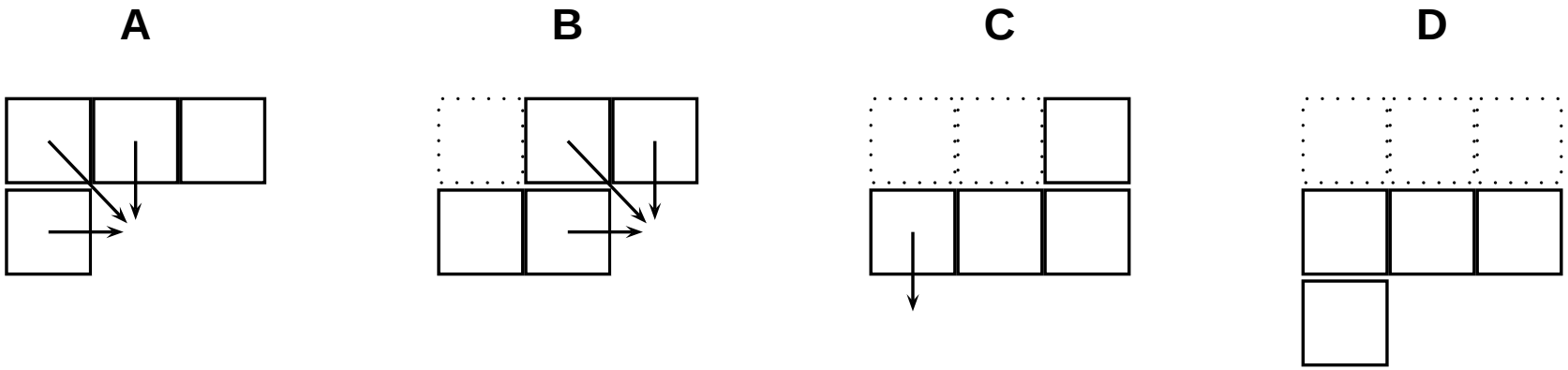
Introduction to Computational Biology; An Evolutionary Approach: Fast Alignment

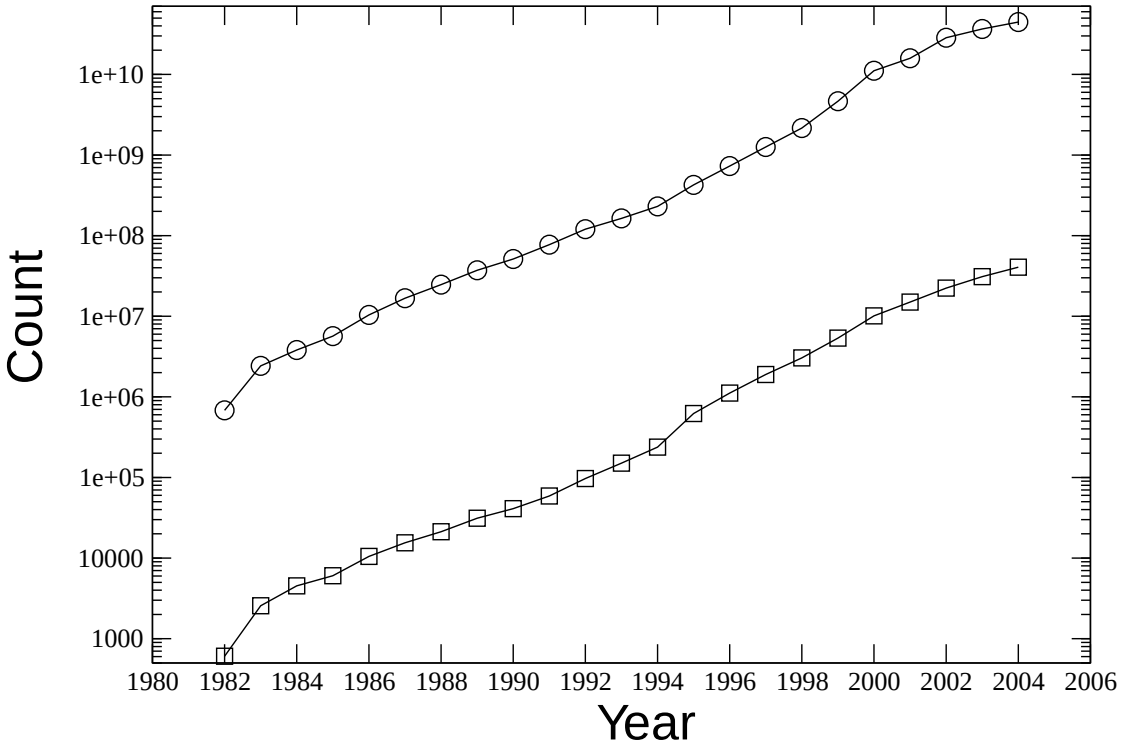
Bernhard Haubold & Thomas Wiehe

Outline

Score Computation in Linear Space



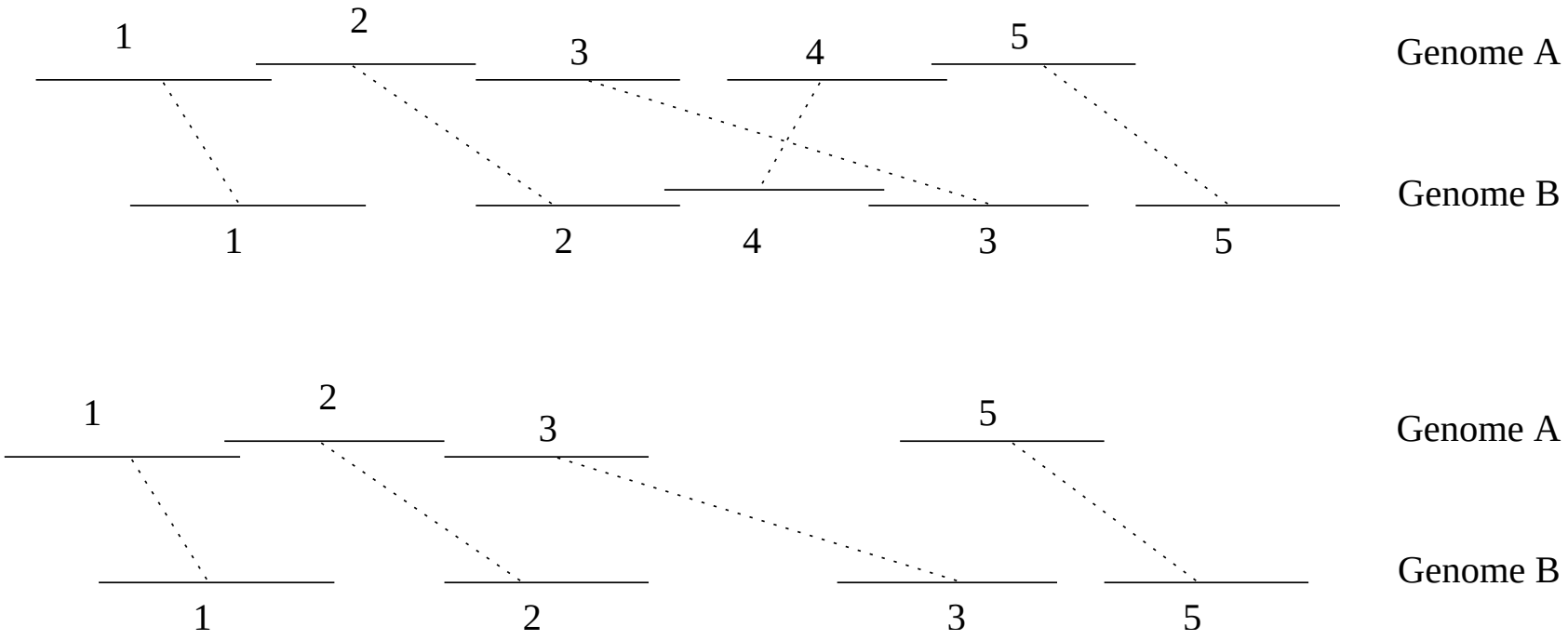
GenBank Size



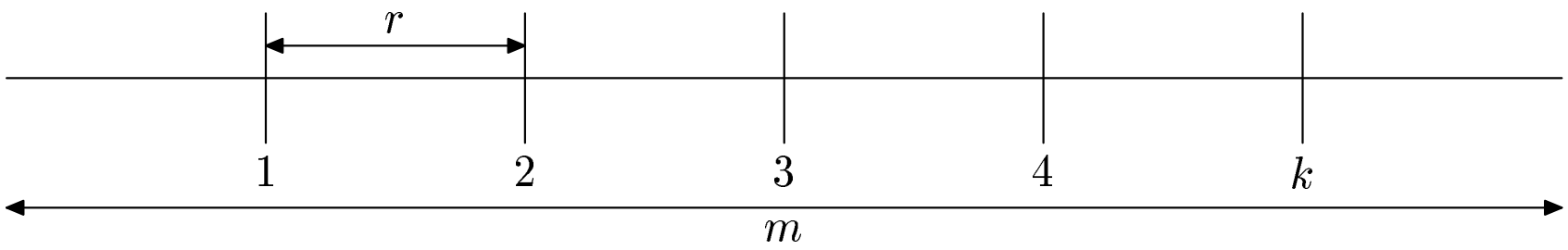
Genome Alignment with `MUMmer`

1. identify all ***m**aximal **u**nique **m**atches* (MUMs) between the two genomes; such repeats cannot be extended and occur only between and not within the genomes;
2. find the longest increasing subsequence MUMs;
3. process the gaps in the resulting alignment.

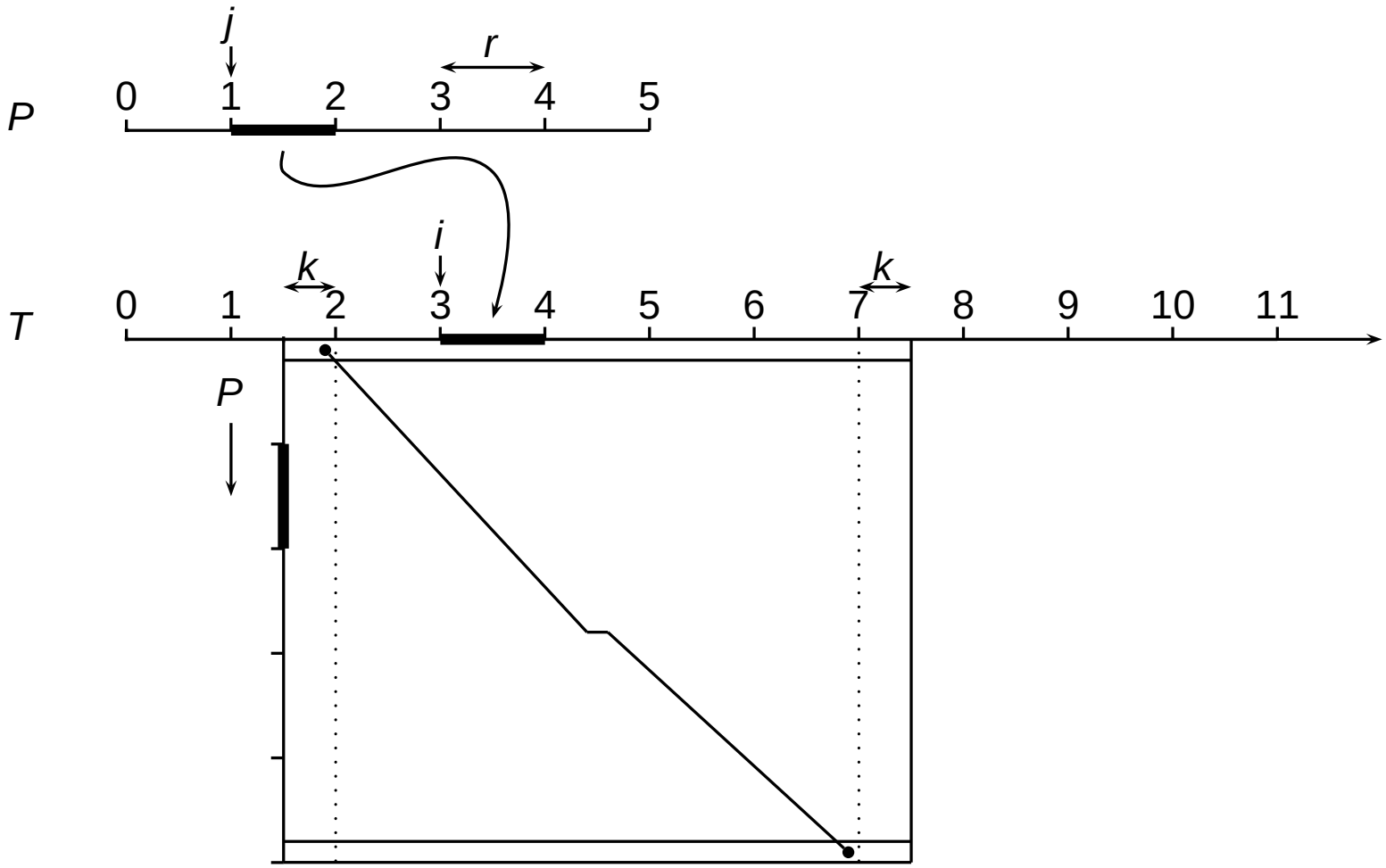
Longest Increasing Subsequence of MUMs



k-Error Match—1



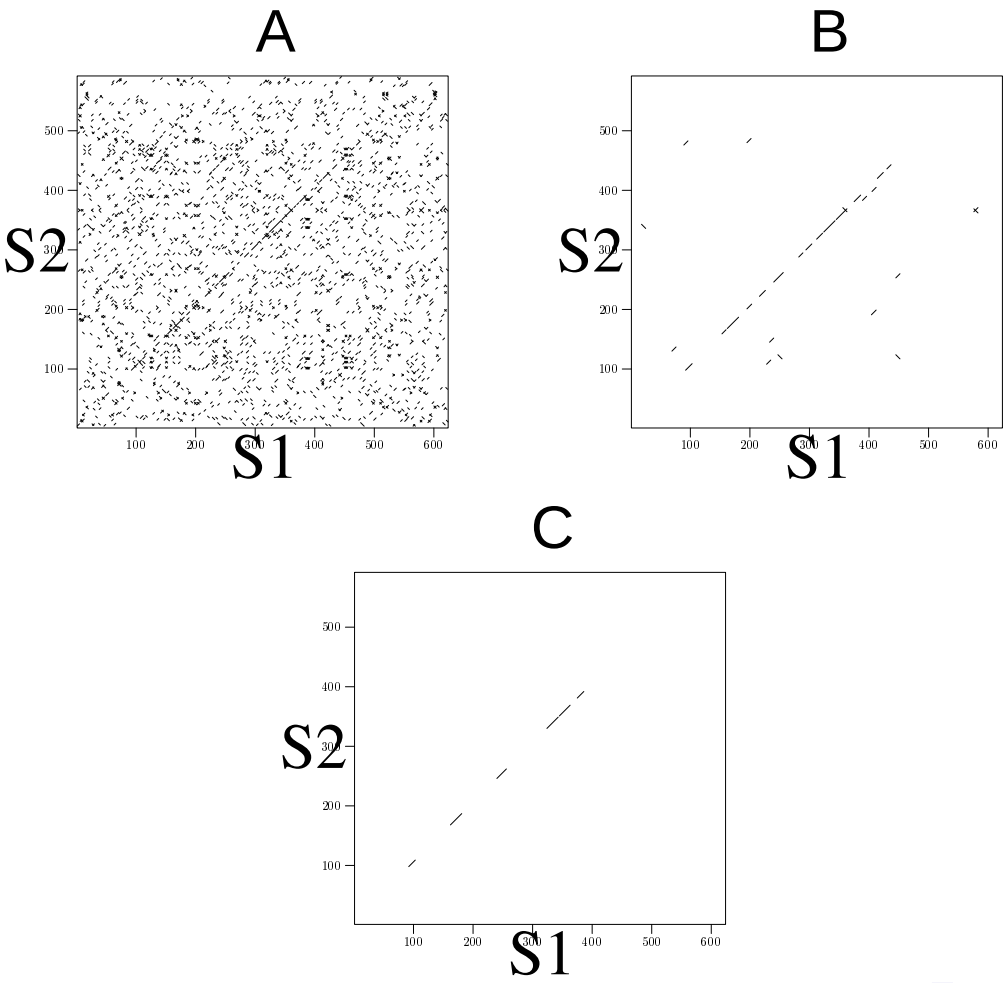
k-Error Match—2



Dotplot

	m	a	d	a	m	i	m	a	d	a	m
m	●				●		●				●
a		●		●				●		●	
d			●						●		
a		●		●				●		●	
m	●				●		●				●
i						●					
m	●				●		●				●
a		●		●				●		●	
d			●						●		
a		●		●				●		●	
m	●				●		●				●

Advanced Dotplot



Hash Table

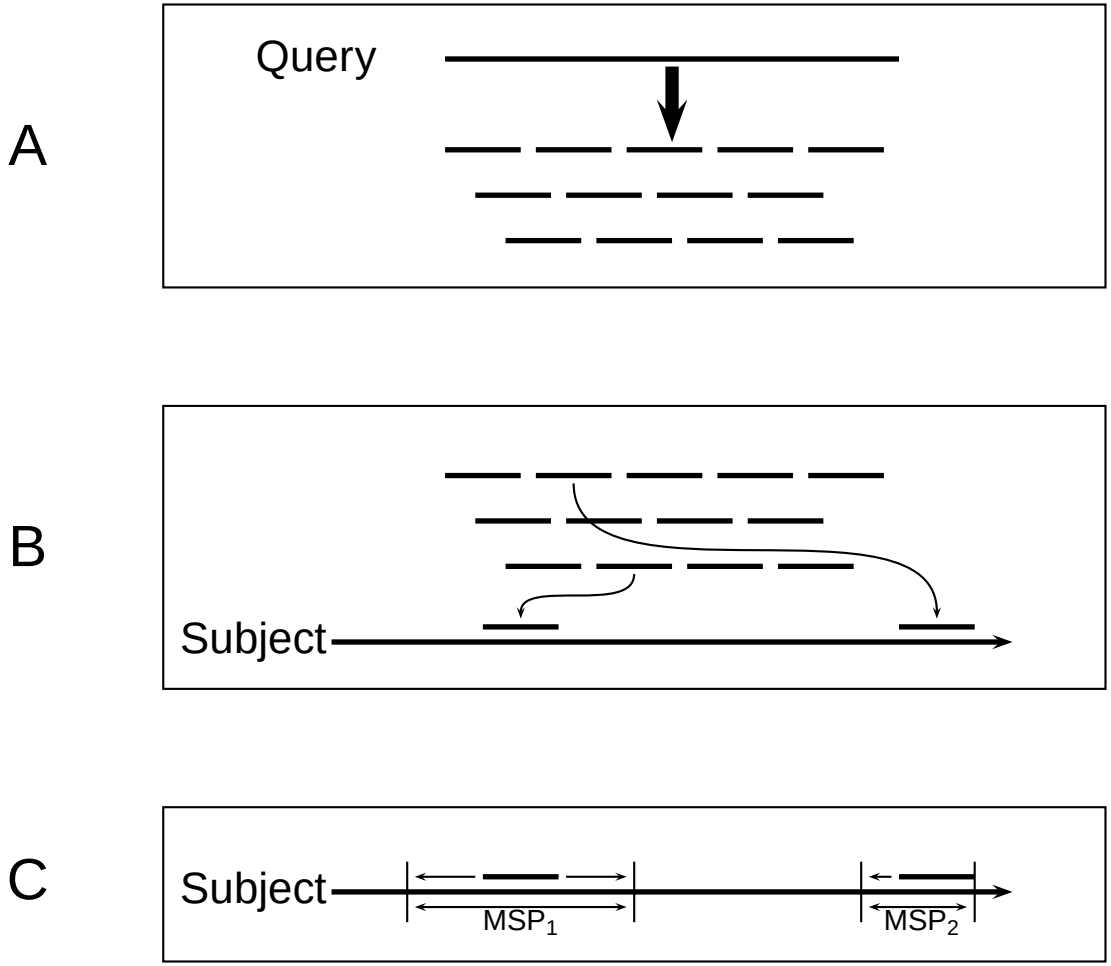
$T = \texttt{ACCAGAGAATT}$

word	position
AC	1
CC	2
CA	3
AG	4, 6
GA	5, 7
AA	8
AT	9
TT	10

FASTA

1. Hash query into words of length *ktup*. Use hash table to localize exact matches in subject. Matches can be imagined as diagonals in dotplot.
2. Rescore 10 diagonals with the highest density of matches.
3. Join diagonals with score $\geq$ threshold into region.
4. Align joined region using dynamic programming.

BLAST



High-Scoring Words

Member of Word List	Match in Query	Score
AH	AH	12
CH	AH	8
GH	AH	8
SH	AH	9
TH	AH	8
VH	AH	8
HY	HY	15
NY	HY	8
YY	HY	9
HF	HY	11
HW	HY	10
YV	YV	11
YI	YV	10
YL	YV	8
YM	YV	8

Detect Protein Families

Proteome

$\{p_1, p_2, p_3, p_4, p_5\}$ $\longrightarrow$

Match Matrix

	p_1	p_2	p_3	p_4	p_5
p_1	0	1	0	0	0
p_2	1	0	0	1	0
p_3	0	0	0	0	1
p_4	0	1	0	0	0
p_5	0	0	1	0	0

$\longrightarrow$

Protein Families

$f_1 = \{p_3, p_5\}$, $f_2 = \{p_1, p_2, p_4\}$

Extreme Value Distribution — 1

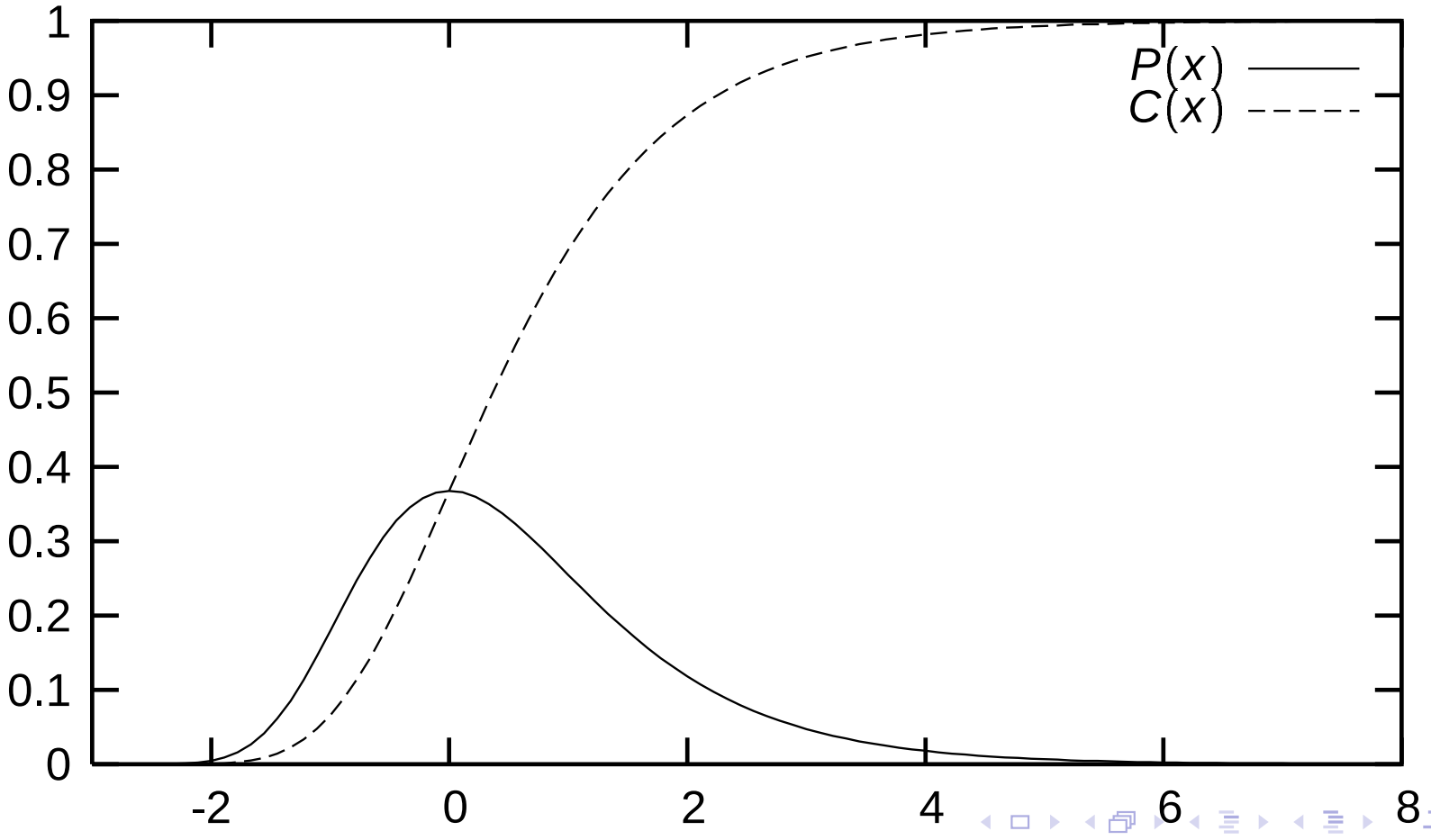
Probability density function:

$$P(x) = \lambda e^{(\mu - x)\lambda - e^{(\mu - x)\lambda}}.$$

Cumulative density function:

$$C(x) = e^{-e^{(\mu - x)\lambda}}.$$

Extreme Value Distribution — 2



Parameter Estimation

To find λ, solve

$$1 = \sum_{i,j} p_i p_j e^{\lambda s_{ij}}.$$

μ is given by

$$\mu = \frac{\ln(Kmn)}{\lambda},$$

where m & n are the *effective* lengths of query & subject.

Example Alignment

Optimal local alignment between human β-hemoglobin (146 aa) & leghemoglobin (153 aa):

```
human β-globin:        50 TPDAVMGNPKVKAHGKKV 67
                          |  .     ||...||   ||
lupine leghemoglobin:  50 TSEVPQNNPELQAHAGKV 67
```

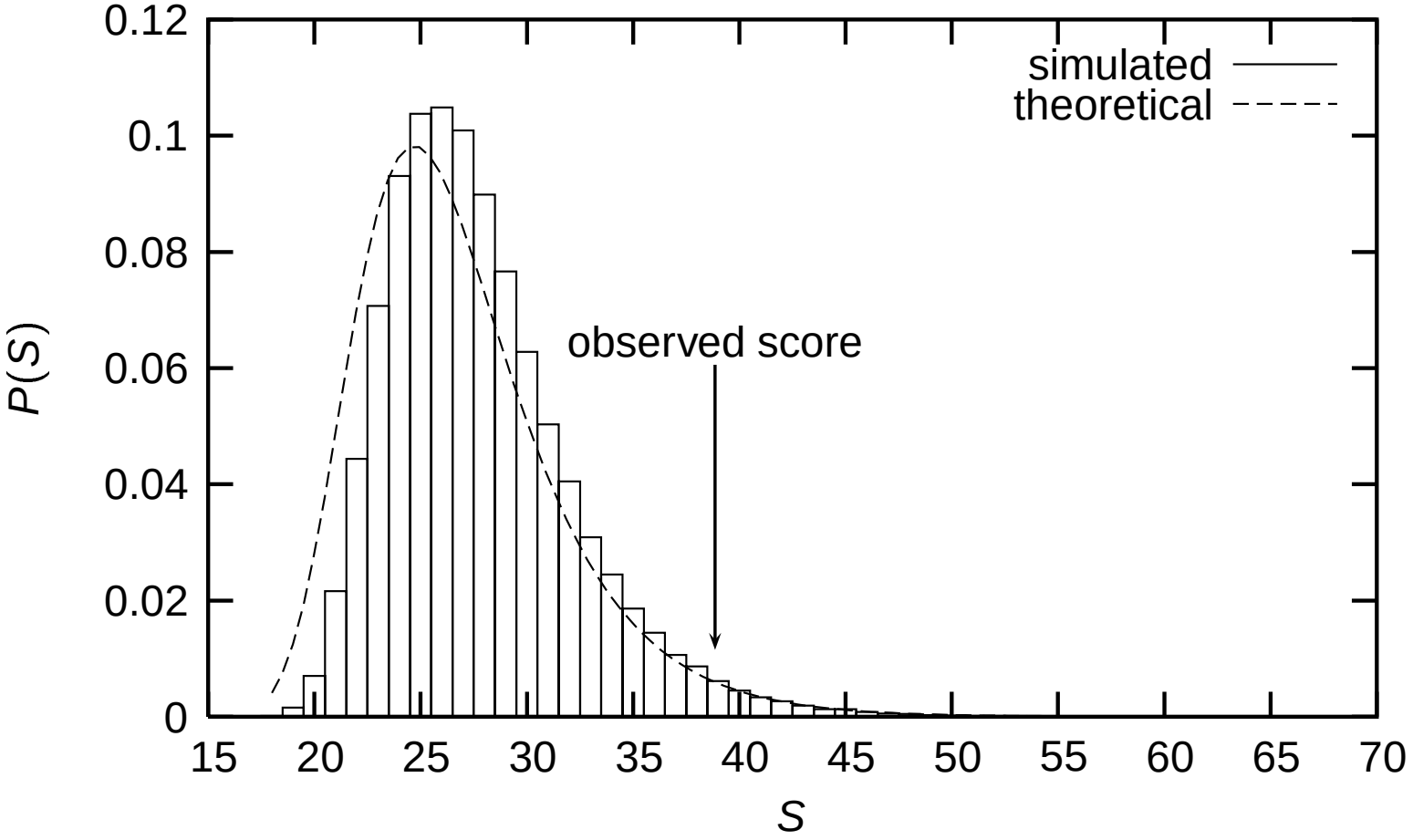
- Parameters: gap opening: -12; gap extension: -1; BLOSUM62
- Score: 39

BLAST Parameters

1. effective length of β-hemoglobin: 130
2. effective length of leghemoglobin 137
3. $\lambda = 0.267$
4. $K = 0.041$

Exercise: Calculate $P(S \geq 39)$

Simulated Null Distribution



Expectation Value

Simulated $P(S \geq 39) = 0.024$; theoretical $P(S \geq 39) = 0.022$.

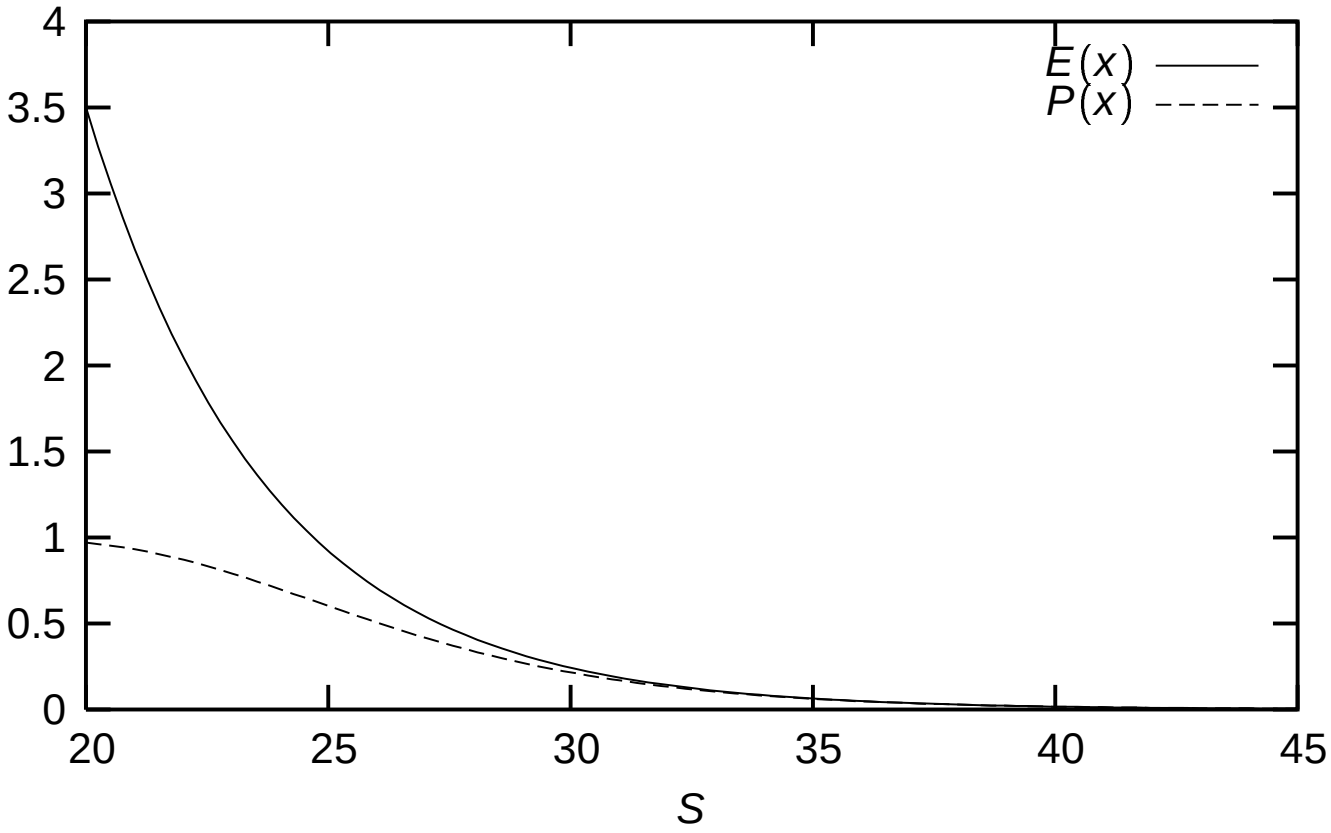
Instead of $P(S \geq x)$ some BLAST-implementations, e.g. NCBI-BLAST return the

expectation-value Number of alignments with $S \geq x$ expected by chance alone

Multiply the complement of the cumulative density function by μ and substitute the definition of μ to get

$$E(S \geq x) = Kmne^{-\lambda x}$$

$P(S \geq x)$ & $E(S \geq x)$



Bounds

- $0 \leq P(S \geq x) \leq 1$
- $0 \leq E(S \geq x) \leq Kmn$

Bit-Scores

$$
\begin{aligned}
S' &= \frac{\lambda S - \ln K}{\ln 2} \\
E(S' \geq x) &= mn2^{-x} \\
P &= 1 - e^{-E}
\end{aligned}
$$