Beoung Hun Lee

CS 199

March 23, 2019

Edwin Solares

In this paper, the researcher proposes new program, GeneTack, that detects frameshift in the protein coding sequences. GeneTack uses Ab initio method instead of comparative method, since comparative method is limited in a sense that it can only search frameshifts in orphan proteins that have known homologs. GeneTack also uses hidden Markov Model to predict protein coding sequences. GeneTack can predict four different states: start and stop, protein coding sequence, non-coding sequence and overlap. Since GeneTack is designed to analyze DNA fragments only, GeneTack-GM is used to analyze the whole genome. GeneTack-GM first needs to use GeneMarkS to break down the whole genome into the fragments. GenemarkS produces adjacent genes since it cannot recognize frameshifts. These adjacent genes are analyzed to find frameshifts and then are filtered for false positives. For high G-C content genome, different probability is used for start state and predicted strand of gene is reassigned.

GeneTack-GM was used on 17 prokaryote genomes and was compared with different programs. For all genomes, GeneTack showed better accuracy than FrameD and FSFind. When FSFind was run with extrinsic method, it performed comparatively with GeneTack-GM, but GeneTack-GM was purely ran in ab initio mode. Another test on different data set showed similar result. GeneTack also successfully recognized programmed frameshifts when tested with annotated programmed frameshifts, as it only missed 5 out of 23 frameshifts, which 5 frameshifts were frameshifts in multiple of three nucleotides. GeneTack was more effective when the frameshift was further away from the start or stop codon. Filter successfully remove most of false positives and did not remove most of true positives, but its accuracy depended on RBS score from the GeneMarkS.

References

1. Antonov, I., Borodovsky, M. (2010) Genetack: Frameshift identification in protein-coding sequences by the Viterbi algorithm. *Journal of Bioinformatics and Computational Biology*. Vol 08. 535-551.