CSE 427
Computational Biology

Genes and Gene Prediction
Gene Finding: Motivation

Sequence data flooding in
What does it mean?
  protein genes, RNA genes, mitochondria, chloroplast, regulation, replication, structure, repeats, transposons, unknown stuff, …

More generally, how do you: learn from complex data in an unknown language, leverage what’s known to help discover what’s not
Protein Coding Nuclear DNA

Focus of these slides
Goal: Automated annotation of new seq 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
Largely done for Human; unlikely for most others
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)
RNA Transcription

(This gene is heavily transcribed, but many are not.)
Translation: mRNA → Protein
Ribosomes

Watson, Gilman, Witkowski, & Zoller, 1992
DNA (thin lines), RNA Pol (Arrow), mRNA with attached Ribosomes (dark circles)

Figure 3-7. Coupled transcription/translation in bacteria is visualized. Oscar Miller and colleagues lysed E. coli cells and immediately collected the cell contents on electron microscope grids. They saw threads of mRNA still associated with DNA (thin lines), and ribosomes—several at a time—were already translating protein along the mRNA. Thus, in bacterial cells, the picture of information recovery and use, at least in broad outline, was complete: mRNA was made on demand; ribosomes recognized the 5’ end of the mRNA, bound, and began protein synthesis even before the mRNA had been completely synthesized. (In this photo, the arrow indicates a presumptive RNA polymerase [the faint disk to the left of the first ribosome]. The DNA thread at the top is being copied into mRNA, but the one at the bottom is not. Both are presumably double stranded.) (Reprinted, with permission, from Miller et al. 1970 [©AAAS].)
## Codons & The Genetic Code

<table>
<thead>
<tr>
<th>First Base</th>
<th>Second Base</th>
<th>Third Base</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U</td>
<td>C</td>
</tr>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td>Met/Start</td>
<td>Thr</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
</tbody>
</table>

- Ala : Alanine
- Arg : Arginine
- Asn : Asparagine
- Asp : Aspartic acid
- Cys : Cysteine
- Gln : Glutamine
- Glu : Glutamic acid
- Gly : Glycine
- His : Histidine
- Ile : Isoleucine
- Leu : Leucine
- Lys : Lysine
- Met : Methionine
- Phe : Phenylalanine
- Pro : Proline
- Ser : Serine
- Thr : Threonine
- Trp : Tryptophane
- Tyr : Tyrosine
- Val : Valine
Idea #1: Find Long ORF’s

Reading frame: which of the 3 possible sequences of triples does the ribosome read?

Open Reading Frame: No internal stop codons

In random DNA

- average ORF $\sim \frac{64}{3} = 21$ triplets
- 300bp ORF once per 36kbp per strand

But average protein $\sim 1000$bp
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
then do it again on the other strand
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 3\(^{rd}\) base

Why? E.g. efficiency, histone, enhancer, splice interactions
Idea #3: Non-Independence

Not only is codon usage biased, but residues (aa or nt) in one position are *not independent* of neighbors.

How to model this? Markov models.
CpG Islands

CpG Islands
More CpG than elsewhere (say, CpG/GpC>50%)
More C & G than elsewhere, too (say, C+G>50%)
Typical length: few 100 to few 1000 bp

Questions
Is a short sequence (say, 200 bp) a CpG island or not?
Given long sequence (say, 10-100kb), find CpG islands?
A sequence $x_1, x_2, \ldots$ of random variables is a \textit{k-th order Markov chain} if, for all $i$, $i^{th}$ value is independent of all but the previous $k$ values:

$$P(x_i \mid x_1, x_2, \ldots, x_{i-1}) = P(x_i \mid x_{i-k}, x_{i-k+1}, \ldots, x_{i-1})$$

\[ k \text{ typically } \ll i-1 \]

Example 1: Uniform random ACGT
Example 2: Weight matrix model
Example 3: ACGT, but $\downarrow \Pr(G \text{ following C})$
A Markov Model (1st order)

States: A,C,G,T
Emissions: corresponding letter
Transitions: \( a_{st} = P(x_i = t \mid x_{i-1} = s) \)
A Markov Model (1st order)

States: A, C, G, T
Emissions: corresponding letter
Transitions: $a_{st} = P(x_i = t | x_{i-1} = s)$
Begin/End states
Pr of emitting sequence $x$

\[
x = x_1 \ x_2 \ldots \ x_n
\]

\[
P(x) = P(x_1, x_2, \ldots, x_n)
\]

\[
= P(x_1) \cdot P(x_2 \mid x_1) \cdots P(x_n \mid x_{n-1}, \ldots, x_1)
\]

\[
= P(x_1) \cdot P(x_2 \mid x_1) \cdots P(x_n \mid x_{n-1})
\]

\[
= P(x_1) \prod_{i=1}^{n-1} a_{x_i, x_{i+1}}
\]

\[
= \prod_{i=0}^{n-1} a_{x_i, x_{i+1}} \quad \text{(with Begin state)}
\]
Training

Max likelihood estimates for transition probabilities are just the frequencies of transitions when emitting the training sequences.

E.g., from 48 CpG islands in 60k bp:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>C</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td>G</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>T</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
<tr>
<td>C</td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
<tr>
<td>G</td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
<tr>
<td>T</td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>

From DEKM
Discrimination/Classification

Log likelihood ratio of CpG model vs background model

\[ S(x) = \log \frac{P(x | \text{model} +)}{P(x | \text{model} -)} = \sum_{i=1}^{L} \log \frac{a_{x_i-1x_i}^+}{a_{x_i-1x_i}^-} = \sum_{i=1}^{L} \beta_{x_i-1x_i} \]

<table>
<thead>
<tr>
<th>( \beta )</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>−0.740</td>
<td>0.419</td>
<td>0.580</td>
<td>−0.803</td>
</tr>
<tr>
<td>C</td>
<td>−0.913</td>
<td>0.302</td>
<td>1.812</td>
<td>−0.685</td>
</tr>
<tr>
<td>G</td>
<td>−0.624</td>
<td>0.461</td>
<td>0.331</td>
<td>−0.730</td>
</tr>
<tr>
<td>T</td>
<td>−1.169</td>
<td>0.573</td>
<td>0.393</td>
<td>−0.679</td>
</tr>
</tbody>
</table>
CpG Island Scores

Figure 3.2 *Histogram of length-normalized scores.*