

Experiment 7: Retrieving DNA from UCSC Genome Browser and Gene Prediction using AUGUSTUS

AIM

To retrieve genomic DNA sequences from UCSC Genome Browser and perform computational gene prediction using AUGUSTUS tool to identify potential genes, their structure, and coding regions in the retrieved sequences.

THEORY

Introduction to UCSC Genome Browser

The UCSC Genome Browser is a comprehensive web-based tool that provides rapid and reliable visualization of genomic data at any scale. Developed and maintained by the University of California, Santa Cruz, it serves as a centralized repository for genomic sequence data from numerous organisms. The browser integrates vast collections of aligned annotation tracks, including known genes, predicted genes, expressed sequence tags (ESTs), mRNAs, CpG islands, single nucleotide polymorphisms (SNPs), and comparative genomics data. Researchers worldwide rely on this tool for visualizing genomic contexts, understanding gene structures, and exploring evolutionary relationships between sequences across species. The browser's strength lies not only in its visualization capabilities but also in its ability to extract specific genomic sequences for downstream analysis, making it an essential starting point for computational genomics projects.

The Table Browser feature of UCSC provides programmatic access to the underlying data, allowing researchers to retrieve sequences, annotations, and other genomic features in various formats. This tool enables precise extraction of genomic regions with customizable flanking sequences, making it ideal for gene prediction exercises where adequate upstream and downstream contexts are crucial for accurate analysis. The browser maintains multiple genome assemblies for each organism, ensuring researchers can access both current and historical versions of genomic data for reproducibility and comparative studies.

Gene Prediction Fundamentals

Gene prediction represents one of the fundamental challenges in computational biology, involving the identification of genomic DNA regions that encode genes. This complex process requires distinguishing coding sequences from non-coding regions, identifying regulatory elements, and determining the precise boundaries of genes. In eukaryotic genomes, this task is particularly challenging due to the presence of introns, alternative splicing, and large intergenic regions. The process must accurately identify protein-coding genes, RNA genes, regulatory regions, splice sites, and translation start and stop signals. The complexity increases further when considering that only about

1-2% of the human genome codes for proteins, while the remaining sequence contains regulatory elements, non-coding RNAs, and repetitive elements that can confound prediction algorithms.

Modern gene prediction employs three main methodological approaches that often work in combination. Ab initio or intrinsic methods use statistical models of gene structure derived from training sets of known genes. These methods analyze sequence composition, codon usage patterns, and splice site signals to identify potential genes without requiring external evidence. Homology-based methods leverage evolutionary conservation by comparing genomic sequences to databases of known genes and proteins from related organisms. Evidence-based methods incorporate experimental data such as expressed sequence tags, RNA-seq data, or protein sequences to guide and validate predictions. The integration of these approaches has significantly improved prediction accuracy, though challenges remain, particularly for genes with unusual structures or those expressed at low levels.

AUGUSTUS Gene Prediction Tool

AUGUSTUS stands as one of the most accurate programs for ab initio gene prediction in eukaryotic genomic sequences. Developed at the University of Greifswald, it employs a sophisticated generalized hidden Markov model (GHMM) that captures the statistical properties of gene structures. The model encompasses various genomic features including exon-intron structures, splice site signals with their consensus sequences, translation initiation and termination signals, branch points for splicing, and both 5' and 3' untranslated regions. This comprehensive modeling allows AUGUSTUS to predict complete gene structures rather than just coding regions, providing researchers with detailed information about gene organization that is crucial for understanding gene function and regulation.

The statistical foundation of AUGUSTUS relies on several key components that work together to identify genes. The program uses Markov chains of different orders to model the distinct sequence composition of coding and non-coding regions, capturing the characteristic triplet periodicity of coding sequences. Position weight matrices are employed to recognize splice sites, with separate models for donor and acceptor sites that account for both canonical and non-canonical splice signals. The length distributions of exons and introns are modeled using empirical distributions derived from training data, allowing the program to favor typical gene structures while still permitting unusual but valid configurations. Additionally, AUGUSTUS considers GC content variations and isochoore structures, which is particularly important for mammalian genomes where gene density correlates with GC content.

One of AUGUSTUS's key strengths lies in its species-specific parameter sets, which have been carefully trained on high-quality gene sets from various organisms. These parameters capture the unique characteristics of gene structures in different species, from the compact genes of fungi to the complex multi-exon genes of vertebrates. The program can also incorporate hints from extrinsic evidence such as RNA-seq data, protein alignments, or conservation information, allowing it to combine the strengths of ab initio and evidence-based approaches. Furthermore, AUGUSTUS can predict alternative splice forms, identify genes on both DNA strands simultaneously, and provide confidence scores for its predictions, making it a versatile tool for genome annotation projects. The

web interface provides accessible entry point for educational purposes, while the command-line version offers additional features for large-scale genomic analyses.

PROCEDURE

Part A: Retrieving DNA from UCSC Genome Browser

1. Access UCSC Genome Browser:

- Open web browser and navigate to <http://genome-asia.ucsc.edu>
- Click on "Genomes" in the top menu bar

2. Select Your Organism and Gene:

- Choose your organism of interest from the available options
- Select the appropriate genome assembly version
- Click on **Genome Browser** and the **Back to Genome Browser**
- In the search tab, enter
 - A gene symbol of your choice
 - OR specific genomic coordinates
- Click "Search" button

3. Navigate to Table Browser:

- Once the genome browser loads and shows your selected region
- Click on "Tools" in the top menu
- Select "Table Browser" from the dropdown menu

4. Configure Table Browser for Sequence Retrieval:

- Verify and adjust these settings:
 - Assembly: (should match your selection)
 - Group: "Genes and Gene Predictions"
 - Track: Select an appropriate annotation track (maybe NCBI RefSeq)
 - Table: Choose suitable gene table (RefSeq All)
- Region settings:
 - Select "position" and verify your coordinates
 - Adjust if needed to capture the complete gene
- Output format: Select "sequence"
- Click "get output"

5. Configure Sequence Options:

- On the next page, configure:
 - Select "Genomic" for DNA sequence
 - Add upstream and downstream padding as needed (suggested: 500 bases)
 - Select "One FASTA record per region"
- Click "get sequence"

6. Save Sequence:

- Copy the FASTA formatted sequence
- Save to a text file with descriptive naming
- Record the sequence length and coordinates for your report

Part B: Gene Prediction using AUGUSTUS

1. Access AUGUSTUS Web Server:

- Navigate to <http://bioinf.uni-greifswald.de/augustus/>
- Click on "Submit a job" or "Web interface" link

2. Upload Your Sequence:

- Upload your FASTA file from Part A
- OR paste your sequence directly in the text box
- Ensure sequence length is appropriate for web server (1kb - 3Mb)

3. Select Prediction Parameters:

- **Species:** Select the most appropriate species model for your organism
- **Strand:** Select "both" to predict on both strands
- **Alternative transcripts:** Configure based on your requirements

4. Submit Job:

- Click "Run AUGUSTUS"
- Note any job ID provided
- Wait for completion

5. Retrieval and Interpret Results:

- **Analyze Gene Predictions:**
- Document the following for each predicted gene:
 - Gene coordinates
 - Number and size of exons
 - Strand orientation
 - Protein length
- Save results screenshots with appropriate naming

DISCUSSION

- 1. Codon Analysis:** What is a codon? From your AUGUSTUS output, identify the exact nucleotide sequences of the start codon and stop codon in your predicted gene.
- 2. Understanding Ab Initio Prediction and Potential Improvements:** Your AUGUSTUS output contains the message "hints for 0 sequences" in the header. Explain what this means about how AUGUSTUS made its predictions. Describe what type of experimental evidence would help improve the gene prediction by AUGUSTUS.