Probabilistic sequence modeling
frequency and profiles

Gene and genes
• Genome: an organism's genetic material
• Gene: discrete units of hereditary information located on the chromosomes and consisting of DNA

Gene prediction: computational challenge
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgcgg
tatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
What’s a gene

• What is a gene, post-ENCODE?
  - Gerstein et al. Genome Res. 2007 17: 669-681
  - ENCODE consortium: characterization of 1% of the human genome by experimental and computational techniques
• Definitions:
  - Definition 1970s–1980s: Gene as open reading frame (ORF) sequence pattern
  - Definition 1990s–2000s: Annotated genomic entity, enumerated in the databanks
  - The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products

Post-ENCODE definition
The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products

Can we still do gene prediction?

• Prokaryotic genes
  - promoter
  - gene
  - start
  - stop
  - terminator

• Eukaryotic genes
  - promoter
  - exon
  - intron
  - intron
  - exon
  - intron
  - exon
  - Open reading frame (ORF)

A simple model for gene prediction: frequency-based DNA modeling

• DNA is a double stranded molecule
  - G-C pair → strong
  - A-T pair → weak

• Coding regions often have higher GC content than non-coding regions

Frequency-based DNA modeling of coding vs. non-coding regions

• To predict coding regions in an organism (e.g. human), collect a set of known coding and non-coding DNA sequences from this organism (training set)

• Compute the frequency distribution of GC pairs in coding and non-coding regions, respectively: f(GC%c), f(GC%nc)

An example from zygomycete Phycomyces blakesleeanus

Table 1. GC content of Phycomyces DNA.

<table>
<thead>
<tr>
<th>DNA type</th>
<th>Sample size</th>
<th>Sample length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding DNA</td>
<td>48 ± 0.6</td>
<td>56</td>
</tr>
<tr>
<td>Total non-coding DNA</td>
<td>30 ± 1.0</td>
<td>49</td>
</tr>
<tr>
<td>Interns</td>
<td>29 ± 1.5</td>
<td>28</td>
</tr>
<tr>
<td>5’-end</td>
<td>34 ± 2.1</td>
<td>10</td>
</tr>
<tr>
<td>3’-end</td>
<td>30 ± 1.1</td>
<td>11</td>
</tr>
</tbody>
</table>

f(GC%)
Simple models for gene prediction:

- Two models: coding vs. non-coding
- Given a DNA sequence, its likelihood of being a coding sequence, based on its GC content (GC%):
  \[ P(c) = \frac{P(GC%) | P(c)}{P(GC%) | P(nc)} = \frac{P(GC%) | P(c) + P(GC%) | P(nc)}{P(GC%) | P(nc)} \]
  \[ P(nc|GC%) = \sum_{i} f(c) \text{ if } GC\% \]
  \[ P(nc|GC%) = \sum_{i} f(nc) \text{ if } GC\% \]

Sliding window approach

A more complicated model: codon usages

- In 1961 Sydney Brenner and Francis Crick discovered frame shift mutations
- Systematically deleted nucleotides from DNA
  - Single and double deletions dramatically altered protein product
  - Effects of triple deletions were minor
- Conclusion: every triplet of nucleotides, each codon, codes for exactly one amino acid in a protein

Translating nucleotides into amino acids

- Codon: 3 consecutive nucleotides
- \(4^3 = 64\) possible codons
- Genetic code is degenerative and redundant
  - Includes start and stop codons
  - An amino acid may be coded by more than one codon (codon degeneracy)
Genetic code and stop codons

UAA, UAG and UGA correspond to 3 Stop codons that (together with Start codon ATG) delineate Open Reading Frames

Testing reading frames

- Create a 64-element hash table and count the frequencies of codons in a reading frame;
- Amino acids typically have more than one codon, but in nature certain codons are more in use;
- Uneven use of the codons may characterize a coding region;

Codon usage in Mouse genome

<table>
<thead>
<tr>
<th>AA codon /1000 freq</th>
<th>AA codon /1000 freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser TCA 11.44 0.14</td>
<td>Ser TCT 15.70 0.19</td>
</tr>
<tr>
<td>Ser TCT 15.70 0.19</td>
<td>Ser TCC 17.92 0.22</td>
</tr>
<tr>
<td>Ser AGT 12.25 0.15</td>
<td>Ser AGC 15.60 0.23</td>
</tr>
<tr>
<td>Ser AGC 19.54 0.24</td>
<td>Ala GCG 15.80 0.23</td>
</tr>
<tr>
<td>Pro CCG 6.33 0.11</td>
<td>Pro CCA 15.10 0.28</td>
</tr>
<tr>
<td>Pro CCT 18.31 0.30</td>
<td>Pro CCA 18.42 0.31</td>
</tr>
</tbody>
</table>

Using codon frequency to find correct reading frame

Consider sequence $x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...$
where $x_i$ is a nucleotide

Let $P_1 = P(x_1 x_2 x_3 | x_4 x_5 x_6 ...)$
$P_2 = P(x_2 x_3 x_4 | x_5 x_6 x_7 ...)$
$P_3 = P(x_3 x_4 x_5 | x_6 x_7 x_8 ...)$

then probability that ith reading frame is the coding frame is:

$$P_i = P_1 + P_2 + P_3$$
Adding the background model: gene finding

- In the previous model, we assume at least one reading frame is the codon sequence → testing reading frames, not gene finding
- Adding a background model
  - \( p_0 = p_1 p_2 p_3 p_4 \ldots \)
  - Based on the nucleotide frequency in the non-coding sequence
  - \( P_r p_i / (p_0 + p_1 + p_2 + p_3) \)
- In practice, this model should be extended to all six reading frames.

Protein secondary structure prediction

- Given a protein sequence, secondary structure prediction aims at predicting the state of each amino acid as being either H (helix), E (extended=strand), or O (other).
- The quality of secondary structure prediction is measured with a “3-state accuracy” score, or \( Q_3 \). \( Q_3 \) is the percent of residues that match “reality” (X-ray structure).

Chou and Fasman: a frequency model

- \( P(\alpha|S)=\prod p(\alpha|S) = \prod p(\alpha|f(s)) \)

  - \( p(\alpha|f(s)) = p(f(s)|\alpha)p(f(s)) \)

  - Similarly for \( \beta \) and turn structures
### Profile model

- The frequency model does not consider the **order** of the training sequences
  - Permuting the training sequences will not change the model
- In some cases, the order is of important biological meaning, e.g. sequence motifs
- Profile model fully constrains the order of the training sequences

### A DNA profile (matrix)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATAA</td>
<td>T</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TATAA</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TATAA</td>
<td>A</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>TATAA</td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Sparse data → pseudo-counts

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATAA</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TATAA</td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TATAA</td>
<td>A</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>TATAA</td>
<td>G</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Testing a motif

- Equivalent to computing the significance of a sequence motif

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATAA</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TATAA</td>
<td>G</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>TATAA</td>
<td>A</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>TATAA</td>
<td>G</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

### Model comparison: relative entropy

- $H(x) = \sum \sum P(x) \log P(x)/P_0(x)$ = $\sum \sum \frac{P(x)}{P_0(x)} \log \frac{P(x)}{P_0(x)}$
  - $b$ → the random background distribution
  - $N$ sequences of length $L$
  - $K → 4$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATAA</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TATAA</td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TATAA</td>
<td>A</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>TATAA</td>
<td>G</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Probability distribution

- What is the probability $P(H|B)$ of getting a matrix with a relative entropy $H$ from the background model $B = \{b_i\}$?
- $p(h)$ → the probability distribution of relative entropy score for the frequency of a single column (can be pre-calculated)
- $P(H) = \sum_{p(x_1), p(x_2), ..., p(x_L)} p(x_1) p(x_2) ... p(x_L)$
  - convolution of function $p(s)$, can be calculated only approximately by Fast Fourier Transformation (FFT)
Searching profiles: inference

- Give a sequence $S$ of length $L$, compute the likelihood ratio of being generated from a profile vs. from background model:

$$R(S|P) = \frac{1}{b} \frac{\prod_{i=1}^{n} (x_i/N)}{b_i}$$

- Searching motifs in a sequence: sliding window approach

Motif finding is difficult

The motif finding problem

- If starting positions $S = (s_1, s_2, \ldots, s_t)$ are given, finding consensus is easy because we can simply construct (and evaluate) the profile to find the motif.
- But... the starting positions $S$ are usually not given. How can we find the “best” profile matrix?
  - Gibbs sampling
  - Expectation-Maximization (EM) algorithm

Finding a motif

- Given a set of DNA sequences:

  - Find the motif in each of the individual sequences

The motif finding problem

- Frequency and profile are two basic models for sequence analysis
- They represent two extreme models in terms of incorporating order information in the sequences
- Model selection should be based on biological ideas