Lecture 17

• Parsing genomes with HMMs
Genome biology overview

• Genomes undergo two fundamental processes (both involve copying!):
  – Replication
  – Transcription

• Genomic functional information is in the form of *sites*:
  – Short (~2 – ~15 base) sequence segments that bind to an *RNA* or *protein* molecule (the *reader*) to help mediate some function
Genome HMMs

• a genome consists of (functionally important) sites within (nonfunctional) background sequence.

• can define an HMM that reflects this:
  – one state per site position, for each type of site
  – background state
  – appropriate topology (allowed transitions)
  – emission & transition probs

and use it to get Viterbi parse & posterior probs
HMM for *C. elegans* 3’ Splice Sites

**CONSENSUS**

```
  W  W  W  W  T  T  T  t  C  A  G  r  w  w
```

**Emission probabilities**

```
A  0.400  0.429  0.282  0.058  0.008  0.092  0.029  1.000  0.000  0.410  0.293  0.307
C  0.118  0.079  0.081  0.029  0.016  0.135  0.834  0.000  0.000  0.156  0.187  0.225
G  0.072  0.070  0.063  0.018  0.005  0.073  0.001  0.000  1.000  0.310  0.159  0.191
T  0.409  0.422  0.574  0.896  0.971  0.700  0.135  0.000  0.000  0.124  0.361  0.276
```

**3’ ss**

Exon  Intron  3’ ss

```
A  3276  3516  2313  476  67  757  240  8192  0  3359  2401  2514
C   970   648   664   236   129  1109  6830  0  0   1277  1533  1847
G   593   575   516   144   39  595   12   0  8192  2539  1301  1567
T  3353  3453  4699  7336  7957  5731  1110  0  0   1017  2957  2264
```

**‘hidden’ states**

0 → 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 11 → 12
• Complication: sites have orientation (top or bottom strand)
  – e.g. from transcription direction

• One strategy: analyze 2 strands separately
  – problem: resolving conflicts

• Better strategy: expand model to allow sites in both orientations, and run on top strand only
  – double # site states
  – bottom strand states have
    • complementary emission probs
    • reversed allowed transitions
• # params does not change
• size of WDAG increases, but only by factor of ~2
  – no transitions *between* top & bottom strand states, except for background
cf. WDAG for 3-state HMM length $n$ sequence (lecture 13)

weights are emission probabilities $e_k(b_i)$ for $i^{th}$ residue $b_i$
weights are transition probabilities $a_{kl}$

$b_{i-1}$ position $i-1$

$b_i$ position $i$

$b_{i+1}$ position $i+1$
Prokaryotes vs eukaryotes

• Such HMMs are most reliable, & most widely used, for prokaryotic genomes, which usually have
  – high site density, homogeneous background
  – relatively simple spatial relationships among sites
  – often relatively little ‘supporting information’ such as
    • protein binding & transcript data
    • closely related genomes
• eukaryotic genomes are less suitable:
  – low site density, heterogeneous background
  – complex site spatial relationships (not well captured by Markov transition model)
  – often much supporting info
    • similar genomes to transfer annotations from;
    • protein binding / RNASeq & other experimental data
    • in principle, some of this could be incorporated into HMM
      – (expanded symbol alphabet)
Prokaryote genomes

• typically a few MB in size
• up to ~80% protein coding
• Typical CDS size ~1 KB
• introns & overlapping CDSs rare
• range of GC contents
ORF analysis

• Translate genome in all 6 reading frames
• In each, find ‘open reading frames’ starting with ATG (or NTG), ending in stop
• Sort ORFs by (decreasing) length
• Work through sorted list, discarding any ORF that
  – overlaps a longer one, or
  – is ‘too short’
• Problems:
  – short CDSs are missed
  – CDSs often have long overlapping fake ORFs on opposite strand
  – poor performance on GC-rich genomes (many long fake ORFs)
• Additional information that is present in real coding sequence (but ignored in ORF analysis) – *cf. lecture 3*
  – amino acid usage
  – synonymous codon bias
• Use this, in a probability model!
The Genetic Code
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Obs/Exp</th>
<th>1&lt;sup&gt;st&lt;/sup&gt; codon base</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt; codon base</th>
<th>3&lt;sup&gt;rd&lt;/sup&gt; codon base</th>
<th># codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1.92</td>
<td>G</td>
<td>A</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>1.80</td>
<td>A</td>
<td>A</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1.62</td>
<td>G</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>M</td>
<td>1.46</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>1.37</td>
<td>A</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>1.25</td>
<td>T</td>
<td>T</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>Q</td>
<td>1.22</td>
<td>C</td>
<td>A</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>I</td>
<td>1.16</td>
<td>A</td>
<td>T</td>
<td>Not G</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>1.14</td>
<td>G</td>
<td>C</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>1.05</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>V</td>
<td>.99</td>
<td>G</td>
<td>T</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>Y</td>
<td>.98</td>
<td>T</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>.95</td>
<td>C(T)</td>
<td>T</td>
<td>N</td>
<td>6</td>
</tr>
<tr>
<td>T</td>
<td>.88</td>
<td>A</td>
<td>C</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>W</td>
<td>.79</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>.74</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>S</td>
<td>.73</td>
<td>T(A)</td>
<td>C(G)</td>
<td>N</td>
<td>6</td>
</tr>
<tr>
<td>H</td>
<td>.67</td>
<td>C</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>.53</td>
<td>C(A)</td>
<td>G</td>
<td>N</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>.52</td>
<td>T</td>
<td>G</td>
<td>Y</td>
<td>2</td>
</tr>
</tbody>
</table>
Synonymous codon bias

• In most organisms, the codons for an amino acid are not used with equal frequency
• For many organisms this may reflect differences in translational efficiency & accuracy
  – more highly expressed genes have stronger biases
• For mammals codon usage mainly reflects the GC content of the region in which the gene is found
  – GC content variation probably reflects GC-biased gene conversion
**Figure 34** The human genetic code and associated tRNA genes. For each of the 64 codons, we show: the corresponding amino acid; the observed frequency of the codon per 10,000 codons; the codon; predicted wobble pairing to a tRNA anticodon (black lines); an unmodified tRNA anticodon sequence; and the number of tRNA genes found with this anticodon. For example, phenylalanine is encoded by UUU or UUC; UUC is seen more frequently, 203 to 171 occurrences per 10,000 total codons; both codons are expected to be decoded by a single tRNA anticodon type, GAA, using a G/U wobble; and there are 14 tRNA genes found with this anticodon. The modified anticodon sequence in the mature tRNA is not shown, even where post-transcriptional modifications can be confidently predicted (for example, when an A is used to decode a U/C third position, the A is almost certainly an inosine in the mature tRNA). The Figure also does not show the number of distinct tRNA species (such as distinct sequence families) for each anticodon; often there is more than one species for each anticodon.
Prokaryote HMMs

• Main types of sites:
  – Codon sites
  – Translation start sites (Shine-Dalgarno)
  – Promoter elements
    • Transcription factor binding sites
  – (RNA genes / RNA folding sites)
  – (replication origin)
• Simple 7-state prokaryote genome model:
  – 1 state for intergenic regions
  – 3 states for codon positions in top-strand genes
  – 3 for codon positions in bottom-strand genes
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Obs/Exp</th>
<th>1(^{st}) codon base</th>
<th>2(^{nd}) codon base</th>
<th>3(^{rd}) codon base</th>
<th># codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1.92</td>
<td>G</td>
<td>A</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>1.80</td>
<td>A</td>
<td>A</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1.62</td>
<td>G</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>M</td>
<td>1.46</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>1.37</td>
<td>A</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>1.25</td>
<td>T</td>
<td>T</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>Q</td>
<td>1.22</td>
<td>C</td>
<td>A</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>I</td>
<td>1.16</td>
<td>A</td>
<td>T</td>
<td>Not G</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>1.14</td>
<td>G</td>
<td>C</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>1.05</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>V</td>
<td>.99</td>
<td>G</td>
<td>T</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>Y</td>
<td>.98</td>
<td>T</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>.95</td>
<td>C(T)</td>
<td>T</td>
<td>N</td>
<td>6</td>
</tr>
<tr>
<td>T</td>
<td>.88</td>
<td>A</td>
<td>C</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>W</td>
<td>.79</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>.74</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>S</td>
<td>.73</td>
<td>T(A)</td>
<td>C(G)</td>
<td>N</td>
<td>6</td>
</tr>
<tr>
<td>H</td>
<td>.67</td>
<td>C</td>
<td>A</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>.53</td>
<td>C(A)</td>
<td>G</td>
<td>N</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>.52</td>
<td>T</td>
<td>G</td>
<td>Y</td>
<td>2</td>
</tr>
</tbody>
</table>
Average codon biases (*lecture 3*)

- At codon position 1,
  - purines (A and G) predominate among over-represented amino acids,
  - pyrimidines (C and T) among under-represented amino acids.
- At codon position 2,
  - A and T predominate among over-represented amino acids,
  - C and G among under-represented amino acids.
- Hypotheses to explain *RWR* codon preference:
  - (Neutralist) Vestige of ancestral code? (Shepherd)
  - (Selectionist) More efficiently translated?
• These biases are somewhat subtle – but strong enough to (often) distinguish
  – coding sequences (of reasonable length)
from
  – background sequence
7-state model for prokaryote genomes

- intergenic
- first codon position – top strand coding sequence
- second codon position – top strand coding sequence
- third codon position – top strand coding sequence
- first codon position – bottom strand coding sequence
- second codon position – bottom strand coding sequence
- third codon position – bottom strand coding sequence

a (very short!) ‘bottom-strand’ gene, in a different region of the genome:
• N.B. the emitted symbols are always \textit{top strand} nucleotides!
A better HMM!

- Amino acid-specific codon blocks
  - Not really ‘sites’ as previously defined – may have more than one tRNA reader
  - Split the three 6-codon amino acids into 2 sites
    \((4 + 2)\)
    - E.g. Leu: CTN and TTR ‘sites’
    - A single YTN site would also emit Phe codons
  - The other 17 aas are each 1 site
- ‘Start’ codon: NTG
  - Part of Shine-Dalgarno
- 2 Stop codon sites: TAR, TGA
- Total codon sites: \(17 + 3 \times 2 + 1 + 2 = 26\)
# The Genetic Code

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
<td>Stop</td>
<td>Stop</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
<td>Stop</td>
<td>Trp</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
<td>Arg</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Met</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
</tr>
</tbody>
</table>
• Total codon states
  = 26 sites × 2 strands × 3 pos = 156
• Transitions within & between codons are the obvious ones
  – Unless one wishes to allow for frameshift sequencing errors!
• Also, states for promoter element sites
  – TF binding sites
• Ignore RNA genes
  – (identify by sequence similarity)
• Ignore replication origins
  – Often can identify after HMM analysis, by orientation biases
• Need more than one background state, to allow memory of where one is in a gene, and strand
• May need additional backgd states if promoter element *order* is important

• Role of ‘memory’ is to *reduce* impact of biologically implausible paths
  – Model may still work without these complications – but with reduced power

• Reasonable to constrain all backgd states to have *same emission probs*
• Use Viterbi or Baum-Welch training
  – (with appropriate top vs bottom constraints etc) to find
  – Codon biases, aa freqs
  – Promoter elements
    • Include sites of size ~6, random initial emission probs
  – Shine-Dalgarno sequence preferences
Complications in Eukaryotes

- 5’ & 3’ splice sites
- poly A sites
- introns
  - Must retain memory of where codon is interrupted!
- 5’ & 3’UTR
- G+C variation
• Not difficult to set up an HMM with states corresponding to the above; *but* complex site spatial relationships are not well captured by Markov transition model:
  – Intron size constraints
  – Enhancers (possibly intronic!)

• Also:
  – alternative splicing
  – alternative promoters
  – overlapping sites

imply any single parse is incomplete