Introduction to Gene Prediction

Gianlucca de Urzêda Alves

2024-09-25

Gene Prediction

Importance

Gene prediction is a fundamental aspect of bioinformatics and computational biology, involving the identification and annotation of genes within large, contiguous sequences of DNA. As a fundamental step in genome analysis, gene prediction enables the discovery of gene products, such as RNA, proteins, and other functional elements like regulatory regions. The gene annotation phase helps us describe and understand the genetic profile of organisms from their bare sequences.

Historically, the gene discovery strategies were based on laboratorial experimentation. Costly and time consuming, such methods have been recently substituted for computational algorithms to compare and predict sequences of genomic structures and functions. With the advent of high–throughput sequencing technologies and advances in computational algorithms and processing power, gene annotation has been transformed into a computational challenge.

The main realms of gene prediction is as follows:

Structural Annotation (Position and Coordinates on Genome)

Functional Annotation (Function of Genes, mRNA and Proteins)

Non-Coding Sequences Annotation (Regulatory Sequences, Non Coding RNA - microRNA, siRNA, small RNAs, Medium and Long RNAs)

The gene annotation tools main functions are to:

1. **Identification of DNA Features:**

    - Gene prediction algorithms can identify and annotate various genomic elements, including:
        - **Genes:** Both protein-coding DNA (mRNA and Proteins) and non-coding DNA (rRNA, tRNA, miRNA, snRNAm snoRNA, lncRNA), which are the fundamental units of heredity and are responsible for various cellular functions.
        - **Introns and Exons:** Identifies the boundaries between introns (non-coding regions) and exons (coding regions), crucial for understanding gene structure and splicing patterns.
        - **Splicing Sites:** Identification of splicing sites is essential for predicting mature mRNA sequences from primary transcripts, impacting protein synthesis.
        - **Regulatory Sites:** Includes promoters, enhancers, and other regulatory elements that control gene expression, which are vital for understanding gene regulation and pathways.

    - **Motifs and ESTs (Expressed Sequence Tags):** Helps in identifying conserved motifs, which are critical for protein function, and ESTs that provide evidence of gene expression.

2. **Distinguishing Coding and Non-Coding Regions:**
    - Gene prediction algorithms help differentiate between coding sequences, which translate into proteins, and non-coding sequences, which may have regulatory roles or unknown functions. This distinction is essential for:
        - **Functional Annotation:** Provides insights into gene function by categorizing sequences as either coding or non-coding.
        - **Comparative Genomics:** Facilitates comparisons between species by aligning coding regions to study evolutionary conservation and divergence.

3. **Predicting Exon-Intron Structures:**
    - Predicting the exon-intron structures of protein-coding genes is important for understanding the gene's full transcript and its protein products. This is important for:
        - **Transcriptome Analysis:** Provides a view of all transcripts expressed in a cell, tissue, or organism.
        - **Alternative Splicing:** Helps identify alternative splicing events, which can generate multiple protein isoforms from a single gene, contributing to protein diversity.

4. **Functional Description of Individual Genes:**
    - Beyond structural annotation, gene prediction also involves the functional annotation of genes, which includes:
        - **Gene Ontology (GO):** Assigns functional terms related to biological processes, molecular functions, and cellular components.
        - **Pathway Analysis:** Links genes to biological pathways, enhancing understanding of metabolic and regulatory networks.

5. **Applications Across Genomics and Related Fields:**
    - Gene prediction has widespread applications across various domains of biology:
        - **Structural Genomics:** Involves mapping and sequencing of genomes, providing a blueprint of gene locations and structures.
        - **Functional Genomics:** Explores gene functions and interactions, aiding in the identification of gene roles in complex traits and diseases.
        - **Metabolomics:** Links gene prediction data to metabolic pathways, supporting the study of metabolite profiles and their biological implications.
        - **Transcriptomics:** Focuses on the complete set of RNA transcripts, facilitated by accurate gene predictions that outline the transcriptome landscape.
        - **Proteomics:** Involves studying the complete set of proteins encoded by the genome, relying on accurate gene predictions to infer protein-coding sequences.
        - **Genome Studies:** Gene prediction underpins comparative genomics, evolutionary studies, and efforts to characterize genomes of new species, supporting biodiversity research and conservation efforts.

Suggested Literature:

- Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329–342. https://doi.org/10.1038/nrg3174
- Jung, H., Ventura, T., Chung, J. S., Kim, W.-J., Nam, B.-H., Kong, H. J., Kim, Y.-O., Jeon, M.-S., & Eyun, S. (2020). Twelve quick steps for genome assembly and annotation in the classroom. PLOS Computational Biology, 16(11), e1008325. https://doi.org/10.1371/journal.pcbi.1008325
- Dong, Y., Li, C., Kim, K., Cui, L., & Liu, X. (2021). Genome annotation of disease-causing microorganisms. *Briefings in Bioinformatics*, *22*(2), 845–854. https://doi.org/10.1093/bib/bbab004