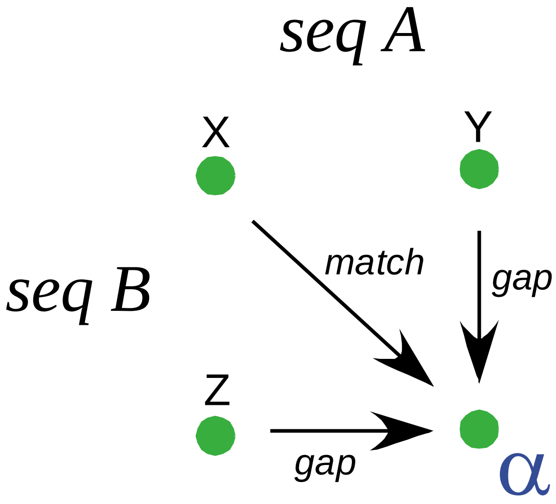
Sequence comparison: Score matrices

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Informal inductive proof of best alignment path

Consider the last step in the best alignment path to node α below. This path must come from one of the three nodes shown, where X, Y, and Z are the cumulative scores of the best alignments up to those nodes. We can reach node α by three possible paths: an A-B match, a gap in sequence A or a gap in sequence B:



The best-scoring path to α is the maximum of:

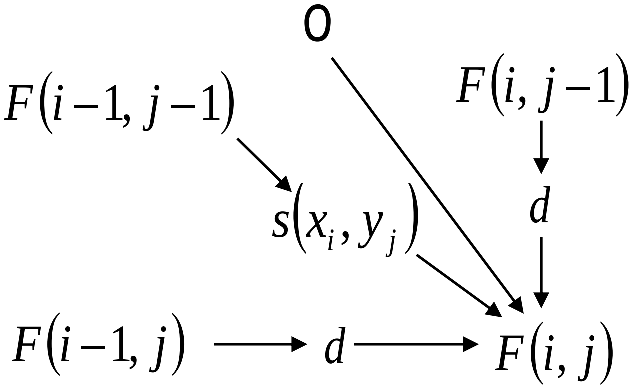
X + match
Y + gap
Z + gap

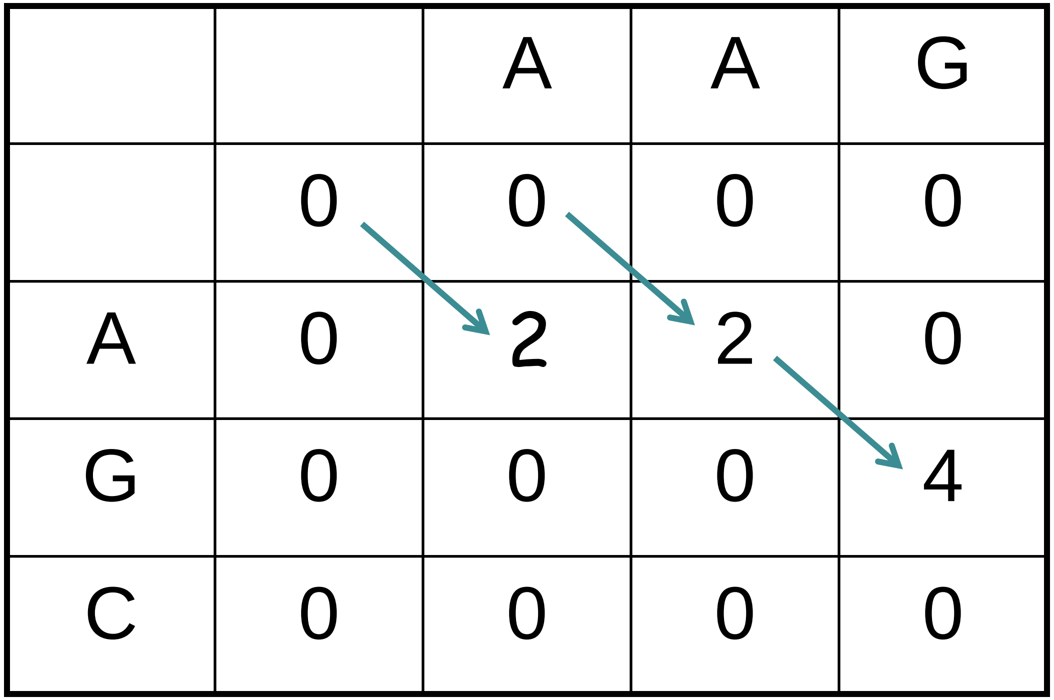
BUT the best paths to X, Y, and Z are analogously the max of their three upstream possibilities, etc. Inductively QED.

Local alignment

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5





		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

(no arrow means no preceding alignment)

Local alignment

- Two differences from global alignment:
  - If a score is negative, replace with 0.
  - Traceback from the highest score in the matrix and continue until you reach 0.
- Global alignment algorithm: *Needleman-Wunsch.*
- Local alignment algorithm: *Smith-Waterman.*

Protein score matrices

- DNA score matrices are much simpler (and are conceptually similar).

- Quantitatively represent the degree of conservation of typical amino acid residues over evolutionary time.

- All possible amino acid changes are represented (matrix of size at least 20 x 20).

- Most commonly used are several different BLOSUM matrices derived for different degrees of evolutionary divergence.

BLOSUM62 Score Matrix

BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
Cluster Percentage: >= 62

regular 20 amino acids

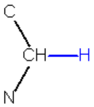
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3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

ambiguity codes
and stop

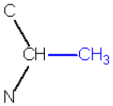
Amino acid structures

Hydrophobic

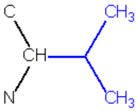
glycine G



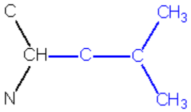
alanine A



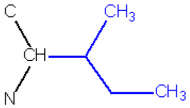
valine V



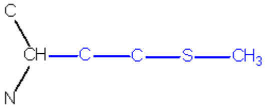
leucine L



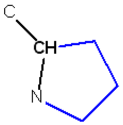
isoleucine I



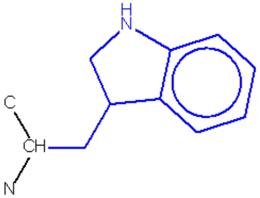
methionine M



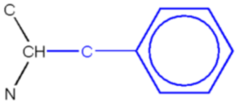
proline P



tryptophan W

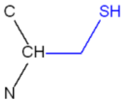


phenylalanine F

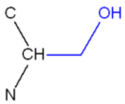


Polar

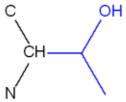
cysteine C



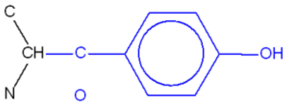
serine S



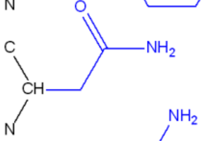
threonine T



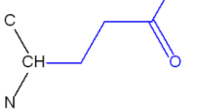
tyrosine Y



asparagine N

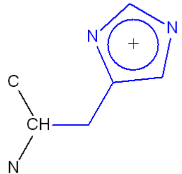


glutamine Q

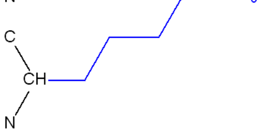


Charged

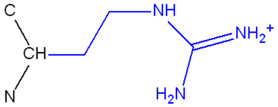
histidine H



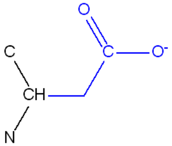
lysine K



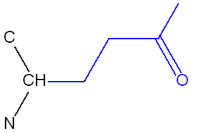
arginine R



aspartate D



glutamate E



BLOSUM62 Score Matrix

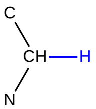
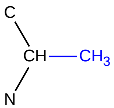
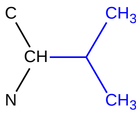
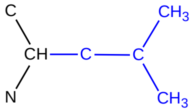
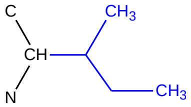
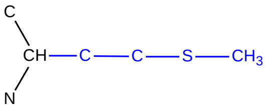
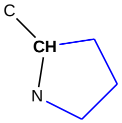
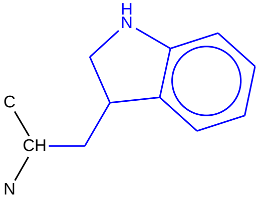
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2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

Good scores – chemically similar

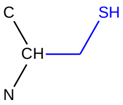
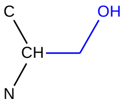
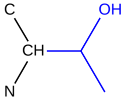
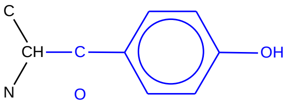
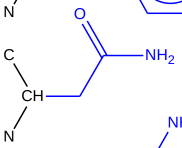
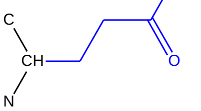
Bad scores – chemically dissimilar

Amino acid structures

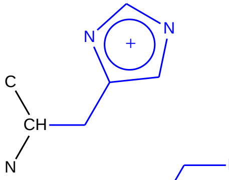
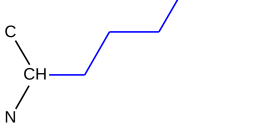
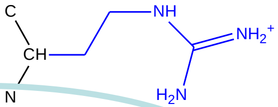
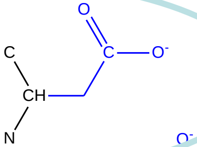
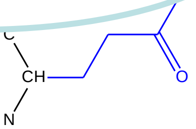
Hydrophobic

glycine	G	
alanine	A	
valine	V	
leucine	L	
isoleucine	I	
methionine	M	
proline	P	
tryptophan	W	

Polar

cysteine	C	
serine	S	
threonine	T	
tyrosine	Y	
asparagine	N	
glutamine	Q	

Charged

histidine	H	
lysine	K	
arginine	R	
aspartate	D	
glutamate	E	

Deriving BLOSUM scores

- Find sets of sequences whose alignment is thought to be correct (this is partly bootstrapped by alignment).

- Measure how often various amino acid pairs occur in the alignments.

- Normalize this to the expected frequency of such pairs randomly in the same set of alignments.

- Derive a log-odds score (often in half bits).

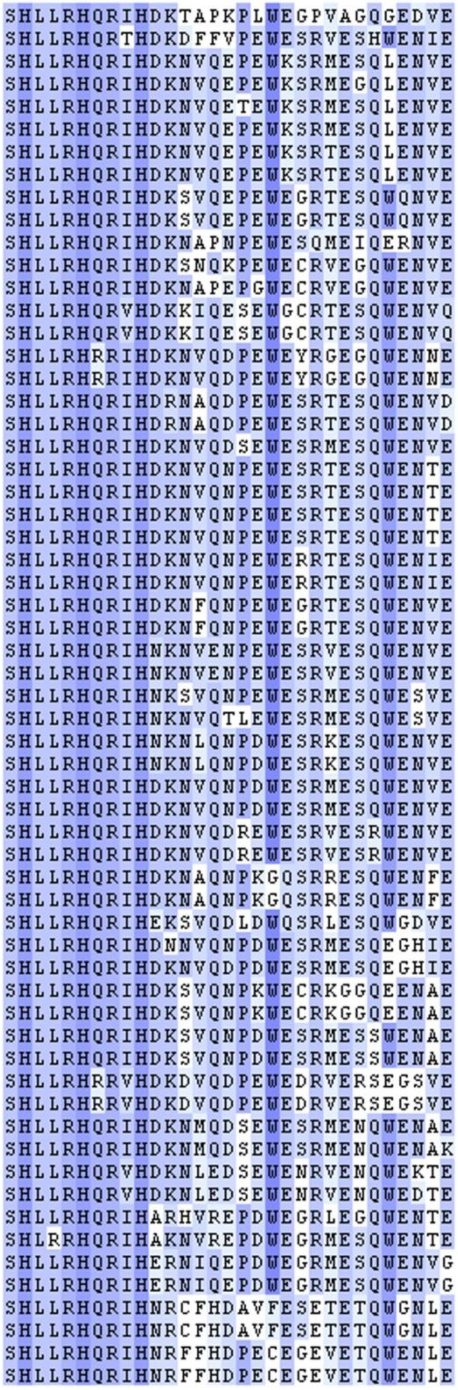
Example of alignment block

31 amino acids (columns)
61 sequences (rows)

- Thousands of such blocks go into computing a single BLOSUM matrix.

- Represent full diversity of sequences.

- Results are summed over all columns of all blocks.



```
SHLLRHQRIHDKTAPKPLWEGPVAGQGEDVE
SHLLRHQRTHDKDFFVPEWESRVESHWENIE
SHLLRHQRIHDKNVQEPEWKSRMESQLENVE
SHLLRHQRIHDKNVQEPEWKSRMEGQLENVE
SHLLRHQRIHDKNVQETEWKSRMESQLENVE
SHLLRHQRIHDKNVQEPEWKSRMESQLENVE
SHLLRHQRIHDKNVQEPEWKSRTESQLENVE
SHLLRHQRIHDKNVQEPEWKSRTESQLENVE
SHLLRHQRIHDKSVQEPEWEGRTESQWQNVE
SHLLRHQRIHDKSVQEPEWEGRTESQWQNVE
SHLLRHQRIHDKNAPNPEWESQMEIQERNVE
SHLLRHQRIHDKSNQKPEWECRVEGQWENVE
SHLLRHQRIHDKNAPEPGWECRVEGQWENVE
SHLLRHQRVHDKKIQESEWGCRTESQWENVQ
SHLLRHQRVHDKKIQESEWGCRTESQWENVQ
SHLLRHRRIHDKNVQDPEWEYRGEGQWENNE
SHLLRHRRIHDKNVQDPEWEYRGEGQWENNE
SHLLRHQRIHDRNAQDPEWESRTESQWENVD
SHLLRHQRIHDRNAQDPEWESRTESQWENVD
SHLLRHQRIHDKNVQDSEWESRMESQWENVE
SHLLRHQRIHDKNVQNPEWESRTESQWENTE
SHLLRHQRIHDKNVQNPEWESRTESQWENTE
SHLLRHQRIHDKNVQNPEWESRTESQWENTE
SHLLRHQRIHDKNVQNPEWESRTESQWENTE
SHLLRHQRIHDKNVQNPEWERRTESQWENIE
SHLLRHQRIHDKNVQNPEWERRTESQWENIE
SHLLRHQRIHDKNFQNPEWEGRTESQWENVE
SHLLRHQRIHDKNFQNPEWEGRTESQWENVE
SHLLRHQRIHNKNVENPEWESRVESQWENVE
SHLLRHQRIHNKNVENPEWESRVESQWENVE
SHLLRHQRIHNKSVQNPEWESRMESQWESVE
SHLLRHQRIHNKNVQTLEWESRMESQWESVE
SHLLRHQRIHNKNLQNPDWESRKESQWENVE
SHLLRHQRIHNKNLQNPDWESRKESQWENVE
SHLLRHQRIHDKNVQNPDWESRMESQWENVE
SHLLRHQRIHDKNVQNPDWESRMESQWENVE
SHLLRHQRIHDKNVQDREWESRVESRWENVE
SHLLRHQRIHDKNVQDREWESRVESRWENVE
SHLLRHQRIHDKNAQNPKGQSRRESQWENFE
SHLLRHQRIHDKNAQNPKGQSRRESQWENFE
SHLLRHQRIHEKSVQDLDWQSRLESQWGDVE
SHLLRHQRIHDNNVQNPDWESRMESQEGHIE
SHLLRHQRIHDKNVQDPDWESRMESQEGHIE
SHLLRHQRIHDKSVQNPKWECRKGGQEENAE
SHLLRHQRIHDKSVQNPKWECRKGGQEENAE
SHLLRHQRIHDKSVQNPDWESRMESSWENAE
SHLLRHQRIHDKSVQNPDWESRMESSWENAE
SHLLRHRRVHDKDVQDPEWEDRVERSEGSVE
SHLLRHRRVHDKDVQDPEWEDRVERSEGSVE
SHLLRHQRIHDKNMQDSEWESRMENQWENAE
SHLLRHQRIHDKNMQDSEWESRMENQWENAK
SHLLRHQRVHDKNLEDSEWENRVENQWEKTE
SHLLRHQRVHDKNLEDSEWENRVENQWEDTE
SHLLRHQRIHARHVREPDWEGRLEGQWENTE
SHLRRHQRIHAKNVREPDWEGRMESQWENTE
SHLLRHQRIHERNIQEPDWEGRMESQWENVG
SHLLRHQRIHERNIQEPDWEGRMESQWENVG
SHLLRHQRIHNRCFHDAVFESETETQWGNLE
SHLLRHQRIHNRCFHDAVFESETETQWGNLE
SHLLRHQRIHNRFFHDPECEGEVETQWENLE
SHLLRHQRIHNRFFHDPECEGEVETQWENLE
```

Pair frequency *vs.* expectation

Actual aligned pair frequency:

$$q_{ij} = \frac{1}{T}\sum c_{ij}$$

where c_{ij} is the count of ij pairs and T is the total pair count.

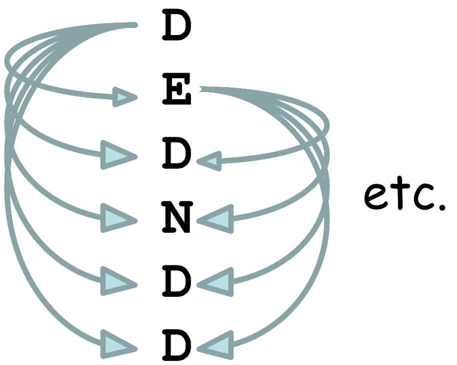
Randomly expected pair frequency:

$$e_{aa} = p_a p_a$$

$$e_{ab} = p_a p_b + p_b p_a = 2 p_a p_b$$

where p_a and p_b are the overall probabilities (frequencies) of specific residues a and b.

Sample column from a multiple alignment:



6 D-D pairs
4 D-E pairs
4 D-N pairs
1 E-N pair

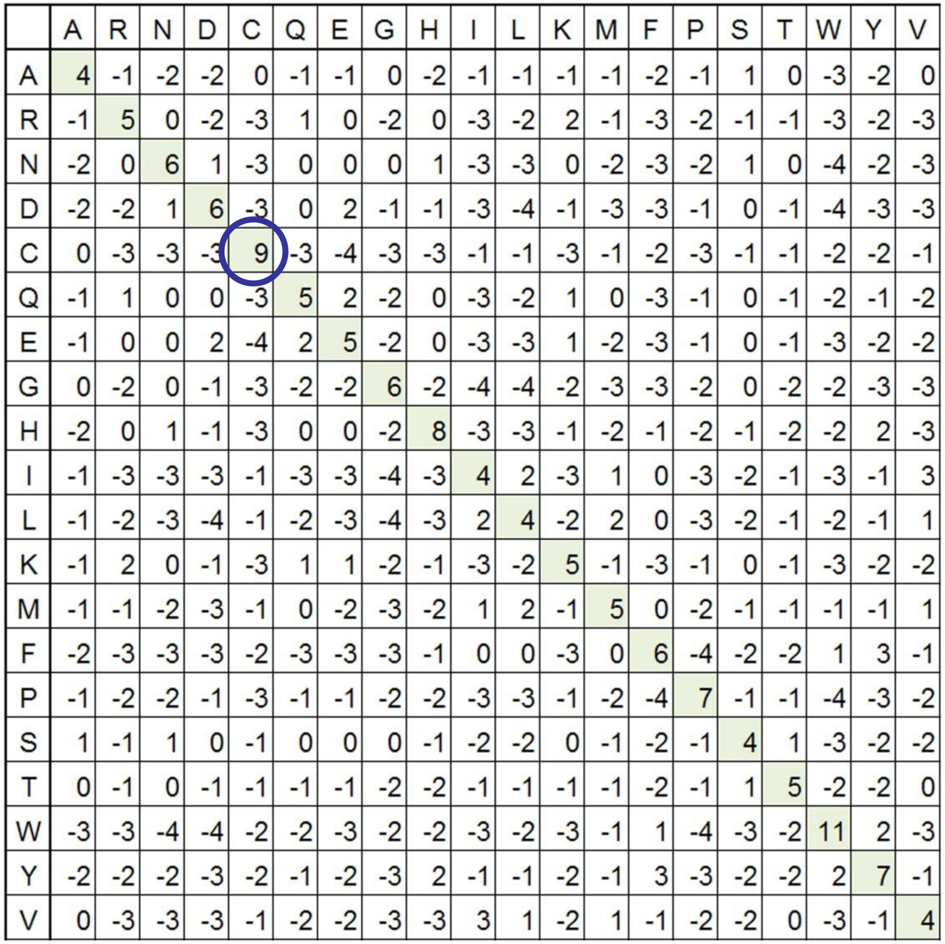
A multiple alignment of N sequences is the equivalent of all the pairwise alignments, which number (N)(N-1)/2.

Log-odds score calculation (so adding scores == multiplying probabilities)

$$s_{ij} = \log_2 \frac{q_{ij}}{e_{ij}}$$

For computational speed often rounded to nearest integer and (to reduce round-off error) they are often multiplied by 2 (or more) first, giving a “half-bit” score:

$$\text{matrixScore} = \text{(rounded) } 2\log_2 \frac{q_{ij}}{e_{ij}}$$



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

BLOSUM62 matrix (half-bit scores)

(9 half-bits = 4.5 bits)

Frequency of C residue over all proteins: 0.0162 (you have to look this up)

Reverse calculation of aligned C–C pair frequency in BLOSUM data set:

C–C $$\frac{q_{cc}}{e_{cc}} = 2^{(4.5)} = 22.63 \qquad e_{cc} = 0.0162 * 0.0162 = 0.000262$$

thus $$q_{cc} = 22.63 * 0.000262 = 0.00594$$

Constructing Blocks

- Blocks are ungapped alignments of multiple sequences, usually 20 to 100 amino acids long.

- Cluster the members of each block according to their percent identity.

- Make pair counts and score matrix from a large collection of similarly clustered blocks.

- Each BLOSUM matrix is named for the percent identity cutoff in step 2 (e.g. BLOSUM70 for 70% identity).

Probabilistic Interpretation of Scores (ungapped)

$$\text{matrixScore} = \text{(rounded) } 2\log_2 \frac{q_{ij}}{e_{ij}}$$ (BLOSUM62)

- By converting scores back to probabilities, we can give a probabilistic interpretation to an alignment score.

- this alignment has a score of 16 (6+2+1+7) by BLOSUM 62, meaning an alignment with this score or more is 2^8 (256) times more likely to be seen in a real alignment than in a random alignment.

```
FIAP
FLSP
```

- this 15 amino acid alignment has a score of 75, meaning that it is ~10^{11} times more likely to be seen in a real alignment than in a random alignment(!!).

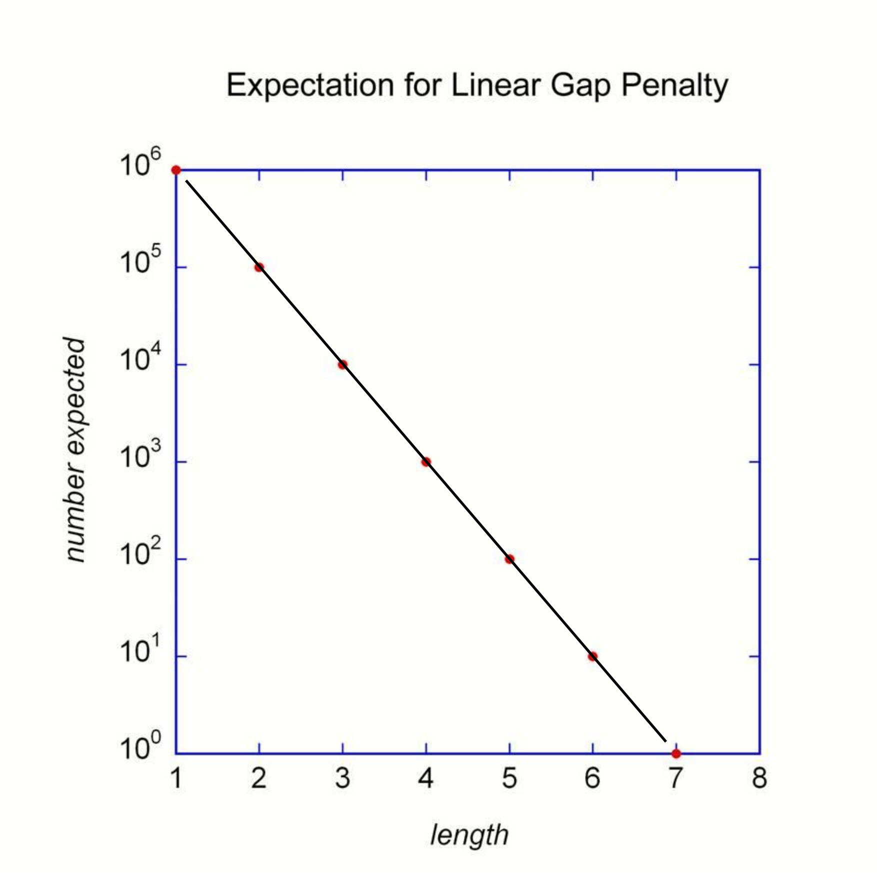
```
VHRDLKPENLLLASK
VHRDLKPENLLLASK
```

(4+8+5+6+4+5+7+5+6+4+4+4+4+4+5)

Randomly Distributed Gaps

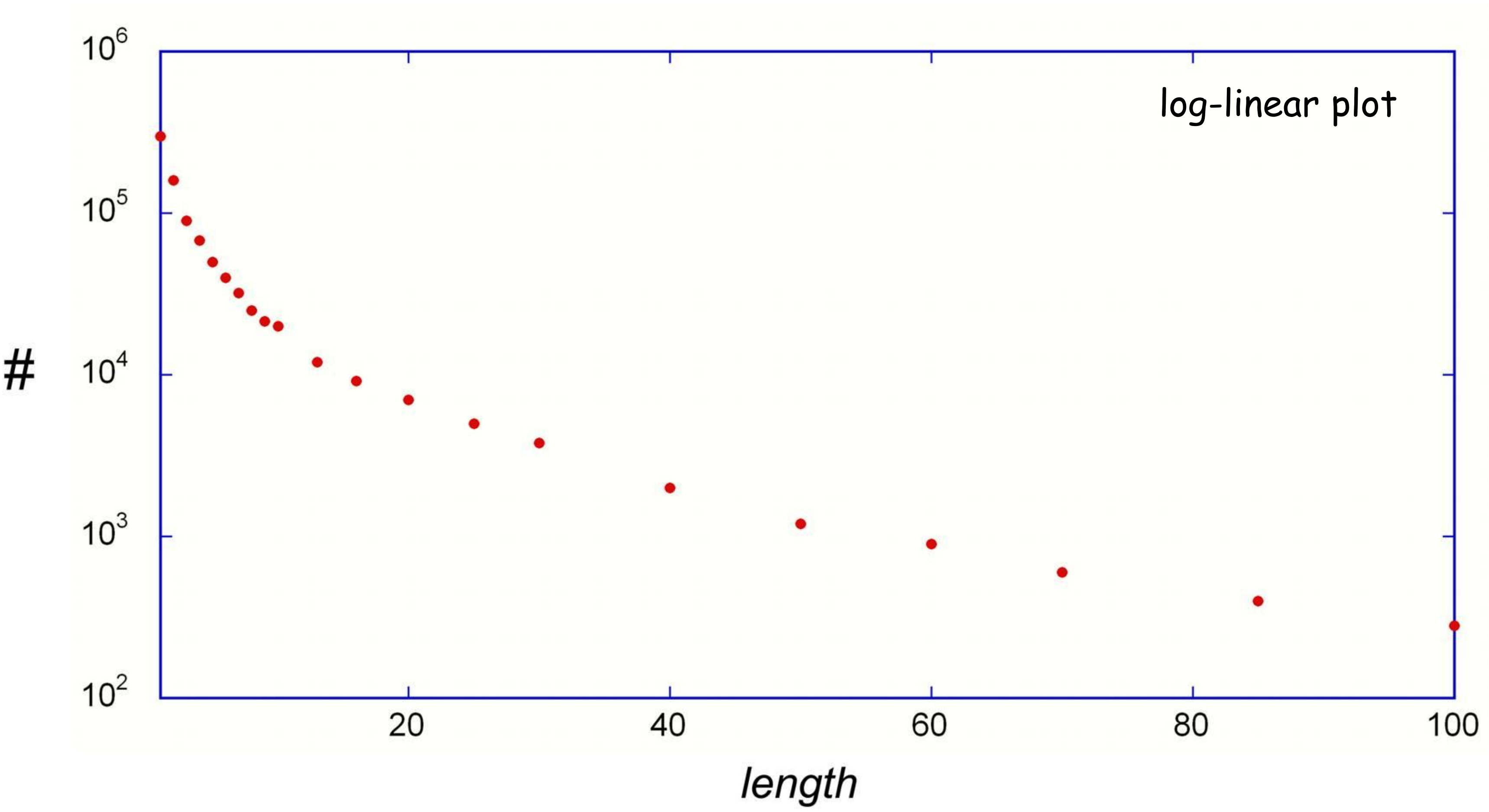
if $p_g = k$ (probability of a gap at each position in the sequence)

then $P(g_1) = k, P(g_2) = k^2, \ldots, P(g_n) = k^n$



[note - the slope of the line on a log-linear plot will vary according to the frequency of gaps, but it will always be linear]

Distribution of alignment gap lengths in large set of structurally-aligned proteins



Summary

- How a score matrix is derived
- What the scores mean probablistically
- Why gap penalties should be affine
- How to use scores in dynamic programming