Detection of frameshifts and improving genome annotation

Ivan Antonov
School of Computational Science and Engineering, Georgia Institute of Technology, 810 Atlantic Drive, Atlanta, GA 30332, USA, ivan.antonov@gatech.edu

Pavel Baranov
Department of Biochemistry, University College Cork, Ireland, brave.oval.pan@gmail.com

Mark Borodovsky
Department of Biomedical Engineering and School of Computational Science and Engineering, Georgia Institute of Technology, 313 Ferst Drive, Atlanta, GA 30332, USA;

Department of Molecular and Biological Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia, borodovsky@gatech.edu

We used an *ab initio* frameshift prediction program GeneTack [1] to screen 1,106 complete prokaryotic genomes; we identified 206,991 genes with frameshifts (fs-genes) [2]. Our final goal was to determine if a frameshift transition was due to (i) a sequencing error, (ii) an indel mutation or (iii) a recoding event. We grouped 102,731 fs-genes into 19,430 clusters based on sequence similarity between protein products, conservation of position of predicted frameshift, and its direction. We classified fs-genes in 4,010 clusters as conserved or hypothetical pseudogenes; on the other hand, in 146 clusters with total of 4,730 fs-genes we detected conserved motifs located near frameshifts characteristic for programmed frameshifts (involved in recoding).

Not only a motif itself but also its phasing with respect to the reading frame is crucial for proper functioning of a programmed ribosomal frameshift. Taking the phasing into account allows identifying the motifs more accurately than it could be done by standard motif searching algorithm, e.g. MEME.

To test the predictions, experiments were performed with cassettes of predicted frameshift-producing sequences of fs-genes residing in 20 clusters.

Programmed ribosomal frameshifting with higher than 10% efficiency was observed for four

clusters (Magnesium chelatase, Spore germination protein, DUF111 and DUF772).

Another interesting finding was the discovery of a novel type of organization of the dnaX gene, where recoding is required for synthesis of the longer protein variant.

We also applied GeneTack to 1,165,799 mRNAs from 100 eukaryotic species and identified 45,295 frameshifts in sequences that produced 4,087 clusters with 12,103 fs-genes [3]. Some clusters contained genes with known programmed frameshift, while several clusters were predicted to contain novel dual coding genes.

We have created a database containing all the fs-genes predicted in prokaryotic genomes and eukaryotic mRNA sequences as well as the web interface. Clusters of fs-genes are characterized with respect to the likely biological origin of frameshifts, such as pseudogenization, phase variation, programmed frameshifts, etc.

All the tools and the GeneTack database of fs-genes clusters are available at http://topaz.gatech.edu/GeneTack/

1. I. Antonov, M. Borodovsky (2010) GeneTack: frameshift identification in protein-coding sequences by the Viterbi algorithm, *Journal of Bioinformatics and Computational Biology*, **8(3):**535-51.

2. I. Antonov, A. Coakley, J.F. Atkins, P.V. Baranov, M. Borodovsky (2013) Identification of the nature of reading frame transitions observed in prokaryotic genomes, *Nucleic Acids Research*, [Epub ahead of print].

3. I. Antonov, P. Baranov, M. Borodovsky (2013) GeneTack database: genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences, *Nucleic Acids Research*, **41:**D152-6.