Bioinformatics and Predictions

Gianlucca de Urzêda Alves

2024-09-25

Bioinformatics and Predictions

Use of experimentation its too expensive and take too long to characterize organisms. Therefore, using bioinformatics software's and algorithms helps speed up the process. The main strategies of gene prediction are: Ab Initio, Homology-Based (Evidence-Based) and Integrated-Based approaches.

Ab Initio Prediction

Ab initio gene prediction is a computational approach that predicts gene structures based solely on the DNA sequence itself, without relying on existing annotations or evidence. This method uses models of gene structure and signals to identify potential genes within genomic sequences.

1. **Principles:**
    - **Signal Sensors:** These are short sequence motifs that are indicative of specific gene features, including:
        - **Splice Sites:** The locations where RNA splicing occurs, separating exons from introns.
        - **Branch Points and Polypyrimidine Tracts:** Key sequences involved in splicing regulation.
        - **Polyadenylation site**: Sequences that indicate a region for the addition of multiple adenine bases as signal for mRNA migration from nucleus to cytosol.
        - **Start and Stop Codons:** Indicators of the beginning and end of a protein-coding region.
    - **Content Sensors:** These analyze the broader sequence context, such as:
        - **Codon Usage Patterns:** Specific to species, these patterns help distinguish coding regions from non-coding regions using statistical models.
        - **Nucleotide Composition:** Variations in GC content, repetitive elements, and sequence complexity that differ between coding and non-coding regions.
2. **Algorithms and Models:**
    - **Dynamic Programming:** A computational approach that breaks down complex problems into simpler sub-problems, used for optimizing gene structure predictions.
    - **Linear Discriminant Analysis:** A statistical method that classifies sequences as coding or non-coding based on multiple features.
    - **Linguistic Methods:** These apply principles of language processing to predict gene structures by recognizing patterns similar to grammatical rules in sequences.

- **Hidden Markov Models (HMMs):** Probabilistic models that capture sequence features and their dependencies, widely used for gene prediction.
- **Neural Networks:** Machine learning models that learn complex patterns in data, allowing for the prediction of gene structures with high accuracy.

3. **Advantages:**

   - **Independence from Known Sequences:** Ab initio methods can predict novel genes that have no homologs in existing databases, making them important tools for annotating newly sequenced genomes.
   - **Versatility:** Applicable to a wide range of organisms, including those with few or no closely related reference genomes.

4. **Limitations:**

   - **Lower Accuracy for Complex Genomes:** Ab initio predictions can be less accurate for genomes with complex structures, such as those with extensive alternative splicing or overlapping genes.
   - **High False Positive Rates:** Without homology evidence, ab initio methods may generate predictions that do not correspond to functional genes, necessitating further validation.

In general, Ab initio gene prediction programs need to be trained on the sequences of target organisms, and the composition of this training set defines the accuracy of the statistical model. Specially on new genomes, without similar datasets to train a model, will lead the Ab initio gene prediction programs to generate a higher rate of false-positive gene identification. The same will happen when these programs are trained into a organism and used in a different organism. Their accuracy is tied to the organism used in the training set.

Homology-Based Approach

The homology-based approach utilizes evolutionary conservation to predict genes by comparing unknown sequences with those from other organisms with known annotations. This method is grounded in the principle that functional regions, like exons, are more conserved across species than non-functional regions, such as intergenic or intronic sequences. By identifying similarities between sequences, this approach infers the structure and potential function of unknown genomic regions.

1. **Principles:**

   - **Sequence Conservation:** Functional regions (e.g., exons) are typically conserved across species, making sequence similarity a indicator of gene presence and structure.
   - **Homology Inference:** By aligning sequences from expressed sequence tags (ESTs), proteins, cDNA, or well-annotated genomes, researchers can extrapolate gene structures in the target genome, effectively transferring annotations from model organisms.

2. **Advantages:**

- **High Accuracy:** This approach achieves high accuracy when homologous sequences are available, enabling precise gene models based on existing annotations.
- **Functional Insights:** By aligning unknown sequences to known references, similarity-based methods not only predict gene structures but also provide functional annotations, such as identifying protein-coding genes, motifs, and other regulatory elements.

3. **Limitations:**

- **Dependency on Known Data:** The success of similarity-based methods is limited by the availability and quality of homologous sequences in databases. Novel or highly divergent genes may be overlooked if no similar sequences are present.
- **Low Resolution in Non-Conserved Regions:** Non-conserved regions, such as species-specific genes or non-coding RNAs, are often missed due to insufficient sequence similarity.

In general, the best approaches for Homology gene prediction is using local alignment methods, using algorithms such as Smith-Waterman and BLAST suite. The majority of genes from an unknown organism can be inferred using this approach, but they do not ensure accurate gene architecture.

Integrated-Based Approaches

To overcome the limitations of each individual method, integrated approaches combine similarity-based and ab initio predictions, along with experimental data (e.g., RNA-Seq, ESTs), to enhance the accuracy and reliability of gene annotations.

- **Hybrid Models:** Tools like MAKER, Ensembl, and NCBI's RefSeq pipelines utilize integrated models that combine ab initio predictions with sequence homology and evidence-based data.
- **Evidence Integration:** Incorporating transcriptome data (e.g., RNA-Seq) provides direct evidence of gene expression, improving the prediction of exon-intron structures and alternative splicing events.
- **Iterative Refinement:** Combining predictions from multiple methods allows for iterative refinement, reducing false positives and enhancing the resolution of gene models.

Gene prediction remains a dynamic and evolving field, with ongoing improvements driven by advances in computational methods, machine learning, and the increasing availability of genomic and transcriptomic data.

Suggested Literature:

Picardi, E., & Pesole, G. (2010). Computational Methods for Ab Initio and Comparative Gene Finding. Em O. Carugo & F. Eisenhaber (Orgs.), *Data Mining Techniques for the Life Sciences* (Vol. 609, p. 269–284). Humana Press. https://doi.org/10.1007/978-1-60327-241-4_16