# Gene Finding in Chimpanzee

Evidence based improvement of *ab initio* gene predictions

Chris Shaffer 06/2009

## Chimp Analysis

- Curriculum, GEP web site and in your binder
- At Washington Univ used after BLAST exercise 1 and 2
  - Exposure to mammalian genomes
  - Practice skills, computational and cognitive
  - Skipped at some schools
  - New drosophila version coming
- Two parts
  - BAC analysis - in class worksheet
  - Chimp chunks - selected regions of the chimp genome are annotated by groups of students (2 or 3) ends with paper and presentation

## Agenda

- Abridged version of Bio4342 lecture (next 5 slides)
- Work together on one chimp feature from "BAC analysis"
- Optional work on chimp chunk individually with help from TA's

## *Basic Strategy for Annotation*

- Use *Ab Initio* prediction to focus attention on genomic features (areas) of interest
- 80% failure rate; where are the mistakes?
- Add as much other evidence as you can to refine the gene model and support your conclusion
- What other evidence is there?
  1. Basic gene structure
  2. Motif information
  3. BLAST homologies: nr, protein, est
  4. Other species or other proteins

## *Chimpanzee annotation*

1. Basic gene structure
   - Only ~15% of known mammalian genes have 1 exon
   - Many pseudogenes are mRNA's that have retro-transposed back into the genome; many of these will appear as a single exon genes
   - Increase vigilance for signs of a pseudogene when considering any single exon gene
   - Alternatively, there may be missing exons

## *Chimpanzee annotation*

2. Motif information
   - Genscan uses statistical methods to predict genes, will tag all apparent ORFs of sufficient length
   - Since genomes are very large, statistical methods will give some false positives

     (sequence looks like a gene simply by chance)
   - If the predicted gene has protein motifs found in other proteins it is much less likely to be false positive and more likely to be a real gene or a real pseudogene

## *Chimpanzee annotation*

3. BLAST homology: nr, protein, EST
   - Homology to known proteins argues against false positive
   - Mammals have many gene families and many pseudogenes (both of these can show high similarity to your predicted gene)
   - Consider length, percent identity when examining alignments. Human vs. chimp orthologs should differ by <1%; most paralogs or homologs will differ by more than this
   - Without good EST evidence you can never be sure; make your best guess and be able to defend it

## *Chimpanzee annotation*

4. Other species or other proteins
   - For any similarity hit, look for even better hits elsewhere in the genome; paralogs and pseudogenes will look similar but will usually have an even better hit somewhere else.
   - If you are convinced you have a gene and it is a member of a multi-gene family, be sure to pick the right ortholog
   - Look at synteny with properly distant species (mouse or rat); evidence for a transposition suggests a pseudogene

## Chimp BAC analysis

- Worksheet in your folder, follow along, ask for help
- Genscan was run on the repeat-masked BAC using the vertebrate parameter set (GENSCAN_ChimpBAC.html)
  - Genscan is a good *ab initio* gene finder
  - 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:
    - http://www.ncbi.nlm.nih.gov/blast/
  - 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
  - NP indicates RefSeq that has been validated by the NCBI staff
- Click on the Score for the second hit in the blastp output (gbIAAH70482.1)
  - 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
- Best Source further evidence human genome

## 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

## Human Genome Browser

- 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 page to use this sequence to search 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

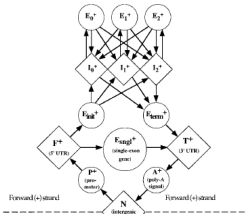
## Questions?

## *ab initio* Gene Finders

- Examples:
  - *Glimmer* for prokaryotic gene predictions
    - (S. Salzberg, A. Delcher, S. Kasif, and O. White 1998)
  - *Genscan* for eukaryotic gene predictions
    - (Burge and Karlin 1997)
- We will use Genscan for our chimpanzee and Drosophila annotations

# Genscan Gene Model

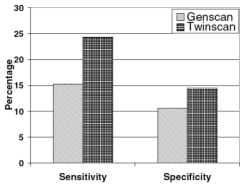
- Genscan considers the following:
  - Promoter signals
  - Polyadenylation signals
  - Splice signals
  - Probability of coding and non-coding DNA
  - Gene, exon and intron length




Chris Burge and Samuel Karlin, *Prediction of Complete Gene Structures in Human Genomic DNA*, JMB. (1997) **268**, 78-94

# 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




Ian Korf, et al. *Integrating genomic homology into gene structure prediction*. Bioinformatics. (2001) **17** S140-S148.

# Gene Annotation System

- All Ensembl gene predictions are based on experimental evidence
- Predictions based on manually curated Uniprot/Swissprot/Refseq databases
- UTRs are annotated only if they are supported by EMBL mRNA records

*Val Curwen, et al. The Ensembl Automatic Gene Annotation System Genome Res., (2004)* ***14*** *942 - 950.*

# UCSC Browser

- 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