Gene finding in prokaryotes

Reading frames

• A protein is coded by groups of three nucleotides (codons):
  ...ACGTACGTACGTACGT...
  ...AC-GTAC-TAC-GTA-CT...
  ...TYVRT...
• There are two other ways in which this sequence can be decomposed into codons:
  ...ACG-TAC-TAC-GTA-CT...
  ...AC-GTA-CGT-ACG-TAC-GT...
• These are the three reading frames
• The complementary strand has three additional reading frames

Coding for a protein

• Three nucleotides (codon) code for an amino acid
• Adjacent codons do not overlap
• Each amino acid has more than one codon that codes for it
• The code is almost universal across organisms
• Codons which code for the same amino acid are similar
• Six reading frames in which a protein can be read. How is the right one determined?

Coding for a protein

• Every gene starts with the codon ATG. This specifies the reading frame and the start of translation site.
• The protein sequence continues until a stop codon (UGA, UAA, UAG) is encountered.
• DNA: TAC GGC GGC TAT TAC TGC CAG GAA GGA ACT
  RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
  Protein: Met Ala Pro Ile Met Thr Val Leu Pro Stop

Open Reading Frames (ORFs)

• An open reading frame is a sequence whose length is a multiple of 3, starts with the start codon, and ends with a stop codon, with no stop codon in the middle.
• How do we determine if an ORF is a protein coding gene? Suppose we see a long run of non-stop codons after a start codon, then it has a low probability of arising by chance.

A model1 for DNA sequences

• Need a probabilistic model assigning a probability to a DNA sequence.
• Simplest model: nucleotides are independent and identically distributed.
  Probability of a sequence $s=s_1s_2...s_n$:
  $P(s) = P(s_1)P(s_2)...P(s_n)$
• Distribution characterized by the numbers $(p_A, p_C, p_G, p_T)$ that satisfy: $p_A+p_c+p_G+p_T=1$ (multinomial distribution)

1”All models are wrong. Some are useful” G.E.P. Box
A method for gene finding

- Under our model, assuming that the four nucleotides have equal frequencies:
  \[ P(\text{run of } k \text{ non-stop codons}) = \left(\frac{61}{64}\right)^k \]
- Choose \( k \) such that
  \[ P(\text{run of } k \text{ non-stop codons}) \]
  is smaller than the rate of false positives we are willing to accept. At the 5% level we get \( k=62 \).

Gene finding

- Amino acids have unequal frequencies of use, as do the different codons coding for each amino acid.
- A simple refinement:
  \[ P(\text{stop}) = P(\text{UAA}) + P(\text{UAG}) + P(\text{UGA}) \]
  \[ P(\text{run of } k \text{ non-stop codons}) = (1 - p(\text{stop}))^k \]

Information we ignored

- Coding regions have different nucleotide usage
- Different statistics of di-nucleotides and 3-mers in coding/non-coding regions
- Promoter region contains sequence signals for binding of TFs and RNA polymerase
- Need better models!
- But works surprisingly well in prokaryotes.