CSE 527, Lecture 13, 11-8-2006
Steve Lewis

Talk after class today: Isochores and symmetry breaking

In the human genome there are some regions which are GC rich, some AT rich
more genes in the GC rich regions

HMMs

PFAM - specifically search for alignment of protein "domains" (regions) which constitute structural or functional "units".

basic transition architecture -
    basic line alphabet or inserted state or deleted state (no emission)

Scoring -
  use Baum-Welch to learn the algorithm

  Globins and training data score high

  scoring for pattern is probability / length since odds fall with every increase in length

  There is better scoring to take the ratio of the odds of data being generated by globin model vs. data generated by background model - compare to background

 Alternative - convert to Z score

    (score - mean) / standard deviation

    Z is a way to correct of length variance -

    mean is mean for proteins of similar length

   variance is variance for proteins of similar length

PFAM initial seeds are hand coded; train from hand alignments

Train with hand selected sample

Now automatically classify from Swiss-Prot

8000 families in rFam covering 75% of proteins

Families

   are hierarchical - globin might be the head of a large family

=====

Pseuocounts - this is the insertion of a small count to handle the case where a residue is simply not seen in some position - probability does not handle 0 well

Pseudocount - represents a Bayesian prior probability

More elaborate pseudocount: Dirchlet mixture Dirchlet priors - adding separate pseudocounts for different regions i.e. hydrophobic region, buried region ...

====

Computational gene prediction

Motivation - lots of sequences - are they genes - are they expressed ???

state of the art 60% accuracy - based on first principles

  80% with similarity training

BUT - predictions are VERY expensive to verify

Basics - there is a start Codon , stop codon and prefix and suffix RNA (UTR)

Transcription - DNA ->RNA

Translation - RNA->Protein

Codon table - reviewed

there is are a few variations in codon meaning

might be a dozen different code tables

mitochondria use a slightly different table

in tetrahymena 2 of the stop codons code for other amino acids sometimes so does the third

Gene finding

First issue: define "reading frame" which base is the start of a codon -

in RNA, 3 possible frames; in DNA, 6 frames (3 per strand)

Open frame - long run without stop codon.

in random DNA a stop happens each 21 triplets, on average (64 codons/3 stops)

So, very low odds of long (several hundred codons) open reading frame. long read frame is a good way to look for genes

Idea2 - compare codon frequency with known codon frequency - look at amino acid sequence

also synonym usage is species dependent - certain species preferentially use certain synonyms (Why? probably regulation - if different tRNAs are present in different levels - genes using codons matching more common tRNAs will be translated more eficiently.)

Markov model - compare likelihoods - 5-6 order Markov - say spanning 2 codons

using likelihood in a virus expressed genes spike in probability however there is a issue in a viral DNA genes might overlap and be read in two different frames

in prokaryotes - most DNA is for coding so open frames look good but
- there are a few short genes
- some genes use abnormal sequences

in eukaryotes -

same signals but MANY interons - Phil Sharp matched the mRNA to DNA saw intron loops

mRNA shorter than DNA with much editing

Biology of splicing

splicing performed by Splicosome - at least 50 proteins and half dozen RNA molecules - RNA is the most conserved portion

splicing MUST preserve reading frame

exon length not a multiple of 3 so any errors must be high fidelity

interons are recognized by binding near the start and end of the interon

why have introns

after introns discovered

in tetrahymena - code for ribosomal RNA has an interon but uses much simpler mechanism where RNA self excises intron region; intron region self catalyzes its own excision

Some interons self excise in mRNAs

might there be uses for discarded interons? - Some introns have short RNA segments which can interfere with pieces of other genes (so-called "microRNAs")

Gene finding in eukaryotes

need to find introns and exons. some sequence signals - need to look at

Some human data:
- internal exons mean 122 median 145 bp
- introns 1000 median mean 3300 - REALLY BIG tail - some really long introns
- genes 14kb median 27kb average, but some big ones, e.g. one gene 2.4mb
- 2.4 MB (dystrophin) gene takes 16 hour to transcribe the gene