CSE 427
Computational Biology

Gene Prediction

A statistical interlude: Fair or biased?

More likely fair or biased?

More likely H0 or H1?

- H0: .5 – .5
- H1: .9 – .1
Quantify likelihood: $H_0$ vs $H_1$

$H H H H T H H T T H$

$H_0: .5 - .5 .5^{10}$

$H_1: .9 - .1 .9^7 .1^3$

Likelihood ratio: $(.5^{10})/(.9^7 .1^3) = .4898$

(i.e., odds favor “biased” by about 2:1)

Gene Finding: Motivation

Sequence data flooding into Genbank
What does it mean?

protein genes, RNA genes, mitochondria, chloroplast, regulation, replication, structure, repeats, transposons, unknown stuff, …

Protein Coding Nuclear DNA

Focus of this lecture
Goal: Automated annotation of new sequence data

State of the Art:
In Eukaryotes:
predictions ~ 60% similar to real proteins
~80% if database similarity used
Prokaryotes
better, but still imperfect
lab verification still needed, still expensive

Biological Basics

Central Dogma:

DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ Protein

Codons: 3 bases code one amino acid
Start codon
Stop codons
3’, 5’ Untranslated Regions (UTR’s)
**Codons & The Genetic Code**

<table>
<thead>
<tr>
<th>First Base</th>
<th>Second Base</th>
<th>Third Base</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
<td>Ala : Alanine</td>
</tr>
<tr>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Stop</td>
<td>Stop</td>
</tr>
<tr>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td>U</td>
</tr>
<tr>
<td>Met/Start</td>
<td>Thr</td>
<td>Lys</td>
<td>Ser</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
</tr>
</tbody>
</table>

**Translation: mRNA → Protein**

**Ribosomes**
Idea #1: Find Long ORF’s

Reading frame: which of the 3 possible sequences of triples does the ribosome read?
Open Reading Frame: No stop codons
In random DNA
  average ORF = 64/3 = 21 triplets
  300bp ORF once per 36kbp per strand
But average protein ~ 1000bp

A Simple ORF finder
start at left end
scan triplet-by-non-overlapping triplet for AUG
then continue scan for STOP
repeat until right end
repeat all starting at offset 1
repeat all starting at offset 2

Scanning for ORFs

* In bacteria, GUG is sometimes a start codon…

Idea #2: Codon Frequency

In random DNA
  Leucine : Alanine : Tryptophan = 6 : 4 : 1
But in real protein, ratios ~ 6.9 : 6.5 : 1
So, coding DNA is not random
Even more: synonym usage is biased (in a species dependant way)
  examples known with 90% AT 3rd base
  Why? E.g. efficiency, histone, enhancer, splice interactions
Recognizing Codon Bias

Assume
Codon usage i.i.d.; abc with freq. f(abc)
\( a_1a_2a_3a_4...a_{3n+2} \) is coding, unknown frame

Calculate
\[ p_1 = \frac{f(a_1a_2a_3)f(a_4a_5a_6)...f(a_{3n-2}a_{3n-1}a_{3n})}{f(a_{3n-2}a_{3n-1}a_{3n})} \]
\[ p_2 = \frac{f(a_2a_3a_4)f(a_5a_6a_7)...f(a_{3n-1}a_{3n}a_{3n+1})}{f(a_{3n-1}a_{3n}a_{3n+1})} \]
\[ p_3 = \frac{f(a_3a_4a_5)f(a_6a_7a_8)...f(a_{3n}a_{3n+1}a_{3n+2})}{f(a_{3n}a_{3n+1}a_{3n+2})} \]
\[ P_i = \frac{p_i}{(p_1+p_2+p_3)} \]

More generally: k-th order Markov model
k=5 or 6 is typical (next lecture)

Codon Usage in \( \Phi x 174 \)