Lecture 3: Probability Models for Sequences

• Probability models
  – Equal frequency & independence assumptions

• ‘Background’ models
  – Failure of equal frequency assumption
    • Neutralist vs selectionist interpretations
  – Failure of independence assumption
    • Markov models

• Assessing significance of sequence patterns
  – Simulations
Probability Models of Sequences

• Sample questions when interpreting genomes:
  – Is this sequence a splice site?
  – Is this sequence part of the coding region of a gene?
  – Are these two sequences evolutionarily related?
  – Does this sequence show evidence of selection?

• Computational analysis can’t answer:
  – only generates hypotheses
    which must ultimately be tested by experiment.

• But hypotheses should
  – have some reasonable chance of being correct, and
  – carry indication of reliability.
• We use *probability models* of sequences to address such questions.

• Not the only approach, but usually the most powerful, because
  – seqs are products of evolutionary process which is *itself* probabilistic
  – want to detect biological “signal” against “noise” of background sequence or mutations
Models: simplicity vs complexity

• “All models are wrong; some models are useful.” – George Box

• “What is simple is always wrong. What is not is unusable.” – Paul Valery

• “Everything should be made as simple as possible, but not simpler.” – Albert Einstein (?)

• Some disadvantages of complexity:
  – computational challenge
  – (lack of) interpretability
  – overfitting
Basic Probability Theory Concepts

• A *sample space* $S$ is set of all possible outcomes of a conceptual, repeatable experiment.
  – $|S| < \infty$ in most of our examples.
  – e.g. $S =$ all possible sequences of a given length.

• Elements of $S$ are called *sample points*.
  – e.g. a particular seq = outcome of “experiment” of extracting seq of specified type from a genome.

• A *probability distribution* $P$ on $S$ assigns non-neg real number $P(s)$ to each $s \in S$, such that
  $$\sum_{s \in S} P(s) = 1$$
  (So $0 \leq P(s) \leq 1 \ \forall s$ )
  – Intuitively, $P(s) =$ fraction of times one would get $s$ as result of the expt, if repeated many times.
• A *probability space* \((S, P)\) is a sample space \(S\) with a prob dist’n \(P\) on \(S\).

• Prob dist’n on \(S\) is sometimes called a *probability model* for \(S\), particularly if several dist’ns are being considered.
  
  – Write models as \(M_1, M_2\), probabilities as \(P(s \mid M_1)\), \(P(s \mid M_2)\).
  
  – e.g.
    
    • \(M_1\) = prob dist’n for splice site seqs,
    • \(M_2\) = prob dist’n for “background” (arbitrary genomic) seqs.
• An event \( E \) is a criterion that is true or false for each \( s \in S \).
  – defines a subset of \( S \) (sometimes also denoted \( E \)).
  – \( P(E) \) is defined to be \( \sum_{s \mid E \text{ is true}} P(s) \).

• Events \( E_1, E_2, \ldots, E_n \) are \textit{mutually exclusive} if no two of them are true for the same point;
  – then \( P(E_1 \text{ or } E_2 \text{ or } \ldots \text{ or } E_n) = \sum_{1 \leq i \leq n} P(E_i) \).

• If \( E_1, E_2, \ldots, E_n \) are also \textit{exhaustive}, i.e. every \( s \) in \( S \) satisfies \( E_i \) for some \( i \), then \( \sum_{1 \leq i \leq n} P(E_i) = 1 \).
• For events $E$ and $H$, the \textit{conditional probability} of $E$ given $H$, is

\[ P(E \mid H) \equiv \frac{P(E \text{ and } H)}{P(H)} \]

(= prob that both $E$ and $H$ are true, given $H$ is true)
– undefined if $P(H) = 0$.

• $E$ and $H$ are (\textit{statistically}) \textit{independent} if

\[ P(E) = P(E \mid H) \]

(i.e. prob. $E$ is true doesn’t depend on whether $H$ is true);
or equivalently

\[ P(E \text{ and } H) = P(E)P(H). \]
Probabilities on Sequences

• Let $S =$ space of DNA or protein sequences of length $n$. Possible assumptions for assigning probabilities to $S$:
  
  – *Equal frequency assumption*: All residues are equally probable at any position;
    
    • $P(E_r^{(i)}) = P(E_q^{(i)})$ for any two residues $r$ and $q$,
      
      – where $E_r^{(i)}$ means residue $r$ occurs at position $i$, then
    
    • Since for fixed $i$ the $E_r^{(i)}$ are mutually exclusive and exhaustive,
      
      $$P(E_r^{(i)}) = 1 / |A|$$
    
      where $A =$ residue alphabet
    
      $$P(E_r^{(i)}) = 1/20 \text{ for proteins, 1/4 for DNA).}$$
  
  – *Independence assumption*: whether or not a residue occurs at a given position is independent of residues at other positions.
Given above assumptions, the probability of the sequence 
\( s = ACGCG \)
(in the space \( S \) of all length 5 sequences) is calculated by considering 5 events:

- Event 1 is that first nuc is A. Probability = .25.
- Event 2 is that 2\(^{\text{nd}}\) nuc is C. Probability = .25.
- Event 3 is that 3\(^{\text{rd}}\) nuc is G. Probability = .25.
- Event 4 is that 4\(^{\text{th}}\) nuc is C. Probability = .25.
- Event 5 is that 5\(^{\text{th}}\) nuc is G. Probability = .25.

By independence assumption, prob of all 5 events occurring is the product \((.25)^5 = 1/1024\).

Since \( s \) is the only sequence satisfying all 5 conditions, \( P(s) = 1/1024 \).
• More generally, under equal freq and indep assumptions,
  \[
  \text{prob of nuc sequence of length } n = .25^n, \\
  \text{prob of protein sequence of length } n = .05^n \\
  \text{in the space } S \text{ of length } n \text{ sequences.}
  \]
‘Background’ models

- ‘Average’ model for genome; contrasted with ‘foreground’ models (for sites & other regions of interest)

- Whole genome vs non-site
Genome background models: Failure of equal frequency assumption

• For most organisms, the genomic nucleotide composition is significantly different from \( .25 \) for each nucleotide, e.g.:
  – \textit{H. influenza} .31 A, .19 C, .19 G, .31 T
  – \textit{P. aeruginosa} .17 A, .33 C, .33 G, .17 T
  – \textit{M. janaschii} .34 A, .16 C, .16 G, .34 T
  – \textit{S. cerevisiae} .31 A, .19 C, .19 G, .31 T
  – \textit{C. elegans} .32 A, .18 C, .18 G, .32 T
  – \textit{H. sapiens} .29 A, .21 C, .21 G, .29 T
• Note approximate symmetry: \( A \approx T, \ C \approx G, \)
  – even though we’re counting nucs on just one strand.
  – Expect *exact* equality when counting both strands
• Explanation:
  – Although individual biological features may have non-
symmetric composition (local *asymmetry*),
  – usually features are distributed approx *randomly* w.r.t.
    strand,
  – so local asymmetries *cancel*, yielding overall
    symmetry.
Thymine (T)  Adenine (A)

Cytosine (C)  Guanine (G)
General Hypotheses Regarding Unequal Frequency

• **Neutralist** hypothesis: *mutation bias*
  – e.g. due to nucleotide pool composition

• **Selectionist** hypothesis: *selection*
  – selection on (many) particular nucleotides
  – selection on mutational bias mechanisms
  – ...
Genome background models: 
Failure of independence assumption

Nucleotide Freqs (C. elegans chr. 1):
A 4575132 (.321); C 2559048 (.179); G 2555862 (.179); T 4582688 (.321)

dinucleotide frequencies (5’ nuc to left, 3’ nuc at top – e.g. obs freq of ApC is .047): (Note “symmetry”!)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>0.135</td>
<td>0.047</td>
<td>0.051</td>
<td>0.088</td>
<td></td>
<td>0.103</td>
<td>0.057</td>
<td>0.057</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.061</td>
<td>0.035</td>
<td>0.033</td>
<td>0.051</td>
<td></td>
<td>0.057</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0.063</td>
<td>0.034</td>
<td>0.034</td>
<td>0.047</td>
<td></td>
<td>0.057</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.061</td>
<td>0.064</td>
<td>0.061</td>
<td>0.135</td>
<td></td>
<td>0.103</td>
<td>0.058</td>
<td>0.057</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>Observed / Expected</td>
<td>1.314</td>
<td>0.818</td>
<td>0.885</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1.055</td>
<td>1.075</td>
<td>1.031</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.106</td>
<td>1.062</td>
<td>1.074</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.597</td>
<td>1.105</td>
<td>1.056</td>
</tr>
</tbody>
</table>
Dinucleotide frequencies

• Underrepresentation of TpA: found in nearly all genomes;
  – reason unknown:
    • neutral (mutation patterns)?
    • selection?


• Unlike mammalian genomes, no underrepresentation of CpG in C. elegans
  – CpG not methylated in C. elegans (or most other non-vertebrates).
## Dinucleotide Freqs – *H. sapiens*Chr.21

### Nucleotide Freqs:

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Frequency</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10032226</td>
<td>0.297</td>
</tr>
<tr>
<td>T</td>
<td>9962530</td>
<td>0.295</td>
</tr>
<tr>
<td>G</td>
<td>6908202</td>
<td>0.204</td>
</tr>
<tr>
<td>C</td>
<td>6921020</td>
<td>0.205</td>
</tr>
</tbody>
</table>

Entrop: 1.976 bits

### Observed Dinuc Freqs

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.099</td>
<td>0.051</td>
<td>0.069</td>
<td>0.078</td>
</tr>
<tr>
<td>C</td>
<td>0.073</td>
<td>0.052</td>
<td>0.011</td>
<td>0.069</td>
</tr>
<tr>
<td>G</td>
<td>0.059</td>
<td>0.043</td>
<td>0.052</td>
<td>0.050</td>
</tr>
<tr>
<td>T</td>
<td>0.066</td>
<td>0.059</td>
<td>0.072</td>
<td>0.098</td>
</tr>
</tbody>
</table>

### Expected (under independence)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.088</td>
<td>0.061</td>
<td>0.061</td>
<td>0.087</td>
</tr>
<tr>
<td>C</td>
<td>0.061</td>
<td>0.042</td>
<td>0.042</td>
<td>0.060</td>
</tr>
<tr>
<td>G</td>
<td>0.061</td>
<td>0.042</td>
<td>0.042</td>
<td>0.060</td>
</tr>
<tr>
<td>T</td>
<td>0.087</td>
<td>0.060</td>
<td>0.060</td>
<td>0.087</td>
</tr>
</tbody>
</table>

### Observed / Expected

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.124</td>
<td>0.839</td>
<td>1.139</td>
<td>0.891</td>
</tr>
<tr>
<td>C</td>
<td>1.204</td>
<td>1.243</td>
<td>0.260</td>
<td>1.139</td>
</tr>
<tr>
<td>G</td>
<td>0.974</td>
<td>1.025</td>
<td>1.245</td>
<td>0.839</td>
</tr>
<tr>
<td>T</td>
<td>0.752</td>
<td>0.976</td>
<td>1.204</td>
<td>1.125</td>
</tr>
</tbody>
</table>
5-methylcytosine ($^{mC}$): the ‘5th base’

- Comprises ~1-6% of mammalian & plant genomes

- Methylation does not affect base-pairing:
• But it *does* affect
  – protein binding, e.g. Sp1, EGR1, CTCF
    ⇒ effects on gene expression, development, cellular differentiation, transposon suppression, embryogenesis, imprinting, X-inactivation, chromatin structure, tumorigenesis
    mouse methyltransferase knockouts are embryonic lethal
  – mutation rate: \(^{\text{m}C}\) is a mutation ‘hotspot’:

• In mammals methylated C’s (nearly) always occur as part of a CpG dinucleotide:
\[ 5' \text mC} \ G \ 3' \]
\[ 3' \ G \text mC} \ 5' \]
• But some Cs not in CpGs are methylated, in some cell types
• as many as 20-30% of all new single-base mutations in mammalian genomes may be at CpGs, judging from
  – analysis of disease-causing mutations,
  – comparison of closely related species
  – polymorphism data
• As a result, CpGs are substantially underrepresented in mammalian DNA:
  – expected frequency \(0.21 \times 0.21 = 0.044\) (in mammalian genomes, G+C freq is about 0.42, A+T about 0.58)
  – only see about 1/5 that many.
• Conversely, TpGs and CpAs are overrepresented
Dinucleotide Freqs – *H. sapiens* Chr.22

Nucleotide Freqs:

A  8745910  0.261; T  8720493  0.261
G  7999585  0.239; C  7997931  0.239

Entropy: 1.999 bits

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Dinuc Freqs</td>
<td>0.077</td>
<td>0.051</td>
<td>0.075</td>
<td>0.058</td>
</tr>
<tr>
<td>Expected (under independence)</td>
<td>0.068</td>
<td>0.062</td>
<td>0.062</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Observed / Expected</td>
<td>1.125</td>
<td>0.817</td>
<td>1.205</td>
<td>0.855</td>
</tr>
</tbody>
</table>

24
Genome background models: Failure of independence assumption

Nucleotide Freqs (C. elegans chr. 1):
A 4575132 (.321); C 2559048 (.179); G 2555862 (.179); T 4582688 (.321)

dinucleotide frequencies (5’ nuc to left, 3’ nuc at top – e.g. obs freq of ApC is .047): (Note “symmetry”!)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.135</td>
<td>0.047</td>
<td>0.051</td>
<td>0.088</td>
<td></td>
<td>0.103</td>
<td>0.057</td>
<td>0.057</td>
<td>0.103</td>
</tr>
<tr>
<td>C</td>
<td>0.061</td>
<td>0.035</td>
<td>0.033</td>
<td>0.051</td>
<td></td>
<td>0.057</td>
<td>0.032</td>
<td>0.032</td>
<td>0.058</td>
</tr>
<tr>
<td>G</td>
<td>0.063</td>
<td>0.034</td>
<td>0.034</td>
<td>0.047</td>
<td></td>
<td>0.057</td>
<td>0.032</td>
<td>0.032</td>
<td>0.057</td>
</tr>
<tr>
<td>T</td>
<td>0.061</td>
<td>0.064</td>
<td>0.061</td>
<td>0.135</td>
<td></td>
<td>0.103</td>
<td>0.058</td>
<td>0.057</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Observed / Expected (under independence)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.314</td>
<td>0.818</td>
<td>0.885</td>
<td>0.853</td>
<td></td>
<td>1.055</td>
<td>1.075</td>
<td>1.031</td>
<td>0.886</td>
</tr>
<tr>
<td>C</td>
<td>1.106</td>
<td>1.062</td>
<td>1.074</td>
<td>0.818</td>
<td></td>
<td>0.597</td>
<td>1.105</td>
<td>1.056</td>
<td>1.313</td>
</tr>
</tbody>
</table>
Conditional probability (in *C. elegans*) of a given nucleotide (top) occurring, given the preceding nucleotide (left)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.421</td>
<td>0.147</td>
<td>0.159</td>
<td>0.274</td>
</tr>
<tr>
<td>C</td>
<td>0.338</td>
<td>0.193</td>
<td>0.185</td>
<td>0.284</td>
</tr>
<tr>
<td>G</td>
<td>0.355</td>
<td>0.190</td>
<td>0.192</td>
<td>0.263</td>
</tr>
<tr>
<td>T</td>
<td>0.191</td>
<td>0.198</td>
<td>0.189</td>
<td>0.421</td>
</tr>
</tbody>
</table>
Markov models

• Such conditional probabilities can be used to define a \textit{first-order Markov model} (or \textit{Markov chain model}) for background sequence probabilities:

\[
P(s_1 s_2 s_3 \cdots s_n) = P(s_1) P(s_2 | s_1) P(s_3 | s_2) \cdots P(s_n | s_{n-1})
\]
Similarly, one can define an a **order-\(k\) Markov model** in which the probability of \(s_i\) is conditional on \(s_{i-k} \cdots s_{i-2} s_{i-1}\) (i.e. the \(k\) preceding residues)

- Note that the required number of parameters is exponential in \(k\)
- **independence model** = **order-0 Markov model**
Assessing significance of sequence patterns

- Problem: Is a particular sequence pattern, e.g.
  - a match between genomes, or
  - a region of a particular composition (e.g. GC-rich) likely to be “biologically significant”, e.g. indicating
    - an evolutionary relationship, or
    - a functional feature
Assessing significance of sequence patterns

• Idea:
  – specify a scoring system for patterns of the given type
  – find the score *distribution* in *negative controls*
    • i.e. sequences not expected to contain the biological feature
  – Scores occurring in real sequence, but not in negative controls, *may* have biological significance

• Caveats:
  – Control may be inadequate in quantity / quality
  – ‘Biologically significant’ ≠ interpretable
    • can’t infer function!!
‘Negative control’ sequences

1. real biological ‘background’ sequences known not to have the feature in question
   – ideal if available – but usually hard to find!

2. simulated sequences
   – requires probability model retaining some features of real sequences
   – Quantity: In general, want multiple such sequences
   – Quality: is the model complex enough?
Theoretical score distributions

- For simple probability models, can sometimes avoid simulations by finding a *theoretical* probability distribution
  - approximate, e.g. Karlin-Altschul for BLAST hits
  - or exact
  for the scores.

- Alternatively, can fit a theoretical distribution to the observed scores for simulated data
  - Avoids need for large number of simulations
Homework 2

• Purpose: Assess significance of HW 1 genomic matches

• Simulate negative controls using two different background sequence models:
  – Order 0 Markov
  – Order 1 Markov

• Then find matches (using HW 1 suffix array method) between real sequence and these control sequences
  – Ideally should do lots of simulations!!