Bioinformatics Algorithms
COS-BIOL-530/630
Lecture03

Days & Times	**Room**	**Meeting Dates**
Tu 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025
Th 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

Bioinformatics Algorithms
COS-BIOL-530/630
Lecture03

Announcements
Week3:
Lecture03
Lab03

- Discussion 3
- Activity 3

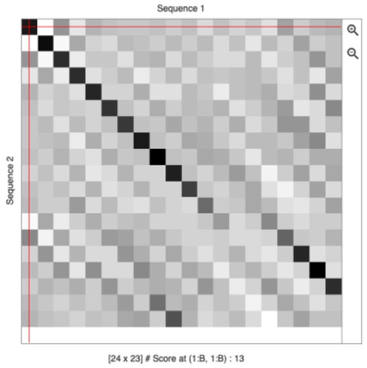
Activity2 (Filamin-A Dotplot) due on Thursday

Quiz 3 opens on Friday. Due on Tuesday 2pm

- Lab00 -> PHYLIP (next week)
- **Exam2: Tuesday April 15th (Schedule updated)**
- Oedipus?

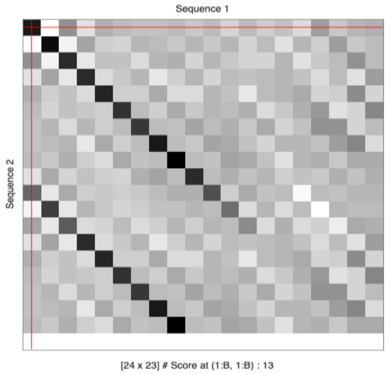
>Sequence1
BIQINFORMATICSALGQRITHMS

>Sequence2
BIQINFORMATICEURITHMICS



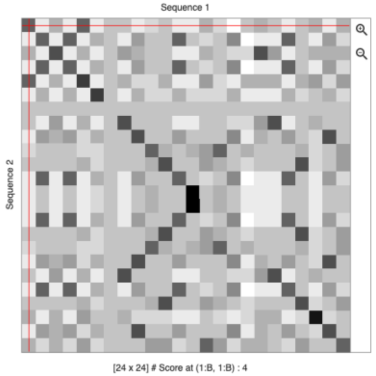
Local alignment
(match)

>Sequence1
BIQINFORMATICSALGQRITHMS
>Sequence2
BIQINFORMATICINFORMATIC



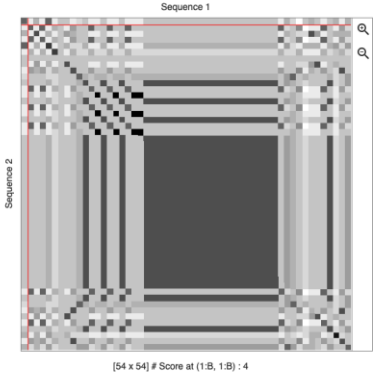
Repeat

>Sequence1
BIQINFORMATICSALGQRITHMS
>Sequence2
BIQINFORMATICCITAMRITHMS



Inverted repeat

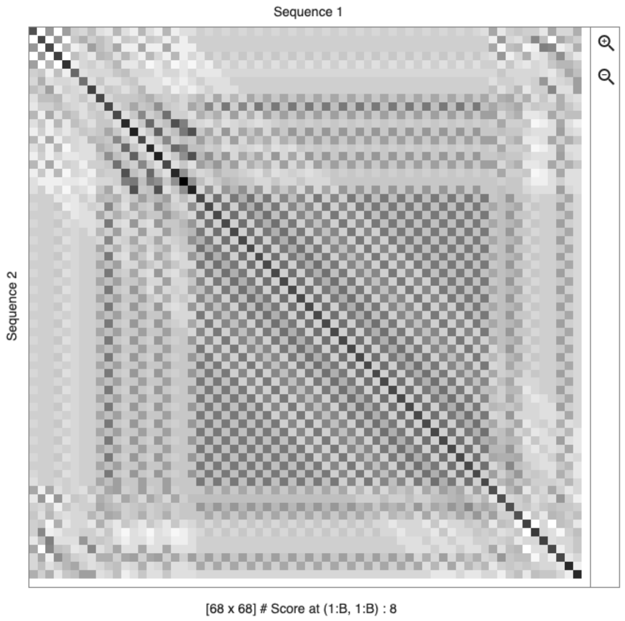
>Sequence1
BIQINFORMATICTICTICCTTTTTTTTTTTTTTTTTTTTTTISALGQRITHMS
>Sequence2
BIQINFORMATICTICTICTTTTTTTTTTTTTTTTTTTTTTTTTITAMRITHMS



Repeat (TIC)

Low complexity (T's)

window size = 1

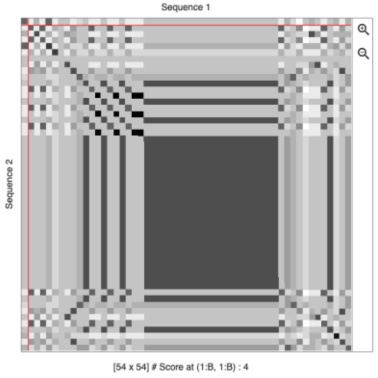


[68 x 68] # Score at (1:B, 1:B) : 8

>Sequence1
BIQINFORMATICTICTICCTATATATATATATATATATATATATATATATATATATATAISALGQRITHMS
>Sequence2
BIQINFORMATICTICTICCTATATATATATATATATATATATATATATATATATATATAISALGQRITHMS

window size = 2

>Sequence1
BIQINFORMATICTICTICCTTTTTTTTTTTTTTTTTTTTTTISALGQRITHMS
>Sequence2
BIQINFORMATICTICTICTTTTTTTTTTTTTTTTTTTTTTTTTITAMRITHMS



[54 x 54] # Score at (1:B, 1:B) : 4

Repeat (TIC)

Low complexity (T’s)

window size = 1

Alignments algorithms
- Lecture02 & Lecture03 -

- Pairwise Alignment
  - Global
  - Local
  - Scoring system

- Multiple Sequence Alignment (MSA)

- Heuristic – Database search

Dynamic programming algorithm (Quiz2 –Q1)

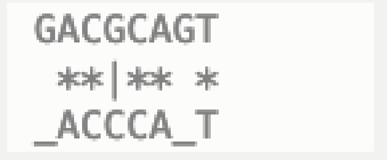
Global alignment: Needleman-Wunsch algorithm

Example from Discussion 2.2

Subject sequence:	GACGCAGT
Query sequence:	ACCCAT

Let's assume the Score Matrix:

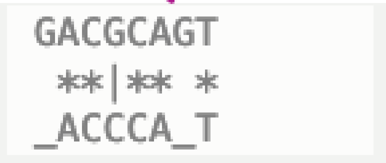
- ❖ Match = 5
- ❖ Mismatch = -2
- ❖ Gap penalty = -6



```
GACGCAGT
 **|** *
_ACCA_T
```

Let's assume the Score Matrix:

- ❖ Match = 4
- ❖ Mismatch = -6
- ❖ Gap penalty = -9



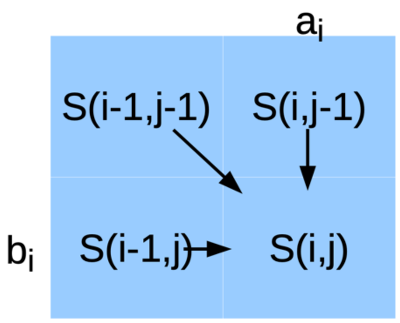
```
GACGCAGT
 **|** *
_ACCA_T
```

Alignment Score?

Dynamic programming algorithm

- Local alignment: **Smith-Waterman algorithm**
  It is just a variation of Global (Needleman-Wunsch algorithm)

- Performed the same way (steps) as global alignment
- To find the best, overall, alignment of subsets of sequences
- Negative scores = 0
- Find the best score within the matrix
- Traceback the best score to the first zero



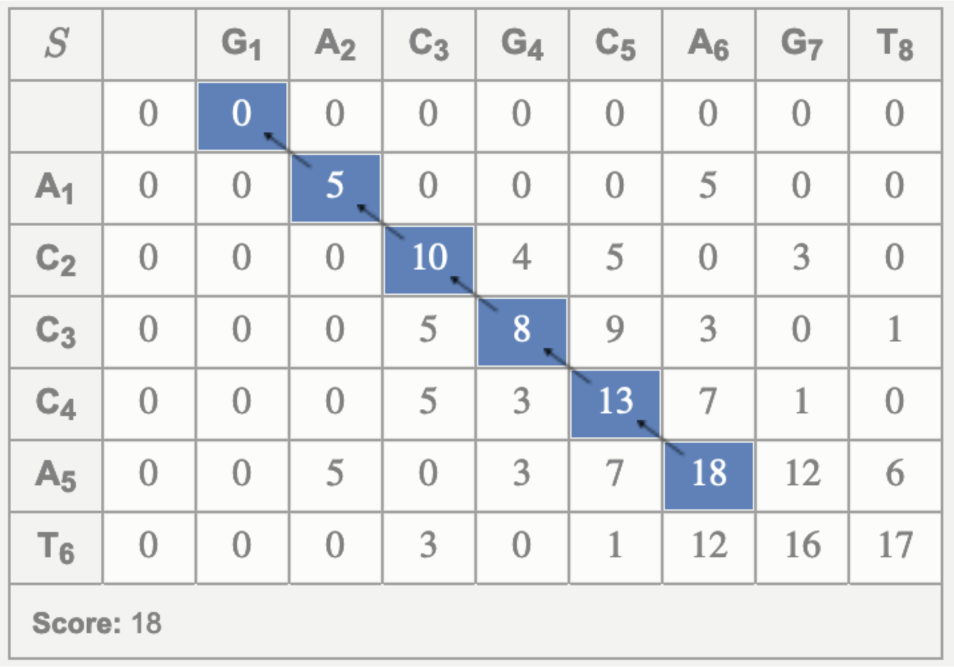
Dynamic programming algorithm

- ➢ Local alignment: **Smith-Waterman algorithm**
  It is just a variation of Global (Needleman-Wunsch algorithm)

Example from Discussion 2.2

Sequence1: GACGCAGT
Sequence2: ACCCAT

Scoring system:
Match: 5
Mismatch: -2
Gap: -6



S		G_1	A_2	C_3	G_4	C_5	A_6	G_7	T_8
	0	0	0	0	0	0	0	0	0
A_1	0	0	5	0	0	0	5	0	0
C_2	0	0	0	10	4	5	0	3	0
C_3	0	0	0	5	8	9	3	0	1
C_4	0	0	0	5	3	13	7	1	0
A_5	0	0	5	0	3	7	18	12	6
T_6	0	0	0	3	0	1	12	16	17

Score: 18

ACCCA
* * | * *
ACGCA

Alignments algorithms
– Lecture02 & Lecture03 -

- Pairwise Alignment
  - Global
  - Local
  - Scoring system

- Multiple Sequence Alignment (MSA)

- Heuristic – Database search

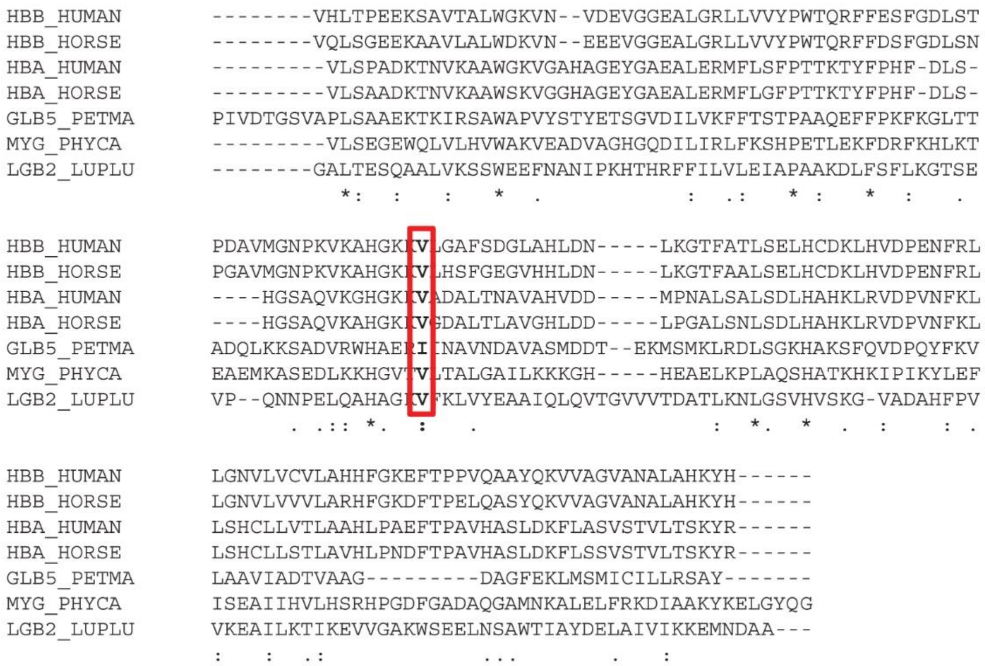
Multiple sequence alignments (MSA)

- Alignments that contain more than two sequences.

- Multiple alignments improve:
  - ➢ Accuracy of alignments between the sequence pairs.
  - ➢ Reveals patterns of conserved residues that are not obvious when looking at only two sequences.

- But:
  - ➢ Required more computing resources.
  - ➢ If alignment is incorrect in an active site residue, any function inference may be wrong.
  - ➢ Sometimes, we need “pre-aligned” sequences to generate multiple alignments.

Multiple sequence alignments



```
HBB_HUMAN       --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
HBB_HORSE       --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
HBA_HUMAN       ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
HBA_HORSE       ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
GLB5_PETMA      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
MYG_PHYCA       ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
LGB2_LUPLU      --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                          *:  :   :   *  .           :  .:   * :   *  :   .

HBB_HUMAN       PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
HBB_HORSE       PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
HBA_HUMAN       ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
HBA_HORSE       ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
GLB5_PETMA      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
MYG_PHYCA       EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
LGB2_LUPLU      VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
                      . .:: *.  :   .                  :  *.  *  .  :    : .

HBB_HUMAN       LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
HBB_HORSE       LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
HBA_HUMAN       LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBA_HORSE       LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
GLB5_PETMA      LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
MYG_PHYCA       ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LGB2_LUPLU      VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                :   :  .:            ...       .   :
```

Figure 8.1 An example multiple sequence alignment of seven globin protein sequences. One position is highlighted.

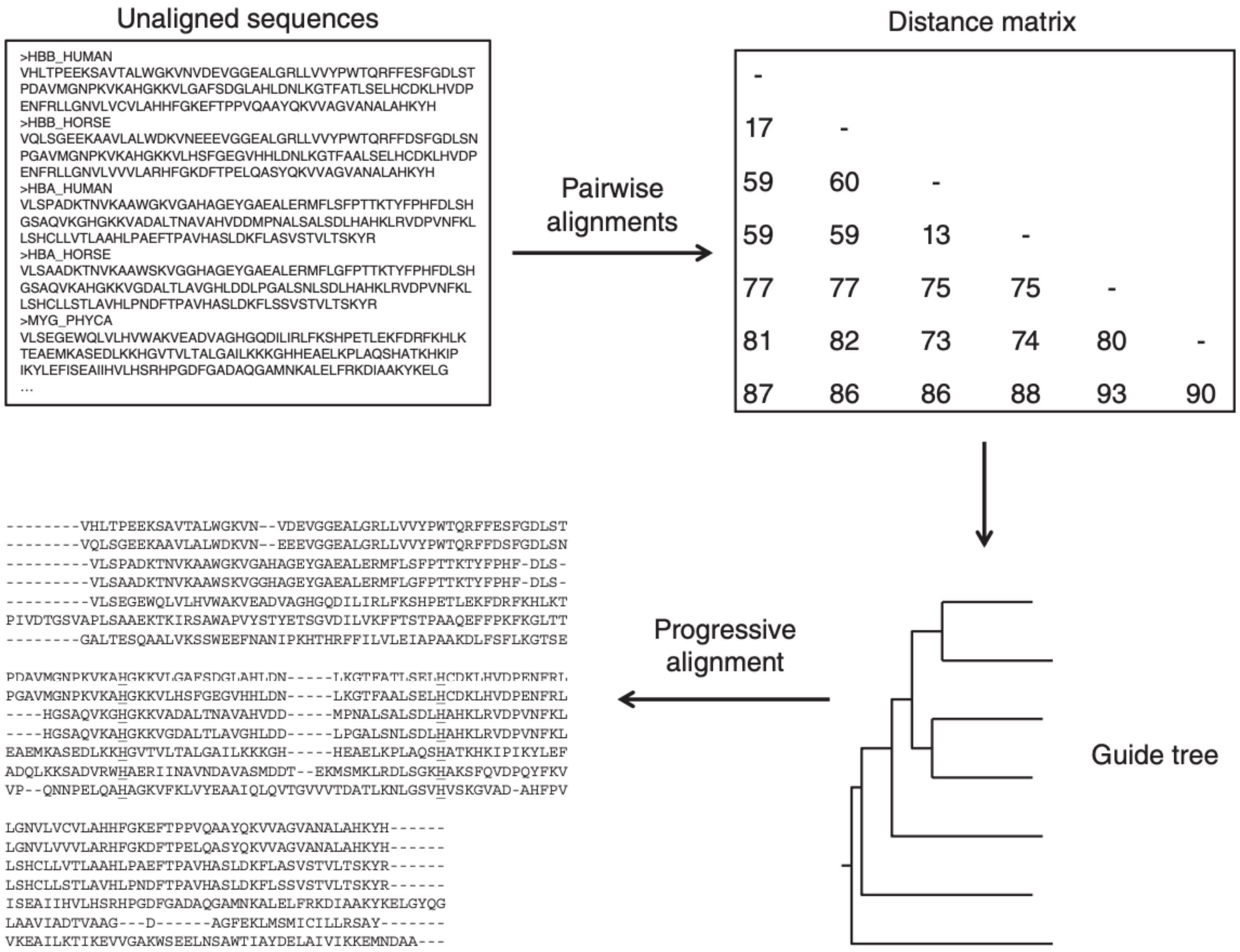
When constructing an MSA, one must also consider insertions and deletions from diverged sequences.

This means that MSA packages must be able to find an arrangement of null characters or "gaps" that maximizes the alignment of homologous residues.

Given a scoring scheme for residue matches and scores for gaps, one can attempt to find an MSA that produces the best overall score (and, thereby, the best overall alignment).

Bioinformatics, Fourth Edition. Edited by Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart.
© 2020 John Wiley & Sons, Inc. Published 2020 by John Wiley & Sons, Inc.
Companion Website: www.wiley.com/go/baxevanis/Bioinformatics_4e

Progressive alignment process



```
>HBB_HUMAN
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP
ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>HBB_HORSE
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
PGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDP
ENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH
>HBA_HUMAN
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
>HBA_HORSE
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSH
GSAQVKAHGKKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKL
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR
>MYG_PHYCA
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLK
TEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIP
IKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELG
...
```

-					
17	-				
59	60	-			
59	59	13	-		
77	77	75	75	-	
81	82	73	74	80	-
87	86	86	88	93	90

```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE

PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV

LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

simple progressive multiple alignment process

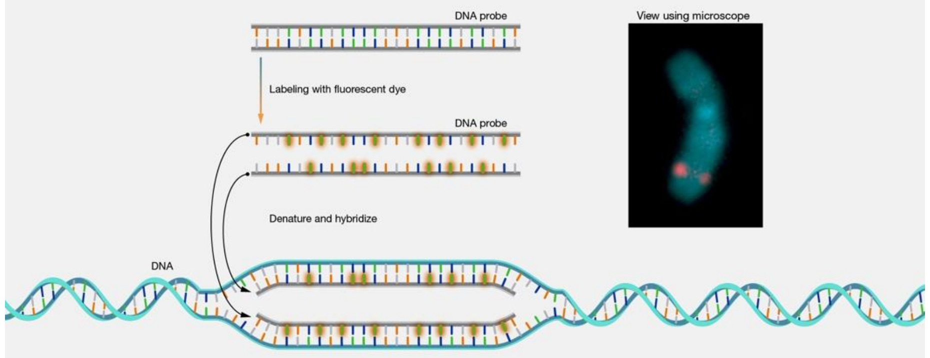
Multiple sequence alignments - Applications

- DNA/Protein: look for functional homology (consensus sequences)

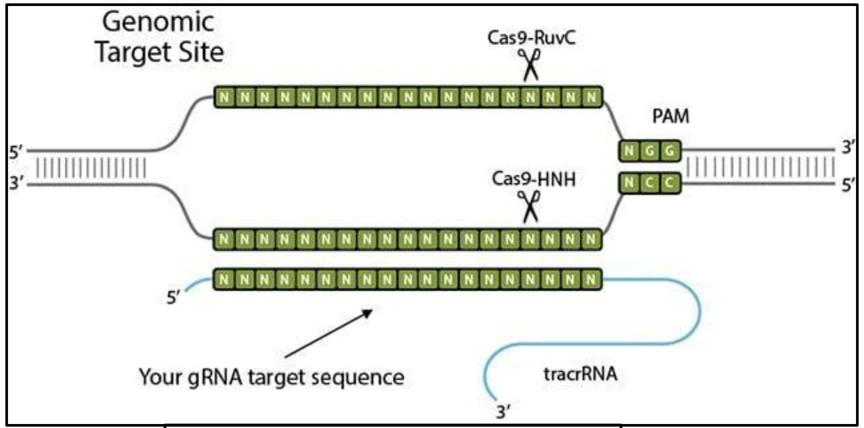
- DNA: highlight conserved regions in the genome
  - ✓ Transcription factor binding sites
  - ✓ Regions to design probes/oligos (fluorescent in situ hybridization - FISH, PCR)

- Proteins: conserved sequence regions (~functionality)
  - ✓ Active sites (eg. cleave site of an endonuclease)
  - ✓ Predicting 3D structure (eg. AlphaFold)
  - ✓ Identify protein family members
  - ✓ Experiment design (CRISPR/Cas modifications)

FISH



genome.gov



CRISPR/Cas system
Source: Sigma-Aldrich

Alignments algorithms
- Lecture02 & Lecture03 -

- Pairwise Alignment
  - Global
  - Local
  - Scoring system

- Multiple Sequence Alignment (MSA)

- Heuristic – Database search

Pairwise alignments & Databases (DBs)

- It all started in the early 1980’s:
  - European Molecular Biology Laboratory - EMBL (1980)
  - GenBank (1982).

  ➢ to establish a central database of DNA sequences rather than have scientists submit sequences to journals.

- By the late 1980’s there were too many sequences.

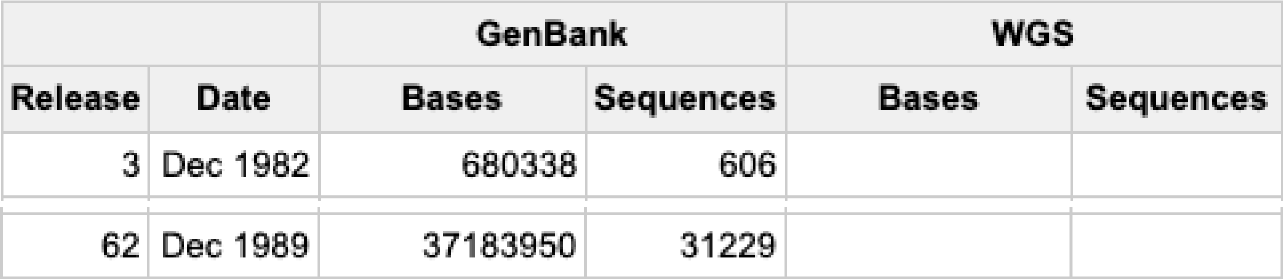
  ➢ Transition to digital format

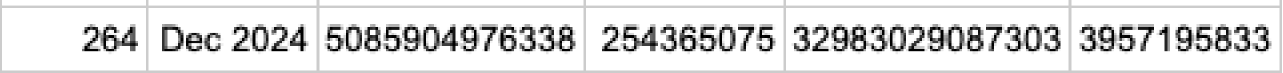
  ➢ But the question was: How to efficiently find related (homologous) sequences in the DBs??

- Development of tools to allow access to sequence databases (FASTP 1984,BLAST 1990, ENTREZ 1992).

Pairwise alignments & Databases (DBs)

WHY not to run pairwise alignments against the DB?





		GenBank		WGS	
Release	Date	Bases	Sequences	Bases	Sequences
3	Dec 1982	680338	606		
62	Dec 1989	37183950	31229		
264	Dec 2024	5085904976338	254365075	32983029087303	3957195833

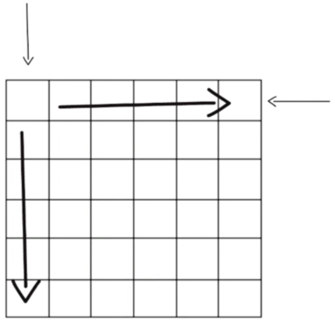
… against over 250 million sequences??

Database searches

- Pairwise alignments are not adequate
- FASTP (Lipman and Pearson 1985)
- FASTA (Pearson and Lipman 1998)
- BLAST (Altschul et al. 1990)
- And others: heuristic methods
  - ➢ Fast approximation to dynamic programming
  - ➢ Contributes to a higher speed execution
  - ➢ But we have to make some assumptions…

Database searches

- Let’s say we want to search a query against the UniProtKB/Swiss-Prot protein knowledgebase (2024):
  - 57,1864 sequence entries
  - 207,016,062 amino acids
  - Query: 50 amino acids
  - How would you do it via dynamic programming?



FAST algorithms

- FASTP (Lipman and Pearson 1985) was the first widely used program designed for database similarity searching.

- FAST (FASTP, FASTA) **assumes that small local areas of similarity are more common in two related sequences than in two unrelated sequences**.

- The FAST algorithms can be divided into four major steps:

  - ➢ Identify regions of identity (words). The word length parameter or k-tuple is called ***ktup*** (equivalent to word size ***W*** in BLAST).

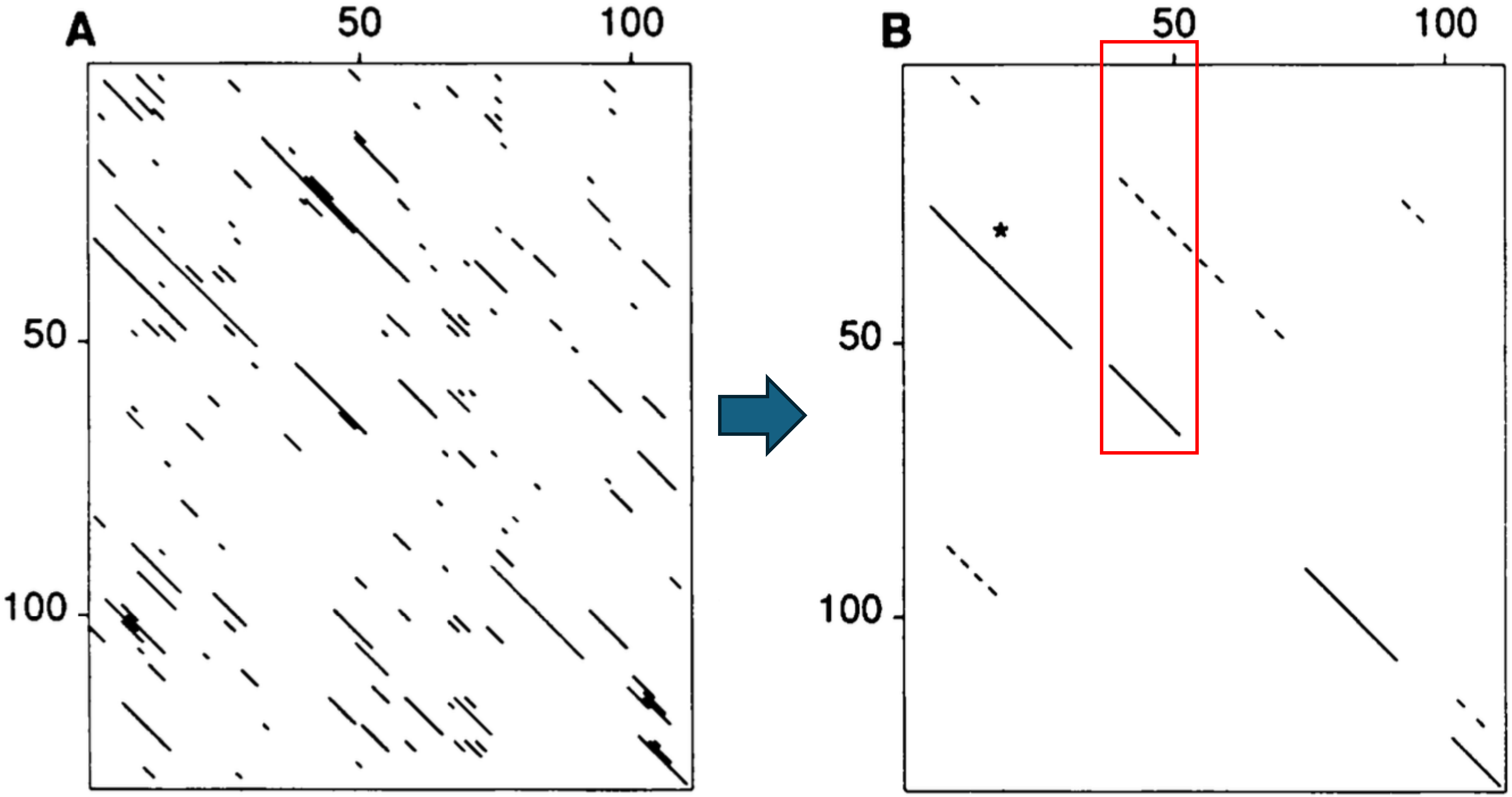
  - ➢ Scan the regions using a scoring matrix and save the best initial regions (using a Threshold)

  - ➢ Join initial regions with scores > Threshold (T)

  - ➢ Optimized initial score

FASTA algorithm

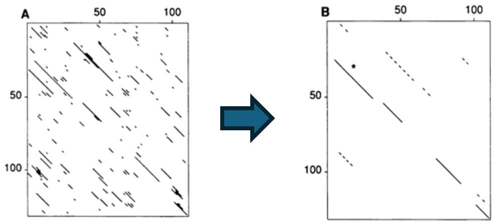
Find all regions of local identity



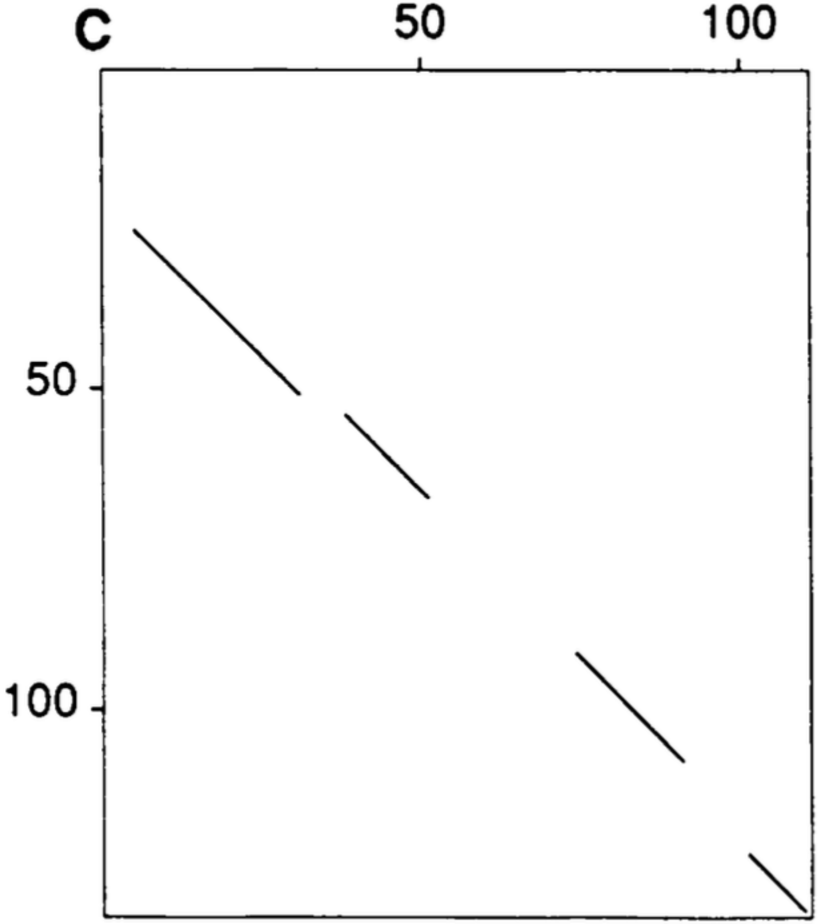
Select the best regions of local identity using a threshold score * (T).

Note that some regions are mutually exclusive (red rectangle).

Pearson and Lipman 1988



FASTA algorithm



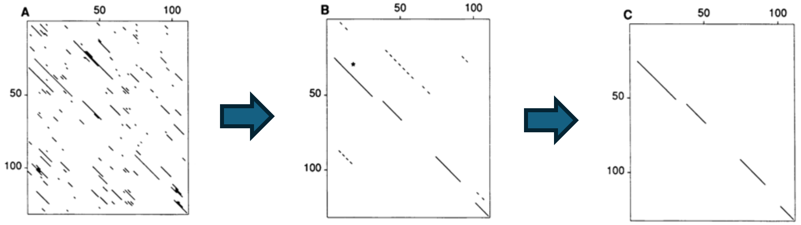
Pearson and Lipman 1988

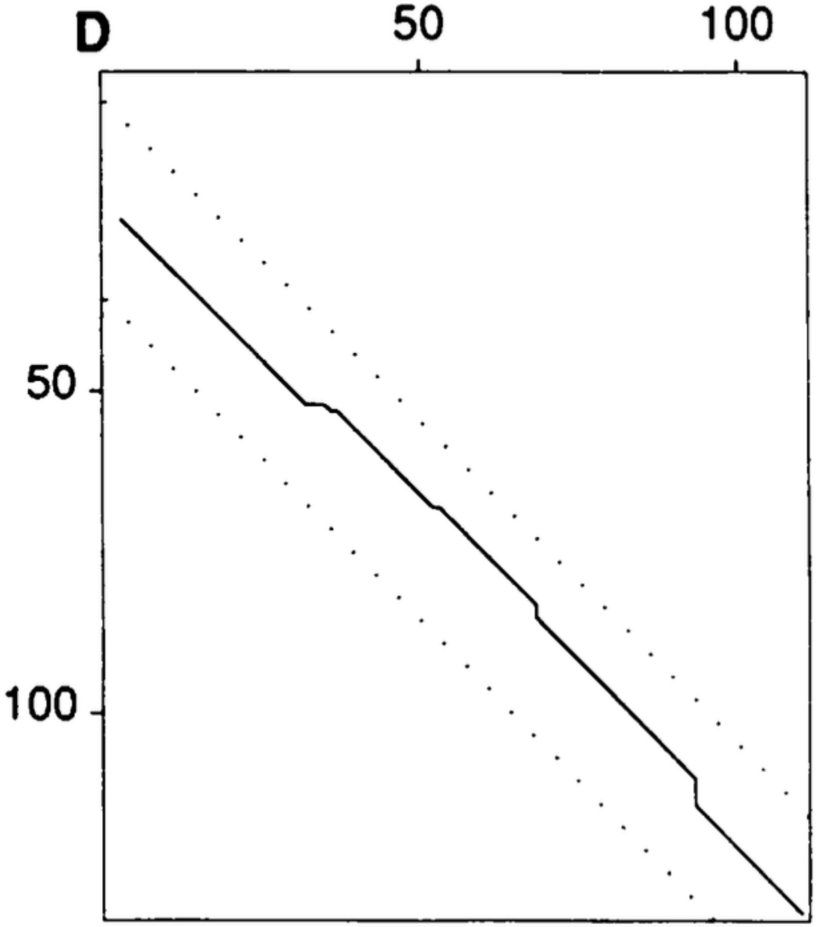
After the initial round of scoring, the top diagonal are selected.

Dashed lines (mutually exclusive) fell off. They were mutually exclusive with the best score.

Now, join the regions of the highest similarity.

FASTA algorithm





Pearson and Lipman 1988

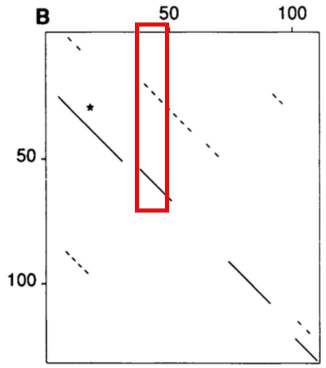
Recalculate an optimized alignment centered around the highest-scoring initial region.

Run dynamic programming along the diagonal to optimize the alignment.

The Smith-Waterman algorithm is applied to yield the optimal pairwise alignments between the two sequences.

FASTA algorithm

- FAST assumption: small areas of local similarity with the highest scores are most likely to be evolutionary conserved regions.
- Do you see any problem with this assumption?



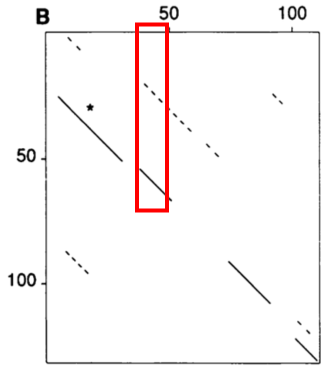
FASTA algorithm

- FAST assumption: small areas of local similarity with the highest scores are most likely to be evolutionary conserved regions.

- But we also assume **duplications** and **gene conversions** are not present (exclusive areas in the dot blot).

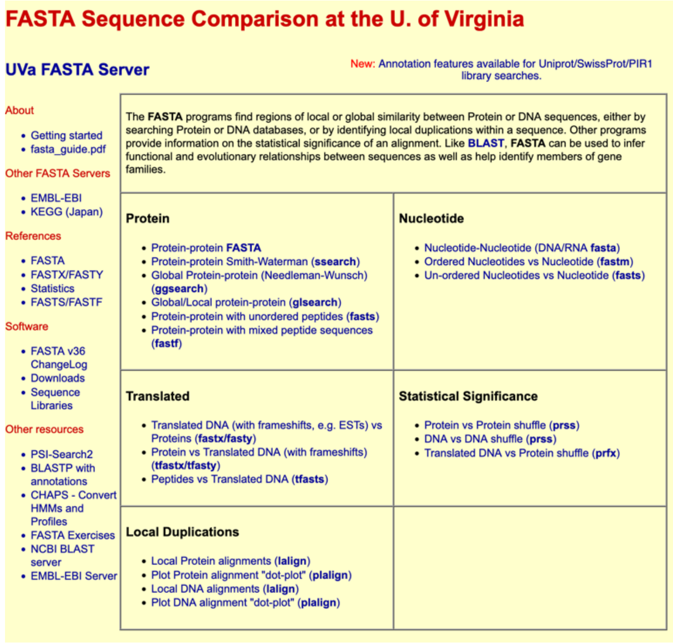
  - Initially, not many sequences deposited carried **duplications** (single copy human genes, mitochondrial genes).

  - More organisms being sequenced: alignments can show tricked scenarios where the FAST algorithm is far from perfect.



FASTA algorithm

- Instead of doing a full dynamic programming query *vs*. database.
- First, check if there is enough k-tuple (***ktup***) matches to be worthy.
- Then, extend and merge the k-tuples.
- It is similar to dynamic programming....but faster.



FASTA Sequence Comparison at the U. of Virginia

UVa FASTA Server

New: Annotation features available for Uniprot/SwissProt/PIR1 library searches.

About

- Getting started
- fasta_guide.pdf

Other FASTA Servers

- EMBL-EBI
- KEGG (Japan)

References

- FASTA
- FASTX/FASTY
- Statistics
- FASTS/FASTF

Software

- FASTA v36 ChangeLog
- Downloads
- Sequence Libraries

Other resources

- PSI-Search2
- BLASTP with annotations
- CHAPS - Convert HMMs and Profiles
- FASTA Exercises
- NCBI BLAST server
- EMBL-EBI Server

The **FASTA** programs find regions of local or global similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like **BLAST**, **FASTA** can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Protein

- Protein-protein **FASTA**
- Protein-protein Smith-Waterman (**ssearch**)
- Global Protein-protein (Needleman-Wunsch) (**ggsearch**)
- Global/Local protein-protein (**glsearch**)
- Protein-protein with unordered peptides (**fasts**)
- Protein-protein with mixed peptide sequences (**fastf**)

Nucleotide

- Nucleotide-Nucleotide (DNA/RNA **fasta**)
- Ordered Nucleotides vs Nucleotide (**fastm**)
- Un-ordered Nucleotides vs Nucleotide (**fasts**)

Translated

- Translated DNA (with frameshifts, e.g. ESTs) vs Proteins (**fastx/fasty**)
- Protein vs Translated DNA (with frameshifts) (**tfastx/tfasty**)
- Peptides vs Translated DNA (**tfasts**)

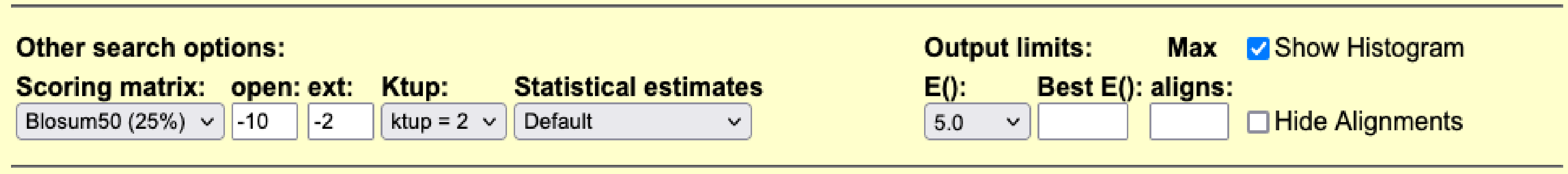
Statistical Significance

- Protein vs Protein shuffle (**prss**)
- DNA vs DNA shuffle (**prss**)
- Translated DNA vs Protein shuffle (**prfx**)

Local Duplications

- Local Protein alignments (**lalign**)
- Plot Protein alignment "dot-plot" (**plalign**)
- Local DNA alignments (**lalign**)
- Plot DNA alignment "dot-plot" (**plalign**)

https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml



Other search options:

Scoring matrix: Blosum50 (25%) **open:** -10 **ext:** -2 **Ktup:** ktup = 2 **Statistical estimates** Default

Output limits: **E():** 5.0 **Best E():** **Max aligns:**

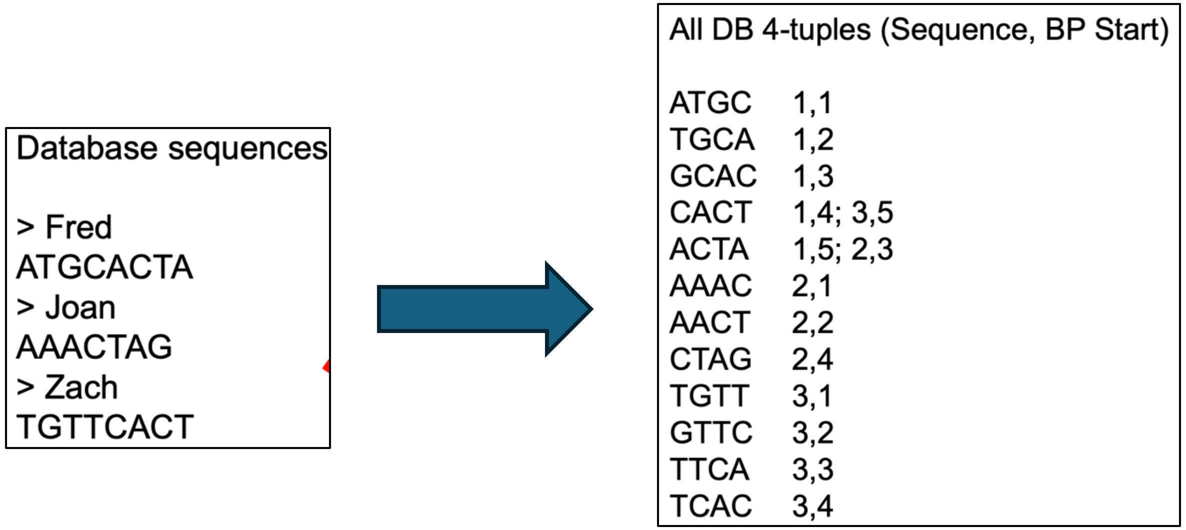
☑ Show Histogram

□ Hide Alignments

Database search

WHAT IF….

- database sequences are preprocessed with a list of all seed of size k.
- Then, we only need to search the query against the indexed database



Database sequences

> Fred
ATGCACTA
> Joan
AAACTAG
> Zach
TGTTCACT

All DB 4-tuples (Sequence, BP Start)

ATGC	1,1
TGCA	1,2
GCAC	1,3
CACT	1,4; 3,5
ACTA	1,5; 2,3
AAAC	2,1
AACT	2,2
CTAG	2,4
TGTT	3,1
GTTC	3,2
TTCA	3,3
TCAC	3,4

BLAST

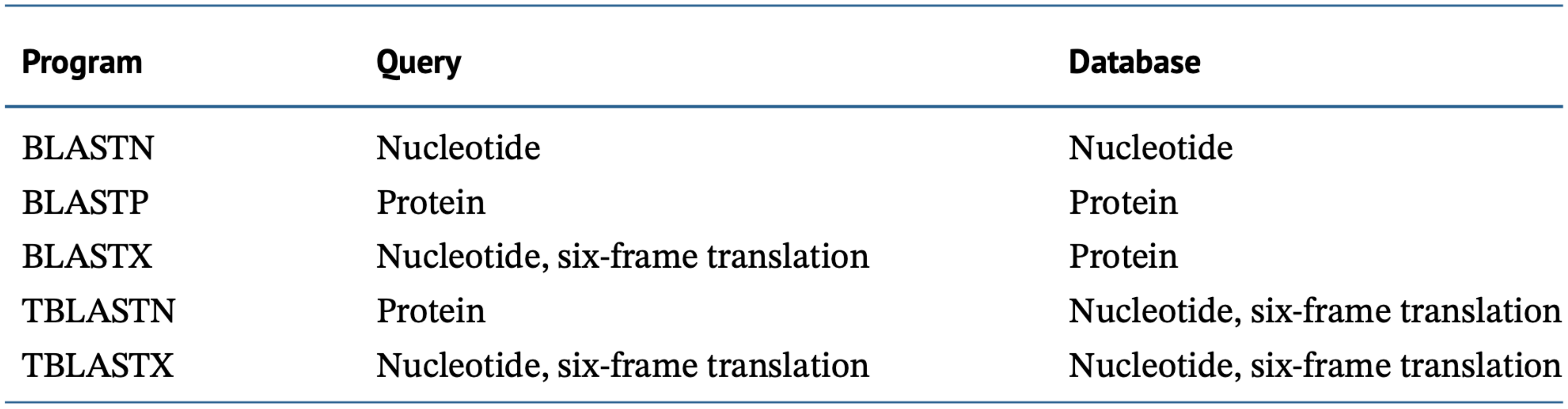
- Basic Local Alignment Search Tool (Altschul et al. 1990) in Perusal.

- BLAST heuristically finds ***high-scoring segment pairs*** (**HSPs**):
    - ➢ Identical length segments from 2 sequences with statistically significant match scores

- Not only the best region of local alignment, but also whether there are other plausible alignments.

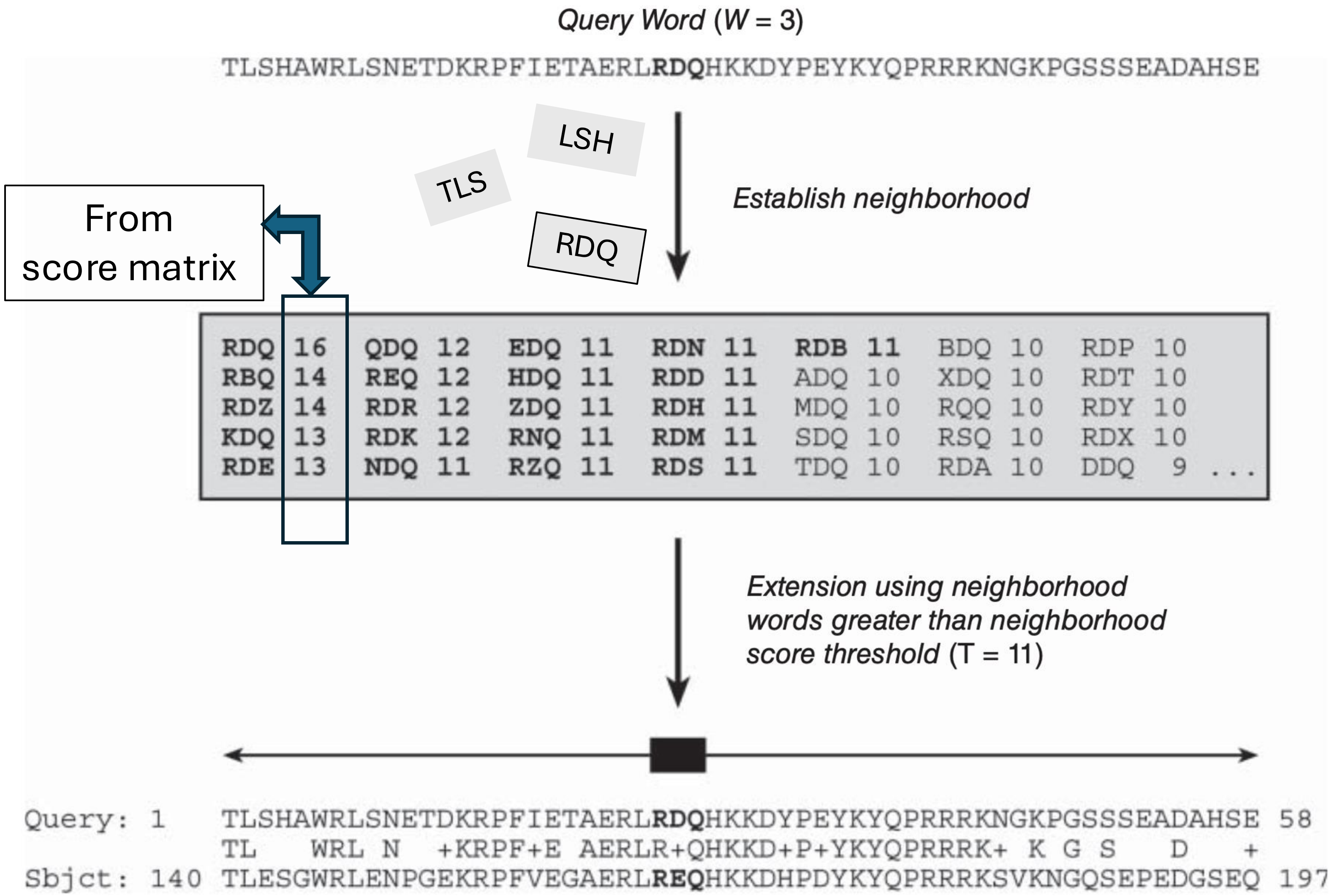
- BLAST method begins by seeding the search with a small subset of letters (query **word**).

BLAST algorithms



Program	Query	Database
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

The initiation of a BLAST search


Query Word (W = 3)
LSH
TLS
RDQ
Establish neighborhood
From score matrix
Extension using neighborhood words greater than neighborhood score threshold (T = 11)

```
TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE
```

RDQ	16	QDQ	12	EDQ	11	RDN	11	RDB	11	BDQ	10	RDP	10
RBQ	14	REQ	12	HDQ	11	RDD	11	ADQ	10	XDQ	10	RDT	10
RDZ	14	RDR	12	ZDQ	11	RDH	11	MDQ	10	RQQ	10	RDY	10
KDQ	13	RDK	12	RNQ	11	RDM	11	SDQ	10	RSQ	10	RDX	10
RDE	13	NDQ	11	RZQ	11	RDS	11	TDQ	10	RDA	10	DDQ	9 ...

```
Query: 1   TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
           TL   WRL N  +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S   D   +
Sbjct: 140 TLESGWRLENPGEKRPFVEGAERLREQHKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197
```

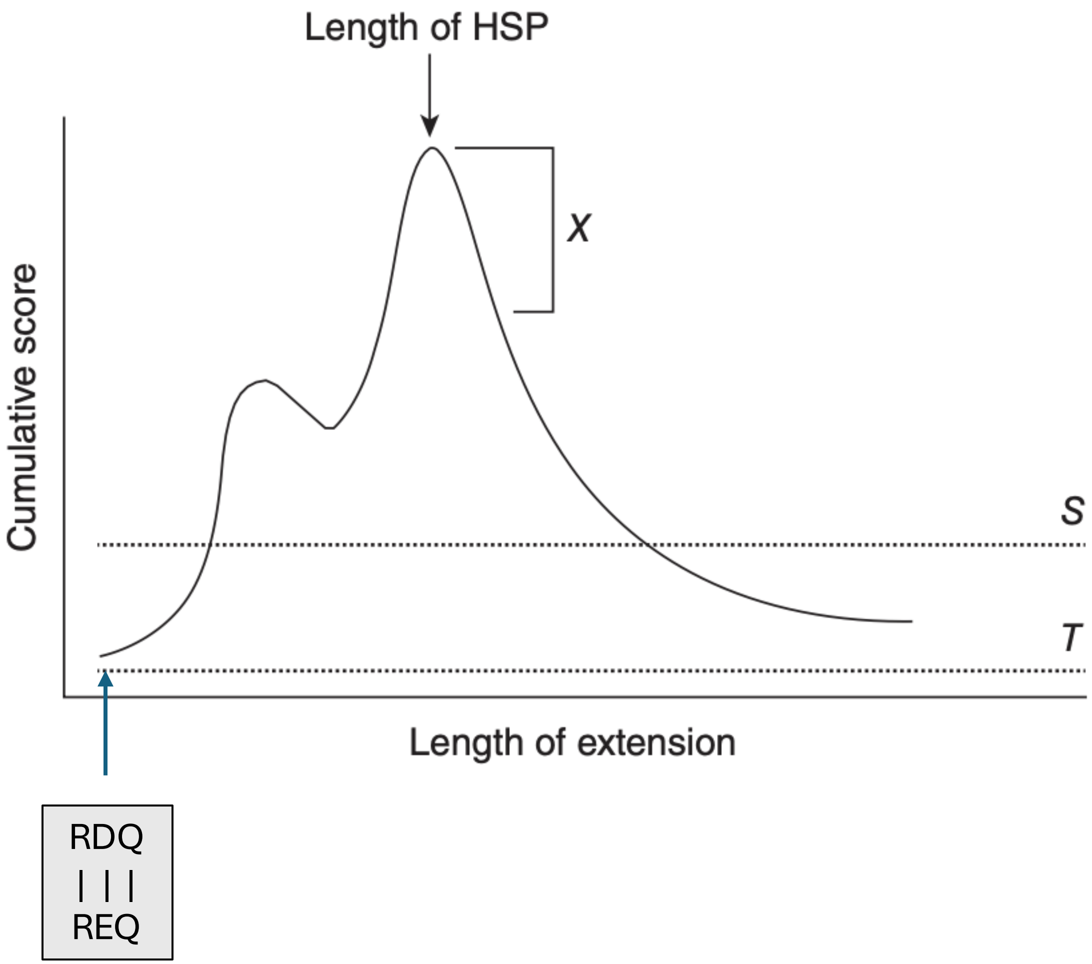
- **Query word** (w=3) is RDQ
- TLS, LSH, ...RDQ

- BLAST search RDQ in the DB, but also related words (with conservative substitutions)
- **Neighborhood**: related words using scoring matrix

- We use a cut-off in the neighborhood: **score threshold** (T)

- Once the query (RDQ) word is aligned with another word (REQ) from DB (with T > 11).
- Now BLAST attempt to extend the alignment in both directions.

How to determine the maximal length of extension



- The number of residues (**extension**) is plotted vs. the **cumulative score** from the alignment.

- As the alignment gets extended:

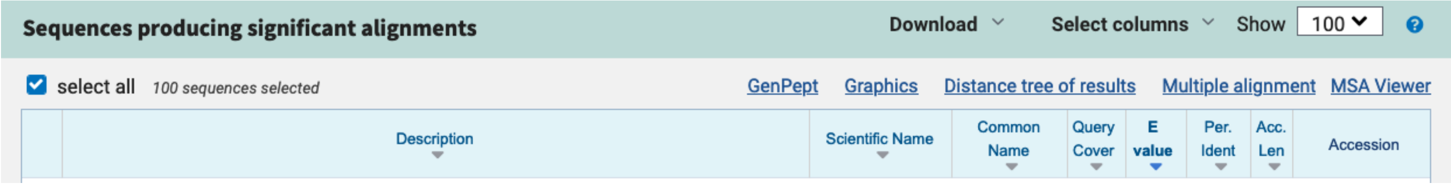
    - matches with a positive score (conservative substitutions): score increase
    - Mismatches and gaps: score decrease

- As soon as the cumulative score breaks the **score threshold S**, the alignment is reported in the BLAST output.

- The resulting alignment is called a high-scoring segment pair, or HSP.

T: neighborhood score threshold
S: minimum score to return a BLAST hit
X: significance decay

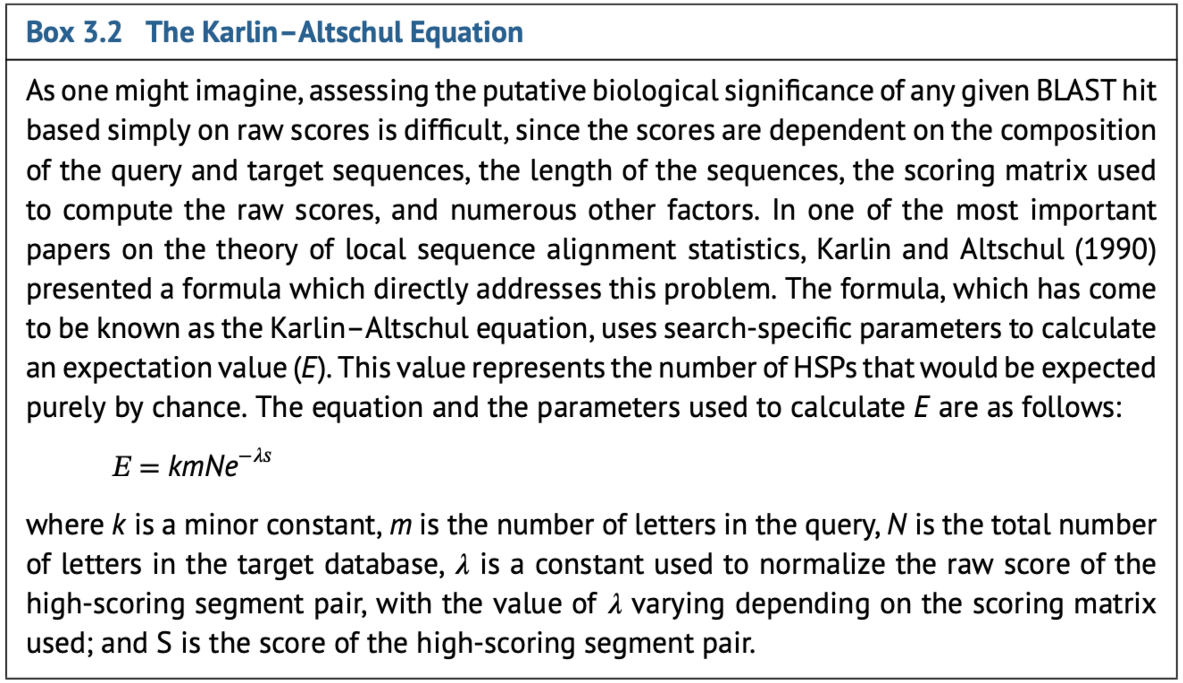
After HSP identification



Sequences producing significant alignments   Download   Select columns   Show 100

☑ select all   *100 sequences selected*   GenPept   Graphics   Distance tree of results   Multiple alignment   MSA Viewer

	Description	Scientific Name	Common Name	Query Cover	E value	Per. Ident	Acc. Len	Accession

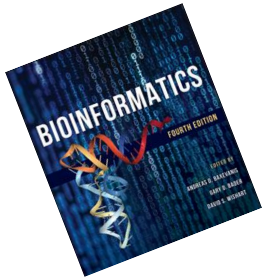


Box 3.2 The Karlin–Altschul Equation

As one might imagine, assessing the putative biological significance of any given BLAST hit based simply on raw scores is difficult, since the scores are dependent on the composition of the query and target sequences, the length of the sequences, the scoring matrix used to compute the raw scores, and numerous other factors. In one of the most important papers on the theory of local sequence alignment statistics, Karlin and Altschul (1990) presented a formula which directly addresses this problem. The formula, which has come to be known as the Karlin–Altschul equation, uses search-specific parameters to calculate an expectation value (*E*). This value represents the number of HSPs that would be expected purely by chance. The equation and the parameters used to calculate *E* are as follows:

$$E = kmNe^{-\lambda s}$$

where *k* is a minor constant, *m* is the number of letters in the query, *N* is the total number of letters in the target database, λ is a constant used to normalize the raw score of the high-scoring segment pair, with the value of λ varying depending on the scoring matrix used; and S is the score of the high-scoring segment pair.

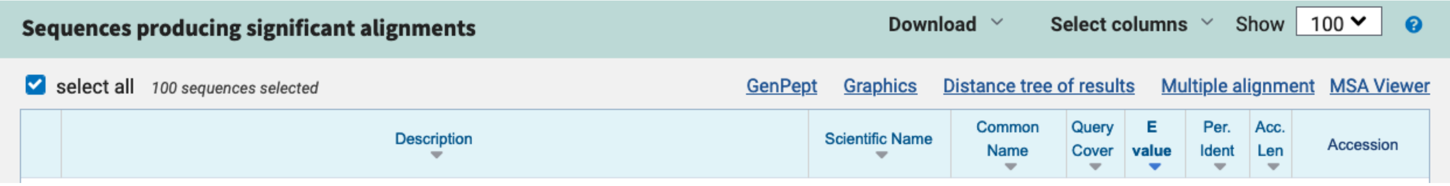


- How to know if the alignment is significant?

- The cumulative score is used, along other parameters, to get a new value: ***E*** value :

  - ➢ ***E*** gives the number of expected HSPs having a score of S or more that BLAST would find purely by chance.
  - ➢ ***E*** provides a measure of whether an HSP is a false positive.
  - ➢ Lower ***E*** values: greater biological significance.
  - ➢ Higher ***E*** values: more likely to get a hit like that by chance.

After HSP identification



Sequences producing significant alignments Download ⌄ Select columns ⌄ Show 100 ⌄ ?

☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Common Name	Query Cover	E value	Per. Ident	Acc. Len	Accession

- How to know if the alignment is significant?

- The cumulative score is used, along other parameters, to get a new value: ***E*** value :

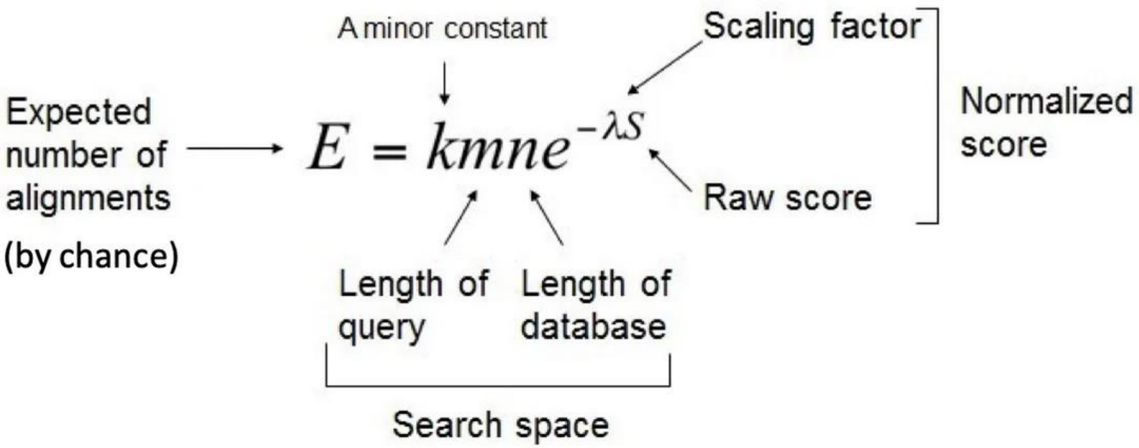
  - ***E*** gives the number of expected HSPs having a score of S or more that BLAST would find purely by chance.

  - ***E*** provides a measure of whether an HSP is a false positive.

  - Lower ***E*** values: greater biological significance.

  - Higher ***E*** values: more likely to get a hit like that by chance.

***E** (HSP with Score > S), also called **E** value.*
*The expected value (**E**) is a parameter that describes the number of hits one can expect to see by chance when searching in a particular database.*



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

Expected number of alignments (by chance)

A minor constant

Scaling factor

Raw score

Normalized score

Length of query

Length of database

Search space

After HSP identification

E value: 1
Expect to get 1 **false positive** in a DB of this size given the alignment score

E value: 0.01
Expect to get 0.01 **false positive** in a DB of this size given the alignment score

- How to know if the alignment is significant?

- The cumulative score is used, along other parameters, to get a new value: ***E*** value :

  - ***E*** gives the number of expected HSPs having a score of S or more that BLAST would find purely by chance.

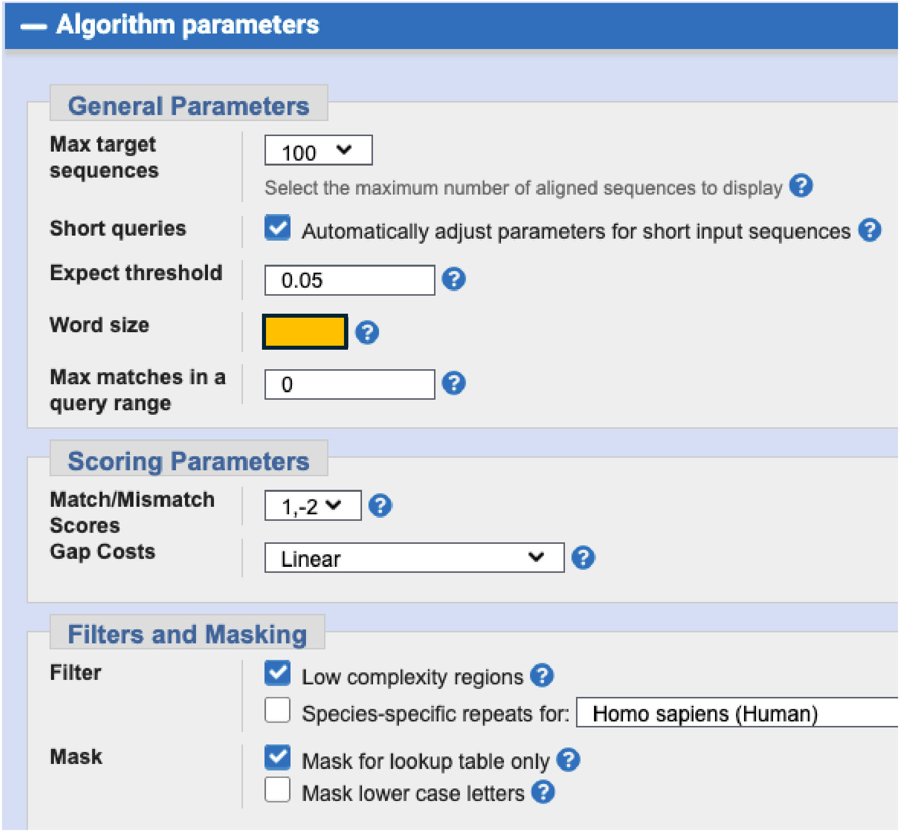
  - ***E*** provides a measure of whether an HSP is a false positive.

  - Lower ***E*** values: greater biological significance.

  - Higher ***E*** values: more likely to get a hit like that by chance.

BLASTN

- Nucleotide query
- Nucleotide database
- BLASTN uses larger word size than protein searches
  - Default word size is .... nucleotides?
  - Increases the likelihood of unique strings
  - Reduces the DB sequences considered



— Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 0.05

Word size: 🟨

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

Filters and Masking

Filter: ☑ Low complexity regions
☐ Species-specific repeats for: Homo sapiens (Human)

Mask: ☑ Mask for lookup table only
☐ Mask lower case letters

BLASTP

- Protein query/database

- Smaller word size:

  - ➢ The default word-size is 5

  - ➢ Amino acid sequences are more likely to be unique than nucleotide sequences(20 *vs*. 4 residues)

  - ➢ Uses substitution matrices (PAM, BLOSUM) to identify more unique w-size strings (neighborhood table).

  - ➢ If word-size decreases: **increase sensitivity / decrease selectivity**
    - ✓ More probable to find distant matches
    - ✓ More time to compute (more hits)

  - ➢ If word-size increase: **increase selectivity / decrease sensitivity**
    - ✓ More probable to find close matches only
    - ✓ Increase compute speed

PSI-BLAST

- Position-Specific Iterated BLAST is particularly well suited for identifying **distantly related proteins** – proteins that may not have been found using the traditional BLASTP method.

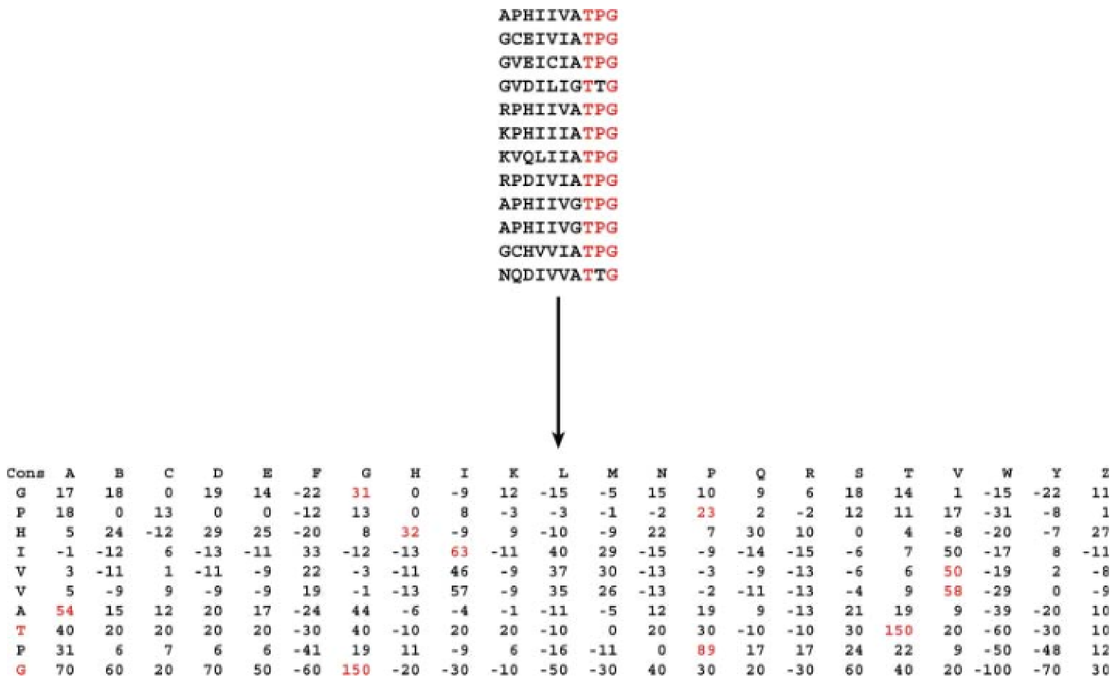
- Use results from BLAST (above specific *E* value) to construct a consensus (align BLAST hits into a multiple sequence alignment - MSA).

- PSI-BLAST relies on the use of position-specific scoring matrices (PSSMs)

- Scoring in PSSMs depends on the frequencies of residues in a specific column of the MSA (not from scores of PAM/BLOSUM)

- Search database with PSSMs matrix (instead of the original query)

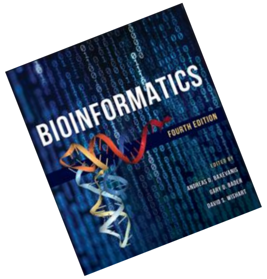
PSI-BLAST



```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

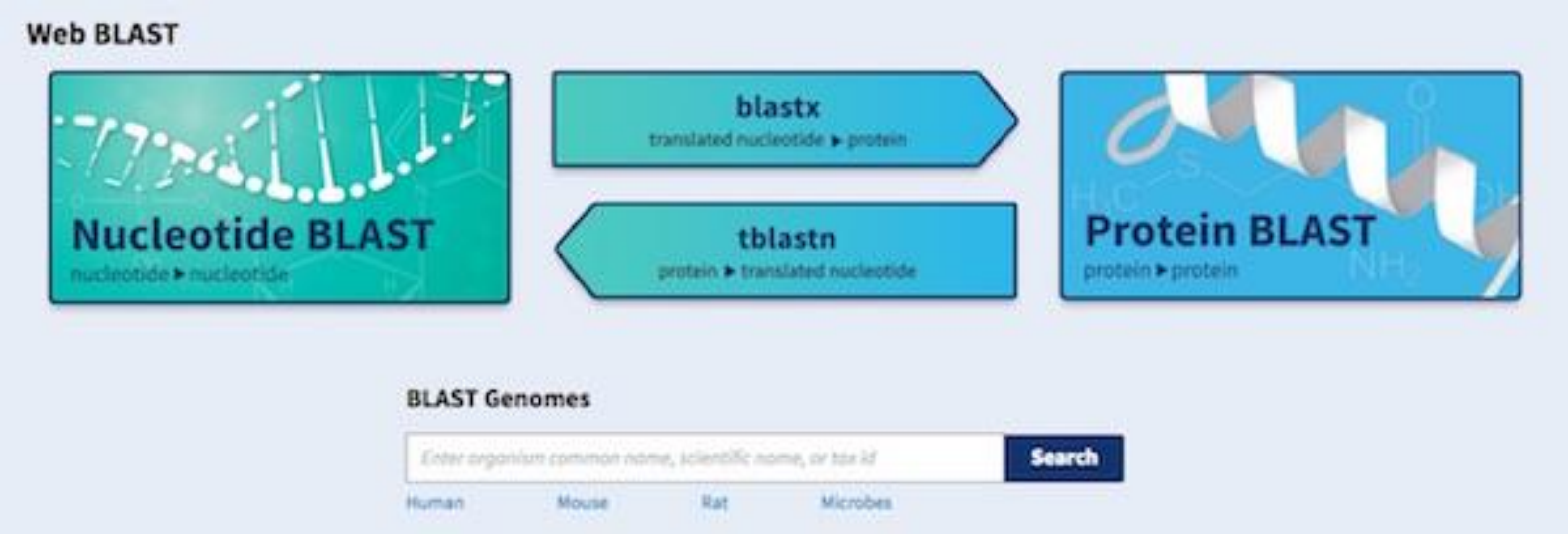
Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

Figure 3.13 Constructing a position-specific scoring matrix (PSSM). In the upper portion of the figure is a multiple sequence alignment of length 10. Using the criteria described in the text, the PSSM corresponding to this multiple sequence alignment is shown in the lower portion of the figure. Each row of the PSSM corresponds to a column in the multiple sequence alignment. Note that position 8 of the alignment always contains a threonine residue (T), whereas position 10 always contains a glycine (G). Looking at the corresponding scores in the matrix, in row 8, the threonine scores 150 points; in row 10, the glycine also scores 150 points. These are the highest values in the row, corresponding to the fact that the multiple sequence alignment shows absolute conservation at those positions. Now, consider position 9, where most of the sequences have a proline (P) at that position. In row 9 of the PSSM, the proline scores 89 points – still the highest value in the row, but not as high a score as would have been conferred if the proline residue was absolutely conserved across all sequences. The first column of the PSSM provides the deduced consensus sequence.



Heuristic methods

- FASTA, BLAST are approximations. Remember, we are trying to find a needle in a haystack.
- Based on heuristics
- If search don’t “seed” well, it can miss evolutionary close sequences
- But they are useful for practical purposes



Heuristic methods

•FASTA begins the search by looking for exact matches of words, while BLAST allows for conservative substitutions in the first step.

•BLAST allows for automatic masking of sequences, while FASTA does not.

•FASTA will return one and only one alignment for a sequence in the hit list, while BLAST can return multiple results for the same sequence, each result representing a distinct HSP.

•Since FASTA uses a version of the more rigorous Smith–Waterman alignment method, it generally produces better final alignments and is more apt to find distantly related sequences than BLAST. For highly similar sequences, their performance is fairly similar.

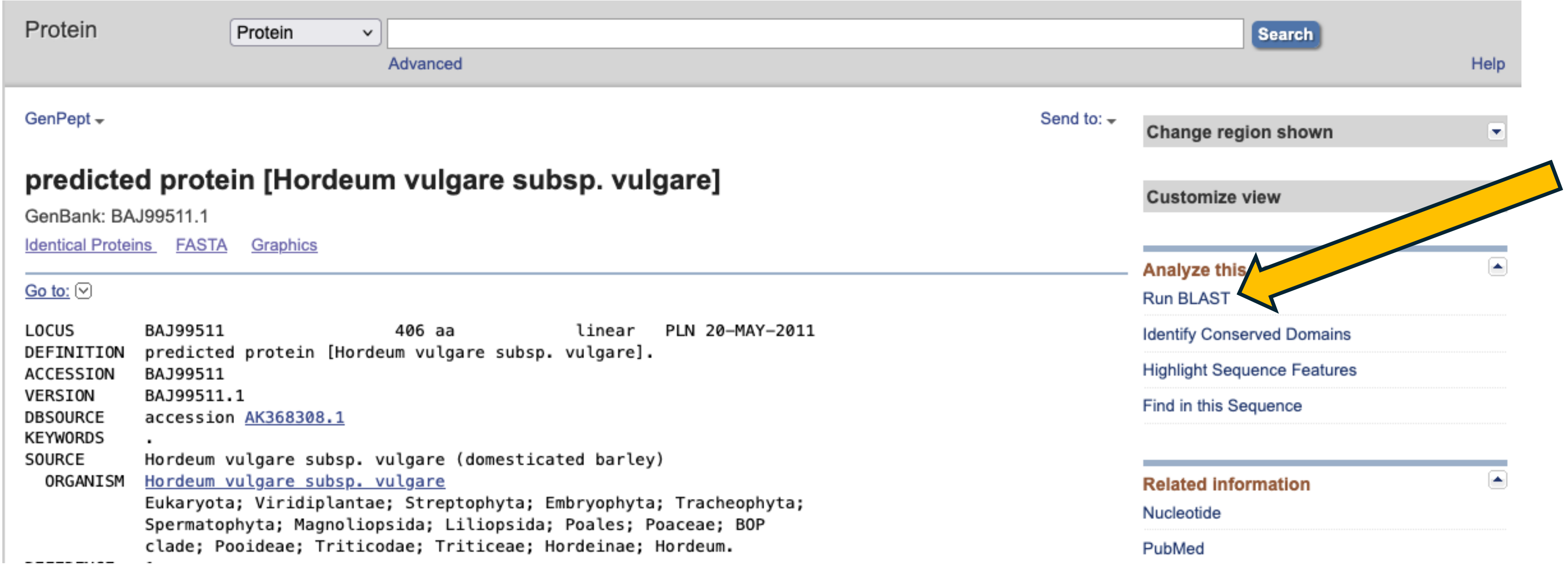
•BLAST runs faster than FASTA, since FASTA is more computationally intensive.

What database to search?
Search the smallest comprehensive database likely to contain your protein

Quick Discussion!

BLASTP
Query ID BAJ99511.1 (Barley)

Source: Wikipedia



Protein

Protein ▾ [ ] Search

Advanced

Help

GenPept ▾

Send to: ▾

Change region shown ▾

Customize view

predicted protein [Hordeum vulgare subsp. vulgare]

GenBank: BAJ99511.1

Identical Proteins FASTA Graphics

Go to: ⌄

```
LOCUS       BAJ99511                 406 aa            linear   PLN 20-MAY-2011
DEFINITION  predicted protein [Hordeum vulgare subsp. vulgare].
ACCESSION   BAJ99511
VERSION     BAJ99511.1
DBSOURCE    accession AK368308.1
KEYWORDS    .
SOURCE      Hordeum vulgare subsp. vulgare (domesticated barley)
  ORGANISM  Hordeum vulgare subsp. vulgare
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; Liliopsida; Poales; Poaceae; BOP
            clade; Pooideae; Triticodae; Triticeae; Hordeinae; Hordeum.
```

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Related information

Nucleotide

PubMed

Run blastp

- What hits do you get?
- Taxonomy?
- Are they biologically significant?