CSE 527
Computational Biology
Lecture Notes
11/16/05
Kelly Stevens, stevensk@u.

Gene Prediction

Claverie JM (1997) "Computational methods for the identification of genes…" Human Molecular Genetics, 6(10) 1735-1744. This is a good article, but note that it is a bit old.

Another paper evaluated a half-dozen leading contenders for gene structure prediction programs of the day. This paper has good results/methodology. "Burset M, Guigo R (1996), Evaluation of gene prediction programs" Genomics 34(3) 353-367.

Additionally, lots of information is flooding into GenBank. The motivation of gene prediction is to decode massive amounts of information as efficiently as possible.

One goal is to create programs that will automatically annotate new sequence data of new organisms. Currently, gene predictions programs can annotate ~60% correctly. If programs are based on known information of another organism (if find something similar to known protein in another organism), annotation can be ~80% correct. Note that 60% and even 80% are far from perfect…there is much room for improvement here! There is also still a need for laboratory verification.

For the purpose of today's lecture, we will assume that a single gene defines a single protein. This is the "one gene one protein" hypothesis. In other words, one gene serves as the DNA template that defines a single protein. But…it now seems that one gene can produce many different proteins (through alternative splicing.) Also, sometimes a gene can code for RNA that does not in turn code for proteins. This is called non-coding RNA and and likely plays regulatory roles. Also, there are 20 amino acids but only 4 bases. Hence need 3 bases to specify all 20 amino acids. So, 3 bases/letters always defines one Amino acid of a protein. Note that ALL possible combinations of 3 bases aren't needed to specify all amino acids…hence some combinations can be used for start and stop signals, etc.

The genetic code
Specific codons (3 base sequences) code for specific amino acids. For example, "AUG" codes methionine. Note that several amino acids have multiple codons (e.g., leucine has 6 codons.)

Translation: mRNA to Protein.
Proteins called ribosomss latch onto an mRNA molecule and walks down the mRNA until it finds the "start codon" AUG (which codes for methionine.) The ribosome then startes coding triplet after triplet of mRNA sequence into amino acid sequence. Note that multiple ribosomes can be present on a given RNA at a given time. Ribosomes allow for

the docking of molecules called transfer RNAs (tRNAs). TRNAs contain a portion of sequence that is the reverse complement of a given codon and also the amino acid that the given codon codes for. Hence, when the proper tRNA is put into place by the ribosome, the proper amino acid is placed onto the growing amino acid peptide chain.

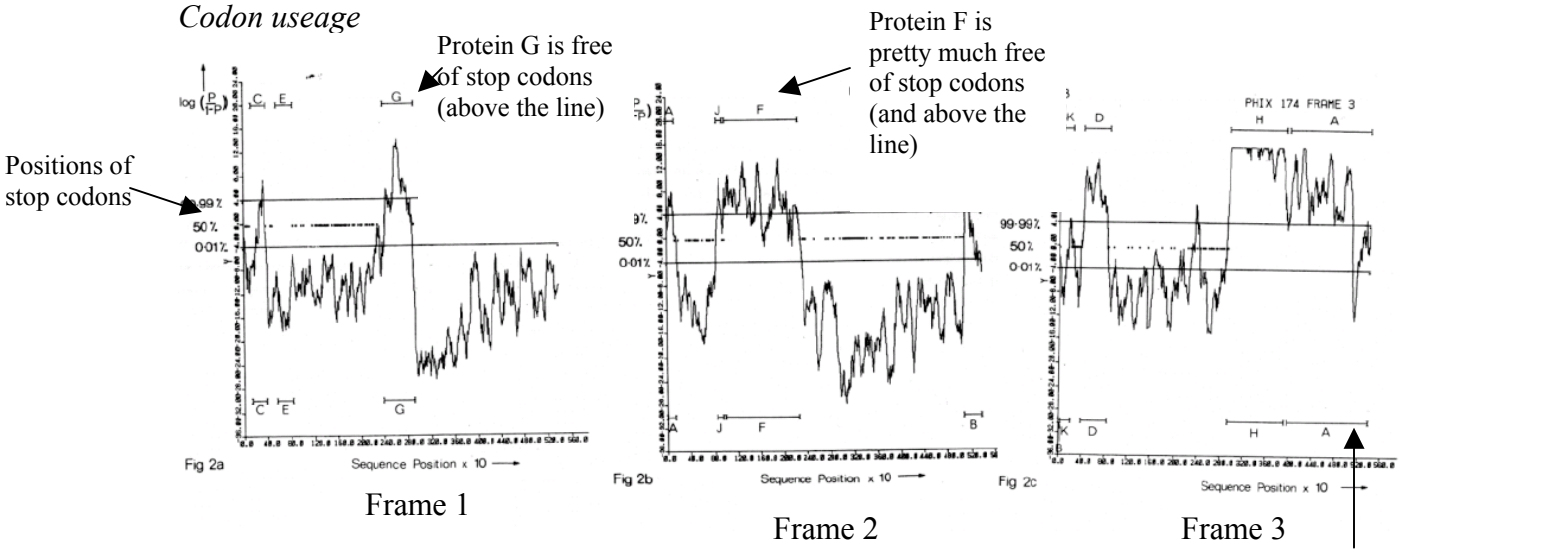
How to find genes??

Idea #1: Long open reading frames
One way to find genes is to look for open reading frames. An "open reading frame" is a sequence of successive codon triplets that does not contain a stop codon until the end of the protein coding region. In other words, it is the "correct frame" of the 3 possible frames (3 possible sequences of triplets) that actually codes the protein. An average open reading frame is 21 codons long in random DNA. A long open reading frame is unlikely to occur by chance alone. Rather, a long open reading frame is usually a protein-coding gene. So then, we could 1) scan the genome, 2) look for every open reading frame, and 3) every long open reading frame would be a candidate for "something interesting."

Idea #2: Codon Frequency
In the coding of real proteins, amino acids are used in frequencies different than that which random change would suggest. For example, there are 6 codons that code for leucine, but the organism does not necessarily have to use them equally. Each organism has bias to which type of codon it prefers to use for coding given amino acids. This is another clue to finding genes, then….if one finds a shortish open reading frame that uses amino acides more like a random ratio rather than the "biased" ratio, it is likely to be "background" non-coding sequence.

Recognizing Codon Bias
How can one evaluate the extent to which codon bias is present? Calculate the probability of that coding sequence in reading frame 1 (P1). Do the same with P2…is the same way only shift over one base. Do the same for P3. Then normalize over P1+P2+P3. Then look at all reading frames compared to background model.

Codon useage



Frame 1

Frame 2

Frame 3

For the most part, sequence above the line corresponds to genes.

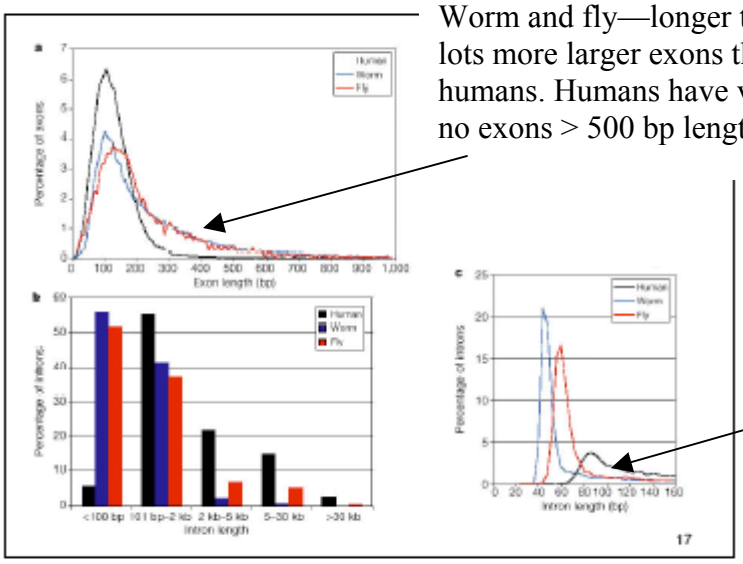
Here, the back end of A looks strange…overlaps with B. This organism actually has the same sequence in different reading frames and codes for two different protreins. This is very rare…it happens most commonly in viiruses

One way to improve gene finding is by modeling promoter regions. For example, could use a WMM model to model the TATA box, a region known to be ~10 bases upstream from transcription start. This works pretty well in prokaryotes.

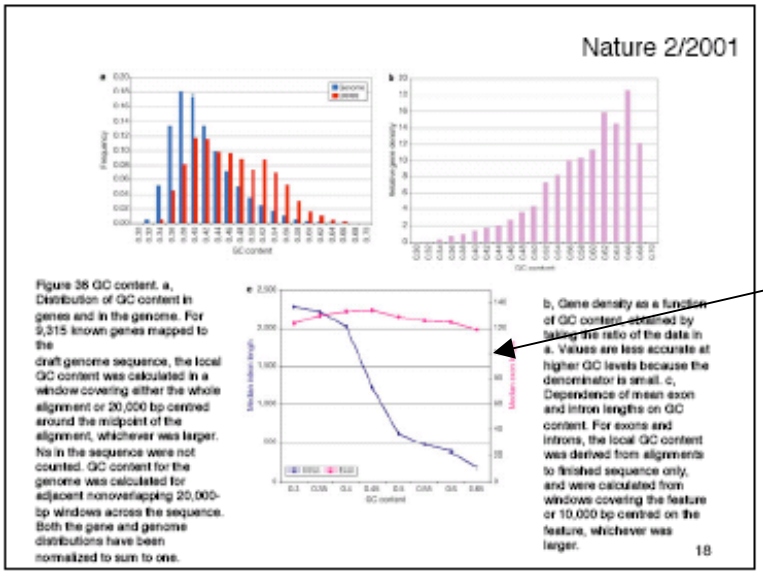
As in prokaryotes, eukaryotes have promoters, transcription start/stop sites, and translation start/stop sites. Eukaryotes also have additional features not found in prokaryotes, including a polyA tail, introns and exons, a branch point signal, and alternative splicing. In eukaryotes, a large fraction of the genes are interrupted with noncoding chunks of DNA. Exons are the coding regions. They tend to be shorter. Introns are the sections of DNA that "interrupt" coding regions. Hence, in eukaryotes, long open reading frames may not be found because exons tend to be short. In eukaryotes, these introns are cut out of "pre-mRNA" by splicosomes prior to transportation of the mRNA out of the nucleus into the cytoplasm. Additionally, a polyA tail gets added onto the mRNA transcript prior to transport to the cytoplasm. Note that consensus sequence patterns exist in DNA at each of the splice sites GT/AT—hence allowing location of the beginning and the end of an intron.

Note also that the majority of eukaryotic genes likely have alternative locations where 3' UTRs end.

Gene size: Many genes are over 100 kb long. The largest gene known is dystrophin, which is 2.4 Mb. It takes ~10 minutes on average to transcribe a gene. It takes hours to transcribe the dystrophin gene.



Humans have only ~25,000 genes! (much less than was originally expected). These genes are likely to be multi-functional. Genes probably produce multiple protein products because of alternative splicing of exons.



Pink line—shows that exon length is not particularly dependent on GC content. But intron length (blue line) varies dramatically with GC content. This is another feature that could be modelined with gene prediction.

For Monday: read the Burge paper (listed in class notes). This paper talks about one particular program for doing gene prediction. It is based on hiddem MM and uses WMM. It exemplifies a nice blend of techniques that we've talked about thus far in class.