



Bioinformatics courses

Principles of Bioinformatics (BSc) &
Fundamentals of Bioinformatics
(MSc)

Lecture 3: (local) alignment and
homology searching

Centre for Integrative Bioinformatics VU (IBIVU)

Faculty of Exact Sciences / Faculty of Earth and Life Sciences

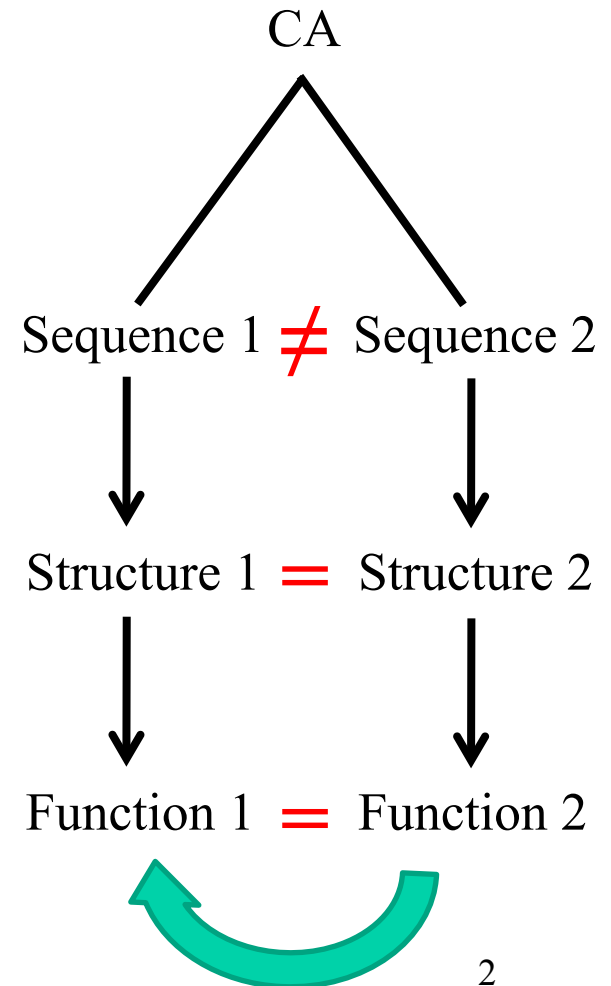
<http://ibi.vu.nl>, heringa@few.vu.nl, 87649 (Heringa), Room P1.28

Divergent evolution

sequence → structure → function

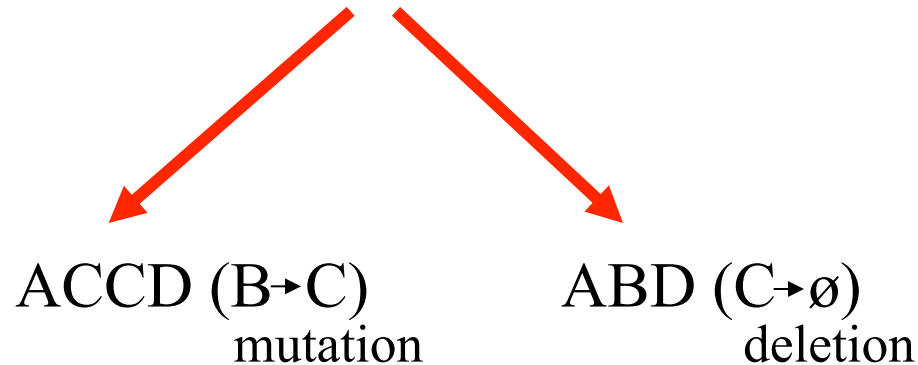
- Common ancestor (CA)
- Sequences change over time
- Protein structures typically remain the same (robust against multiple mutations)
- Therefore, function normally is preserved within orthologous families

“Structure more conserved than sequence”



Reconstructing divergent evolution

Ancestral sequence: ABCD



ACCD
AB—D

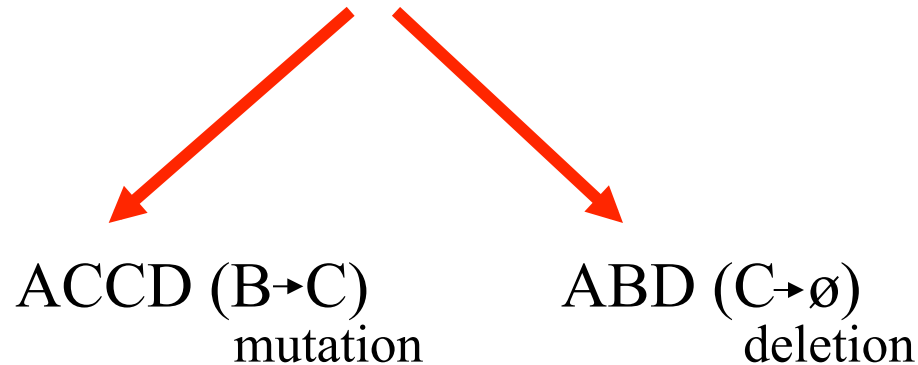
or

ACCD
A—BD

Pairwise Alignment

Reconstructing divergent evolution

Ancestral sequence: ABCD



ACCD
AB—D

or

ACCD
A—BD

Pairwise Alignment

true alignment

Pairwise alignment examples

A protein sequence alignment

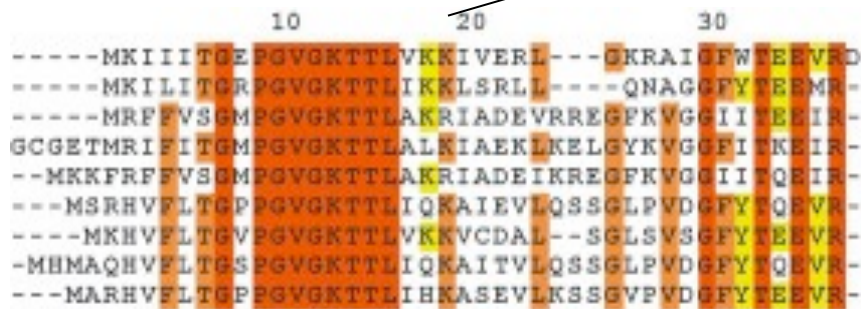
```
MSTGAVLIY--TSILIKECHAMPAGNE-----  
---GGILLFHRTHELIKESHAMANDEGGSNNS  
*      *      *      * * * *      * * *
```

A DNA sequence alignment

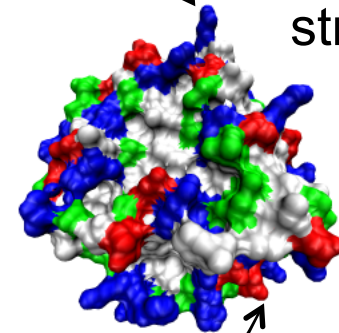
```
attcgttggcaaatcgcccctatccgggccttaa  
att---tggcggatcg-cctctacggggcc-----  
* * *      * * * *      * * * *      * *      * * * * * *
```

Evolution and three-dimensional protein structure information

Multiple alignment



Protein structure



What do we see if we colour code the space-filling (CPK) protein model?

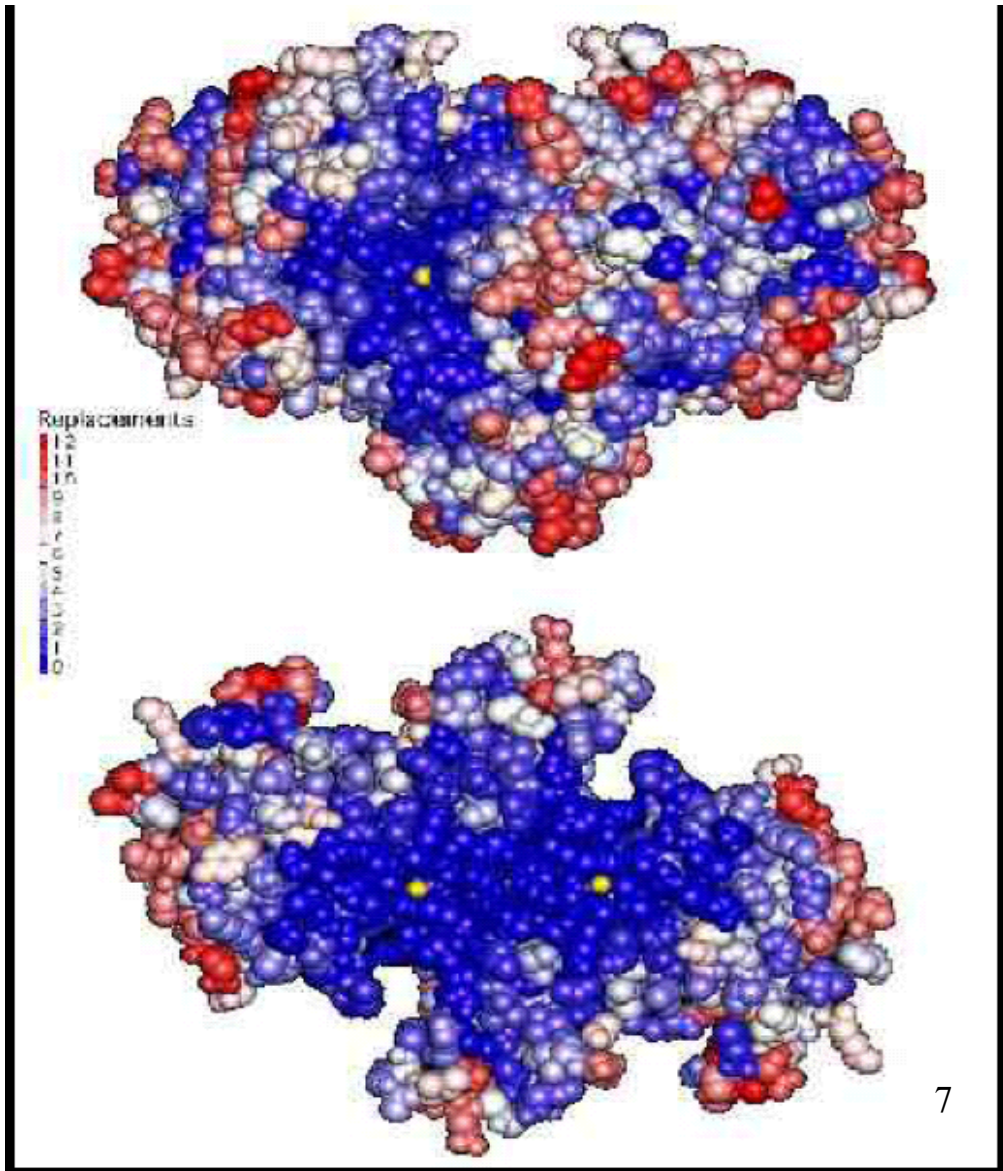
- E.g., red for conserved alignment positions to blue for variable (unconserved) positions.

Evolution and three-dimensional protein structure information

Isocitrate dehydrogenase:

The distance from the active site (in yellow) determines the rate of evolution (red = fast evolution, blue = slow evolution)

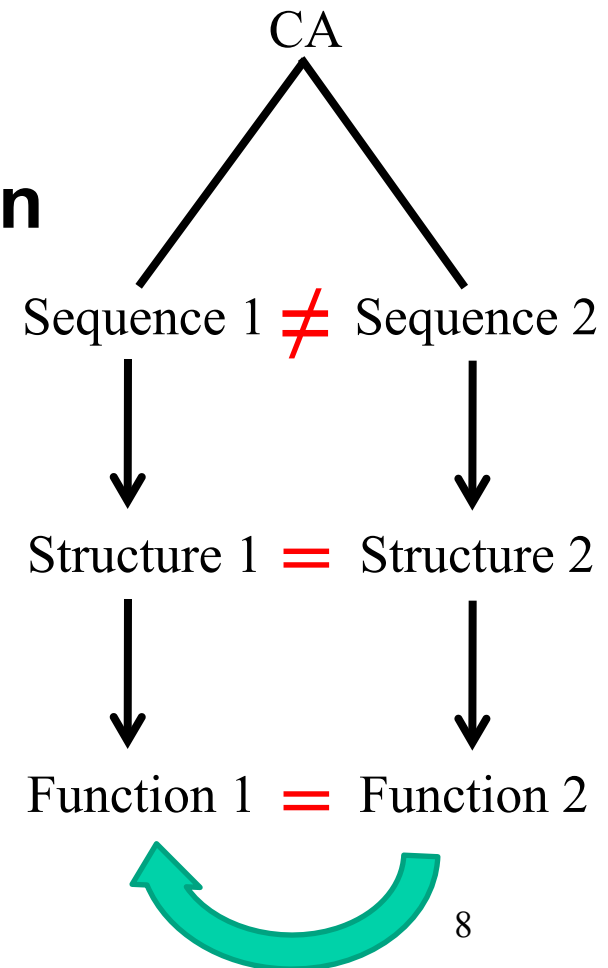
Dean, A. M. and G. B. Golding: *Pacific Symposium on Bioinformatics 2000*



Can we just transfer information about structure and/or function?

- **Structure (and function) more conserved than sequence**
- **Sequence → structure → function**

- So, if the sequences already tell us it's the same thing (homolog), then certainly the structures and functions are supposed to be the same.
- This works most of the time, but there are cases where likely homology does not bear out.

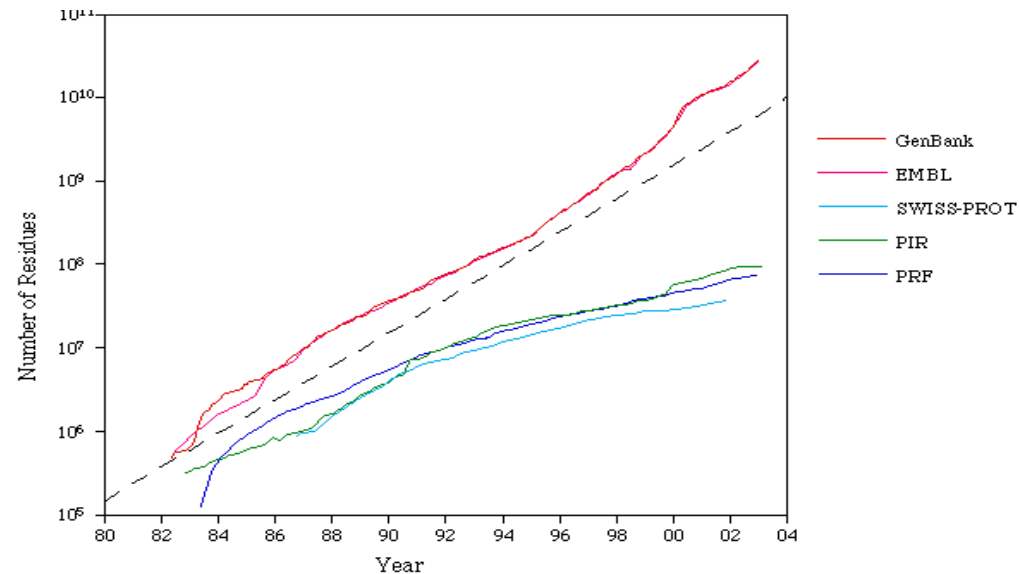
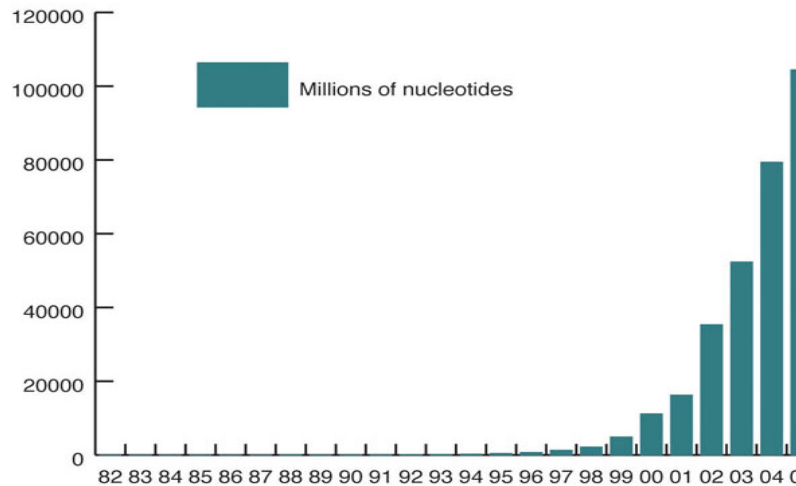


What function does your gene have

- We are going to use the homology principle
- We are going to seriously search through sequence databases
 - Non-redundant (NR) database > 7 million sequences
 - Each and every sequence should be considered

Sequence searching - challenges

- Exponential growth of databases



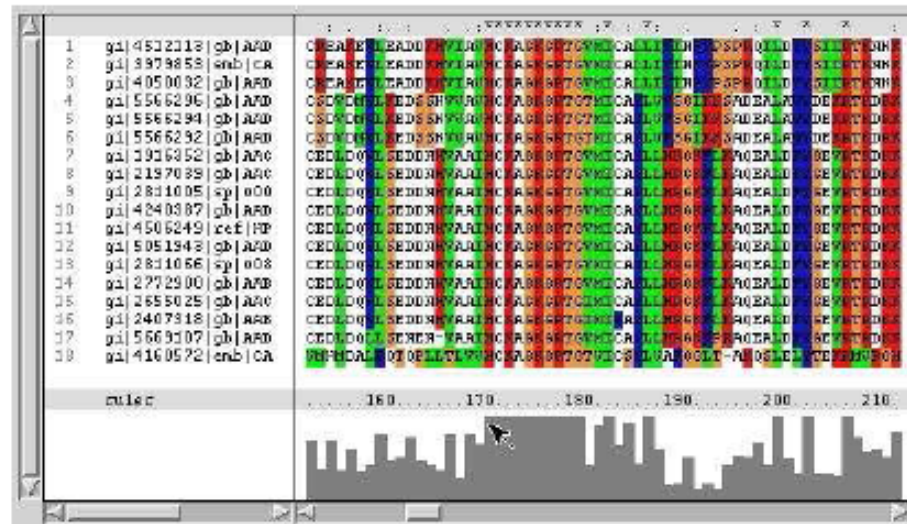
*Log function: Straight line
implies exponential growth*

Bioinformatics justification

- ***“Mind the Gap”***
- There are far more sequence data than structural/functional data
- We need to fill this gap by analysis and prediction pipelines



Alignments are useful ...



Conserved
patterns

Evolutionary
analysis

Structure
prediction

Motifs

Function
prediction

PRALINE web-interface

PRALINE multiple sequence alignment



[Advanced Interface](#)

[Options Help](#)

[PRALINE sample output](#)

[References and FAQs](#)

PRALINE is a multiple sequence alignment program with many options to optimise the information for each of the input sequences; e.g. global or local preprocessing, predicted secondary structure information and iteration capabilities.

Paste in your sequences in FASTA format (MAX 500 sequences, length 2000):

Or Upload a FASTA file (MAX 500 sequences, length 2000):

Enter a name for your job

Options

Exchange weights matrix: **BLOSUM62** [Help](#) Associated gap penalties: Open Extension [Help](#)

Global progressive alignment strategy: [Help](#)

- ☐ Standard progressive strategy
- ☐ Pre-profile global processing Iterations Score Cutt-off
- ☐ Pre-profile local processing Iterations Score Cutt-off
- ☒ PSI-BLAST pre-profile processing (Homology-extended alignment) (new option)
- PSI-BLAST Iterations Start e-value Cutt-off at DB

Secondary structure prediction

 [Help](#)

DSSP-defined secondary structure search

☐ YES ☒ NO [Help](#)

Tree representation of the final alignment

☐ YES ☒ NO [Help](#)

Customize alignment representation colours

☐ YES ☒ NO [Help](#)

Final alignment file format

☐ NO FILE ☒ MSF ☐ FASTA [Help](#)

E-mail

If you would like to be notified when your job has completed, please **tick the box below** and enter the e-mail address the notification should be sent to:

☐ I want to be notified when my job is done at

Submit

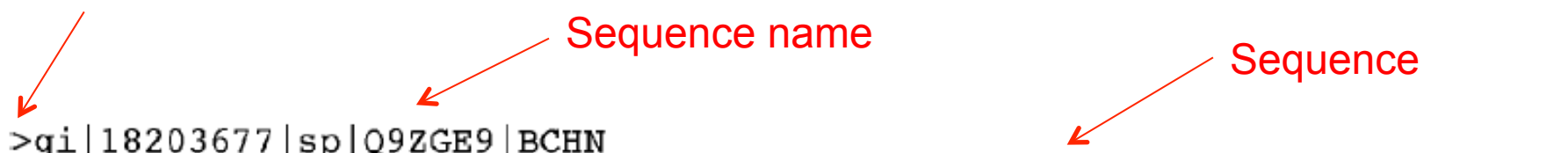
Frequently used (input) format to describe protein sequences:

Fasta Format

Sequence start
indicator '>'

Sequence name

Sequence



```
>gi|18203677|sp|Q9ZGE9|BCHN
MERVERENGCFHTFCPIASVAWLHRKIKDSFFLIVGTHHTCAHFIQTALDVMVYAHSRFGFAVLEESDLVS
ASPTEELGKVVQQVVDEWHPKVIFVLSTCSVDILKMDLEVSCCKDLSTRFGFPVLPASTSGIDRSFTQGED
AVLHALLPFVPKEAPAVEPVEEKKPRWFSFGKESEKEKAEPARNLVLIGAVTDSTIQQLQWELKQLGLPK
VDVFPDGDIRKMPVINEQTVVVPLQPYLNDTLATIRRERRAKVLSTVFPDGTARFLEAICLEFGLDT
SRIKEKEAQAWRDLEPQLQILRGKKIMFLGDNLLELPLARFLTSCDVQVVEAGTPYIHSKDLQOELELLK
ERDVRIVESPDFTKQLQRMQEYKPDLLVAGLGICNPLEAMGFTTAWSIETFAQIHGFVNAIDLKLFK
PLLKRQALMEHGWAEAGWLE
```

Fasta files can contain many sequences starting with a '>' symbol. The '>' symbol each time signifies a new sequence.

A protein sequence alignment

```
MSTGAVLIY--TSILIKECHAMPAGNE-----  
---GGILLFHRTHEHATECHAMPAGNEGGSNNS  
      *      *      *      * * * * * * * * *
```

A DNA sequence alignment

```
attcgttggcaaatcgcccctatccgggccttaa  
att---tggcggatcg-cctctacggggcc----  
***      ****      *****      **      * * * * *
```

DISCLAIMER: Alignment should only be applied to (putative) homologous sequences!! All sequences are supposed to derive from a common ancestor. Ideally, an *orthologous* set of sequences gets aligned.

How many pair-wise alignments

T	D	W	V	T	A	L	K
T	D	W	L	-	-	I	K

Combinatorial explosion

- 1 gap in 1 sequence: $n+1$ possibilities
- 2 gaps in 1 sequence: $(n+1)n$
- 3 gaps in 1 sequence: $(n+1)n(n-1)$, etc.

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \sim \frac{2^{2n}}{\sqrt{\pi n}}$$

→ 2 sequences of 300 a.a.: $\sim 10^{88}$ alignments

2 sequences of 1000 a.a.: $\sim 10^{600}$ alignments!

Technique to overcome the alignment combinatorial explosion:

Dynamic Programming (DP)

- Break alignment problem up in smaller subproblems and solve these iteratively
- Alignment is simulated as a Markov process, all sequence positions are seen as independent and identically distributed (i.i.d).
- Chances of sequence events are independent
 - Therefore, probabilities per aligned position are multiplied
 - Amino acid matrices contain so-called log-odds values (\log_{10} of the probabilities), so probabilities can be summed [$\log(ab)=\log(a)+\log(b)$]

History of Dynamic Programming algorithm

1970 Needleman-Wunsch global pair-wise alignment

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J Mol Biol.* 48(3):443-53.

- Align sequences in their entirety

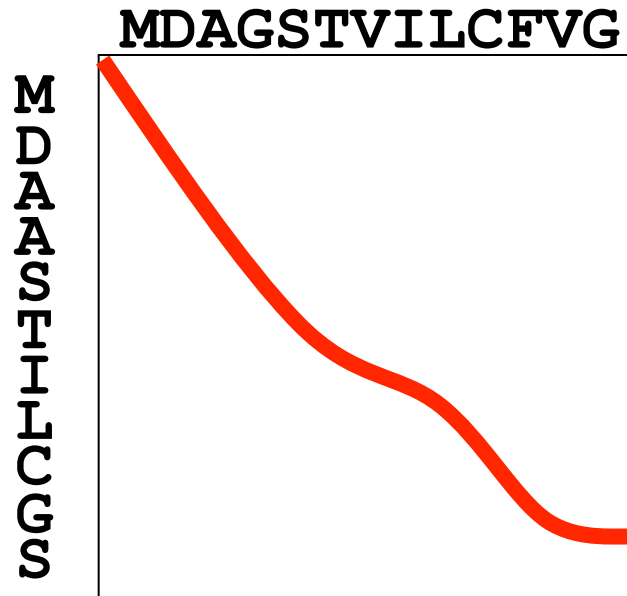
1981 Smith-Waterman local pair-wise alignment

Smith, TF, Waterman, MS (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.

- Only align subsequences with sufficient evolutionary memory (conservation)
- BLAST incorporates an heuristic version of local pairwise alignment

Pairwise sequence alignment

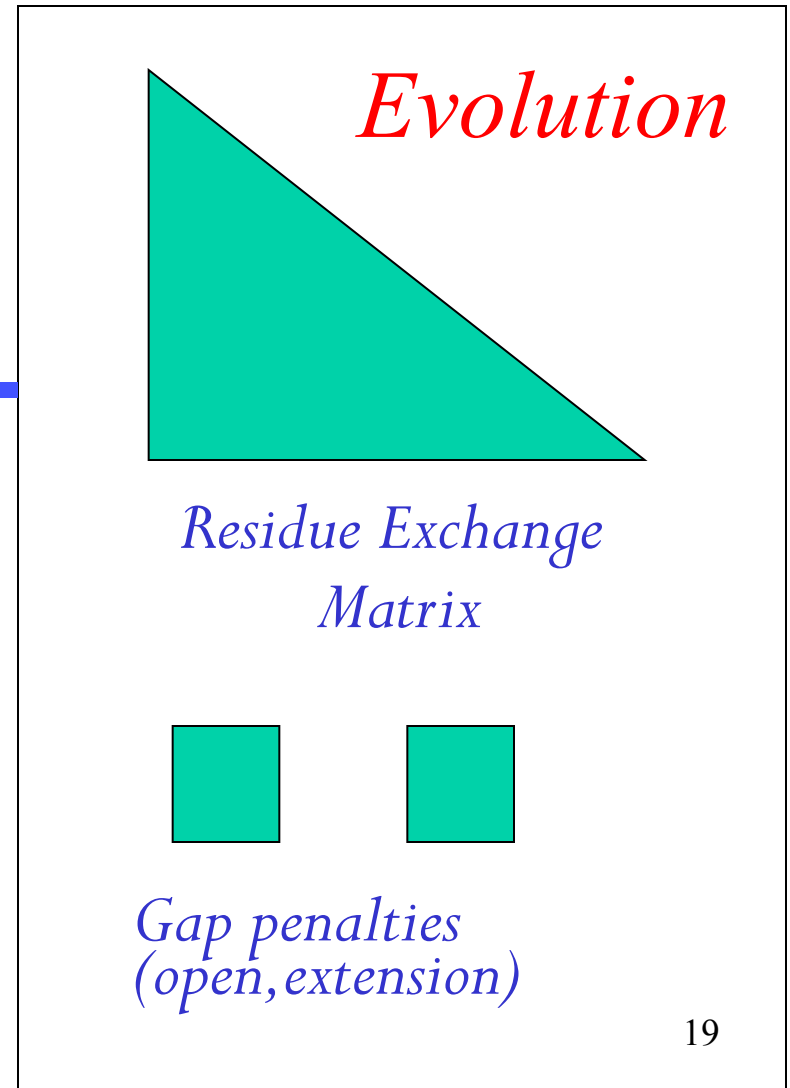
Global dynamic programming (DP)



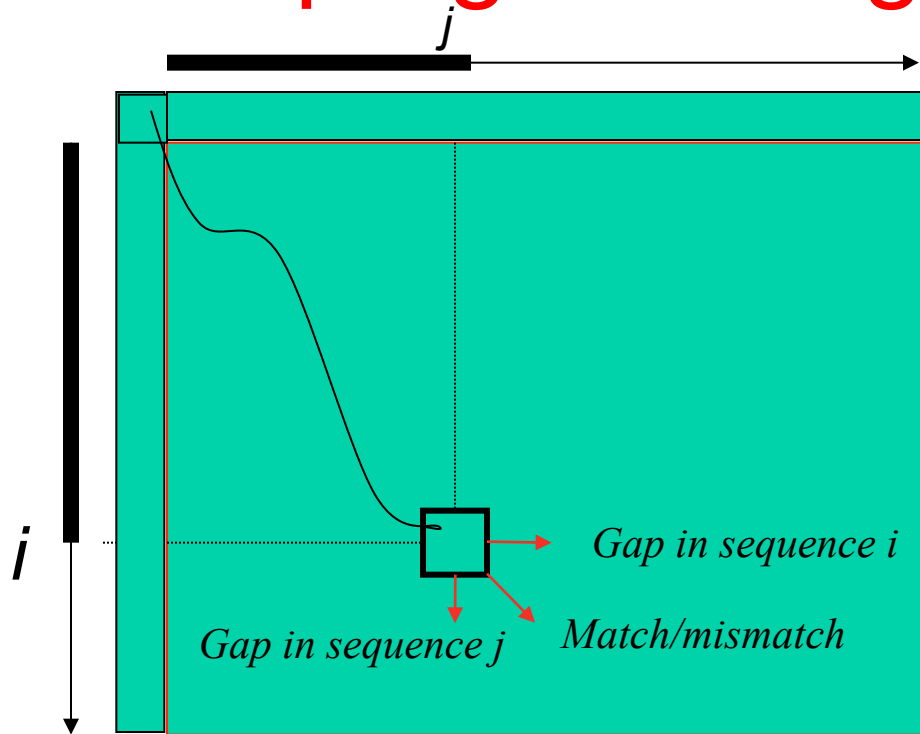
Search matrix



MDAGSTVILCFVG-
MDAAST-ILC--GS



Dynamic programming matrix

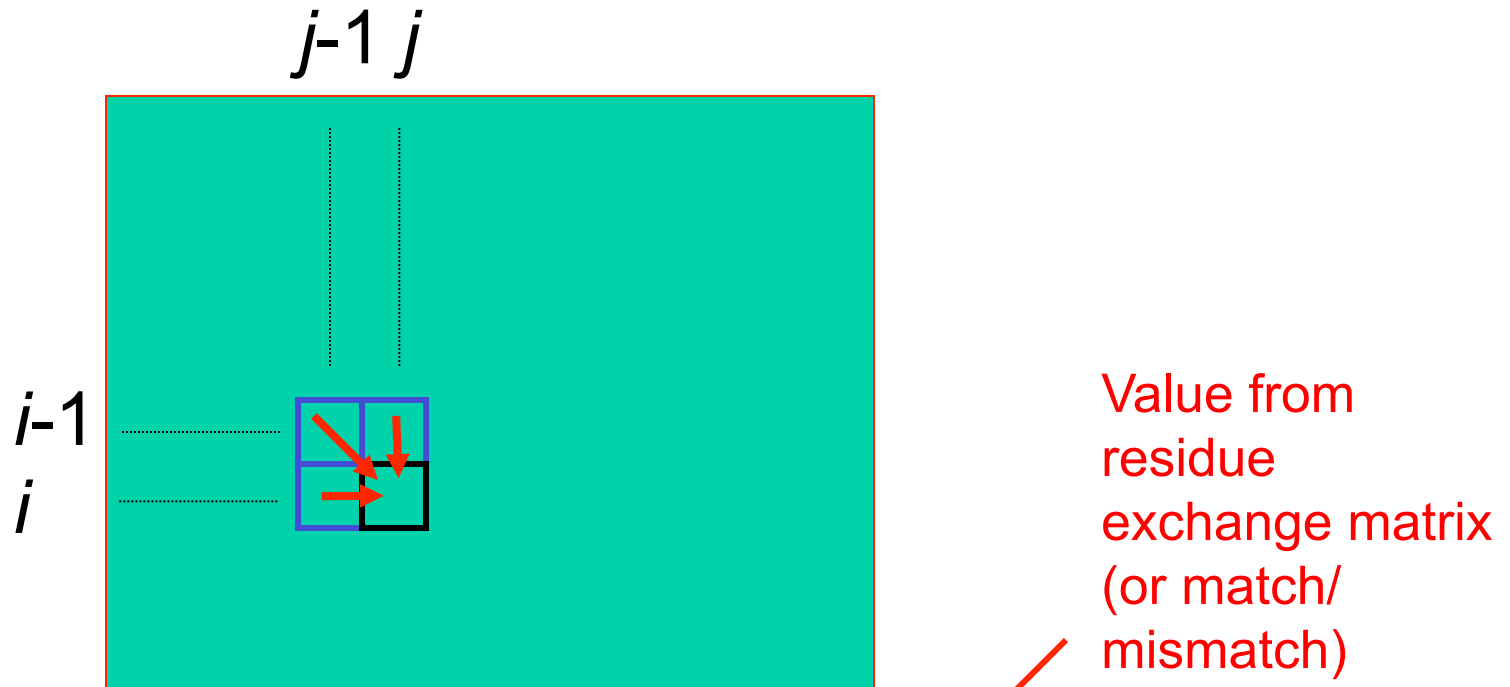


The cell $[i, j]$ contains the alignment score of the best scoring alignment of subsequence $1..i$ and $1..j$, that is, the subsequences up to $[i, j]$

Cell $[i, j]$ does not 'know' what that best scoring alignment is (it is one or a number of alternatives) out of very many possibilities, leading to $[i, j]$

By going through the matrix in row-wise fashion, each time extend alignment from cell $[i, j]$

Global dynamic programming



Value from
residue
exchange matrix
(or match/
mismatch)

$$H(i,j) = \text{Max} \begin{cases} H(i-1,j-1) + s(i,j) & \text{diagonal} \\ H(i-1,j) - g & \text{vertical} \\ H(i,j-1) - g & \text{horizontal} \end{cases}$$

*This is a recursive
formula*

Gap penalty

Substitution matrices for a.a.

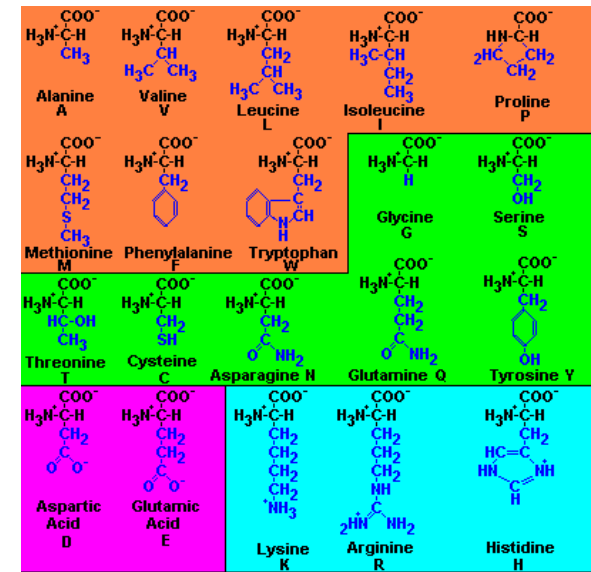
■ Amino acids are **not** equal:

1. Some are **similar** and easily substituted:

- biochemical properties
- structure

2. Some mutations occur more often due to **similar codons**

■ The two above give us **substitution matrices**



<http://www.people.virginia.edu/~rjh9u/aminacid.html>

orange: nonpolar and hydrophobic.
green: polar and hydrophilic
magenta box are acidic
light blue box are basic

		2nd base in codon				3rd base in codon
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	

<http://www.cimr.cam.ac.uk/links/codon.htm>

BLOSUM 62 substitution matrix

Henikoff & Henikoff, PNAS 89:10915; 1993

	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	

Positive values
-Preferred substitution

Negative values
-Avoided substitution

Zero values
-Randomly expected

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + S(x[i], y[j]) \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$


Gap Penalty

Substitution Matrices: DNA

define a score for match/mismatch of letters



Simple:

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


$$M[i, j] = \max \begin{cases} M[i-1, j-1] \pm 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$

Used in genome alignments:

	A	C	G	T
A	91	-114	-31	-123
C	-114	100	-125	-31
G	-31	-125	100	-114
T	-123	-31	-114	91


$$M[i, j] = \max \begin{cases} M[i-1, j-1] + S(x[i], y[j]) \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$


This is how the substitution scores are used

Example:

global alignment of two sequences

- Align two DNA sequences:
 - GAGTGA
 - GAGGCGA (note the length difference)
- Parameters of the algorithm:
 - Match: $\text{score}(A,A) = 1$
 - Mismatch: $\text{score}(A,T) = -1$
 - Gap: $g = 2$

$$M[i,j] = \max \begin{cases} M[i-1,j-1] \pm 1 \\ M[i,j-1] - 2 \\ M[i-1,j] - 2 \end{cases}$$

The algorithm. Step 1: init

- Create the matrix
- Initiation
 - 0 at [0,0]
 - Apply the equation...

$$M[i, j] = \max \begin{cases} M[i-1, j-1] \pm 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$

	j→	0	1	2	3	4	5	6
i↓		-	G	A	G	T	G	A
0	-	0						
1	G							
2	A							
3	G							
4	G							
5	C							
6	G							
7	A							

The algorithm. Step 1: init

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$



- Initiation of the matrix:
 - 0 at pos [0,0]
 - Fill in the first row using the “→” rule
 - Fill in the first column using “↓”

j

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2						
A	-4						
G	-6						
G	-8						
C	-10						
G	-12						
A	-14						

i

The algorithm. Step 2: fill in

$$M[i,j] = \max \begin{cases} M[i-1,j-1] + 1 \\ M[i,j-1] - 2 \\ M[i-1,j] - 2 \end{cases}$$



- Continue filling in the matrix, remembering from which cell the result comes (arrows)

j

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3			
A	-4	-1	2				
G	-6						
G	-8						
C	-10						
G	-12						
A	-14						

i

The algorithm. Step 2: fill in

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$



- We are done...
- Where's the result?

j

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3	-5	-7	-9
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	1	2	2	0
C	-10	-7	-4	-1	0	1	1
G	-12	-9	-6	-3	-2	1	0
A	-14	-11	-8	-5	-4	-1	2

i

The algorithm. Step 2: fill in

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$



- We are done...
- Where's the result?

The lowest-rightmost cell

j

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3	-5	-7	-9
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	1	2	2	0
C	-10	-7	-4	-1	0	1	1
G	-12	-9	-6	-3	-2	1	0
A	-14	-11	-8	-5	-4	-1	2

i

The algorithm. Step 3: trace-back

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \end{cases}$$

j

- Start at the last cell of the matrix
- Go against the direction of arrows
- Sometimes the value may be obtained from more than one cell (which one?)

i

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3	-5	-7	-9
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	1	2	2	0
C	-10	-7	-4	-1	0	1	1
G	-12	-9	-6	-3	-2	1	0
A	-14	-11	-8	-5	-4	-1	2

Trace-back: follow arrows back -- with vertical/horizontal arrows the aligned cell is at base of arrow (top/leftmost point)

The algorithm. Step 3: trace-back

High road and low road

- Extract the alignments

a) *high road*

GAGT–GA

GAGGCGA

b) *low road*

GA–GTGA

GAGGCGA

- c) *‘middle road’*

GAG–TGA

GAGGCGA

j

	-	G	A	G	T	G	A
-	-2	-4	-6	-8	-10	-12	
G	-2	-1	-3	-5	-7	-9	
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	1	2	2	0
C	-10	-7	-4	-1	0	1	1
G	-12	-9	-6	-3	-2	1	0
A	-14	-11	-8	-5	-4	-1	2

i

The image displays a sequence alignment matrix (dynamic programming table) for aligning two sequences. The columns represent the sequence being aligned (top sequence), and the rows represent the reference sequence (bottom sequence). The sequences are: Top: -, G, A, G, T, G, A; Bottom: -, G, A, G, G, C, G, A. The matrix cells contain alignment scores. A red line indicates the optimal alignment path, starting from the top-left cell (score -2) and ending at the bottom-right cell (score 2). The path follows the sequence: (0,0) → (1,1) → (2,2) → (3,3) → (4,4) → (5,5) → (6,6) → (7,7). A blue 'a' is positioned above the first cell, and a blue 'b' is positioned to the left of the second cell. A red arrow points down from the cell (4,4) to the cell (5,4).

How can one change a ‘high road’ algorithm into a ‘low road’ one, or vice versa?

Complexity of the basic DP algorithm

- The **time complexity** is N^2 (speed of the algorithm)
- The **memory complexity** is N^2 (amount of memory needed)

Time complexity becomes a problem when large numbers of sequences should be aligned, such as in aligning a sequence against a database of sequences

Why can we solve this in $O(NM)$?

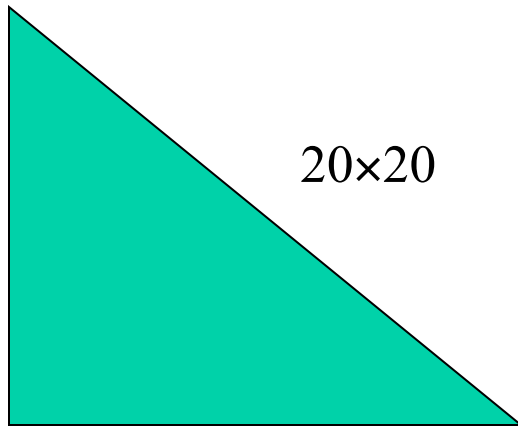
- Dynamic Programming:
solve sub problems, and combine them

	-	G	A	G	T	G	A
-	-	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3	-5	-7	-9
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	2	2	0	
C	-10	-7	-4	0	1	1	
G	-12	-9	-6	-2	-1	0	
A	-14	-11	-8	-4	-1	2	

Changing alignment here, will not alter earlier alignment steps

Alignment input parameters

Scoring alignments



20x20

Amino Acid Exchange Matrix

10

1

*Gap penalties (open,
extension)*

A number of different schemes have been developed to compile residue exchange matrices

However, there are no formal concepts to calculate corresponding gap penalties

Empirically determined values are therefore recommended; e.g. PAM250, BLOSUM62, etc.

Dynamic programming

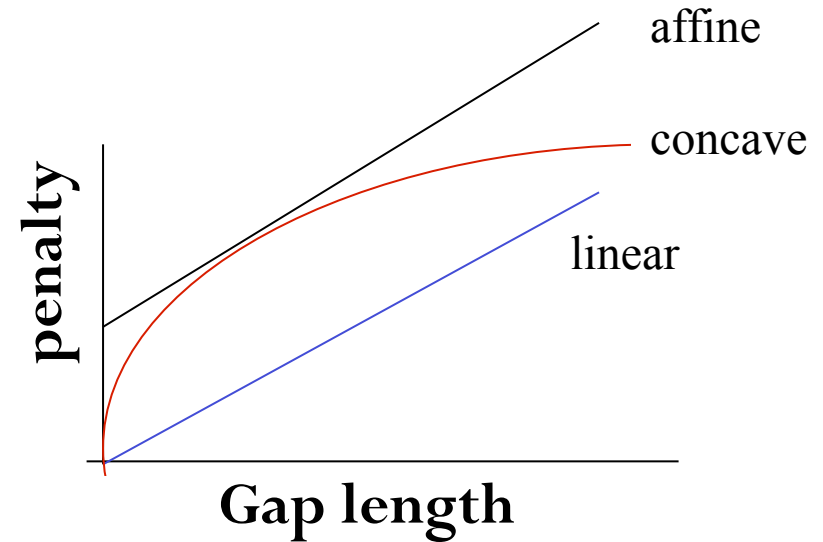
Scoring alignments

– Substitution (or match/mismatch) scores

- DNA
- proteins

– Gap penalty

- Linear: $gp(k)=ak$
- Affine: $gp(k)=b+ak$
- Concave, e.g.: $gp(k)=\log(k)$



The score for an alignment is the sum of the scores of all alignment columns (including gaps)

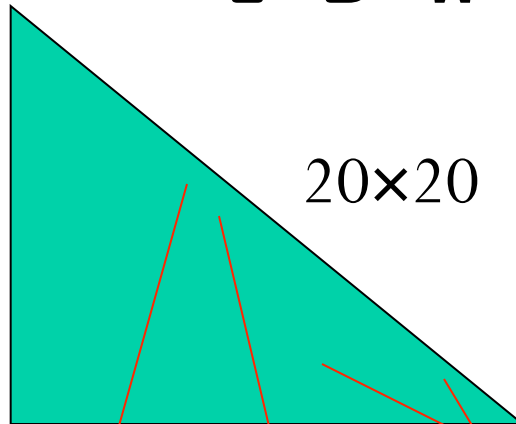
General alignment score:

$$S_{a,b} = \sum_i s(a_i, b_j) - \sum_k N_k \cdot gp(k)$$

Dynamic programming

Scoring alignments

T	D	W	V	T	A	L	K
T	D	W	L	-	-	I	K



Amino Acid Exchange Matrix

10

1

Affine gap penalties (open, extension)

Score: $s(T,T) + s(D,D) + s(W,W) + s(V,L) - P_o - 2P_x +$
 $+ s(L,I) + s(K,K)$

Take home

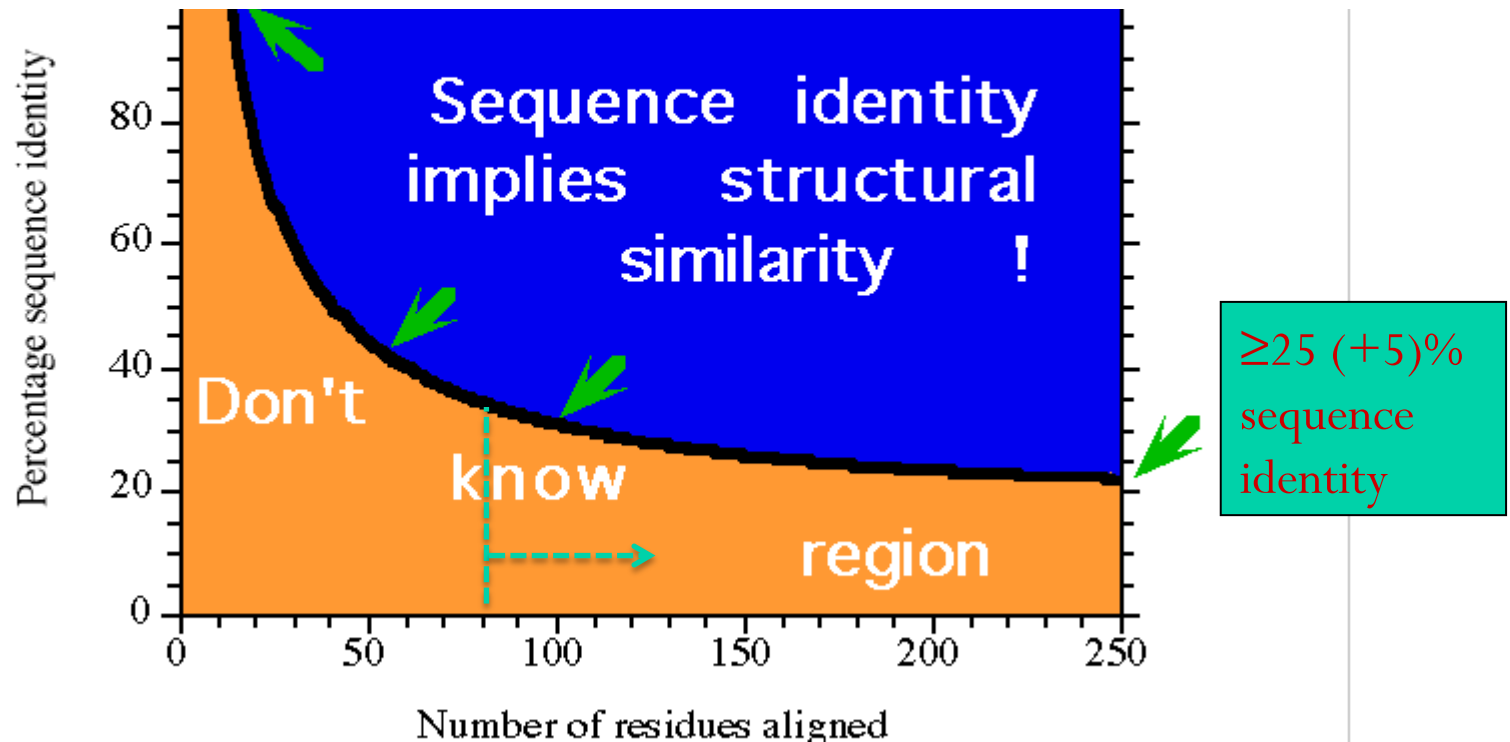
- Alignments represent divergent evolution with columns 'representing' a position in the common ancestral sequence
- Dynamic programming (DP) is the main technique to perform alignments
 - It balances match and mismatch scores against gap penalties to obtain the optimal (or highest scoring) alignment
 - The input to the DP algorithm is an evolutionary model comprising residue exchange scores and gap-penalties.
 - Many different optimal alignments may exist, but only one is typically shown

Measuring Sequence Similarity using an alignment

- **Sequence identity** (number of identical exchanges per unit length)
- **Raw alignment score** (using mutation probabilities as in the previous examples)
- **Sequence similarity** (alignment score normalised to a maximum possible)
- **Alignment score normalised** to a randomly expected situation (database/homology searching)

What can sequence alignment tell us about structure

HSSP Sander & Schneider 1991



The non-linear curve is interpreted in practice as a straight horizontal line for alignments of length 80 or more. This means that a general cutoff value for inferring homology is 25% sequence identity (this is a crude measure). So, sequence alignments with >30% SeqID (5% safety margin over 25%) are “in the blue” and imply homology (and hence structural similarity)

Outline – PART I

- DP Needleman-Wunsch + example
 - handout (15 mins)
- Number of possible alignments -> complexity
- Dynamic Programming – show intuitive proof
- Global / local alignment
- Project picture

Now practise:

Finding a global alignment

- Sequences:
- 1) TGATT
- 2) TGGAGT
- What else do you need?

Global Alignment Exercise ***

$$H(i,j) = \text{Max} \begin{cases} H(i-1,j-1) + S(i,j) & \nwarrow \\ H(i-1,j) - g & \uparrow \\ H(i,j-1) - g & \leftarrow \end{cases}$$

$$S(i,j) = \begin{cases} 1 & \text{for match} \\ -1 & \text{for mismatch} \end{cases}$$

$$g = 2$$

H(i,j) matrix j →

	-	T	G	A	T	T
-	0	-2				-10
T	-2	1				-7
G						
G						-3
A						-2
G	-10	-7				
T	-12	-9			0	

traceback matrix

	-	T	G	A	T	T
-	.	↖	↖	.	.	↖
T	↑	↖	.	.	.	↖↖
G	↑	↖
G	↑	.	.	.	↖↖	↖↖
A	↖↖
G
T	↑	↖↑	↑	↑	.	.

Output

- Alignment: T-GATT
 TGGAGT

OR ???????

- Score: 1

Time and memory complexity of DP

- The *time complexity* is **$O(n^2)$** : for aligning two sequences of n residues, you need to perform n^2 algorithmic steps (square search matrix has n^2 cells that need to be filled)
- The *memory (space) complexity* is also **$O(n^2)$** : for aligning two sequences of n residues, you need a square search matrix of n by n containing n^2 cells

Domains - example

Immunoglobulin domain

Representative ig domain proteins

[1A01_GORGO](#) [Gorilla gorilla gorilla (lowland gorilla)] class i histocompatibility antigen, gogo-a0101 alpha chain precursor



[ALC1_GORGO](#) [Gorilla gorilla gorilla (lowland gorilla)] ig alpha-1 chain c region



[AMAL_DROME](#) [Drosophila melanogaster (fruit fly)] amalgam protein precursor



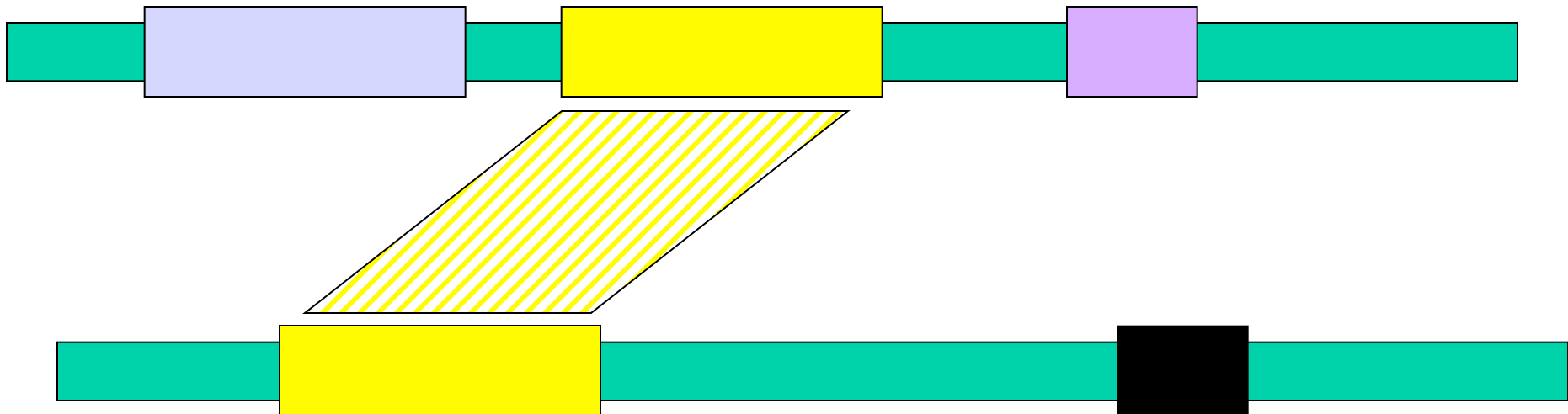
[B2MG_BOVIN](#) [Bos taurus (bovine)] beta-2-microglobulin precursor (lactollin)



What would you like to align between these sequences?

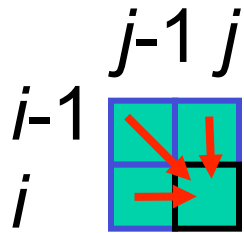
Local pairwise alignment

- Smith-Waterman local alignments are locally optimal segments from either sequence, such that their alignment score is optimal over all possible local alignments
- Multiple local alignments are possible, but only the optimal (highest scoring) one is selected using the Smith-Waterman algorithm



Local dynamic programming (Smith and Waterman, 1981)

basic algorithm



$$H(i,j) = \text{Max} \left\{ \begin{array}{ll} H(i-1,j-1) + S(i,j) & \text{diagonal} \\ H(i-1, j) - g & \text{vertical} \\ H(i, j-1) - g & \text{horizontal} \\ \mathbf{0} & \end{array} \right.$$

The algorithm. Step 2: fill in

- No initialising row/column
- Find the highest cell anywhere in the matrix
- Trace back from highest cell to '0' cell (not including) or beginning of sequence

$$M[i, j] = \max \begin{cases} M[i-1, j-1] \pm 1 \\ M[i, j-1] - 2 \\ M[i-1, j] - 2 \\ 0 \end{cases}$$

	$j \rightarrow$	1	2	3	4	5	6
$i \downarrow$		G	A	G	T	G	A
1	G	1	0	1	0	1	0
2	A	0	2	0	0	0	2
3	G	1	0	3	1	1	0
4	G	1	0	1	2	2	0
5	C	0	0	0	0	1	1
6	G	1	0	1	0	1	0
7	A	0	2	0	0	0	2

Project Overview

Running BLAST and PSI-BLAST

Sequence Database

all against all

BLAST

PSI -BLAST

reference database

GO

PFAM

SCOP

Which is better: BLAST or
PSI-BLAST?

The BLAST suite

- ❑ Computer program for homology searching
- ❑ BLAST is a fast heuristic local alignment tool
- ❑ Given a protein query sequence (for which the function is unknown), the program searches through a non-redundant sequence database (NR) of >7 million sequences
- ❑ BLAST aligns a given query sequence with each database sequence and calculates similarity
- ❑ If sequence similarity is high enough (low BLAST e-value), the query and database sequence are deemed homologous, so that the database sequence's function (if known) can be transferred to the query.

The BLAST suite (2)

- ❑ BLAST is based upon an intricate and very fast algorithm to search through the NR database.
 - It operates under the assumption that the vast majority of the DB sequences can be easily excluded as putative homologs (sequences that are clearly too different for safely assuming homology)

BLAST entry page

Protein BLAST: search protein databases using a protein query - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRA... blast

Getting Started Latest Headlines

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help My NCBI Sign In Register

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#) Query subrange [From](#) [To](#)

Or, upload file [Browse...](#)

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database

Organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query

Enter an Entrez query to limit search

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)

☐ Show results in a new window

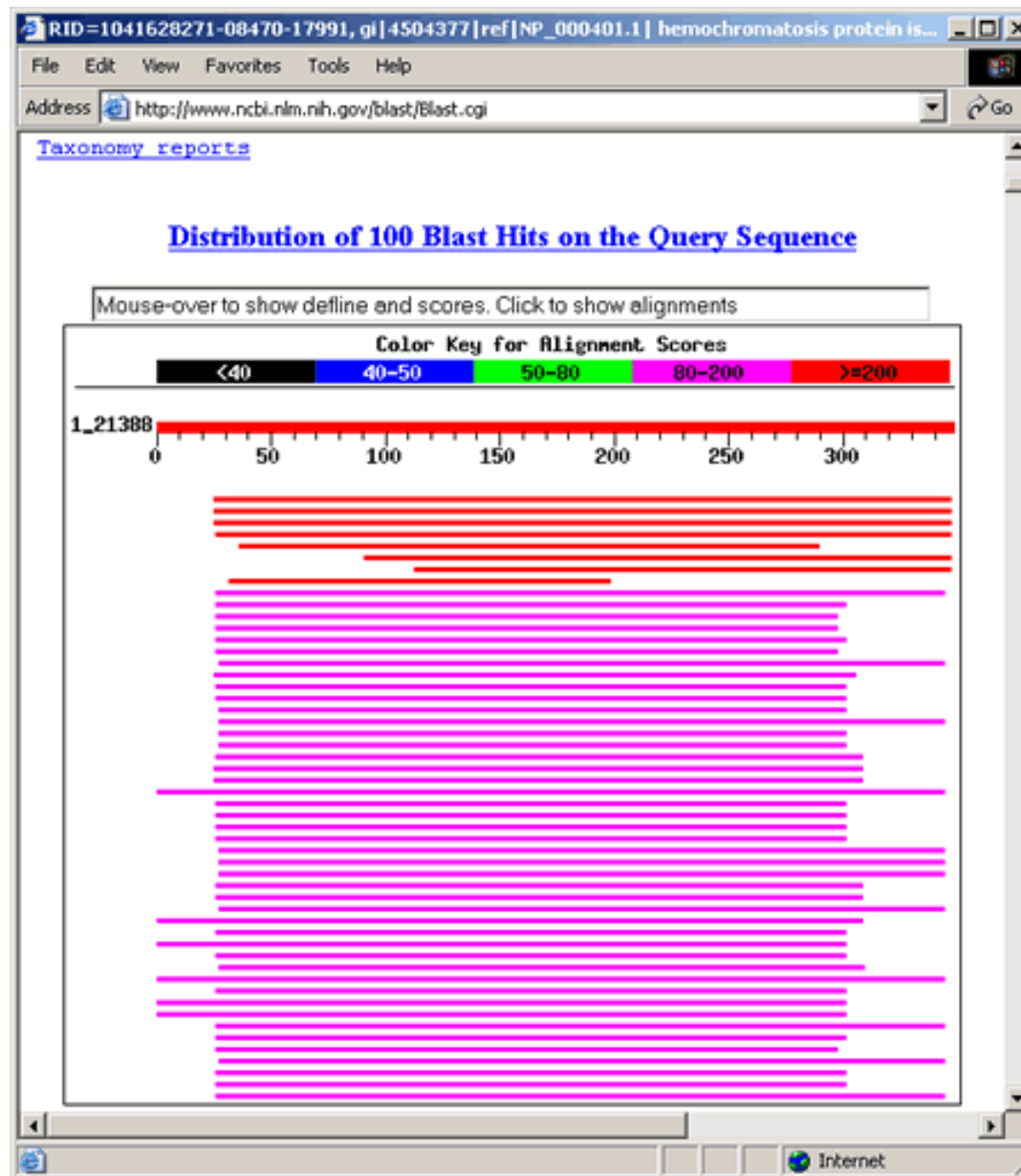
Algorithm parameters

Done

2 Windows Ex... Microsoft Powe... Protein BLAST: ... heringa@flits:/... ref - heringa@fl... NL 0:36

Paste your
query
sequence

Choose the
BLAST
program you
want



Related Structures

Sequences producing significant alignments:

			2	3	
			Score (bits)	E Value	
1	gi 6754190 ref NP_034554.1 	hemochromatosis [Mus musculus] ...	419	e-117	L - 4
	gi 26354116 dbj BAC40688.1 	unnamed protein product [Mus mu...	419	e-117	
	gi 12844463 dbj BAB26373.1 	unnamed protein product [Mus mu...	419	e-117	L
	gi 25742831 ref NP_445753.1 	hemochromatosis [Rattus norveg...	412	e-115	L
	gi 2624957 gb AAB86597.1 	hereditary hemochromatosis protei...	366	e-101	L
	gi 2072657 emb CAA73197.1 	HFE (HLA-H) [Mus musculus]	345	7e-95	L
	gi 5734363 gb AAD49965.1 AF176534.1	hemochromatosis gene pr...	303	2e-82	L
	gi 1930010 gb AAB51504.1 	hereditary haemochromatosis prote...	247	1e-65	L
	gi 2225995 emb CAA74333.1 	MHC class I alpha chain [Rattus ...	173	4e-43	L
	gi 2851391 sp P16391 HA12_RAT	RT1 class I histocompatibilit...	171	1e-42	L

- 1 - This portion of each description links to the sequence record for a particular hit.
- 2 - Score or bit score is a value calculated from the number of gaps and substitutions associated with each aligned sequence. The higher the score, the more significant the alignment. Each score links to the corresponding pairwise alignment between query sequence and hit sequence (also referred to as subject or target sequence).
- 3 - E Value (Expect Value) describes the likelihood that a sequence with a similar score will occur in the database by chance. The smaller the E Value, the more significant the alignment. For example, the first alignment has a very low E value of e^{-117} meaning that a sequence with a similar score is very unlikely to occur simply by chance.
- 4 - These links provide the user with direct access from BLAST results to related entries in other databases. 'L' links to LocusLink records and 'S' links to structure records in NCBI's Molecular Modeling DataBase.

```

RID=1041628271-08470-17991, gi|4504377|ref|NP_000401.1| hemochromatosis protein isoform 1 precu - Micro...
File Edit View Favorites Tools Help
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#6754190 Go

>gi|6754190|ref|NP_034554.1| L hemochromatosis [Mus musculus]
gi|3219802|sp|P70387|HFE_MOUSE Hereditary hemochromatosis protein homolog precursor
gi|7439228|pir|JC5382 hereditary hemochromatosis protein precursor - mouse
gi|1519485|gb|AAB07525.1| L hereditary haemochromatosis homolog [Mus musculus]
gi|2897948|gb|AAC03447.1| L HFE [Mus musculus]
Length = 359

Score = 419 bits (1078), Expect = e-117
Identities = 204/331 (61%), Positives = 236/331 (71%), Gaps = 9/331 (2%)

Query: 26 RSHSLHYLFMGASEQDLGLSLFEALGYVDDQLFVYFDHESRRVEPRTPWVSSRIXXXXXX 85
RSHSL YLFMGASE DLGL LFEA GYVDDQLFV Y+HESRR EPR PW+ +
Sbjct: 30 RSHSLRYLFMGASEPDLGLPLFEARGYVDDQLFVSYNHESRRAEPRAPWILEQTSSQLWL 89

Query: 86 XXXXXXKQWDHMFVDFWTFIMENHNHNSK-----ESHTLQVILGCEMQEDNSTEGYWK 137
KQWD+MF VDFWTFIM N+NHNSK ESH LQV+LGCE+ EDNST G+W+
Sbjct: 90 HLSQSLKQWDYMFIVDFWTFINGNYNHNSKVTKLGVVSESHILQVVLGCEVHEDNSTSGFWR 149

Query: 138 YGYDGGDHLEFCPDTLDWRAAEPRAPWPTKLEWERHKIRARQNRAYLERDCPAQLQOLLEL 197
YGYDGGDHLEFCP TL+W AAEP AW TK+EW+ HKIRA+QNR YLE+DCP QL++LLEL
Sbjct: 150 YGYDGGDHLEFCPRTLNWSAEPGAWATKVEWDEHKIRAKQNRDYLEKDCPEQLKRLEL 209

Query: 198 GRGVLDQQVPPLXXXXXXXXXXXXXLRCLALNYYPNITHKWLKDKQPMDAKEFEPKDV 257
GRGVLD QQVP L LRC+AL+++PNITH+WLKD QP+DAK+ P+ VL
Sbjct: 210 GRGVLDQQVPPTLVKVRHWASTGTSRLCQALDFFPNITHRWLKDQPLDAKDVNPEKVL 269

Query: 258 PNGDGTYYQGWITLAVPPGEEQRYTCQVEHPGLDQPLIVIEWPSPSGTLVIGVISXXXXXX 317
PNGD TYQGW+TLAV PG+E R+TCQVEHPGLDQPL WEP S ++IG+IS
Sbjct: 270 PNGDGTYYQGWITLAVAPGDETRFTCQVEHPGLDQPLTASWEPLQSQAMIIIGIIS-GVTIC 328

Query: 318 XXXXXXXXXXXXRRKRGSGRGAMGHYVLAERE 348
RRK+ S G MG YVL + E
Sbjct: 329 AIFLVGILFLILRRKASGGTNGGYVLTDC 359

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&list_uids=6754190&do Internet

```

'X' residues denote low-complexity sequence fragments that are ignored

BLAST 'flavours'

- **blastp** compares an amino acid query sequence against a protein sequence database
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares the six-frame conceptual protein translation products of a nucleotide query sequence against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database translated in six reading frames
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Some homologous protein families are distant

- Homologous sequences can be closely or distantly related
 - Histone family (protecting DNA) is completely conserved from bacteria to human
 - Hemoglobin family (transporting oxygen through blood) can be 90% different between close organisms
 - Some distant homologous families have sequences that have alignment scores below random (homology cannot be identified statistically)
- How to find distantly related family members in a homology search?

Outline (2)

- BLAST
 - E-values in detail
 - Why use BLAST (homology, function, structure)
 - Question on databases specific searching (k-mer)
 - Word sizes (DNA vs protein)
- PSI-BLAST
 - Profiles
 - Key idea (+ question on databases)
 - Show Sequence Log⁵⁰o's

Why BLAST?

Finding homologous sequences

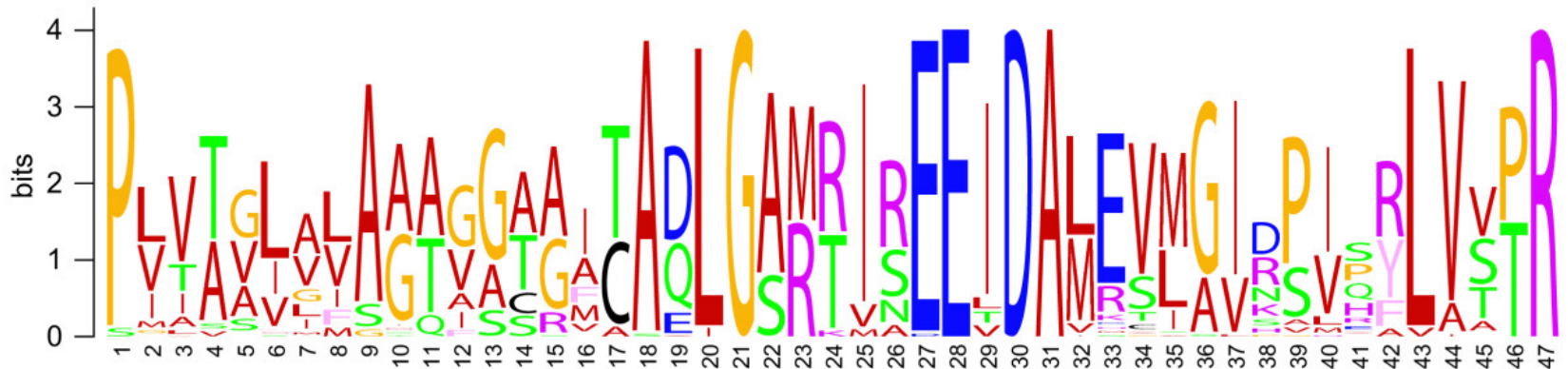
- **Homology**- similarity between sequences that result from a common ancestor.
- Sequence identity
- Use homology
- **Save time!** - exploit the knowledge you have about your homologues, and conclude about your query.

More than:
25% for proteins
70% for nucleotides
will be considered as indicating homology

Why BLAST?

Finding homologous sequences

Identify sequence motifs



*Sequence
logo*

Why BLAST?

Finding homologous sequences

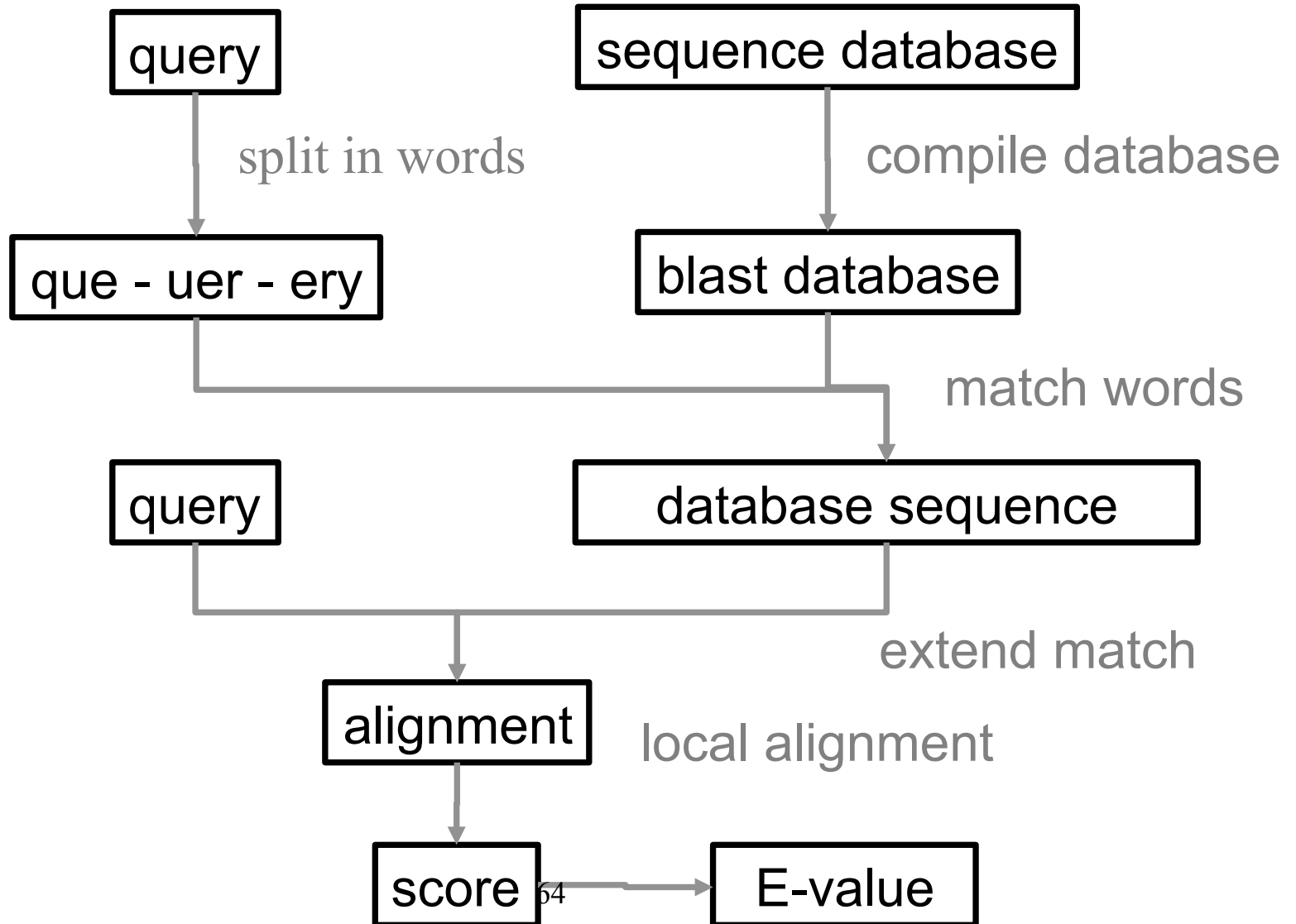
Find out which regions are evolutionary conserved
→ important for function and/or structure



Heuristic Alignment Motivation

- dynamic programming (DP) has performance $O(mn)$ which is too slow for large databases with high query traffic
 - Consider 7 M database sequences of 300 a.a. against query of 333 a.a. makes $7 * 10^{12}$ comparisons
 - Many searches per day
- heuristic methods do fast approximation to dynamic programming
 - FASTA [Pearson & Lipman, 1988]
 - **BLAST** [Altschul *et al.*, 1990]
 - **PSI-BLAST** [Altschul *et al.*, 1997]

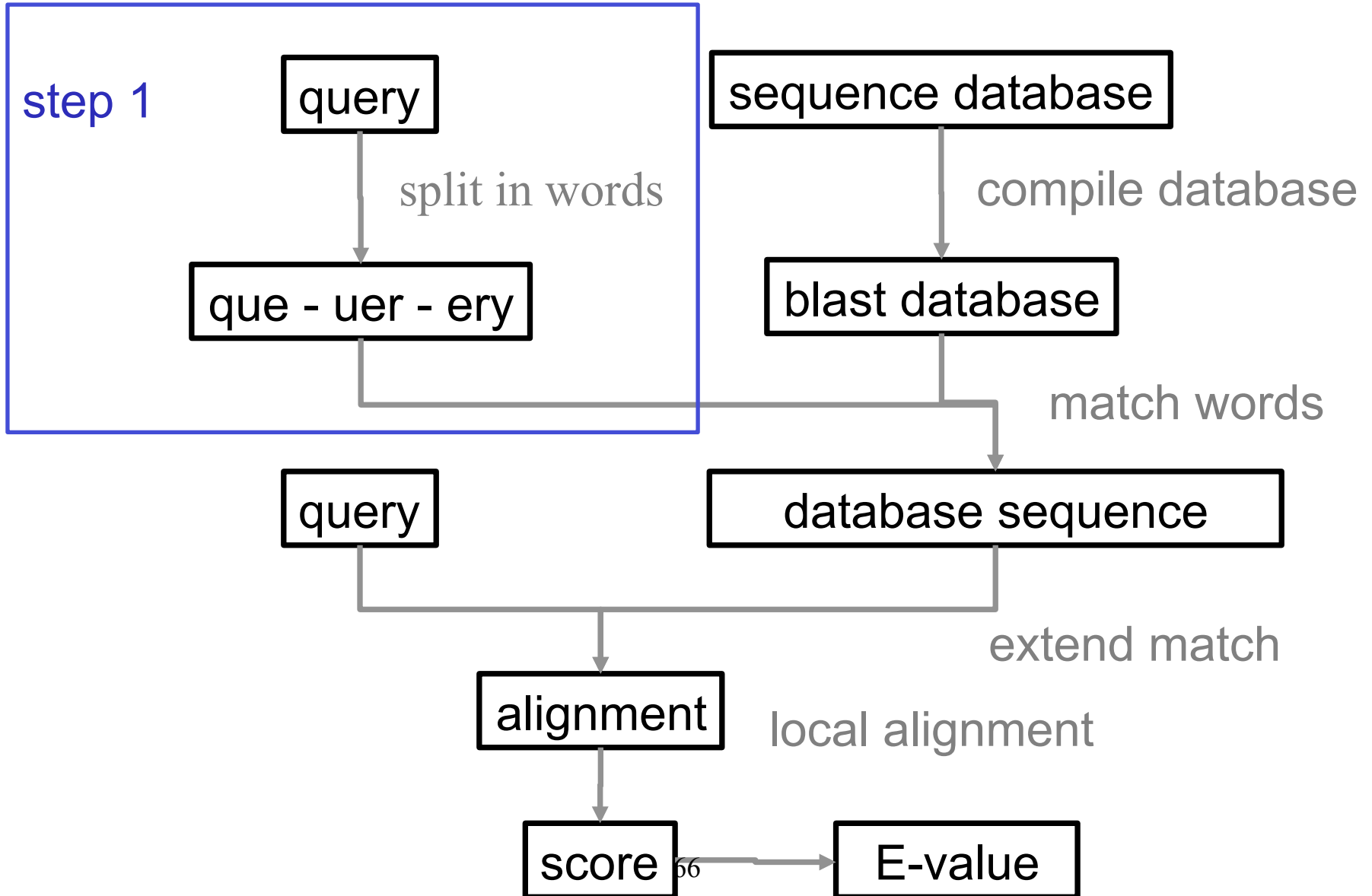
BLAST - overview



BLAST – overview in words

1. **Preprocess** the query sequence using protein 3-words (3-mers)
 - i. Make word lists using T-cutoff value
2. Use word lists to rapidly find gapless alignment regions (**diagonals**) between query and each DB sequence
3. **Two-hit method**: Try and find (small) diagonals on same matrix diagonal at a small-enough separation
 - i. If found: determine seed position and do two-directional local alignment (using DP – this is slow)
=> Report the alignment (putative homologous relationship)
 - ii. If not found: discard DB sequence as being unrelated (there is a risk involved that it might still be distantly related). Until this moment things have been very fast.

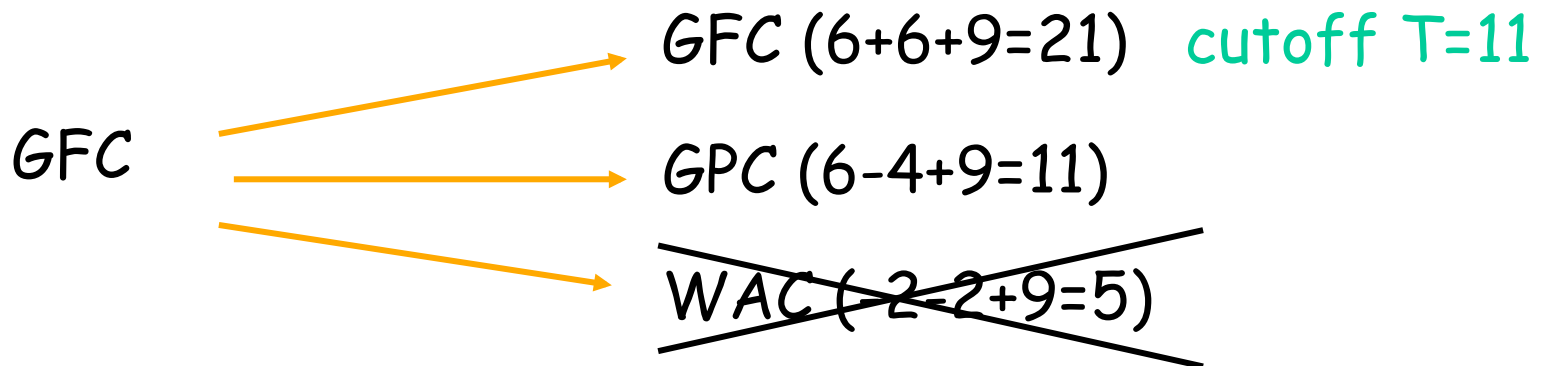
BLAST - overview



How does BLAST work?

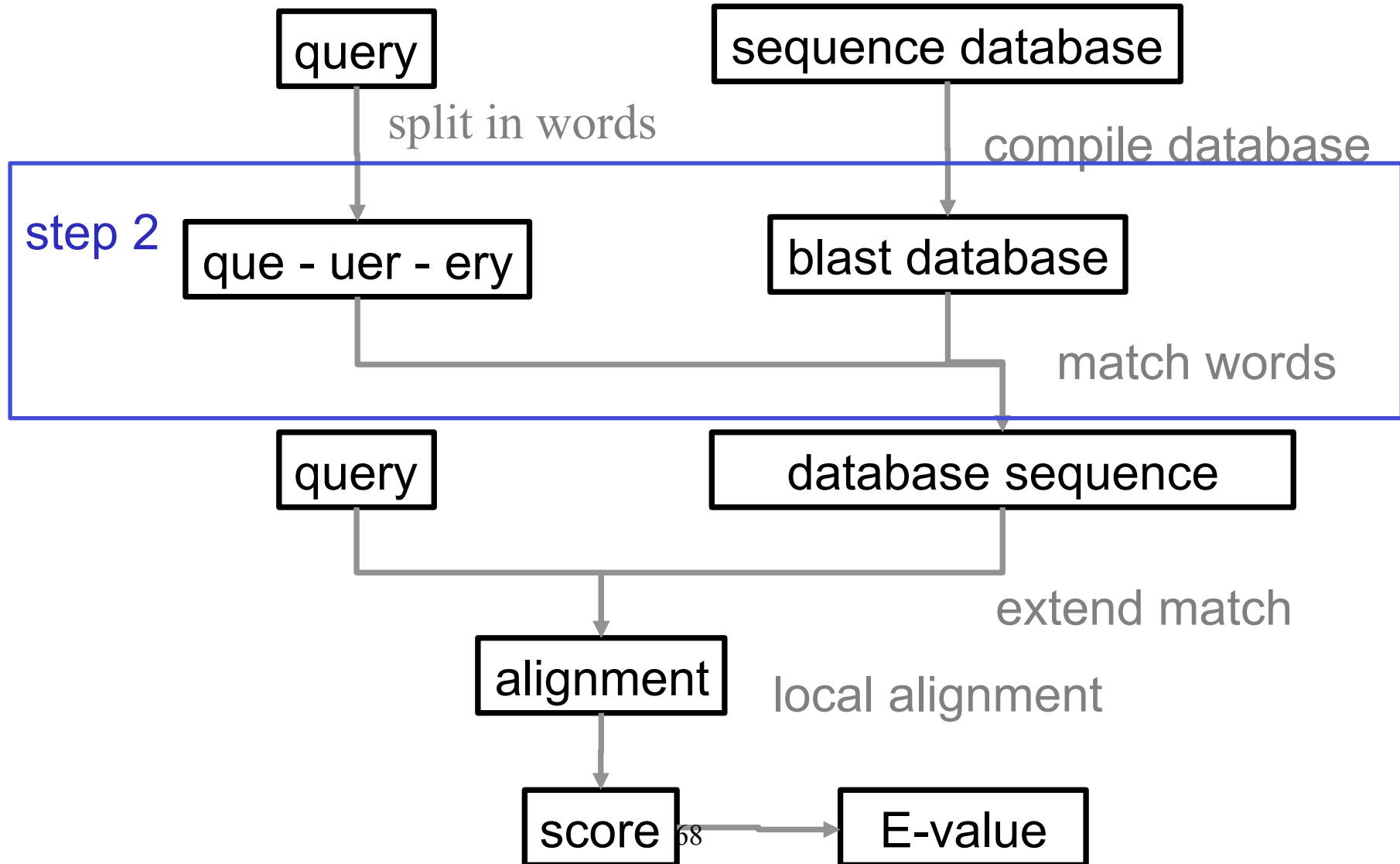
BLAST: Preprocessing the query sequence

- When the user enters a **query sequence**, it is divided into 3-words: GFCPLAV \rightarrow GFC FCP CPL PLA LAV
- For each 3-word, a **list of T-similar words** is defined according to a scoring matrix (e.g., BLOSUM62 for proteins) with a certain cutoff level

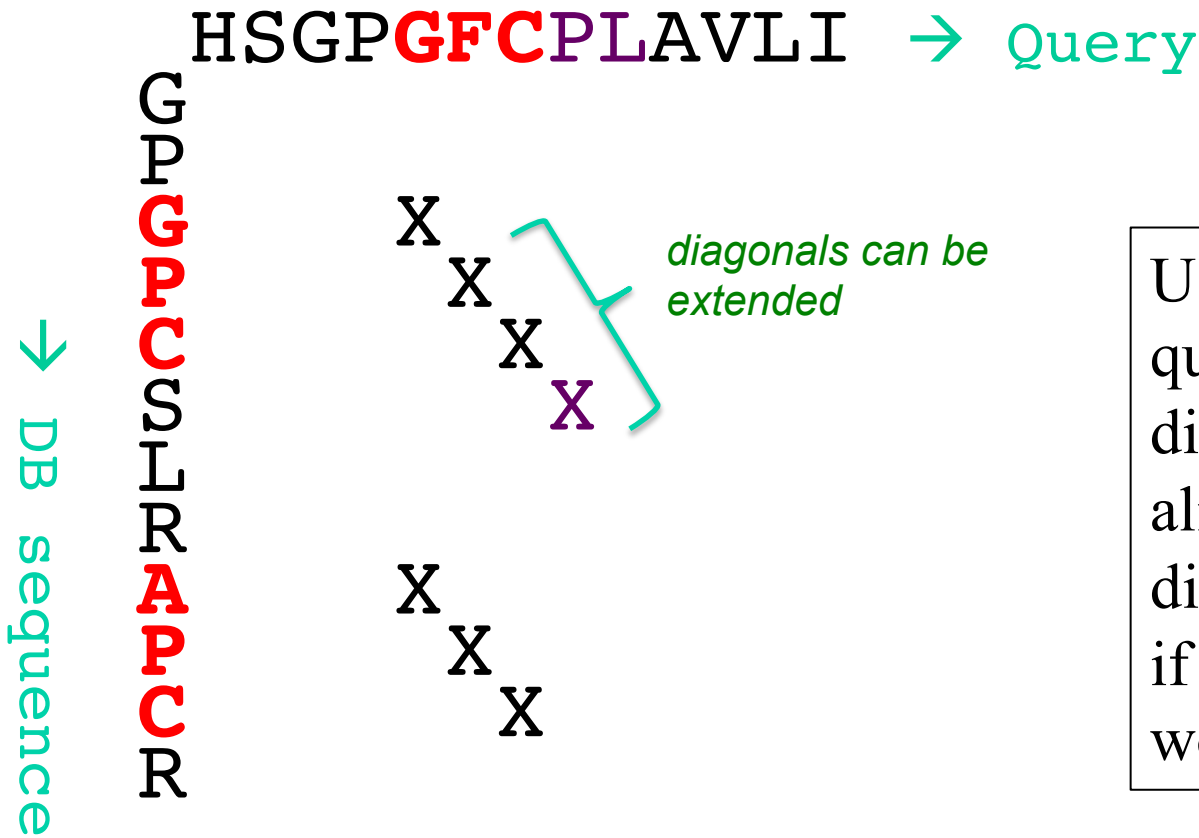


- For each 3-word, search the **database sequence** for consecutive neighboring words

BLAST - overview



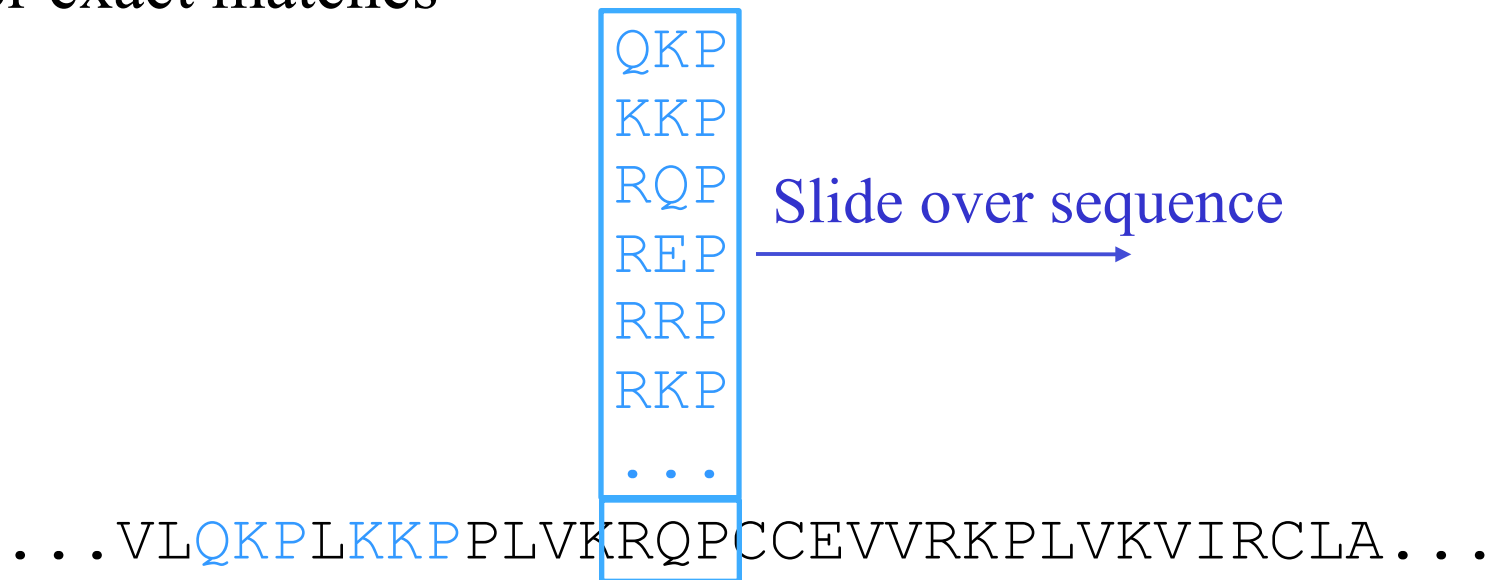
BLAST step 2: Make rapid gapless alignment



Use the word lists to quickly fill in diagonals (gapless alignments) – diagonals can grow if consecutive three-words are matched

BLAST, Step 2: Find “near-exact” matches with scanning

- Use all the T -similar k -words to build a Finite State Machine (to make scanning very fast)
- Scan for exact matches



Wrapping up BLAST so far

The parameters-

W : Word size - find W-mers in target/query
3 for aa, 11 for nucleotides.

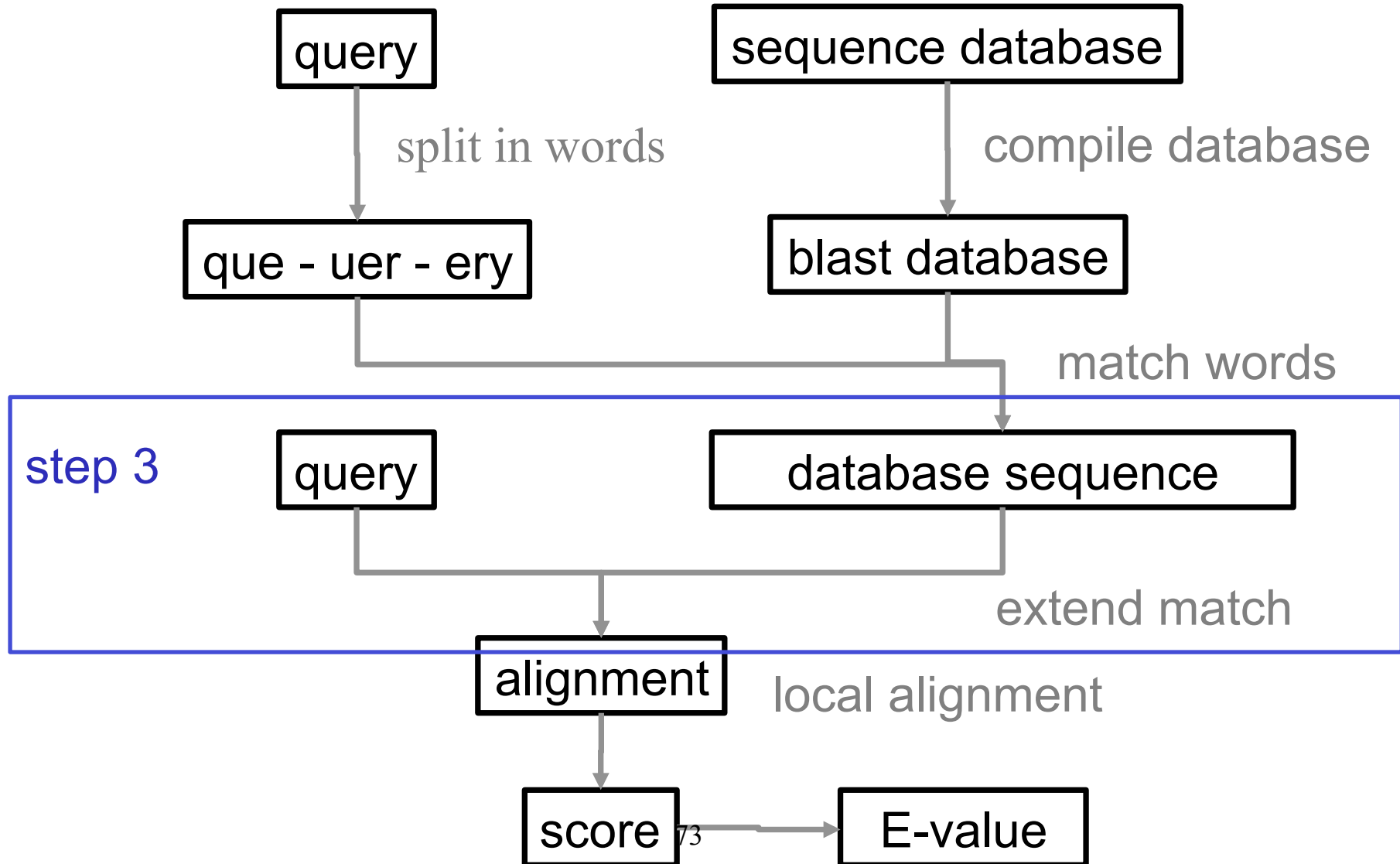
T : Threshold - focus on pairs scoring $>T$
usually 11-13 (T-similar words)

S : Score - the final score of segment pair

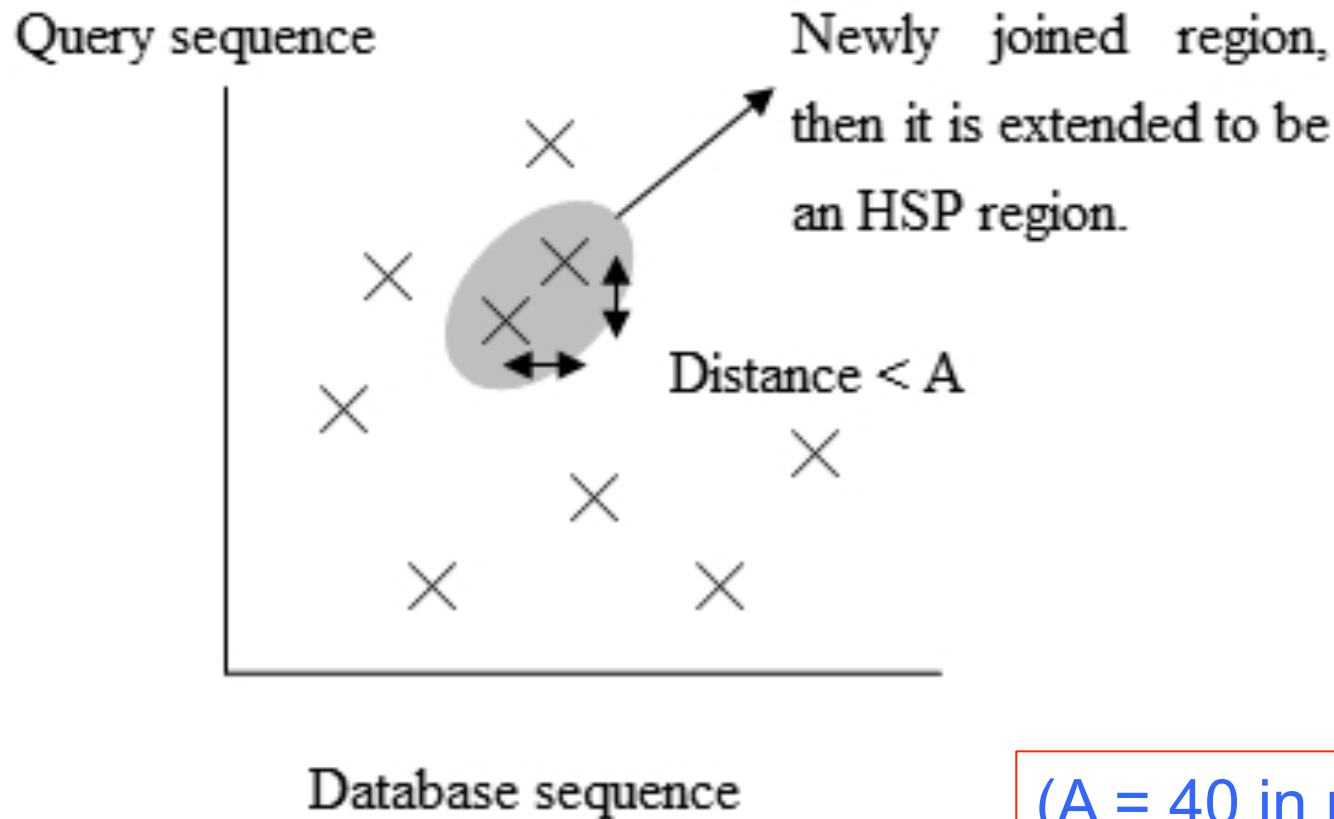
QUIZ: Word sizes – DNA vs protein *

- How many possible words for:
 - 3 letter protein code
 - 11 letter DNA code
- (If) these are standard settings:
 - Why would this be sensible? Name two reasons:
 - 1
 - 2
 - DNA words typically need to match exactly

BLAST - overview

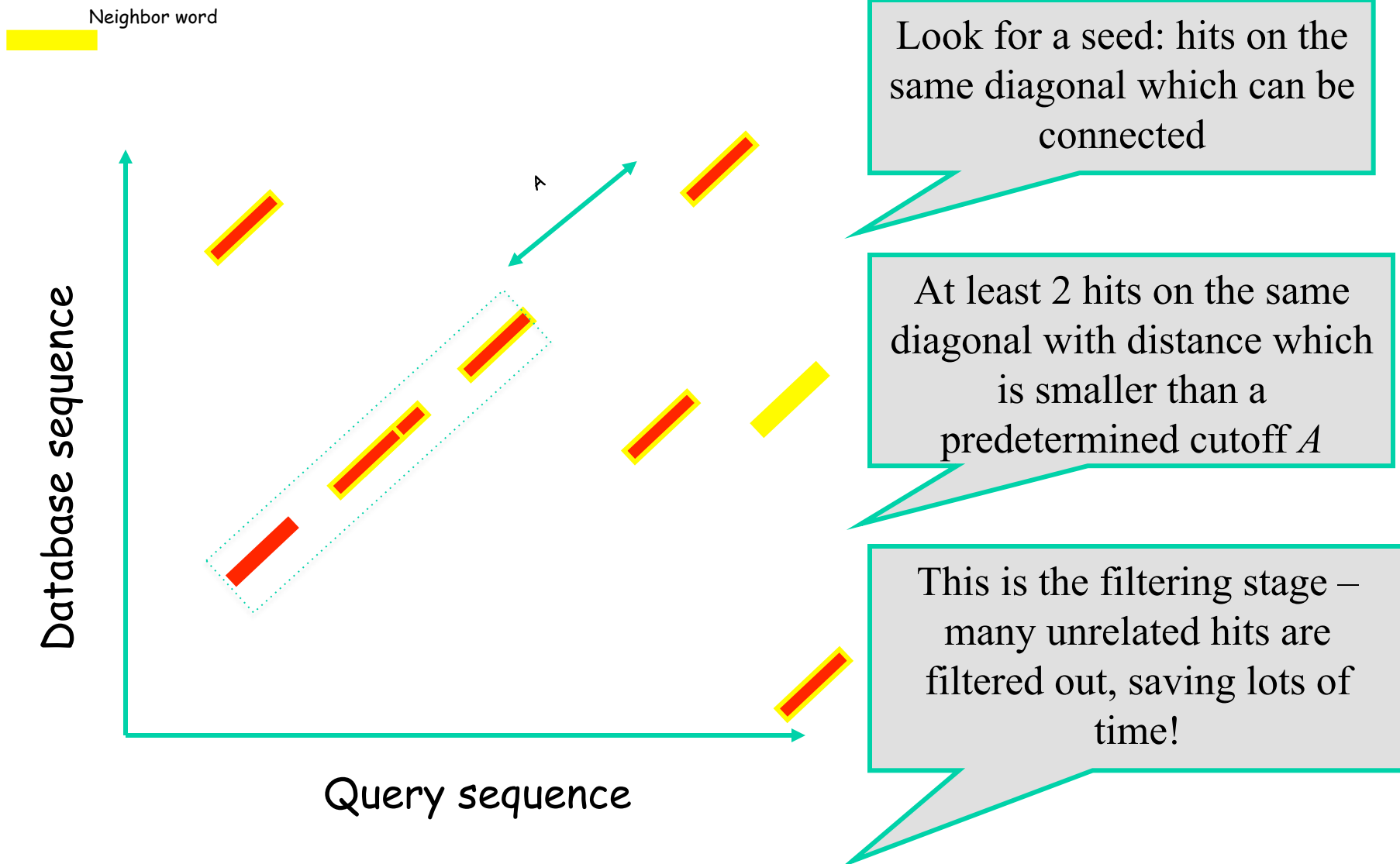


BLAST step 3 (extension): The Two-Hit Method

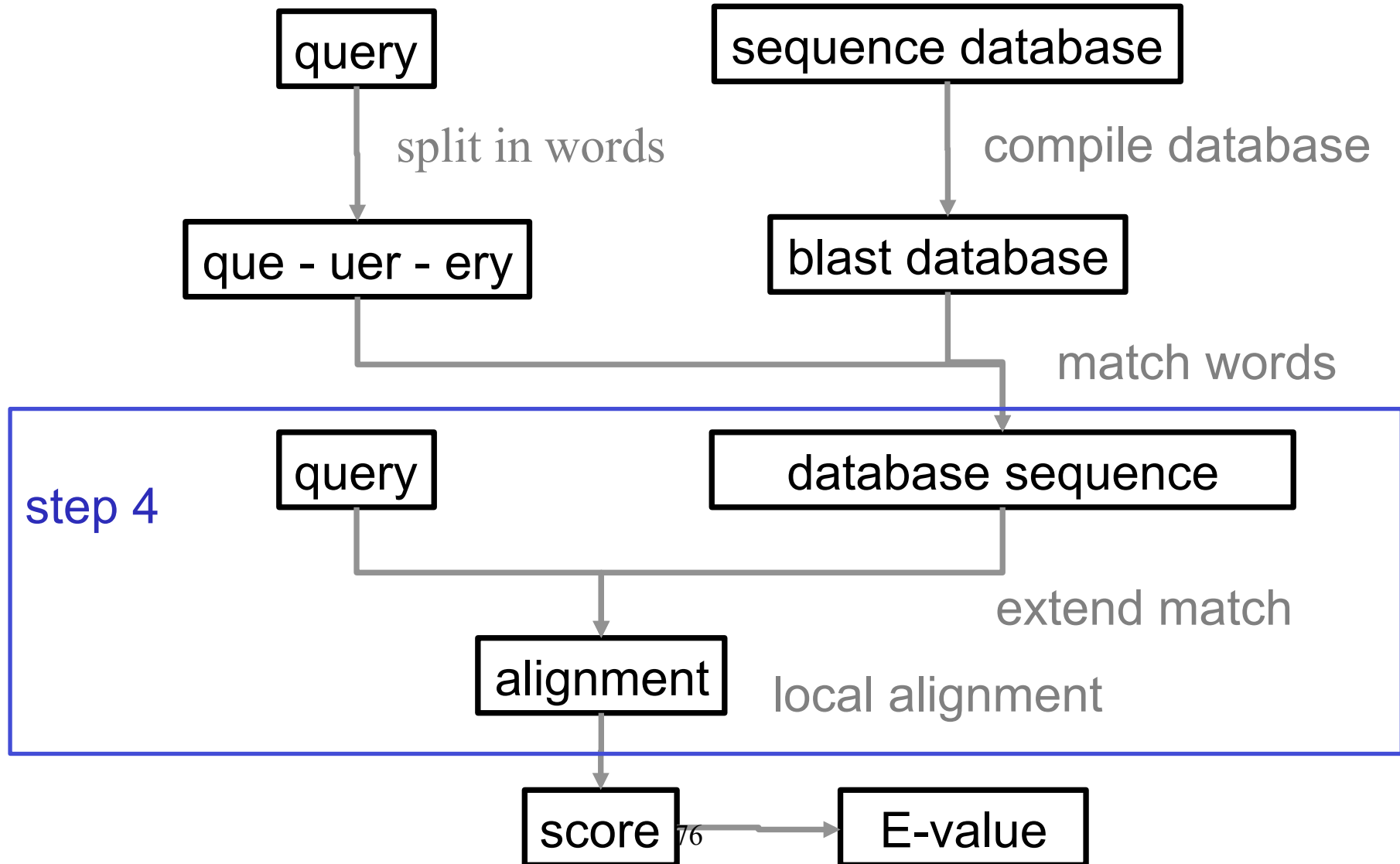


(A = 40 in many
BLAST
implementations)

BLAST two-hit extension

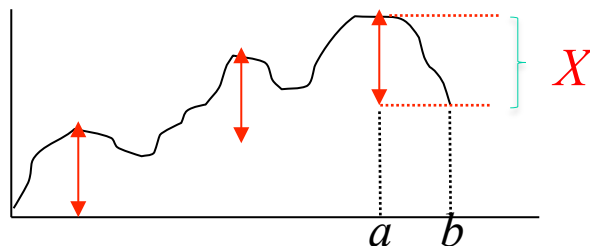


BLAST - overview



BLAST: Extension by local alignment

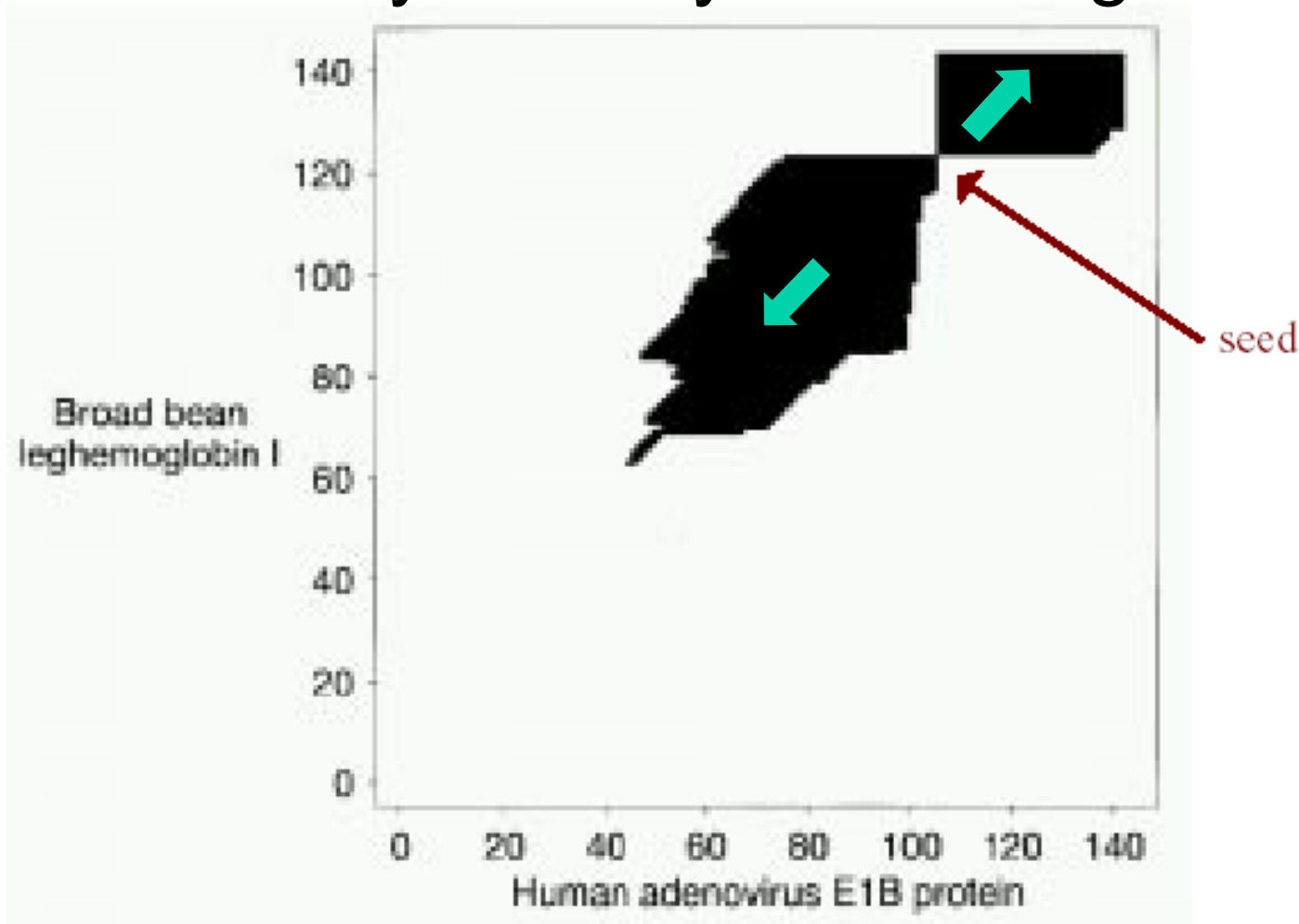
- trigger gapped alignment if two-hit extension has a sufficiently high score
 - So you first need to have two-hits on diagonal close enough together
- find length-11 segment with highest score (slide 11-window); use central pair in this segment (central position in corresponding window) as seed
- run DP process both forward & backward from seed
- prune cells when local alignment score falls a certain distance below best score yet



Continue as long as
 $\text{Score}(b) > \text{Score}(a) - X$

BLAST

2-way local Dynamic Programming



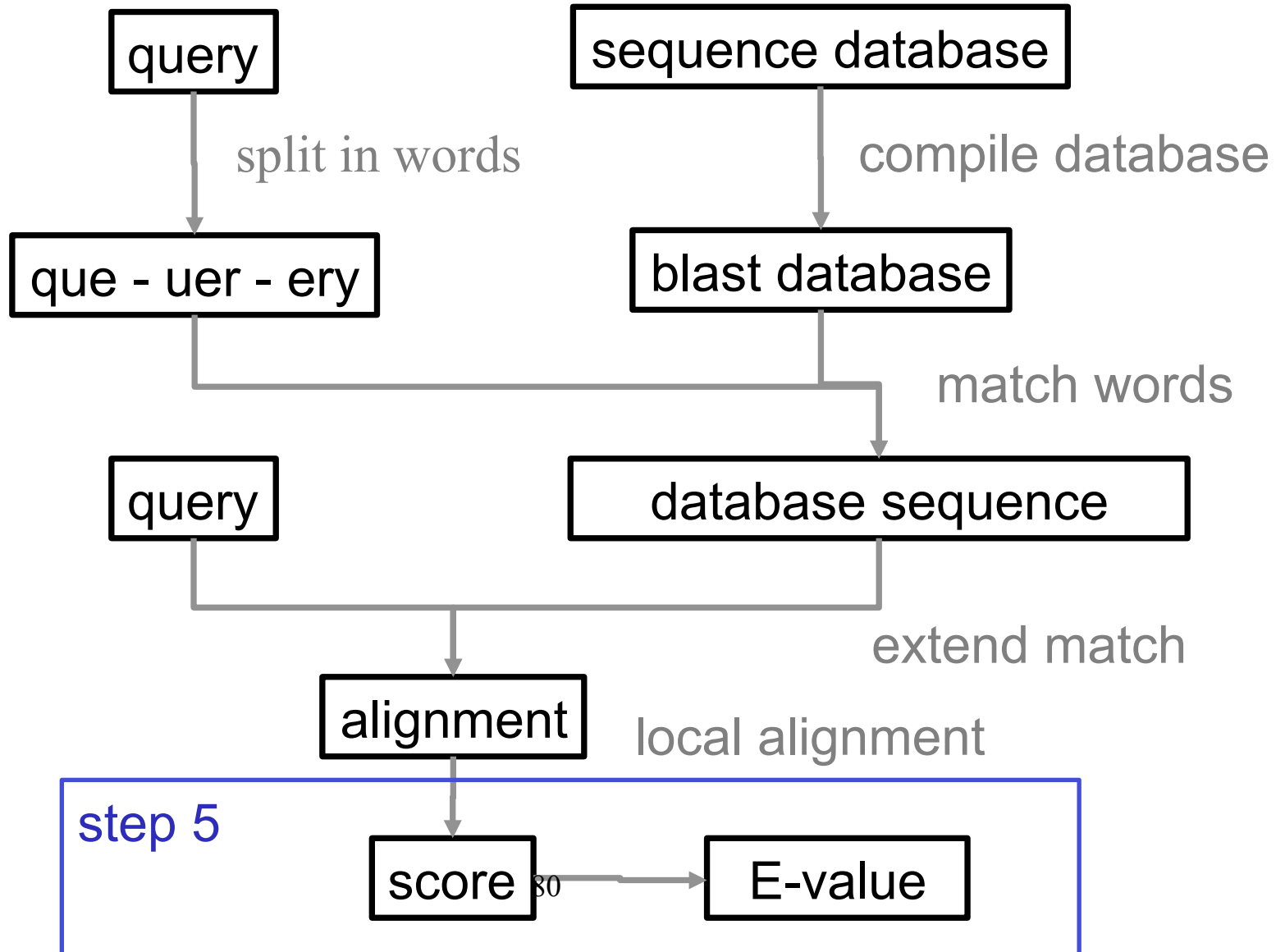
Wrapping up BLAST search

Two-Hit Gapped BLAST

The gapped BLAST algorithm:

1. Start with the two hit method-
 - (a) find two hits of score higher than T , within a distance A .
2. If the HSP generated has an expected score:
 - (a) Trigger a **gapped** extension
 - (b) If the final score has a significant E-value - report the gapped alignment.

BLAST - overview



Score to E-value

How can we compare alignment scores between different sequences?

Differentiate between random sequence similarity and homology.

		j→					
		1	2	3	4	5	6
i↓		G	A	G	T	G	A
1	G	1	0	1	0	1	0
2	A	0	2	0	0	0	2
3	G	1	0	3	1	1	0
4	G	1	0	1	2	2	0
5	C	0	0	0	0	1	1
6	G	1	0	1	0	1	0
		0	2				

Raw alignment score S — *how to convert this to a meaningful statistical score?*

Remainder of this lecture:

- Statistical scoring of database hits
- PSI-BLAST
- Performance evaluation using standard of truth

Scoring BLAST alignments

- Score should optimise the chance to select proper hits (True Positives)
- Scoring alignments is dependent on
 - The scoring system used (residue exchange matrix and gap penalty regime)
 - Characteristics of the sequence database (size, residue composition)
- The BLAST way of scoring has been adopted by other methods as well; e.g., some implementations of FASTA, etc.
 - **Bit-score**
 - **E-value**

Normalized score - Alignment Bit Score

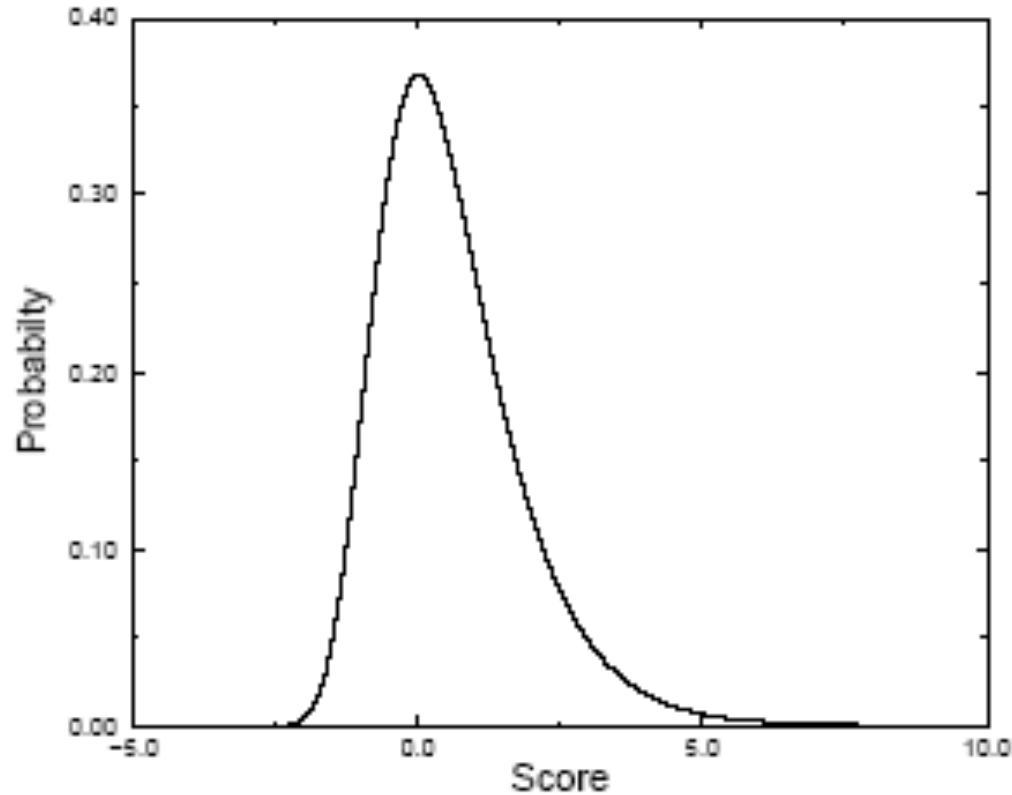
$$B = (\lambda S - \ln K) / \ln 2$$

- S is the raw alignment score
- The bit score ('bits') B has a standard set of units
- B is calculated from the number of gaps and substitutions associated with each aligned sequence pair. The higher the score, the more significant the alignment
- λ and K are statistical parameters associated with a given scoring system (e.g. BLOSUM62 in Blast)
 - See Altschul and Gish (1996) for a collection of values for λ and K over a set of widely used scoring matrices.
- Because bit scores are normalized with respect to the scoring system, they can be used to compare alignment scores from different searches based on different scoring schemes (a.a. exchange matrices and gap penalties)

The BLAST model for database searching score probabilities

- Scores resulting from searching with a query sequence against a database follow the Extreme Value Distribution (EVD) (Gumbel, 1955).
- Using the EVD, the raw alignment scores are converted to a statistical score (E value) that keeps track of the database amino acid composition and the scoring scheme (a.a. exchange matrix)

Extreme Value Distribution (EVD)



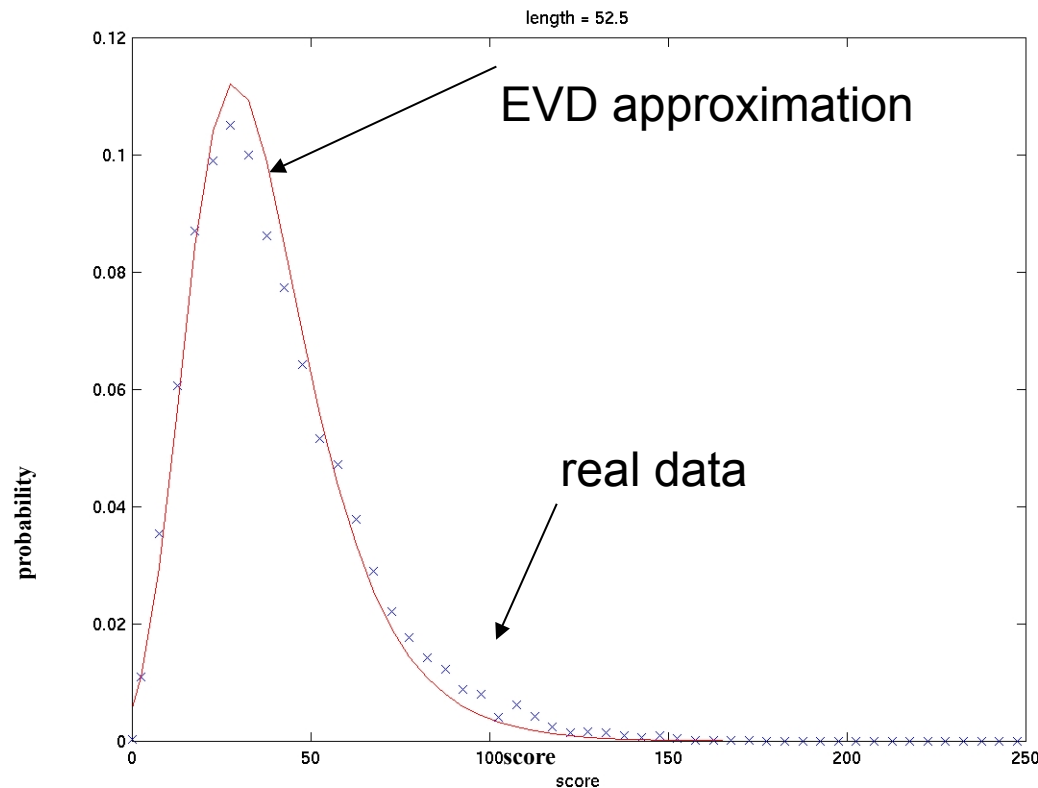
$$y = 1 - \exp(-e^{\lambda(x-\mu)})$$

Probability density function for the extreme value distribution resulting from parameter values $\mu = 0$ and $\lambda = 1$

$[y = 1 - \exp(-e^{-x})]$, where μ is the characteristic value (where the EVD peaks) and λ is the decay constant.

Extreme Value Distribution (EVD)

The optimal gapped local alignment **scores** produced by the Smith-Waterman algorithm or approximated by BLAST follow an extreme value distribution.



Compared to using the normal distribution, when using the EVD an alignment has to score further away from the expected mean value to become a significant hit.

Extreme Value Distribution (EVD)

The probability of a score S to be larger than a given value x can be approximated following the EVD as:

$$P(S \geq x) = 1 - \exp(-Kmn e^{-\lambda x})$$

- **P-value** for S local pairwise alignment score
- m : length sequence 1
- n : length sequence 2
- K and λ are fitted, to BLAST output – gaps, scoring scheme, amino acid composition
- Or: How likely is a pairwise hit with score S by chance? <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html>

E-value

Expected number of alignments with score at least S

$$E = Kmn e^{-\lambda S}$$

Used by BLAST as an approximation

- **E-value** for score S from a local pairwise alignment
- m : sequence length 1
- n : sequence length 2
- K and λ are fitted, to BLAST output – gaps, scoring scheme, amino acid composition (see bit score – earlier slide)
- Or: How often do we expect a pairwise hit with a score S by chance?

Database size

Or: How likely is a pairwise hit with score S by chance given the **database**?

- remember that a larger database makes fortuitous (random) hits more likely

- $E' = E \cdot D$

- Multiply the e-value (E) by database factor (D)

- $D = N$ (Fasta)

- $D = N/n$ (BLAST)

- N : size of the database

- n : size of the query sequence

- (longer sequences lower chance of random hit)

Normalised sequence similarity

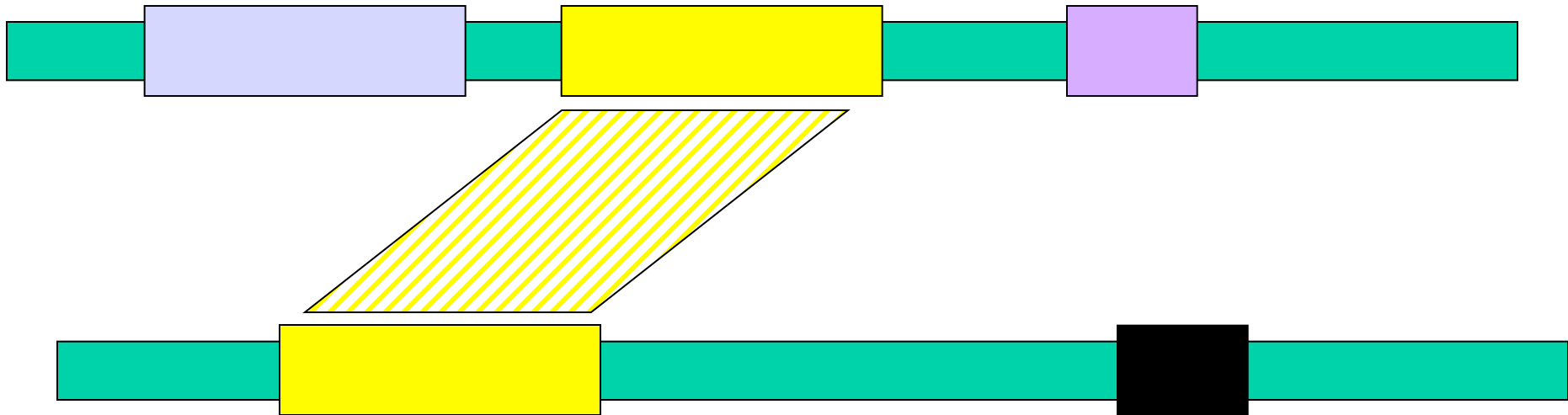
E-value: statistical significance

- The E-value of a BLAST alignment denotes the chance that the associated query and database sequence are *unrelated*
- Database searching is commonly performed using an E-value in between 0.1 and 0.001.
- Lower (i.e. stringent) E-values decrease the number of **false positives** in a database search (hit sequences that turn out not to be homologous), but increase the number of **false negatives** (homologous sequences not making it as hits), thereby lowering the sensitivity of the search (see later slides).

BLAST

The result - local alignment

- The result of BLAST will be a series of **local gapped alignments** between the query and the different hits found



Why use BLAST? *

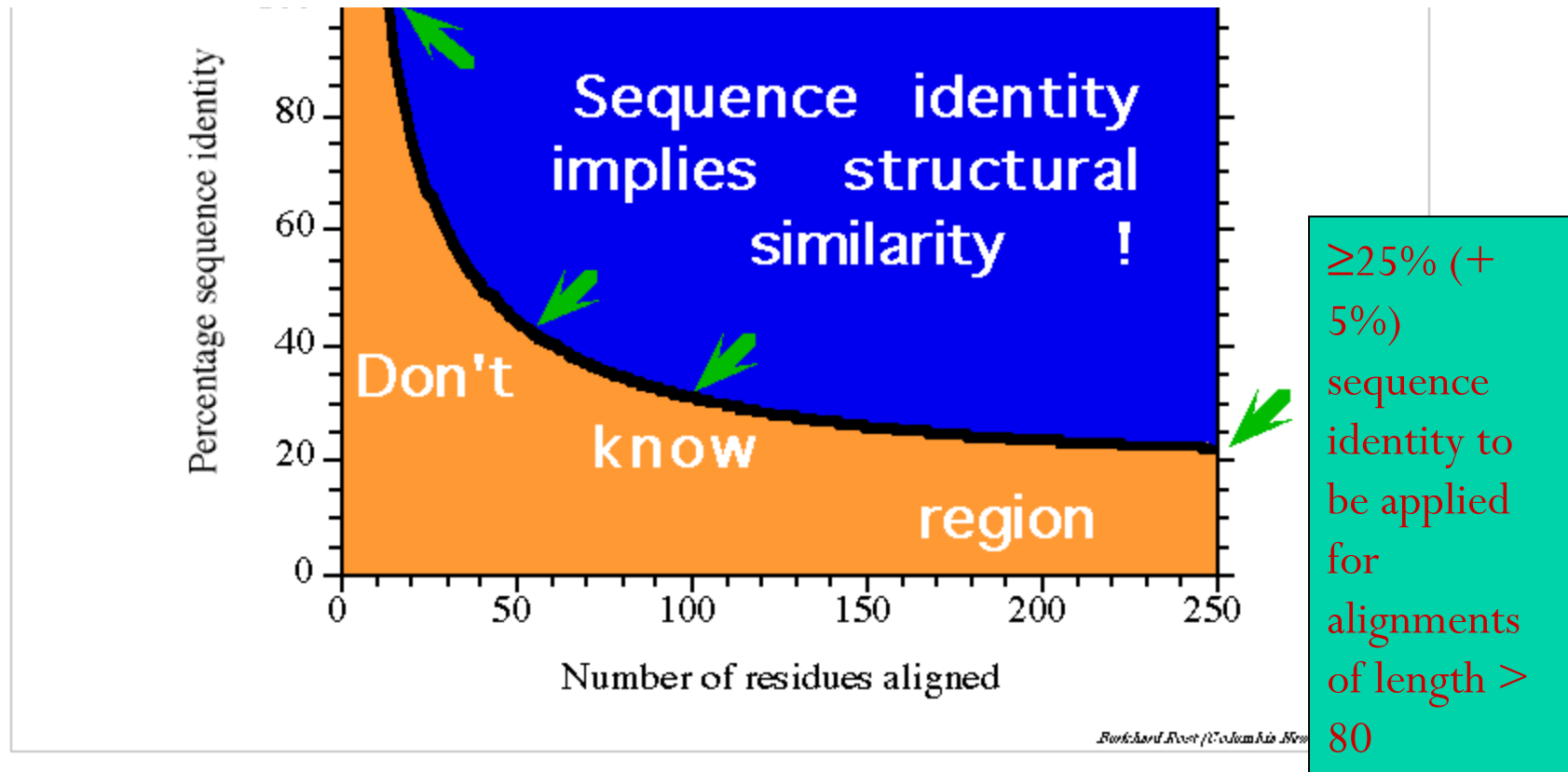
- We have found a hit, protein B, to our query protein A, with e-value of 0.001 . We know very little about protein A. What kind of information may you infer from this BLAST result?

- 1)
- 2)
- 3)

- Understanding this will be key to practicals / project!

What can sequence alignment tell us about structure (Recap)

HSSP Sander & Schneider, 1991



Some homologous protein families are distant

- Homologous sequences can be closely or distantly related
 - Histone family (protecting DNA) is completely conserved from bacteria to human
 - Hemoglobin family (transporting oxygen through blood) can be 90% different between close organisms
 - Some distant homologous families have sequences that have alignment scores below random (homology cannot be identified statistically anymore)
- How to find distantly related family members in a homology search?
- How to separate signal from noise?

Iterative homology searching using PSI-BLAST (Position-specific Iterated BLAST)

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997).

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25(17):3389-402.

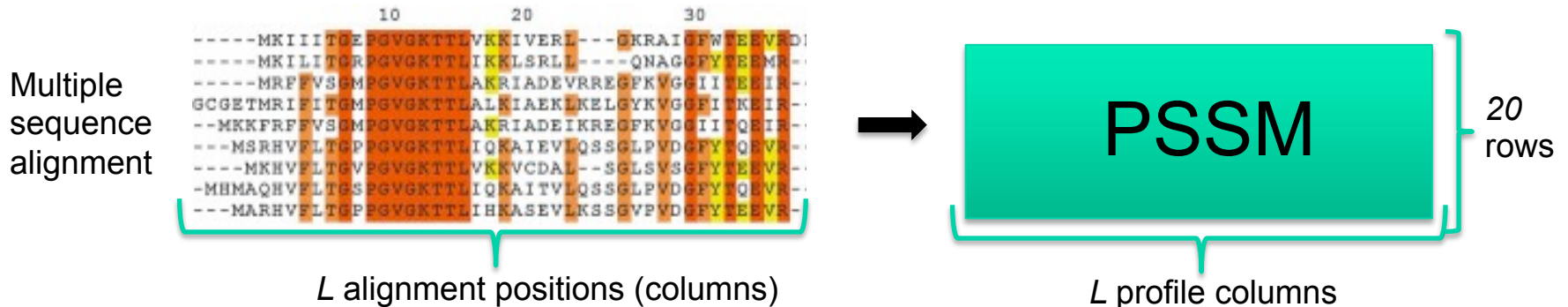
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

PSI (*Position Specific Iterated*) BLAST

- basic idea
 - Run gapped-BLAST, get putative homologs for query sequence and use the information to search better in a next iteration
 - This is done using results (hits) from initial BLAST search to construct a *profile matrix*
 - search database again with profile instead of query sequence (optimized search image – ‘put on better glasses’)
- Iterate
 - Hopefully at each iteration we aim to get more divergent family members (distant homologs)

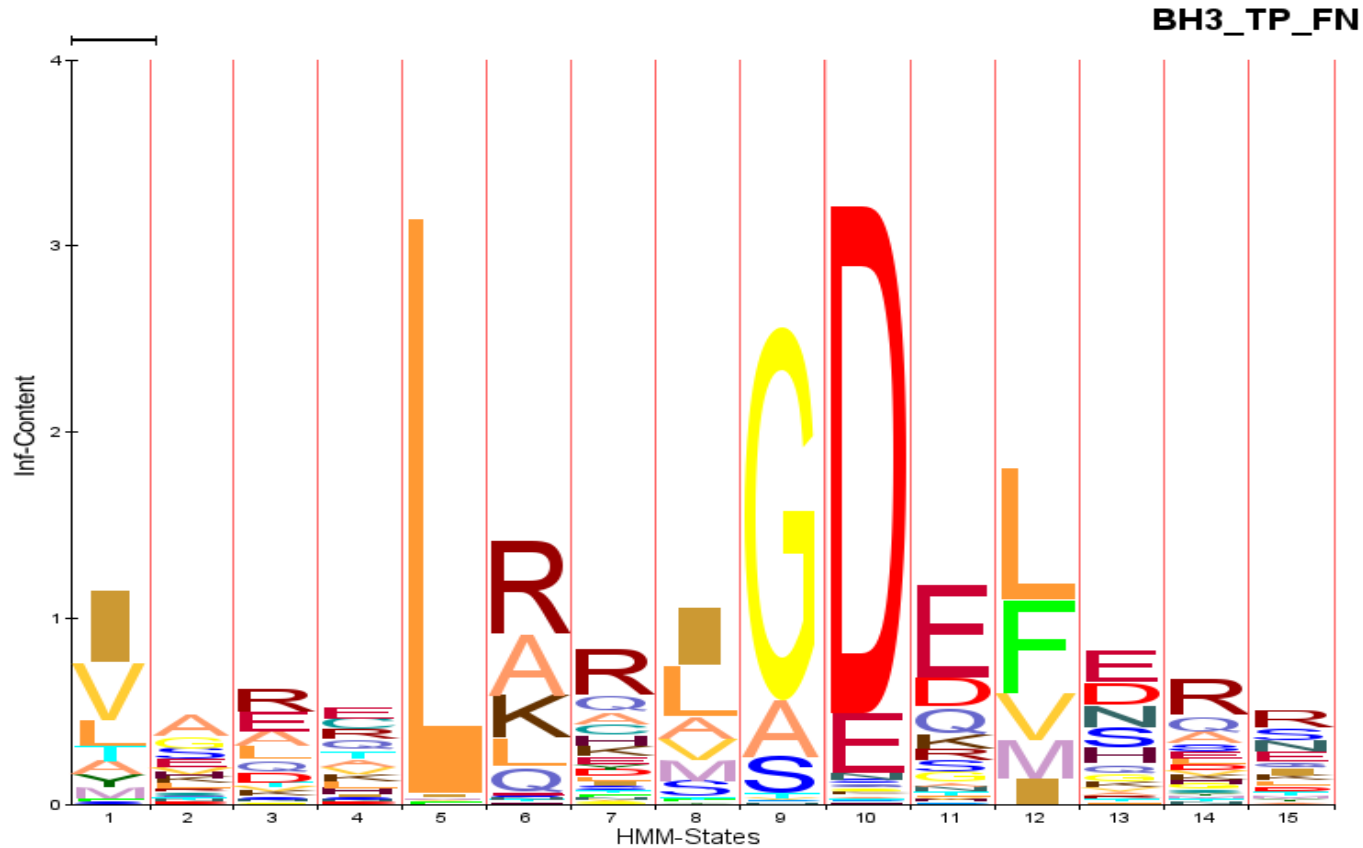
What is a sequence profile matrix?

- A sequence profile is a frequency-based scoring matrix that approximates the likelihood of occurrence of an amino acid at a given (multiple) alignment position
 - The mathematics to convert the a.a. frequencies to probabilities may differ (BLAST uses log conversion)
 - Basically, the scoring matrix has dimension $L * 20$, with L the number of columns (positions) in the multiple alignment (or BLAST alignment)
 - Profiles may have an extra column (21st position) to describe position-specific gap penalties.
 - In BLAST a profile is called **Position-Specific Scoring Matrix (PSSM)** and has only 20 columns.

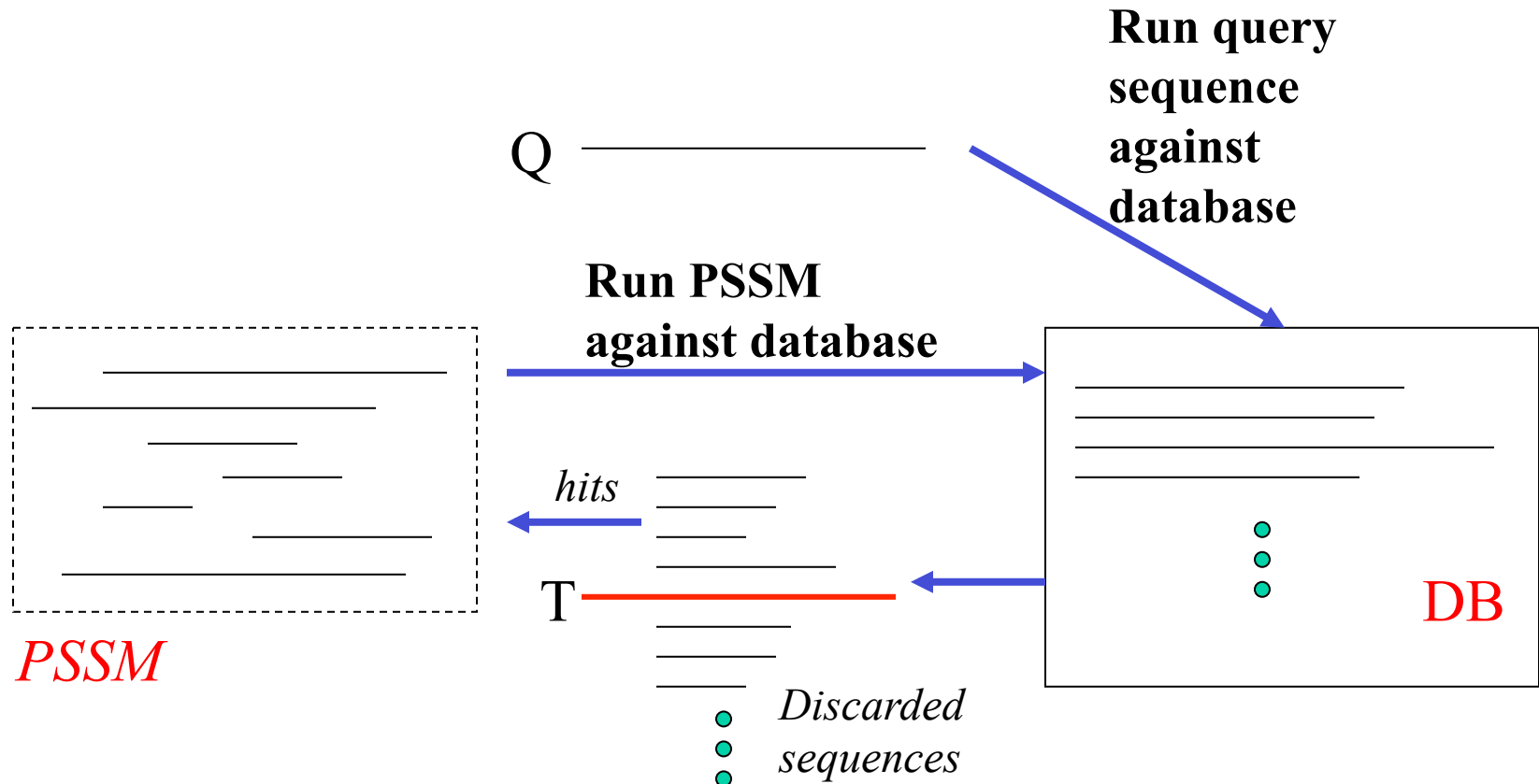


Homework: **

- How may a HMM logo / sequence profile help to find distant homologs?

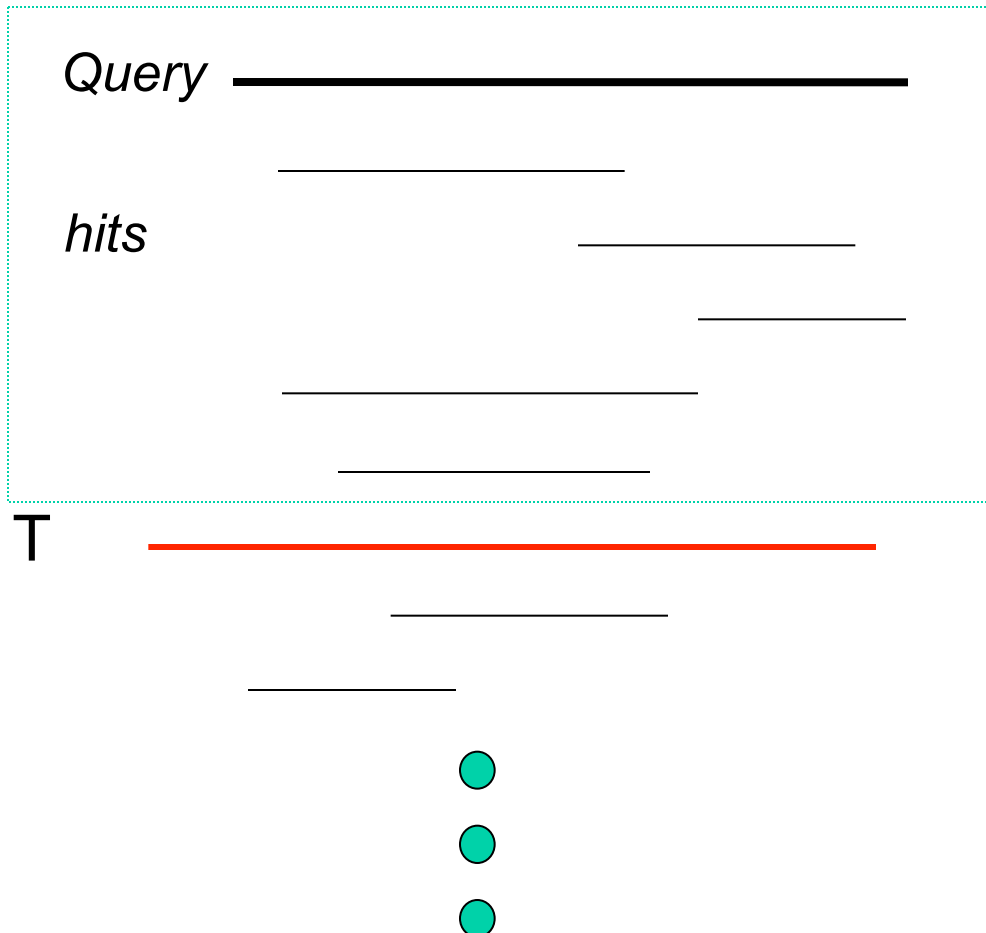


PSI-BLAST iteration scheme



The first iteration is running (gapped) BLAST, from second iteration onward a profile is used to search through database

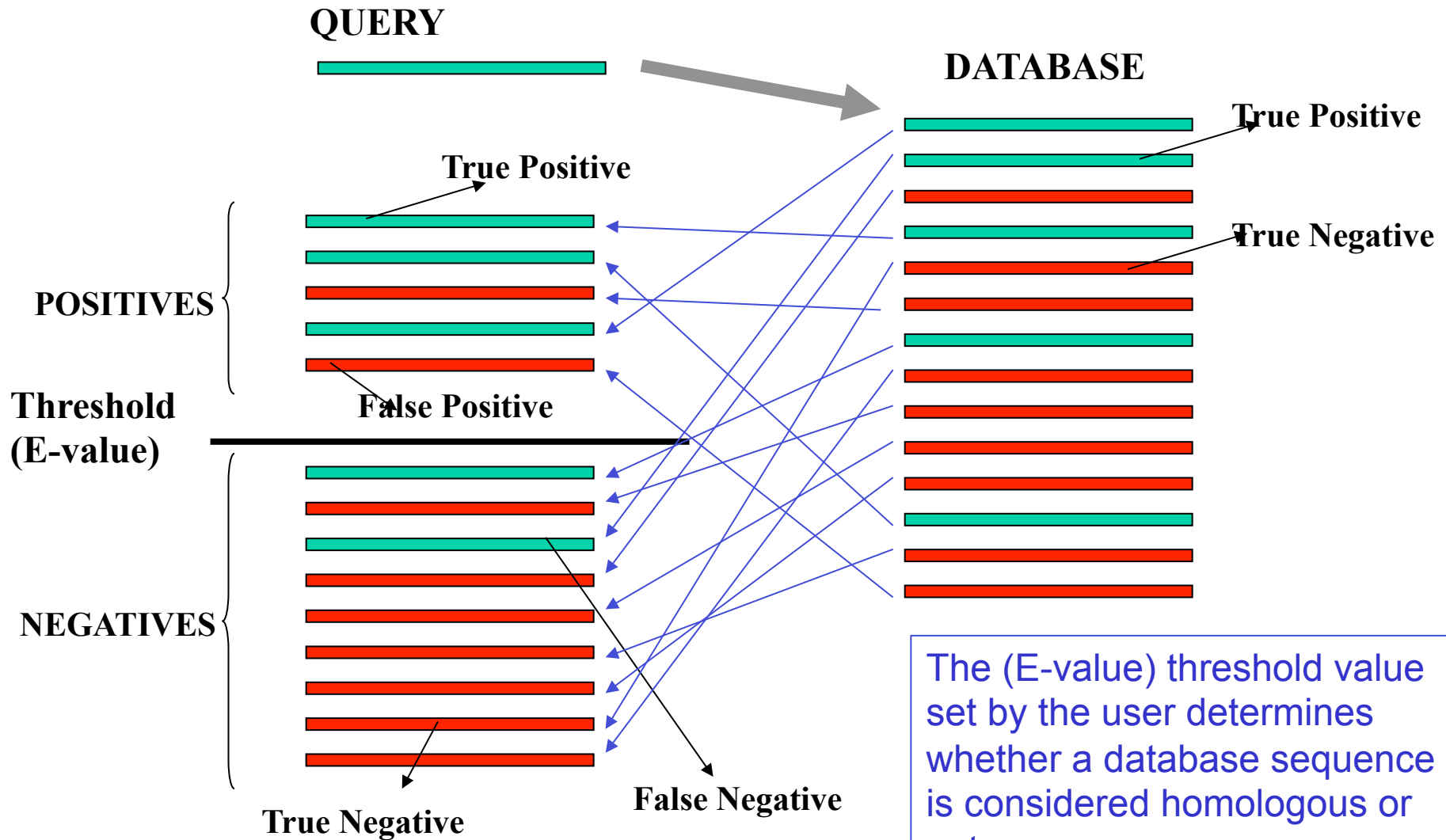
PSI-BLAST iteration



During iteration, new hits can come in and hits can drop out of the hit-list

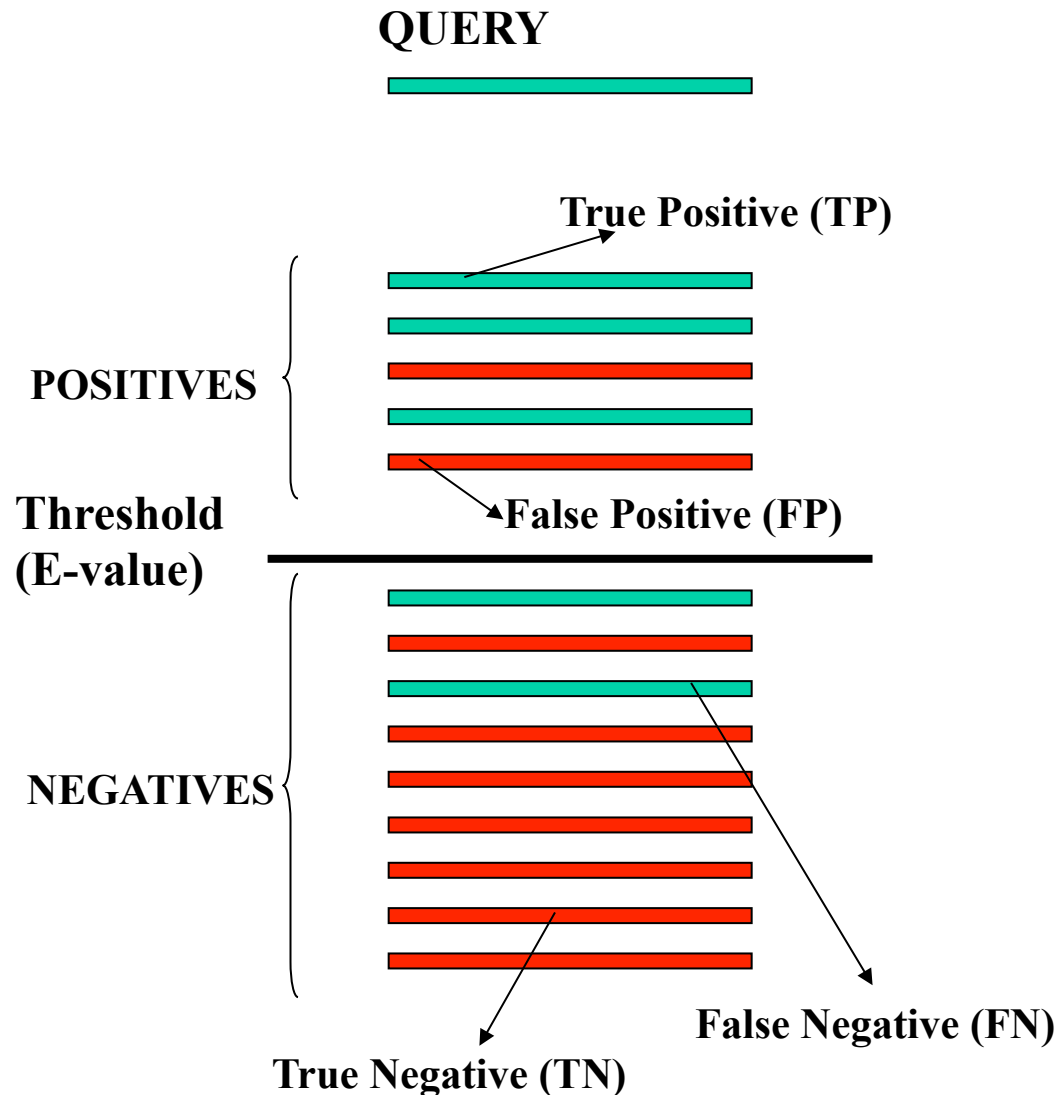
At each iteration a new profile is made of the master-slave alignment (query seq is the master)

Sequence searching: BLAST performance



The (E-value) threshold value set by the user determines whether a database sequence is considered homologous or not

Sequence searching: BLAST performance



$$\text{Coverage (recall)} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision (positive predictive value)} = \frac{TP}{TP + FP}$$

The (E-value) threshold value set by the user determines whether a database sequence is considered homologous or not

A note on (PSI-)BLAST

- ❑ BLAST and PSI-BLAST score alignments by E-values, and allow varying the E-value threshold. This will influence at which sequence similarity level (between query and a given database sequence) homology is declared.
 - Note that the statistical scoring scheme in (PSI-)BLAST is dependent upon a number of parameters, including size of the database, its residue composition, the residue exchange matrix, etc. This means that NR database updates (happening every week) will lead to a *changed statistical score* for the same query-DB sequence pair (upsetting some who don't expect this..) ¹⁰⁴

Take home

Part I

- Homology and alignment principles
- Needleman-Wunsch global alignment – **know and be able to execute the algorithm by hand!**
- Smith-Waterman local alignment (the famous zero)
- Local vs Global alignment

Part II

- Homology principle: transfer of structural and functional information
- BLAST and PSI-BLAST are heuristic methods
- BLAST: T-similar words
- PSI-BLAST: PSSM & iteration
- Statistical scoring: E-value & Database size

Take home (cont.)

- Alignments represent divergent evolution with alignment columns (positions) representing the common ancestor
- Dynamic programming (DP) is the main technique to perform alignments
- DP is too slow to perform database searching
- BLAST is the most important method for homology searching
- PSI-BLAST is the most sensitive technique in the BLAST suite (it iterates the search using sequence profiles)
- BLAST alignment statistics (E-values) are based upon the Extreme Value Distribution (EVD) which approximates the statistical score distribution – it is a fast way to get a statistical significance score

Take home (cont.)

- Global (Needleman-Wunsch) and local (Smith-Waterman) alignment with Dynamic Programming (DP)
- How does gapped BLAST work?
 - K-words; neighbor-words (T-similar words); two-hit method; gapped extension using DP; x-drop method
- How does PSI-BLAST work?
 - Iteration; PSSM (Position-specific scoring matrix)
 - This allows using conservation patterns at specific alignment positions (master-slave alignments) which increases sensitivity
- BLAST issue: Low-complexity filtering (discussed during project)
- PSI-BLAST pitfalls: Profile-wander / premature convergence (discussed during project)

END of Lecture 3