Aalto University
School of Science

CS-E5865 Computational genomics

Autumn 2020

Lecturer: Pekka Marttinen

Assistants: Alejandro Ponce de León, Zeinab Yousefi, Onur Poyraz

Lecture 1, 2020

Course logistics

- Lecturer: Pekka Marttinen, firstname.lastname@aalto.fi
- Teachine assistants (TAs):
  - Alejandro Ponce de León, Zeinab Yousefi, Onur Poyraz
- Course webpage in MyCourses
- Schedule:
  - See *comp_gen_timetable_2020.pdf* in myCourses
- Course exam: Tuesday, Oct 20th, 9:00-13:00

NOTE: the exam time is tentative, check the final time from Oodi!

Online implementation in 2020

- The lectures are recorded and released in advance.
- Students can post questions about the lectures in **Slack**.
- Each lecture is followed by an online Q&A session in Zoom. The lecturer will go through questions related to the lecture posted in the Slack and the students can also ask additional questions.
- Links to Slack and Zoom will be posted in MyCourses.

Exercises

- 5 sets of assignments
- Assignments are released on Fridays. Students return their answers in MyCourses as a single PDF one week later, on Fridays at **23:55**.
- Getting help:
  - Write a question in a dedicated Slack channel. The TAs will answer them at the times of the exercise sessions (possibly also at other times, see the details in MyCourses).
  - TAs will be present in a Zoom meeting during the exercise sessions and can provide help for getting started with assignments.
  - ***Students are welcome to comment and give hints to each other's questions in Slack; however, do not reveal the full answer.***
- The due date and the time of the related exercise session are written on the exercise sheet.

Computer exercises

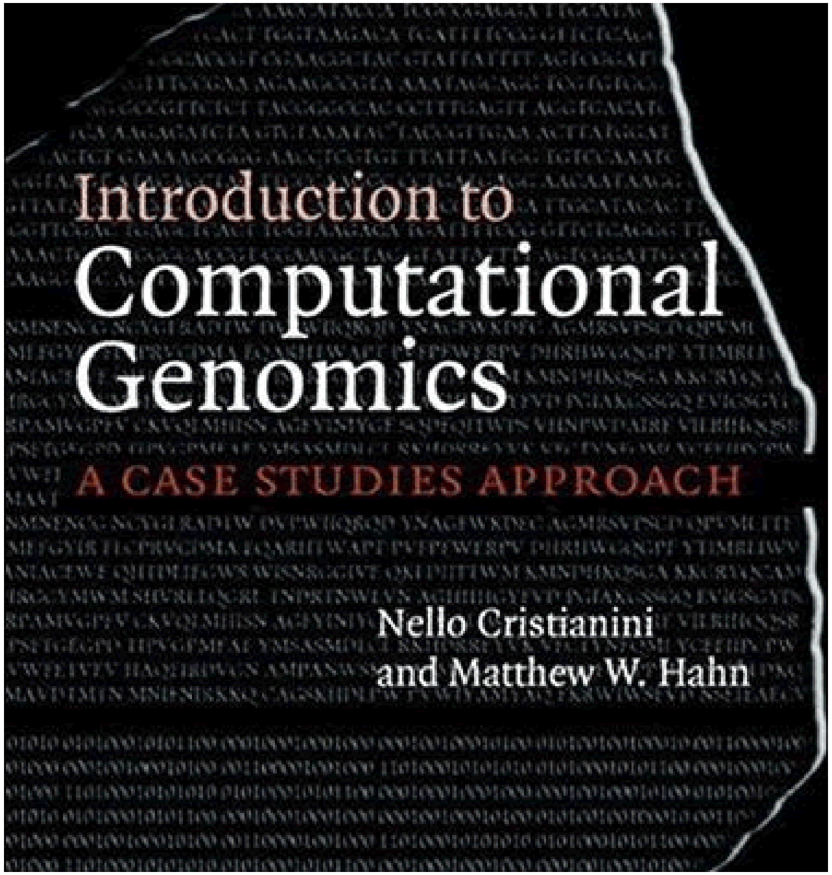
- These are like the “regular” exercises but are done with a computer, and usually consist of programming assignments.
- The students must return the required files (often code) in MyCourses.
- The language is R. If a student wants to use some other language, that’s allowed, but there will not be any support.
- Computer exercise 1 (on 1st week) consists of an introduction to R.

Completing the course

- **Exam** is graded from 0 to 5
  - Arranged online, more details will be provided later.
- **Exercises** (both regular and computer)
  - Graded by the TAs. Points per problem, for example: 0p (not done or completely wrong), 1p (reasonable, somewhat correct), 2p (mostly correct)
- **Final grade** is a weighted average of
  - Exam, weight 35%
  - Exercises, 30%
  - Computer exercises, 35 %

Course Book

- Lectures and exercises follow the Cristianini & Hanh book (more or less)
- Aalto Library: https://alli.linneanet.fi/vwebv/holdingsInfo?searchId=291&recCount=10&recPointer=0&bibId=608709
- From Book stores: suomalainen.com, amazon.co.uk, amazon.com
- Accompanying web site (material for computer exercises): http://www.computational-genomics.net/


Introduction to Computational Genomics

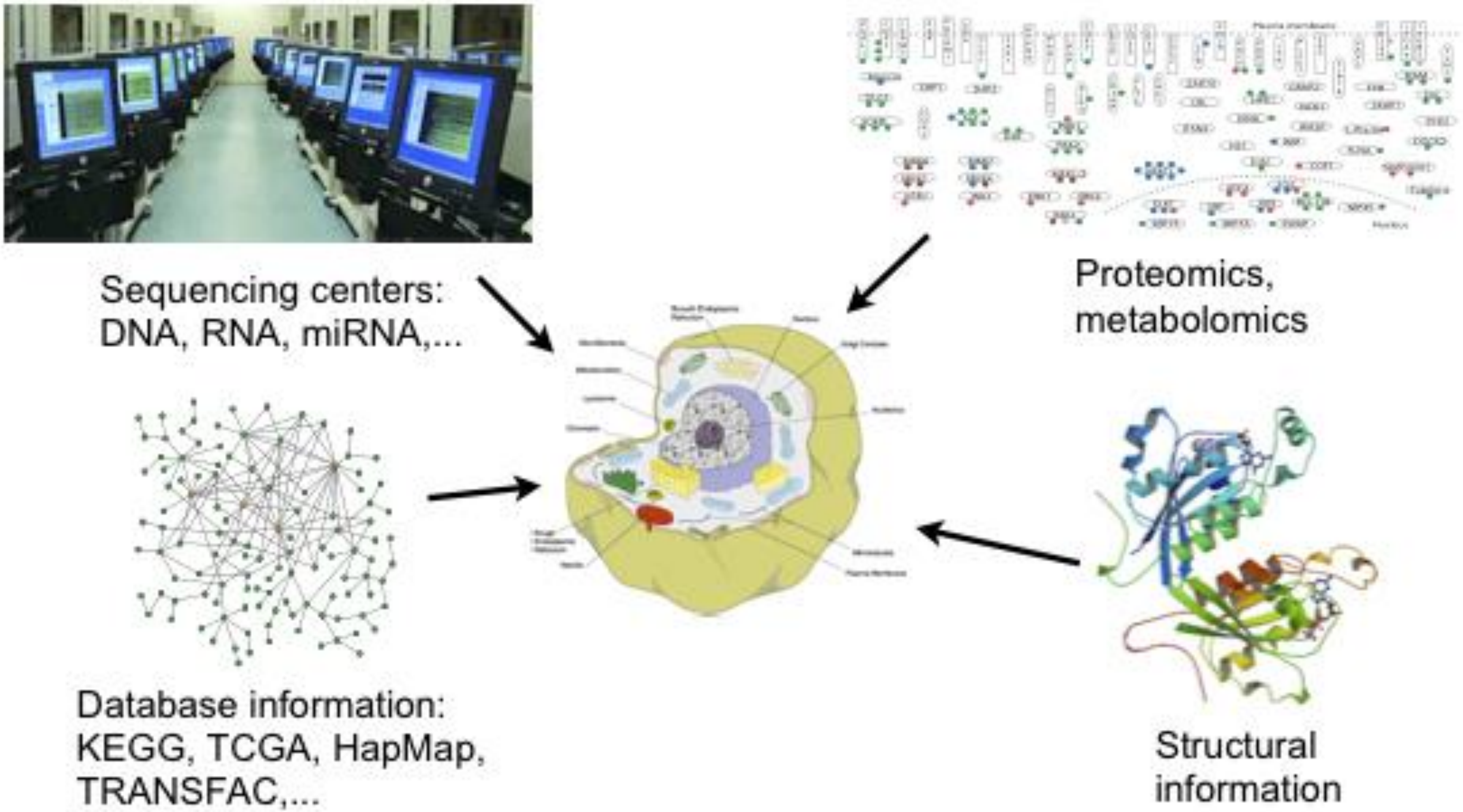
A CASE STUDIES APPROACH

Nello Cristianini
and Matthew W. Hahn

Topics to be covered

- Sequence statistics
- Gene finding
- Sequence alignment
- Hidden Markov Models
- Genome Variation
- Phylogenetic analysis
- Whole-genome comparisons

Biological challenge



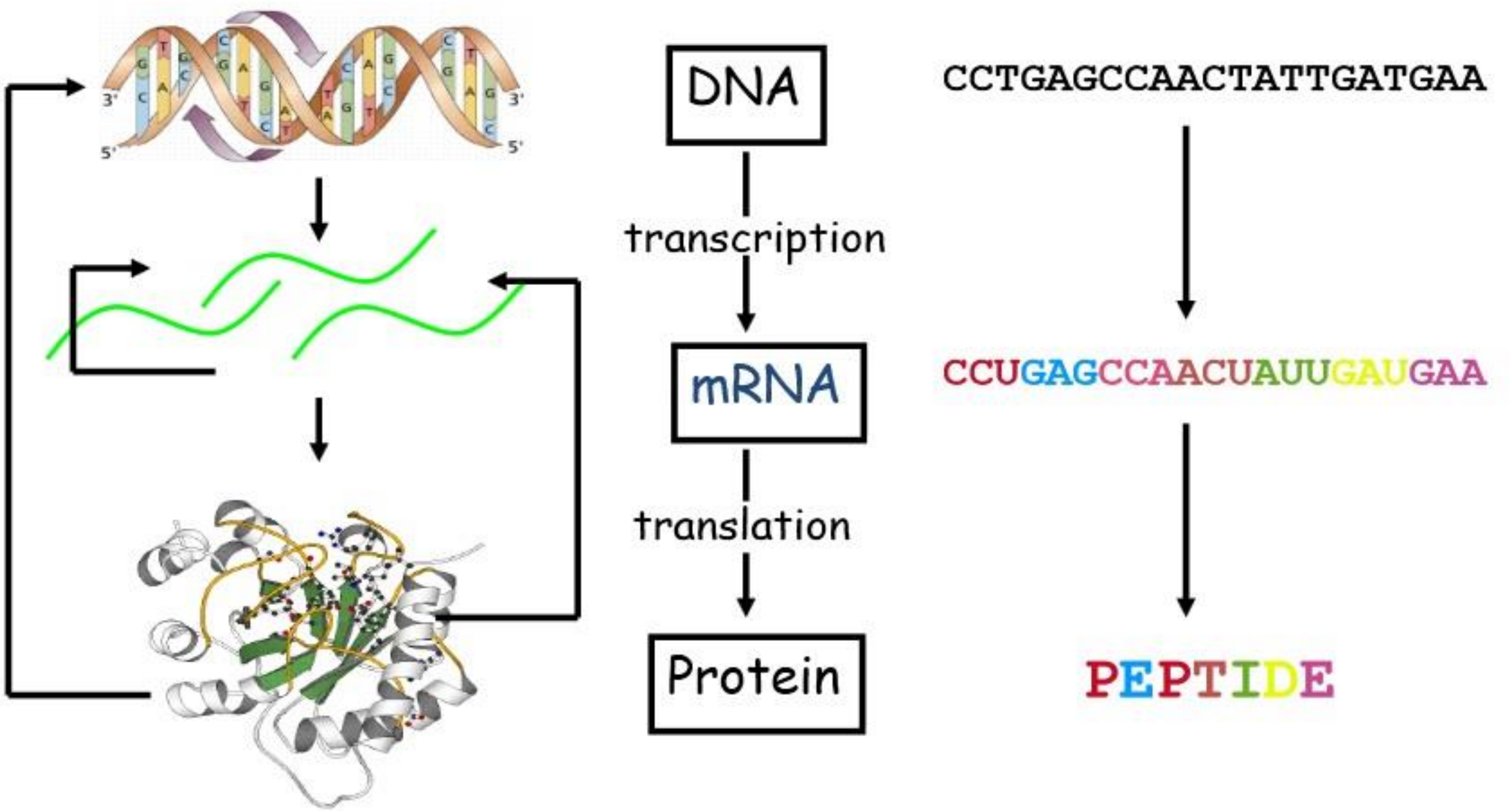
Sequencing centers: DNA, RNA, miRNA,...

Proteomics, metabolomics

Database information: KEGG, TCGA, HapMap, TRANSFAC,...

Structural information

Central dogma of molecular biology



DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

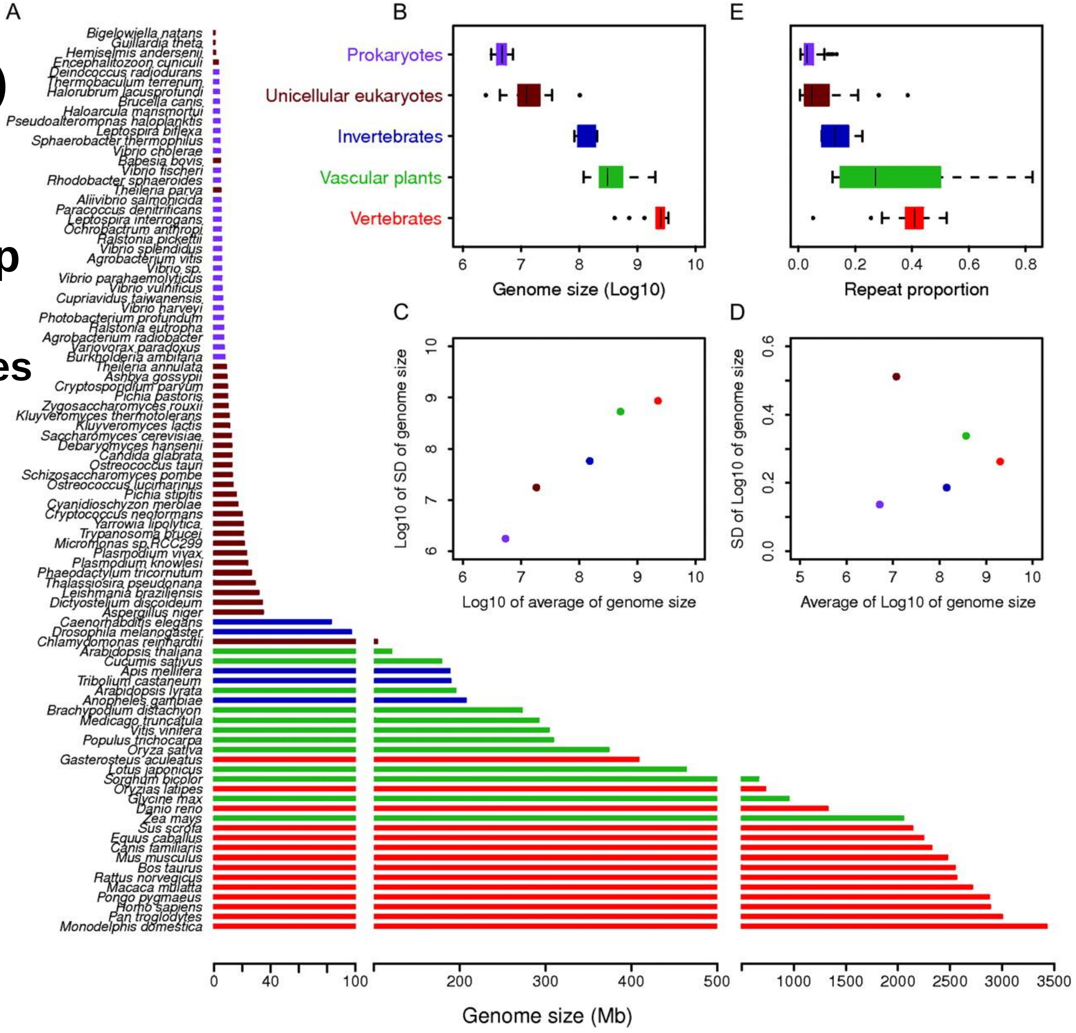
PEPTIDE

Genome

- A genome is an organism's complete set of DNA (including its genes).
- In humans, less than 2% of the genome encodes for genes.
- However, a much larger % of the genome is transcribed (miRNAs, lncRNAs, ...)
- And a large part of the rest of the genome serves as a control regions.

Genome sizes

- Prokaryotes < 10^7 base pairs (bp)
  - bacteria and archea
  - cell without nucleaus
- Unicellular eukaryotes: 10^7-10^8 bp
  - yeasts
  - have nucleus and other organelles
- Invertebrates: ca. 10^8 bp
  - worms, insects, ...
  - organisms without spine
- Vascular planta: 10^8-10^9 bp
  - trees, flowering plants,..
- Vertebrates: > 10^9 bp mostly
  - organisms with spine
  - mammals, fish, ...



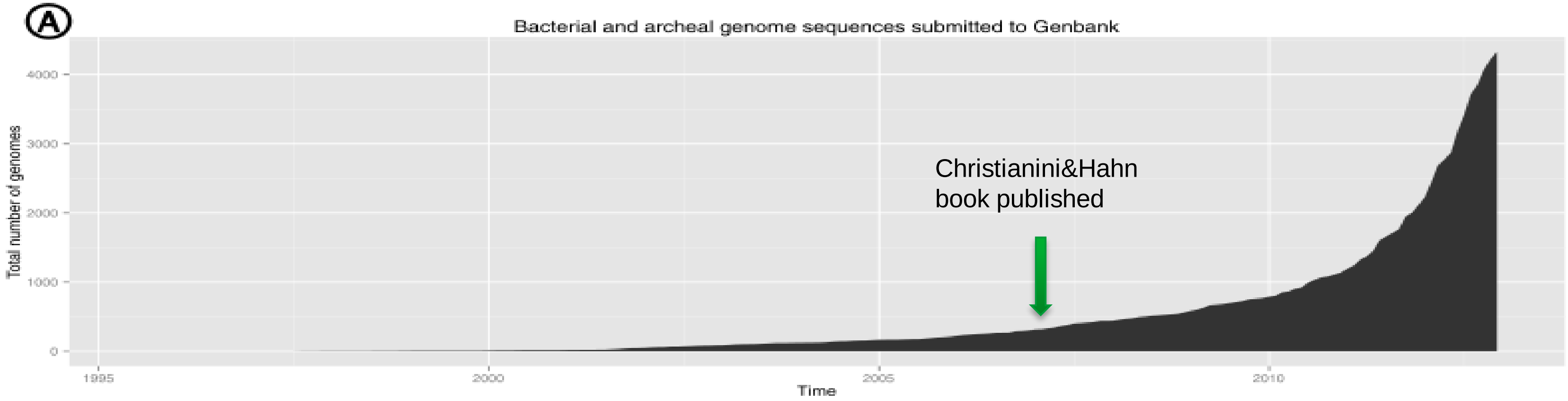
Li X et al. Mol Biol Evol 2011;28:1901-1911

Aalto University
School of Science

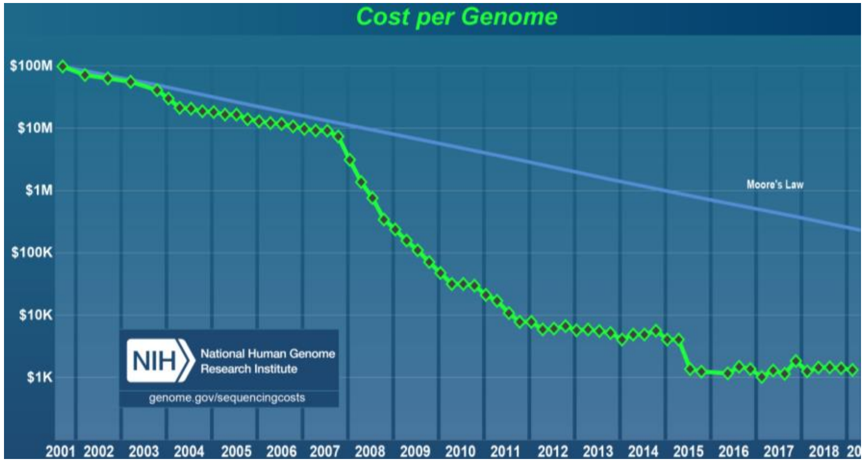
Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution 2011.

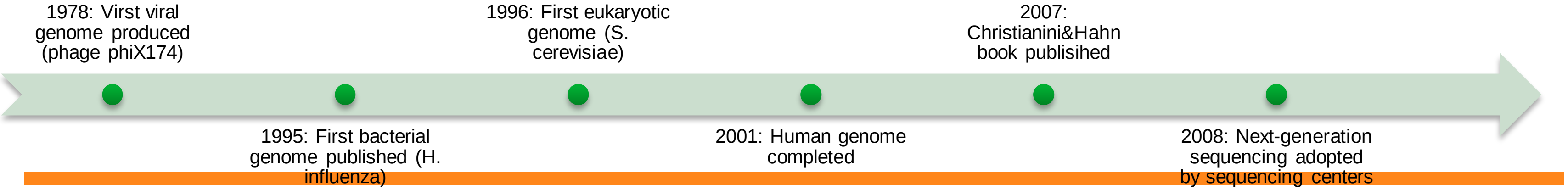
MBE MOLECULAR BIOLOGY AND EVOLUTION

The genomic explosion



Christianini&Hahn book published





1978: Virst viral genome produced (phage phiX174)

1995: First bacterial genome published (H. influenza)

1996: First eukaryotic genome (S. cerevisiae)

2001: Human genome completed

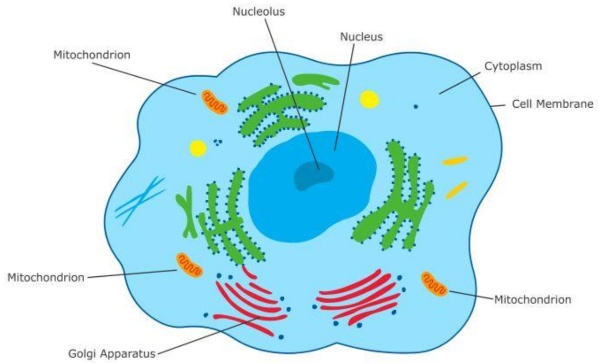
2007: Christianini&Hahn book publisihed

2008: Next-generation sequencing adopted by sequencing centers

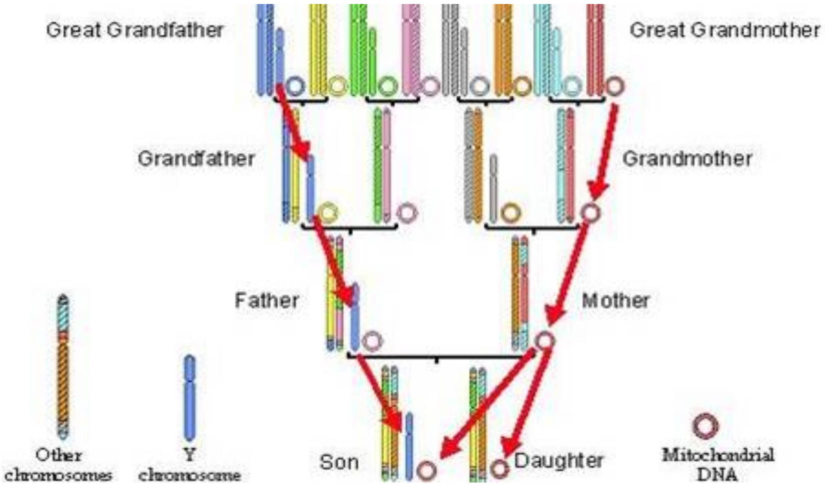
Aalto University
School of Science

Organelle genomes

- In eukaryotic organisms, not all DNA resides within the nucleus
- In addition, organelles contain their own DNA
  - Mitochondria (in most eukaryotes)
  - Plastids (in plants and algae)
- The organelle DNA is replicated independently from the nuclear DNA
  - significance in human genetics studies as it is only inherited from mother



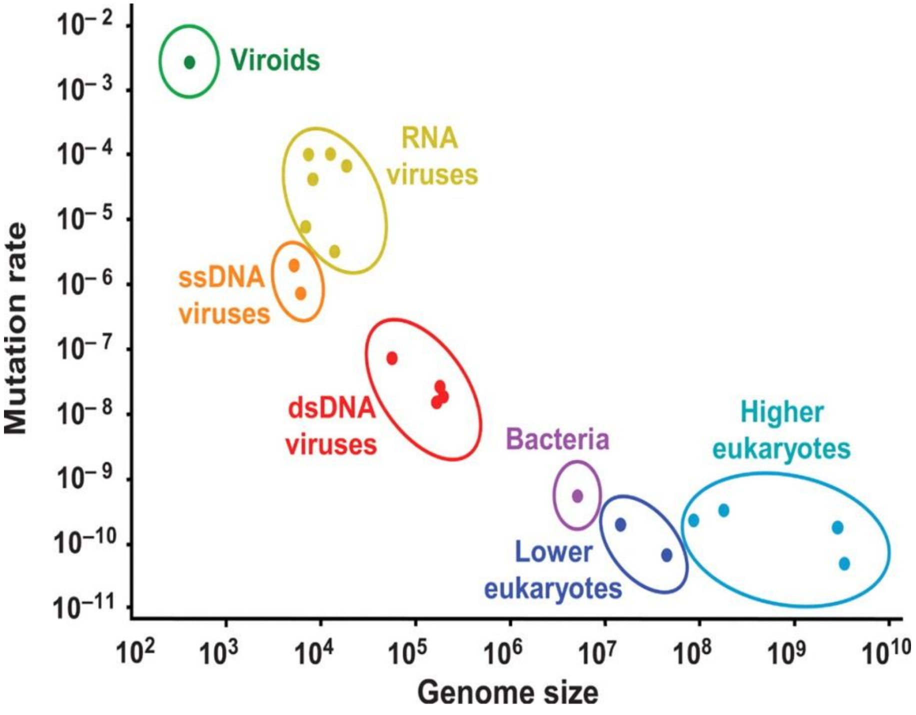
© 2007-2010 The University of Waikato I www.sciencelearn.org.nz



Aalto University
School of Science

Viral genomes

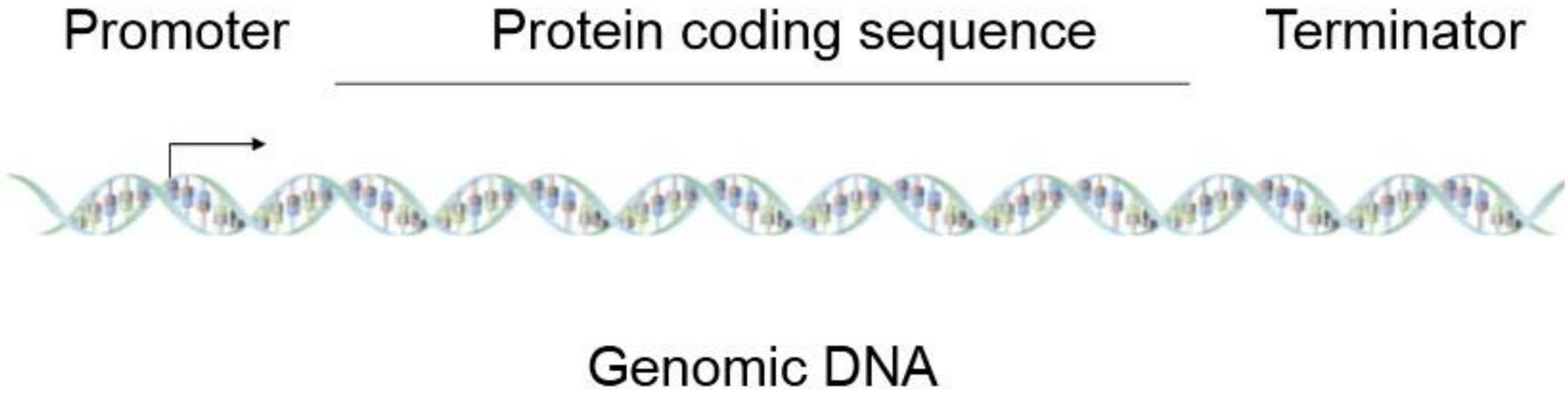
- Viruses are infectious agents that rely on living cells for replication
  - Much smaller genomes and much faster mutation rates than cellular organisms
- Viruses consist of 2 or 3 parts:
  i. the genetic material made from either DNA or RNA
  ii. a protein coat that protects these genes
  iii. in some cases also an envelope of lipids that surrounds the protein coat when they are outside a cell.
- Currently 9,228 viruses have been sequenced (Sep. 3, 2019, NCBI Viral Genome Browser)



Selma Gago, Santiago F. Elena, Ricardo Flores, and Rafael Sanjuán
Science 6 March 2009: **323** (5919), 1308
http://www.sciencemag.org/content/323/5919/1308/F1.expansion.html

Genes

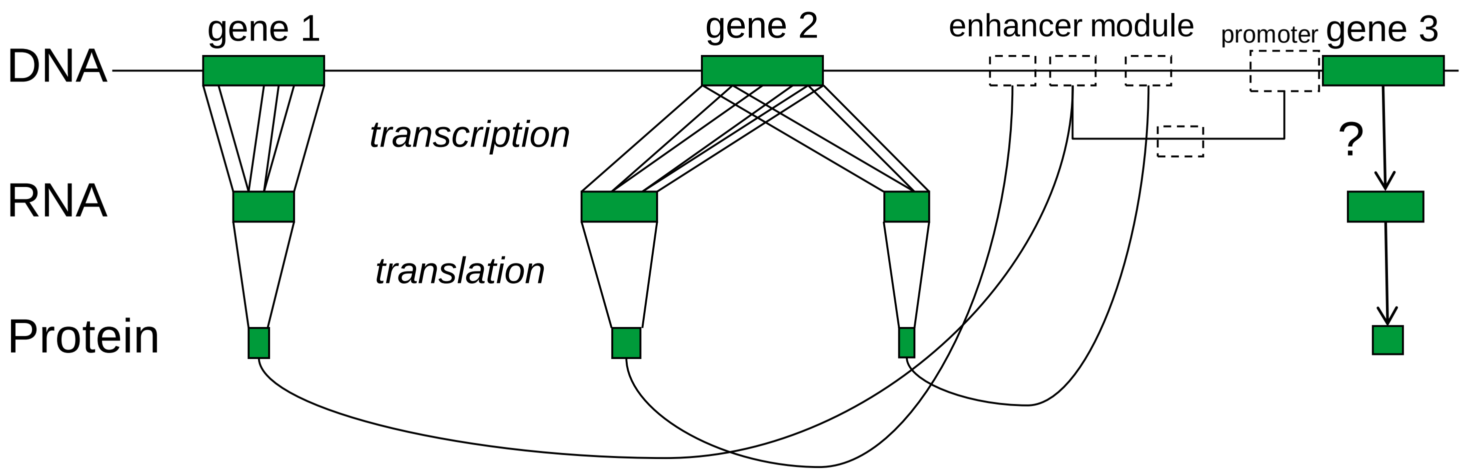
- What is a gene?



Gene structure

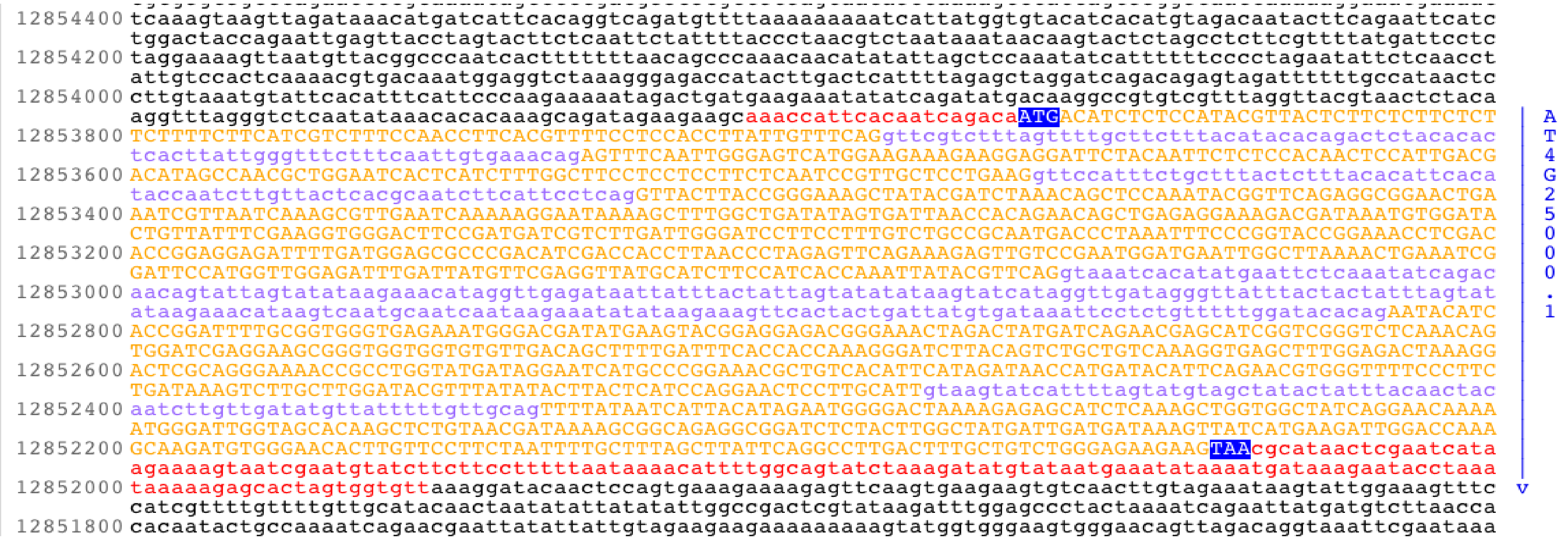
- **Genes**
  - start and stop codons
  - Introns and exons (in eukaryotic organisms)

- **Promoter regions**
  - binding sites for regulatory proteins



Typical eukaryotic gene

- ATG –start codon, TAA –stop codon
- yellow: exons, blue: introns, red: untranslated region
- black: upstream (promoter) and downstream regions

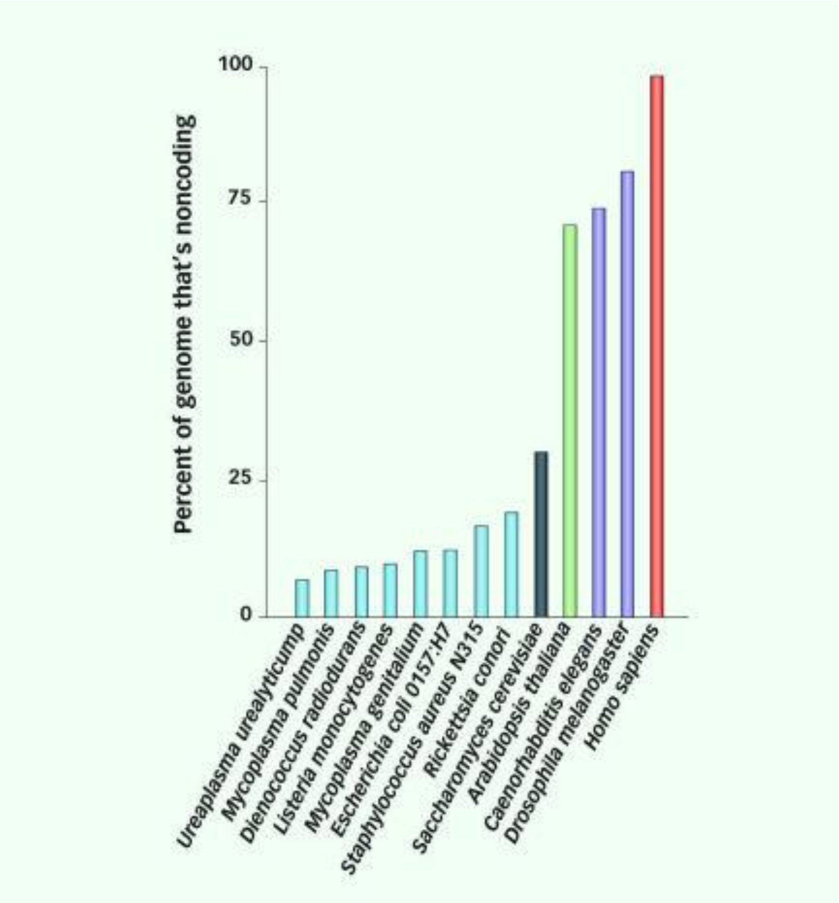


```
12854400 tcaaagtaagttagataaacatgatcattcacaggtcagatgttttaaaaaaaaatcattatggtgtacatcacatgtagacaatacttcagaattcatc
         tggactaccagaattgagttacctagtacttctcaattctattttaccctaacgtctaataaataacaagtactctagcctcttcgttttatgattcctc
12854200 taggaaaagttaatgttacggcccaatcactttttttaacagcccaaacaacatatattagctccaaatatcatttttttcccctagaatattctcaacct
         attgtccactcaaaacgtgacaaatggaggtctaaagggagaccatacttgactcattttagagctaggatcagacagagtagattttttgccataactc
12854000 cttgtaaatgtattcacatttcattcccaagaaaaatagactgatgaagaaatatatcagatatgacaaggccgtgtcgtttaggttacgtaactctaca
         aggtttagggtctcaatataaacacacaaagcagatagaagaagcaaaccattcacaatcagacaATGACATCTCTCCATACGTTACTCTTCTCTTCTCT
12853800 TCTTTTCTTCATCGTCTTTCCAACCTTCACGTTTTCCTCCACCTTATTGTTTCAGgttcgtctttagttttgcttctttacatacacagactctacacac
         tcacttattgggtttctttcaattgtgaaacagAGTTTCAATTGGGAGTCATGGAAGAAAGAAGGAGGATTCTACAATTCTCTCCACAACTCCATTGACG
12853600 ACATAGCCAACGCTGGAATCACTCATCTTTGGCTTCCTCCTCCTTCTCAATCCGTTGCTCCTGAAGgttccatttctgctttactctttacacattcaca
         taccaatcttgttactcacgcaatcttcattcctcagGTTACTTACCGGGAAAGCTATACGATCTAAACAGCTCCAAATACGGTTCAGAGGCGGAACTGA
12853400 AATCGTTAATCAAAGCGTTGAATCAAAAAGGAATAAAAGCTTTGGCTGATATAGTGATTAACCACAGAACAGCTGAGAGGAAAGACGATAAATGTGGATA
         CTGTTATTTCGAAGGTGGGACTTCCGATGATCGTCTTGATTGGGATCCTTCCTTTGTCTGCCGCAATGACCCTAAATTTCCCGGTACCGGAAACCTCGAC
12853200 ACCGGAGGAGATTTTGATGGAGCGCCCGACATCGACCACCTTAACCCTAGAGTTCAGAAAGAGTTGTCCGAATGGATGAATTGGCTTAAAACTGAAATCG
         GATTCCATGGTTGGAGATTTGATTATGTTCGAGGTTATGCATCTTCCATCACCAAATTATACGTTCAGgtaaatcacatatgaattctcaaatatcagac
12853000 aacagtattagtatataagaaacataggttgagataattatttactattagtatatataagtatcataggttgatagggttatttactactatttagtat
         ataagaaacataagtcaatgcaatcaataagaaatatataagaaagttcactactgattatgtgataaattcctctgtttttggatacacagAATACATC
12852800 ACCGGATTTTGCGGTGGGTGAGAAATGGGACGATATGAAGTACGGAGGAGACGGGAAACTAGACTATGATCAGAACGAGCATCGGTCGGGTCTCAAACAG
         TGGATCGAGGAAGCGGGTGGTGGTGTGTTGACAGCTTTTGATTTCACCACCAAAGGGATCTTACAGTCTGCTGTCAAAGGTGAGCTTTGGAGACTAAAGG
12852600 ACTCGCAGGGAAAACCGCCTGGTATGATAGGAATCATGCCCGGAAACGCTGTCACATTCATAGATAACCATGATACATTCAGAACGTGGGTTTTCCCTTC
         TGATAAAGTCTTGCTTGGATACGTTTATATACTTACTCATCCAGGAACTCCTTGCATTgtaagtatcattttagtatgtagctatactatttacaactac
12852400 aatcttgttgatatgttattttttgttgcagTTTATAATCATTACATAGAATGGGGACTAAAAGAGAGCATCTCAAAGCTGGTGGCTATCAGGAACAAAA
         ATGGGATTGGTAGCACAAGCTCTGTAACGATAAAAGCGGCAGAGGCGGATCTCTACTTGGCTATGATTGATGATAAAGTTATCATGAAGATTGGACCAAA
12852200 GCAAGATGTGGGAACACTTGTTCCTTCTAATTTTGCTTTAGCTTATTCAGGCCTTGACTTTGCTGTCTGGGAGAAGAAGTAAcgcataactcgaatcata
         agaaaagtaatcgaatgtatcttcttcctttttaataaaacatttggcagtatctaaagatatgtataatgaaatataaaatgataaagaatacctaaa
12852000 taaaaagagcactagtggtgttaaaggatacaactccagtgaaagaaaagagttcaagtgaagaagtgtcaacttgtagaaataagtattggaaagtttc
         catcgttttgttttgttgcatacaactaatatattatatattggccgactcgtataagatttggagccctactaaaatcagaattatgatgtcttaacca
12851800 cacaatactgccaaaatcagaacgaattatattattgtagaagaagaaaaaaaaagtatggtgggaagtgggaacagttagacaggtaaattcgaataaa
```

http://en.wikipedia.org/wiki/File:AMY1gene.png

Non-coding DNA

- Non-coding DNA includes all segments of the genome that do not get translated to proteins
- In higher organisms, most of the DNA is non-coding
  - In humans, over 98% of the genome is non-coding



Types of non-coding DNA

Noncoding functional RNA, RNA genes

- Functional RNA molecules that are not translated into protein.

Introns

- Regions inside the coding region that are not transcribed into mRNA
- Common in higher organisms

Regulatory elements

- Binding sites of special proteins called transcription factors
- Typically within in the promotor region of the gene or within the introns
- Carry important function

Pseudogenes

- Genes that have lost their protein coding ability
- Thought to be non-functional

Repeat sequences

- Simple repeats, CpG islands
- DNA satellites
- Mobile sequences (transposons)
- Possible role in epigenetics

‘Junk DNA’

- DNA with no function
- Open question: How much of that is there?

Sequence statistics

DNA sequences formally

- Alphabet of nucleotide symbols: $\aleph = \{A, C, G, T\}$
- DNA sequence: $s = s_1 s_2 \ldots s_n \in \aleph^n$
- A Genome is a set of DNA sequences
- **Subsequence** $s(K) = s_{k_1} s_{k_2} \ldots s_{k_r}$ collects the elements inside the index set $K = (k_1, k_2, \ldots, k_r)$
- **(Sub)string** is a *contiguous (sub)sequence*, we use shorthand $K(i{:}j) = (i, i+1, \ldots, j-1, j)$ for accessing substrings
- Example: s = ATATGTCGTGCA,
  - s(3:6) = ATGT is both a subsequence and a substring of s
  - s(8,10) = GG is a subsequence but not a substring of s

Aalto University
School of Science

Other alphabets

- RNA alphabet

$$\mathcal{N}_{RNA} = \{A, C, G, U\}$$

- Amino acid alphabet (20 standard amino acids)

$$\mathcal{A} = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$$

- Codon alphabet

$$\mathcal{C} = \{AAA, \ldots, TTT\}$$

- When the alphabet does not matter, e.g. the method can use any alphabet, we use a generic symbol $\sum$
- $\sum^n$ denotes the set of strings of length n from alphabet $\sum$

Aalto University
School of Science

Multinomial sequence model

- The simplest model for DNA sequences
- Assumes that nucleotides appear independently from each other and with a fixed probability, according to a given distribution (i.i.d assumption)

$$p = (p_A, p_C, p_G, p_T)$$

- The probability of observing a nucleotide x on position i in sequence s is independent of the position

$$p_x = p(\mathbf{s}(i) = x)$$

- Probability of a sequence s is obtained by multiplying the observed nucleotide probabilities

$$P(s) = \prod_{i=1}^{n} p(\mathbf{s}(i)) = \prod_{x \in \mathcal{N}} p_x^{n(x,s)}$$

where n(x,s) denotes the number of occurrences of x in s

Aalto University
School of Science

Uses of probabilistic sequence models

- Modeling DNA with a random i.i.d model may not always seem appropriate
- However, comparing observed data against the expectation given by a suitable random model may be very useful.
  - For instance, if the nucleotide distribution of a genomic region deviates from the expected distribution given by the model, this may mean that the region contains some elements of biological significance

Example: GC content

- The frequency of G and C bases or GC content

$$GC(s) = (n(G,s) + n(C,s))/n$$

is a simple statistics for describing genomes

  - One value is enough to characterize all nucleotide frequencies n(A,s)/n, n(C,s)/n, n(G,s)/n and n(T,s)/n for double stranded DNA.
  - Why?
    - The content of G and C is often very similar (just like the content of A and T)
    - The sum of all four frequencies has to be 1.

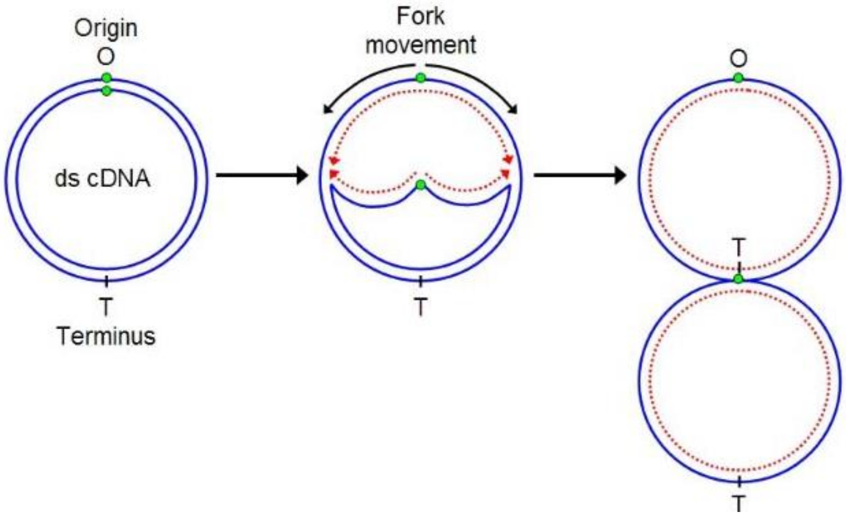
- Potential uses for GC content
  - Tell the difference between genomes of different organisms
  - Tell the difference between coding and non-coding regions

GC content and genome sizes (in megabasepairs, Mb) for various organisms

Mycoplasma genitalium	31.6%	0.585
Escherichia coli K-12	50.7%	4.693
Pseudomonas aeruginosa PAO1	66.4%	6.264
Pyrococcus abyssi	44.6%	1.765
Thermoplasma volcanium	39.9%	1.585
Caenorhabditis elegans	36%	97
Arabidopsis thaliana	35%	125
Homo sapiens	41%	3080

DNA replication fork

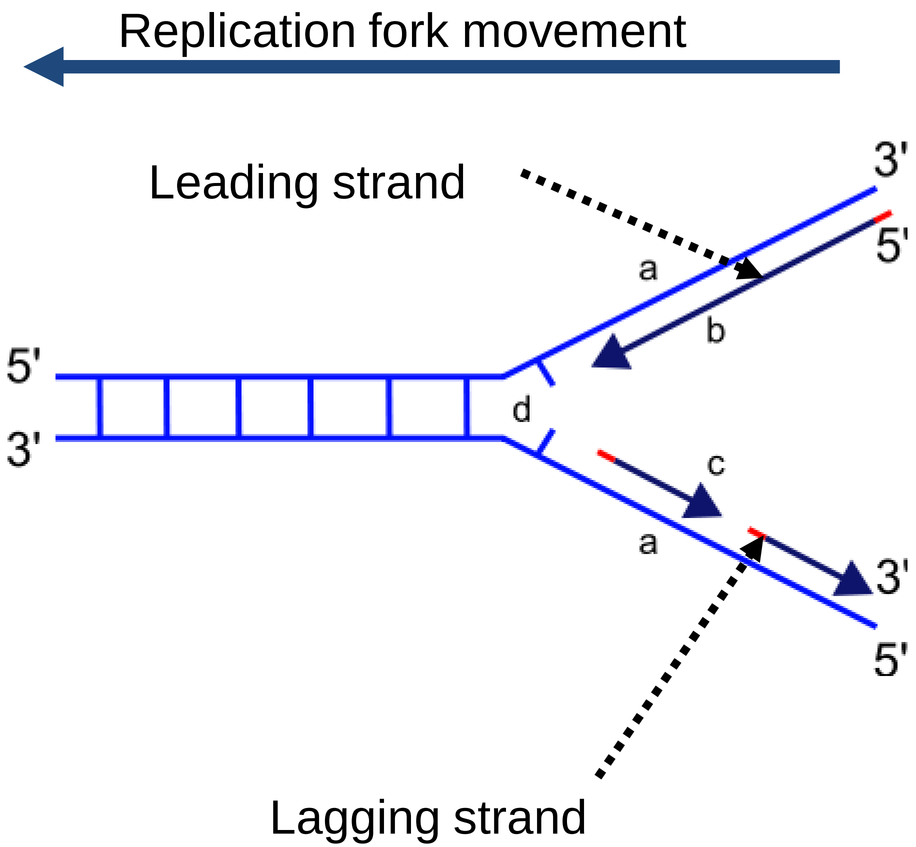
- When DNA is replicated, the molecule takes the *replication fork* form
- New complementary DNA is synthesised at both strands of the ”fork”
- This process has specific starting points in genome (*origins of replication*)



http://cronodon.com/BioTech

DNA replication fork

- New strand in 5’-3’ direction corresponding to replication fork movement is called *leading strand* and the other *lagging strand*
- Observation: leading strand is enriched in Guanine (G) and Thymine (T)
- This can be described by *GC skew* statistics



Replication fork movement

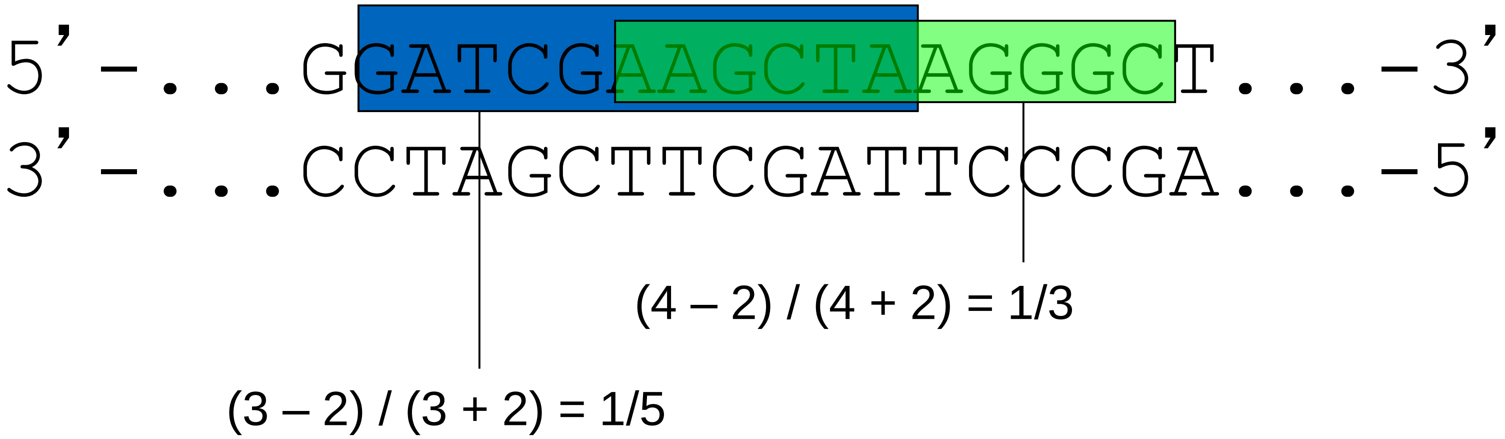
Leading strand

Lagging strand

Replication fork

GC skew

- GC skew is defined as (#G - #C) / (#G + #C)
- It is calculated at successive positions in intervals (windows) of specific width



5’-...GGATCGAAGCTAAGGGCT...-3’

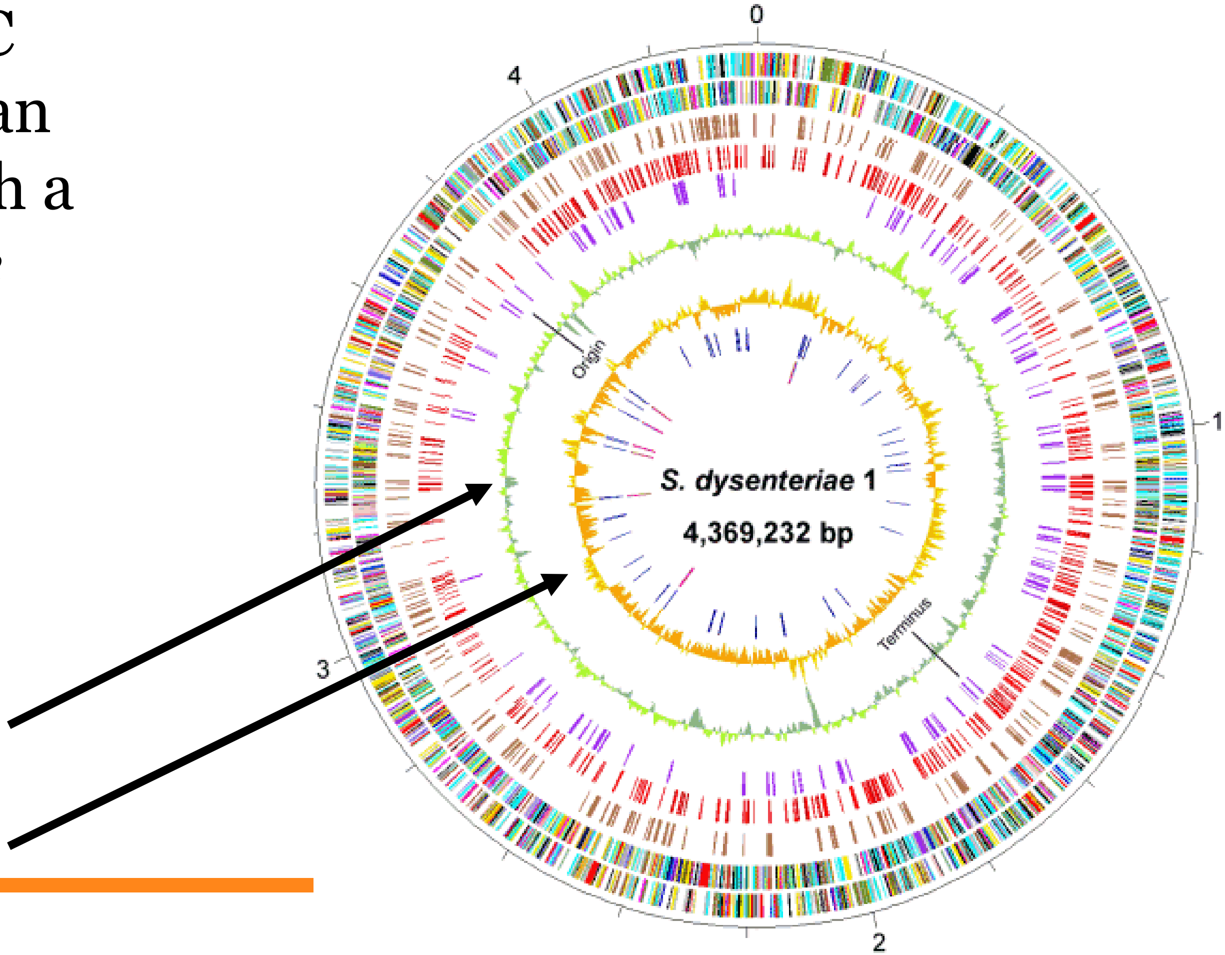
3’-...CCTAGCTTCGATTCCCGA...-5’

(4 – 2) / (4 + 2) = 1/3

(3 – 2) / (3 + 2) = 1/5

GC content & GC skew

- GC content & GC skew statistics can be displayed with a *circular genome map*



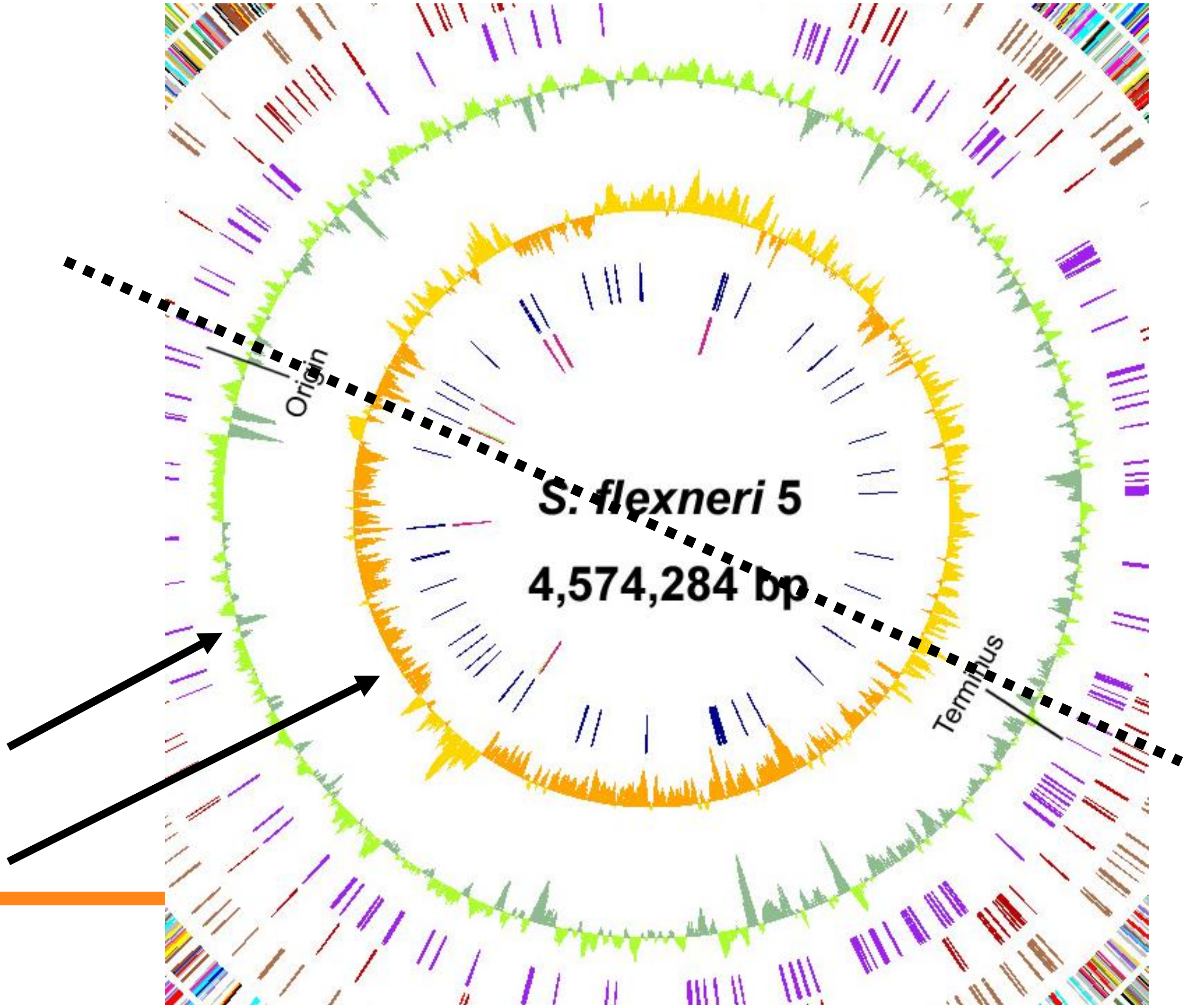
GC content

GC skew
(10kb window size)

Chromosome map of *S. dysenteriae*, the nine rings describe different properties of the genome
http://www.mgc.ac.cn/ShiBASE/circular_Sd197.htm

GC skew

- GC skew often changes sign at origin and terminus of replication



G+C content

GC skew
(10kb window size)

Nie et al., BMC Genomics, 2006

Aalto University
School of Science

Refining the i.i.d. model

- i.i.d. model describes some organisms well but fails to characterize many others
- We can refine the model by defining probabilities of k-mers, substrings of k bases
  - 1-mers: individual nucleotides (bases) – our i.i.d model!
  - 2-mers: dinucleotides (AA, AC, AG, AT, CA, ...)
  - 3-mers: codons (AAA, AAC, ...)
  - 4-mers and beyond

Over- and underrepresented k-mers

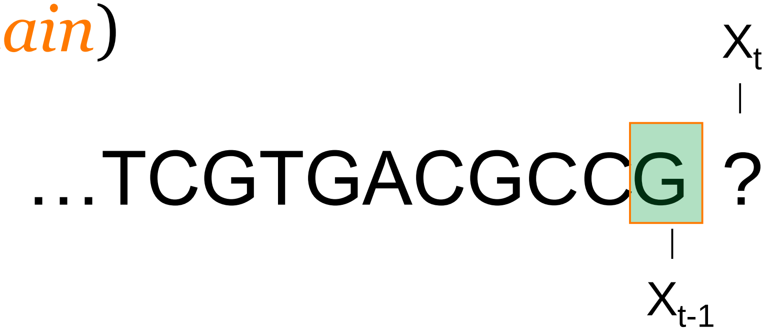
- A simple and useful way to find interesting sections of DNA is to compute the level of over- or under-representation of a k-mer in a sequence
- Compare the frequency of the k-mer against the expected frequency if the k-mer is a random combination of l-mers, where 1<l<k
- Odds ratio is a typical measure: for a dinucleotide AG

$$oddsratio = \frac{fr(AG,s)}{fr(A,s)fr(G,s)}$$

- fr(X,s) = n(X,s)/n is the (relative) frequency of X in s
- If the sequence has been generated by a multinomial model, the ratio should be 1
- Any significant deviation from 1 signals the fact that ’AG’ is either over or under represented
  - This might indicate that ‘AG’ may have biological significance in sequence s

First-order Markov chains

- Let’s assume that in sequence X the letter at position t, X_t, depends only on the previous letter X_{t-1} (*first-order markov chain*)



X_t

…TCGTGACGCCG ?

X_{t-1}

- Probability of letter b occuring at position t given X_{t-1} = a is $p_{ab} = P(X_t = b \mid X_{t-1} = a)$
- We consider *homogeneous* markov chains: probability p_{ab} is independent of position t

Estimating p_{ab}

- We can estimate conditional probabilities p_{ab} (”the probability that b follows a”) from observed dinucleotide frequencies fr_{ab} (≈joint probabilities)

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

Frequency of dinucleotide AT in sequence

Base frequency π(C)

…the values p_{AA}, p_{AC}, ..., p_{TG}, p_{TT} sum to 1

Aalto University
School of Science

Estimating p_{ab}

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

Probability of transition a -> b

Base frequency of nucleotide a, π(a)

Dinucleotide frequency

The base frequencies are: π=(0.345, 0.158, 0.159, 0.337)

0.052 / 0.345 ≈ 0.151

	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	0.028	0.051
T	0.087	0.047	0.063	0.140

$P(X_t = b, X_{t-1} = a)$

	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

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

Simulating a DNA sequence

- From a transition matrix, it is easy to generate a DNA sequence of length n:
  - First, choose the starting base randomly according to the base frequency distribution π=(0.345, 0.158, 0.159, 0.337)
  - Then, choose next base according to the distribution $P(x_t \mid x_{t-1})$ until n bases have been chosen

T T C T T C A 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

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

Simulating a DNA sequence

- Now we can quickly generate sequences of arbitrary length…

```
ttcttcaaaataaggatagtgattcttattggcttaagggataacaatttagatcttttttcatgaatcatgtatgtcaacgttaaaagttgaactgcaataagttc
ttacacacgattgtttatctgcgtgcgaagcatttcactacatttgccgatgcagccaaaagtatttaacatttggtaaacaaattgacttaaatcgcgcacttaga
gtttgacgtttcatagttgatgcgtgtctaacaattacttttagtttttaaatgcgtttgtctacaatcattaatcagctctggaaaaacattaatgcatttaaac
cacaatggataattagttacttattttaaaattcacaaagtaattattcgaatagtgccctaagagagtactggggttaatggcaaagaaaattactgtagtgaaga
ttaagcctgttattatcacctgggtactctggtgaatgcacataagcaaatgctacttcagtgtcaaagcaaaaaaatttactgataggactaaaaaccctttattt
ttagaatttgtaaaaatgtgacctcttgcttataacatcatatttattgggtcgttctaggacactgtgattgccttctaactcttatttagcaaaaaattgtcata
gctttgaggtcagacaaacaagtgaatggaagacagaaaaagctcagcctagaattagcatgttttgagtggggaattacttggttaactaaagtgttcatgactgt
tcagcatatgattgttggtgagcactacaaagatagaagagttaaactaggtagtggtgatttcgctaacacagttttcatacaagttctattttctcaatggtttt
ggataagaaaacagcaaacaaatttagtattattttcctagtaaaagcaaacatcaaggagaaattggaagctgcttgttcagtttgcattaaattaaaaatttat
ttgaagtattcgagcaatgttgacagtctgcgttcttcaaataagcagcaaatcccctcaaaattgggcaaaaacctaccctggcttcttttaaaaaaccaagaaa
agtcctatataagcaacaaatttcaaaccttttgttaaaaattctgctgctgaataaataggcattacagcaatgcaattaggtgcaaaaaaggccatcctctttct
tttttgtacaattgttcaagcaactttgaatttgcagattttaacccactgtctatatgggacttcgaattaaattgactggtctgcatcacaaatttcaactgcc
caatgtaatcatattctagagtattaaaaatacaaaaagtacaattagttatgcccattggcctggcaatttatttactccactttccacgttttggggatattta
acttgaatagttcacaatcaaaacataggaaggatctactgctaaaagcaaaagcgtattggaatgataaaaaactttgatgtttaaaaaactacaaccttaatgaa
ttaaagttgaaaaaatattcaaaaaaagaaattcagttcttggcgagtaatattttgatgtttgagatcagggttacaaaataagtgcatgagattaactcttcaa
atataaactgatttaagtgtatttgctaataacattttcgaaaaggaatattatggtaagaattcataaaaatgtttaatactgatacaactttcttttatatcctc
catttggccagaatactgttgcacacaactaattggaaaaaaaatagaacgggtcaatctcagtgggaggagaagaaaaaagttggtgcaggaaatagtttctacta
acctggtataaaaacatcaagtaacattcaaattgcaaatgaaaactaaccgatctaagcattgattgatttttctcatgcctttcgcctagttttaataaacgcgc
cccaactctcatcttcggttcaaatgatctattgtatttatgcactaacgtgcttttatgttagcatttttcaccctgaagttccgagtcattggcgtcactcacaa
atgacattacaatttttctatgttttgttctgttgagtcaaagtgcatgcctacaattctttcttatatagaactagacaaaatagaaaaaggcacttttggagtct
gaatgtcccttagtttcaaaaaggaaattgttgaattttttgtggttagttaaattttgaacaaactagtatagtggtgacaaacgatcaccttgagtcggtgacta
taaaagaaaaaggagattaaaaatacctgcggtgccacattttttgttacgggcatttaaggtttgcatgtgttgagcaattgaaacctacaactcaataagtcatg
ttaagtcacttctttgaaaaaaaaaaagaccctttaagcaagctc
```

Simulating a DNA sequence

	Dinucleotide frequencies	
	Simulated	Observed
aa	0.145	0.146
ac	0.050	0.052
ag	0.055	0.058
at	0.092	0.089
ca	0.065	0.063
cc	0.028	0.029
cg	0.011	0.010
ct	0.058	0.056
ga	0.048	0.050
gc	0.032	0.030
gg	0.029	0.028
gt	0.050	0.051
ta	0.084	0.086
tc	0.052	0.047
tg	0.064	0.063
tt	0.138	0.0140

n = 10000

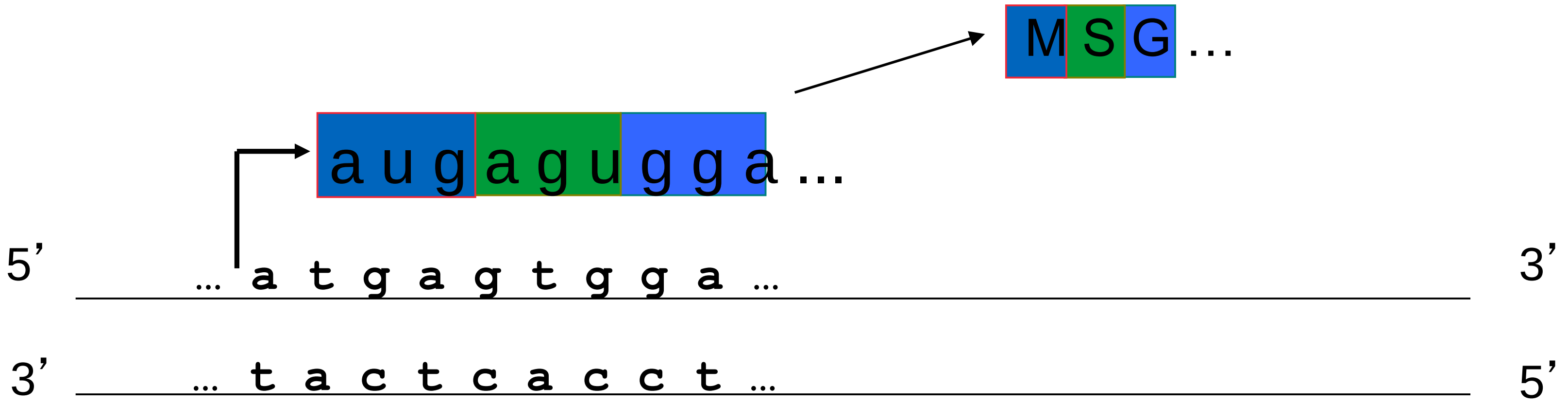
Simulating a DNA sequence

- The model is able to generate correct proportions of 1- and 2-mers in genomes…
- …but fails with k=3 and beyond.

```
ttcttcaaaataaggatagtgattcttattggcttaagggataacaatttagatcttttttcatgaatcatgtatgtcaacgttaaaagttgaactgcaataagttc
ttacacacgattgtttatctgcgtgcgaagcatttcactacatttgccgatgcagccaaaagtatttaacatttggtaaacaaattgacttaaatcgcgcacttaga
gtttgacgtttcatagttgatgcgtgtctaacaattacttttagttttttaaatgcgtttgtctacaatcattaatcagctctggaaaaacattaatgcatttaaac
cacaatggataattagttacttattttaaaattcacaaagtaattattcgaatagtgccctaagagagtactggggttaatggcaaagaaaattactgtagtgaaga
ttaagcctgttattatcacctgggtactctggtgaatgcacataagcaaatgctacttcagtgtcaaagcaaaaaaatttactgataggactaaaaaccctttattt
ttagaatttgtaaaaatgtgacctcttgcttataacatcatatttattgggtcgttctaggacactgtgattgccttctaactcttatttagcaaaaaattgtcata
gctttgaggtcagacaaacaagtgaatggaagacagaaaaagctcagcctagaattagcatgttttgagtggggaattacttggttaactaaagtgttcatgactgt
tcagcatatgattgttggtgagcactacaaagatagaagagttaaactaggtagtggtgatttcgctaacacagttttcatacaagttctattttctcaatggtttt
ggataagaaaacagcaaacaaatttagtattattttcctagtaaaaagcaaacatcaaggagaaattggaagctgcttgttcagtttgcattaaattaaaaatttat
ttgaagtattcgagcaatgttgacagtctgcgttcttcaaataagcagcaaatcccctcaaaattgggcaaaaacctaccctggcttctttttaaaaaaccaagaaa
agtcctatataagcaacaaatttcaaacctttttgttaaaaattctgctgctgaataaataggcattacagcaatgcaattaggtgcaaaaaaggccatcctctttct
ttttttgtacaattgttcaagcaactttgaatttgcagattttaacccactgtctatatgggacttcgaattaaattgactggtctgcatcacaaatttcaactgcc
caatgtaatcatattctagagtattaaaaatacaaaaagtacaattagttatgcccattggcctggcaatttatttactccactttccacgttttggggatattta
acttgaatagttcacaatcaaaacataggaaggatctactgctaaaagcaaaagcgtattggaatgataaaaaactttgatgtttaaaaaactacaaccttaatgaa
ttaaagttgaaaaaatattcaaaaaaagaaattcagttcttggcgagtaatatttttgatgtttgagatcagggttacaaaataagtgcatgagattaactcttcaa
atataaactgatttaagtgtatttgctaataacattttcgaaaaggaatattatggtaagaattcataaaaatgtttaatactgatacaactttcttttatatcctc
catttggccagaatactgttgcacacaactaattggaaaaaaaatagaacgggtcaatctcagtgggaggagaagaaaaaagttggtgcaggaaatagtttctacta
acctggtataaaaacatcaagtaacattcaaattgcaaatgaaaactaaccgatctaagcattgattgatttttctcatgcctttcgcctagttttaataaacgcgc
cccaactctcatcttcggttcaaatgatctattgtatttatgcactaacgtgcttttatgttagcatttttcaccctgaagttccgagtcattggcgtcactcacaa
atgacattacaatttttctatgttttgttctgttgagtcaaagtgcatgcctacaattctttcttatatagaactagacaaaatagaaaaaggcacttttggagtct
gaatgtcccttagtttcaaaaaggaaattgttgaattttttgtggttagttaaattttgaacaaactagtatagtggtgacaaacgatcaccttgagtcggtgacta
taaaagaaaaaggagattaaaaatacctgcggtgccacattttttgttacgggcatttaaggtttgcatgtgttgagcaattgaaacctacaactcaataagtcatg
ttaagtcacttctttgaaaaaaaaaaagaccctttaagcaagctc
```

3-mers: codons

- We can extend the previous method to 3-mers
- k=3 is an important case in study of DNA sequences because of genetic code



3-mers in Escherichia coli genome

Word	Count	Observed	Expected
AAA	108924	**0.02348**	**0.01492**
AAC	82582	0.01780	0.01541
AAG	63369	0.01366	0.01537
AAT	82995	0.01789	0.01490
ACA	58637	0.01264	0.01541
ACC	74897	0.01614	0.01591
ACG	73263	0.01579	0.01588
ACT	49865	0.01075	0.01539
AGA	56621	0.01220	0.01537
AGC	80860	0.01743	0.01588
AGG	50624	0.01091	0.01584
AGT	49772	0.01073	0.01536
ATA	63697	0.01373	0.01490
ATC	86486	0.01864	0.01539
ATG	76238	0.01643	0.01536
ATT	83398	0.01797	0.01489

Word	Count	Observed	Expected
CAA	76614	0.01651	0.01541
CAC	66751	0.01439	0.01591
CAG	104799	0.02259	0.01588
CAT	76985	0.01659	0.01539
CCA	86436	0.01863	0.01591
CCC	47775	0.01030	0.01643
CCG	87036	0.01876	0.01640
CCT	50426	0.01087	0.01589
CGA	70938	0.01529	0.01588
CGC	115695	**0.02494**	**0.01640**
CGG	86877	0.01872	0.01636
CGT	73160	0.01577	0.01586
CTA	26764	**0.00577**	**0.01539**
CTC	42733	0.00921	0.01589
CTG	102909	0.02218	0.01586
CTT	63655	0.01372	0.01537

3-mers in Escherichia coli genome

Word	Count	Observed	Expected
GAA	83494	0.01800	0.01537
GAC	54737	0.01180	0.01588
GAG	42465	0.00915	0.01584
GAT	86551	0.01865	0.01536
GCA	96028	0.02070	0.01588
GCC	92973	0.02004	0.01640
GCG	114632	**0.02471**	**0.01636**
GCT	80298	0.01731	0.01586
GGA	56197	0.01211	0.01584
GGC	92144	0.01986	0.01636
GGG	47495	0.01024	0.01632
GGT	74301	0.01601	0.01582
GTA	52672	0.01135	0.01536
GTC	54221	0.01169	0.01586
GTG	66117	0.01425	0.01582
GTT	82598	0.01780	0.01534

Word	Count	Observed	Expected
TAA	68838	0.01484	0.01490
TAC	52592	0.01134	0.01539
TAG	27243	**0.00587**	**0.01536**
TAT	63288	0.01364	0.01489
TCA	84048	0.01812	0.01539
TCC	56028	0.01208	0.01589
TCG	71739	0.01546	0.01586
TCT	55472	0.01196	0.01537
TGA	83491	0.01800	0.01536
TGC	95232	0.02053	0.01586
TGG	85141	0.01835	0.01582
TGT	58375	0.01258	0.01534
TTA	68828	0.01483	0.01489
TTC	83848	0.01807	0.01537
TTG	76975	0.01659	0.01534
TTT	109831	0.02367	0.01487

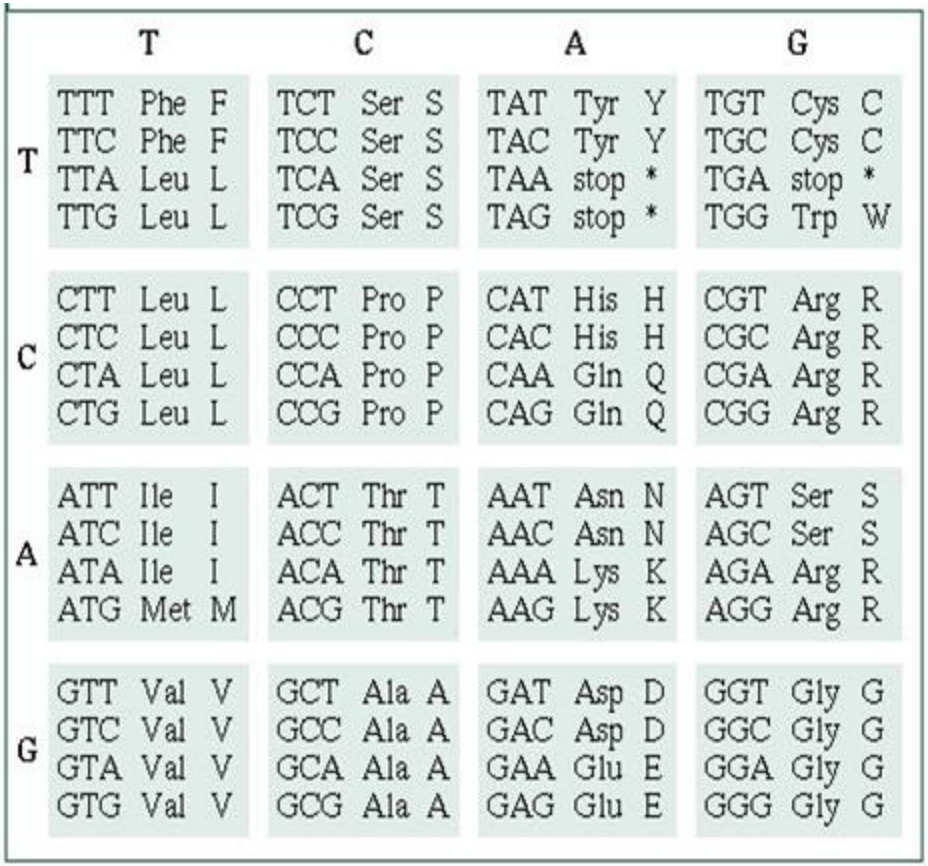
2nd order Markov Chains

- Markov chains readily generalise to higher orders
- In 2nd order markov chain, position t depends on positions t-1 and t-2
- Transition matrix:

	A	C	G	T
AA				
AC				
AG				
AT				
CA				
…				

Codon translation table

- 61 codons that specify amino acids and three stop codons.
- ATG which encodes Methionine (M) is the start codon
- There are 20 common amino acids => most amino acids are specified by more than one codon.
- This has led to the use of a number of statistics to summarize the “bias” in codon usage.



	T	C	A	G
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys C
	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys C
	TTA Leu L	TCA Ser S	TAA stop *	TGA stop *
	TTG Leu L	TCG Ser S	TAG stop *	TGG Trp W
C	CTT Leu L	CCT Pro P	CAT His H	CGT Arg R
	CTC Leu L	CCC Pro P	CAC His H	CGC Arg R
	CTA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R
	CTG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R
A	ATT Ile I	ACT Thr T	AAT Asn N	AGT Ser S
	ATC Ile I	ACC Thr T	AAC Asn N	AGC Ser S
	ATA Ile I	ACA Thr T	AAA Lys K	AGA Arg R
	ATG Met M	ACG Thr T	AAG Lys K	AGG Arg R
G	GTT Val V	GCT Ala A	GAT Asp D	GGT Gly G
	GTC Val V	GCC Ala A	GAC Asp D	GGC Gly G
	GTA Val V	GCA Ala A	GAA Glu E	GGA Gly G
	GTG Val V	GCG Ala A	GAG Glu E	GGG Gly G

Codon Adaptation Index (CAI)

- CAI compares the distribution of codons in a given gene with the preferred codons in a reference set of genes, usually highly expressed genes.
- Observation: cells prefer certain codons in highly expressed genes

Moderately expressed

Amino acid	Codon	Predicted	Gene class I	Gene class II
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

Highly expressed

Codon frequencies for some genes in E. coli

Codon Adaptation Index (CAI)

- Consider an amino acid sequence $X = x_1x_2 \ldots x_n$ where x_k represents the amino acid residue corresponding to codon k in the gene.
- Let p_k be the probability that codon k is used to code amino acid x_k in highly expressed genes
- Let q_k be the highest probability of codons coding the same amino acid in highly expressed genes
  - For example, if codon k is ”GCC”, the corresponding amino acid is Alanine (see genetic code table; also GCT, GCA, GCG code for Alanine)
  - Assume that $p_{GCC} = 0.164$, $p_{GCT} = 0.275$, $p_{GCA} = 0.240$, $p_{GCG} =$ **0.323**
  - Now $q_{GCC} = q_{GCT} = q_{GCA} = q_{GCG} =$ **0.323**

Codon Adaptation Index (CAI)

- CAI is defined as

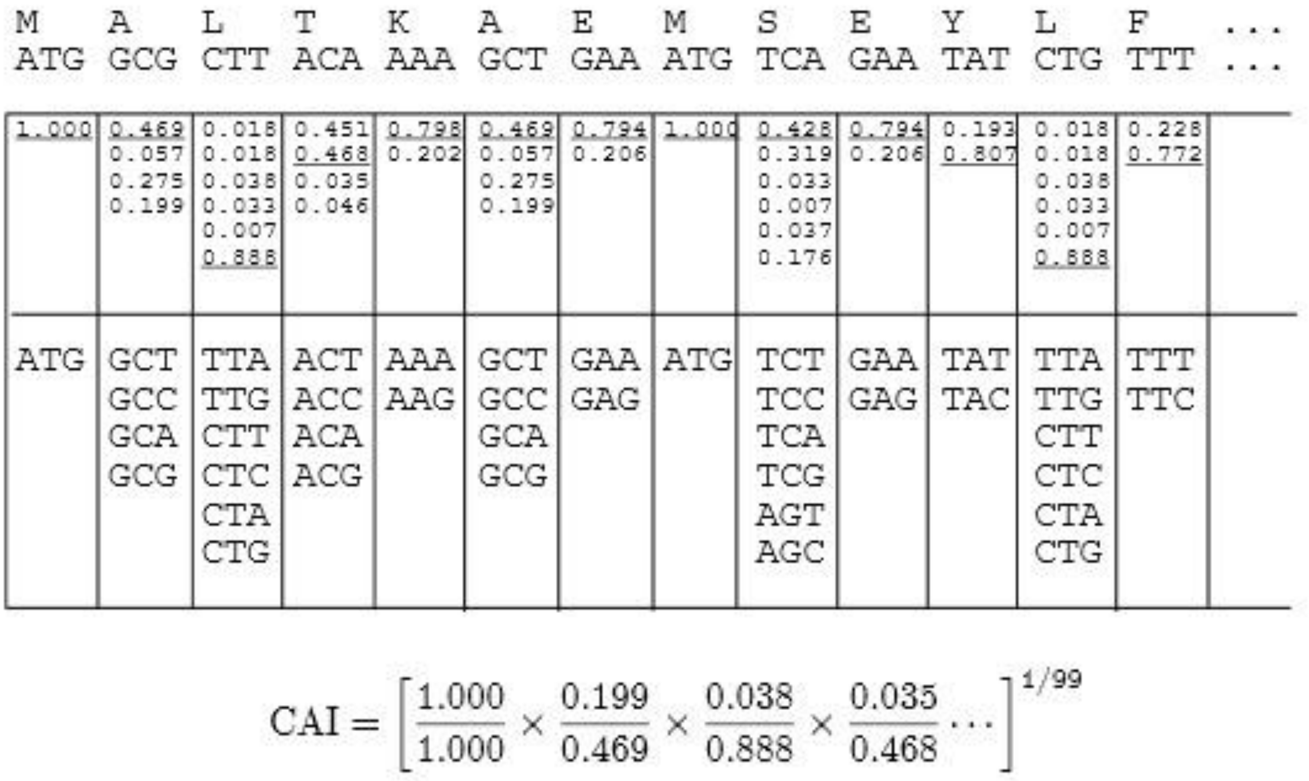
$$\mathrm{CAI} = \left(\prod_{k=1}^{n} p_k / q_k\right)^{1/n}$$

- CAI can be given also in *log-odds* form
  - Log-odds used to avoid numerical problems:

$$\log(\mathrm{CAI}) = (1/n) \sum_{k=1}^{n} \log(p_k / q_k)$$

CAI: example with an E. coli gene

The amino acid sequence from the amino terminal end of the himA gene of E. coli. Below are the probabilities of the different codons for the same amino acid, and the corresponding codons. The maximum probabilities (the q_k) are underlined.



M	A	L	T	K	A	E	M	S	E	Y	L	F	...
ATG	GCG	CTT	ACA	AAA	GCT	GAA	ATG	TCA	GAA	TAT	CTG	TTT	...
1.000	0.469 0.057 0.275 0.199	0.018 0.018 0.038 0.033 0.007 0.888	0.451 0.468 0.035 0.046	0.798 0.202	0.469 0.057 0.275 0.199	0.794 0.206	1.000	0.428 0.319 0.033 0.007 0.037 0.176	0.794 0.206	0.193 0.807	0.018 0.018 0.038 0.033 0.007 0.888	0.228 0.772	
ATG	GCT GCC GCA GCG	TTA TTG CTT CTC CTA CTG	ACT ACC ACA ACG	AAA AAG	GCT GCC GCA GCG	GAA GAG	ATG	TCT TCC TCA TCG AGT AGC	GAA GAG	TAT TAC	TTA TTG CTT CTC CTA CTG	TTT TTC	

$$\mathrm{CAI} = \left[\frac{1.000}{1.000} \times \frac{0.199}{0.469} \times \frac{0.038}{0.888} \times \frac{0.035}{0.468} \cdots\right]^{1/99}$$

Aalto University
School of Science

CAI: properties

- CAI = 1.0 : each codon in the gene under consideration was equal to the most frequently used codon in the reference set of highly expressed genes
- In a sample of E.coli genes, CAI ranged from 0.2 to 0.85
- CAI correlates with mRNA levels: it can be used to predict expression levels for new genes

Biological words: summary

- Simple 1-, 2- and 3-mer models can describe interesting properties of DNA sequences
  - GC skew can identify DNA replication origins
  - It can also reveal *genome rearrangement* events and *lateral transfer* of DNA
  - GC content can be used to locate genes: human genes are comparably GC-rich
  - CAI predicts high gene expression levels