach better demarcates non-coding RNA from random decoy than RNALfold and other window-based approaches, while preserving time and space consumption.

Non-coding RNAs have much more subtle signal than protein-coding genes, thus energy-based approaches are inappropriate as ncRNA-finder tools. However, highly accurate determination of structured intervals is useful for several purposes:

- de novo regulatory and non-coding RNA search preprocessing
- accurate ncRNA bound definition
- correlation of genome-wide dataset of structural potential with another genomic features (such as gene bounds, ribosome profiling, transcriptome data)

1. Clote P, Ferre' F, Kranakis E, Krizanc D (2005) Structural RNA has lower free energy than random RNA of the same dinucleotide frequency, *RNA*, **11:**578-591.

2. Washietl S, Hofacker I, Stadler P (2005) Fast and reliable prediction of noncoding RNAs, *Proc Natl Acad Sci USA*, **102(7):**2454-2459.