GeneMarkS-2: Powerful Gene Prediction Tool Advances Prokaryotic Genome Annotation (#8281)

Flexible modeling of DNA sequence patterns for identifying leaderless transcription and atypical genes with a new level of accuracy

Georgia Tech researchers have developed a novel bioinformatics algorithm for gene prediction implemented as a software tool. This new technology, called GeneMarkS-2, utilizes a multi-model approach for finding both native genes as well as horizontally transferred genes that are more difficult to detect. GeneMarkS-2 leverages a self-training algorithm that works in iterations by (1) segmenting the genome into protein-coding and non-coding regions and (2) recalibrating the model’s parameters based on the patterns it learns from the segmentation. This intelligent algorithm identifies several types of signals involved in gene expression control, including those for leaderless transcription. In fact, application of this tool led to the discovery that leaderless transcription is widespread in prokaryotes.

Benefits/Advantages

- **Accurate**: Produces fewer false-negative and false-positive gene predictions than other commonly used methods
- **Robust**: Employs two large sets of atypical gene models instead of one as in GeneMarkS
- **Intelligent**: Utilizes an unsupervised training algorithm, which increases convenience for users
- **Advanced**: Predicts the type of transcript—leadered or leaderless—and therefore provides information about regulatory motifs important for genetic research

Potential Commercial Applications

- Genomics research
- Pharmaceuticals research and development

Background/Context for This Invention

Ab initio gene prediction is the process of identifying genes in a DNA sequence using patterns developed in the evolution of protein-coding regions. Improving the accuracy of gene prediction is important, as sequencing of the Earth’s enormous number of microbial species continues to produce a volume of sequence data containing genes not detectable by mapping ortholog proteins. Thanks to its high accuracy and robustness, GeneMarkS-2 was used by the researchers of the National Center of Biotechnology Information as a part of the pipeline to annotate more than 250,000 genomes of prokaryotes (bacteria and archaea) available in the GenBank® genetic sequence database.

GenBank is a registered trademark of the National Library of Medicine.
Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Research, May 17, 2018

For more information about this technology, please visit:
https://industry.gatech.edu/technology/genemarks-2-powerful-gene-prediction-tool-advances-prokaryotic-genome-annotation