582746 Modelling and Analysis in Bioinformatics

Lecture 1: Genomic k-Mer Statistics

Juha Kärkkäinen

06.09.2016

Outline

Outline

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

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 $\longleftrightarrow$ T
  - C $\longleftrightarrow$ G
- The strands have opposite directions marked by the 5'- and 3'-prime endings

5'– ... GGATCGAAGCTAAGGGCT ... –3'
3'– ... CCTAGCTTCGATTCCCGT ... –5'

Outline

1-Mer Frequencies

5'– GGATCGAAGCTAAGGGCT –3'
3'– CCTAGCTTCGATTCCCGT –5'

- Top strand: $\text{fr}(\text{G}) = 7/18$, $\text{fr}(\text{C}) = 3/18$, $\text{fr}(\text{A}) = 5/18$, $\text{fr}(\text{T}) = 3/18$
- Bottom strand: $\text{fr}(\text{G}) = 3/18$, $\text{fr}(\text{C}) = 7/18$, $\text{fr}(\text{A}) = 3/18$, $\text{fr}(\text{T}) = 5/18$
- Both strands: $\text{fr}(\text{G}) = 10/36$, $\text{fr}(\text{C}) = 10/36$, $\text{fr}(\text{A}) = 8/36$, $\text{fr}(\text{T}) = 8/36$

G+C or GC Content

5'– GGATCGAAGCTAAGGGCT –3'
3'– CCTAGCTTCGATTCCCGT –5'

- 1-mer frequencies can be summarized as one number, the G+C or GC content
  - fr(G+C) = fr(G) + fr(C) = 10/36 + 10/36 = 20/36
- For double stranded case, other frequencies can be computed from the G+C content
  - fr(G) = fr(C) = fr(G+C)/2 = 10/36
  - fr(A+T) = 1 − fr(G+C) = 16/36
  - fr(A) = fr(T) = fr(A+T)/2 = 8/36
- G+C content can be computed from a single strand too
  - fr(G+C) = 10/18 = 20/36

G+C content and genome size

Species	G+C content	genome size (Mb)
Mycoplasma genitalium	31.6%	0.585
Thermoplasma volcanium	39.9%	1.585
Pyrococcus abyssi	44.6%	1.765
Escherichia coli K-12	50.7%	4.693
Pseudomonas aeruginosa PAO1	66.4%	6.264
Caenorhabditis elegans	36%	97
Arabidopsis thaliana	35%	125
Homo sapiens	41%	3080

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 $(\#\mathrm{G} - \#\mathrm{C})/(\#\mathrm{G} + \#\mathrm{C})$
- It is calculated for windows (intervals) of specific width

... A GGATCGAA TCTAAG ... (3-1)/(3+1) = 2/4

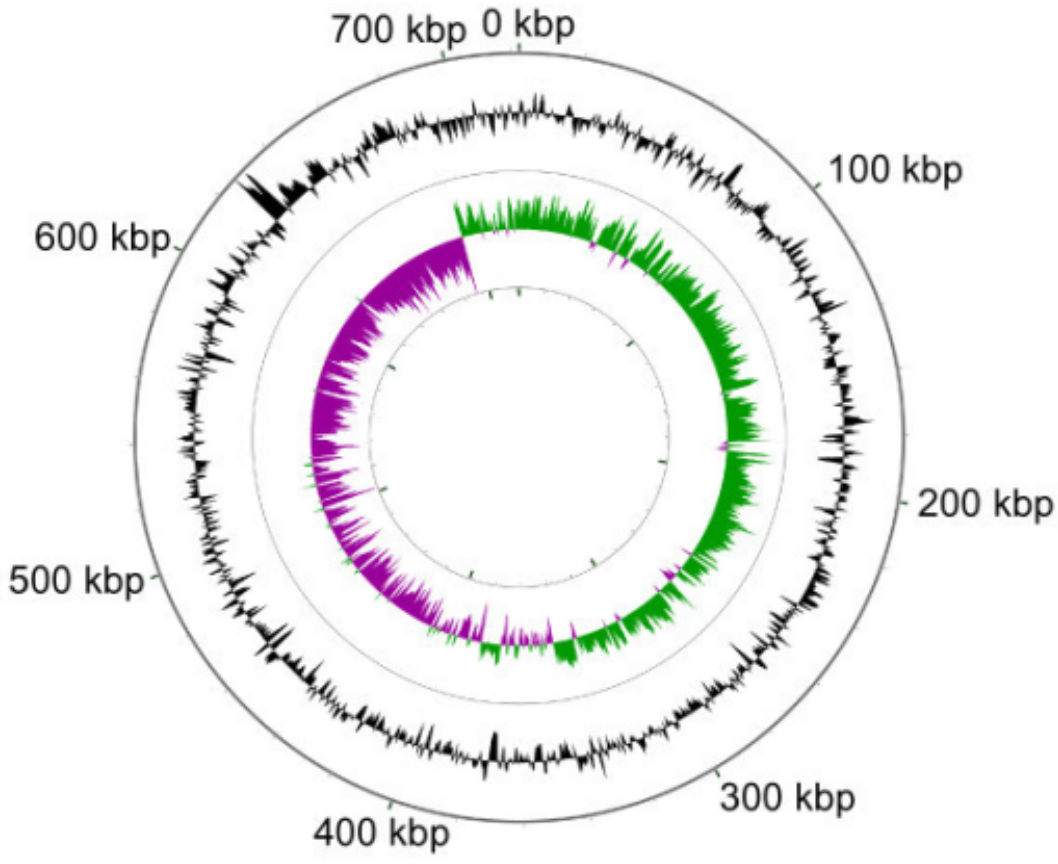
... AG GATCGAAT CTAAG ... (2-1)/(2+1) = 1/3

... AGG ATCGAATC TAAG ... (1-2)/(1+2) = −1/4

- GC skew can be different in different regions of a genome

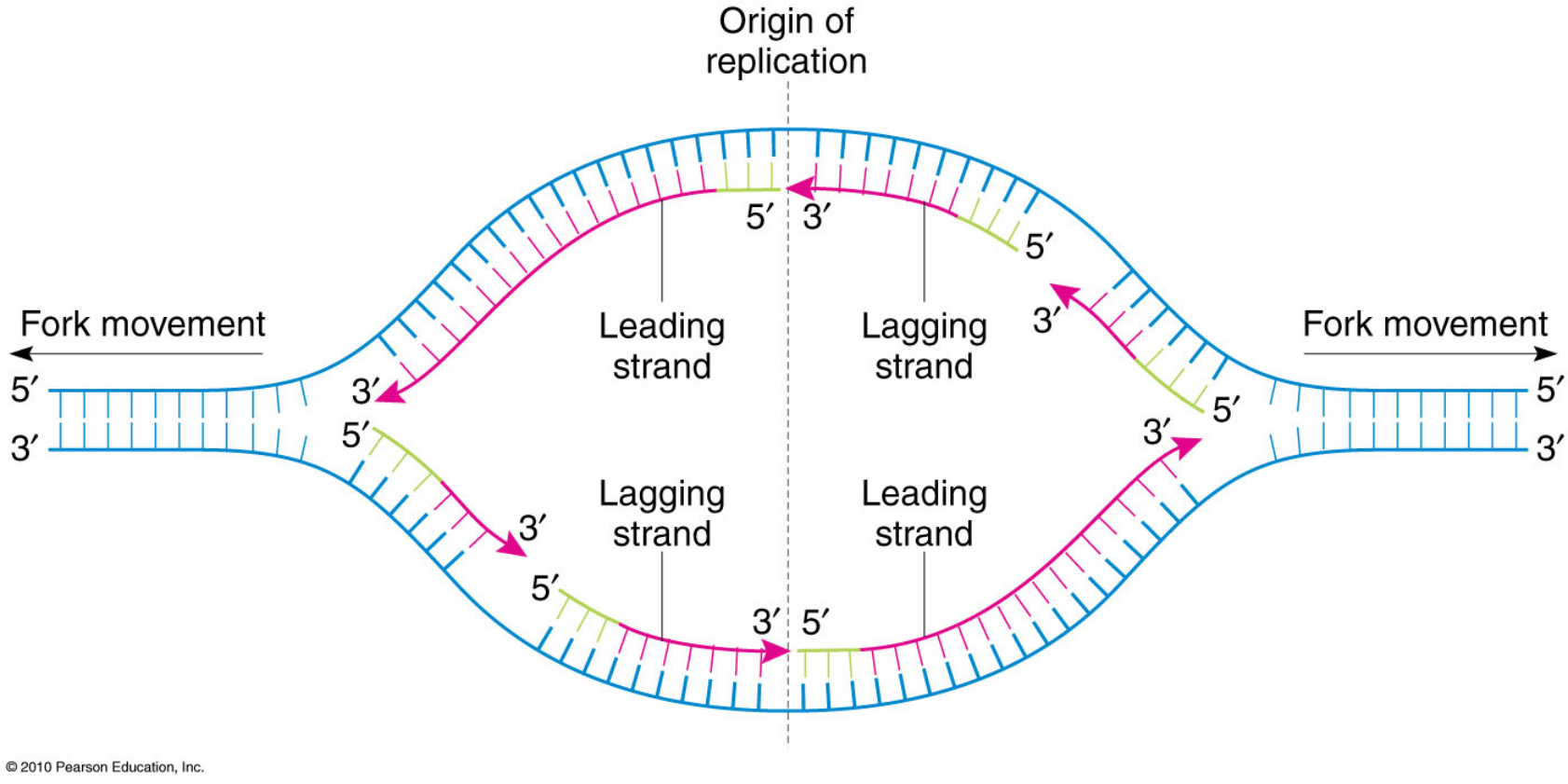
GC Skew in A Genome

- 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

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 that same at each position t
- Transition matrix

	A	C	G	T
A	p_{AA}	p_{AC}	p_{AG}	p_{AT}
C	p_{CA}	p_{CC}	p_{CG}	p_{CT}
G	p_{GA}	p_{GC}	p_{GG}	p_{GT}
T	p_{TA}	p_{TC}	p_{TG}	p_{TT}

- 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}

	A	C	G	T
A	p_{AA}	p_{AC}	p_{AG}	p_{AT}
C	p_{CA}	p_{CC}	p_{CG}	p_{CT}
G	p_{GA}	p_{GC}	p_{GG}	p_{GT}
T	p_{TA}	p_{TC}	p_{TG}	p_{TT}

Estimating the Probabilities

- Use observed frequencies: fr(A), fr(C), ..., fr(AA), fr(AC), ...
- Base probabilities: $p_A = \text{fr(A)}$, $p_C = \text{fr(C)}$, ...
- Note that $\text{fr(A)} \approx \text{fr(AA)} + \text{fr(AC)} + \text{fr(AG)} + \text{fr(AT)}$
- Transition probabilities

$$p_{ab} = P(X_t = b \mid X_{t-1} = a) = \frac{P(X_t = b,\ X_{t-1} = a)}{P(X_{t-1} = a)} = \frac{fr(ab)}{fr(a)}$$

Estimating the Probabilities

$\text{fr}(ab) = P(X_t = b, X_{t-1} = a)$

	A	C	G	T
A	0.146	0.052	0.058	0.089
C	0.063	0.029	0.010	0.056
G	0.050	0.030	0.030	0.051
T	0.086	0.047	0.063	0.140

$p_{ab} = P(X_t = b \mid X_{t-1} = a)$

	A	C	G	T
A	0.423	**0.151**	0.168	0.258
C	0.399	0.184	0.063	0.354
G	0.314	0.189	0.176	0.321
T	0.258	0.138	0.187	0.415

Example: computing p_{AC}

- $p_A = \text{fr}(\text{A}) = \text{fr}(\text{AA}) + \text{fr}(\text{AC}) + \text{fr}(\text{AG}) + \text{fr}(\text{AT})$
  $= 0.146 + 0.052 + 0.058 + 0.089 = 0.345$
- $p_{AC} = \text{fr}(\text{AC}) \;/\; \text{fr}(\text{A}) = 0.052 \;/\; 0.345 \approx 0.151$

CpG Suppression

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

Outline

2nd Order Markov Model

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

	A	C	G	T
AA	p_{AAA}	p_{AAC}	p_{AAG}	p_{AAT}
AC	p_{ACA}	p_{ACC}	p_{ACG}	p_{ACT}
AG	p_{AGA}	p_{AGC}	p_{AGG}	p_{AGT}
AT	p_{ATA}	p_{ATC}	p_{ATG}	p_{ATT}
CA	p_{CAA}	p_{CAC}	p_{CAG}	p_{CAT}
...				

- Can be estimated from 3-mer frequencies

Codons

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

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

- 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

			Gene expression	
Amino acid	Codon	Predicted	Moderate	High
Phe	TTT	0.493	0.551	0.291
	TTC	0.507	0.449	0.709
Ala	GCT	0.246	0.145	0.275
	GCC	0.254	0.276	0.164
	GCA	0.246	0.196	0.240
	GCG	0.254	0.382	0.323
Asn	AAT	0.493	0.409	0.172
	AAC	0.507	0.591	0.828

Codon Adaptation Index (CAI)

- Let $X = x_1 x_2 \ldots x_n$ 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(GCT)=0.275, p(GCC)=0.164, p(GCA)=0.240, p(GCG)=0.323
  - q(GCT) = q(GCC) = q(GCA) = q(GCG) = 0.323
- CAI is defined as

$$\text{CAI} = \left(\prod_{i=1}^{n} p(x_i)/q(x_i) \right)^{1/n}$$

CAI: Example

Amino acid	Ala	Tyr	Met	Ser	$\prod$	CAI
Codon	GCT	GAA	ATG	TCA		
p	0.47	0.19	1.00	0.03		
q	0.47	0.81	1.00	0.43		
p/q	1.00	0.23	1.00	0.07	0.016	$0.016^{1/4}$

- 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

$$\log\text{CAI} = (1/n)\sum_{i=1}^{n}\log(p(x_i)/q(x_i))$$

- CAI can be used for predicting expression levels

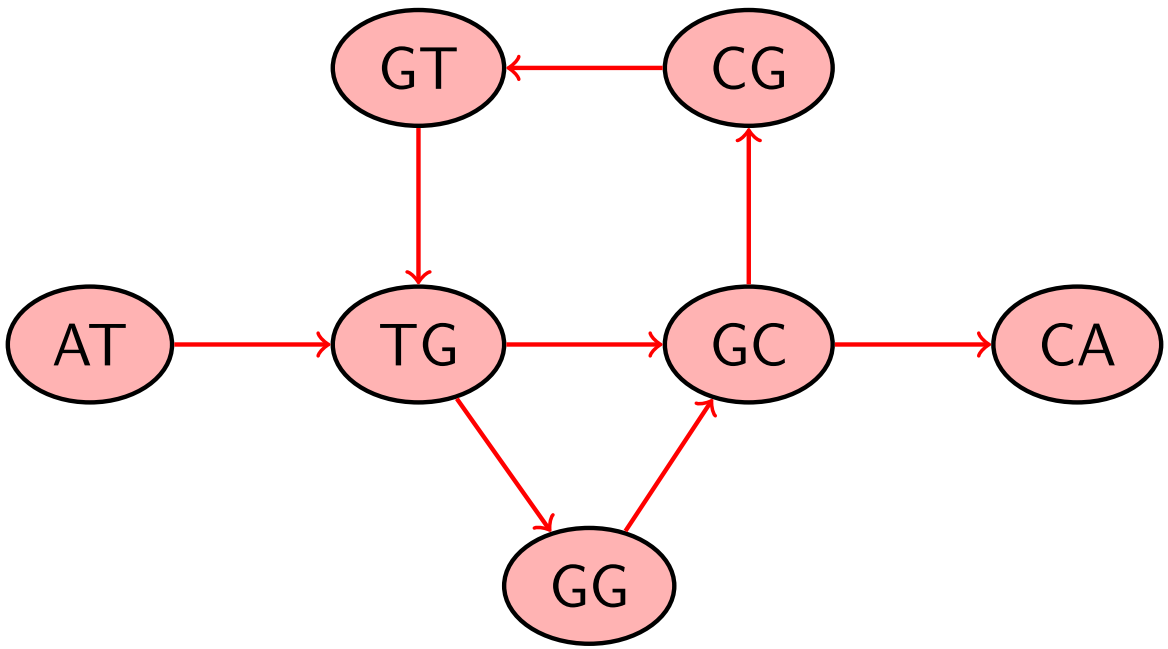
Outline

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 $=$
  $\mathrm{ATG}, \mathrm{TGC}, \mathrm{GTG}, \mathrm{TGG}, \mathrm{GGC}, \mathrm{GCA}, \mathrm{GCG}, \mathrm{CGT}$



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 interesting
  - Why are they over- or under-abundant?
  - Are some under-abundant k-mers "poisonous"?

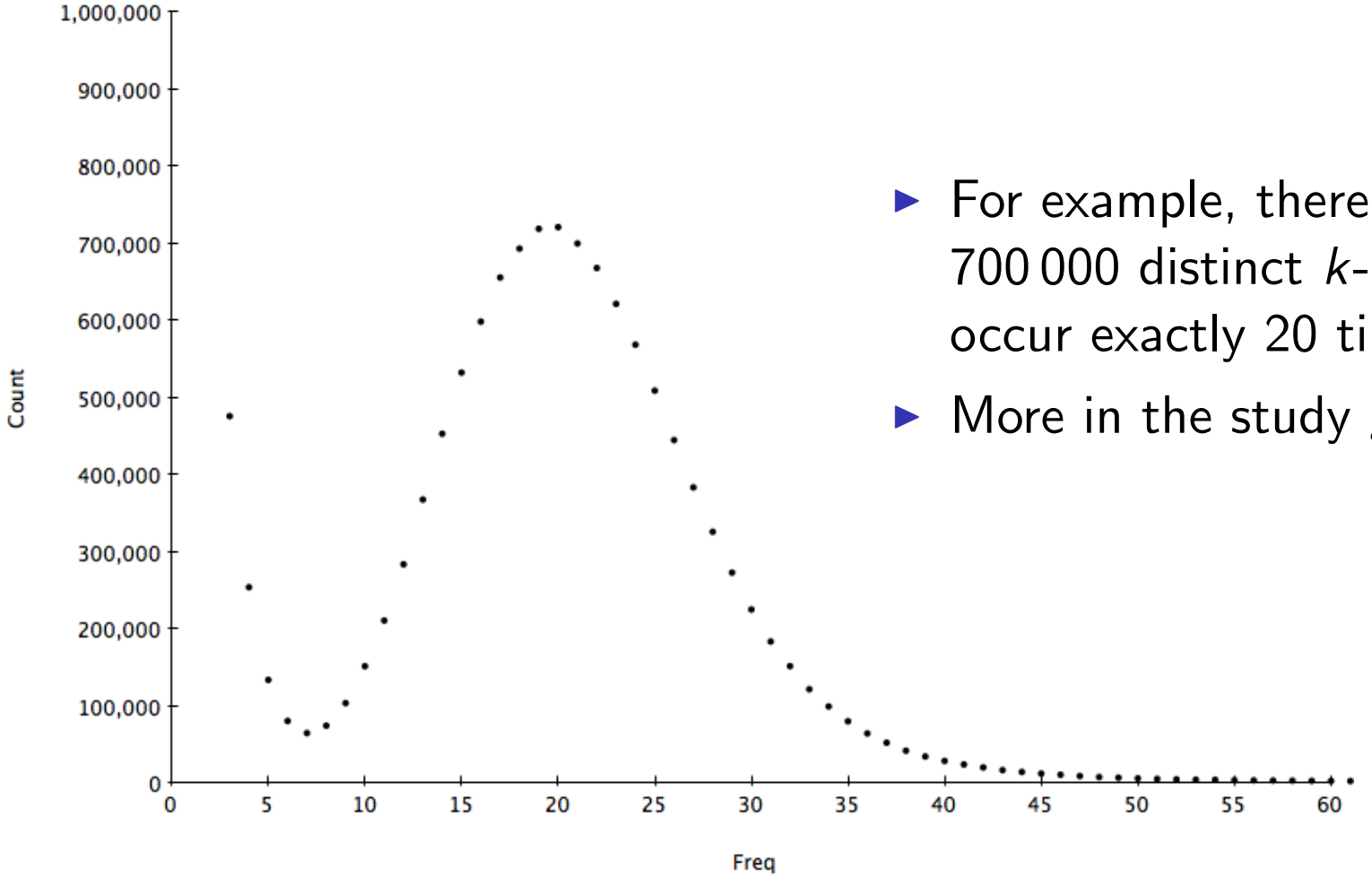
Most Over-Abundant 11-mers in Human Genome

k-mer	Count	Expected	OR
cgcgcgcgcgc	2448	0.061	40098
gcgcgcgcgcg	2437	0.060	39949
cgccgccgccg	4642	0.481	9650
cggcggcggcg	4601	0.480	9566
ccgcgcccggc	19701	2.541	7753
gccgggcgcgg	19556	2.539	7703
caccgcgcccg	19907	2.895	6876
cgggcgcggtg	19848	2.888	6873
accgcgcccgg	19655	3.002	6547
ccgggcgcggt	19539	2.996	6522

Most Under-Abundant 11-mers in Human Genome

k-mer	Count	Expected	OR
tcgaaattcgc	0	46.0	0.011
ccccccccctat	9	817.2	0.012
attgcgaacga	0	41.9	0.012
tcgcgagttaa	0	34.2	0.015
atcttcgcgag	0	33.9	0.015
tcagggggggg	15	1003.5	0.015
atcgcaacgga	0	32.3	0.015
tatgtttcgcg	0	31.9	0.016
tgcaacgatcg	0	31.1	0.016
agtccgcgcaa	0	30.3	0.016

k-Mer Spectra



- 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