

# Bioinformatics — Lecture 2

## Sequence analysis

(EG Ch. 5, 12; MM Ch. 12)

Krzysztof Bartoszek

Linköping University

*krzysztof.bartoszek@liu.se*

13 XI 2023 (U7)

# Today

## A DNA sequence

- Human genome

- A gene

- TATA box

- Nucleic acid codes

- TATA box

- GC content

## Sequencing

- Timeline

- Shotgun sequencing

- Poisson model

- Shortest common superstring

## Modelling DNA

- PWM

- Markov chains

- HMMs

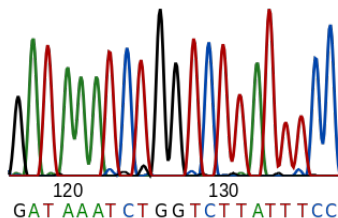
# The human genome

$2 \times 23$  chromosomes

3.23483 Giga-basepairs (on single chromosome copy),

6.46983 Giga-basepairs (on both chromosome copies),

$$\text{genome} \in \{A, C, G, T\}^{6.46983 \cdot 10^9}$$



[https://en.wikipedia.org/wiki/Nucleic\\_acid\\_sequence](https://en.wikipedia.org/wiki/Nucleic_acid_sequence) (graphic by Sjef, public domain)

# The human genome [https://en.wikipedia.org/wiki/Human\\_genome](https://en.wikipedia.org/wiki/Human_genome)

$\approx$  20000 protein coding genes

protein coding genes are  $\approx$  1.5% of genome

median size of protein coding gene 26288bp

98.5%: *non-coding* RNA, *regulatory sequences*, *introns*,  
*retrotransposon* (DNA that is transcribed into RNA, then reverse  
transcribed into DNA and can be re-inserted into genome),

*transposable elements* (can change position in genome)

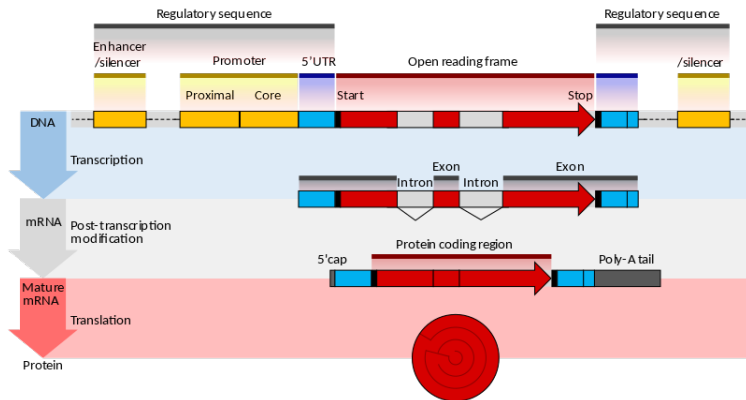
and *unknown* sequences

*pseudogenes* (inactive copies of gene, often generated by gene  
duplication),

*repetitive DNA sequences* (50% of genome, DNA occurring in  
multiple copies)

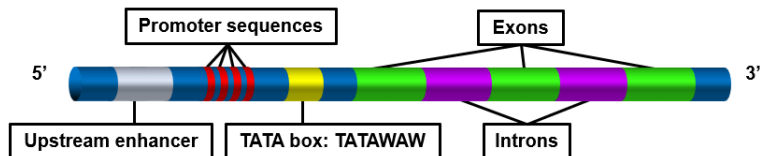
SNPs:  $\approx$  1 in 1000bp but also CNV (copy number variation)

# A gene



T. Shafee, R. Lowe (2017). "Eukaryotic and prokaryotic gene structure". WikiJournal of Medicine 4(1). DOI:10.15347/wjm/2017.002 CC BY 4.0

# TATA box (description)



[https://en.wikipedia.org/wiki/TATA\\_box](https://en.wikipedia.org/wiki/TATA_box) (graphic by Luttysar, CC BY-SA 4.0)

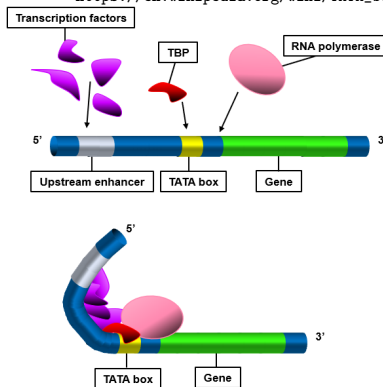
$$W \in \{A, T\}$$

# IUB/IUPAC nucleic acid codes ([https://en.wikipedia.org/wiki/FASTA\\_format](https://en.wikipedia.org/wiki/FASTA_format))

Nucleic Acid Code ↕	Meaning ↕	Mnemonic ↕
A	A	<b>A</b> denine
C	C	<b>C</b> ytosine
G	G	<b>G</b> uanine
T	T	<b>T</b> hymine
U	U	<b>U</b> racil
R	A or G	<b>p</b> u <b>R</b> ine
Y	C, T or U	<b>p</b> Yrimidines
K	G, T or U	bases which are <b>K</b> etones
M	A or C	bases with <b>a</b> Mino groups
S	C or G	<b>S</b> trong interaction
W	A, T or U	<b>W</b> eak interaction
B	not A (i.e. C, G, T or U)	<b>B</b> comes after A
D	not C (i.e. A, G, T or U)	<b>D</b> comes after C
H	not G (i.e., A, C, T or U)	<b>H</b> comes after G
V	neither T nor U (i.e. A, C or G)	<b>V</b> comes after U
N	A C G T U	<b>N</b> ucleic acid
-	gap of indeterminate length	

# TATA box (mechanism)

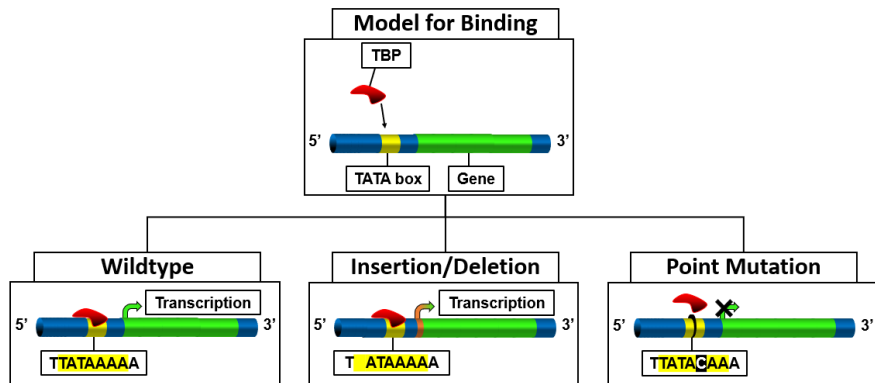
[https://en.wikipedia.org/wiki/TATA\\_box](https://en.wikipedia.org/wiki/TATA_box) (graphic by Luttysar, CC BY-SA 4.0)



TBP: TATA binding protein  
needs to bind for first step in transcription initiation



# TATA box (mutation mechanism)



[https://en.wikipedia.org/wiki/TATA\\_box](https://en.wikipedia.org/wiki/TATA_box) (graphic by Luttysar, CC BY-SA 4.0)

# CpG islands

CpG: (5'-C-phosphate-G-3')

phosphate (P) links any two nucleotides in DNA

Regions with “high frequency” of CpG sites

region: at least 200bp

$(\#C + \#G)/(|\text{sequence}|) > 50\%$  (GC%) and

observed-to-expected CpG ratio greater than 60

observed :  $\#CpG$

expected :  $\frac{\#C \cdot \#G}{|\text{sequence}|}$  or  $\frac{((\#C + \#G)/2)^2}{|\text{sequence}|}$  (\* for next slide)

many (vertebrate) genomes have CpG islands near start

(*promoter*) of transcribed region of genes

(esp. *housekeeping*: maintenance of basic cellular function)

[https://en.wikipedia.org/wiki/CpG\\_site](https://en.wikipedia.org/wiki/CpG_site)

[https://en.wikipedia.org/wiki/Housekeeping\\_gene](https://en.wikipedia.org/wiki/Housekeeping_gene)

10/42

# Human genome composition

A : 29.3%, C : 20.0%, G : 20.7%, T : 20%, GC : 40.7%  
(content is chromosome specific)

[https://en.wikipedia.org/wiki/Chargaff%27s\\_rules](https://en.wikipedia.org/wiki/Chargaff%27s_rules): %A = %T, %G = %C

rule 1: on double strand, rule 2: on both strands separately

background probability  $\text{CpG} = 0.2 \cdot 0.207 = 0.0414$

Table 1. Overview of CpG distribution in the human genome

Subset	Length, Mb	GC content	Observed CpG fraction	Normalized CpG fraction
Whole genome	3.1*	0.38	0.009	0.25
1 kb upstream regions	15	0.53	0.042	0.60
1 kb downstream regions	15	0.45	0.013	0.26
Transcription units	930	0.42	0.011	0.26
Exons	45	0.50	0.028	0.45
Introns	880	0.41	0.010	0.24

observed/expected(\*)

Length refers to the total length of DNA examined.

\*Length given in gigabases.

S. Saxonov, P. Berg, D. L. Brutlag (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters". PNAS 103(5) : 1412–1417, doi:10.1073/pnas.0510310103.

Copyright (2006) National Academy of Sciences.

# Human gene content

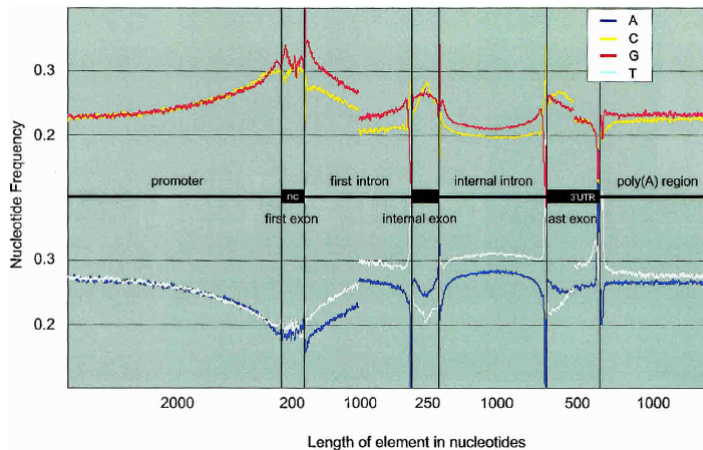


Fig. 1 in E. Louie, J. Ott, J. Majewski (2003). "Nucleotide Frequency Variation Across Human Genes". Genome Research 13 : 2594–2601. Usage permissions, CC BY-NC 4.0 .

# Genome content (GC sorted)

Organism ↕	Taxon ↕	%A ↕	%G ↕	%C ↕	%T ↕	A / T ↕	G / C ↕	%GC ▼	%AT ↕
<i>E. coli</i>	<i>Escherichia</i>	24.7	26.0	25.7	23.6	1.05	1.01	51.7	48.3
Maize	<i>Zea</i>	26.8	22.8	23.2	27.2	0.99	0.98	46.1	54.0
Wheat	<i>Triticum</i>	27.3	22.7	22.8	27.1	1.01	1.00	45.5	54.4
φX174	<i>PhiX174</i>	24.0	23.3	21.5	31.2	0.77	1.08	44.8	55.2
Chicken	<i>Gallus</i>	28.0	22.0	21.6	28.4	0.99	1.02	43.7	56.4
Rat	<i>Rattus</i>	28.6	21.4	20.5	28.4	1.01	1.00	42.9	57.0
Grasshopper	Orthoptera	29.3	20.5	20.7	29.3	1.00	0.99	41.2	58.6
Human	<i>Homo</i>	29.3	20.7	20.0	30.0	0.98	1.04	40.7	59.3
Yeast	<i>Saccharomyces</i>	31.3	18.7	17.1	32.9	0.95	1.09	35.8	64.4
Octopus	<i>Octopus</i>	33.2	17.6	17.6	31.6	1.05	1.00	35.2	64.8
Sea Urchin	<i>Echinacea</i>	32.8	17.7	17.3	32.1	1.02	1.02	35.0	64.9

[https://en.wikipedia.org/wiki/Chargaff%27s\\_rules](https://en.wikipedia.org/wiki/Chargaff%27s_rules)

# Sequencing projects historical timeline

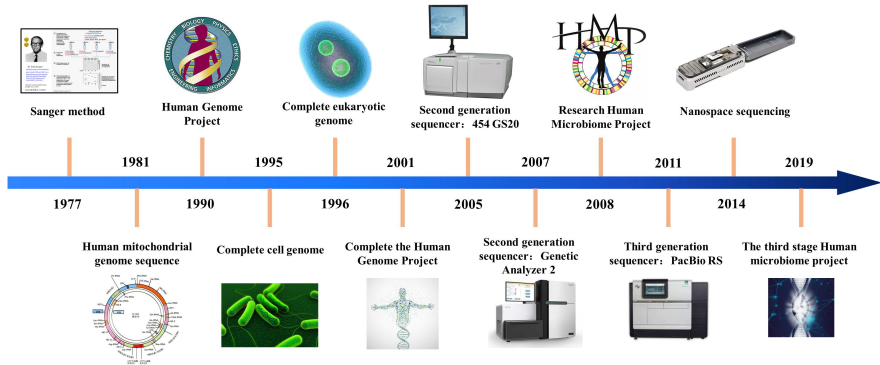
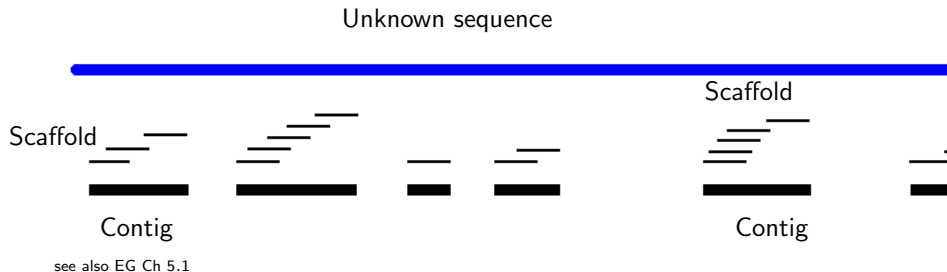


Fig. 1 in A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han and L. Zhang (2020) Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Front. Bioeng. Biotechnol. 8. doi:10.3389/fbioe.2020.01032 CC BY

see also Fig. 1 in International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409 : 860–921. doi:10.1038/35057062

# Shotgun sequencing (EG Ch. 5.1): Contigs



In reality *locations* and *orientation* of contigs is unknown.  
How does one assemble then?

# Poisson model

$N$  fragments of length  $L$ , from genome of length  $G \gg L$   
(end effects ignored)

Coverage:  $a := NL/G$

fragments taken at random:

left-hand ends i.i.d.  $X \sim \text{Unif}[0, G]$  (continuous approximation)

$$P(X \in (x, x + h)) = h/G$$

$$\#X \in (x, x + h) \sim \text{Binomial}(N, hG) \Rightarrow E[\#X] = Nh/G$$

$N$  large,  $h$  small  $\Rightarrow \text{Binomial}(N, hG) \approx \text{Poisson}(\text{mean} = Nh/G)$



# Poisson model

$Y \sim \text{Poisson}(a)$ : # of fragments with left-hand end in a given interval of length  $L$

$$\text{Prob}(Y = 0) = e^{-a}$$

Mean proportion of genome covered:  $1 - e^{-a}$   
 point chosen at random is covered by at least one fragment:  
 at least one fragment has its left-hand end in the  $L$ -interval  
 immediately to the left of this point

mean number of contigs:  $Ne^{-a} = Ne^{-NL/G}$   
 $N$  times probability that a fragment is the rightmost member of a  
 contig, i.e. no other fragment has its left-hand end inside it

## Exercises (EG Ch. 5.1)

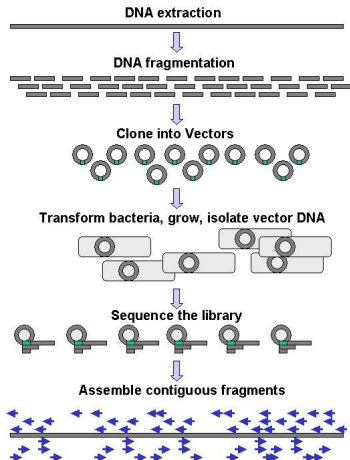
**Exercise:** study  $1 - e^{-a}$  is there a point beyond which increasing  $a$  is pointless?

**Exercise:** Plot  $Ne^{-a}$  as a function of  $N$ , explain

**Read:** What is the mean contig size?

**Read:** What is the mean number of fragments covering a point (base)?

# Shortest common superstring (NPc)



[https://en.wikipedia.org/wiki/DNA\\_sequencing](https://en.wikipedia.org/wiki/DNA_sequencing)  
(graphic by Abizar Lakdawalla, public domain)

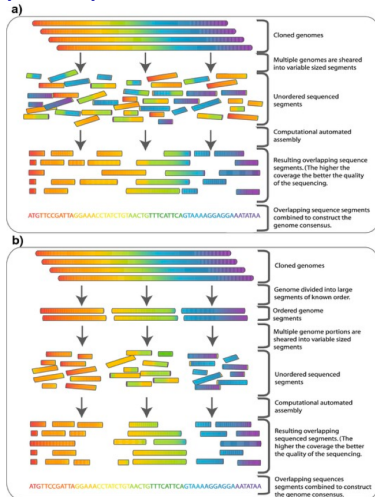
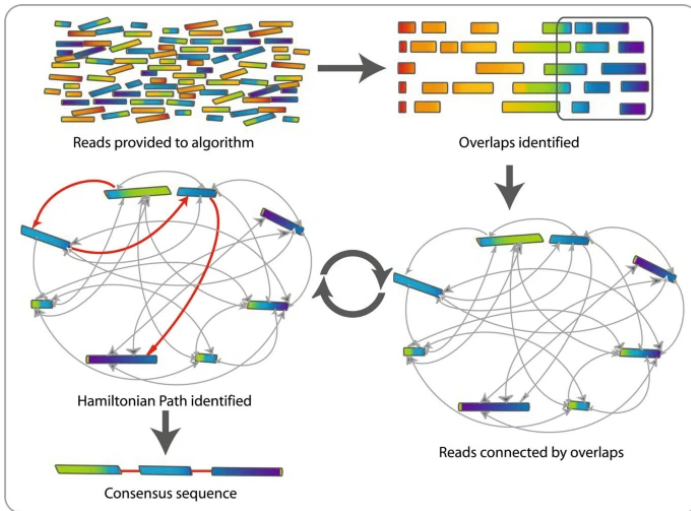


Fig. 1 in J. Commins, C. Toft, M.A. Fares (2009). Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. Biol. Proc. Online 11:52-78. doi:10.1007/s12575-009-9004-1. CC BY 2.0 / 42

# Shortest common superstring (NPc)



Overlap should be  
**end to end**  
not in the middle

Fig. 2 in J. Commins, C. Toft, M.A. Fares (2009). Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. *Biol. Proc. Online* 11:52-78. doi:10.1007/s12575-009-9004-1. CC BY 2.0

# Weighted set cover problem (NPc)

We are given

a universe  $\mathcal{U}$

family of sets  $\mathcal{S} \ni s \subseteq \mathcal{U}$

with weights  $w(s)$ ,

**AIM:** Find a *cover*  $\mathcal{C} \subseteq \mathcal{S}$  with minimal weight

i.e.

minimize  $\sum_{c \in \mathcal{C}} w(c)$  over all covers of  $\mathcal{U}$  contained in  $\mathcal{S}$

*Cover:*  $\bigcup_{c \in \mathcal{C}} c = \mathcal{U}$

# Shortest common superstring reduction

$S = \{s_1, \dots, s_n\}$  (collection of strings for SCS)

$M = \emptyset$  and then “for each pair of strings  $s_i$  and  $s_j$ , if the last  $k$  symbols of  $s_i$  are the same as the first  $k$  symbols of  $s_j$ , then add a string to  $M$  that consists of the concatenation with maximal overlap of  $s_i$  with  $s_j$ ”

$\mathcal{U} = S$

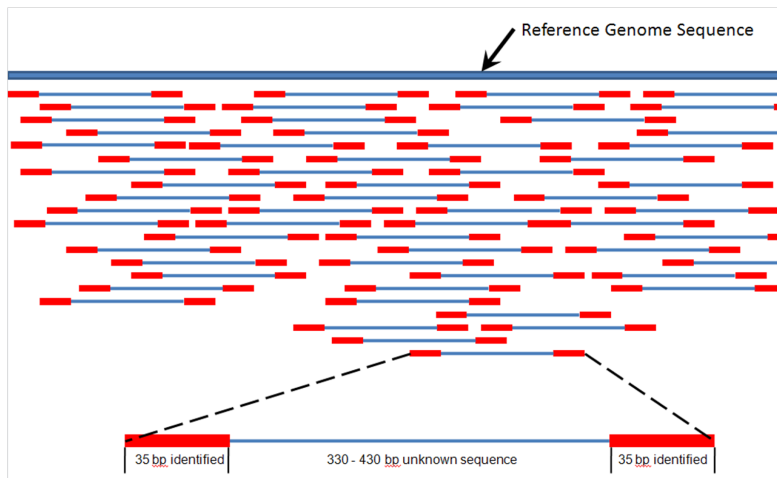
$\mathcal{S} = \{P(x) : x \in S \cup M\}$  ( $P(x)$ : set of all substrings of  $x$ )  
 $w(P(x)) = |x|$

solve by weighted set cover algorithm and SCS is arbitrary concatenation of  $x$ s from chosen  $P(x)$  sets

Set-cover is NP-complete but has  $O(\log(n))$ -approximation algorithm,  $w \leq \log(n)w_{opt}$

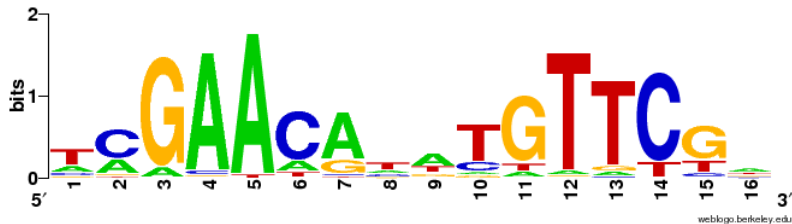
[https://en.wikipedia.org/wiki/Shortest\\_common\\_supersequence\\_problem](https://en.wikipedia.org/wiki/Shortest_common_supersequence_problem)

# Reference genome (alignment, BLAST)



[https://en.wikipedia.org/wiki/DNA\\_sequencing](https://en.wikipedia.org/wiki/DNA_sequencing) (graphic by Suspencewl, public domain) 23/ 42

# Genome variability



[https://en.wikipedia.org/wiki/Position\\_weight\\_matrix](https://en.wikipedia.org/wiki/Position_weight_matrix) (graphic by Gnomehacker, CC BY-SA 3.0)



# Position weight matrix (PWM, EG Ch. 5.3.2)

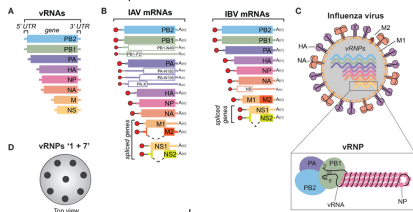
Positions assumed independent

	position				
	1	2	3	4	...
A	<i>PA1</i>	<i>PA2</i>	<i>PA3</i>	<i>PA4</i>	...
C	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	...
G	<i>PG1</i>	<i>PG2</i>	<i>PG3</i>	<i>PG4</i>	...
T	<i>PT1</i>	<i>PT2</i>	<i>PT3</i>	<i>PT4</i>	...
Total	1	1	1	1	...

Avian flu: PA gene segment

GIPL amino acids (synonymous changes)

Sequence	# strains
GGT ATA CCG TTA	6
GGG ATA CCG CTG	19
GGT ATA CCG CTA	27
GGG ATA CCG CTA	1
	14



K. Bartoszek, Bayesian Variable Selection Applied to the Assessment of Pathogenicity of Avian Flu, 2008, MPHil Dissertation, Cambridge Univ.

Fig. 1 in D. Dou, R. Revol, H. Östbye, H. Wang, R. Daniels, 2018. Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. Front. Immunol. 9:1581. doi: 10.3389/fimmu.2018.01581. CC BY

	position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0.79	0	0.79	0	0	0	0	0	0.51
C	0	0	0	0	0	0	0.79	0.79	0	0.7	0	0
G	0.79	0.79	0.3	0	0	0	0	0	0.79	0	0	0.28
T	0	0	0.49	0	0.79	0	0	0	0	0.09	0.79	0
—	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
Total	1	1	1	1	1	1	1	1	1	1	1	1

# Markov chains (EG Ch. 4.5–4.10, MM Ch. 12)

$S = \{A, C, G, T\} \ni X_n$  nucleotide at position  $n$  of chromosome/genome/DNA fragment/e.t.c.

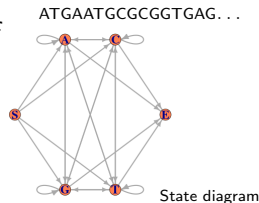
*transition probability:*

$$p_{ij} = P(X_{n+1} = j | X_n = i) = P(X_{n+1} = j | X_n = i, \dots, X_0 = i_0)$$

*Markov chain of order  $k$ :*

$$P(X_{n+1} = j | X_n = i, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i, \dots, X_{n-k+1} = i_{n-k+1})$$

**Remember:** 3 nucleotides per amino acid



# Markov chains

*transition matrix:*  $P = [p_{ij}] \in [0, 1]^{|S| \times |S|}$

*2-step transition probability:*  $p_{ij}^{(2)} = \sum_{k \in S} p_{ik} p_{kj}$ , i.e.  $P^2$

*n-step transition probability:*  $P^n$

*absorbing state:*  $i$  s.t.  $p_{ii} = 1$  (or  $p_{ij} = 0$ ,  $i \neq j$ )

if  $X_0 \sim \phi_0$  then  $X_1 \sim \phi_1^T = \phi_0^T P$

*stationary distribution:*  $\phi^T = \phi^T P$

# Markov chains

	position				
	A	C	G	T	Total
A	$p_{AA}$	$p_{AC}$	$p_{AG}$	$p_{AT}$	1
C	$p_{CA}$	$p_{CC}$	$p_{CG}$	$p_{CT}$	1
G	$p_{GA}$	$p_{GC}$	$p_{GG}$	$p_{GT}$	1
T	$p_{TA}$	$p_{TC}$	$p_{TG}$	$p_{TT}$	1
Total	$p_{\rightarrow A}$	$p_{\rightarrow C}$	$p_{\rightarrow G}$	$p_{\rightarrow T}$	

stochastic matrix

$$p_{\rightarrow A} = p_{\rightarrow C} = p_{\rightarrow G} = p_{\rightarrow T} = 1$$

doubly stochastic matrix

Higher order: transition matrices large, parameter rich  
(data can be insufficient)

Maximal dependence decomposition for group of *aligned* sequences

Ch 5.3.4 and T.-Y. Lee et. al. (2011). "Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences". Bioinformatics 27(13) : 1780–1787.

# Repeats (EG Ch. 5.4, MM p. 223)

Does a given nucleotide (say  $A$ ) repeat itself more often than expected by chance in a *long* DNA sequence of length  $N$ ?

$p$ : probability of observing  $A$ ,

$Y$ : number of repeats of  $A$ ,  $P(Y = y) = (1 - p)p^y$

$n$  such sequences

$Y_{\max}$  longest:  $P(Y_{\max} \geq y) = 1 - (1 - p^y)^n$

a success ( $A$ ) has to be preceded by a failure,

$\approx (1 - p)N$  failures implying  $\approx (1 - p)N$  successes

take  $n \approx (1 - p)N$

# Repeats

$Y_{\max}$  test statistic with p-value (EG Eq. 5.15):

$$1 - (1 - p^{y_{\max}})^{(1-p)N} \approx 1 - \exp(-(1 - p)Np^{y_{\max}})$$

$N$  large

$$(1 - p)Np^{y_{\max}} < 1$$

**remember:**  $(1 + x/n)^n \rightarrow e^x$ .

Is  $(1 - p)N$  a good approximation for number of failures?

Average of binomial with parameters  $N$  and  $1 - p$ .

# Patterns (EG Ch. 5.6–5.8)

Counting specific repeats: e.g. *GCGC*,

A T G C G C G C A A G C G C T T

**2** or **3** ?

Different models and results depending if *overlaps* counted.

# Motifs (EG Ch. 5.9)

Many short sequences serve specific functions and do not tolerate many mutations.

*transcription factor binding sites, splice junction signals*

*Motif*: collection of  $m$  different (uncontained) sequences, similar and of same length

## Example:

$M = \{GATGGTGG, GCTGGTGG, GGTGGTGG, GTTGGTGG\}$   
(crossover hotspot initiator in *Hemophilus influenzae*, bacteria)



# Motifs (EG Ch. 5.9)

[https://en.wikipedia.org/wiki/Position\\_weight\\_matrix](https://en.wikipedia.org/wiki/Position_weight_matrix)

A motif can be represented as a PWM

$$M = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix} \end{matrix}$$

GAGGTAAAC  
TCCGTAAGT  
CAGGTTGGA  
ACAGTCAGT  
TAGGTCATT  
TAGGTAAGT  
ATGGAAGT  
CAGGTATAC  
TGTGTGAGT  
AAGGTAAGT

Probability of observing a motif

$$P(M) = \sum_{u \in M} P(u)$$

$P(u)$ : independent positions, product over them,

$$P(\text{GAGGTAAAC}) = 0.1 \cdot 0.6 \cdot 0.7 \cdot 1 \cdot 1 \cdot 0.6 \cdot 0.7 \cdot 0.2 \cdot 0.2$$

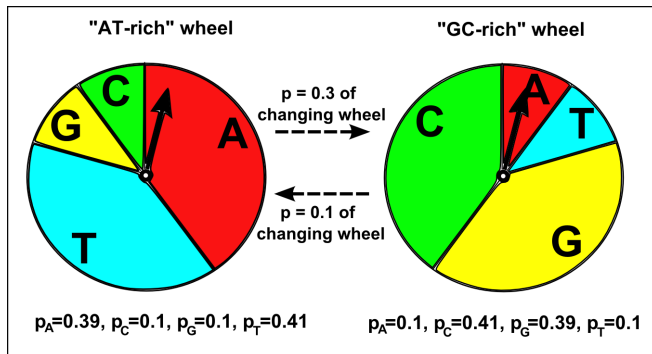
$N$ : sequence length

mean number of motif occurrences:  $(N - |M| + 1)P(M)$

variance with or without overlaps, Eq. (5.91) or (5.92)

test if occurs as often as expected

# Hidden Markov Models (EG Ch. 12, MM Ch. 12)

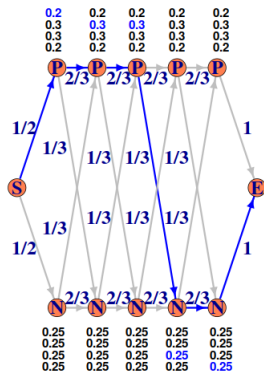


A. Coghlan, (2011) A Little Book of R For Bioinformatics

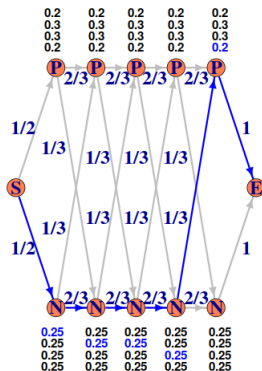
<https://a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/> CC BY

## Hidden Markov Models

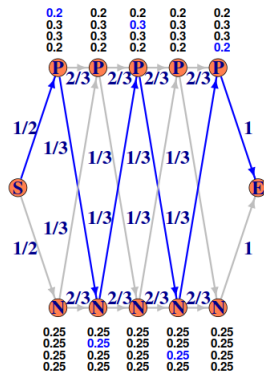
Sequence: **ACCGT**, P: protein coding, N: non-coding



$$\frac{1}{2} \frac{1}{5} \frac{2}{3} \frac{3}{10} \frac{2}{3} \frac{3}{10} \frac{1}{3} \frac{1}{4} \frac{2}{3} \frac{1}{4} \approx 0.000192$$



$$\frac{1}{2} \frac{1}{4} \frac{2}{3} \frac{1}{4} \frac{2}{3} \frac{1}{4} \frac{2}{3} \frac{1}{4} \frac{1}{3} \frac{1}{5} \approx 3.858 \cdot 10^{-5}$$



$$\frac{1}{2} \frac{1}{5} \frac{1}{3} \frac{1}{4} \frac{1}{3} \frac{3}{10} \frac{1}{3} \frac{1}{4} \frac{1}{3} \frac{1}{5} \approx 4.63 \cdot 10^{-6}$$

# Hidden Markov Models

Observed outputs:  $\mathcal{O} = \mathcal{O}_1, \dots, \mathcal{O}_T$

Hidden sequence:  $Q = q_1, \dots, q_T$

Values of hidden states:  $S_1, \dots, S_N$

Model parameters:  $\lambda$

Chain homogeneous in time, i.e. in  $1, \dots, T$

Notation:

$$\pi_i = P(q_1 = S_i)$$

$$b_i(a) = P(S_i \text{ emits } a)$$

$p_{ij}$ : transition probability of hidden chain

# Hidden Markov Models

## Aims:

1. find  $P(\mathcal{O}|\lambda)$
2. find hidden sequence  $Q$ , i.e.

$$\arg \max_Q P(Q|\mathcal{O})$$

3. find  $\lambda$ , *heuristic*: Baum–Welch algorithm

$$\arg \max_{\lambda} P(\mathcal{O}|\lambda)$$

Number of possible paths make exact algorithm computationally impossible.

# The likelihood, forward algorithm $O(TN^2)$ , (EG p. 411)

$$\alpha(t, i) := P(\mathcal{O}_1, \dots, \mathcal{O}_t, q_t = S_i)$$

$$P(\mathcal{O}) = \sum_{i=1}^N \alpha(T, i)$$

*initialization:*  $\alpha(1, i) = P(q_1 = S_i)P(S_i \text{ emits } \mathcal{O}_1) = \pi_i b_i(\mathcal{O}_1)$

*induction:*

$$\alpha(t+1, i) = \sum_{j=1}^N \alpha(t, j) p_{ji} b_i(\mathcal{O}_{t+1})$$

because

$$\alpha(t+1, i) = \sum_{j=1}^N P(\mathcal{O}_1, \dots, \mathcal{O}_{t+1}, q_{t+1} = S_i, q_t = S_j)$$

remember:  $p_{ji} = P(S_j \rightarrow S_i)$

# The likelihood, backward algorithm, (EG p. 412)

**Aim:** calculate

$$\beta(t, i) := P(\mathcal{O}_{t+1}, \dots, \mathcal{O}_T | q_t = S_i)$$

*initialization:*  $\beta(T, j) \equiv 1$  for all  $j$

*recursion:*

$$\beta(t-1, i) = \sum_{j=1}^N p_{ij} b_j(\mathcal{O}_t) \beta(t, j)$$

Viterbi algorithm  $O(TN^2)$ , (EG p. 413)

$$\delta_t(i) = \max_{q_1, \dots, q_{t-1}} P(q_1, q_2, \dots, q_{t-1}, q_t = S_i, \mathcal{O}_1, \dots, \mathcal{O}_t)$$

$\delta_t(i)$ : max. probability to end in state  $S_i$  and time  $t$  with observations  $\mathcal{O}$

*initialization*:  $\delta_1(i) = \pi_i b_i(\mathcal{O}_1)$

*induction*: for  $2 \leq t \leq T$ ,  $1 \leq j \leq N$

$$\delta_t(j) = \max_{1 \leq i \leq N} \delta_{t-1}(i) p_{ij} b_j(\mathcal{O}_t)$$

Recover hidden states:

$$\psi_T := \arg \max_{1 \leq i \leq N} \delta_T(i), \text{ put } q_T = S_{\psi_T}$$

and for  $t \leq T - 1$

$$\psi_t := \arg \max_{1 \leq i \leq N} \delta_t(i) p_{i\psi_{t+1}} \text{ and put } q_t = S_{\psi_t}$$



# Baum–Welch algorithm, (EG p. 414, MM p. 324)

Data: multiple sequences  $\{\mathcal{O}^{(d)}\} = \{\mathcal{O}^{(1)}, \dots, \mathcal{O}^{(n)}\}$

*Initialize:*  $\pi_i, p_{jk}, b_i(a)$  at some value

*Update*

$\bar{\pi}_i$ : as expected proportion of times  $q_1 = S_i$  given  $\{\mathcal{O}^{(d)}\}$

$\bar{p}_{jk} = E [N_{jk} | \{\mathcal{O}^{(d)}\}] / E [N_j | \{\mathcal{O}^{(d)}\}]$

$\bar{b}_i(a) = E [N_i(a) | \{\mathcal{O}^{(d)}\}] / E [N_i | \{\mathcal{O}^{(d)}\}]$

where

$N_{jk}$ : (random) number of transitions  $q. = S_j$  to  $q_{+1} = S_k$

$N_i$ : (random) number of  $q. = S_i$

$N_i(a)$ : (random) number of emissions of  $a$  by  $S_i$

for a **single random** sequence

*Iterate* until convergence

Exact computational details: EG p. 415, 416

# Questions?