Sequence Models

Dynamic Programming

Sequence Alignment

Sequence Models (1)

eiid: independently indentically distributed with **equal** probabilities:

$$p_a = p_c = p_g = p_t = 1/4.$$

No parameter.

niid: independently identically distributed with **non-equal** probabilities:

$$p_a + p_c + p_g + p_t = 1.$$

3 parameters.

iid models are often used as reference for other models.

Applications: distinguish E. coli DNA from others.

Sequence models (2)

niid under Chargaff Parity Rule II:
$p_c \approx p_g$, $p_a \approx p_t$. Take $\approx$ to be $=$:

$$2p_c + 2p_a = 1$$

One parameter, e.g., the $G+C$ or simply GC content.

Note: The Chargaff Rules:

1. Parity Rule I (1948-1950): $p_c = p_g$, $p_a = p_t$ in **double**-strand DNA.
2. Parity Rule II (1968): $p_c \approx p_g$, $p_a \approx p_t$ in **single**-strand DNA.
3. Cluster Rule (1963): 60% of Y runs together.
4. GC Rule (1951, 1979): GC is a constant in a species.

Erwin Chargaff, How students got a chemical education, *Annals N.Y. Acad. Sci.* **325** (1979) 345 – 361.

Sequence Models (3)

MMn — Markov Chain Model of order n:

$n = 1$

4 initial probabilities: p_a, p_c, p_g, p_t
independent on position in sequence

16 transfer probabilities calculated from 2-tuple frequencies:

$$T_{\alpha\beta} = \begin{pmatrix} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{pmatrix}$$

5 normalization conditions:

$$\sum_{\alpha \in \{a,c,g,t\}} p_\alpha = 1,$$

$$\sum_{\beta \in \{a,c,g,t\}} p_{\alpha\beta} = 1, \quad \forall \alpha \in \{a, c, g, t\}$$

4+16-5=15 parameters.

Sequence Models (4)

MMn — Markov Chain Model of order n:

$n = 2$

16 initial probabilities, nearest neighbor correlation taken into account, no other positional dependence.

64 transfer probabilities calculated from 3-tuple frequencies.

17 normalization conditions: 1 for initial probabilities, 16 for each row of the transfer matrix.

$4^2 + 4^3 - (1 + 4^2) = 63$ parameters.

MM3 and lower are not capable to account for DNAs (known since 1980s)

MM5 are widely used in gene-finding programs for introns and intergenic regions. A simplest inhomogeneous Markov Model, namely, the period 3 MM5 model is used for exons.

Home work: how many parameters are there in a period 3 MM5 model?

Sequence Models (5)

Weight Matrix — position-dependent but no correlation:

Example:

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9
a	0.37	0.49	0.94	0.92	0.27	0.31	0.37	0.21	0.11
c	0.12	0.21	0.0	0.01	0.22	0.11	0.02	0.25	0.02
g	0.11	0.20	0.0	0.02	0.23	0.34	0.44	0.25	0.02
t	0.40	0.10	0.06	0.05	0.28	0.24	0.17	0.29	0.85

A 4×9 matrix. 27 parameters.

A simpler position-dependent model for the above example, the **consensus sequence**:

$WWAANWRNW$

Drawback of both: no correlations of adjacent nucleotides taken into account. One may combine MMn with Weight Matrix to build more complicated sequence models. Hard to call them Markovian due to limited length of the signal.

Known dataset:

Training set: parameter fitting

Test set: check for expected results.

FP: False positives
FN: False negatives
TP: True positives
TN: True negatives

SN: Sensitivity
$$SN = \frac{TP}{TP + FN}$$

SP: Specificity
$$SP = \frac{TP}{TP + FP}$$

Unknown dataset

The 70% hurdle of all kinds of predictions in bioinformatics.

G+C Content Domains in Genomic Sequences

Isochores of about 300kbp (G. Bernardi, 1985). 4 types of isochores in human genome.

From training data: $p_a^i, p_c^i, p_g^i, p_t^i,\ \ i = 1, 2, 3, 4$ and a **niid** reference set of $p_a^0, p_c^0, p_g^0, p_t^0$ from all sequences in the training set.

For a given unknown sequence calculate the likelihood or odd-ratio:

$$r = \frac{\prod_{\{a,c,g,t\}} (p_\alpha^i)^{N_\alpha}}{\prod_{\{a,c,g,t\}} (p_\alpha^0)^{N_\alpha}}$$

for all i.

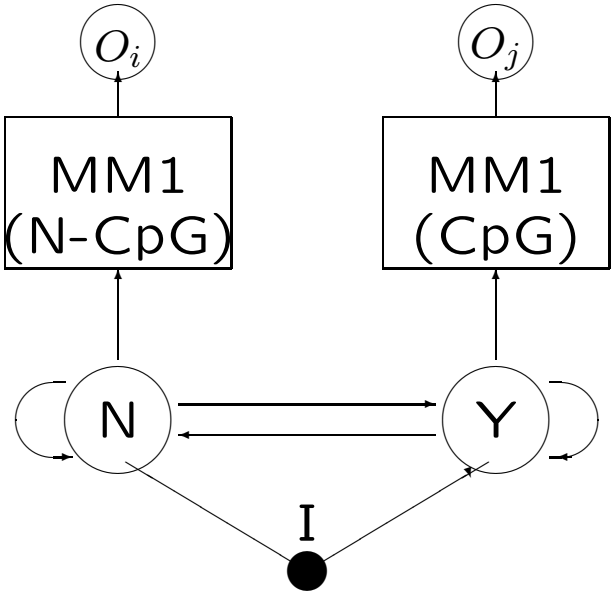
Logarithmic odd-ratio: $\log r$.

Human genome: no isochores, "GC content domains" (E. S. Lander *et al.*, 2001)

CpG Islands

1. niid model: not good as dinucleotide frequencies are concerned.
2. MM1 model: better, but not as good as HMM.
3. HMM model: HMM = Hidden Markov Model

HMM for CpG Islands



Prediction without Training Data (1)

Example: iterative method (Zheng Wei-mou)

Given a contig of 120 000bp. Determine the coding (CDS) and non-coding (NCDS) segments in the contig.

Divide the contig into many non-overlapping segments of 120bp. Label these segments as C and N in an arbitrary way.

Calculate $\{p_\alpha^C\}$,$\{p_\alpha^N\}$, $\{p_\alpha^0\}$ for the collection of C, N, and all segments.

Calculate the likelihood. Keep or change the labels according to the likelihood.

Iterate.

If it converges, the contig is divided into a collection of two different composition groups.

Many possible refinements and modifications.

Prediction without Training Data (2)

See Zheng Wei-mou's recent papers:

1. "Genomic signal enhancement by clustering", *Commun. Theor. Phys.* **39** (2003), 631.

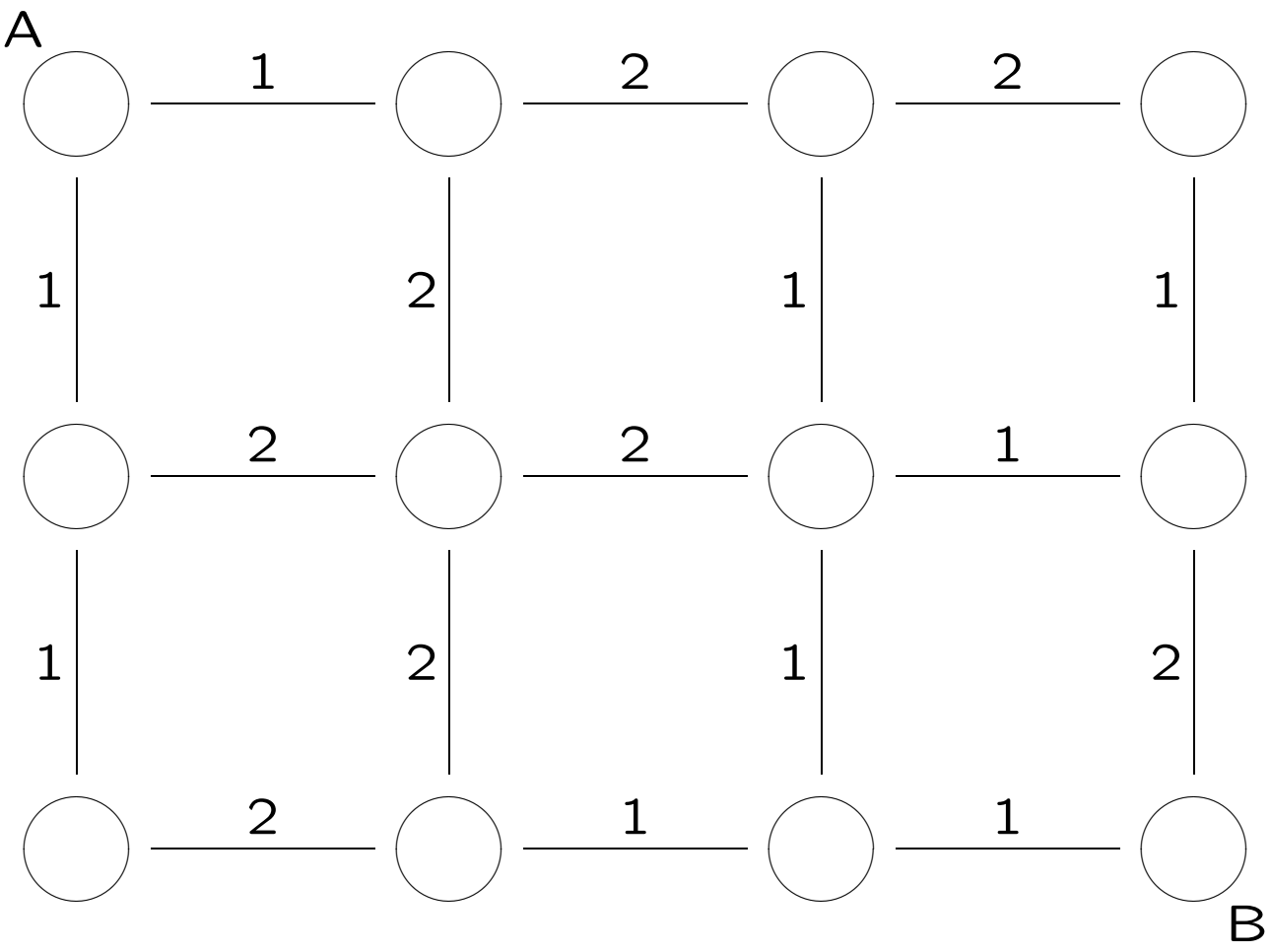
2. "Genomic signal search by dynamic programming", *Commun. Theor. Phys.* **39** (2003), 761.

3. "Finding signals for plant promoters", *Genomics, Proteomics & Bioinformatics* **1** (2003) 68.

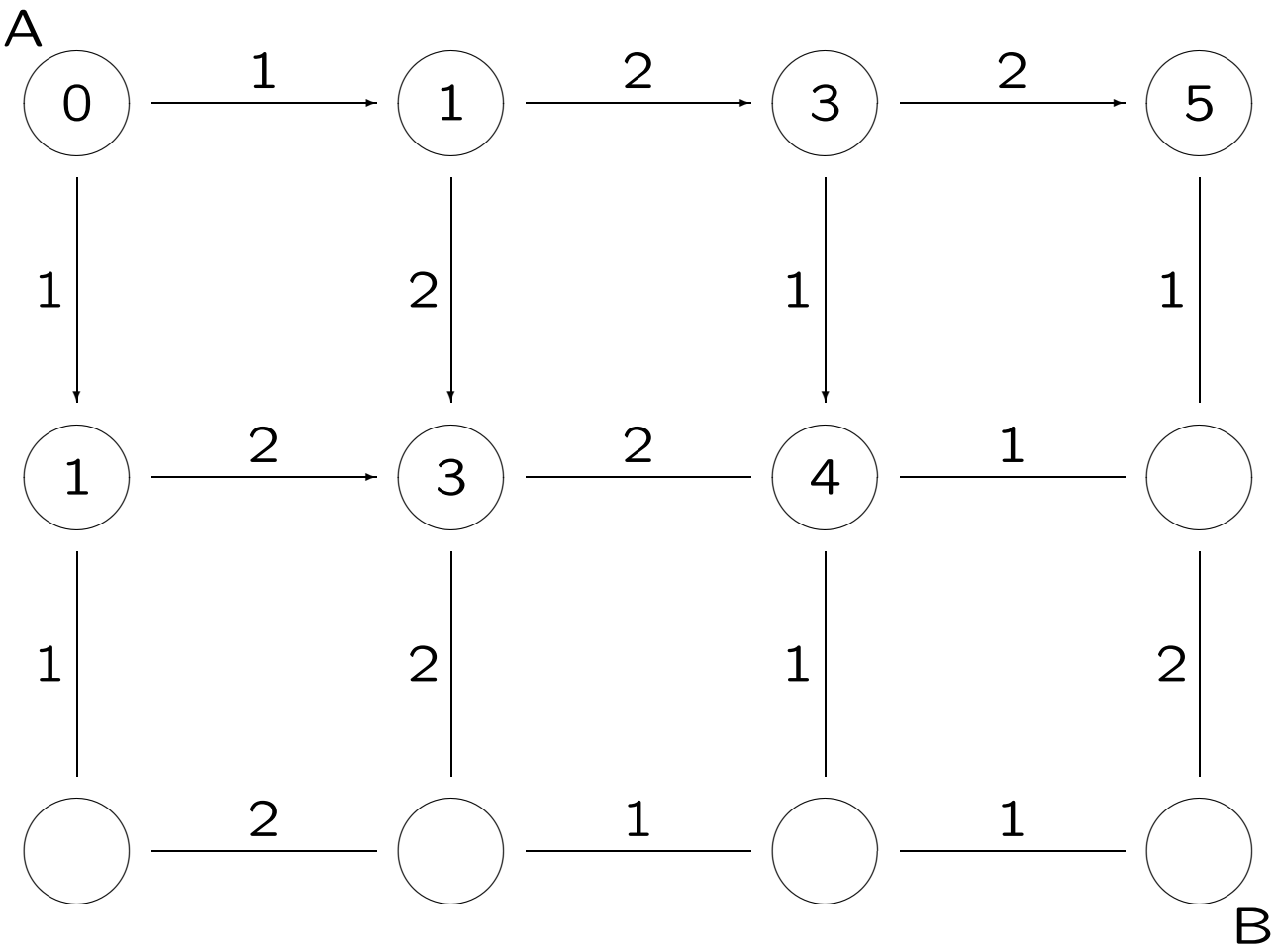
Dynamic Programming

Invented by Richard Bellman in the 1950s

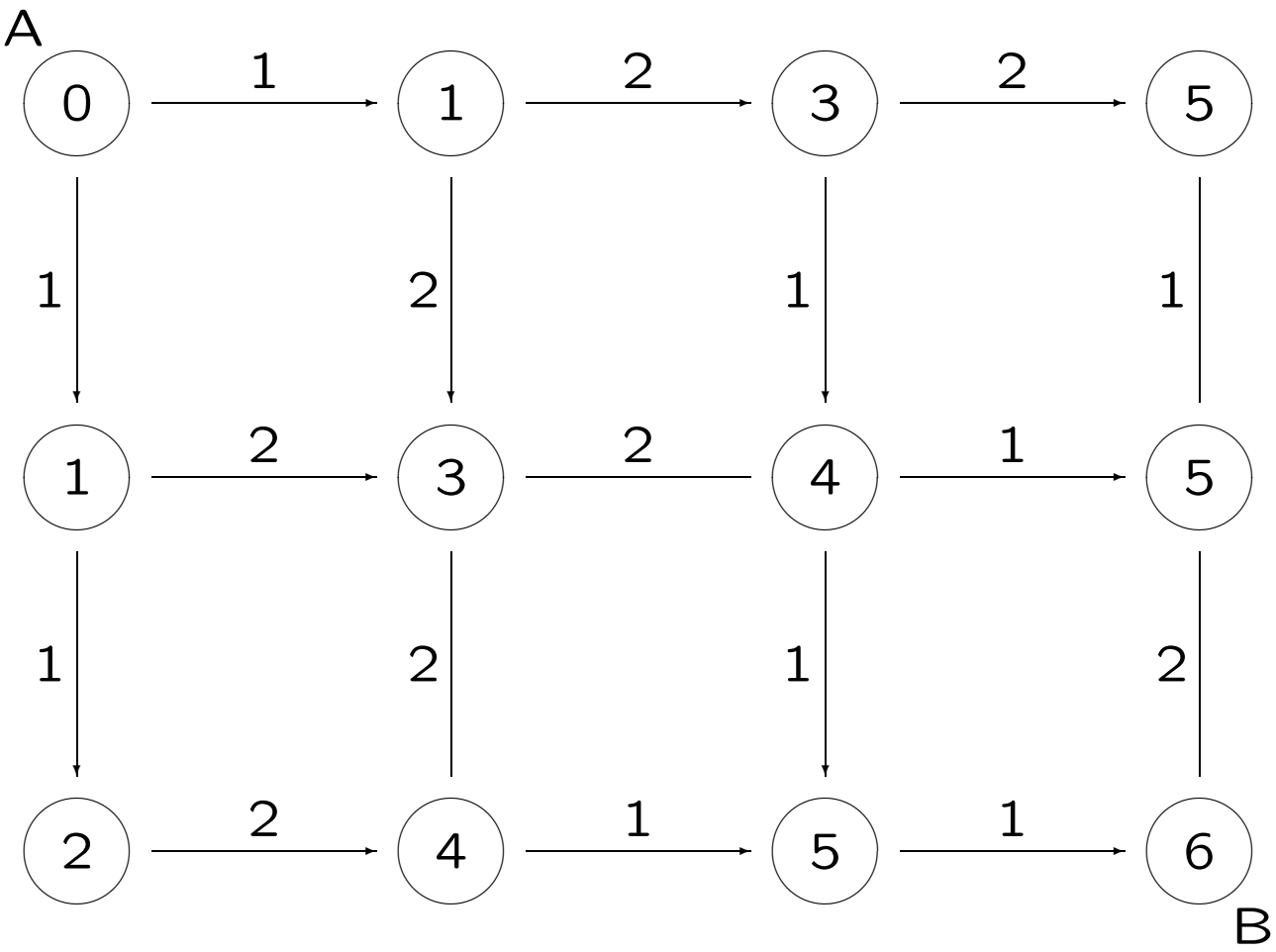
Simple example: find the minimal-toll path from A to B:



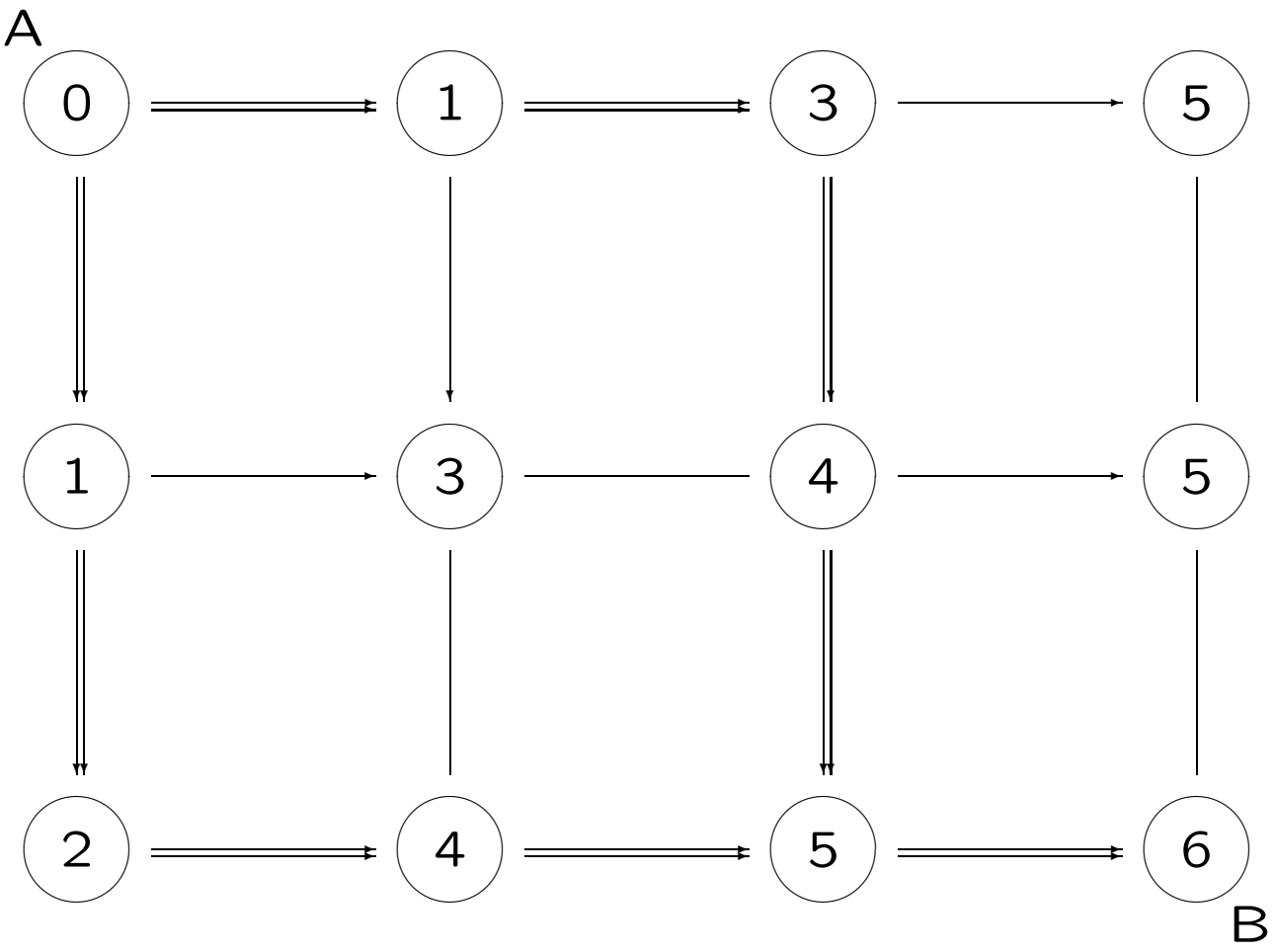
Forward calculation — at intermediate stage:



Forward calculation — final situation:



Tracing back:



Pairwise Sequence Aligment

Given two symbolic sequences of
length N and M, respectively:

$\Sigma = \sigma_1\sigma_2\cdots\sigma_N$

and

$\Pi = \pi_1\pi_2\cdots\pi_M,$

where σ_i and π_j are symbols from
the same alphabet.

Suppose that there is a scoring scheme:

	Score
Exact match $\sigma_i \equiv \pi_j$	$\delta > 0$
Mismatch $\sigma_i \neq \pi_j$	$\beta < \delta$
Gap penalty	$\gamma < 0$

Goal: obtain the alignment(s) with the
largest additive score.

Alignment by Dynamic Programming

Start from position $i = 0$ and $j = 0$ with initial score $S(0, 0) = 0$.

Suppose that we have obtained the best alignment up to position i in sequence Σ and position j in sequence Π with total (additive) score $S(i, j)$. The next score is generated as follows:

$$S(i+1, j+1) = \max \begin{bmatrix} S(i,j) + \delta(\text{if match}), \\ S(i,j) + \beta(\text{if dismatch}), \\ S(i, j+1) + \gamma, \\ S(i+1, j) + \gamma. \end{bmatrix}$$

Mark the path whereby the actual $S(i + 1, j + 1)$ was obtained.

Repeat until we reach $i = N$ and $j = M$ with the highest total score $S(N, M)$. Trace back to figure out the best path which gives the required alignment. There may be more than one alignment with the same highest score.

Home Work

Given two sequences:

ALGORITHMIC

ARITHMETIC

Please obtain the best alignment(s) of these two sequences according to the scoring scheme:

	Score
Match	2
Mismatch	-1
Gap	-2

Scoring Matrix for DNA Sequences

	a	c	g	t
a	1	0	0	0
c	0	1	0	0
g	0	0	1	0
t	0	0	0	1

	a	c	g	t
a	0.9	-0.1	-0.1	-0.1
c	-0.1	0.9	-0.1	-0.1
g	-0.1	-0.1	0.9	-0.1
t	-0.1	-0.1	-0.1	0.9

Scoring matrix used in BLASTN

	a	c	g	t
a	M	N	N	N
c	N	M	N	N
g	N	N	M	N
t	N	N	N	M

$M > 0,\ N < 0$ (default: $M = 5,\ N = -2$)

The PAM250 Scoring Matrix for Amino Acids

PAM = Point Accepted Mutation

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	−2	6																		
N	0	0	2																	
D	0	−1	2	4																
C	−2	−4	−4	−5	12															
Q	0	1	1	2	−5	4														
E	0	−1	1	3	−5	2	4													
G	1	−3	0	1	−3	−1	0	5												
H	−1	2	2	1	−3	3	1	−2	6											
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5										
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6									
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4

R. M. Schwartz, M. O. Dayhoff, in *Atlas of Protein Sequence and Structure*, ed. by M. O. Dayhoff, 345–352, 353–358, 1978.

The BLOSUM62 Scoring Matrix for Amino Acids

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	−1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Comparison of PAM and BLOSUM Scoring Matrices

	PAM	BLOSUM
Obtained from	Global alignment of closely related seqs	Local alignment of distantly related seqs
Numbering	The greater the farther	The greater the closer
Values by	Extrapolation from closely related seqs	direct calculation
Default	PAM250	BLOSUM62

General Form of Scoring Matrices (Karlin and Altschul 1990)

Under two quite general conditions:

1. At least one of the matrix elements $\{s_{ij}\}$ is greater than zero.
2. The expectation of all the matrix elements is negative:

$$\sum_{ij} p_i p_j s_{ij} < 0$$

any scoring matrix must be of the form

$$s_{ij} = \frac{1}{\lambda} \log \frac{q_{ij}}{p_i p_j},$$

where λ is a scaling factor and all the biology is contained in q_{ij}.

Proof

Define an auxiliary function

$$f(x) = \sum_{ij} p_i p_j e^{s_{ij} x},$$

whose behavior is determined by

1. $f(0) = \sum_{ij} p_i p_j = 1.$
2. $f'(x) = \sum_{ij} p_i p_j s_{ij} e^{s_{ij} x},$
   $f'(0) = \sum_{ij} p_i p_j s_{ij} < 0.$
3. $f'' = \sum_{ij} p_i p_j s_{ij}^2 e^{s_{ij} x} > 0$ everywhere.

Therefore, $f(x)$ is a concave up function. $f(x)$ decreases from 1 near $x = 0$. However, since there is at least one $s_{ij} > 0$, $f(x)$ must diverse for x big enough. There must be an $x = \lambda$ where the equality holds:

$$f(x = \lambda) = \sum_{ij} p_i p_j e^{s_{ij} \lambda} = 1.$$

Let $q_{ij} \equiv p_i p_j e^{s_{ij} \lambda}$ **QED**

λ and q_{ij} have important meaning for scoring systems:

1. It does not matter if one multiply all s_{ij} by a factor, say, 10. Nothing changes if $\lambda \rightarrow 1/\lambda$. Therefore, λ is a scale factor and $\lambda s_{ij} = \text{const}_{ij}$. In particular, ln may take any base and s_{ij} acquires a unit. There was a time when the only searching tool was FASTA, people used scores without mentioning their units. Base e leads to *nits*, $\log_2$ leads to *bits*. In practice, all s_{ij} are integers, because the calculated floating numbers have been rounded off to integers.

2. What is the highest score expected? The result was simple, but the proof took S. Karlin and Demko much work (*Ann. Prob.* **22**, 2022-2039). Their paper was almost unreadable.

   Before considering the highest score, let us determine the number of distinct local aligments that have a score larger than

a preset value x. On what factors this number will depend?

- It depends on the length of the two sequences to be compared, m and n, usually we use search space size $N = m \times n$. $E(x) \propto mn$.
- It depends on the scores s_{ij}.
- It should decay exponentially with growing x: $E(x) \propto e^{-\lambda x}$ We would like to have this constant be the λ defined before. It was indeed proved by Karlin and Demko.

Therefore, we have

$$E(x) = KNe^{-\lambda x}.$$

This is an asymptotic relation for

$$m, n \gg 1.$$

Karlin and Demko have formula to calculate K. It is expressed by an infinite series with geometric convergence. For many scoring schemes K is of the order of 1/10. We see that x obeys a Poisson distribution.