Chapter III: Pairwise Sequence Alignment Part 2

Presentations use info from:

Jonathan Pevsner, Ph.D.
http://bioinfbook.org
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(3rd edition, ©2015 John Wiley & Sons, Ltd.)
You may use this PowerPoint for teaching purposes

dr. Stanislav Kolenčík
stanislav.kolencik@famnit.upr.si

What will you learn?

To define homology as well as orthologs and paralogs

To explain how PAM (accepted point mutation) matrices are derived

To contrast the utility of PAM and BLOSUM scoring matrices.

To define dynamic programming and explain how global and local pairwise alignments are performed

To perform pairwise alignment of protein or DNA sequences at the NCBI website

Outline

Introduction
- Protein alignment: often more informative than DNA alignment
- Definitions: homology, similarity, identity
- Gaps
- Pairwise alignment, homology, and evolution of life

Scoring matrices
- Dayhoff model: 7 steps
- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local
- Global sequence alignment: algorithm of Needleman and Wunsch
- Local sequence alignment: Smith and Waterman algorithm
- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST
- Basic Local Alignment Search Tool (BLAST)
- Pairwise alignment with dotplots

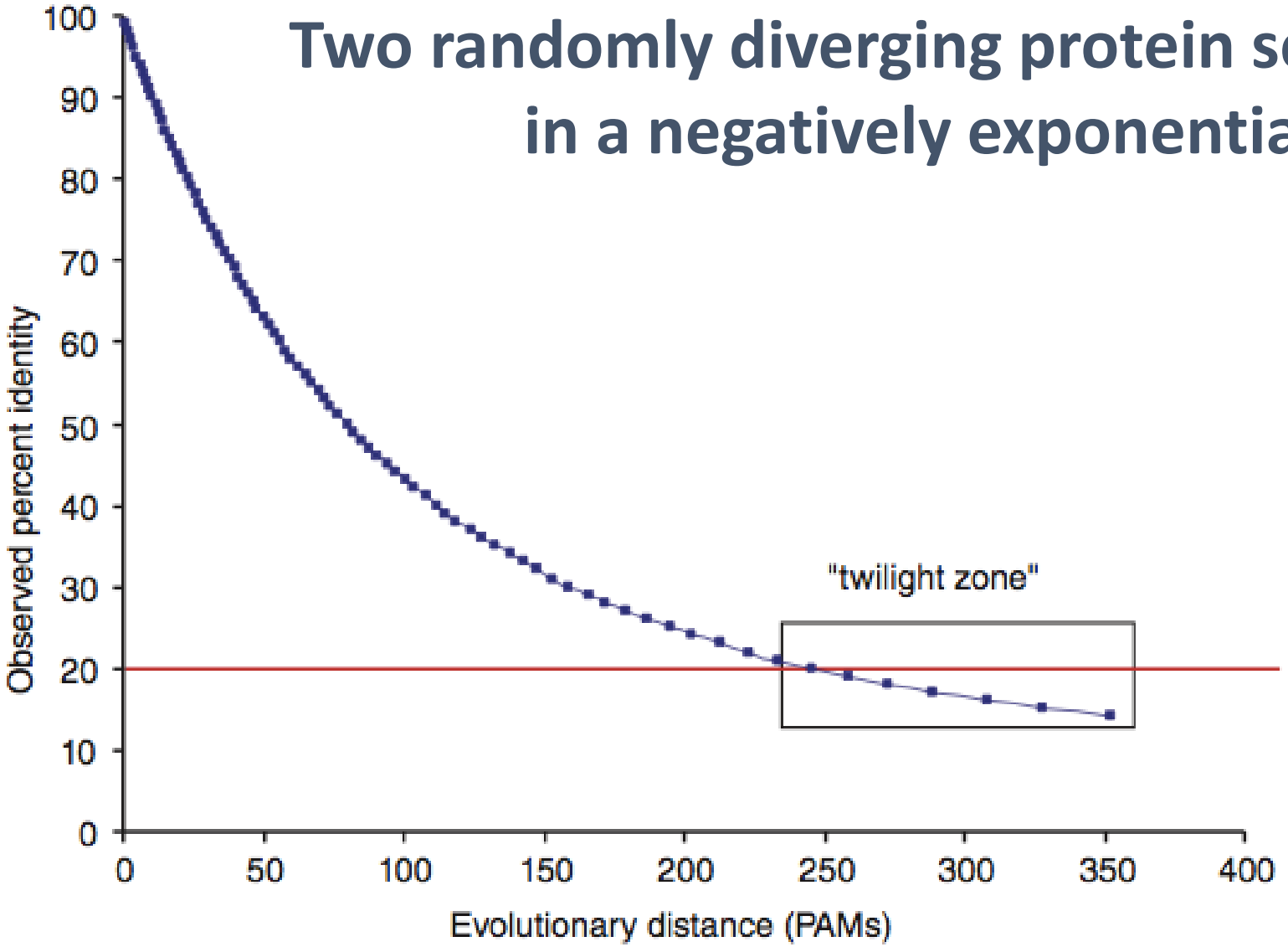
The statistical significance of pairwise alignments
- Statistical significance of global alignments
- Percent identity and relative entropy

Perspective

PAM matrices:
Point-accepted mutations

- **PAM matrices** are based on global alignments of closely related proteins.
- **The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.** At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.
- Other PAM matrices are extrapolated from PAM1. For PAM250, 250 changes have occurred for two proteins over a length of 100 amino acids.
- All the PAM data come from closely related proteins (>85% amino acid identity).

Two randomly diverging protein sequences change in a negatively exponential fashion



- At PAM1, two proteins are 99% identical
- At PAM10.7, there are 10 differences per 100 residues
- At PAM80, there are 50 differences per 100 residues
- At PAM250, there are 80 differences per 100 residues

Correspondence between observed differences (per 100 residues) and evolutionary distance (in PAMs)

Observed differences in 100 residues	Evolutionary distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

PAM 80

PAM 250

Outline

Introduction
- Protein alignment: often more informative than DNA alignment
- Definitions: homology, similarity, identity
- Gaps
- Pairwise alignment, homology, and evolution of life

Scoring matrices
- Dayhoff model: 7 steps
- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local
- **Global sequence alignment: algorithm of Needleman and Wunsch**
- **Local sequence alignment: Smith and Waterman algorithm**
- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST
- Basic Local Alignment Search Tool (BLAST)
- Pairwise alignment with dotplots

The statistical significance of pairwise alignments
- Statistical significance of global alignments
- Percent identity and relative entropy

Perspective

Two kinds of sequence alignment: global and local

- We will first consider the global alignment algorithm of Needleman and Wunsch (1970).
- We will then explore the local alignment algorithm of Smith and Waterman (1981).
- Finally, we will consider BLAST, a heuristic version of Smith-Waterman.

Global alignment with the algorithm of Needleman and Wunsch (1970)

- Two sequences can be compared in a matrix along x- and y-axes.
- If they are identical, a path along a diagonal can be drawn.
- Find the optimal subpaths and add them up to achieve the best score.
  This involves:
  - -adding gaps when needed;
  - -allowing for conservative substitutions;
  - -choosing a scoring system (simple or complicated).
- N-W is guaranteed to find optimal alignment(s).

Three steps to global alignment with the Needleman-Wunsch algorithm

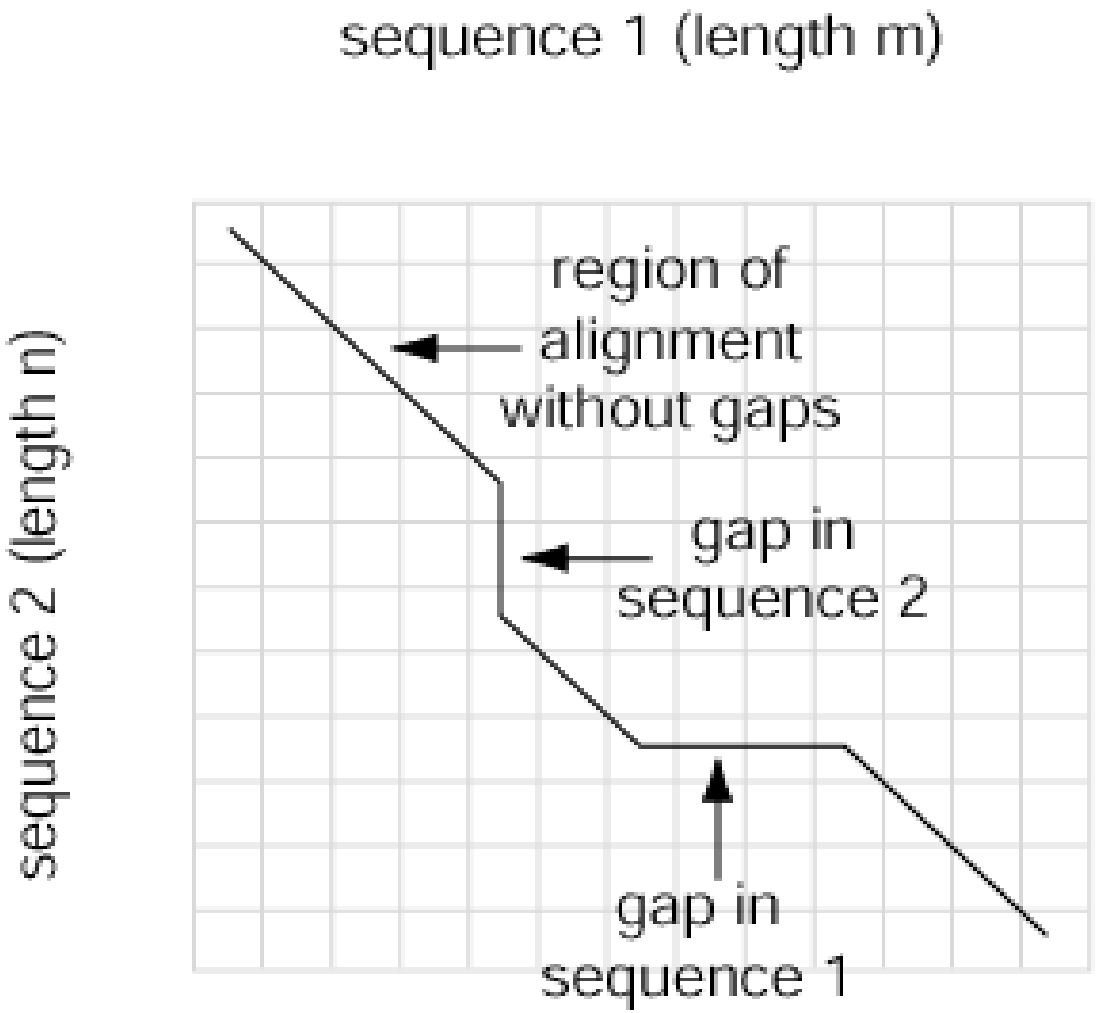
---

[1] set up a matrix

[2] score the matrix

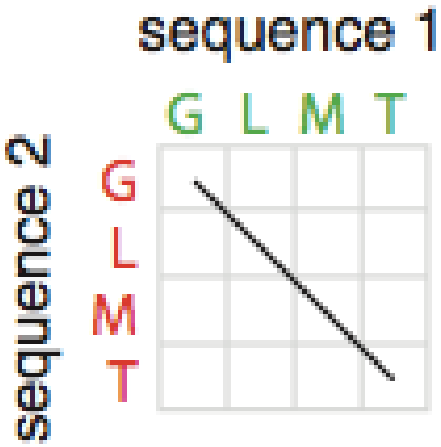
[3] identify the optimal alignment(s)

Four possible outcomes in aligning two sequences

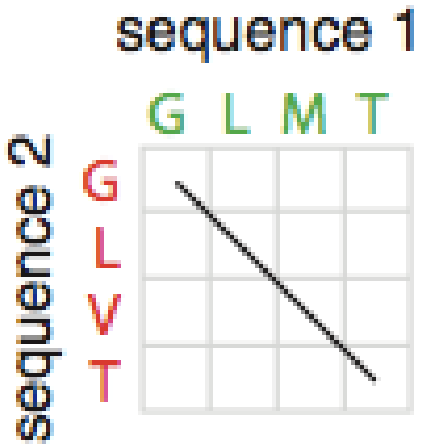


[1] identity (stay along a diagonal)
[2] mismatch (stay along a diagonal)
[3] gap in one sequence (move vertically!)
[4] gap in the other sequence (move horizontally!)

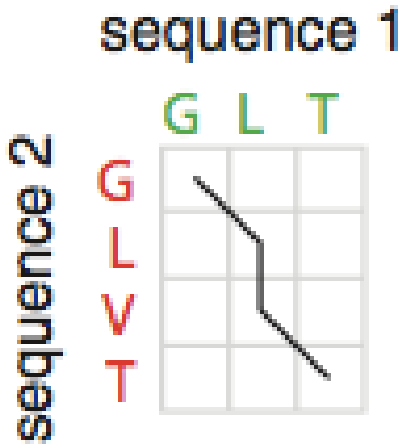
Four possible outcomes in aligning two sequences



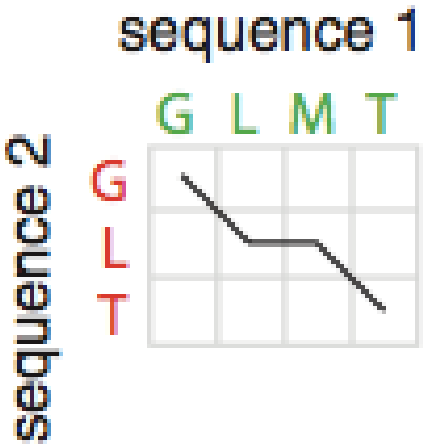
1 GLMT
2 GLMT



1 GLMT
2 GLVT



1 GL-T
2 GLVT

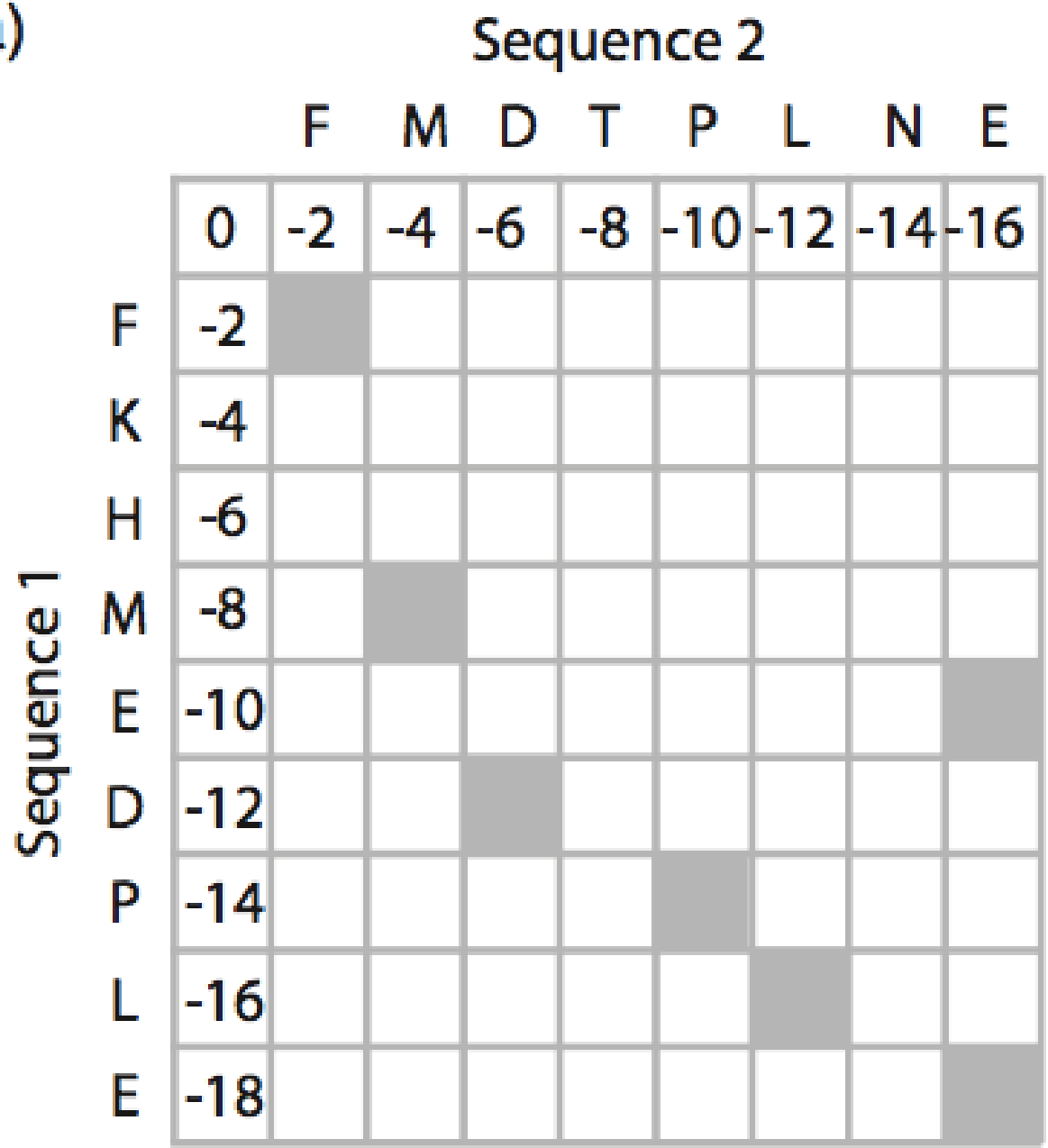


1 GLMT
2 GL-T

Global pairwise alignment using Needleman-Wunsch

Identify positions
of identity (shaded gray).

(a)



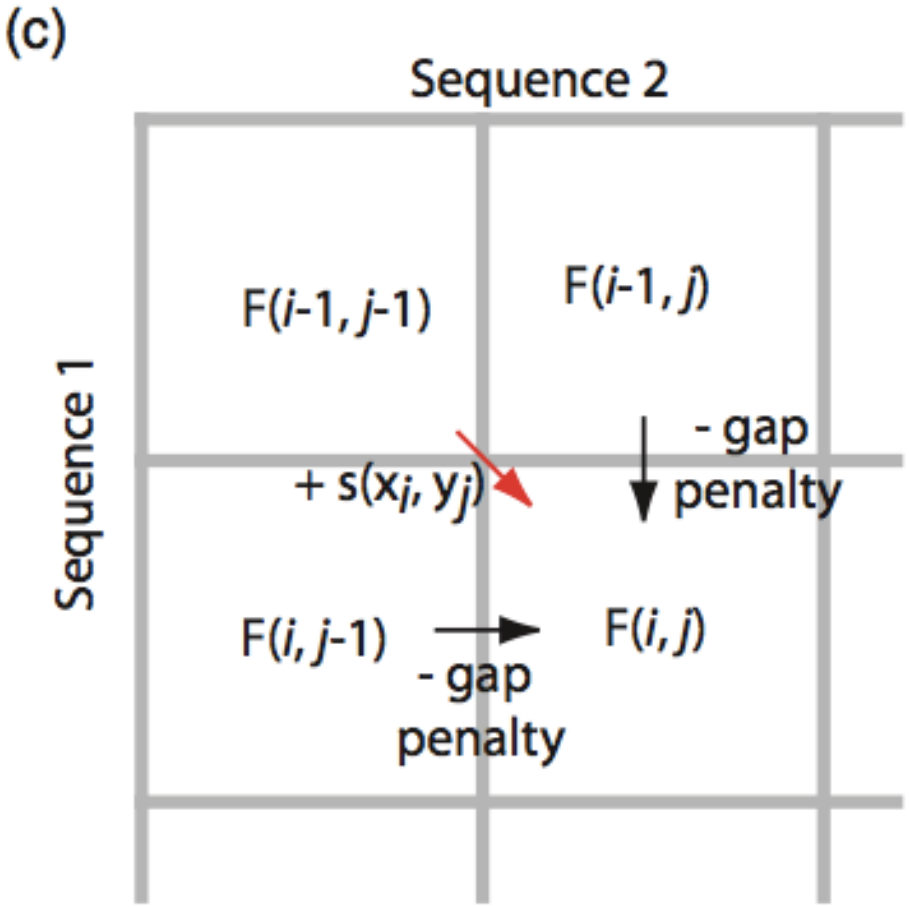
Global pairwise alignment using Needleman-Wunsch

$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
-2 (mismatch)
-2 (gap penalty)

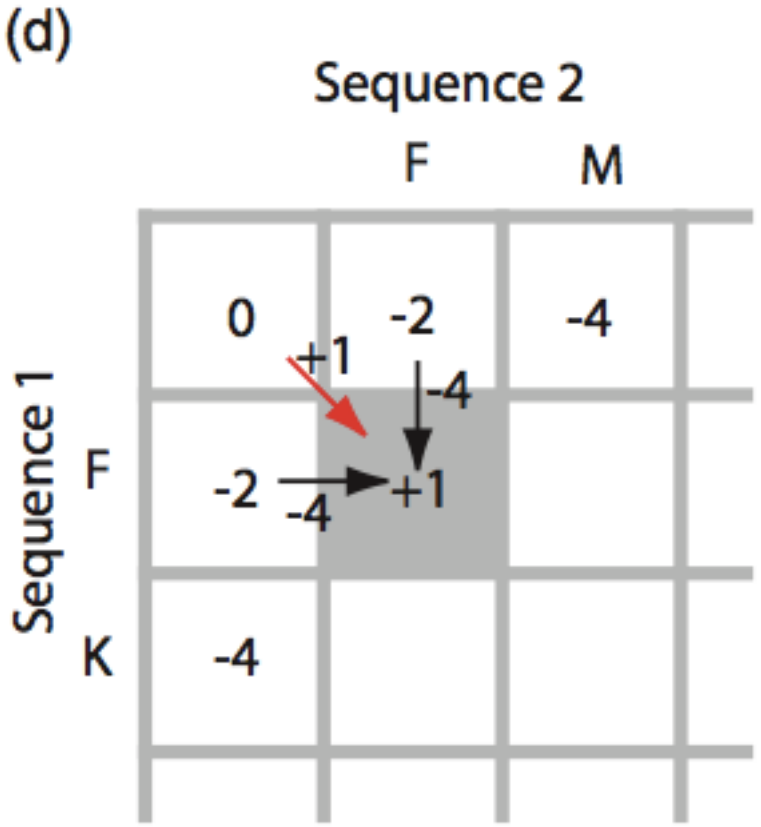
Define an overall score that maximizes cumulative scores at each position of the pairwise alignment, allowing for substitutions and gaps in either sequence.

Global pairwise alignment using Needleman-Wunsch



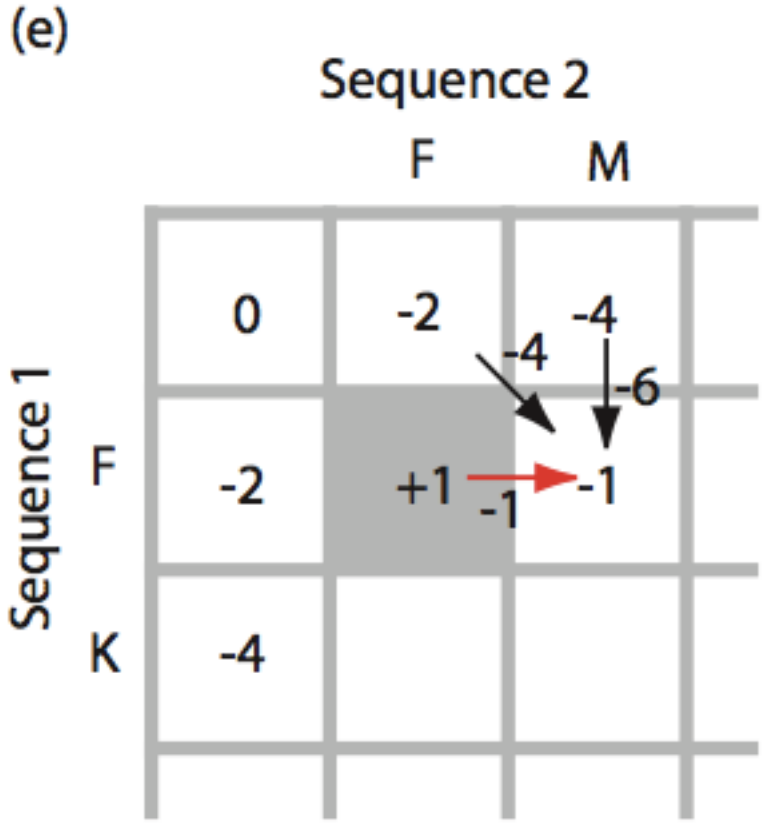
To decide how to align sequences 1 and 2 in the box at lower right, decide what the scores are beginning at upper left (not requiring a gap), or beginning from the left or top (each requiring a gap penalty).

Global pairwise alignment using Needleman-Wunsch



Here the best score involves +1 (proceed from upper left to gray, lower right square).
If we instead select an alignment involving a gap the score would be worse (-4).

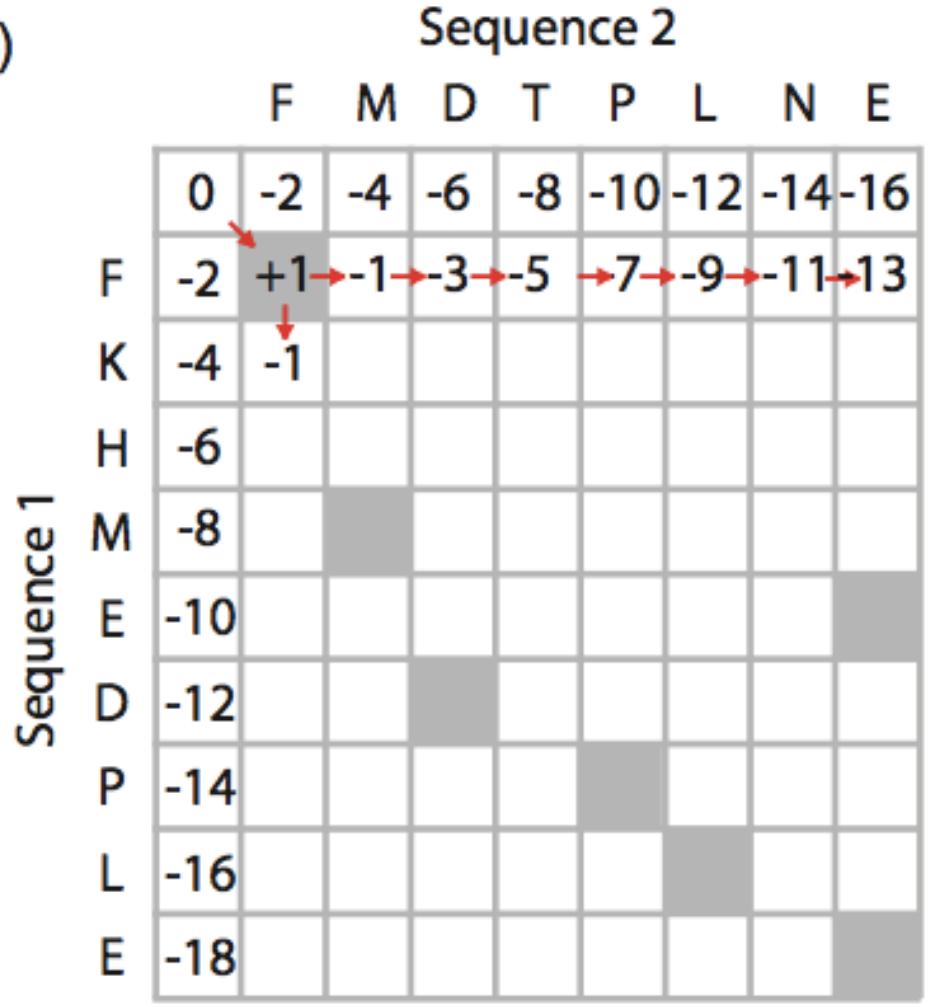
Global pairwise alignment using Needleman-Wunsch



Proceed to calculate the optimal score for the next position.

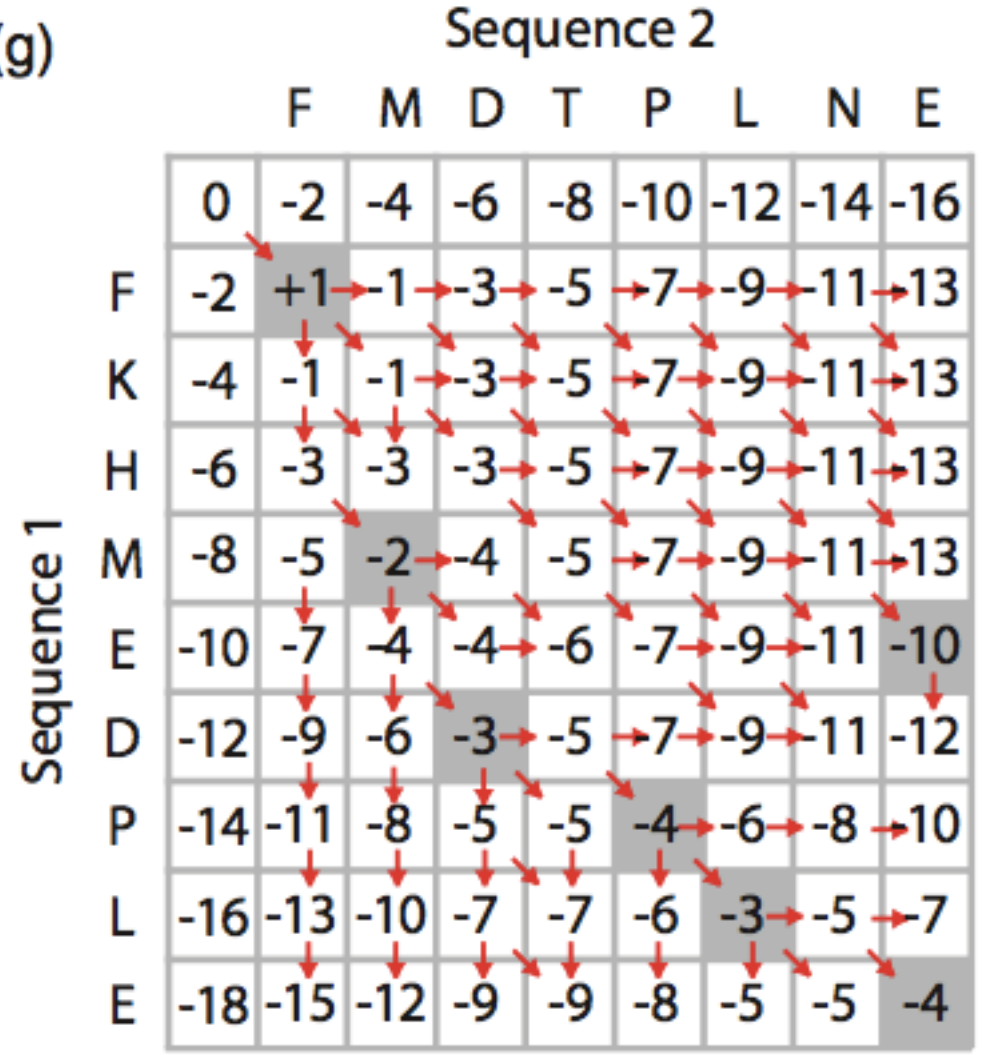
Global pairwise alignment using Needleman-Wunsch

(f)



Continue filling in the matrix.

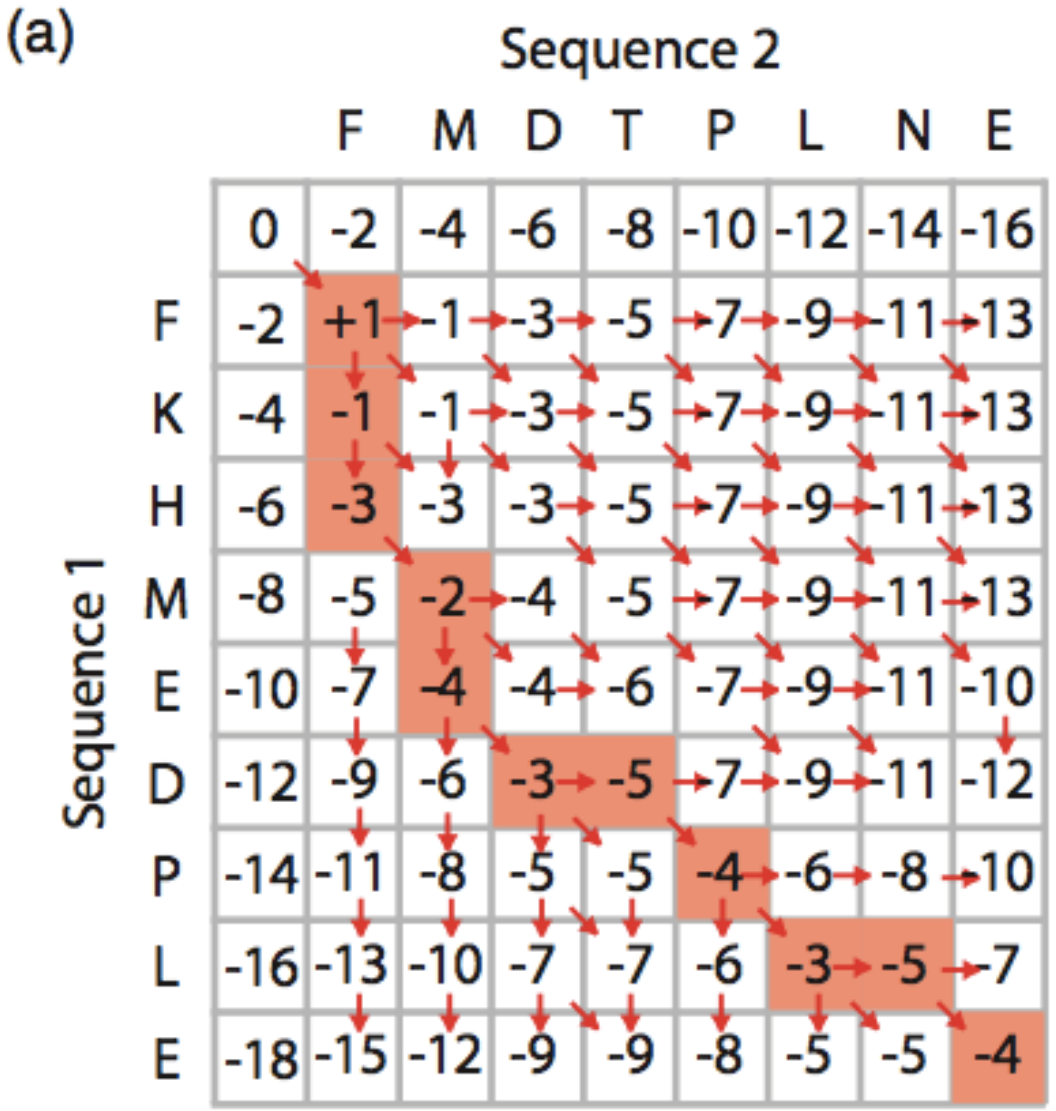
Global pairwise alignment using Needleman-Wunsch



Complete filling in the matrix.

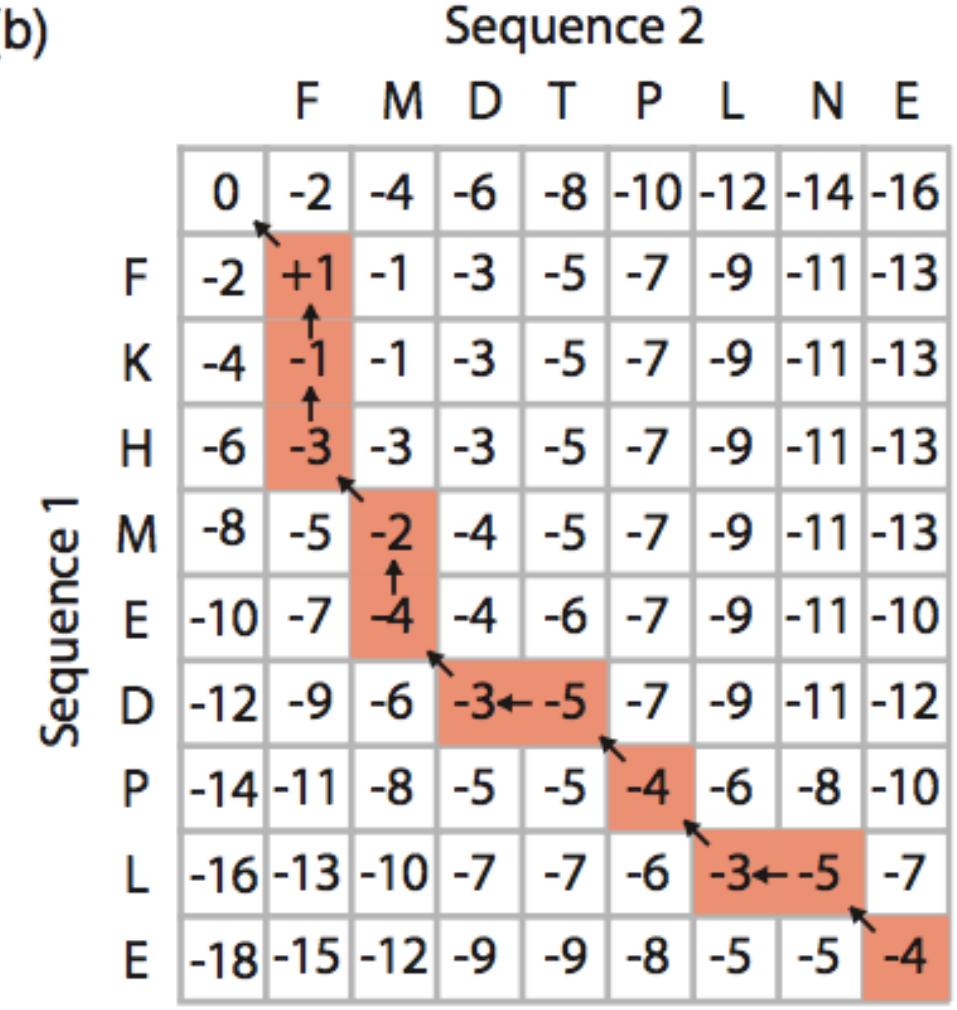
Global pairwise alignment using Needleman-Wunsch

Highlighted cells indicate the optimal path (best scores), indicating how the two sequences should be aligned.



Global pairwise alignment using Needleman-Wunsch

(b)



B&FG 3e
Fig. 3-22
Page 100

Equivalent representation, showing the traceback procedure: begin at the lower right cell and proceed back to the start.

Global pairwise alignment using Needleman-Wunsch

Equivalent representation, showing the traceback procedure: begin at the lower right cell and proceed back to the start.

(b)

Sequence 1 \ Sequence 2		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

	+1	-1	-3	-2	-4	-3	-5	-4	-3	-5	-4
Sequence 1	F	K	H	M	E	D	-	P	L	-	E
Sequence 2	F	-	-	M	-	D	T	P	L	N	E

Needleman-Wunsch: dynamic programming

N-W is guaranteed to find optimal alignments, although the algorithm does not search all possible alignments.

It is an example of a dynamic programming algorithm: an optimal path (alignment) is identified by incrementally extending optimal subpaths.
Thus, a series of decisions is made at each step of the alignment to find the pair of residues with the best score.

Global alignment versus local alignment

Global alignment (Needleman-Wunsch) extends from one end of each sequence to the other.

Local alignment finds optimally matching regions within two sequences ("subsequences").

Local alignment is almost always used for database searches such as **BLAST.** It is useful to find domains (or limited regions of homology) within sequences.

Smith and Waterman (1981) solved the problem of performing optimal local sequence alignment. Other methods (BLAST, FASTA) are faster but less thorough.

Global alignment (top) includes matches ignored by local alignment (bottom)

Global:
15% identity

Local:
30% identity

NP_824492, NP_337032

Outline

- Introduction
  - Protein alignment: often more informative than DNA alignment
  - Definitions: homology, similarity, identity
  - Gaps
  - Pairwise alignment, homology, and evolution of life
- Scoring matrices
  - Dayhoff model: 7 steps
  - Pairwise alignment and limits of detection: the “twilight zone”
- Alignment algorithms: global and local
  - Global sequence alignment: algorithm of Needleman and **Wunsch**
  - Local sequence alignment: Smith and Waterman algorithm
  - Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST
  - Basic Local Alignment Search Tool (BLAST)
  - Pairwise alignment with dotplots
- The statistical significance of pairwise alignments
  - Statistical significance of global alignments
  - Percent identity and relative entropy
- Perspective

How the Smith-Waterman algorithm works

(a)

Sequence 2		Sequence 1												
		C	A	G	C	C	U	C	G	C	U	U	A	G
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A		0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A		0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U		0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G		0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C		0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C		0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A		0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U		0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U		0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G		0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
A		0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C		0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G		0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G		0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

positions of nucleotide identity are shaded gray

(b)

```
sequence 1    GCC-UCG
sequence 2    GCCAUUG
```

(c)

```
sequence 1    CA-GCC-UCGCUUAG
sequence 2    AAUGCCAUUGACG-G
```

They scoring system here **is +1 for a match, minus one-third for a mismatch, and a gap penalty of the difference between a match and a mismatch (−1.3 for a gap of length one).** The matrix is scored based on finding the maximum of four possible non-negative values. The **highest value in the matrix (3.3)** corresponds to the beginning of the optimal local alignment, and the aligned residues (green font) extend up and to the left until a value of zero is reached. (b) The local alignment derived from this matrix is shown. Note that this alignment includes identities, a mismatch, and a gap. (c) A global alignment of the two sequences is shown for comparison to the local alignment. Note that it encompasses the entirety of both sequences.

Global alignment (top) includes matches ignored by local alignment (bottom)

Global:
15% identity

Local:
30% identity

NP_824492, NP_337032

Where to use the Smith-Waterman algorithm

[1] Galaxy offers “needle” and “water” EMBOSS programs.

[2] EBI offers needle and water. http://www.ebi.ac.uk/Tools/psa/

[3] Try using **SSEARCH** to perform a rigorous Smith-Waterman local alignment: http://fasta.bioch.virginia.edu/

[4] Next-generation sequence aligners incorporate Smith-Waterman in some specialized steps.

Rapid, heuristic versions of Smith-Waterman: FASTA and BLAST

- Smith-Waterman is very rigorous, and it is guaranteed to find an optimal alignment.

- But Smith-Waterman is slow. It requires computer space and time proportional to the product of the two sequences being aligned (or the product of a query against an entire database).

- Gotoh (1982) and Myers and Miller (1988) improved the algorithms so both global and local alignment require less time and space.

- **FASTA and BLAST provide rapid alternatives to S-W**.

Outline

Introduction
- Protein alignment: often more informative than DNA alignment
- Definitions: homology, similarity, identity
- Gaps
- Pairwise alignment, homology, and evolution of life

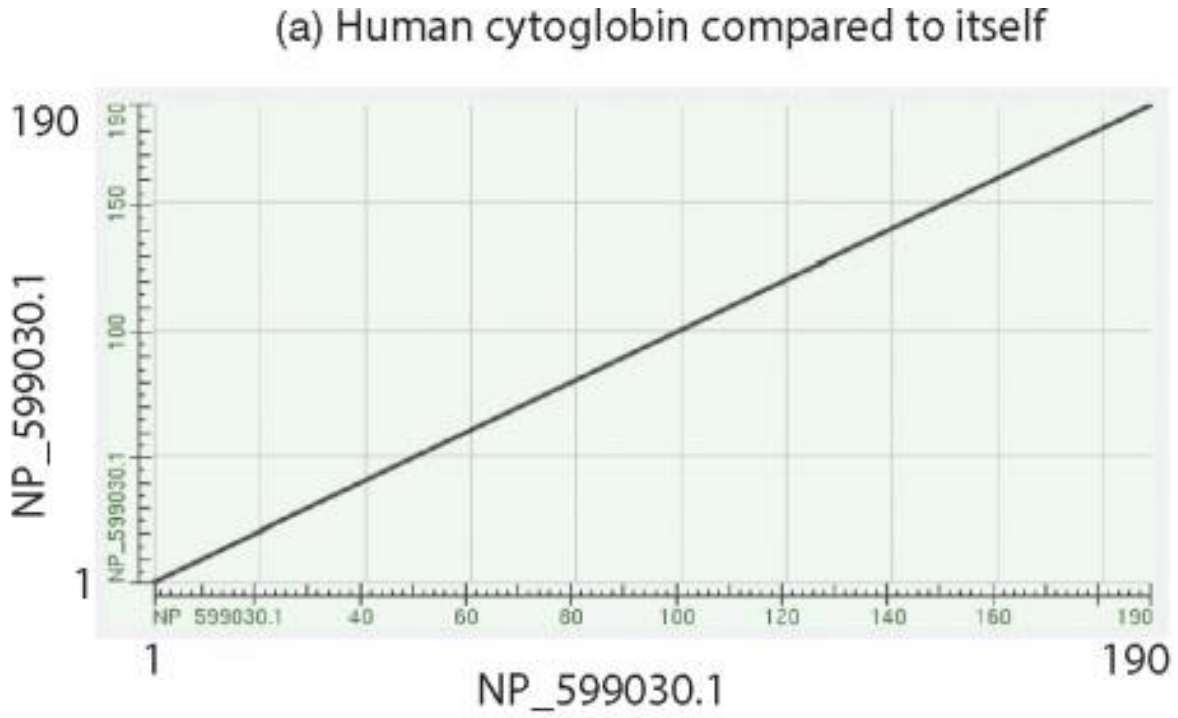
Scoring matrices
- Dayhoff model: 7 steps
- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local
- Global sequence alignment: algorithm of Needleman and Wunsch
- Local sequence alignment: Smith and Waterman algorithm
- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST
- **Basic Local Alignment Search Tool (BLAST)**
- Pairwise alignment with dotplots

The statistical significance of pairwise alignments
- Statistical significance of global alignments
- Percent identity and relative entropy

Perspective

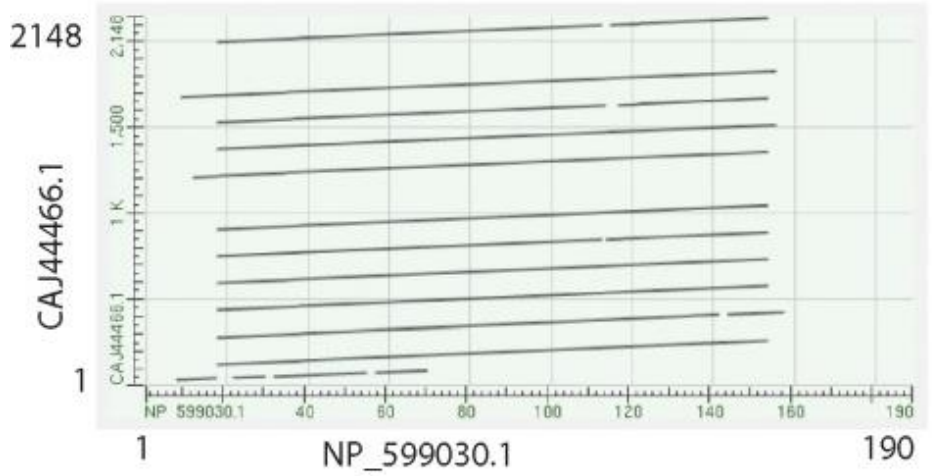
Pairwise alignment with dotplots



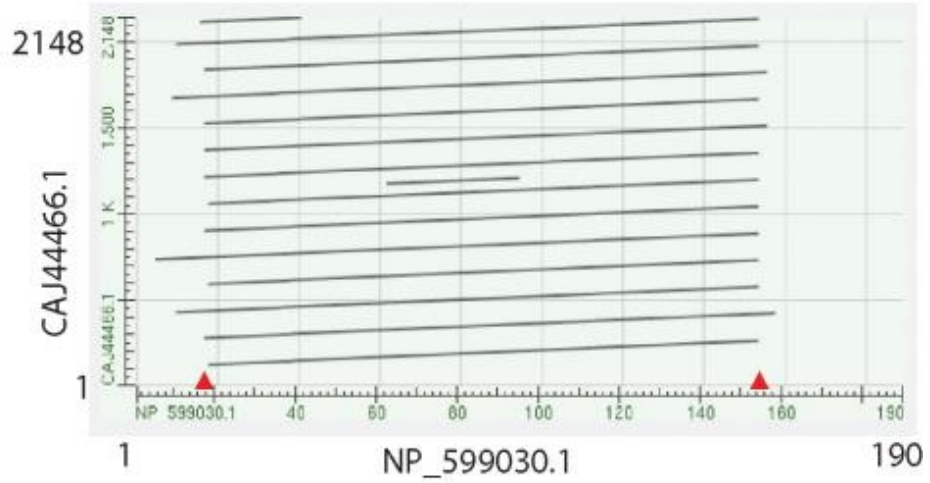
A human globin searched against itself produces a unit diagonal on a dot plot (NCBI BLASTP, aligning 2 sequences).

Pairwise alignment with dotplots

(b) Cytoglobin compared to a snail globin (BLOSUM62)



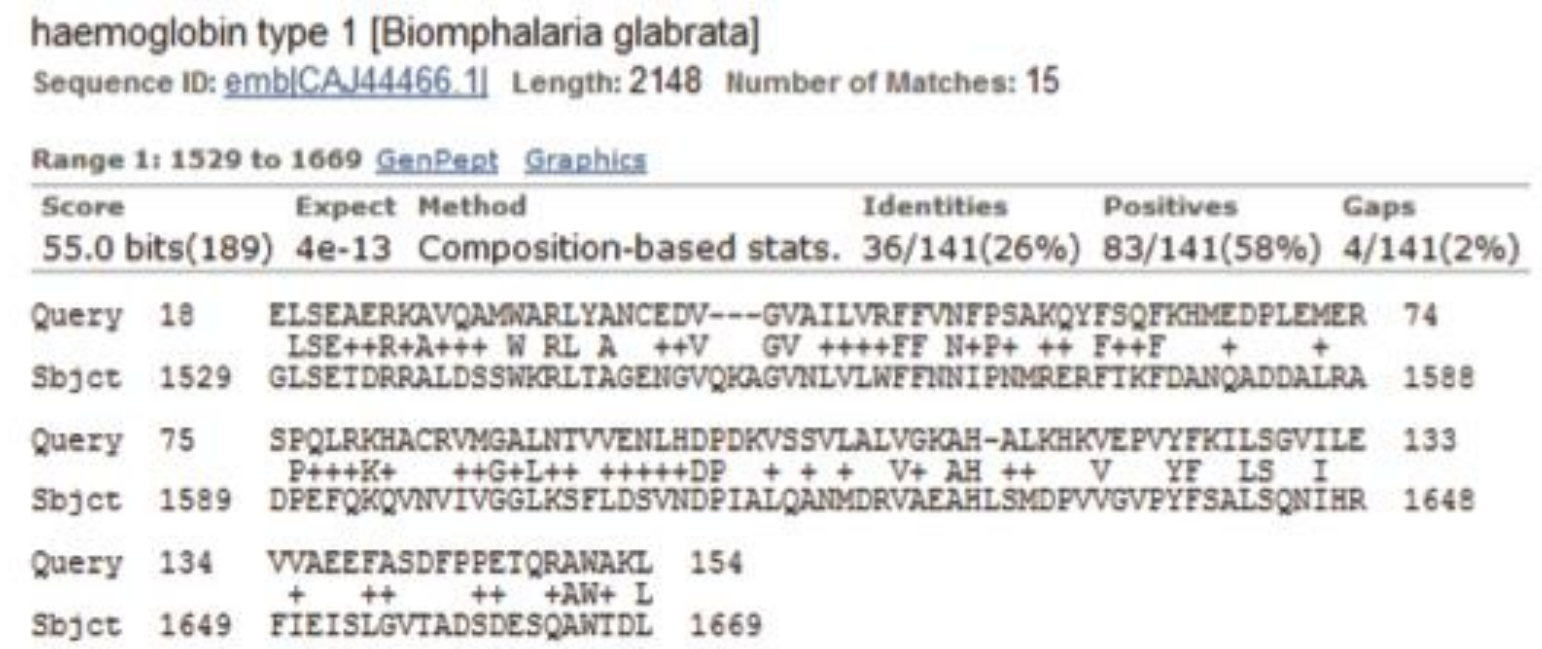
(c) Cytoglobin compared to a snail globin (PAM250)



Search human cytoglobin against a large snail globin (having many globin repeats). More repeats are observed using PAM250 than BLOSUM62.

To “read” this plot note that cytoglobin (x-axis) matches the snail globin (y-axis) at about a dozen locations across the snail protein. Red arrows indicate that the first few and last few amino acids of cytoglobin do not participate in this repeat structure.

Pairwise alignment with dotplots



haemoglobin type 1 [Biomphalaria glabrata]

Sequence ID: emb|CAJ44466.1| Length: 2148 Number of Matches: 15

Range 1: 1529 to 1669 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
55.0 bits(189)	4e-13	Composition-based stats.	36/141(26%)	83/141(58%)	4/141(2%)

```
Query  18    ELSEAERKAVQAMWARLYANCEDV---GVAILVRFFVNFPSAKQYFSQFKHMEDPLEMER  74
             LSE++R+A+++ W RL A  ++V   GV ++++FF N+P+ ++ F++F   +    +
Sbjct  1529  GLSETDRRALDSSWKRLIAGENGVQKAGVNLVLWFFNNIPNMRERFTKFDANQADDALRA  1588

Query  75    SPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAH-ALKHKVEPVYFKILSGVILE  133
             P+++K+   ++G+L++ +++++DP  + + +  V+ AH ++   V   YF  LS  I
Sbjct  1589  DPEFQKQVNVIVGGLKSFLDSVNDPIALQANMDRVAEAHLSMDPVVGVPYFSALSQNIHR  1648

Query  134   VVAEEFASDFPPETQRANAKL  154
              +   ++    ++  +AW+ L
Sbjct  1649  FIEISLGVIADSDESQAWIDL  1669
```

BLASTP output includes the various sequence alignments. One is shown here: human cytoglobin (residues 18-154) aligns to the snail globin (at residues 1529-1669). The expect value is convincing (4e-13), and this is one of a dozen sequence alignments.

Conclusion: the dotplot is an excellent way to visualize complex repeats.

Outline

Introduction
- Protein alignment: often more informative than DNA alignment
- Definitions: homology, similarity, identity
- Gaps
- Pairwise alignment, homology, and evolution of life

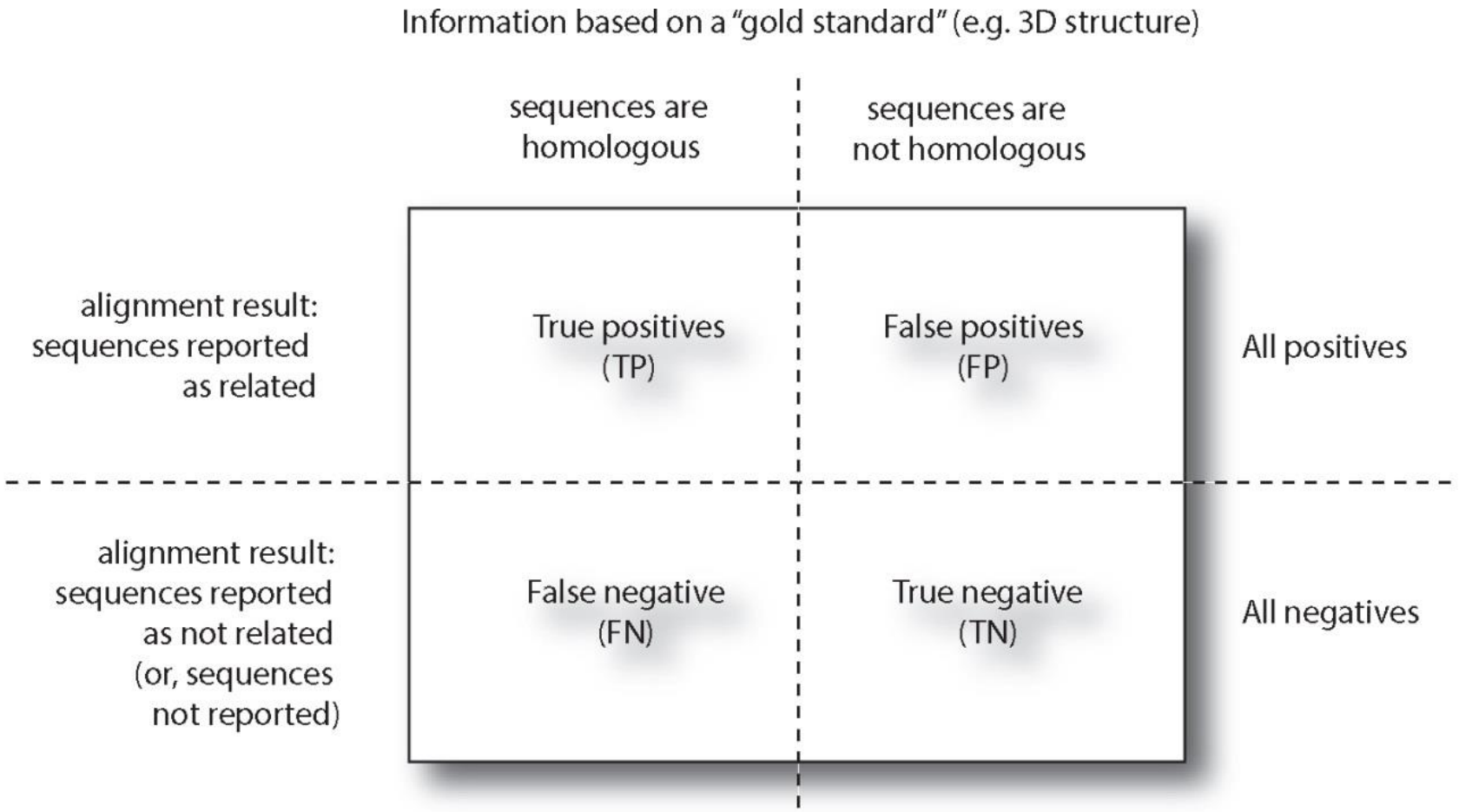
Scoring matrices
- Dayhoff model: 7 steps
- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local
- Global sequence alignment: algorithm of Needleman and Wunsch
- Local sequence alignment: Smith and Waterman algorithm
- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST
- Basic Local Alignment Search Tool (BLAST)
- **Pairwise alignment with dotplots**

The statistical significance of pairwise alignments
- Statistical significance of global alignments
- Percent identity and relative entropy

Perspective

Statistical significance of pairwise alignments



Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)

Statistical significance of pairwise alignments

The statistical significance of global alignments is not well described.
We can apply a z-score.

$$Z = \frac{x - \mu}{s}$$

For local alignment the statistical significance is thoroughly understood.
See Chapter 4 (BLAST).

http://pevsnerlab.kennedykrieger.org/wiley/chapter3/Webdocument_3-7_prss_shuffle.htm

Outline

Introduction
- Protein alignment: often more informative than DNA alignment
- Definitions: homology, similarity, identity
- Gaps
- Pairwise alignment, homology, and evolution of life

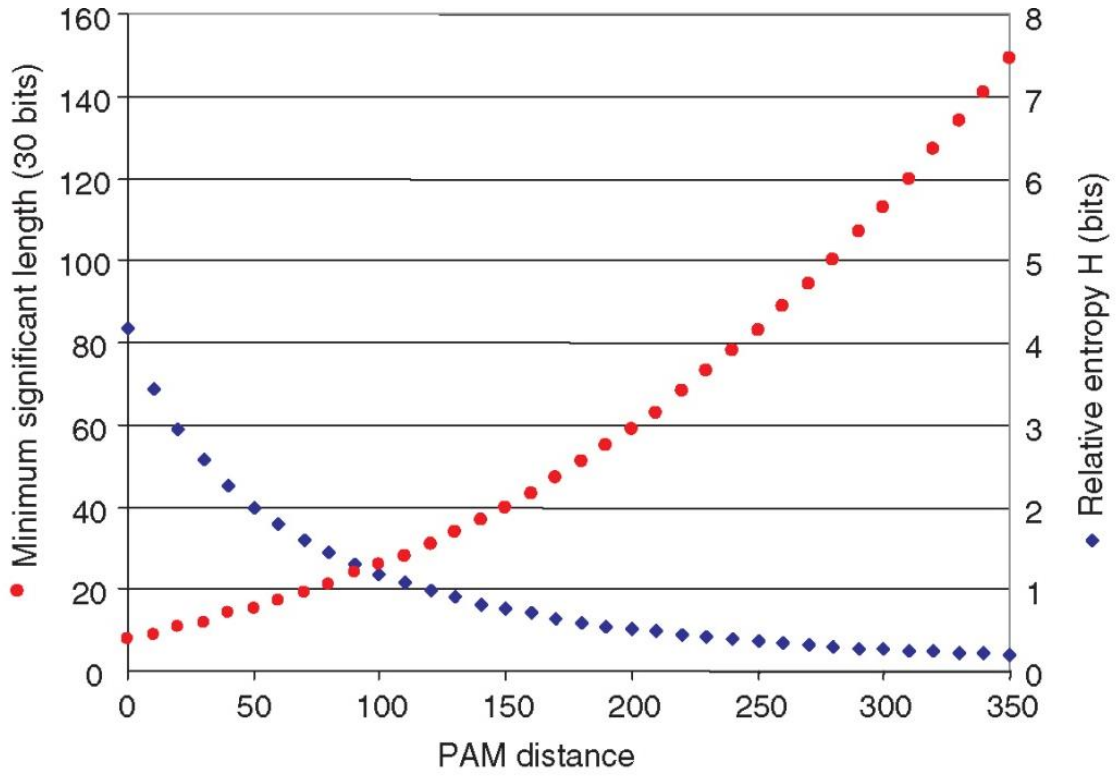
Scoring matrices
- Dayhoff model: 7 steps
- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local
- Global sequence alignment: algorithm of Needleman and Wunsch
- Local sequence alignment: Smith and Waterman algorithm
- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST
- Basic Local Alignment Search Tool (BLAST)
- Pairwise alignment with dotplots

The statistical significance of pairwise alignments
- **Statistical significance of global alignments**
- Percent identity and relative entropy

Perspective

Relative entropy (*H*) as a function of PAM distance



For **PAM** matrices with low values (e.g. **PAM10**) the relative entropy (in bits) is high, and the minimum alignment length needed to detect a significant pairwise alignment is short. Relative entropy relates to the information content.

For **PAM250** and similar matrices the relative entropy is low. It is necessary to have a longer region of amino acids aligned (e.g. 80 residues) to detect significant pairwise relatedness.

Perspective: Pairwise alignment is a fundamental problem in bioinformatics.