UCSC Genome Browser

Introduction to \textit{ab initio} and evidence-based gene finding

Wilson Leung 06/2006

Outline

- Introduction to annotation
- \textit{ab initio} gene finding
- Basics of the UCSC Browser
- Evidence-based gene finding using the UCSC Browser
AAACAACATCTAAAATAGAGGAAGTTTTCGGAATATACGATAAGTGAAAT
ATCGTTCTTAAAAAAAGACGAAGACGCTTTAAACCTTGAATAAAACAGATTATT
CCAAATAGCGTAAATGCAATGAGCGATGCCGAGGCGGCTTTTCTC
TCAGGATGCTGGGATGAGCAATTCGCTAAATCGGATTTGAGGAGGT
GCCCGATTAGCGCCCGGTTTGGACGCTGTCATCCAAATATGGCCGAGATGGCG
CCGCGCAATACCATCAGTCGAGCTGGGGATGCCAGCACGTTAATCAGGATACCAATTGAGGAGGT
GCCCRAGCTCACCEAGAGCCGCCGCAAATAGGACCCCATCGGCGGGGCCGCTTA
TGTGGAAGCCCAAAAATATAACATAGCAACCGATTTGTGGAATATGCAAATTT
AAGAACCGCGTACGCCACCCGCTCAACAAAGTGCCCAAAGCCATCTTTGGG
GCTAATGGCCCTCATCAATTTGGGGCGGAACCTTGGGGCGAGAAGCATAGTGCC
GCCGATAGACACCAGCTTTGGACGCTGTCATCCAHATAECAACAGT
CTGGTGTGCAGTCGTGTGCTAATGCGCTGCTGGTGGCGCGCTGTCGTTGCC
CCGATGGGAATATCAATGCGCGCAGCAACAGGACCTGCGTGGCGTCGCTGC
GCCGAGGATTTATATTTAGGAGCTGGGAGTAGTCAGAATTTGAGGAGGT
GCCGACAGTTGGGGACATCATCCTCAAGATGTGCTCAAAATATGGCCGAGATGGCG
CCGCGCAATACCATCAGTCGAGCTGGGGATGCCAGCACGTTAATCAGGATACCAATTGAGGAGGT
GCCCRAGCTCACCEAGAGCCGCCGCAAATAGGACCCCATCGGCGGGGCCGCTTA
TGTGGAAGCCCAAAAATATAACATAGCAACCGATTTGTGGAATATGCAAATTT
AAGAACCGCGTACGCCACCCGCTCAACAAAGTGCCCAAAGCCATCTTTGGG
GCTAATGGCCCTCATCAATTTGGGGCGGAACCTTGGGGCGAGAAGCATAGTGCC
GCCGATAGACACCAGCTTTGGACGCTGTCATCCAHATAECAACAGT
CTGGTGTGCAGTCGTGTGCTAATGCGCTGCTGGTGGCGCGCTGTCGTTGCC
CCGATGGGAATATCAATGCGCGCAGCAACAGGACCTGCGTGGCGTCGCTGC
GCCGAGGATTTATATTTAGGAGCTGGGAGTAGTCAGAATTTGAGGAGGT
GCCGACAGTTGGGGACATCATCCTCAAGATGTGCTCAAAATATGGCCGAGATGGCG
CCGCGCAATACCATCAGTCGAGCTGGGGATGCCAGCACGTTAATCAGGATACCAATTGAGGAGGT
GCCCRAGCTCACCEAGAGCCGCCGCAAATAGGACCCCATCGGCGGGGCCGCTTA

**What to Annotate?**

- **Genes**
  - Novel genes, known genes, pseudogenes
- **Non-coding RNAs**
  - tRNAs, miRNAs, siRNAs
- **Regulatory Elements**
  - Promoters, enhancers, silencers
- **Repeats**
  - Transposable elements, simple repeats
**ab initio** Gene Prediction

*ab initio* = From the Beginning

- Gene prediction using only the genomic DNA sequence
  - Search for “signals” of protein coding regions
  - Typically uses a probabilistic model
    - Hidden Markov Model (HMM)
- Requires external evidence to support predictions (mRNA, ESTs)

**Performance of Gene Finders**

- Most gene-finders can predict prokaryotic genes accurately
- However, gene-finders do a poor job of predicting genes in eukaryotes
  - Not as much is known about the general properties of eukaryotic genes
  - Splice site recognition, isoforms
"ab initio" Gene Predictions

- Examples:
  - *Glimmer* for prokaryotic gene predictions
  - *Genscan* for eukaryotic gene predictions
    - (Burge and Karlin 1997)

*Genscan* is the gene finder we will use for our chimpanzee and Drosophila annotations.

Genscan Gene Model

- Genscan considers the following:
  - Promoter signals
  - Polyadenylation signals
  - Splice signals
  - Probability of coding and non-coding DNA
  - Gene, exon and intron length
Common Problems

- Common problems with gene finders
  - Fusing neighboring genes
  - Spliting a single gene
  - Missing exons or entire genes
  - Overpredicting exons or genes

- Other challenges
  - Nested genes
  - Noncanonical splice sites
  - Pseudogenes
  - Different isoforms of same gene

How to Improve Predictions?

- New gene finders use additional evidence to generate better predictions:
  - *Twinscan* extends model in Genscan by using homology between two related species
  - Separate model used for exons, introns, splice sites, UTR’s

How to Improve Predictions?

- **Manual annotation**
  - Collect evidence from multiple biological and computational sources to create gene models
  - This method still generates the best annotations

- **Need a place to collate all the different lines of evidence available**
  - UCSC Browser

---

**Introduction to the UCSC Browser**

**About the UCSC Genome Bioinformatics Site**

This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides a portal to the ENCODE project.

During development of the Genome Browser, researchers and software developers were given access to a collection of tools that can be used to test and develop new approaches to genome analysis.

The Genome Browser currently supports over 30,000 batches of data from a variety of sources, including the ENCODE project. These data include gene expression data, protein-protein interactions, transcription factor binding sites, and other types of genomic data.

**UCSC Genome Browser**

The UCSC Genome Browser is a web-based tool that allows researchers to explore and analyze genomics data. The UCSC Genome Browser provides a powerful and intuitive interface for exploring and analyzing genomics data, and it is available to researchers around the world.

**http://genome.ucsc.edu**
UCSC Browser Developers

UCSC Browser is created by the Genome Bioinformatics Group of UC Santa Cruz

Development team: http://genome.ucsc.edu/staff.html
- Led by Jim Kent and David Haussler

UCSC Browser was initially created for the human genome project
- It has since been adapted for many other organisms

We have set up local version of the UCSC Browser for Bio 4342

Functions of UCSC Browser

- Functionalities of UCSC Browser
  - Genome Browser - views of genomic regions
  - BLAT - BLAST-Like Alignment Tool
  - Table Browser - SQL access to genomic data

- Training section on the UCSC web site
    - Pre-recorded tutorial (presentation slide set)
    - Reference cards
Chimp BAC Analysis

Goal: Annotation of one of the features in a 170 kb chimpanzee BAC

A more detailed walkthrough is available

Genscan was run on the repeat-masked BAC using the vertebrate parameter set (GENSCAN_ChimpBAC.html)
- Predicts 8 genes within this BAC
- By default, Genscan also predicts promoter and poly-A sites; however, these are generally unreliable
- Output consists of map, summary table, peptide and coding sequences of the predicted genes

Chimp BAC Analysis

Analysis of Gene 1 (423 coding bases):
- Use the predicted peptide sequence to evaluate the validity of Genscan prediction

blastp of predicted peptide against the nr database
- Typically uses the NCBI BLAST page:
- Choose blastp and search against nr
- For the purpose of this tutorial, open blastpGene1.html
Interpreting blastp Output

- Many significant hits to the nr database that cover the entire length of the predicted protein
- Do not rely on hits that have accession numbers starting with XP
  - XP indicates RefSeq without experimental confirmation
- Click on the Score for the second hit in the blastp output (gb|AAH70482.1)
  - Indicates hit to human HMGB3 protein

Investigating HMGB3 Alignment

- The full HMGB3 protein has length of 200 aa
  - However, our predicted peptide only has 140 aa
- Possible explanations:
  1. Genscan mispredicted the gene
     - Missed part of the real chimp protein
  2. Genscan predicted the gene correctly
     - Pseudogene that has acquired an in-frame stop codon
     - Functional protein in chimp that lacks one or more functional domains when compared to the human version
Analysis using UCSC Browser

- Go back to Genscan output page and copy the first predicted coding sequence

- Navigate to UCSC browser @ http://genome.ucsc.edu

- Click on “BLAT”
  - Select the human genome (May 2004 assembly)
  - Paste the coding sequence into the text box
  - Click “submit”

Human Blat Results

- Predicted sequence matches to many places in the human genome
  - Top hit shows sequence identity of 99.1% between our sequence and the human sequence
  - Next best match has identity of 93.6%, below what we expect for human / chimp orthologs (98.5% identical)

- Click on “browser” for the top hit (on chromosome 7)
  - The genome browser for this region in human chromosome 7 should now appear
  - Navigation buttons are on the top menu bar

- Reinitialize the browser by clicking on “hide all”
Adjusting Display Options

- Adjust following tracks to “pack”
  - Under “Mapping and Sequencing Tracks”:
    - Blat Sequence

- Adjust following tracks to “dense”
  - Under “Gene and Gene Prediction Tracks”:
    - Known Genes, RefSeq, Ensembl Genes, Twinscan, SGP Genes, Genscan Genes
  - Under “mRNA and EST Tracks”:
    - Human mRNAs, Spliced ESTs, Human ESTs, Other ESTs
  - Under “Comparative Genomics”:
    - Mouse Net
  - Under “Variation and Repeats”:
    - RepeatMasker

Human Genome Browser

- Hit “refresh” and look at new image; zoom out 3x to get a broader view.

- There are no known genes in this region
  - Only evidence is from hypothetical genes predicted by SGP and Genscan
  - SGP predicted a larger gene with two exons
  - There are also no known human mRNA or human ESTs in the aligned region
  - However, there are ESTs from other organisms
Investigate Partial Match

- Go to GenBank record for the human HMGB3 protein (using the BLAST result)

- Click on the “Display” button and select “FASTA” to obtain the sequence

- Go back to the BLAT search to search this sequence against the human genome assembly (May 2004)

BLAT search of human HMGB3

- Notice the match to part of human chromosome 7 we observed previously is only the 7th best match (identity of 88%)
  - Consistent with one of our hypotheses that our predicted protein is a paralog

- Click on “browser” to see corresponding sequence on human chromosome 7
  - BLAT results overlap Genscan prediction but extend both ends
  - Why would Genscan predict a shorter gene?
Examining Alignment

Now we need to examine the alignment:
- Go back to previous page and click on "details"

In general, the alignment looks good except for a few changes
- However, when examining some of the unmatched (black) regions, notice there is a "tag" - a stop codon.

Confirm predicted protein is in frame relative to human chromosome 7 by
- Looking at the side-by-side alignment

Confirming Pseudogene

Side-by-side alignment color scheme
- Lines = match
- Green = similar amino acids
- Red = dissimilar amino acids

We noticed a red “X” (stop codon) aligning to a “Y” (tyrosine) in the human sequence
Confirming Pseudogene

 Alignment after stop codon showed no deterioration in similarity suggest our prediction is a recently retrotransposed pseudogene.

 To confirm hypothesis, go back to BLAT results and get the top hit (100% identity on chromosome X).

 The real HMGB3 gene in human is a 4-exon gene!

Conclusions

 Based on evidence accumulated:

- As a cDNA, the four-exon HMGB3 gene was retrotransposed.
- It then acquired a stop codon mutation prior to the split of the chimpanzee and human lineages.
- Retrotransposition event is relatively recent.
  - Pseudogene still retains 88.8% sequence identity to source protein.
Questions?