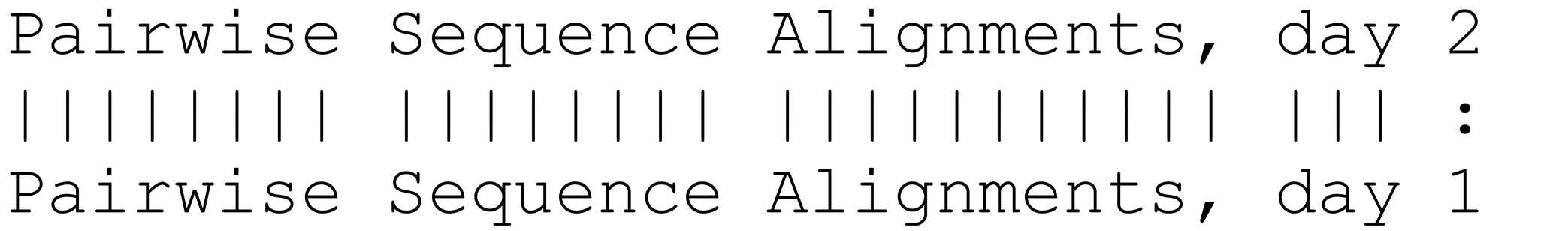
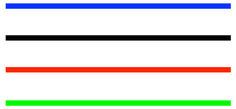


Mark Stenglein, MIP 280A4

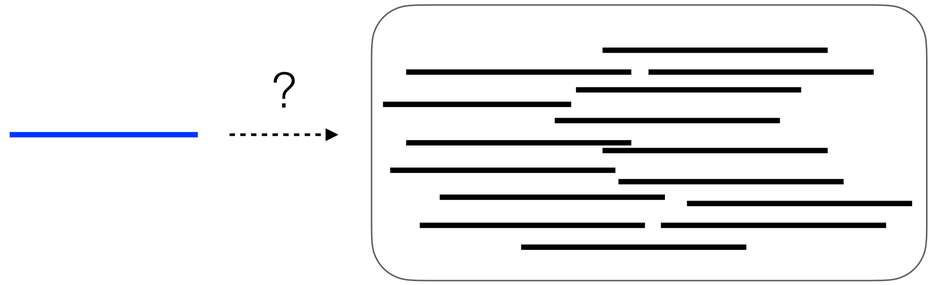
Today we'll continue to learn about pairwise alignments

Pairwise alignment

Multiple sequence alignment

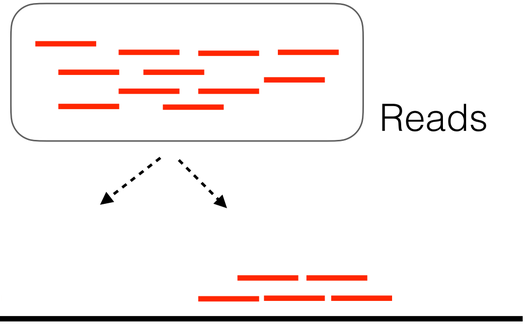


Alignment-based database searches



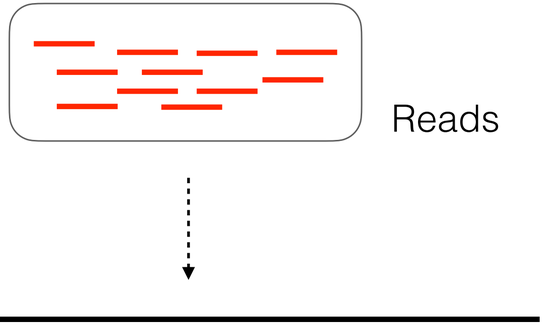
Database

Alignment to reference (mapping)



Existing Reference sequence

Assembly



New Reference sequence

The “optimal” pairwise alignment is a function of:

- The algorithm used
- Whether it’s a **global** or **local** alignment
- The **scoring system** used
- How **gaps** are handled

"Brute force" solutions to sequence alignment, where you create and score all possible alignments, are effectively impossible for longer sequences.

The number of possible global alignments for 2 sequences of length N is: $\frac{2^{2N}}{\sqrt{\pi N}}$

So, for 2 sequences of length 150, there are $\frac{2^{300}}{\sqrt{\pi 150}}$ ~ **10^{88}** possible pairwise alignments!

(There are an estimated 10^{78} atoms in the universe)

Credit for these formulas/numbers: Vladimir Likic

Brute force approaches aren't practical,
so clever algorithms are necessary

Needleman-Wunsch is the name of the standard global alignment algorithm

Muhammad ibn Musa
al-Khwarizmi

Definition of *algorithm*

: a procedure for solving a mathematical problem (as of finding the greatest common divisor) in a finite number of steps that frequently involves repetition of an operation

broadly : a step-by-step procedure for solving a problem or accomplishing some end

The Needleman-Wunsch algorithm:

- Is guaranteed to find the highest scoring global alignment between two sequences
- Tells you the score of the highest scoring alignment
- Works for different scoring systems
- Works for protein or nucleotide sequences
- For two sequences of length N, has a run time proportional to NxN (can be slow but much much better than brute force)
- Is something you can program a computer to do.
- Solves the same easy problem repeatedly instead of trying to solve one difficult problem once (“dynamic programming”)

J. Mol. Biol. (1970) **48**, 443–453

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH

Department of Biochemistry, Northwestern University, and Nuclear Medicine Service, V. A. Research Hospital Chicago, Ill. 60611, U.S.A.

(*Received 21 July 1969*)

A computer adaptable method for finding similarities in the amino acid sequences of two proteins has been developed. From these findings it is possible to determine whether significant homology exists between the proteins. This information is used to trace their possible evolutionary development.

Needleman-Wunsch exercise!

Online N-W demo

https://bioboot.github.io/bimm143_W20/class-material/nw/

Global Alignment App

Here we present an interactive example of the Needleman-Wunsch global alignment algorithm from BIMM-143 **Class 2**. The purpose of this app is to vis
based on user defined **Match**, **Mismatch** and **Gap Scores**. Note that for improving your understanding of this algorithm there is no substitute for workin

Experiment by changing the various **Scores**, altering the two **Sequences** and noting how the alignment matrix values, **trace back alignment path** (ir

▸ **Details**:

Sequence 1 ACG

Sequence 2 ACTG

Match Score	Mismatch Score	Gap Score
1	-1	-1

Compute Optimal Alignment | Clear Path | Custom Path

A	C	T	G
A	C	-	G

Score = 2

		A	C	T	G
	0	-1	-2	-3	-4
A	-1	1	0	-1	-2
C	-2	0	2	1	0
G	-3	-1	1	1	2

▸ **Reference**:

The Smith-Waterman algorithm is the canonical local alignment algorithm

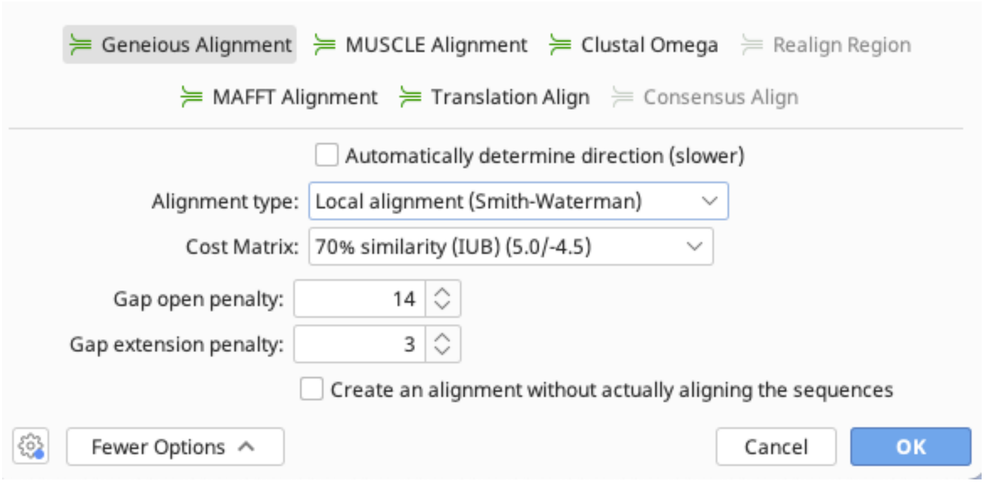
S-W has similar properties to N-W, including a guarantee to find the highest scoring (in this case possibly local) alignment

J. Mol. Biol. (1981), **147**, 195–197

Identification of Common Molecular Subsequences

T. F. Smith and M. S. Waterman

The identification of maximally homologous subsequences among sets of long sequences is an important problem in molecular sequence analysis. The problem is straightforward only if one restricts consideration to contiguous subsequences (segments) containing no internal deletions or insertions. The more general problem has its solution in an extension of sequence metrics (Sellers 1974; Waterman *et al.*, 1976) developed to measure the minimum number of "events" required to convert one sequence into another.

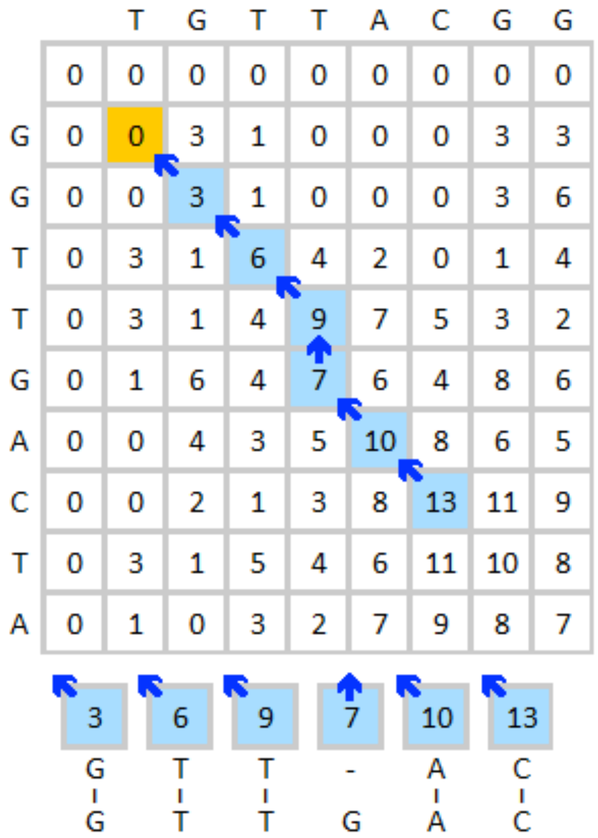


The Smith-Waterman algorithm is a variant of the N-W algorithm

Rule changes:

1) Set negative cell values to 0

2) Start traceback at the highest score in the matrix, and stop when you get to a 0-valued cell.



The first base and last two bases of each sequence mismatch, and are not included in the alignment

Image credit: Yz cs5160 CC-BY-SA-4.0, link

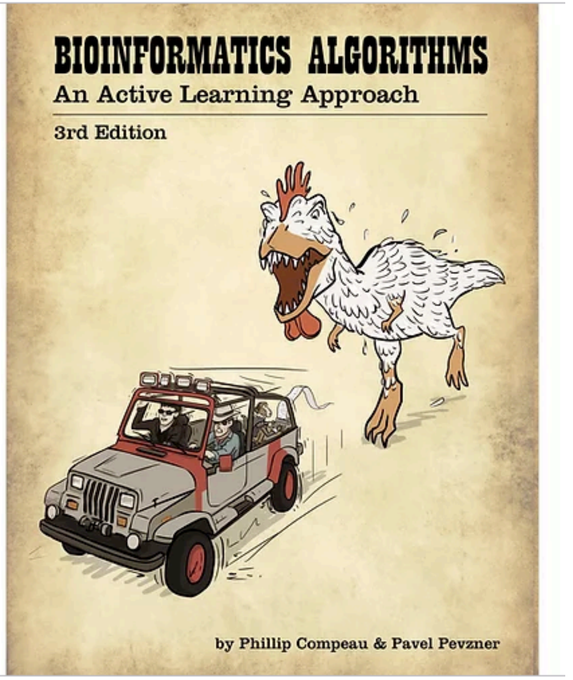
Major categories of sequence alignment

Alignment type	Purpose	Commonly-used software
Pairwise alignment	Identify the similarities or differences between two sequences	Needle (global alignment) Water (local alignment)
Multiple sequence alignment	Identify the similarities or differences between >2 sequences. Input to tree building.	MAFFT
Alignment-based search	Find the most closely related sequence in a database of sequences	BLAST
Mapping (alignment to reference)	Determine the most likely location in a reference sequence from which a shorter sequence (a read) derives	BWA Bowtie2
Assembly	Create a new reference sequence using overlapping reads	SPAdes

This course is not going to go in depth into algorithms

CS 425 Introduction to Bioinformatics Algorithms Credits: 4 (3-2-0)
Course Description: Algorithms for analysis of large scale biological data.
Prerequisite: (BZ 360 with a minimum grade of C or CS 320 with a minimum grade of C) and (CS 345 with a minimum grade of C).
Registration Information: Must register for lecture and laboratory.
Term Offered: Fall.
Grade Mode: Traditional.
Special Course Fee: No.

https://www.bioinformaticsalgorithms.org



There are more complicated nucleotide alignment scoring systems That give different rewards or costs for all possible substitutions

A simple substitution matrix

	A	C	G	T
A	1			
C	-1	1		
G	-1	-1	1	
T	-1	-1	-1	1

Equivalent to
Match: +1
Mismatch: -1

This substitution matrix penalizes transitions less than transversions

	A	C	G	T
A	1			
C	-1	1		
G	-0.5	-1	1	
T	-1	-0.5	-1	1

Only showing bottom-left half of matrices because they are symmetric across diagonal

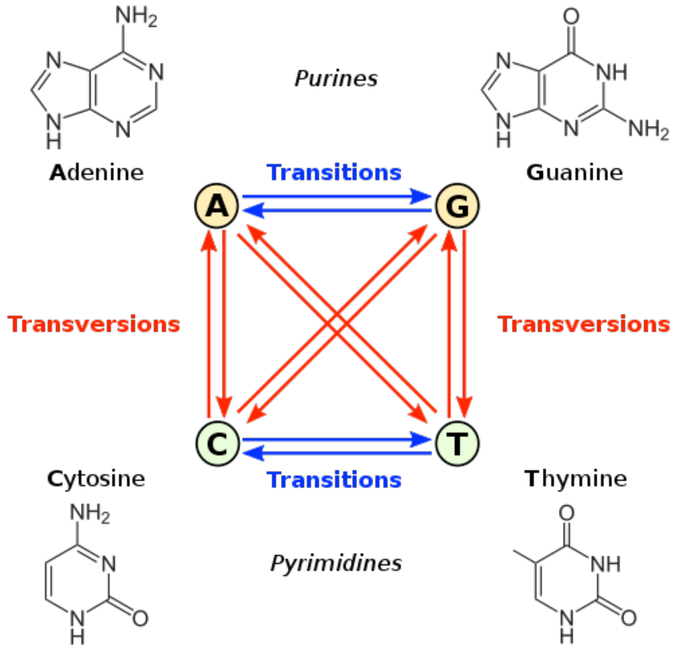
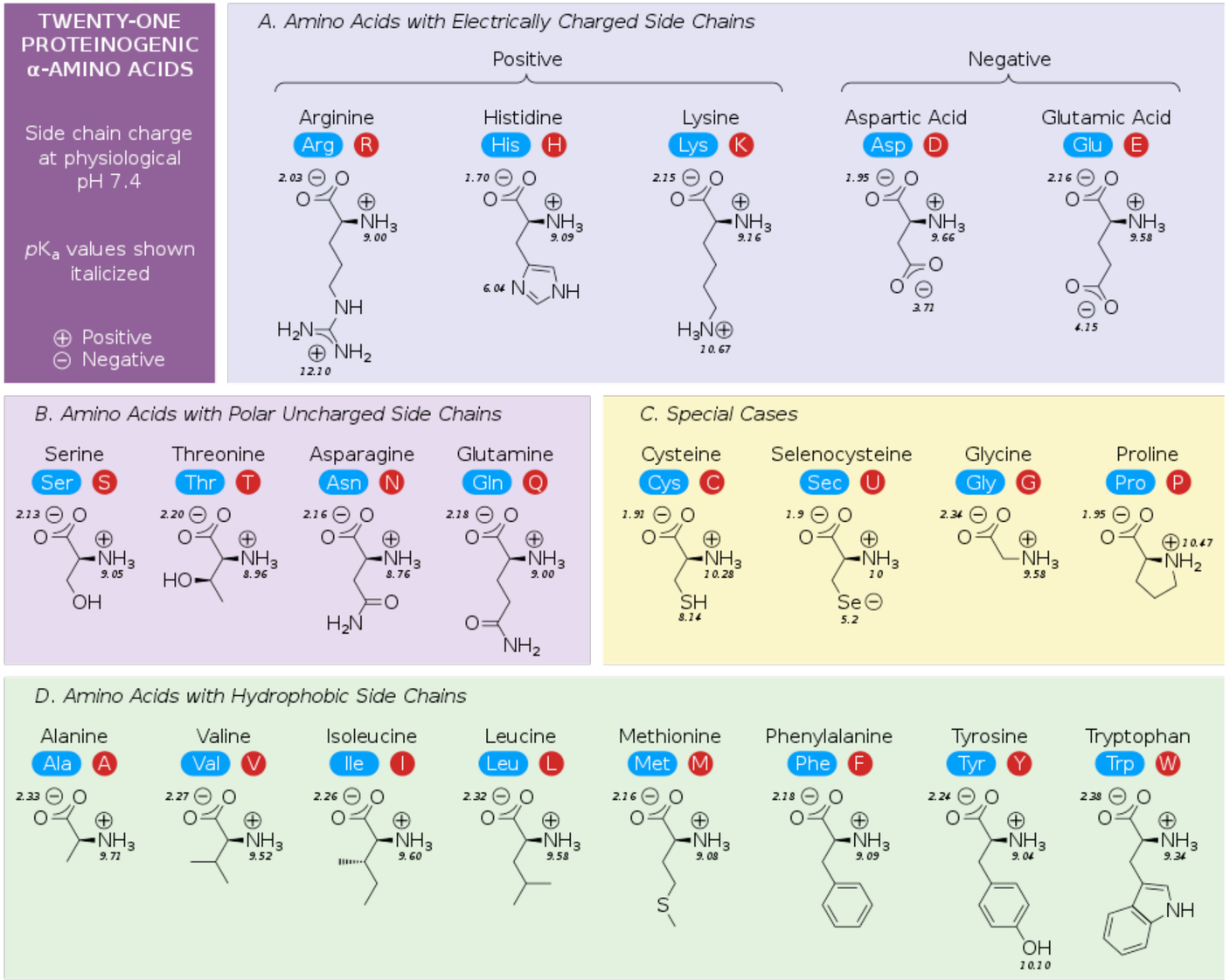


Image credit: Manudouz/Petulda, CC-BY-SA-4.0, link

Groups of amino acids have similar chemical properties



Question: how could this information be incorporated into protein alignment scoring systems?

Protein-protein alignments are scored using matrices, like BLOSUM, that are calculated using observed substitution patterns from real alignments

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

The BLOSUM62 matrix is based on many alignments of proteins with ~62% pairwise identity

This 20x20 matrix assigns a penalty - or a reward - to every possible amino acid mismatch

You could make scoring matrices from words with shared ancestry

```
   Pater
   Father
   Vader
   Padre
   Père
   Paternal
Jupiter
```

P, F, and V often substitute for each other in Indo-European languages because they are similar sounds

Similarly, T and D often substitute for each other in Indo-European words

George Lucas was foreshadowing key information with the name Darth Vader

Mismatches involving proline are always penalized

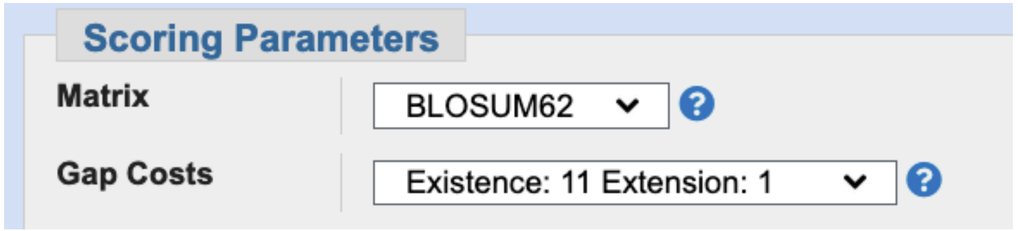
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

Mismatches involving large hydrophobic residues are actually rewarded and not penalized

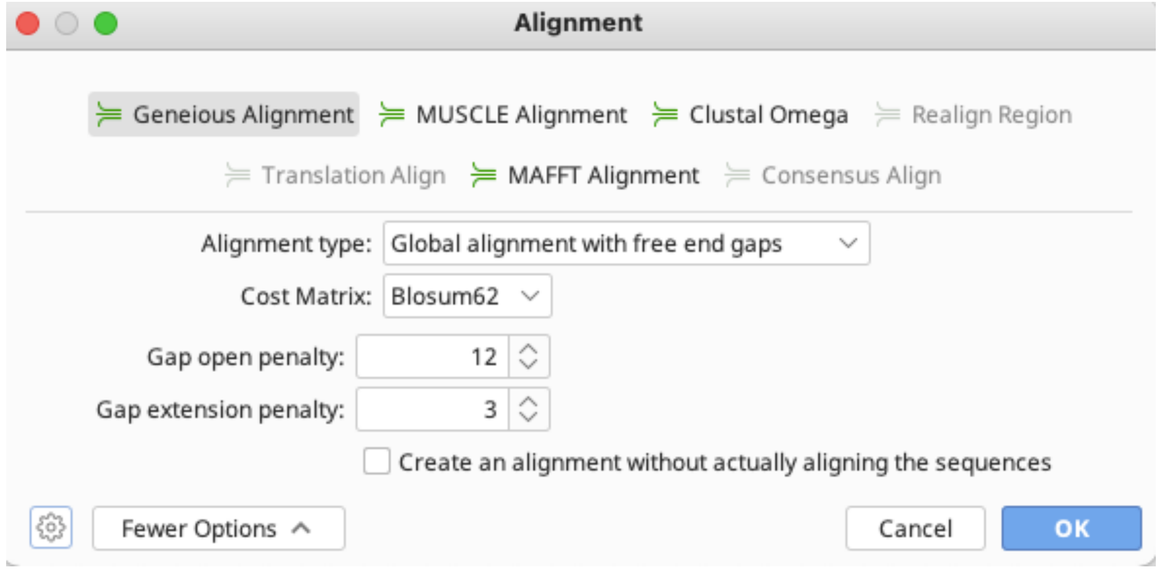
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

BLOSUM62 scoring matrix is frequently the default scoring scheme

A protein BLAST search (BLASTP)



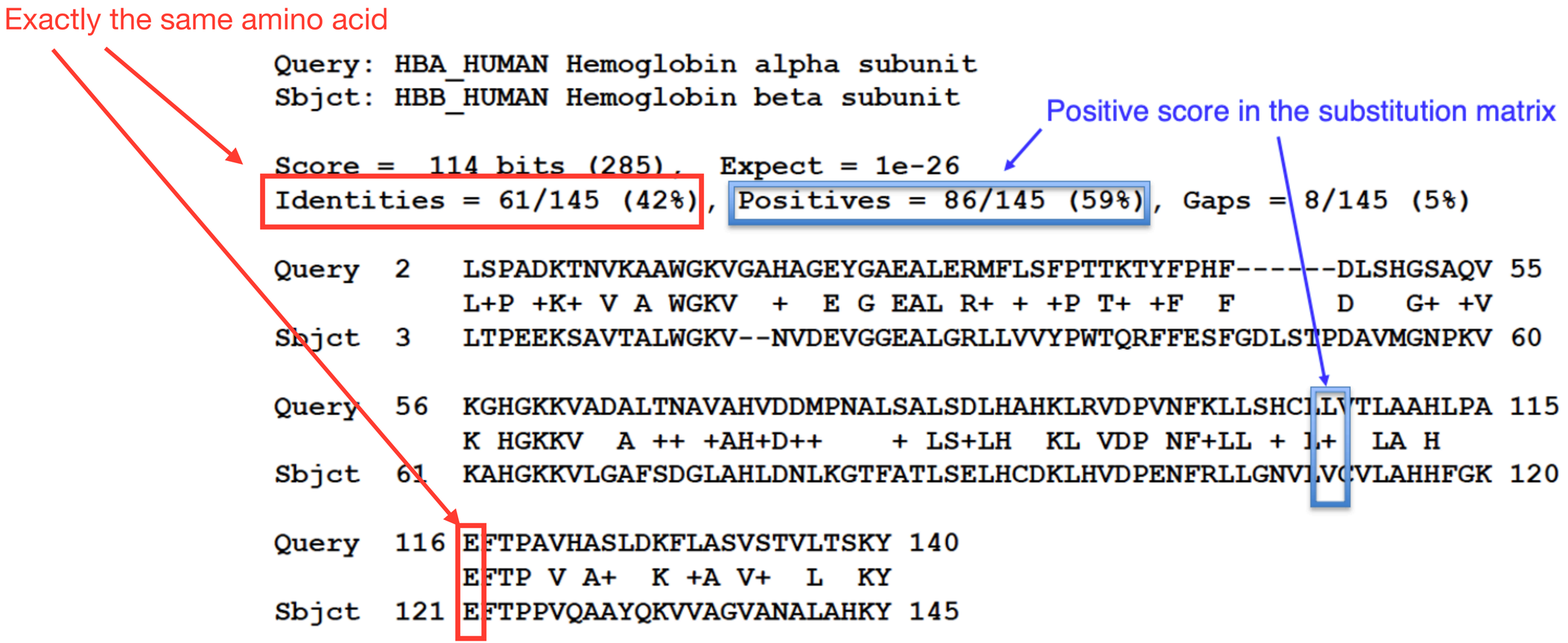
In Geneious pairwise aligner



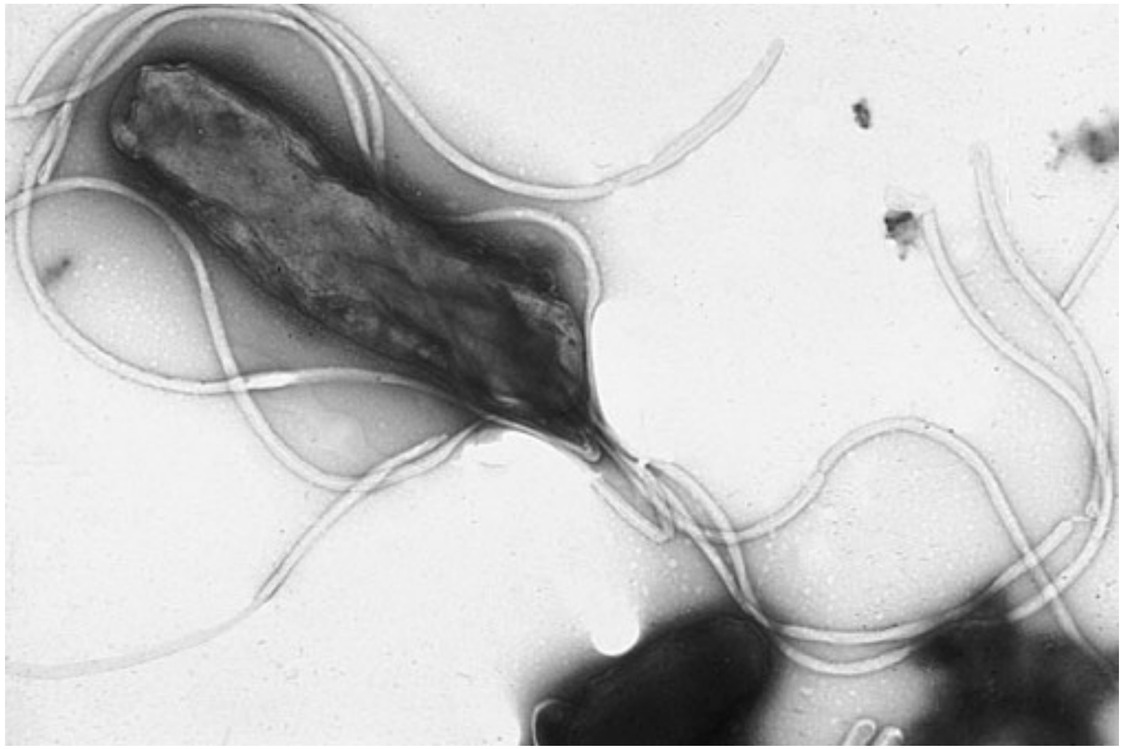
BLOSUM exercise!

The output of protein BLAST alignments provides information about mismatches that are actually rewarded not penalized. These are “similar” amino acids.

Alignment “similarity” vs. “identity”



Protein alignment exercise!



Electron micrograph of *H. pylori* possessing multiple flagella, image: Yutaka Tsutsumi (link)