Probabilistic sequence modeling
frequency and profiles

Genome 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

```
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaa
tgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgcgg
```

Gene prediction: computational challenge

```
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaa
tgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgc
```

Gene prediction: computational challenge

```
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaa
tgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
```
A simple model for gene prediction: frequency-based DNA modeling

- DNA is a double strand molecule
  - G-C pair \( \rightarrow \) strong
  - A-T pair \( \rightarrow \) 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: \( f(GC\%|c) \), \( f(GC\%|nc) \)

### A example from zygomycete Phycomyces blakesleeanus

Table 1. GC content of Phycomyces DNA.

<table>
<thead>
<tr>
<th>DNA type</th>
<th>GC</th>
<th>Sample size</th>
<th>Sample length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding DNA</td>
<td>48 ± 6.6</td>
<td>56</td>
<td>29,862</td>
</tr>
<tr>
<td>Total non-coding DNA</td>
<td>30 ± 1.0</td>
<td>49</td>
<td>13,088</td>
</tr>
<tr>
<td>Introns</td>
<td>29 ± 1.5</td>
<td>28</td>
<td>2877</td>
</tr>
<tr>
<td>5’-end</td>
<td>34 ± 3.1</td>
<td>10</td>
<td>6182</td>
</tr>
<tr>
<td>3’-end</td>
<td>30 ± 1.1</td>
<td>11</td>
<td>4069</td>
</tr>
</tbody>
</table>

Model comparison

- Two models: coding vs. non-coding
- Given a DNA sequence, its likelihood of being a coding sequence, based on its GC content (GC%)

\[
P(c|GC%) = \frac{P(GC%)|c)}{P(GC%)|nc)} = \frac{f(GC\%|c)}{f(GC\%|nc)}
\]

\[
P(nc|GC%) = \frac{P(GC%)|nc)}{P(GC%)|c)} = \frac{f(GC\%|nc)}{f(GC\%|c)}
\]

Assume \( P(c) = P(nc) \),

\[
R = \log \frac{P(GC\%|c)}{P(GC\%|nc)}
\]

### Sliding window approach

- Sliding window approach
Protein

RNA

DNA

transcription

transcription

translation

translation

Protein

PEPTIDE

A more complicated model: codon usages

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)

Codons

- In 1961 Sydney Brenner and Francis Crick discovered **frameshift 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

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;

Six frames in a DNA sequence

- stop codons – TAA, TAG, TGA
- start codons - ATG
Codon usage in Human genome

<table>
<thead>
<tr>
<th>AA codon</th>
<th>% frequency</th>
<th>frac</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser TCG</td>
<td>4.31</td>
<td>0.05</td>
</tr>
<tr>
<td>Ser TCA</td>
<td>11.44</td>
<td>0.14</td>
</tr>
<tr>
<td>Ser TCT</td>
<td>15.70</td>
<td>0.19</td>
</tr>
<tr>
<td>Ser TCC</td>
<td>17.92</td>
<td>0.22</td>
</tr>
<tr>
<td>Ser AGT</td>
<td>12.25</td>
<td>0.15</td>
</tr>
<tr>
<td>Ser AGC</td>
<td>19.54</td>
<td>0.24</td>
</tr>
<tr>
<td>Pro CCG</td>
<td>6.33</td>
<td>0.11</td>
</tr>
<tr>
<td>Pro CCA</td>
<td>17.10</td>
<td>0.28</td>
</tr>
<tr>
<td>Pro CCT</td>
<td>18.31</td>
<td>0.30</td>
</tr>
<tr>
<td>Pro CCC</td>
<td>18.42</td>
<td>0.31</td>
</tr>
<tr>
<td>Leu CTG</td>
<td>39.95</td>
<td>0.40</td>
</tr>
<tr>
<td>Leu CTA</td>
<td>7.89</td>
<td>0.08</td>
</tr>
<tr>
<td>Leu CTT</td>
<td>12.97</td>
<td>0.13</td>
</tr>
<tr>
<td>Leu CTC</td>
<td>20.04</td>
<td>0.20</td>
</tr>
<tr>
<td>Ala GCG</td>
<td>6.72</td>
<td>0.10</td>
</tr>
<tr>
<td>Ala GCA</td>
<td>15.80</td>
<td>0.23</td>
</tr>
<tr>
<td>Ala GCT</td>
<td>20.12</td>
<td>0.29</td>
</tr>
<tr>
<td>Ala GCC</td>
<td>26.51</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Codon usage in Mouse genome

<table>
<thead>
<tr>
<th>AA codon</th>
<th>% frequency</th>
<th>frac</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser TCA</td>
<td>11.44</td>
<td>0.14</td>
</tr>
<tr>
<td>Ser TCT</td>
<td>15.70</td>
<td>0.19</td>
</tr>
<tr>
<td>Ser TCC</td>
<td>17.92</td>
<td>0.22</td>
</tr>
<tr>
<td>Ser AGT</td>
<td>12.25</td>
<td>0.15</td>
</tr>
<tr>
<td>Ser AGC</td>
<td>19.54</td>
<td>0.24</td>
</tr>
<tr>
<td>Pro CCG</td>
<td>6.33</td>
<td>0.11</td>
</tr>
<tr>
<td>Pro CCA</td>
<td>17.10</td>
<td>0.28</td>
</tr>
<tr>
<td>Pro CCT</td>
<td>18.31</td>
<td>0.30</td>
</tr>
<tr>
<td>Pro CCC</td>
<td>18.42</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Using codon frequency to find correct reading frame

Consider sequence \(x_1 x_2 x_3 x_4 x_5 x_6 \ldots\)
where \(x_i\) is a nucleotide

Let \(p_1 = p_{x1 \_x2 \_x3} p_{x4 \_x5 \_x6} \ldots\)
\(p_2 = p_{x2 \_x3 \_x4} p_{x5 \_x6 \_x7} \ldots\)
\(p_3 = p_{x3 \_x4 \_x5} p_{x6 \_x7 \_x8} \ldots\)

Then probability that \(i\)th reading frame is the coding frame is:

\[
P_i = \frac{p_i}{p_1 + p_2 + p_3}
\]
slide a window along the sequence and compute \(P_i\)

Adding the background model: gene finding

- In the previous model, we assume at least one reading frame is the codon sequence \(\Rightarrow\) testing reading frames, not gene finding
- Adding a background model
  - \(p_0 = p_{x1} p_{x2} p_{x3} \ldots\)
  - Based on the nucleotide frequency in the non-coding sequence
  - \(P_i = p_i / (p_0 + p_1 + p_2 + p_3)\)
- In practice, this model should be extended to all six reading frames.

Protein secondary structure prediction

Amino acid sequence
NLKTEWELV6KG5V6EE
AKVIQ10D7K6PAEQ11I5V
PVG7T7V7M7EY5MIDVR
LVFDKL7N5IA5EVPRVG

Basic structural units of proteins: Secondary structure

- \(\alpha\)-helix
- \(\beta\)-sheet
- Secondary structures, \(\alpha\)-helix and \(\beta\)-sheet, have regular hydrogen-bonding patterns.
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(α|S)=Π_{s}p(α|s)= Π_{s}p(α|f(s))
- p(α|f(s))~p(f(s)|α)/p(f(s))
- Similarly for β and turn structures

Chou and Fasman: frequency model

| Amino Acid | α-Helix | β-Sheet | Turn | Favors
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>1.29</td>
<td>0.99</td>
<td>0.70</td>
<td>α-Helix</td>
</tr>
<tr>
<td>Cys</td>
<td>1.10</td>
<td>0.98</td>
<td>0.80</td>
<td>β-Sheet</td>
</tr>
<tr>
<td>Leu</td>
<td>1.47</td>
<td>0.97</td>
<td>0.99</td>
<td>Turn</td>
</tr>
<tr>
<td>Glu</td>
<td>1.44</td>
<td>0.73</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>1.27</td>
<td>0.80</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Chou and Fasman: frequency model

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

Profile / PSSM

- DNA / protein segments of the same length L
- Often represented as positional frequency matrix

A DNA profile (matrix)

<table>
<thead>
<tr>
<th>DNA / protein segments</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATAAA</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>TATAAT</td>
<td>T</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TATAAG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TATAAA</td>
<td>A</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>TATATA</td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sparse data → pseudo-counts</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DNA / protein segments</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>TATAAA</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TATAAT</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TATAAG</td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TATAAA</td>
<td>A</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>TATATA</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

<table>
<thead>
<tr>
<th>TATAAA</th>
<th>TCGAT</th>
<th>TATAAA</th>
<th>GCATT</th>
<th>TATAAA</th>
<th>ACTTTA</th>
<th>TATAAA</th>
<th>GCCTGC</th>
<th>TATAAA</th>
<th>AACGCG</th>
<th>TATAAA</th>
<th>CGTATC</th>
<th>TATAAA</th>
<th>CCAAGT</th>
<th>TATAAA</th>
<th>GACCTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>T</td>
<td>9 2</td>
<td>7 2</td>
<td>1 2</td>
<td>1 2</td>
<td>5 3</td>
<td>4</td>
<td>2 3</td>
<td>5 3</td>
<td>5 3</td>
<td>1 2</td>
<td>2 3</td>
<td>5 3</td>
<td>1 2</td>
<td>2 3</td>
<td>1 2</td>
</tr>
<tr>
<td>C</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Equivalent to computing the significance of a sequence motif

Model comparison: relative entropy

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

| T       | 9 2   | 7 2    | 1 2   | 5 3    | 4      | 2 3    | 5 3    | 5 3    | 1 2    | 2 3    | 5 3    | 1 2    | 2 3    | 1 2    |
| C       | 1 1   | 1 1    | 1 1   | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    | 1 1    |
| A       | 1     | 8      | 2     | 8      | 9      | 8      | 9      | 3      | 5      | 3      | 4      | 3      | 5      | 3      | 4      |
| G       | 1     | 1      | 2     | 1      | 1      | 1      | 1      | 2      | 1      | 1      | 1      | 2      | 1      | 1      | 1      |

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)
- $P(H) = \sum_{h \in S} p(h) 1 \ldots 1$  (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:

$$R(S,P) = \prod_{i=1}^{L} \left( \frac{P(x)_{S_i}}{P(x)_{B_i}} \right)$$

- Searching motifs in a sequence: sliding window approach

Finding a motif

Motif finding is difficult
The motif finding problem

- Given a set of DNA sequences:
  
  - Find the motif in each of the individual sequences

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

Conclusion

- 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