Gene Finding in Chimpanzee

Evidence-based improvement of
ab initio gene predictions

Last Update: 08/2021 Chris Shaffer 06/2009

1

Chimp Analysis

- Prerequisites (*BLAST* exercises):
  - Detecting and Interpreting Genetic Homology
  - Using mRNA and EST Evidence in Annotation
- Learning objectives:
  - Exposure to mammalian genomes
  - Practice computational and cognitive skills
- Two parts:
  - BAC analysis — in class worksheet
  - Chimp chunks — selected regions of the chimp genome are annotated by groups of 2–3 students; ends with paper and presentation

2

Agenda

- Abridged version of Bio 4342 lecture (next 5 slides)
- Work together on one chimp feature from "BAC analysis"
- Optional work on chimp chunk individually with help from TA's

3

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, ESTs
  4. Other species or other proteins

4

Chimpanzee Annotation

1. Basic gene structure
   - Only **~15%** of known mammalian genes have **one exon**
   - Many **pseudogenes** are mRNAs that have **retrotransposed** 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

5

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 a false positive and more likely to be a real gene or a real pseudogene

6

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 can show high sequence similarity to your predicted gene
   - Consider **length** and **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 or RNA-Seq evidence you can never be sure; make your best guess and be able to defend it

7

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

8

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

9

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:
    - https://blast.ncbi.nlm.nih.gov/Blast.cgi
    - Click on the "Protein *BLAST*" image
    - Select the "*blastp*" algorithm
    - Search against the nr database
  - For the purpose of this tutorial, open blastpGene1.txt

10

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 "Description" for the best curated RefSeq hit in the *blastp* output (**NP_001288157.1**)
  - Indicates hit to human HMGB3 protein

11

Investigating HMGB3 Alignment

- The full HMGB3 protein has length of **200 aa**
  - However, our predicted peptide only has **140 aa**
- Possible explanations:
  - *Genscan* mispredicted the gene
    - Missed part of the real chimp protein
  - *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 from the **human genome**

12

Analysis Using the *UCSC Genome Browser*

- Go back to *Genscan* output page and copy the first predicted coding sequence
- Navigate to the *UCSC Genome Browser* at https://genome.ucsc.edu
- Click on the "***BLAT***" link (under "Our tools")
  - Select the "**Human**" genome
  - Select the "**Mar. 2006 (NCBI36/hg18)**" assembly
  - Paste the coding sequence into the text box
  - Click "Submit"

13

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

14

Human *UCSC 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 mRNAs or human ESTs in the aligned region
  - However, there are ESTs from other organisms

15

Investigate Partial Match

- Go to GenBank record for the human HMGB3 protein (using the *BLAST* result)
- Click on the "**FASTA**" link to obtain the sequence
- Go back to the *BLAT* search page to use this sequence to search the human genome assembly
  - **Mar. 2006 (NCBI36/hg18)**

16

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 **88.8%**
  - 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?

17

Examining Alignment

- Now we need to examine the alignment:
  - Go back to previous page and click on "**details**"
- 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
- Examine the **side-by-side alignment** to confirm that the "TAG" sequence is an in-frame stop codon on human chromosome 7
  - This in-frame stop codon caused *Genscan* to predict a shorter gene

18

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

19

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 has four coding exons!

20

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
  - The retrotransposition event is relatively recent
    - Pseudogene still retains 88.8% sequence identity to the source protein

21

Questions?

22