Alignment ▶ Database Searching ▶ HT Mapping

university of groningen

umcg

Sequence Alignments: Intro

- Why?
  - Similarity searching
  - Infer knowledge
    - Similar structures often have similar functions

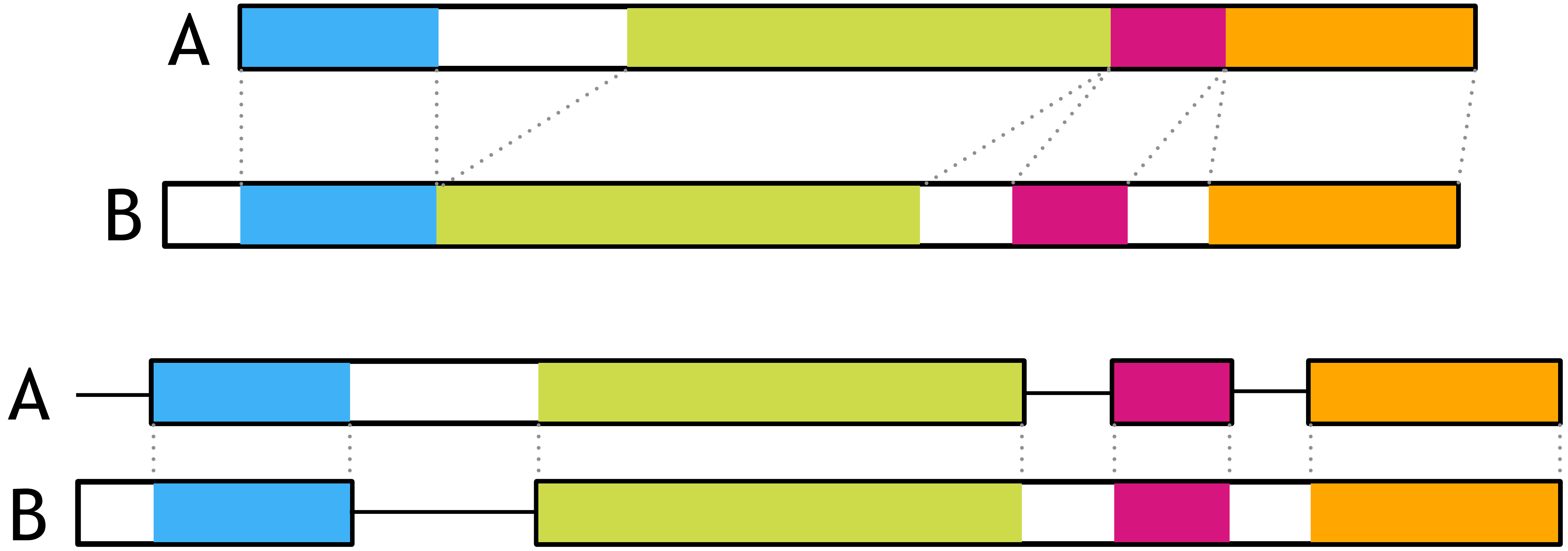
Sequence Alignments: Intro

- Burkhart Rost:
  - Identity > 30 % ➙ 90 % similar structures
  - Identity < 25 % ➙ 10 % similar structures

% Identity	Homology ?
> 30	Presume Homology
20 - 30	Twilight zone
< 20	Midnight zone

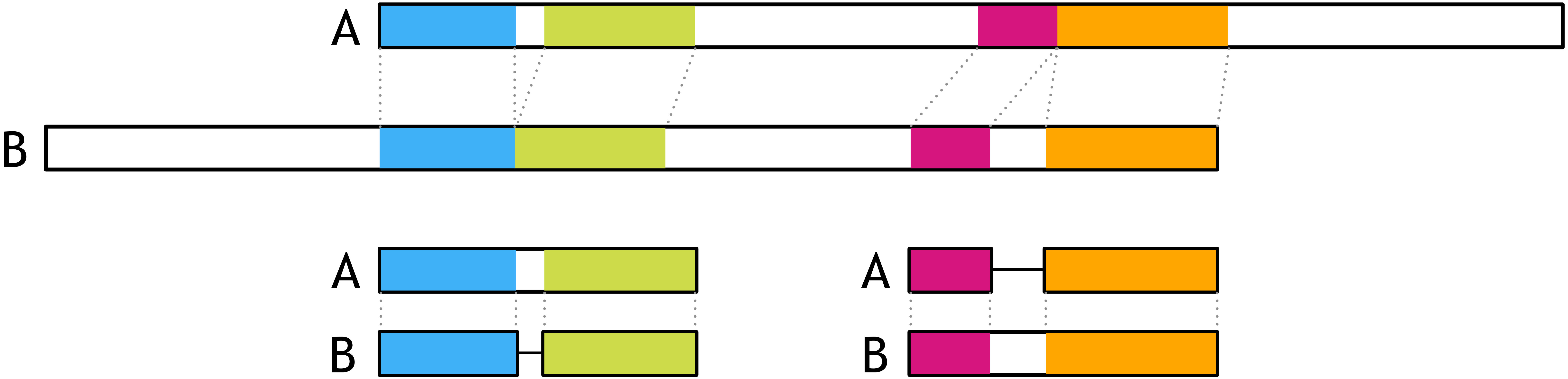
Sequence Alignments: Intro

- Global
  - Try to align entire sequences
  - Most useful for highly similar sequences



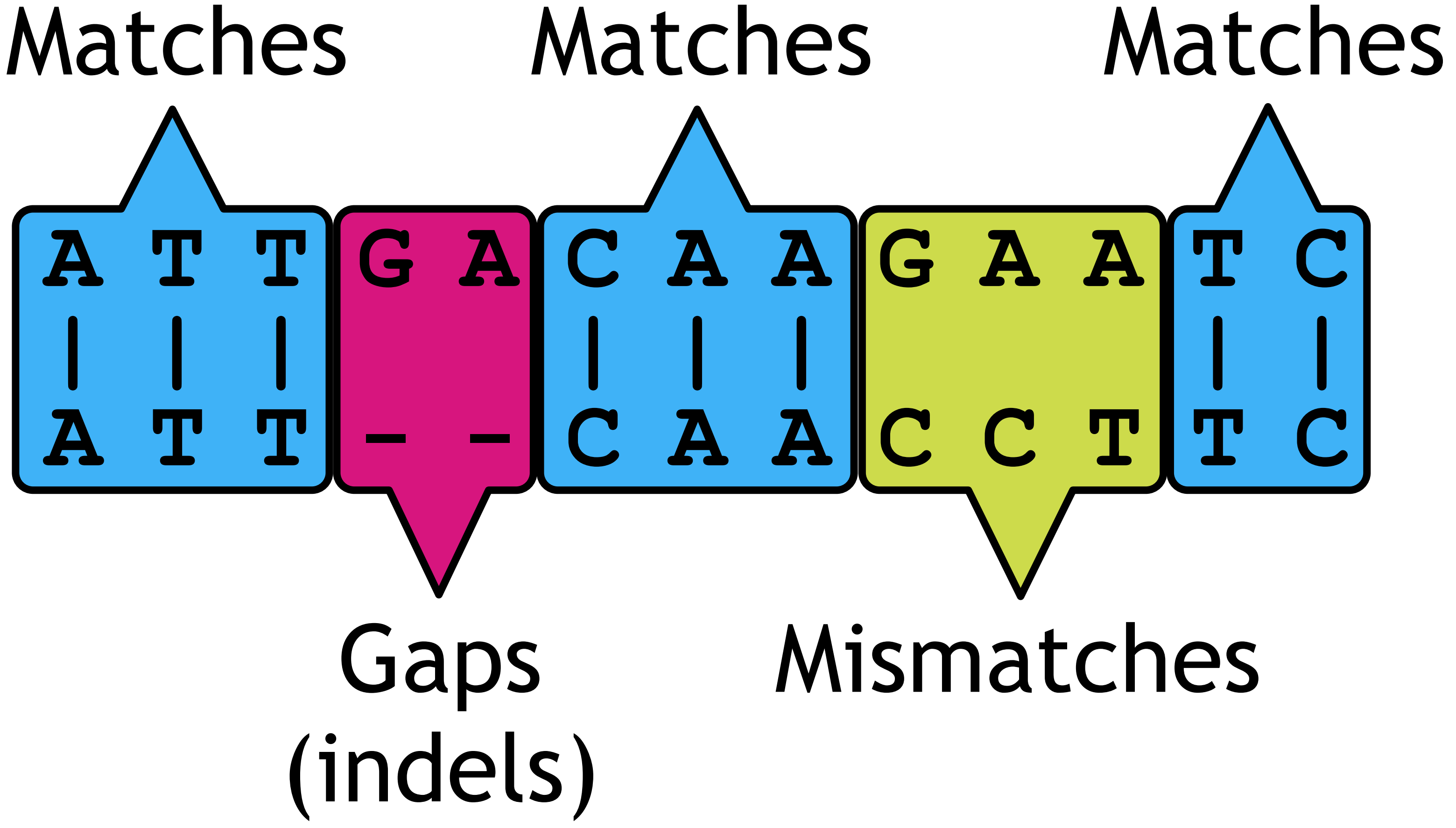
Sequence Alignments: Intro

- Local
  - Try to create optimal alignments for regions
  - Most useful when only parts of the sequences are conserved



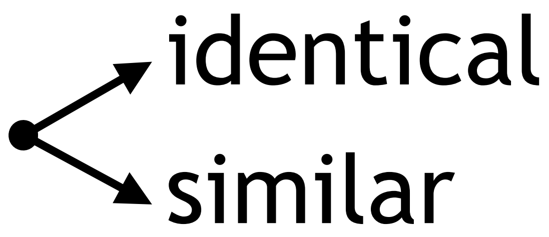
Sequence Alignments: Scoring

- How-to?
  - Scoring
    - Find best alignment by finding alignment with highest Score



Sequence Alignments: Scoring

- Scoring
  - Gap penalties
    - Open gap
    - Extend gap
  - Substitution matrices
    - mismatches + matches



identical

similar

Sequence Alignments: Substitution Matrices

- PAM
  - **P**oint **A**ccepted **M**utation matrix
  - PAM++ ➙ Mutations++ ➙ Evolutionairy distance++

- BLOSUM
  - **BLO**cks **SU**bstitution **M**atrix
  - BLOSUM++ ➙ Conservation++ ➙ Evolutionairy distance--

Sequence Alignments: Substitution Matrices

Protein PAM 250

Cys	C	**12**																			
Gly	G	-3	**5**																		
Pro	P	-3	-1	**6**																	
Ser	S	0	1	1	**2**																
Ala	A	-2	1	1	1	**2**															
Thr	T	-2	0	0	1	1	**3**														
Asp	D	-5	1	-1	0	0	0	**4**													
Glu	E	-5	0	-1	0	0	0	3	**4**												
Asn	N	-4	0	-1	1	0	0	2	1	**2**											
Gln	Q	-5	-1	0	-1	0	-1	2	2	1	**4**										
His	H	-3	-2	0	-1	-1	-1	1	1	2	3	**6**									
Lys	K	-5	-2	-1	0	-1	0	0	0	1	1	0	**5**								
Arg	R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	**6**							
Val	V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	**4**						
Met	M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	**6**					
Ile	I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	**5**				
Leu	L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	**6**			
Phe	F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	**9**		
Tyr	Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	**10**	
Trp	W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	**17**
		C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W

Sequence Alignments: PAM 250

Cys	C	12																			
Gly	G	-3	5																		
Pro	P	-3	-1	6																	
Ser	S	0	1	1	2																
Ala	A	-2	1	1	1	2															
Thr	T	-2	0	0	1	1	3														
Asp	D	-5	1	-1	0	0	0	4													
Glu	E	-5	0	-1	0	0	0	3	4												
Asn	N	-4	0	-1	1	0	0	2	1	2											
Gln	Q	-5	-1	0	-1	0	-1	2	2	1	4										
His	H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
		C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W

Other hydrophilic

Acid / Acid-amide

Basic

Hydrophobic

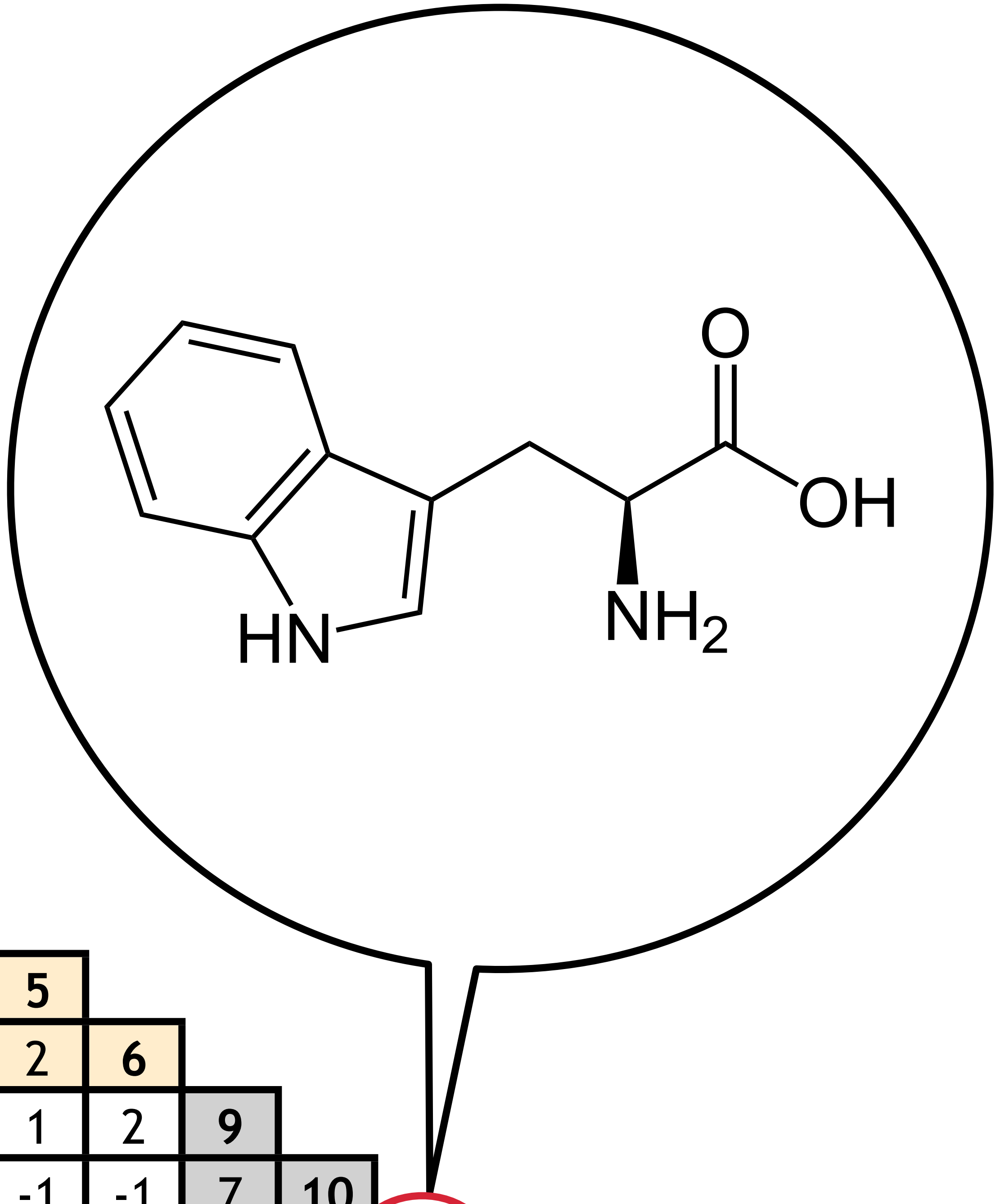
Aromatic

Sequence Alignments: PAM 250

Cys	C	12																			
Gly	G	-3	5																		
Pro	P	-3	-1	6																	
Ser	S	0	1	1	2																
Ala	A	-2	1	1	1	2															
Thr	T	-2	0	0	1	1	3														
Asp	D	-5	1	-1	0	0	0	4													
Glu	E	-5	0	-1	0	0	0	3	4												
Asn	N	-4	0	-1	1	0	0	2	1	2											
Gln	Q	-5	-1	0	-1	0	-1	2	2	1	4										
His	H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
		C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W

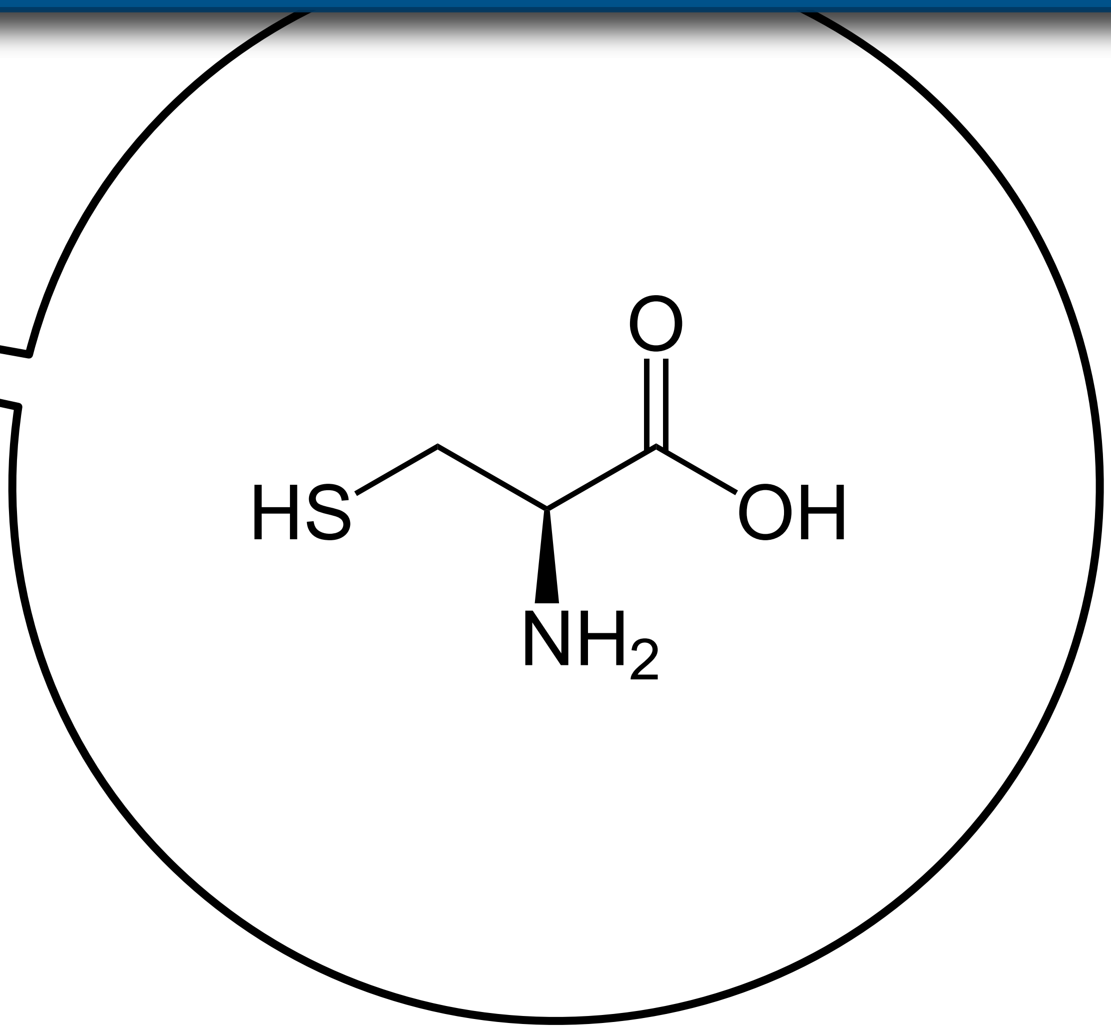
Sequence Alignments: PAM 250

Cys	C	12																			
Gly	G	-3	5																		
Pro	P	-3	-1	6																	
Ser	S	0	1	1	2																
Ala	A	-2	1	1	1	2															
Thr	T	-2	0	0	1	1	3														
Asp	D	-5	1	-1	0	0	0	4													
Glu	E	-5	0	-1	0	0	0	3	4												
Asn	N	-4	0	-1	1	0	0	2	1	2											
Gln	Q	-5	-1	0	-1	0	-1	2	2	1	4										
His	H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
		C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W



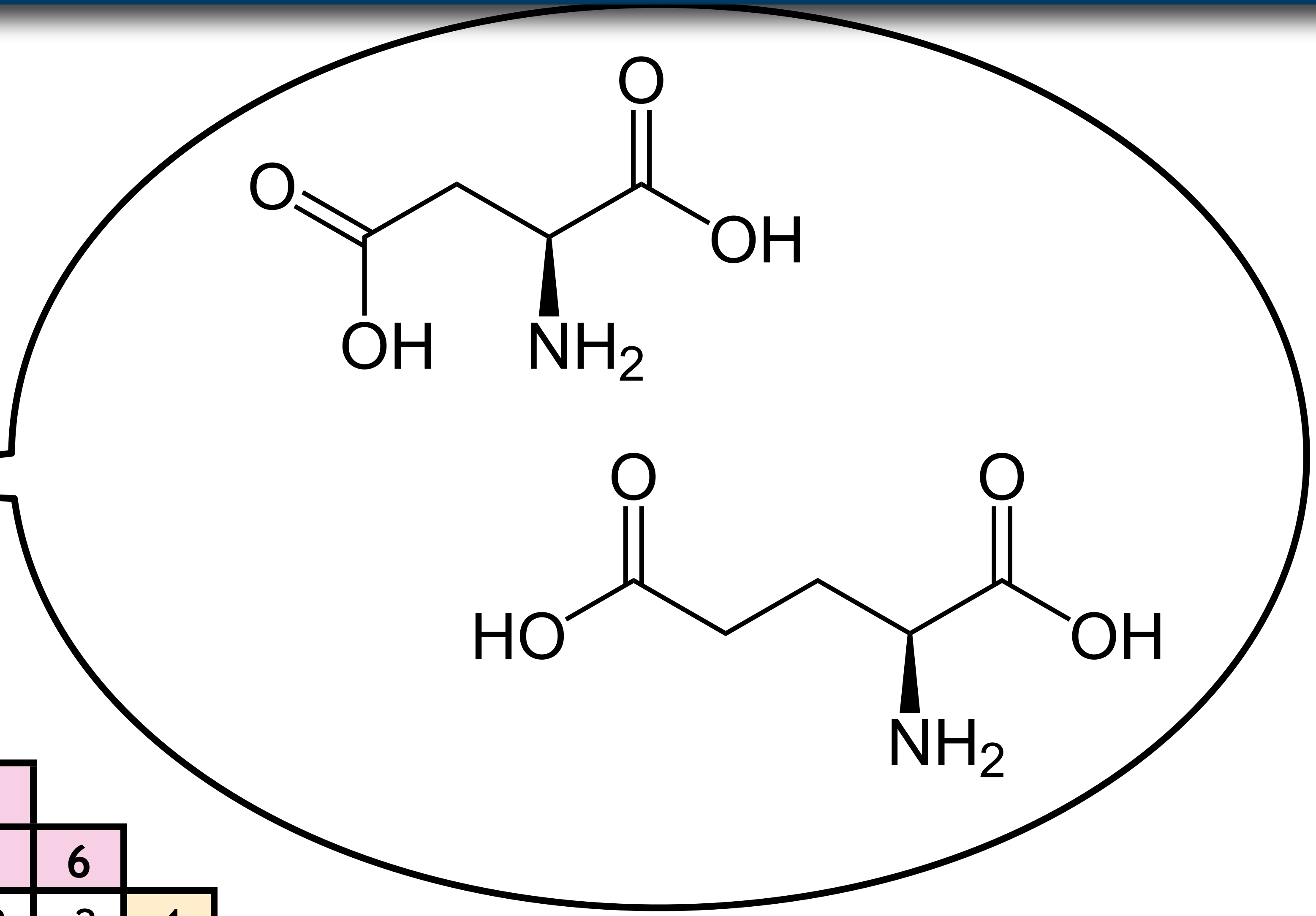
Sequence Alignments: PAM 250

Cys	C	12																			
Gly	G	-3	5																		
Pro	P	-3	-1	6																	
Ser	S	0	1	1	2																
Ala	A	-2	1	1	1	2															
Thr	T	-2	0	0	1	1	3														
Asp	D	-5	1	-1	0	0	0	4													
Glu	E	-5	0	-1	0	0	0	3	4												
Asn	N	-4	0	-1	1	0	0	2	1	2											
Gln	Q	-5	-1	0	-1	0	-1	2	2	1	4										
His	H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
		C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W



Sequence Alignments: PAM 250

Cys	C	12																			
Gly	G	-3	5																		
Pro	P	-3	-1	6																	
Ser	S	0	1	1	2																
Ala	A	-2	1	1	1	2															
Thr	T	-2	0	0	1	1	3														
Asp	D	-5	1	-1	0	0	0	4													
Glu	E	-5	0	-1	0	0	0	3	4												
Asn	N	-4	0	-1	1	0	0	2	1	2											
Gln	Q	-5	-1	0	-1	0	-1	2	2	1	4										
His	H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
		C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W

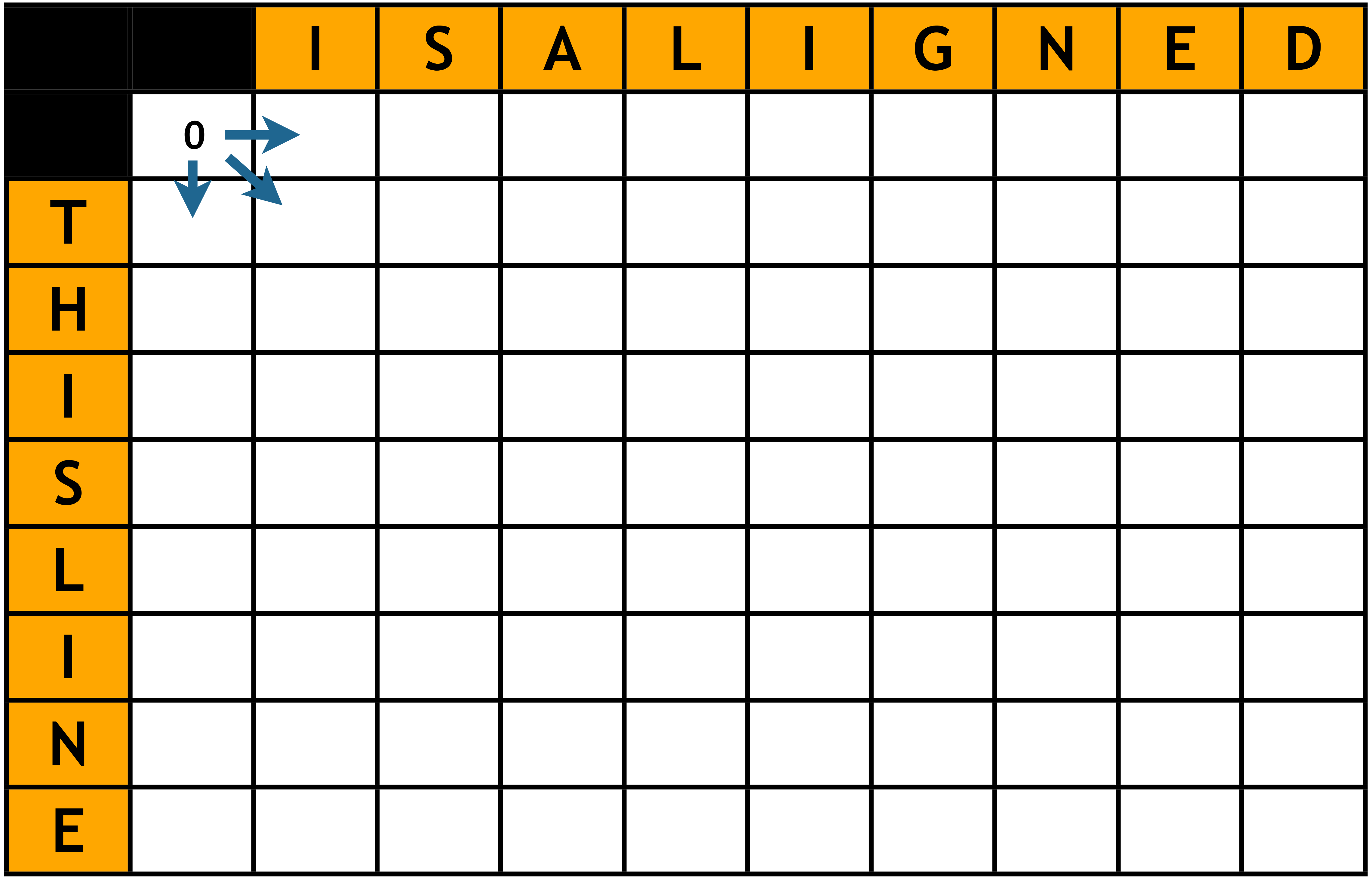


Sequence Alignments: Dynamic Programming

- Dynamic programming
  - Turn complex calculation into recursive series of simpler calculations
  - Guaranteed to find the optimal local alignment
    - (With respect to the scoring matrix used)
  - Relatively slow

- Needleman & Wunsch (global alignments)
- Smith & Waterman (local alignments)

Sequence Alignments: Needleman & Wunsch



		I	S	A	L	I	G	N	E	D
	0									
T										
H										
I										
S										
L										
I										
N										
E										

Create a Matrix

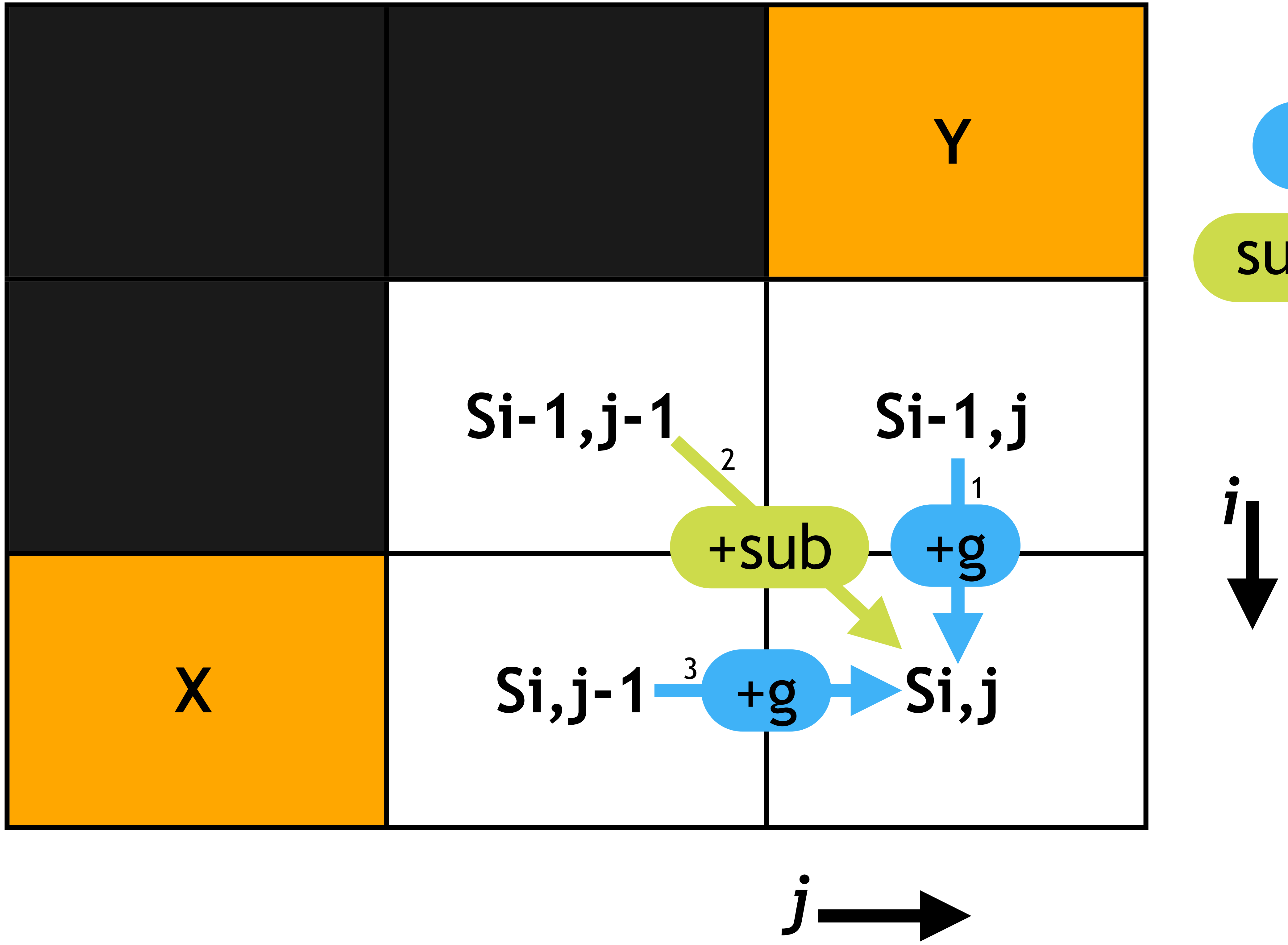
Fill the Matrix starting from the top left corner

Sequence Alignments: Needleman & Wunsch

		I	S	A	L	I	G	N	E	D
	0	-8	-16	-24	-32	-40	-48	-56	etc.	$S_{0,j}$
T	-8	?								
H	-16									
I	-24									
S	etc									
L	$S_{i,0}$									
I										
N										
E										

Gap penalties
at the borders

Sequence Alignments: Needleman & Wunsch

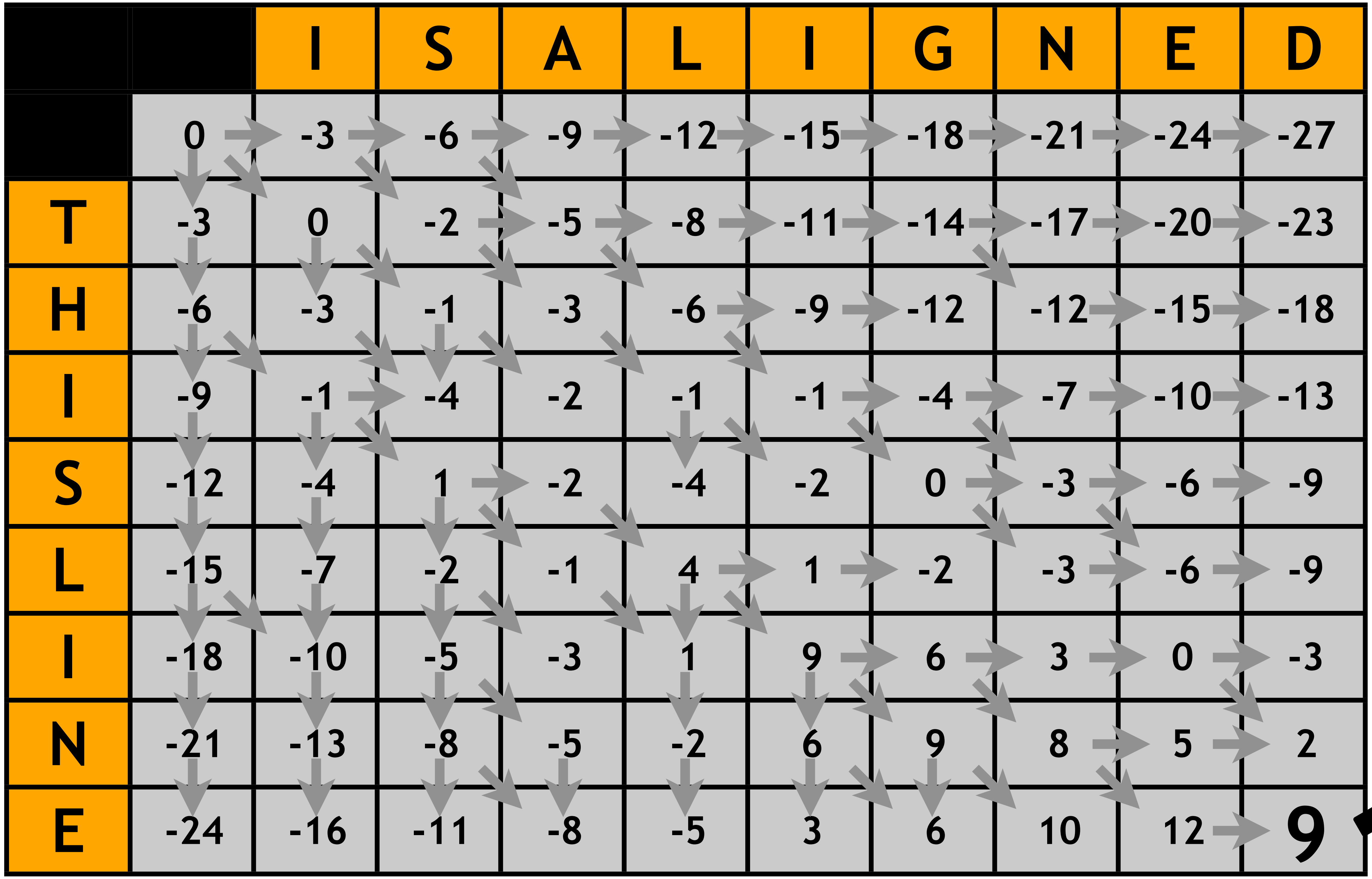


g = Gap penalty

sub = Substitution (Xi,Yj)

3 options:
Choose highest score
Mark chosen path

Sequence Alignments: Needleman & Wunsch

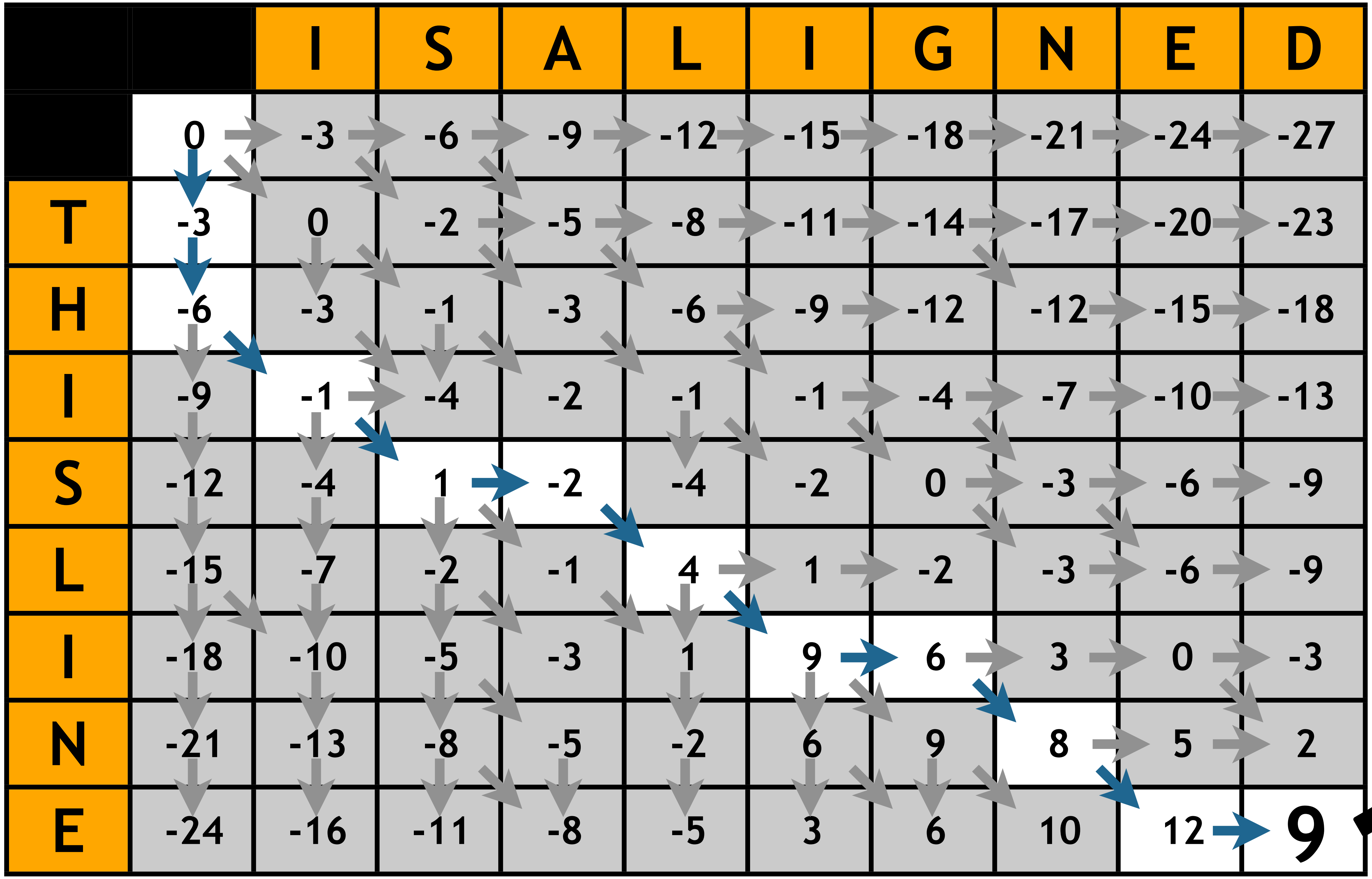


		I	S	A	L	I	G	N	E	D
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27
T	-3	0	-2	-5	-8	-11	-14	-17	-20	-23
H	-6	-3	-1	-3	-6	-9	-12	-12	-15	-18
I	-9	-1	-4	-2	-1	-1	-4	-7	-10	-13
S	-12	-4	1	-2	-4	-2	0	-3	-6	-9
L	-15	-7	-2	-1	4	1	-2	-3	-6	-9
I	-18	-10	-5	-3	1	9	6	3	0	-3
N	-21	-13	-8	-5	-2	6	9	8	5	2
E	-24	-16	-11	-8	-5	3	6	10	12	**9**

PAM 250
Gap penalty = -3

Score = 9

Sequence Alignments: Needleman & Wunsch



		I	S	A	L	I	G	N	E	D
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27
T	-3	0	-2	-5	-8	-11	-14	-17	-20	-23
H	-6	-3	-1	-3	-6	-9	-12	-12	-15	-18
I	-9	-1	-4	-2	-1	-1	-4	-7	-10	-13
S	-12	-4	1	-2	-4	-2	0	-3	-6	-9
L	-15	-7	-2	-1	4	1	-2	-3	-6	-9
I	-18	-10	-5	-3	1	9	6	3	0	-3
N	-21	-13	-8	-5	-2	6	9	8	5	2
E	-24	-16	-11	-8	-5	3	6	10	12	9

PAM 250
Gap penalty = -3

```
THIS-LI-NE-
  || || ||
--ISALIGNED
```

Score = 9

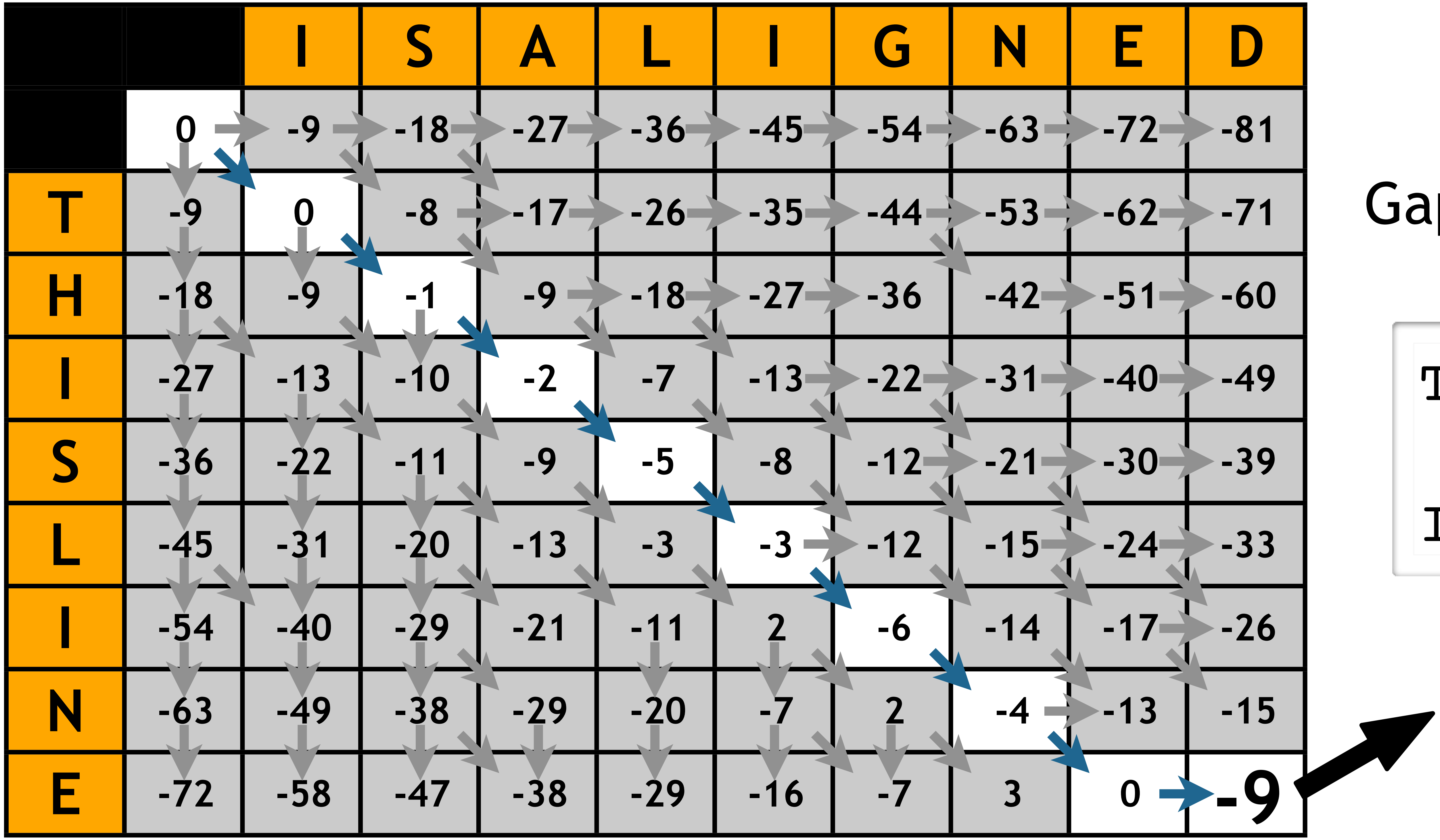
Sequence Alignments: Needleman & Wunsch

		I	S	A	L	I	G	N	E	D
	0	-9	-18	-27	-36	-45	-54	-63	-72	-81
T	-9	0	-8	-17	-26	-35	-44	-53	-62	-71
H	-18	-9	-1	-9	-18	-27	-36	-42	-51	-60
I	-27	-13	-10	-2	-7	-13	-22	-31	-40	-49
S	-36	-22	-11	-9	-5	-8	-12	-21	-30	-39
L	-45	-31	-20	-13	-3	-3	-12	-15	-24	-33
I	-54	-40	-29	-21	-11	2	-6	-14	-17	-26
N	-63	-49	-38	-29	-20	-7	2	-4	-13	-15
E	-72	-58	-47	-38	-29	-16	-7	3	0	**-9**

PAM 250
Gap penalty = -9

Score = -9

Sequence Alignments: Needleman & Wunsch



		I	S	A	L	I	G	N	E	D
	0	-9	-18	-27	-36	-45	-54	-63	-72	-81
T	-9	0	-8	-17	-26	-35	-44	-53	-62	-71
H	-18	-9	-1	-9	-18	-27	-36	-42	-51	-60
I	-27	-13	-10	-2	-7	-13	-22	-31	-40	-49
S	-36	-22	-11	-9	-5	-8	-12	-21	-30	-39
L	-45	-31	-20	-13	-3	-3	-12	-15	-24	-33
I	-54	-40	-29	-21	-11	2	-6	-14	-17	-26
N	-63	-49	-38	-29	-20	-7	2	-4	-13	-15
E	-72	-58	-47	-38	-29	-16	-7	3	0	-9

PAM 250
Gap penalty = -9

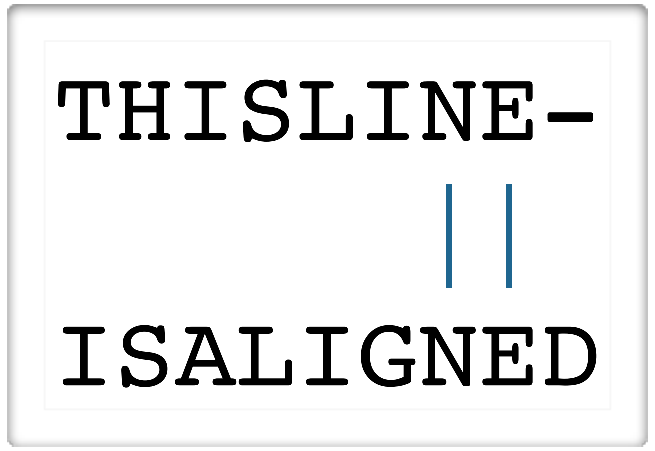
```
THISLINE-
      ||
ISALIGNED
```

Score = -9

Sequence Alignments: Needleman & Wunsch

- Gap penalty not tuned for matrix and problem
  - Crap alignment

PAM 250
Gap penalty = -9



```
THISLINE-
      ||
ISALIGNED
```

Score = -9

Sequence Alignments: Scoring

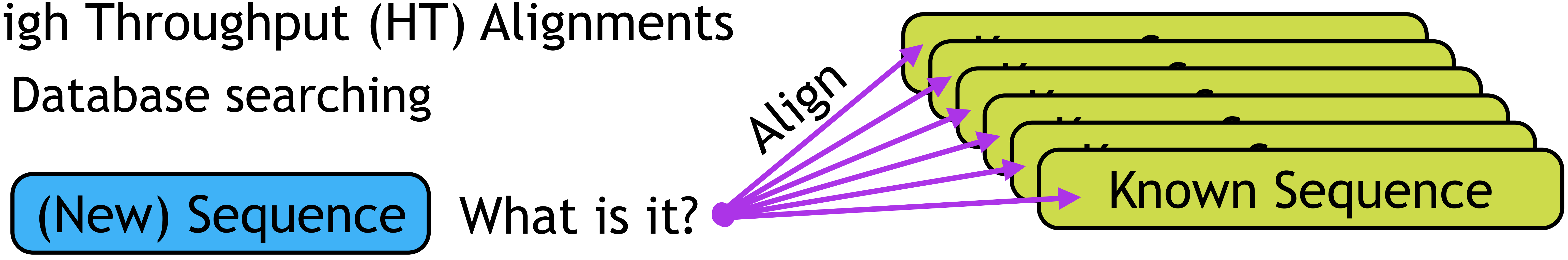
- Tune matrix for problem
- Tune gap penalties for matrix

- Most programs: reasonable defaults

- Usually you don’t know how well your sequences were conserved
  - Try defaults first for raw result / big picture
  - Try different matrices and gap penalties for fine tuned analysis

High Throughput Alignments: Intro

- High Throughput (HT) Alignments
  - Database searching



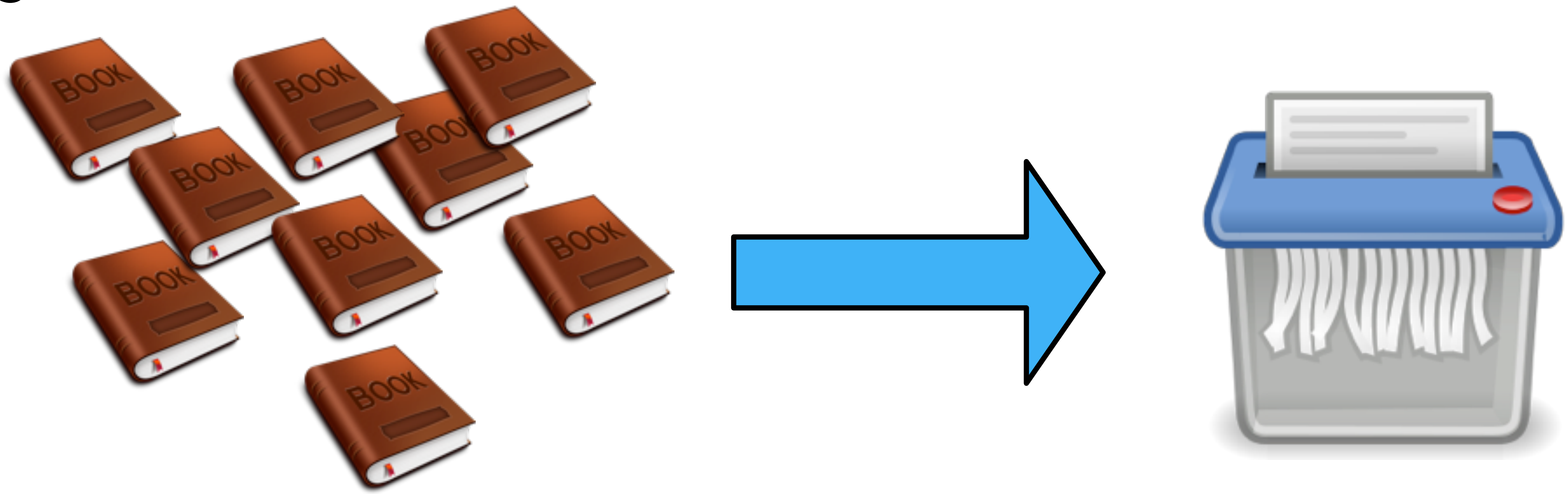
(New) Sequence

What is it?

Align

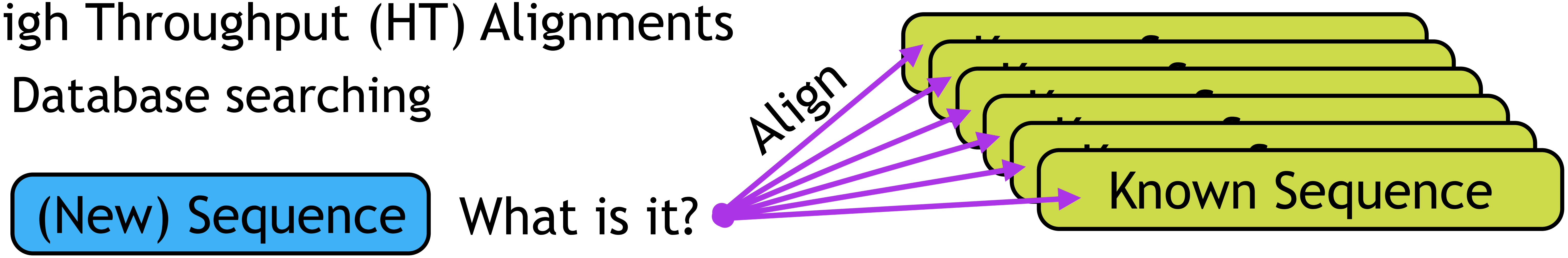
Known Sequence

  - Mapping of NGS reads to reference

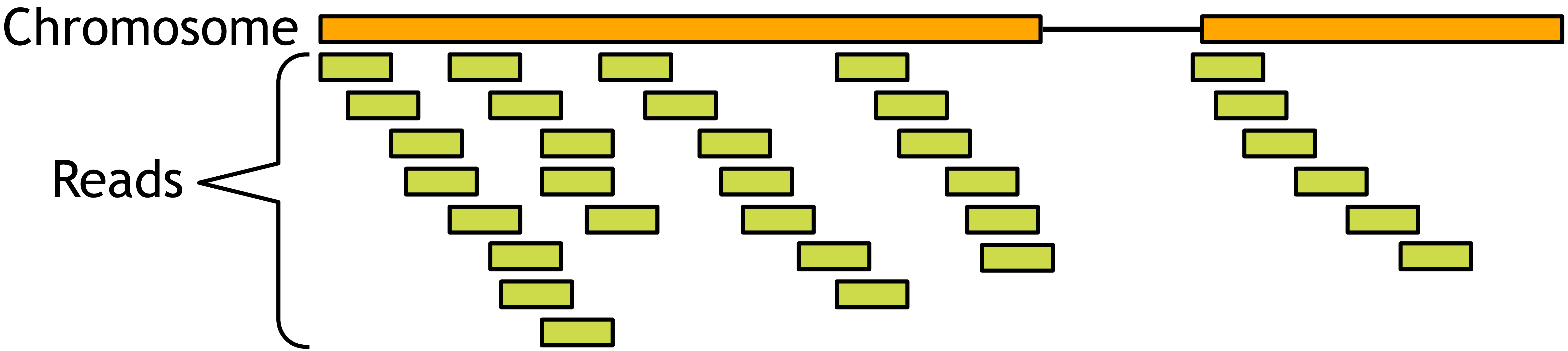


High Throughput Alignments: Intro

- High Throughput (HT) Alignments
  - Database searching



  - Mapping of NGS reads to reference



High Throughput Alignments: Intro

- Requires making lots of alignments
- HT: Requires Efficiency
  - Efficient algorithms / software
    - Exact → Approximate algorithms
  - Dedicated hardware
  - Reduce search space
    - Use NR (Non Redundant) databases
    - Remove uninteresting parts from query or database*
      - Repeat Masking | Low Complexity Filtering

* Uninteresting for most != all Biologists

Sequence Database Searching: Intro

- Repeat Masking | Low Complexity Filtering
  - For example HTT gene ➙ Huntington disease

```
  1 matleklmka feslksfqqq qqqqqqqqqq qqqqqqqqqq pppppppppp pqlpqpppqa
 61 qpllpqpqpp ppppppppgp avaeeplhrp kkelsatkkd rvnhcltice nivaqsvrns
121 pefqkllgia melfllcsdd aesdvrmvad eclnkvikal mdsnlprlql elykeikkng
181 ...
```

Repeat Length	Phenotype
6-35	Healthy
36-40	Healthy | Diseased
41+	Diseased

Sequence Database Searching: Intro

- Repeats | Low Complexity Regions
  - In genes: Rare
  - In the rest of the genome: Abundant
    - Centromeres
    - Telomeres
    - (DNA) Transposons ± 3 %
    - Retrotransposons ± 42 %

Human

- Masking | Filtering usually enabled by default

Sequence Database Searching: Intro

- Terminology
  - Query: sequence we use to search in a database
  - Hit (Subject): similar sequence we found in a database

Sequence Database Searching: Intro

- Exact algorithms
    - Use local alignments to compare the query to *all* database entries
    - Highest sensitivity, but slooowww

- Approximate or *Heuristic* algorithms
    - Use short-cut to **skip** alignment of query versus **entire** database
    - Less sensitive, but fst
        - i.e. BLAST
            - Find small identical (or high scoring) stretches: “*words*” / “*seeds*”
            - Use “*words*” to initiate local alignment of surrounding region

Sequence Database Searching: Intro

- Approximate or *Heuristic* algorithms
  - FastA
  - BLAST
  - BLAT
  - SSAHA2
  - ... BWA-SW

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
```

Sequence Database Searching: BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Heuristic approach based on Smith-Waterman
  - Most widely used alignment tool
  - Most widely used bioinformatics tool
  - All combinations possible

Query	**Database**	**Blast**
Protein	Protein	blastp
DNA	DNA	blastn
Translated	Protein	blastx
Protein	Translated	tblastn
Translated	Translated	tblastx

Sequence Database Searching: BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Assumes
    - Random sequences
    - Constant composition
  - Reports
    - Alignments surprisingly different from expectation (with expectation based on the assumptions above)

Sequence Database Searching: BLAST

- 1. Find words (seeds) in query
  - DNA Q: ATTCGCGTGAGTGCCCGGTGTGAGAGAC

```
ATTCGCGTGAG
 TTCGCGTGAGT
  TCGCGTGAGTG
   CGCGTGAGTGC
    GCGTGAGTGCC
     CGTGAGTGCCC
```

Default word size = 11

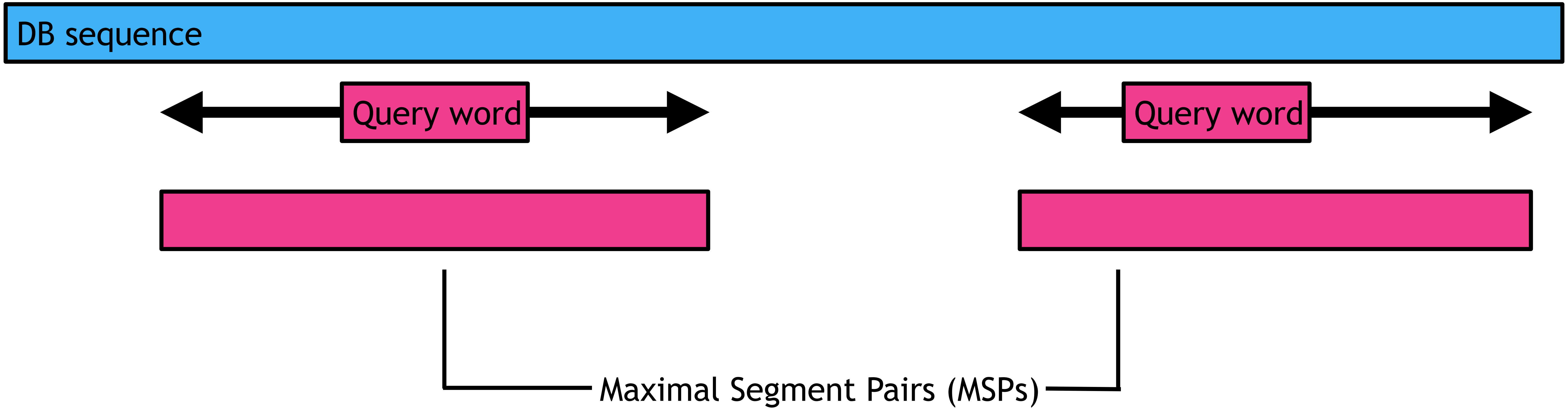
et cetera

Sequence Database Searching: BLAST

- 2. Search databases for words from query
  - DNA hit requires at least one exact matching word
  - Protein hit requires at least two words (exact matches or neighborhood words)

Sequence Database Searching: BLAST

- 3. Extend alignment in both directions
  - Find ungapped alignments with score above a threshold

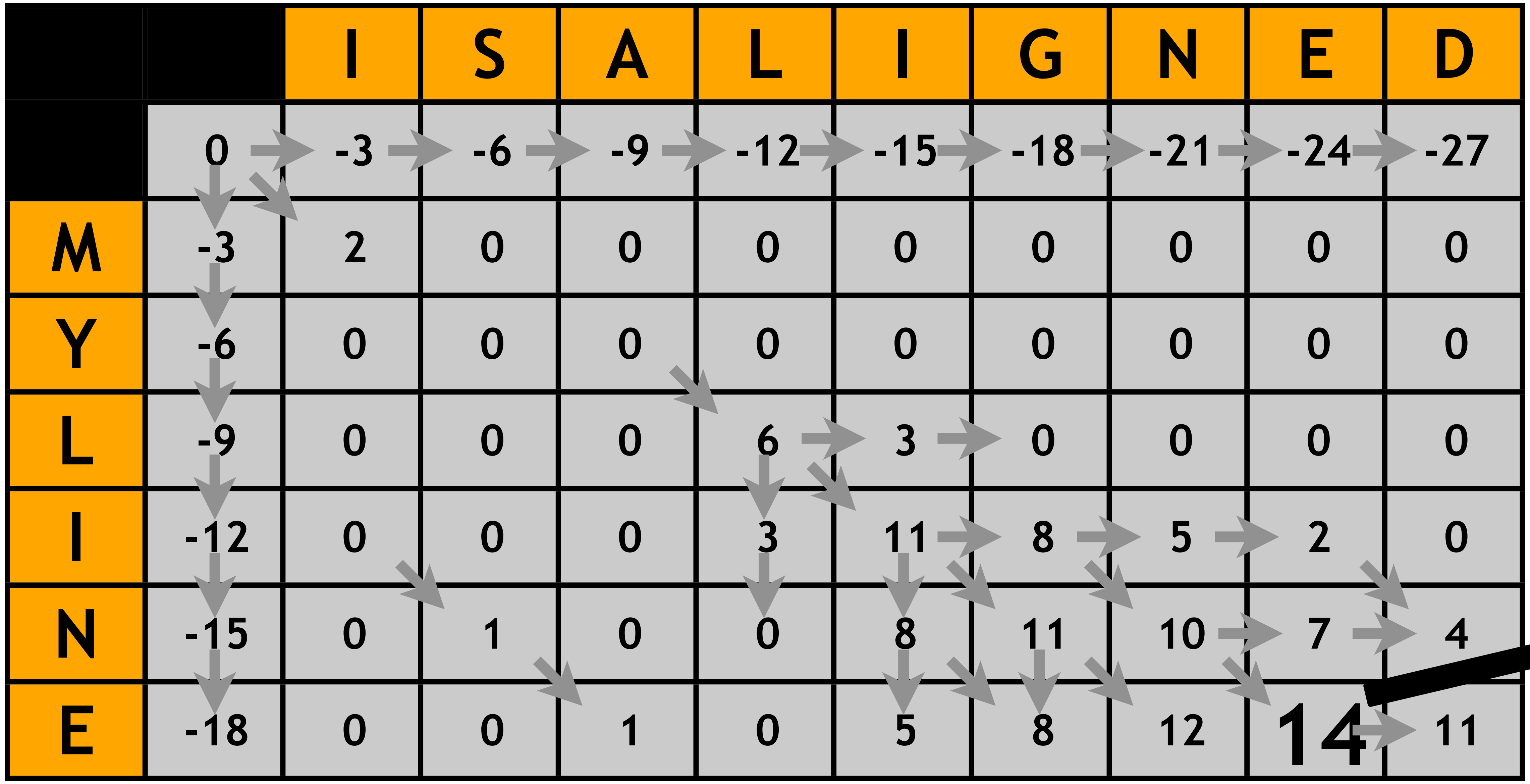


Sequence Database Searching: BLAST

- Ungapped local alignments
  - Pair of equal length segments: one from query; other from hit
  - Modified Smith-Waterman or Sellers algorithms find:
    - Maximal Segment Pairs (MSPs):
      Pairs whose scores cannot be improved by extension or trimming
    - High-scoring Segment Pairs (HSPs):
      MSP with score above a certain threshold

Sequence Alignments: Smith & Waterman

- Set negative cell values to zero
- Trace back from highest cell value to zero



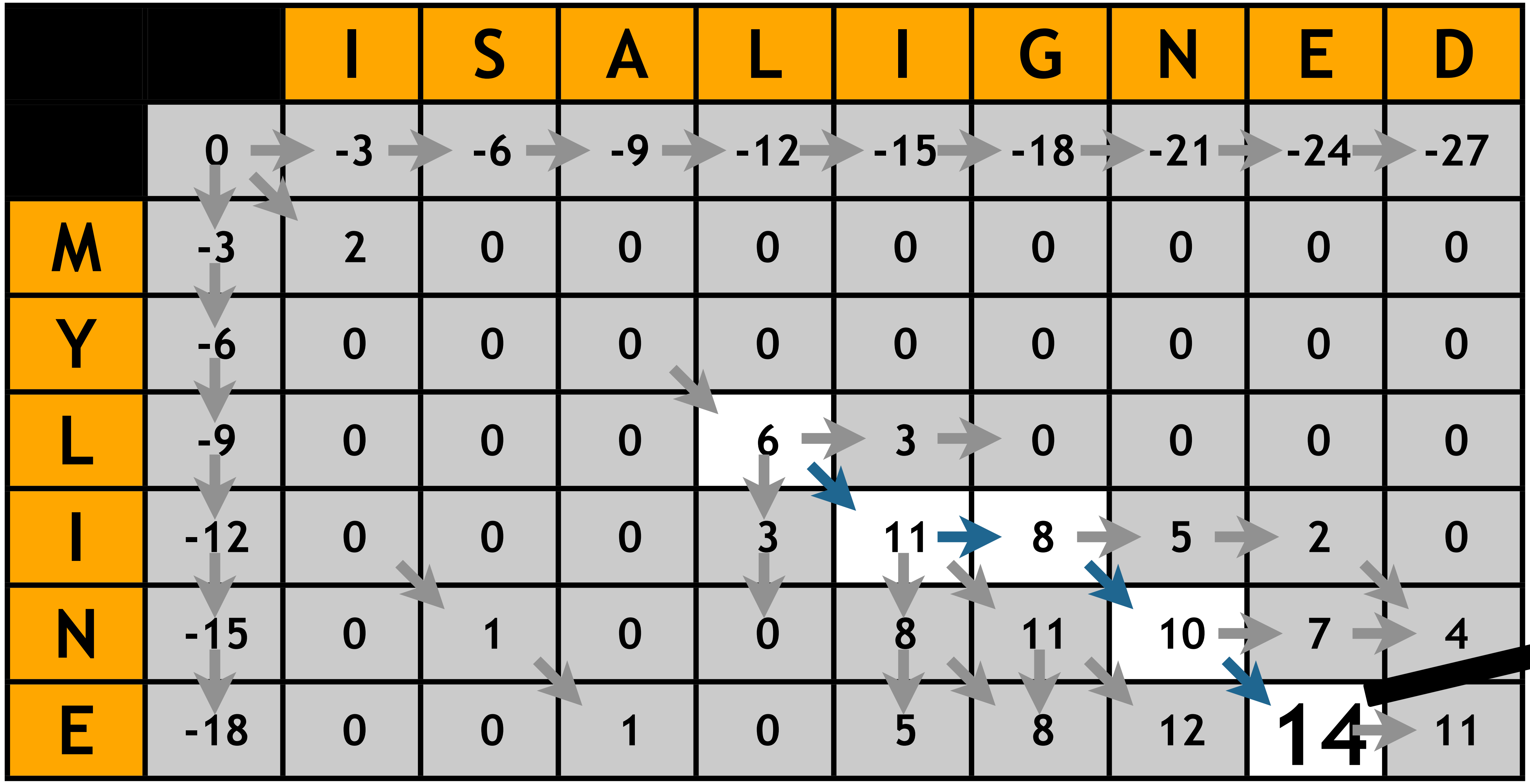
		I	S	A	L	I	G	N	E	D
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27
M	-3	2	0	0	0	0	0	0	0	0
Y	-6	0	0	0	0	0	0	0	0	0
L	-9	0	0	0	6	3	0	0	0	0
I	-12	0	0	0	3	11	8	5	2	0
N	-15	0	1	0	0	8	11	10	7	4
E	-18	0	0	1	0	5	8	12	14	11

PAM 250
Gap penalty = -3

Score = 14

Sequence Alignments: Smith & Waterman

- Requires substitution matrix with:
  Positive scores for good and negative scores for bad matches



		I	S	A	L	I	G	N	E	D
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27
M	-3	2	0	0	0	0	0	0	0	0
Y	-6	0	0	0	0	0	0	0	0	0
L	-9	0	0	0	6	3	0	0	0	0
I	-12	0	0	0	3	11	8	5	2	0
N	-15	0	1	0	0	8	11	10	7	4
E	-18	0	0	1	0	5	8	12	14	11

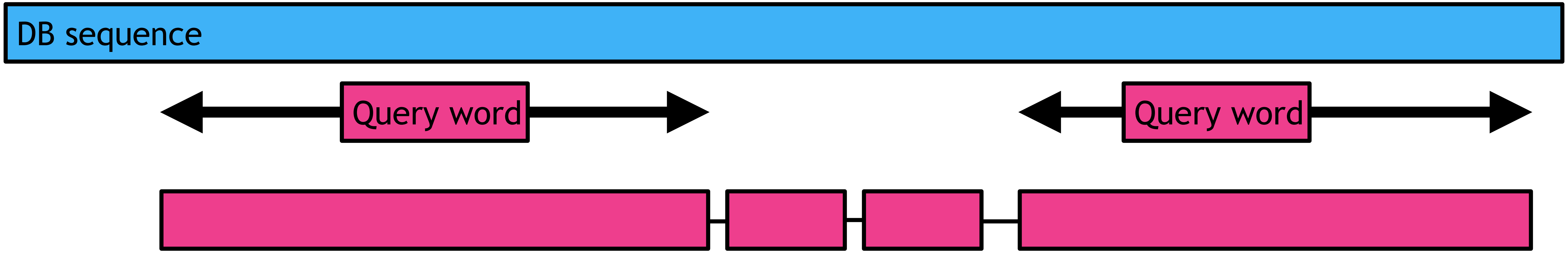
PAM 250
Gap penalty = -3

```
 MYLI-NE
   || ||
ISALIGNED
```

Score = 14

Sequence Database Searching: BLAST

- 4. Join MSPs if they are close together
  - May introduce gaps for a gapped alignment



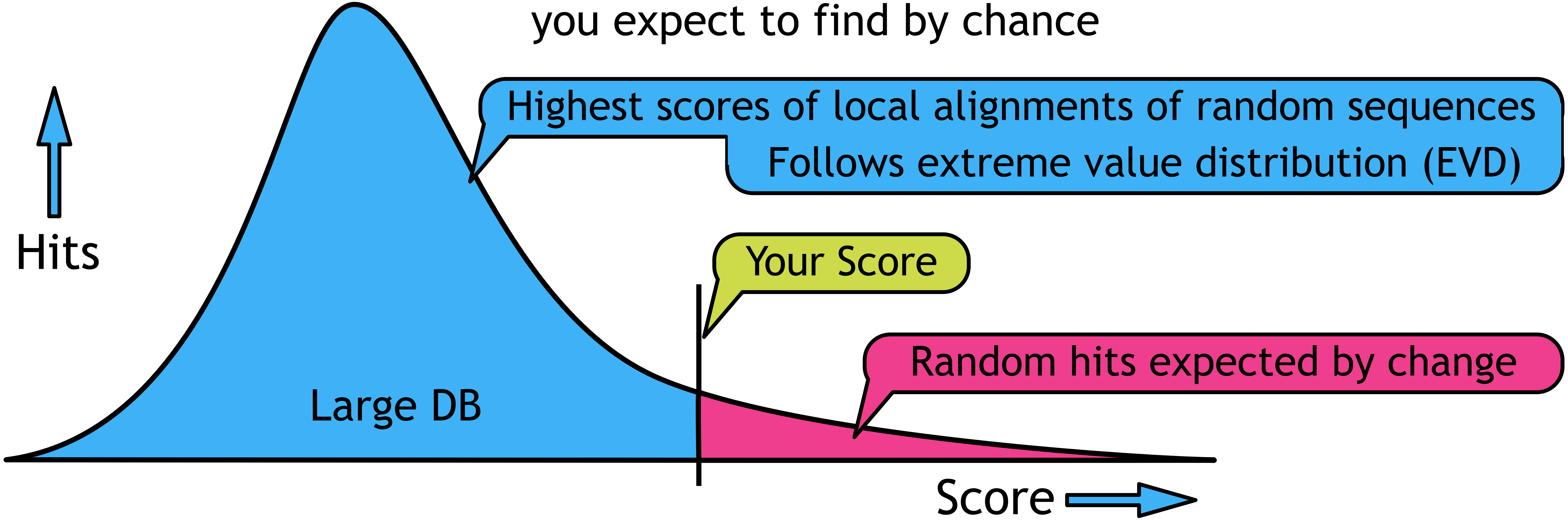
Sequence Database Searching: Q1

- The following alignment has 45/60 (75%) identities / matches
  - Not a bad score, but...
  - BLAST will not be able to find this alignment.
  - Why?

```
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGACGTTT
|||| ||| | ||| |||||| ||| ||||  |||  ||||| | | | ||||| |||||
AATGTGTATCGGTGAGAATCAGATGCGCGCTGTACTAGTCAGCGACTGAACCTGCCGTTT
```

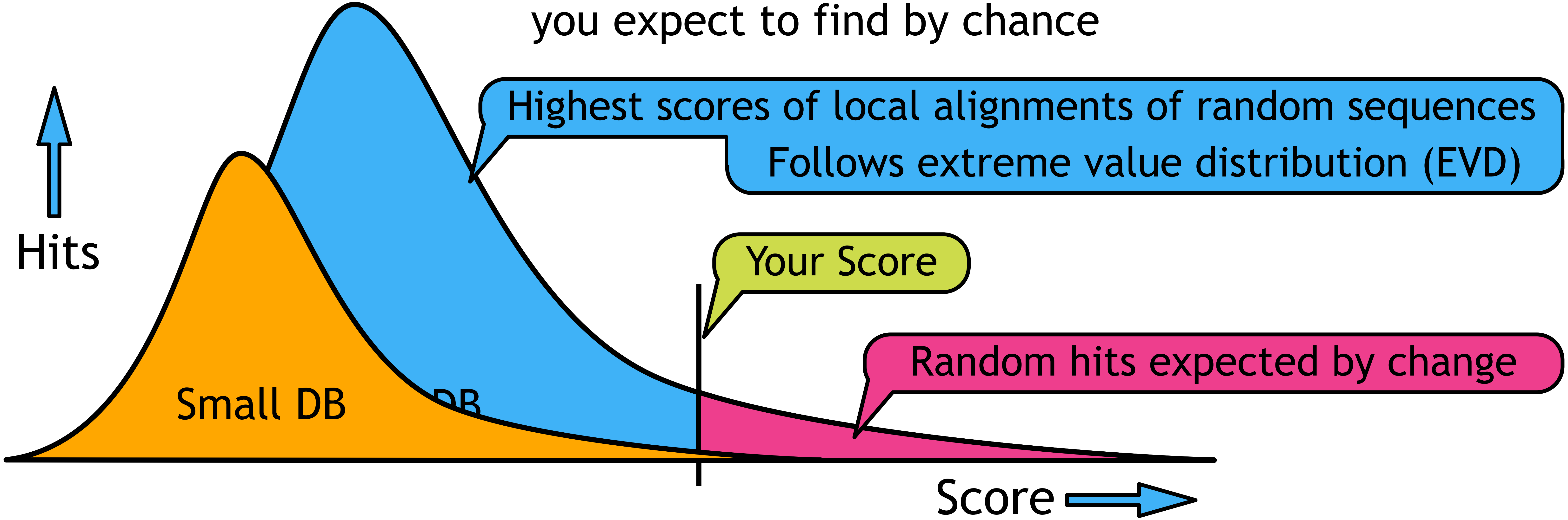
Sequence Database Searching: BLAST

- Statistics
  - Expect value (E)
    - E-value for score S = number of hits with score S or better you expect to find by chance


Highest scores of local alignments of random sequences
Follows extreme value distribution (EVD)
Hits
Your Score
Random hits expected by change
Large DB
Score

Sequence Database Searching: BLAST

- Statistics
  - Expect value (E)
    - E-value for score S = number of hits with score S or better you expect to find by chance



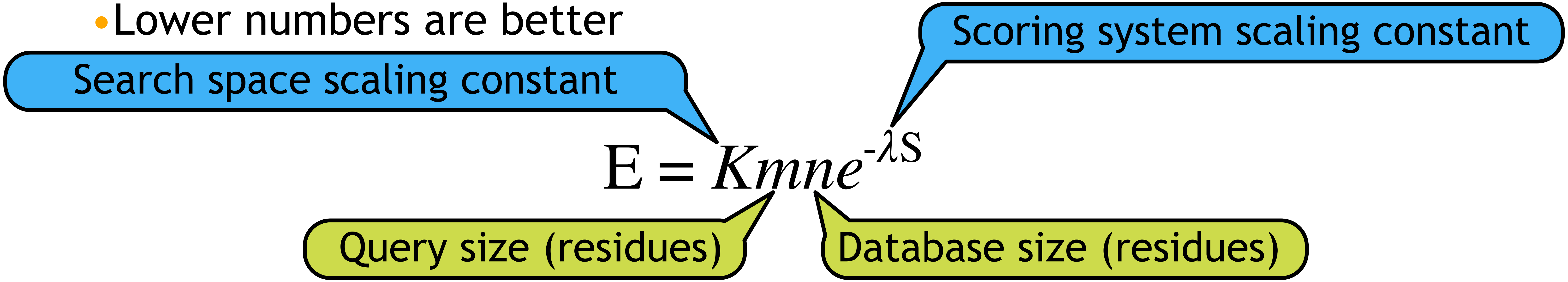
Highest scores of local alignments of random sequences
Follows extreme value distribution (EVD)

Your Score

Random hits expected by change

Sequence Database Searching: BLAST

- Statistics
  - Expect value (E)
    - E-value for score S = number of hits with score S or better you expect to find by chance
    - Depends on search space (mostly DB size)
    - Depends on the scale of the scoring system
    - Lower numbers are better



Scoring system scaling constant

Search space scaling constant

$$E = Kmne^{-\lambda S}$$

Query size (residues)

Database size (residues)

Sequence Database Searching: BLAST

- Statistics
  - Bit score (S’)
    - Normalised score
    - Can be used to compare scores from different searches
    - Higher numbers are better

Scoring system scaling constant

Search space scaling constant

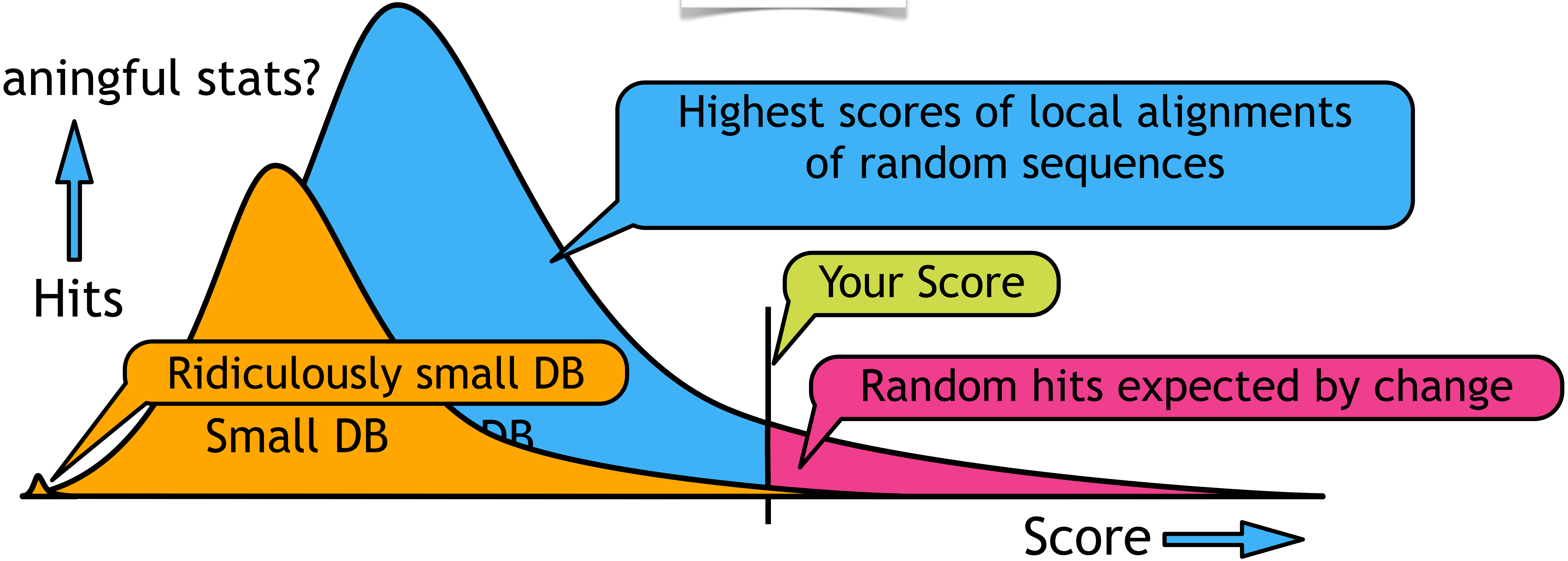
$$S' = \frac{\lambda S - \ln K}{\ln 2} \longrightarrow E = mn2^{-S'}$$

Query size (residues)

Database size (residues)

Sequence Database Searching: Q2

- Can/should you use BLAST to align two sequences? (DB with n=1)
    - Quality for spent resources a.k.a for your ?
    - Meaningful stats?


Hits
Highest scores of local alignments of random sequences
Your Score
Random hits expected by change
Ridiculously small DB
Small DB
DB
Score

Rules of thumb for DB searching

- Searching for distant homologs
  - Large evolutionairy distance
  - Align **proteins** for **increased resolution** of similarities using
    - Silent mutations
    - Similar amino acids

DNA		Protein
AUA	→	Ile
CUC	→	Leu
UUG	→	Leu

Rules of thumb for DB searching

- Searching for distant homologs
  - Large evolutionairy distance
  - Align **proteins** for **increased resolution** of similarities using
    - Translated searches: Check genetic code!

Vertebrate DNA	Codon	AA
Nuclear	AUA	Ile
Mitochondrial	AUA	Met
Nuclear	UGA	Stop
Mitochondrial	UGA	Trp

Rules of thumb for DB searching

- Too many hits
  - Decrease E-value
  - Change scoring matrix
    - PAM-- | BLOSUM++
  - Enable “low complexity filtering” | “repeat masking”
  - Split the sequence (in domains, genes, ...)
    - Is your query sequence a fusion protein or large genomic fragment?

Rules of thumb for DB searching

- No hits or only a few
  - Decrease word size
  - Increase E-value
  - Change scoring matrix
    - PAM++ | BLOSUM--
  - Disable “low complexity filtering” | “repeat masking”
    - Usually enabled by default
    - Are you looking for repeats or transposons?

Rules of thumb for DB searching

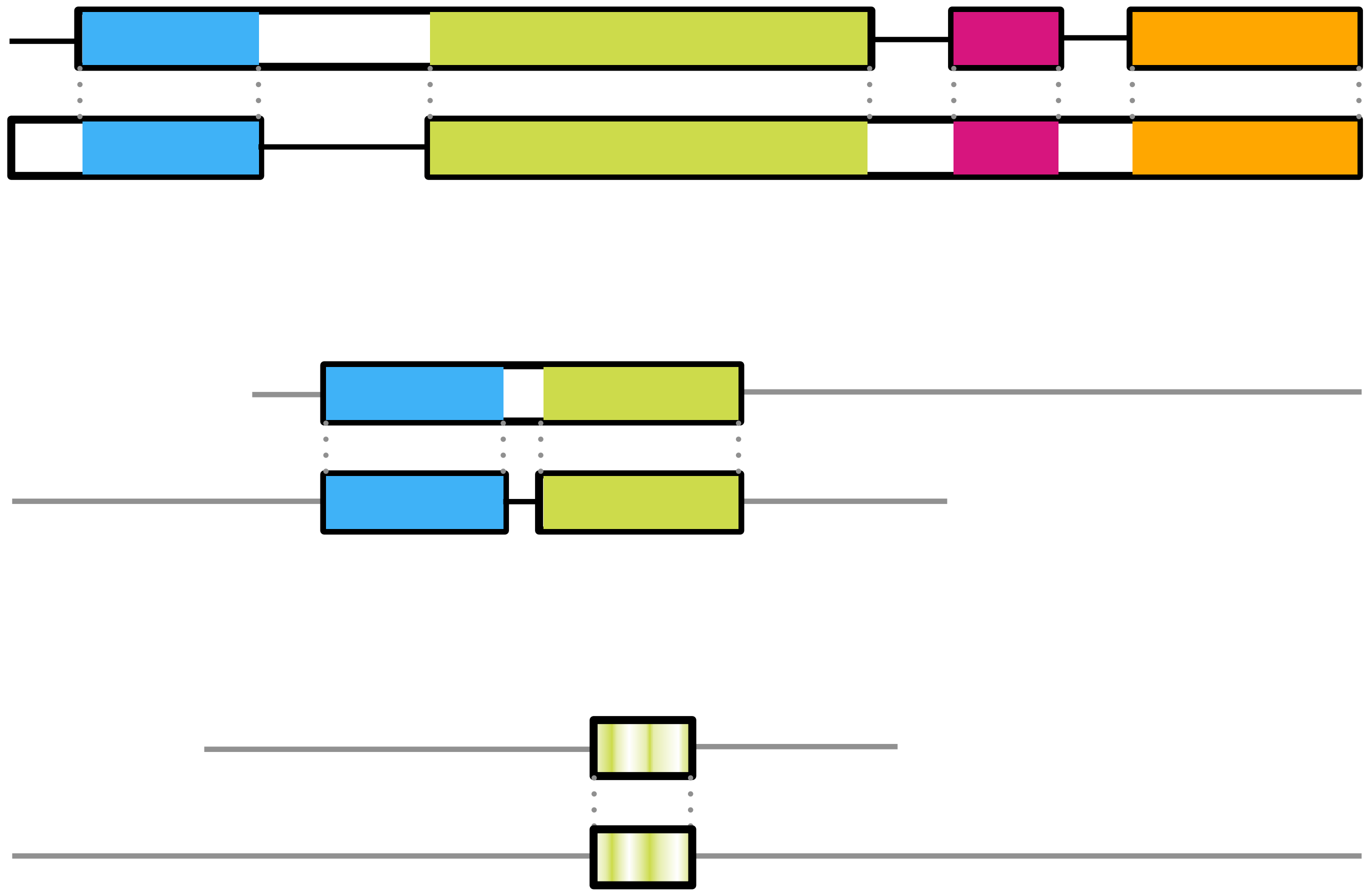
- Still no hits or only a few
  - BLAST not sensitive enough
  - Switch to motif searching
    - Pickup more distant homologs

Rules of thumb for similarity searching

Global Alignment

Local Alignments
(DB Searching)

Sequence Motifs
(Motif Searching)



Sequences	Homology
few	high
many	medium
many	low

Sequence Database Searching: Q3

- Why not to increase the word size in case you get too many hits?
  - Example: word size = 20 nucleotides

```
                                                              Found
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
||||||||||||||||||||||||||||||||||||||||||||||||||||||||        ✓
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC

TTACCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
    ||||||||||||||||||||                                        ✓
AATGCGTAACCGTGCGAATCAAATATAGTAACGTTGCCATGCCCATTCGTGACACG

AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
||||||||||||||| ||||||||||||||| ||||||||||||||| ||||||||        ✗
AATGCGTAACCGTGCCAATCAAATGGGCGCGATACATGTCAGAGTCATTACCTGAC
```

  - May miss relevant hits!

Take Home Message

Speed ↔ Accuracy

Heuristic ↔ Complete

Check if heuristics are compatible with your research question!

Sequence Database Searching: Exercises

- 4. The following alignment has 45/60 (75%) identities / matches
  - Not a bad score, but...
  - Neither BLAST nor BLAT will be able to find this alignment.
  - Explain why.

```
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGACGTTT
|||| ||| | ||| |||||| ||| ||||  |||  ||||| | | | ||||| |||||
AATGTGTATCGGTGAGAATCAGATGCGCGCTGTACTAGTCAGCGACTGAACCTGCCGTTT
```

Sequence Database Searching: Exercises

- 5. If you get too many hits
  - Should you increase the word size? Why?
- 6. Find out what this human sequence represents
  - Sequence: TACTACTACTGCTGCTGCTGCTGCT
  - Try BLAST @ www.ncbi.nlm.nih.gov
    - Go to section *BLAST Assembled RefSeq Genomes* → human
    - Paste the query sequence, use defaults for everything else and hit BLAST
  - Try BLAT (default) @ www.ensembl.org
  - Try Google...

Sequence Database Searching: Exercises

- 3. Myoglobin vs. Hemoglobin with Hemoglobin as “DB” (bl2seq)

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha GN=HBA1
Length=142

 Score = 51.2 bits (121),  Expect = 6e-12, Method: Compositional matrix adjust.
 Identities = 41/149 (27%), Positives = 62/149 (41%), Gaps = 8/149 (5%)

Query  1    MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60
            M LS  +   V   WGKV A    +G E L R+F   P T   F  F      D    S
Sbjct  1    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSA  54

Query  61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSK  119
             +K HG  V  AL   +         +  L+  HA K ++ PV + + +S C++  L +
Sbjct  55   QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNF-KLLSHCLLVTLAAH  113

Query  120  HPGDFGADAQGAMNKALELFRKDMASNYK  148
             P +F      +++K L      + S Y+
Sbjct  114  LPAEFTPAVHASLDKFLASVSTVLTSKYR  142
```

Sequence Database Searching: Exercises

- 3. Myoglobin vs. Hemoglobin with UniProtKB Human as DB

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha GN=HBA1
Length=142

 Score = 51.2 bits (121),  Expect = 2e-07, Method: Compositional matrix adjust.
 Identities = 41/149 (27%), Positives = 62/149 (41%), Gaps = 8/149 (5%)

Query  1    MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60
            M LS  +   V   WGKV A    +G E L R+F   P T   F  F      D    S
Sbjct  1    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSA  54

Query  61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSK  119
             +K HG  V  AL   +         +  L+  HA K ++ PV + + +S C++  L +
Sbjct  55   QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNF-KLLSHCLLVTLAAH  113

Query  120  HPGDFGADAQGAMNKALELFRKDMASNYK  148
             P +F      +++K L      + S Y+
Sbjct  114  LPAEFTPAVHASLDKFLASVSTVLTSKYR  142
```

Sequence Database Searching: Exercises

- 5. Effect of increasing the word size in case you get too many hits
  - Example: word size = 20 nucleotides

Found

```
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
```

✓

```
TTACCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
    ||||||||||||||||||||
AATGCGTAACCGTGCGAATCAAATATAGTAACGTTGCCATGCCCATTCGTGACACG
```

✓

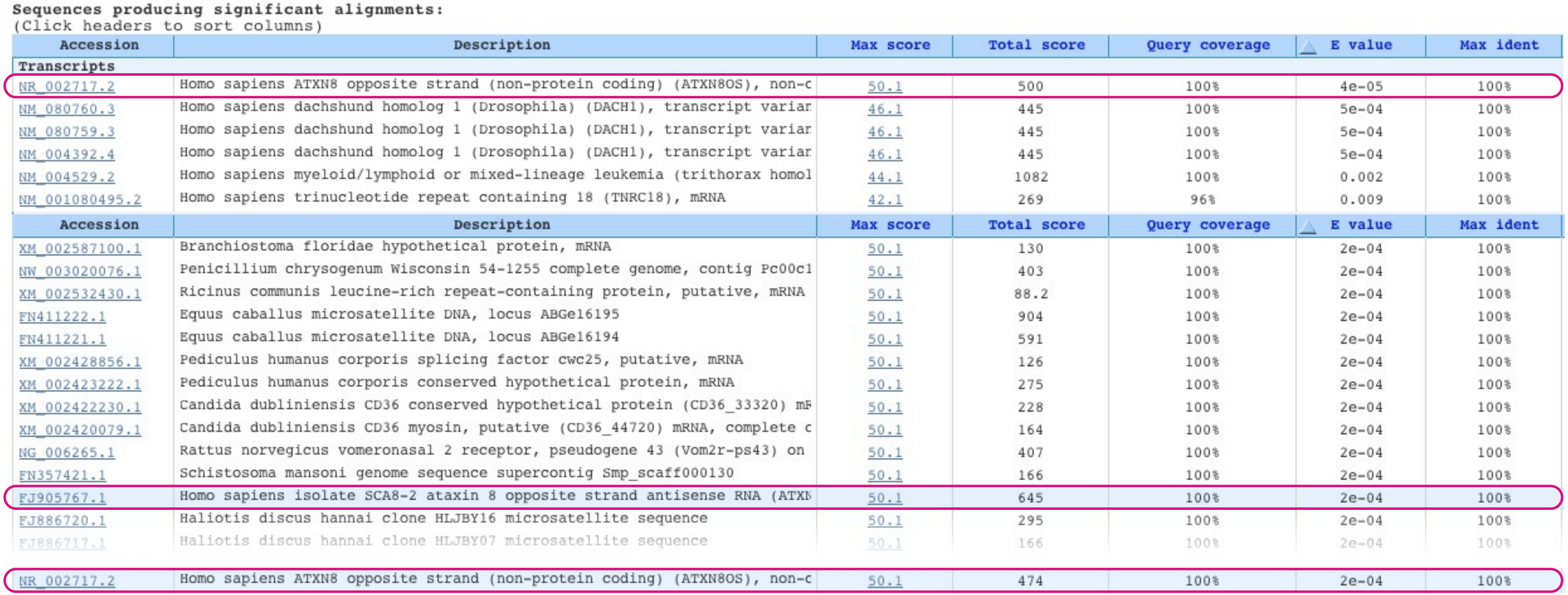
```
AATGCGTAACCGTGCGAATCAAATGGGCGCGTTACATGTCAGAGTCAGTACCTGAC
||||||||||||||| ||||||||||||||| ||||||||||||||| ||||||||
AATGCGTAACCGTGCCAATCAAATGGGCGCGATACATGTCAGAGTCATTACCTGAC
```

✗

  - May miss relevant hits!

Sequence Database Searching: Exercises

- 6. TACTACTACTGCTGCTGCTGCTGCT
  Homo sapiens ATXN8 opposite strand (ATXN8OS) (no protein)



Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
Transcripts						
NR_002717.2	Homo sapiens ATXN8 opposite strand (non-protein coding) (ATXN8OS), non-c	50.1	500	100%	4e-05	100%
NM_080760.3	Homo sapiens dachshund homolog 1 (Drosophila) (DACH1), transcript variar	46.1	445	100%	5e-04	100%
NM_080759.3	Homo sapiens dachshund homolog 1 (Drosophila) (DACH1), transcript variar	46.1	445	100%	5e-04	100%
NM_004392.4	Homo sapiens dachshund homolog 1 (Drosophila) (DACH1), transcript variar	46.1	445	100%	5e-04	100%
NM_004529.2	Homo sapiens myeloid/lymphoid or mixed-lineage leukemia (trithorax homol	44.1	1082	100%	0.002	100%
NM_001080495.2	Homo sapiens trinucleotide repeat containing 18 (TNRC18), mRNA	42.1	269	96%	0.009	100%

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
XM_002587100.1	Branchiostoma floridae hypothetical protein, mRNA	50.1	130	100%	2e-04	100%
NW_003020076.1	Penicillium chrysogenum Wisconsin 54-1255 complete genome, contig Pc00c1	50.1	403	100%	2e-04	100%
XM_002532430.1	Ricinus communis leucine-rich repeat-containing protein, putative, mRNA	50.1	88.2	100%	2e-04	100%
FN411222.1	Equus caballus microsatellite DNA, locus ABGe16195	50.1	904	100%	2e-04	100%
FN411221.1	Equus caballus microsatellite DNA, locus ABGe16194	50.1	591	100%	2e-04	100%
XM_002428856.1	Pediculus humanus corporis splicing factor cwc25, putative, mRNA	50.1	126	100%	2e-04	100%
XM_002423222.1	Pediculus humanus corporis conserved hypothetical protein, mRNA	50.1	275	100%	2e-04	100%
XM_002422230.1	Candida dubliniensis CD36 conserved hypothetical protein (CD36_33320) mF	50.1	228	100%	2e-04	100%
XM_002420079.1	Candida dubliniensis CD36 myosin, putative (CD36_44720) mRNA, complete c	50.1	164	100%	2e-04	100%
NG_006265.1	Rattus norvegicus vomeronasal 2 receptor, pseudogene 43 (Vom2r-ps43) on	50.1	407	100%	2e-04	100%
FN357421.1	Schistosoma mansoni genome sequence supercontig Smp_scaff000130	50.1	166	100%	2e-04	100%
FJ905767.1	Homo sapiens isolate SCA8-2 ataxin 8 opposite strand antisense RNA (ATXN	50.1	645	100%	2e-04	100%
FJ886720.1	Haliotis discus hannai clone HLJBY16 microsatellite sequence	50.1	295	100%	2e-04	100%
FJ886717.1	Haliotis discus hannai clone HLJBY07 microsatellite sequence	50.1	166	100%	2e-04	100%

NR_002717.2	Homo sapiens ATXN8 opposite strand (non-protein coding) (ATXN8OS), non-c	50.1	474	100%	2e-04	100%