Genes that Make You Go HMM: Errors in Protein Sequences Propagated by Heuristic Conveniences Fixed with Probabilistic Models

Liu, Qijia (School: Westview High School)

The sequences of most proteins are inferred from the longest open reading frame (ORF) found in their parent mRNA. While the rule is simple and convenient, it does not model the underlying biology, such as composition of sequences and the Kozak consensus. As genomes and their gene parts are in immense quantity and lay the foundation of modern biology, even a small percentage of potential misannotations is critical. To improve the accuracy of gene annotation, hidden Markov models of mRNA structure are built to identify the most probable protein sequences. Results show that approximately 4.132% of the proteins do not follow the longest ORF rule and may be misannotated; in addition, for 39.7% of all the differences in gene annotation, the hidden Markov model predicted a longer ORF than the longest ORF model due to outdated data. Further examinations of a subset of highly conserved proteins corroborate this interpretation.