Introduction to Bioinformatics

JTMS-19

Marc-Thorsten Hütt mhuett@constructor.university

Felix Jonas fjonas@constructor.university

session Wed, September 18, 2024

Application examples for probability models; first concepts of alignment algorithms

What is this session about?

Several application of the probability concepts from the previous session are discussed (simple probability estimation, Bayes' theorem, Markov chains). First steps towards pairwise sequence alignment are described (scoring functions, Needleman-Wunsch algorithm).

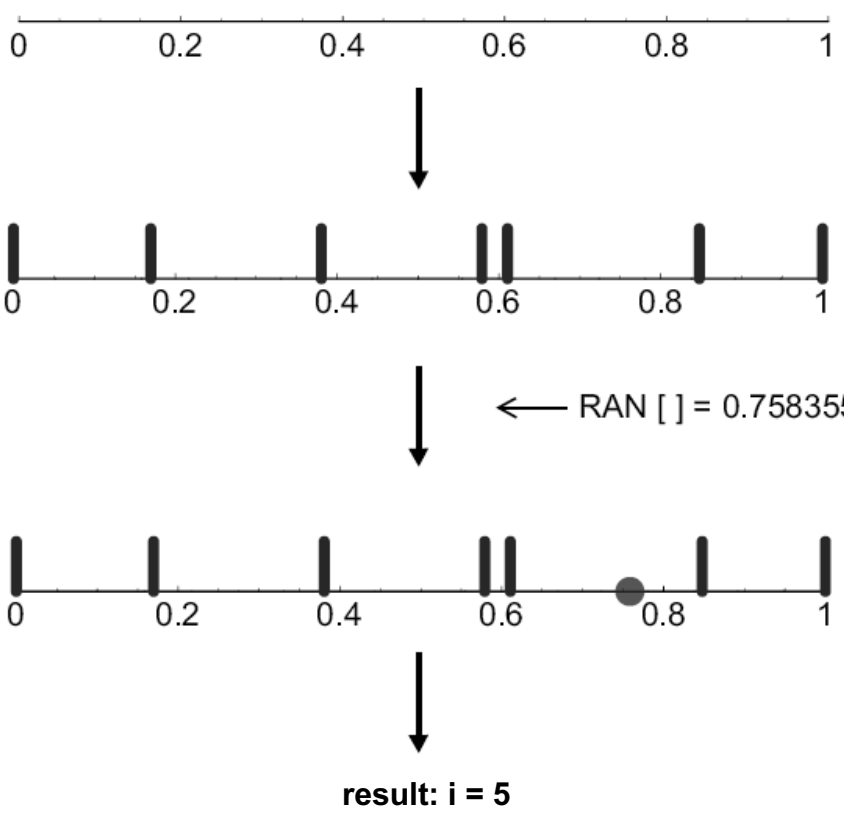
How can you revise the material after the session?

Read Durbin et al. chapters 3.1 and 2.1– 2.3

alternative reading: Hütt/Dehnert chapters 2.6 – 2.7 and 3.1.1-3.1.2

Probability models

Implementing a discrete probability distribution

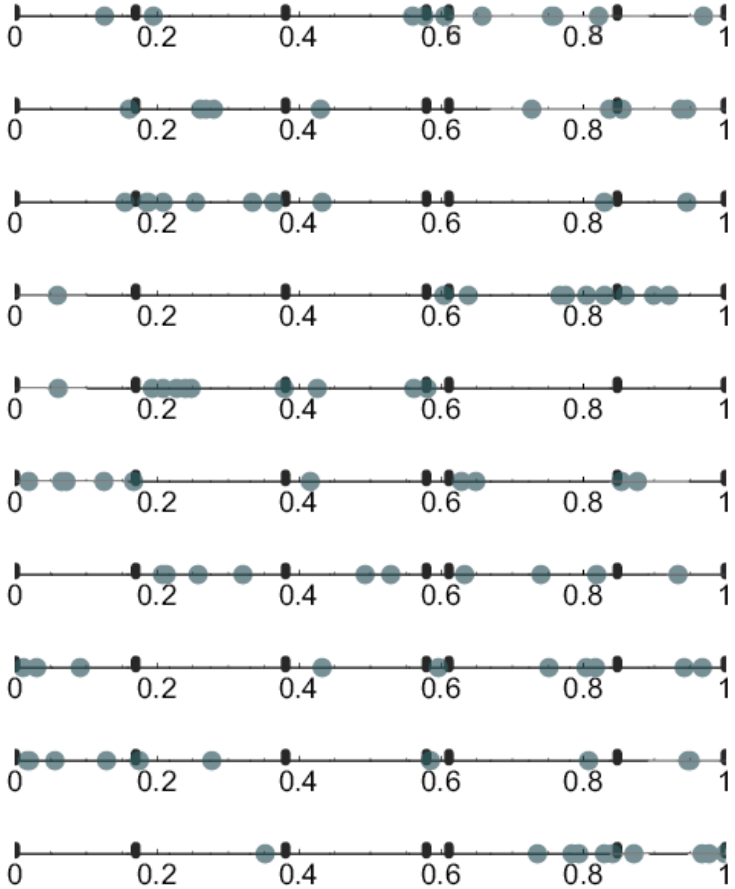


range [0, 1]

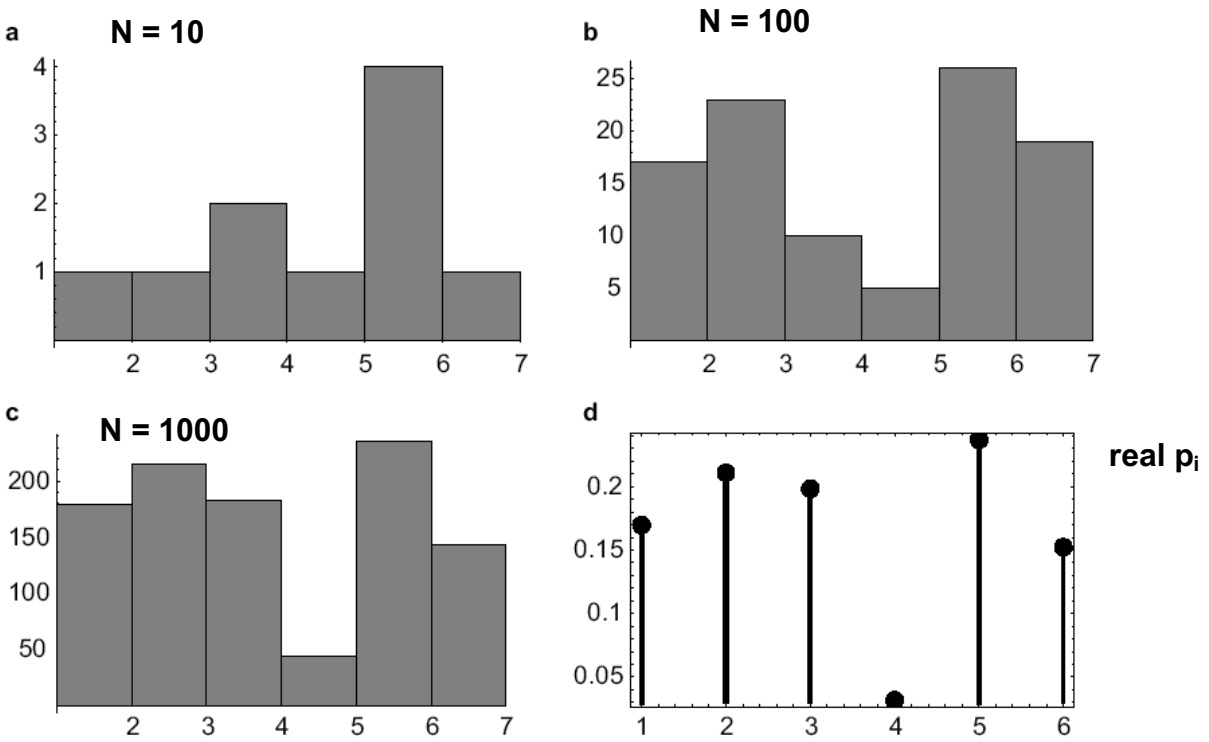
insertion of the p_i

insertion of a random number

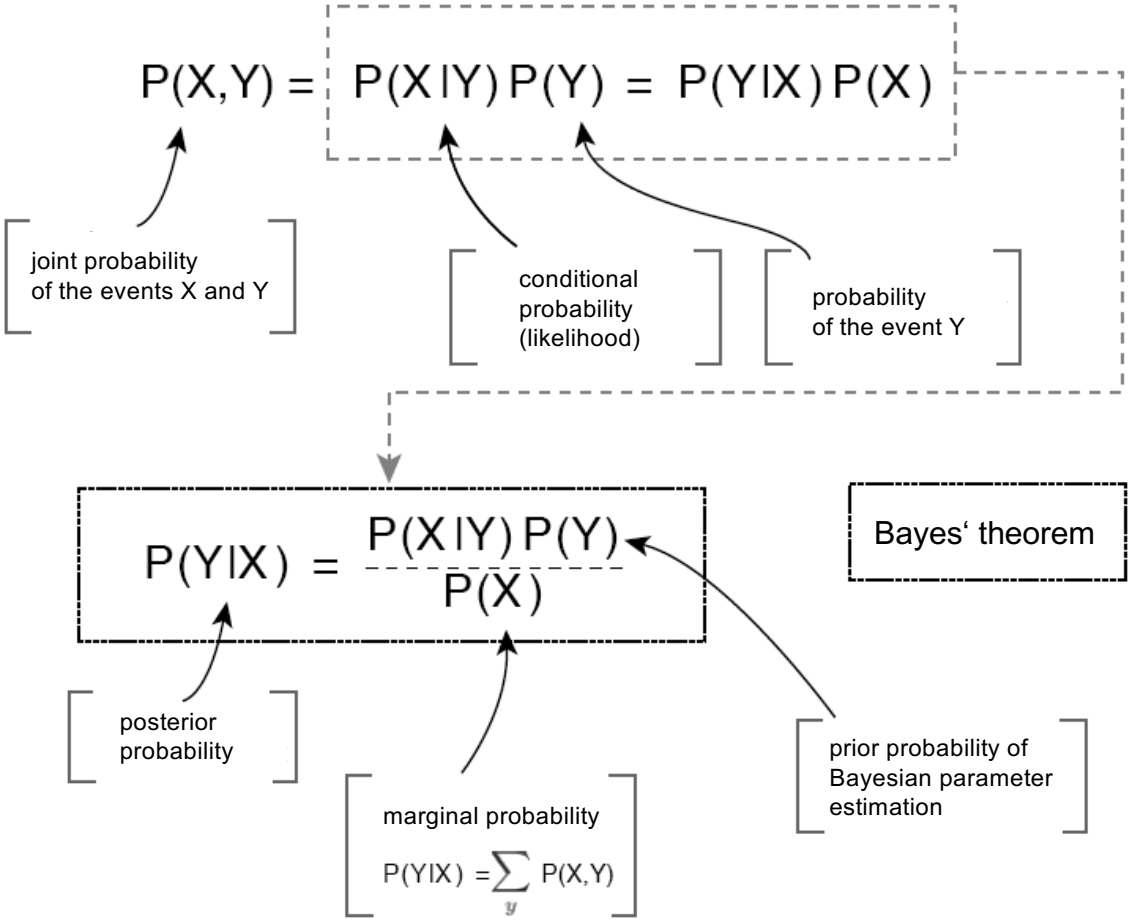
Implementing a discrete probability distribution



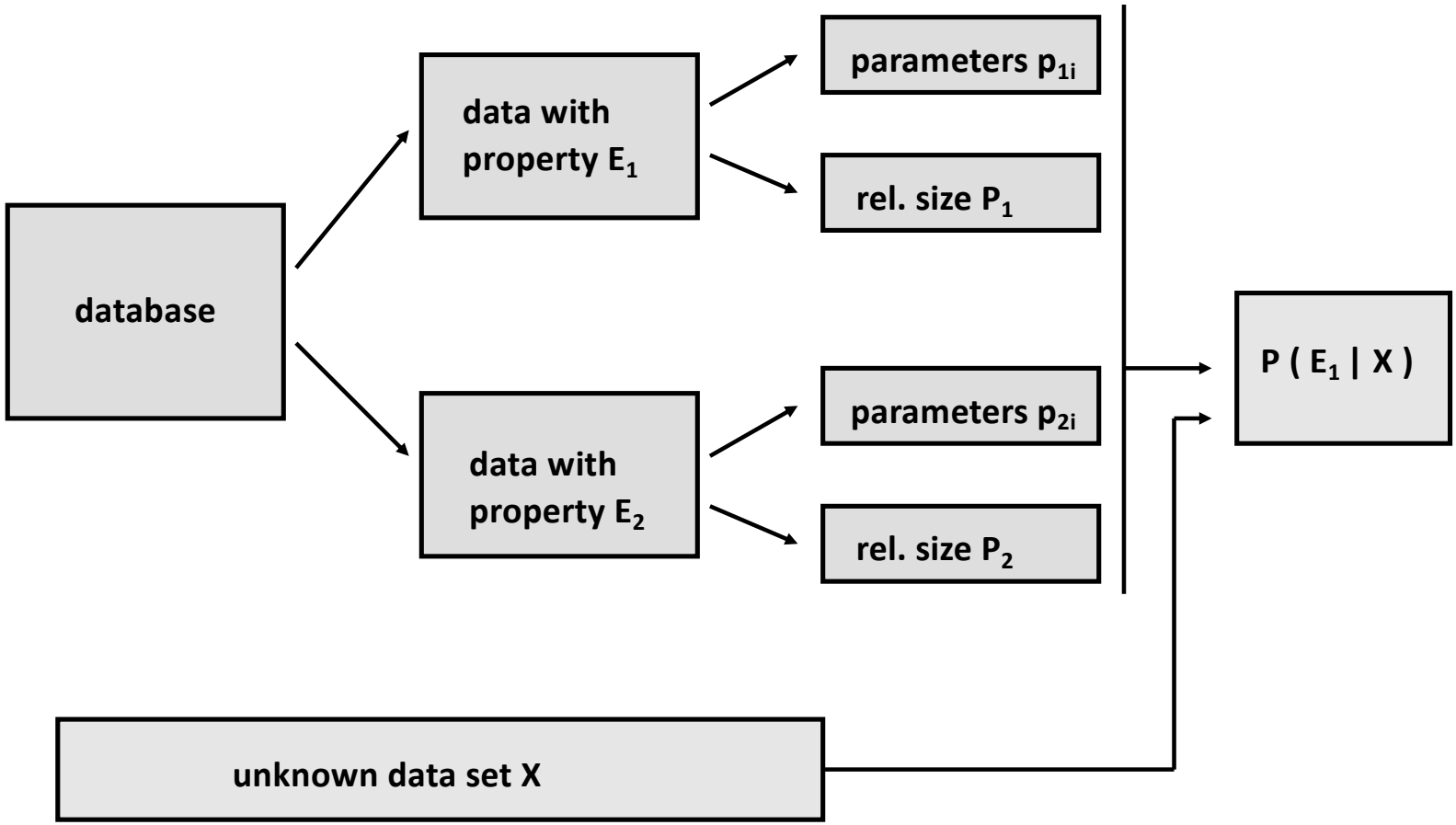
Implementing a discrete probability distribution



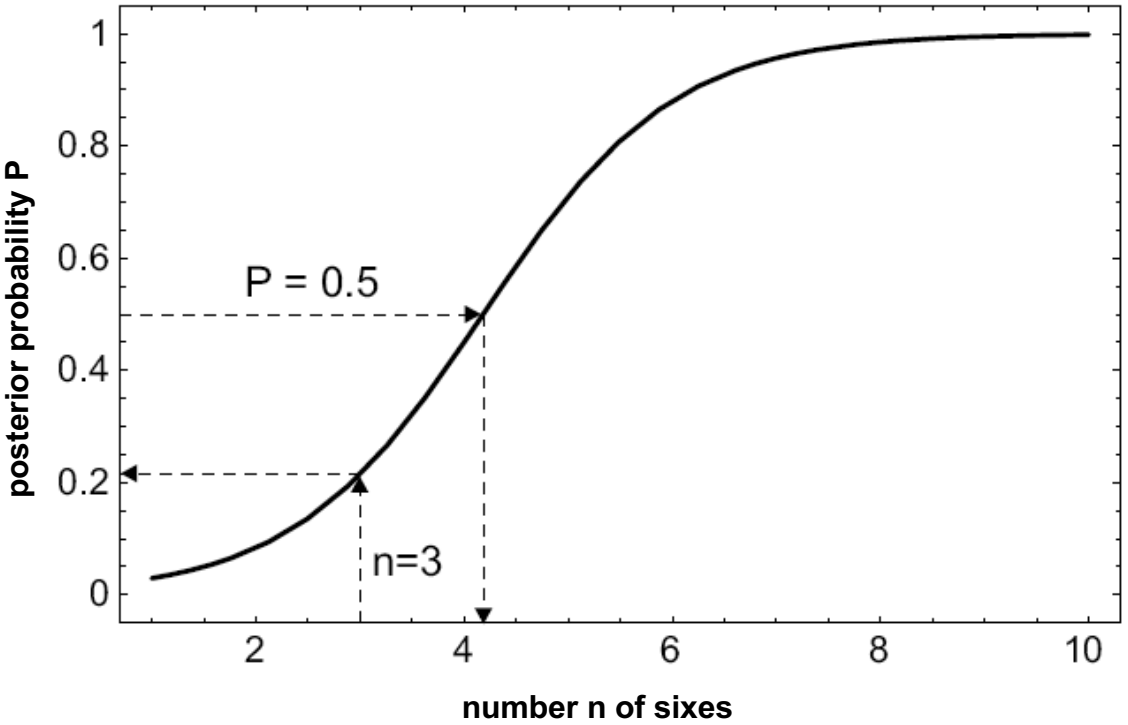
How to quantify the match between data and a probability model?



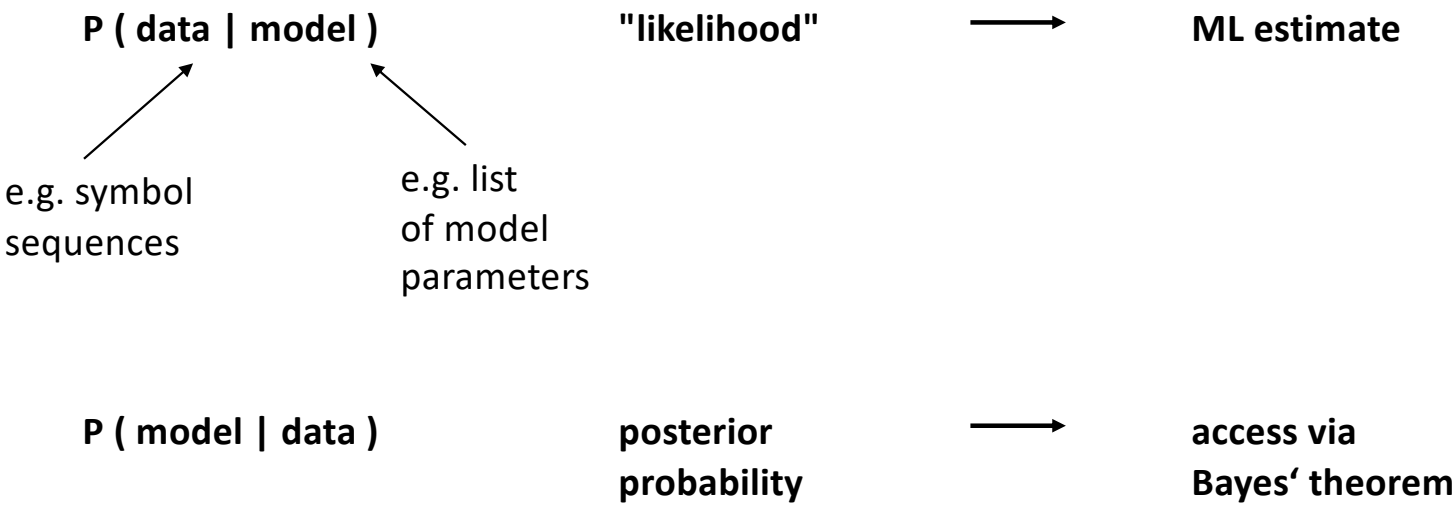
How to use probability models in practice?



Numerical example: occasionally dishonest casino
(Durbin et al. 1998)



How to quantify the match between data and a probability model?



concept of CpG islands

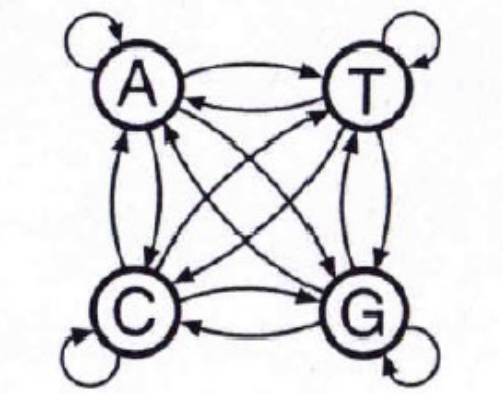
In many genomes CpG dinucleotides are highly suppressed,

C → methylated-C → T,

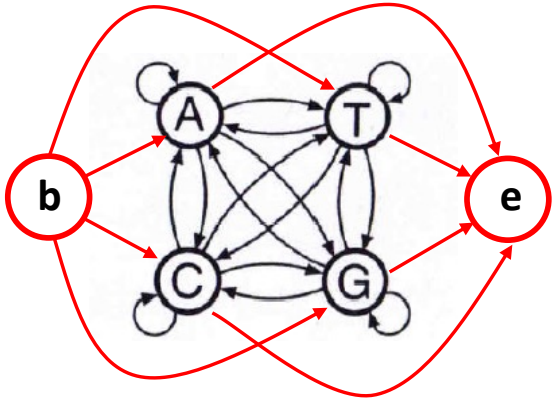
except close to promotors and the regulatory regions of genes

→ CpG islands

Markov chains as a tool for studying CpG islands



Transition probabilities as model parameters



Introduction of start and end states

Markov chains as a tool for studying CpG islands

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

$$S(x) = \log\left(\frac{P(x \mid \text{model } +)}{P(x \mid \text{model } -)}\right) = \log\left(\frac{P(B)\prod_{i=1}^{L} a^{+}_{x_{i-1}x_i}}{P(B)\prod_{i=1}^{L} a^{-}_{x_{i-1}x_i}}\right) = \sum_{i=1}^{L} \log\left(\frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}\right) = \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

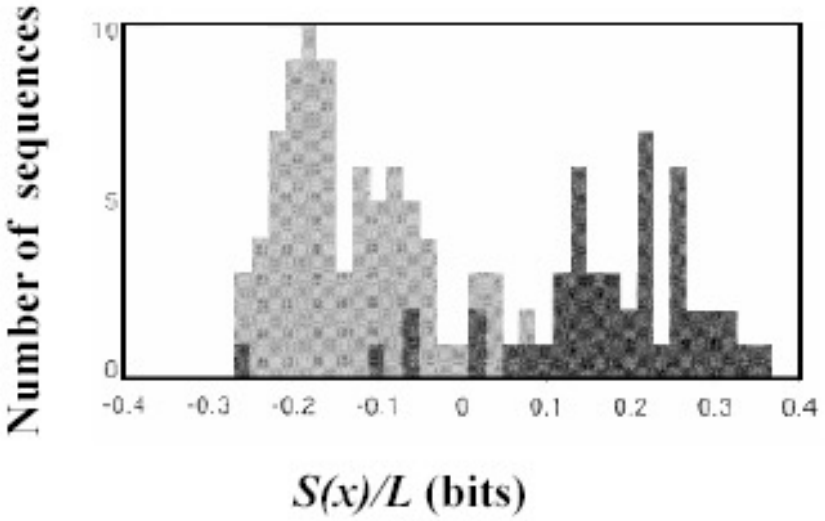
a number for each sequence x
→ histogram of score values S(x)
for many sequences x

a number for each dinucleotide
→ table of "log-likelihoods"

Markov chains as a tool for studying CpG islands

ß(log$_2$)	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.0685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

table of "log-likelihoods"



histogram of scores

Pairwise sequence alignment

Taken from: Lesk, Introduction to Bioinformatics (2002) Oxford Univ. Press

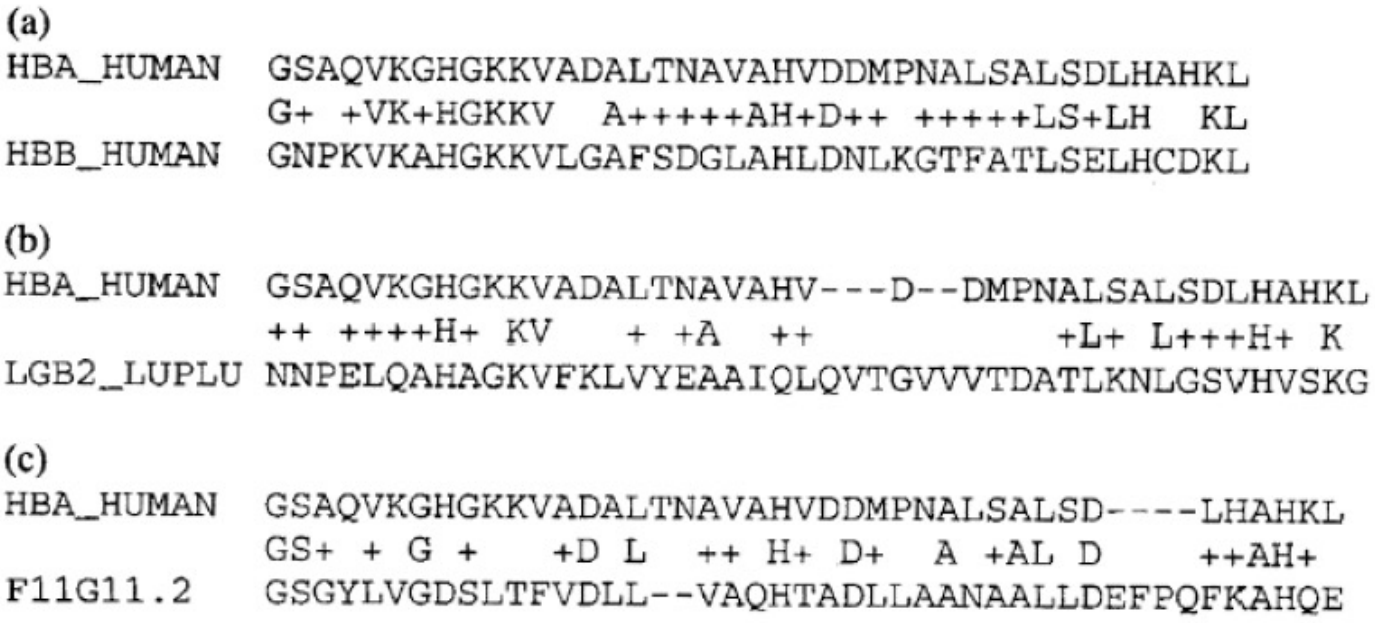


Figure 2.1 *Three sequence alignments to a fragment of human alpha globin. (a) Clear similarity to human beta globin. (b) A structurally plausible alignment to leghaemoglobin from yellow lupin. (c) A spurious high-scoring alignment to a nematode glutathione S-transferase homologue named F11G11.2.*

Questions

- How does one pair the symbols of two given sequences? Does one allow gaps and shifts?
- How are dissimilarities evaluated?
- When are observed similarities systematic? When are they just random?

Tasks

(1) quantitative evaluation of similarities

(2) systematic (algorithm-based) search for an appropriate (or even optimal) alignment

- Uninformative:

```
-------gctgaacg
ctataatc-------
```

- Without gaps:

```
gctgaacg
ctataatc
```

- With gaps:

```
gctga-a--cg
--ct-ataatc
```

- Another one:

```
gctg-aa-cg
-ctataatc-
```

[next three slides:
M. Schroeder, TU Dresden]

- Example
  - match +1
  - mismatch -1
  - Gap opening -3
  - Gap extension -1

- Uninformative: 0%, score= -21

```
-------gctgaacg
ctataatc-------
```

- Without gaps:25%,score= -4

```
gctgaacg
ctataatc
```

- With gaps: 0%, score= -23

```
gctga-a--cg
--ct-ataatc
```

- Another one: 50%, score=-12

```
gctg-aa-cg
-ctataatc-
```

- Example
  - match +2
  - mismatch -1
  - Gap opening -1
  - Gap extension -1

- Uninformative: 0%, score= -17

```
-------gctgaacg
ctataatc-------
```

- Without gaps:25%,score= -2

```
gctgaacg
ctataatc
```

- With gaps: 0%, score= -15

```
gctga-a--cg
--ct-ataatc
```

- Another one: 50%, score=0

```
gctg-aa-cg
-ctataatc-
```

a

```
CGATC-CTGT
|  |||  |   |
C-ATCGCCTT
```

b

```
CGAT--CCTGT
|  ||   |||  |
C-ATCGCCT-T
```

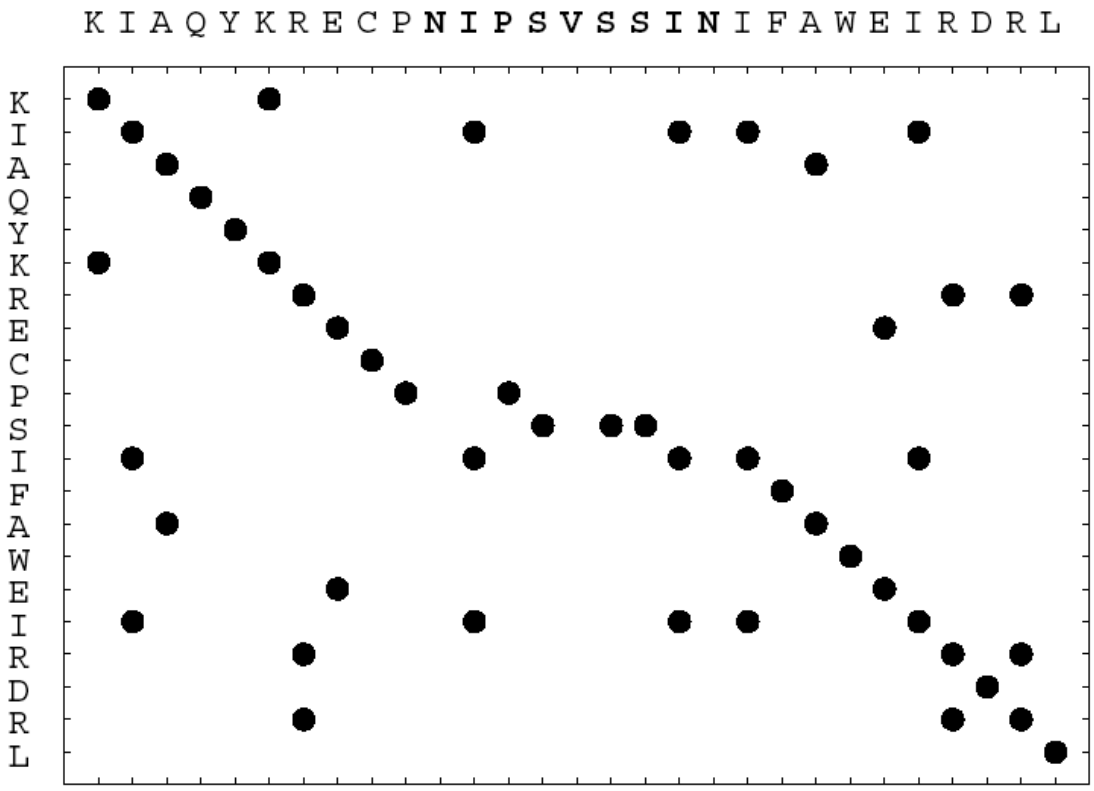
c

```
CGATCCTGT
|     |   |
CATCGCCTT
```

Parameterset	A	B	C
Match	1	1	1
Mismatch	1	0	1
Gap-opening	3	0	5
Gap-extension	1	0	1

	Ali. 1	Ali. 2	Ali. 3
A	-2	-3	-3
B	6	7	3
C	-6	-10	-3

Sequence alignment as a path in the dotplot plane


K I A Q Y K R E C P N I P S V S S I N I F A W E I R D R L
K
I
A
Q
Y
K
R
E
C
P
S
I
F
A
W
E
I
R
D
R
L

DNA sequences: the case of global sequence alignment

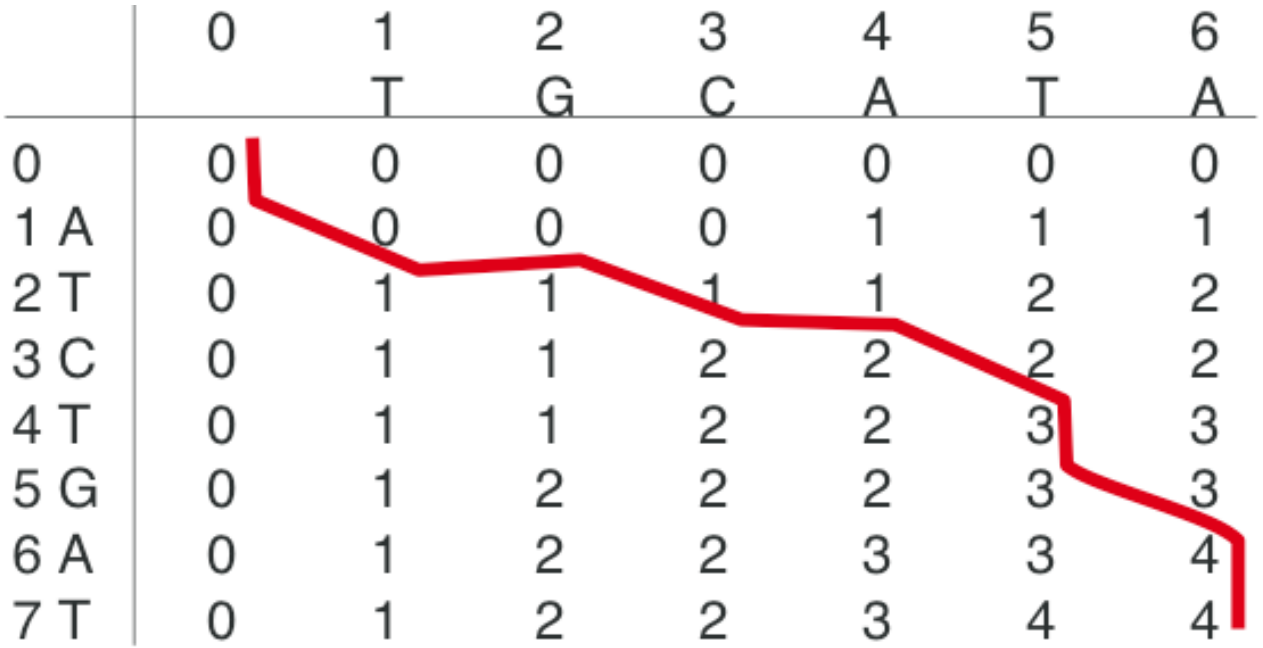
	0	1 T	2 G	3 C	4 A	5 T	6 A
0	0	0	0	0	0	0	0
1 A	0						
2 T	0						
3 C	0						
4 T	0						
5 G	0						
6 A	0						
7 T	0						

Insert a row 0 and column 0 initialised with 0

		0	1 T	2 G	3 C	4 A	5 T	6 A
0		0	0	0	0	0	0	0
1	A	0	0	0	0	1	1	1
2	T	0						
3	C	0						
4	T	0						
5	G	0						
6	A	0						
7	T	0						

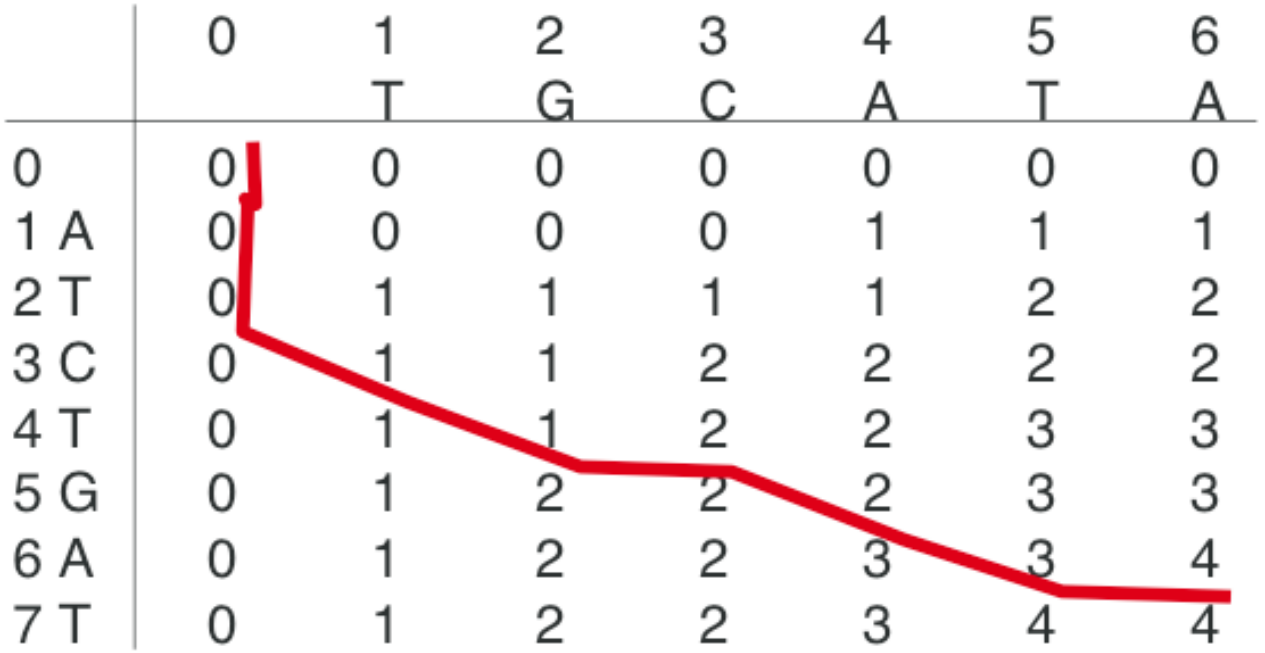
- Consider
  - Value north
  - Value west
  - Value northwest if the row/column character **mismatch**
  - 1 + value northwest if the row/column character **match**
- Put down the **maximum of these values** for current cell

	0	1 T	2 G	3 C	4 A	5 T	6 A
0	0	0	0	0	0	0	0
1 A	0	0	0	0	**1**	1	**1**
2 T	0	**1**	1	1	1	2	2
3 C	0	1	1	**2**	2	2	2
4 T	0	**1**	1	2	2	**3**	3
5 G	0	1	**2**	2	2	3	3
6 A	0	1	2	2	**3**	3	**4**
7 T	0	**1**	2	2	3	**4**	4



		0	1 T	2 G	3 C	4 A	5 T	6 A
0		0	0	0	0	0	0	0
1	A	0	0	0	0	1	1	1
2	T	0	1	1	1	1	2	2
3	C	0	1	1	2	2	2	2
4	T	0	1	1	2	2	3	3
5	G	0	1	2	2	2	3	3
6	A	0	1	2	2	3	3	4
7	T	0	1	2	2	3	4	4

```
-tgcat-a-
at-c-tgat
```



		0	1 T	2 G	3 C	4 A	5 T	6 A
0		0	0	0	0	0	0	0
1	A	0	0	0	0	1	1	1
2	T	0	1	1	1	1	2	2
3	C	0	1	1	2	2	2	2
4	T	0	1	1	2	2	3	3
5	G	0	1	2	2	2	3	3
6	A	0	1	2	2	3	3	4
7	T	0	1	2	2	3	4	4

```
---tgcata
atctg-at-
```