582746 Modelling and Analysis in Bioinformatics

Lecture 1: Genomic k-Mer Statistics

Juha Kärkkäinen

06.09.2016
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
Course Topics

- Sequence analysis (Juha Kärkkäinen)
- Network models (Leena Salmela)
- Network inference (Antti Honkela)
Schedule

- **First six weeks**
  - Tuesday 12–14: Lecture to introduce the topic
  - Thursday 10–12: Study group to deepen the knowledge
  - Thursday 12–14: Exercise session to get started on exercises
  - Exercise solutions must be returned about week later

- **Week 7: Guest lectures**
  - Tuesday and Thursday 12–14
How to Pass the Course

- Attending study groups on Thursday mornings is mandatory.
- Attending invited lectures on Tuesday 18.10. and Thursday 20.10. is mandatory.
- Submit the exercises
  - At least 6 points for each three exercise set (sequence analysis, network models, network inference)
  - At least 30 points total
- No exam
- Not possible to pass with separate exam
- If you miss a study group or a visiting lecture, contact the lecturers for an alternative assignment.
Grading

- Grading is based on submitted exercises
- 60 points available
- 30 points is required for passing
- 50 points gives highest grade 5
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
Genomic $k$-mers

- DNA sequence is a sequence of nucleotides or bases ‘A’, ‘C’, ‘G’, and ‘T’
- $k$-mer is sequence of $k$ bases
  - 1-mers: A, C, G, T
  - 2-mers: AA, AC, AG, AT, CA, CC, ...
  - 3-mers (codons): AAA, AAC, ...
  - 4-mers and beyond
- We are interested in the frequencies of $k$-mers in a genome
  - $fr(A)$, $fr(CA)$, $fr(AAC)$, ...
Double Stranded DNA

- DNA is typically double stranded
- In the complementary strand:
  - A ↔ T
  - C ↔ G
- The strands have opposite directions marked by the 5’- and 3’-prime endings

5’– ... GGATCGAAGCTAAGGGCT ... –3’
3’– ... CCTAGCTTCGATTCCCGT ... –5’
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
1-Mer Frequencies

5’– GGATCGAAGCTAAGGGCT –3’
3’– CCTAGCTTCGATTCCCGT –5’

- Top strand: \( \text{fr}(G) = \frac{7}{18}, \text{fr}(C) = \frac{3}{18}, \text{fr}(A) = \frac{5}{18}, \text{fr}(T) = \frac{3}{18} \)
- Bottom strand: \( \text{fr}(G) = \frac{3}{18}, \text{fr}(C) = \frac{7}{18}, \text{fr}(A) = \frac{3}{18}, \text{fr}(T) = \frac{5}{18} \)
- Both strands: \( \text{fr}(G) = \frac{10}{36}, \text{fr}(C) = \frac{10}{36}, \text{fr}(A) = \frac{8}{36}, \text{fr}(T) = \frac{8}{36} \)
G+C or GC Content

5’– GGATCGAAGCTAAGGGCT –3’
3’– CCTAGCTTTCGATTCCCGT –5’

▶ 1-mer frequencies can be summarized as one number, the G+C or GC content
  ▶ \( \text{fr}(G+C) = \text{fr}(G) + \text{fr}(C) = \frac{10}{36} + \frac{10}{36} = \frac{20}{36} \)
  ▶ For double stranded case, other frequencies can be computed from the G+C content
    ▶ \( \text{fr}(G) = \text{fr}(C) = \frac{\text{fr}(G+C)}{2} = \frac{10}{36} \)
    ▶ \( \text{fr}(A+T) = 1 - \text{fr}(G+C) = \frac{16}{36} \)
    ▶ \( \text{fr}(A) = \text{fr}(T) = \frac{\text{fr}(A+T)}{2} = \frac{8}{36} \)
  ▶ G+C content can be computed from a single strand too
    ▶ \( \text{fr}(G+C) = \frac{10}{18} = \frac{20}{36} \)
## G+C content and genome size

<table>
<thead>
<tr>
<th>Species</th>
<th>G+C content</th>
<th>Genome size (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>31.6%</td>
<td>0.585</td>
</tr>
<tr>
<td>Thermoplasma volcanium</td>
<td>39.9%</td>
<td>1.585</td>
</tr>
<tr>
<td>Pyrococcus abyssi</td>
<td>44.6%</td>
<td>1.765</td>
</tr>
<tr>
<td>Escherichia coli K-12</td>
<td>50.7%</td>
<td>4.693</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa PAO1</td>
<td>66.4%</td>
<td>6.264</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>36%</td>
<td>97</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>35%</td>
<td>125</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>41%</td>
<td>3080</td>
</tr>
</tbody>
</table>
Bernoulli or i.i.d. Model for DNA

- i.i.d. = independent and identically distributed
- Generate a sequence so that the base at each position is chosen randomly
  - independently of other positions
  - using identical distribution for all positions
- In the simplest case, we can use a uniform distribution:
  \[ P(A) = P(C) = P(G) = P(T) = 0.25 \]
- If we know the desired G+C content, we can use the base frequencies derived from the G+C content as the probabilities
- Bernoulli is a reasonable model for some organisms but not for others
GC Skew

- GC skew is defined by \((\#G - \#C)/(\#G + \#C)\)
- It is calculated for windows (intervals) of specific width

... AG\underline{GATCGAAT}\underline{GATCGAATC}TAAG ...
   \((2-1)/(2+1) = 1/3\)

... AG\underline{GATCGAAT}\underline{GATCGAATC}TAAG ...
   \((1-2)/(1+2) = -1/4\)

- GC skew can be different in different regions of a genome
GC Skew in A Genome

- Circular genome of Blochmannia vafer
- G+C content is black
- GC skew is purple/green

- The GC skew switch point is *origin of replication*
- More complex organisms have multiple origins of replication and less striking GC skew statistics
DNA Replication

- Asymmetry in replication (continuous vs. piecewise)
- Leading strand tends to have a different GC skew than lagging strand
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
First Order Markov Model

- In Bernoulli model, each base is independent of others
- In a first order Markov model, each base depends on the preceding base
- The probabilities can be obtained from 2-mer frequencies
- Often much more realistic than Bernoulli model
Markov Chain Probabilities

- Let $X$ be a sequence and $X_t$ the base at position $t$
  - $X$ and $X_t$ are random variables
- Probability that $X_t = b$ given that $X_{t-1} = a$:
  \[ p_{ab} = P(X_t = b \mid X_{t-1} = a) \]
- $p_{ab}$ is the same at each position $t$

<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>$p_{AA}$</td>
<td>$p_{AC}$</td>
<td>$p_{AG}$</td>
<td>$p_{AT}$</td>
</tr>
<tr>
<td>C</td>
<td>$p_{CA}$</td>
<td>$p_{CC}$</td>
<td>$p_{CG}$</td>
<td>$p_{CT}$</td>
</tr>
<tr>
<td>G</td>
<td>$p_{GA}$</td>
<td>$p_{GC}$</td>
<td>$p_{GG}$</td>
<td>$p_{GT}$</td>
</tr>
<tr>
<td>T</td>
<td>$p_{TA}$</td>
<td>$p_{TC}$</td>
<td>$p_{TG}$</td>
<td>$p_{TT}$</td>
</tr>
</tbody>
</table>

- Transition matrix

- We also need the probabilities of the first base:
  \[ p_A, p_C, p_G, p_T \]
Generating a Sequence

1. Choose the starting base by the distribution $p_A, p_C, p_G, p_T$
2. Then choose each base using a distribution defined by a row of the transition matrix

   ▶ For example if $X_{t-1} = C$ then $X_t$ is chosen by the distribution
   $p_{CA}, p_{CC}, p_{CG}, p_{CT}$

<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>$p_{AA}$</td>
<td>$p_{AC}$</td>
<td>$p_{AG}$</td>
<td>$p_{AT}$</td>
</tr>
<tr>
<td>C</td>
<td>$p_{CA}$</td>
<td>$p_{CC}$</td>
<td>$p_{CG}$</td>
<td>$p_{CT}$</td>
</tr>
<tr>
<td>G</td>
<td>$p_{GA}$</td>
<td>$p_{GC}$</td>
<td>$p_{GG}$</td>
<td>$p_{GT}$</td>
</tr>
<tr>
<td>T</td>
<td>$p_{TA}$</td>
<td>$p_{TC}$</td>
<td>$p_{TG}$</td>
<td>$p_{TT}$</td>
</tr>
</tbody>
</table>
Estimating the Probabilities

- Use observed frequencies: \( f_r(A), f_r(C), \ldots, f_r(AA), f_r(AC), \ldots \)
- Base probabilities: \( p_A = f_r(A), p_C = f_r(C), \ldots \)
- Note that \( f_r(A) \approx f_r(AA) + f_r(AC) + f_r(AG) + f_r(AT) \)
- Transition probabilities

\[
p_{ab} = P(X_t = b \mid X_{t-1} = a) = \frac{P(X_t = b, X_{t-1} = a)}{P(X_{t-1} = a)} = \frac{f_r(ab)}{f_r(a)}
\]
Estimating the Probabilities

\[ fr(ab) = P(X_t = b, X_{t-1} = a) \quad \text{and} \quad p_{ab} = P(X_t = b \mid X_{t-1} = a) \]

<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>0.146</td>
<td>0.052</td>
<td>0.058</td>
<td>0.089</td>
</tr>
<tr>
<td>C</td>
<td>0.063</td>
<td>0.029</td>
<td>0.010</td>
<td>0.056</td>
</tr>
<tr>
<td>G</td>
<td>0.050</td>
<td>0.030</td>
<td>0.030</td>
<td>0.051</td>
</tr>
<tr>
<td>T</td>
<td>0.086</td>
<td>0.047</td>
<td>0.063</td>
<td>0.140</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>A</td>
<td>0.423</td>
<td>0.151</td>
<td>0.168</td>
<td>0.258</td>
</tr>
<tr>
<td>C</td>
<td>0.399</td>
<td>0.184</td>
<td>0.063</td>
<td>0.354</td>
</tr>
<tr>
<td>G</td>
<td>0.314</td>
<td>0.189</td>
<td>0.176</td>
<td>0.321</td>
</tr>
<tr>
<td>T</td>
<td>0.258</td>
<td>0.138</td>
<td>0.187</td>
<td>0.415</td>
</tr>
</tbody>
</table>

Example: computing \( p_{AC} \)

\[ p_A = fr(A) = fr(AA) + fr(AC) + fr(AG) + fr(AT) \]
\[ = 0.146 + 0.052 + 0.058 + 0.089 = 0.345 \]

\[ p_{AC} = fr(AC) / fr(A) = 0.052 / 0.345 \approx 0.151 \]
CpG Suppression

- Low frequency of CG is common (in vertebrates)
- Human genome: \( \text{fr(CG)} \approx 0.01 \)
  - Bernoulli model: \( \text{fr(C+G)} = 0.41 \Rightarrow \text{fr(CG)} \approx 0.04 \)
- This is called CG or CpG suppression
  - \( (\text{CpG} = \text{C-phosphate-G}) \)
- Caused by the tendency of CG to mutate into TG
- Some regions, called CpG islands, have a higher CG frequency
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
2nd Order Markov Model

- Markov model generalizes to higher orders
- In 2nd order Markov model, position $t$ depends on positions $t-1$ and $t-2$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$p_{AAA}$</td>
<td>$p_{AAC}$</td>
<td>$p_{AAG}$</td>
<td>$p_{AAT}$</td>
</tr>
<tr>
<td>AC</td>
<td>$p_{ACA}$</td>
<td>$p_{ACC}$</td>
<td>$p_{ACG}$</td>
<td>$p_{ACT}$</td>
</tr>
<tr>
<td>AG</td>
<td>$p_{AGA}$</td>
<td>$p_{AGC}$</td>
<td>$p_{AGG}$</td>
<td>$p_{AGT}$</td>
</tr>
<tr>
<td>AT</td>
<td>$p_{ATA}$</td>
<td>$p_{ATC}$</td>
<td>$p_{ATG}$</td>
<td>$p_{ATT}$</td>
</tr>
<tr>
<td>CA</td>
<td>$p_{CAA}$</td>
<td>$p_{CAC}$</td>
<td>$p_{CAG}$</td>
<td>$p_{CAT}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Can be estimated from 3-mer frequencies
Codons

- In genes, amino acids are encoded by 3-mers called codons.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala/A</td>
<td>GCT, GCC, GCA, GCG</td>
</tr>
<tr>
<td>Arg/R</td>
<td>CGT, CGC, CGA, CGG, AGA, AGG</td>
</tr>
<tr>
<td>Asn/N</td>
<td>AAT, AAC</td>
</tr>
<tr>
<td>Asp/D</td>
<td>GAT, GAC</td>
</tr>
<tr>
<td>Cys/C</td>
<td>TGT, TGC</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- DNA sequence can be divided into codons in three different ways called reading frames:
  
  AGG TGA CAC CGC AAG CCT TAT ATT AGC A
  A GGT GAC ACC GCA AGC CTT ATA TTA GCA
  AG GTG ACA CCG CAA GCC TTA TAT TAG CA

- 3-mer models can help in identifying reading frames.
**Codon Usage Bias**

- Highly expressed genes prefer certain codons among synonymous ones

### Codon frequencies for some genes in E. coli

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codon</th>
<th>Predicted</th>
<th>Gene expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>TTT</td>
<td>0.493</td>
<td>0.551 0.291</td>
</tr>
<tr>
<td></td>
<td>TTC</td>
<td>0.507</td>
<td>0.449 0.709</td>
</tr>
<tr>
<td>Ala</td>
<td>GCT</td>
<td>0.246</td>
<td>0.145 0.275</td>
</tr>
<tr>
<td></td>
<td>GCC</td>
<td>0.254</td>
<td>0.276 0.164</td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td>0.246</td>
<td>0.196 0.240</td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>0.254</td>
<td>0.382 0.323</td>
</tr>
<tr>
<td>Asn</td>
<td>AAT</td>
<td>0.493</td>
<td>0.409 0.172</td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>0.507</td>
<td>0.591 0.828</td>
</tr>
</tbody>
</table>
Codon Adaptation Index (CAI)

- Let \( X = x_1x_2\ldots x_n \) be a sequence of codons in a gene
- \( p(x_i) \) is the probability of \( x_i \) among synonymous codons in a highly expressed gene
- \( q(x_i) \) is the highest probability among synonymous codons
- Example: Alanine in E. coli
  - \( p(\text{GCT})=0.275, p(\text{GCC})=0.164, p(\text{GCA})=0.240, p(\text{GCG})=0.323 \)
  - \( q(\text{GCT}) = q(\text{GCC}) = q(\text{GCA}) = q(\text{GCG}) = 0.323 \)
- CAI is defined as

\[
\text{CAI} = \left( \prod_{i=1}^{n} \frac{p(x_i)}{q(x_i)} \right)^{1/n}
\]
## CAI: Example

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Ala</th>
<th>Tyr</th>
<th>Met</th>
<th>Ser</th>
<th>$\Pi$</th>
<th>CAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codon</td>
<td>GCT</td>
<td>GAA</td>
<td>ATG</td>
<td>TCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>0.47</td>
<td>0.19</td>
<td>1.00</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$q$</td>
<td>0.47</td>
<td>0.81</td>
<td>1.00</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p/q$</td>
<td>1.00</td>
<td>0.23</td>
<td>1.00</td>
<td>0.07</td>
<td>0.016</td>
<td>0.016^{1/4}</td>
</tr>
</tbody>
</table>

» CAI = \(0.016^{1/4} = 0.36\)
CAI: Properties

- CAI = 1.0: each codon is the most probable one
- In a sample E. coli genes CAI ranged from 0.2 to 0.85
- As the product can be a very small number, numerical problems could be avoided by using a log-odds form
  \[
  \log\text{CAI} = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{p(x_i)}{q(x_i)}\right)
  \]
- CAI can be used for predicting expression levels
Outline

Course introduction

Genomic $k$-Mers

1-Mers

2-Mers

3-Mers

$k$-Mers for Larger $k$
**k-Mers for Larger k**

- \((k - 1)\)th order Markov models
- For large \(k\), summary statistics can be more useful
  - Surprising (over- or under-abundant) \(k\)-mers
  - \(k\)-mer spectra
- For large enough \(k\), most frequencies are zero
  - The set of \(k\)-mers is widely used in genome assembly in the form of De Bruijn graphs
De Bruijn Graph

- Directed graph
- Vertices are \((k - 1)\)-mers
- Edges are \(k\)-mers
  - Connects prefix to suffix
- Example: \(k\)-mers = ATG, TGC, GTG, TGG, GGC, GCA, GCG, CGT
Surprising $k$-mers

- $fr =$ observed frequency of a $k$-mer
- $p =$ expected frequency using a lower order model
- odds-ratio
  $$OR = \frac{(1 - fr + 0.5)/(fr + 0.5)}{(1 - p + 0.5)/(p + 0.5)}$$
  + 0.5 prevents division by zero
- if $OR \gg 1$, the $k$-mer is over-abundant
- if $OR \ll 1$, the $k$-mer is under-abundant
- Over- and under-abundant $k$-mers are more likely to be interesting
  - Why are they over- or under-abundant?
  - Are some under-abundant $k$-mers “poisonous”?
Most Over-Abundant 11-mers in Human Genome

<table>
<thead>
<tr>
<th>k-mer</th>
<th>Count</th>
<th>Expected</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>cgccgccccggc</td>
<td>2448</td>
<td>0.061</td>
<td>40098</td>
</tr>
<tr>
<td>gcgcgccccggc</td>
<td>2437</td>
<td>0.060</td>
<td>39949</td>
</tr>
<tr>
<td>cgccgccccggc</td>
<td>4642</td>
<td>0.481</td>
<td>9650</td>
</tr>
<tr>
<td>cgccgccccggc</td>
<td>4601</td>
<td>0.480</td>
<td>9566</td>
</tr>
<tr>
<td>ccgcgccccggc</td>
<td>19701</td>
<td>2.541</td>
<td>7753</td>
</tr>
<tr>
<td>ggcgggccgcggg</td>
<td>19556</td>
<td>2.539</td>
<td>7703</td>
</tr>
<tr>
<td>caccgccgcggc</td>
<td>19907</td>
<td>2.895</td>
<td>6876</td>
</tr>
<tr>
<td>ccgggggcgggtg</td>
<td>19848</td>
<td>2.888</td>
<td>6873</td>
</tr>
<tr>
<td>accgcgcgccccg</td>
<td>19655</td>
<td>3.002</td>
<td>6547</td>
</tr>
<tr>
<td>cgccgggcgggt</td>
<td>19539</td>
<td>2.996</td>
<td>6522</td>
</tr>
</tbody>
</table>
## Most Under-Abundant 11-mers in Human Genome

<table>
<thead>
<tr>
<th>k-mer</th>
<th>Count</th>
<th>Expected</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>tcgaaattcgc</td>
<td>0</td>
<td>46.0</td>
<td>0.011</td>
</tr>
<tr>
<td>cccccccccctat</td>
<td>9</td>
<td>817.2</td>
<td>0.012</td>
</tr>
<tr>
<td>attgcgaacga</td>
<td>0</td>
<td>41.9</td>
<td>0.012</td>
</tr>
<tr>
<td>tcgcgagttaa</td>
<td>0</td>
<td>34.2</td>
<td>0.015</td>
</tr>
<tr>
<td>atcttgcgcgag</td>
<td>0</td>
<td>33.9</td>
<td>0.015</td>
</tr>
<tr>
<td>tcagggggggggg</td>
<td>15</td>
<td>1003.5</td>
<td>0.015</td>
</tr>
<tr>
<td>atcgcaacgga</td>
<td>0</td>
<td>32.3</td>
<td>0.015</td>
</tr>
<tr>
<td>tatgtttcgcg</td>
<td>0</td>
<td>31.9</td>
<td>0.016</td>
</tr>
<tr>
<td>tgcaacgatcg</td>
<td>0</td>
<td>31.1</td>
<td>0.016</td>
</tr>
<tr>
<td>agtccgcgcaa</td>
<td>0</td>
<td>30.3</td>
<td>0.016</td>
</tr>
</tbody>
</table>
For example, there are over 700,000 distinct $k$-mers that occur exactly 20 times.

More in the study groups.
What Next?

- Thu 10-12: study groups
  - Assignments on the course home page
  - Read the assigned material in advance
  - Mandatory attendance

- Thu 12-14: exercise session
  - Exercise problems will be added to the course home page
  - No advance preparation required
  - Solutions must be returned by 15.9.
  - No mandatory attendance