Introduction to BLAST

BIO392

BLAST is an algorithm used for comparing biological sequences, either amino-acids for comparing proteins or nucleotides for DNA and RNA. It can help find similar sequences.

Why do we use sequence similarity search tools

- Find the function of an unknown protein by comparing with very similar proteins
- Check the specificity of primers and probes *in-silico*
- Select data for phylogenic tree construction as well as define a related but different sequence as outgroup
- Identify host contamination in metagenomic data
- Describe the taxonomic profile of viral metagenomes

Where can we find sequence information?

Nucleic acid database

INSDC (International Nucleotide Sequence Database Collaboration)

DDBJ ENA NCBI

These databases are synchronised meaning that they share the same information after synchronisation.

Protein database

Uniprot

EMBL-EBI PIR SIB

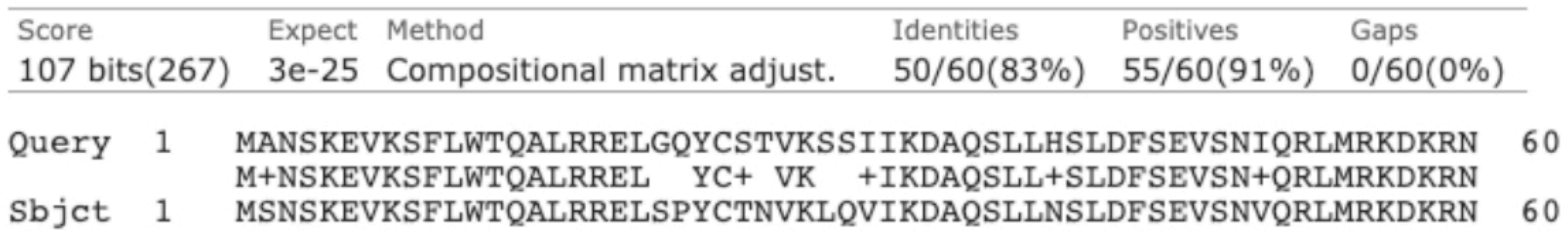
- **Unreviewed entries:** Presents the sequence in the way it was submitted, plus automatic annotation. There are many errors or missing features, and these data are poorly updated.
- **Manually reviewed entries**: The sequence annotation has been curated by reviewers, with addition of biological knowledge. These data are updated.

Databases	**Unreviewed data**	**Manually reviewed data**
Nucleotides	INSDC (GenBank, EMBL-EBI, DDBJ)	NCBI Reference Sequences (RefSeq)
Proteins	UniprotKB/TrEMBL	UniprotKB/Swiss-Prot

Similarity & Homology

Similarity

- It refers to the "likeness" or percentage of identity between 2 sequences
- It can be quantified by calculating a shared statistically significant number of bases or amino acids



Score	Expect	Method	Identities	Positives	Gaps
107 bits(267)	3e-25	Compositional matrix adjust.	50/60(83%)	55/60(91%)	0/60(0%)

```
Query  1    MANSKEVKSFLWTQALRRELGQYCSTVKSSIIKDAQSLLHSLDFSEVSNIQRLMRKDKRN  60
            M+NSKEVKSFLWTQALRREL  YC+ VK  +IKDAQSLL+SLDFSEVSN+QRLMRKDKRN
Sbjct  1    MSNSKEVKSFLWTQALRRELSPYCTNVKLQVIKDAQSLLNSLDFSEVSNVQRLMRKDKRN  60
```

The figure above shows an alignment of two protein sequences

Amino acids represent identical amino acids between both sequences. ‘+’ represents two amino acids with similar chemical properties.

50 identical amino acids out of 60 amino acids mean that these sequences are 83% identical.

Homology

- Most of the time, users will perform sequence searches on databases to identify genes that have an evolutionary relationship with the input sequence.

- This is **homology** : two sequences are said to be homologous if they are derived from a common ancestor. So either they are homologous or not.

- Homology usually implies similarity and cannot be quantified

Search algorithms

Exhaustive vs Heuristic Search Strategies

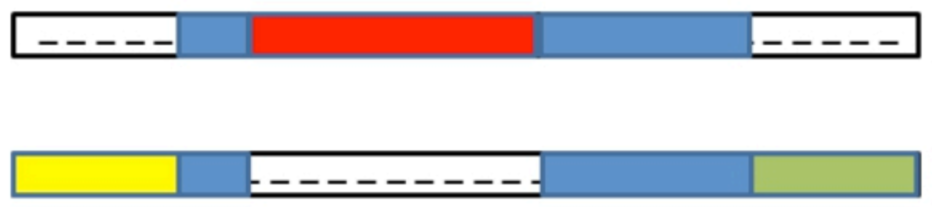
- An **exhaustive search** is a search process enumerating all possible candidates for the solution and checking whether each candidate provides a possible best match.
  - It becomes problematic since the number of comparisons required grow exponentially with the database size.
  - Such as Needleman–Wunsch algorithm (global alignment) and Smith-Waterman algorithm (local alignment)

- A **heuristic search** is to solve a problem in a faster and more efficient fashion, but not necessarily optimal for a difficult optimisation problem.
  - Such as BLAST

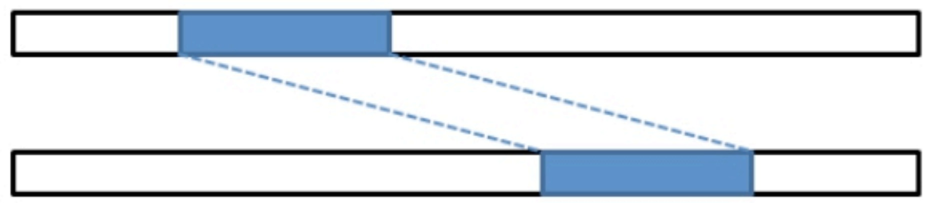
Local vs Global Alignment

- **Global alignment** algorithms considers the entire sequence, adding gaps when necessary.

- **Local alignment** algorithms find the region (or regions) of highest similarity between two sequences regardless of the other lengths of sequences. BLAST is based on local alignment.



Global Alignment



Local Alignment

Question 1

Where can you find nucleotide sequences?

i. NCBI GenBank

ii. Uniprot

Question 2

If you want to find local regions with the highest level of similarity between sequences, which alignment strategy is preferred?

i. Local alignment

ii. Global alignment

Scoring system

Nucleotide

Identity matrix is used to examine the alignment between query and database hit sequence. Each nucleotide identity or mismatch corresponds to a score. The score for each nucleotide is added, resulting in the alignment raw score.

The value itself is meaningless, but allows the comparison of sequence similarity with regards to the query. Therefore the scoring system is not fixed and the user can decide the values for a match or a mismatch.

In this example Match= +1 Mismatch= -3 Gap= -3

```
CAGGTAGCAAGCTTGCATGTCA
|| ||||||||||||  |||||
CACGTAGCAAGCTTG-GTGTCA
```

	A	G	C	T
A	1	-3	-3	-3
G	-3	1	-3	-3
C	-3	-3	1	-3
T	-3	-3	-3	1

The raw score is the sum: 19 (*1) matches - 2 (*3) mismatches and -1 (*3) Gap => 19-6-3= score of 10

Scoring system

Protein

- Unlike nucleotides, mutations in proteins do not all have the same weight in term of functionality.

  For example, an alanine could be replaced by a valine without major consequence, but replacing it with a proline could be disastrous.

- An ideal scoring matrix should reflect the biological phenomena that the alignment seeks to expose.

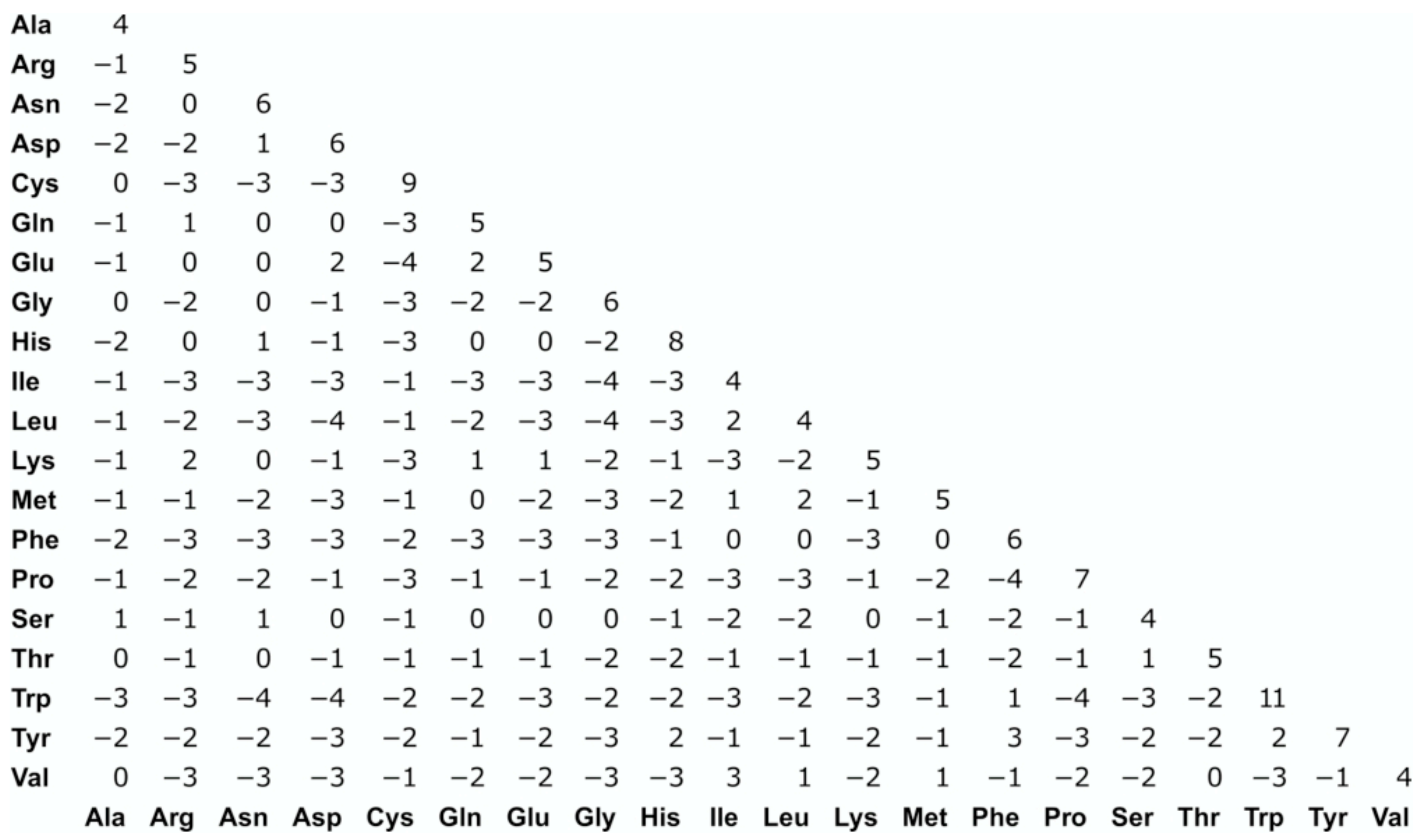
  People use all-purpose matrices called PAM and BLOSUM

BLOSUM (**BLO**cks **SU**bstitution **M**atrix) matrix is a scoring matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences.

- BLOSUM 30, 62, 80, 100

The choice of the matrix used depends on the similarity of the proteins you are considering. To compare closely related sequences, BLOSUM matrices with higher numbers are created, e.g. BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.

BLOSUM62 is BLAST default matrix.



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Scoring system

Gap score

- Gaps indicate an absence of alignment and therefore cannot be scored in terms of similarity. Still, the presence of gaps must be considered when scoring alignments.
- The method used the most in BLAST is called affine gap-penalty. The penalty is composed of two parts: a penalty for the existence of a gap (gap open), and a further length-dependent penalty (gap extension). $O+E^{*}(L-1)$

Final score of an alignment

The quality of the alignment is represented by the score, which is the sum of scores for each position, minus gap penalties. It should be noted that different matrices produce different scores

Question 3

What is the default matrix chosen by BLAST?

i. BLOSUM-80
ii. BLOSUM-62
iii. BLOSUM-45

BLAST= Basic Local Alignment Search Tool

It is a heuristic algorithm based on local alignment

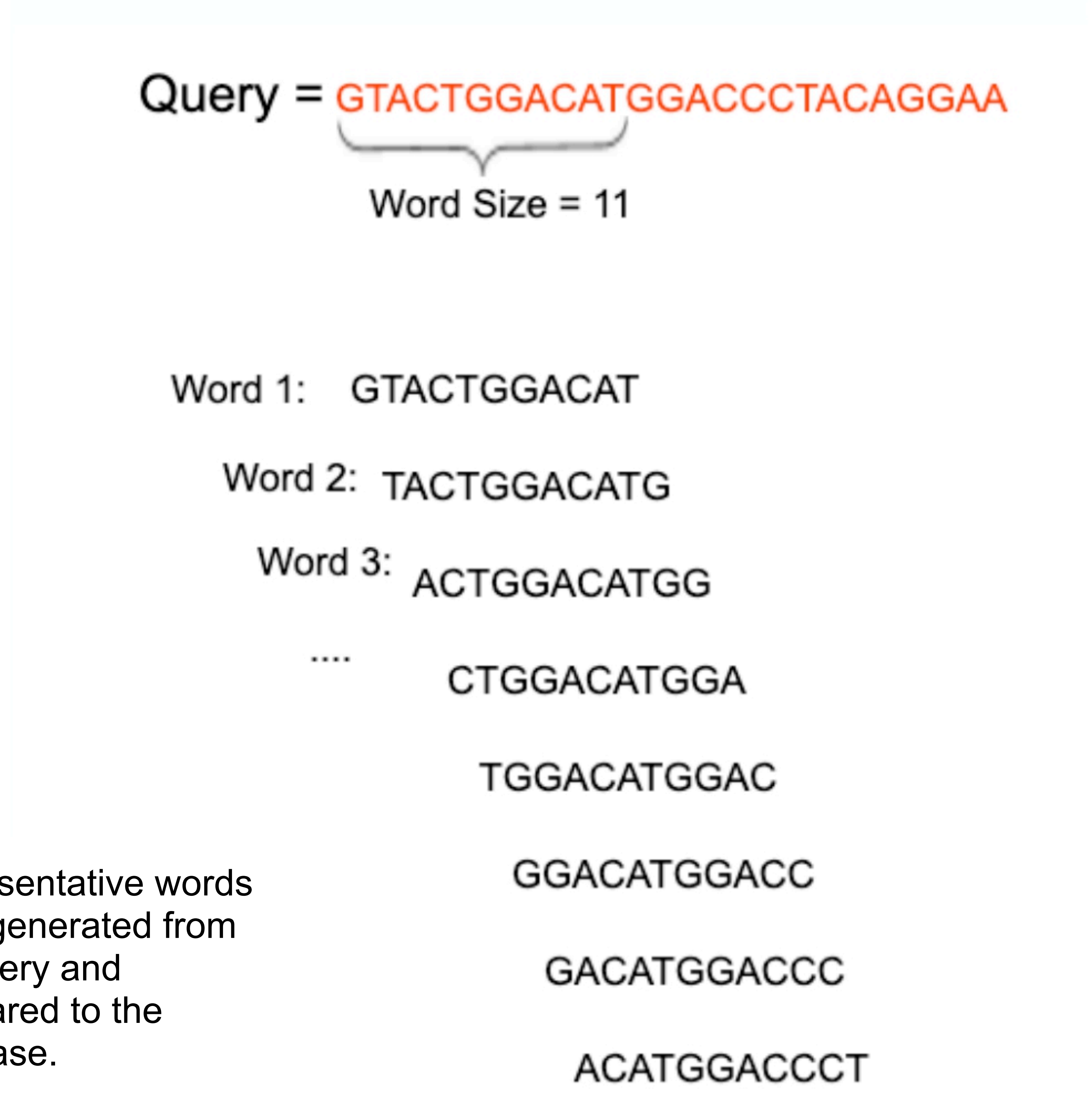
BLAST finds similar sequences by: 1) searching for matching “words” rather than individual residues. 2) using statistics to determine if a match might have occurred by chance

Steps

I. The query sequence is divided into small units, called words
II. Words are matched with database sequences
III. Pairwise alignments are created between matching and query sequences
IV. Each pairwise alignment is scored and the result is sorted on the basis of these scores

Words in BLAST

Nucleotide words



Query = GTACTGGACATGGACCCTACAGGAA

Word Size = 11

Word 1: GTACTGGACAT

Word 2: TACTGGACATG

Word 3: ACTGGACATGG

.... CTGGACATGGA

TGGACATGGAC

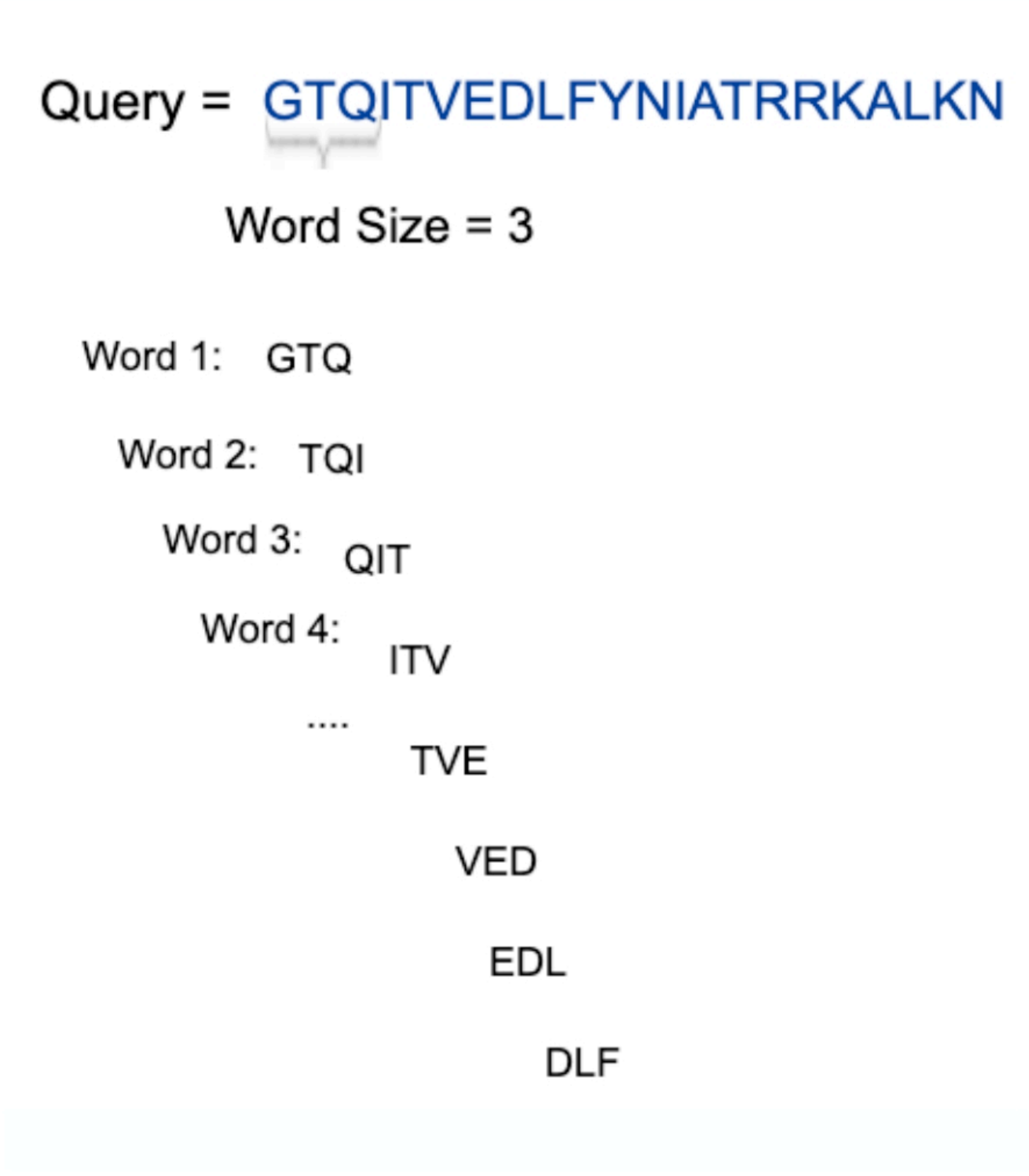
GGACATGGACC

GACATGGACCC

ACATGGACCCT

Representative words were generated from the query and compared to the database.

Protein words



Query = GTQITVEDLFYNIATRRKALKN

Word Size = 3

Word 1: GTQ

Word 2: TQI

Word 3: QIT

Word 4: ITV

.... TVE

VED

EDL

DLF

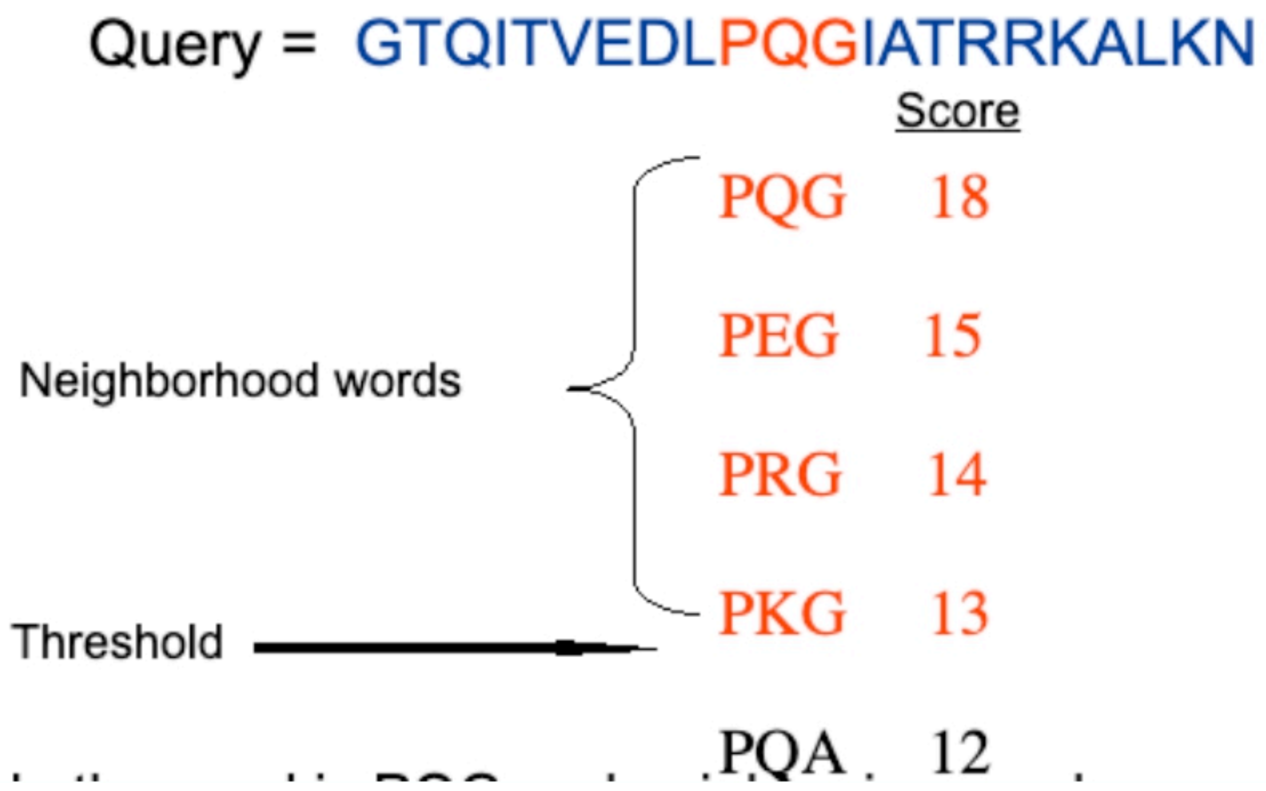
The word size is adjustable

- In BLAST nucleotide, it can be reduced from the default value of 11 to a minimum of 7
- In BLAST protein, it can be reduced from the default value of 3 to a minimum of 2
- The use of short words will increase sensitivity but the task will take longer in that there are more words to compare.

Words in BLAST

Neighborhood words

When comparing two sequences, BLAST searches for exact word matches called word *Hits*. Some alignments do not contain identical words.The neighborhood of a word contains the word itself and all the words whose score is significant when compared to a scoring matrix.



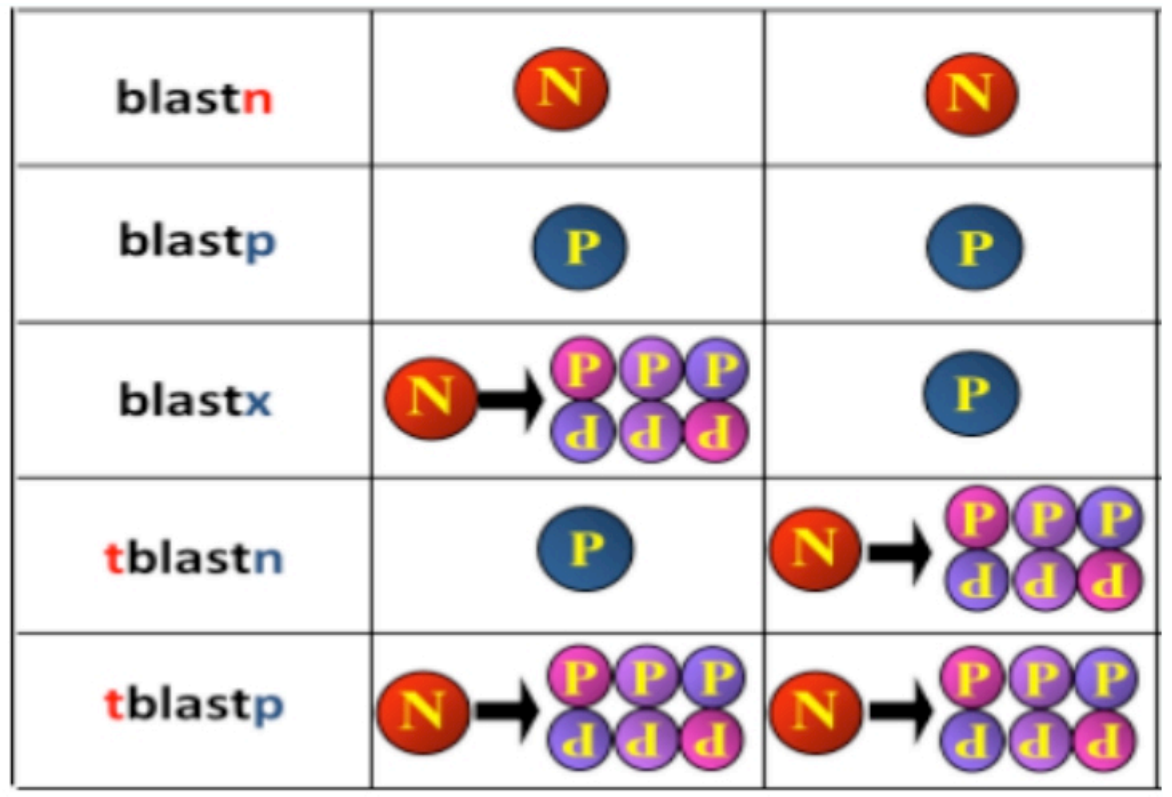
PKG is a neighboring word, PQA is not.

Minimum requirements for a Hit

Nucleotide BLAST requires one exact word match

Protein BLAST requires two neighboring matches within 40 residues

Type of BLAST



blastn	N	N
blastp	P	P
blastx	N → P P P d d d	P
tblastn	P	N → P P P d d d
tblastp	N → P P P d d d	N → P P P d d d

BLASTn: search nucleotide sequences against nucleotide data
Used to find nucleotide similar sequences

BLASTp: search protein sequences against protein data
Find similar protein sequences and information about protein function

BLASTx: search nucleotide sequence to protein data
Used to identify coding regions in a nucleotide sequence

tBLASTn: search protein sequence to nucleotide data
Used to compare a protein sequence on nucleotide data, to find similar proteins even if they have not been annotated

tBLASTx: search translated nucleotide to translated nucleotide data
Used as a gene prediction tool used for unannotated sets of genomes

BLASTn and BLASTp are the most widely used

Question 4

What does the BLAST algorithm search for?

I. Individual nucleotides/amino acids
II. Words

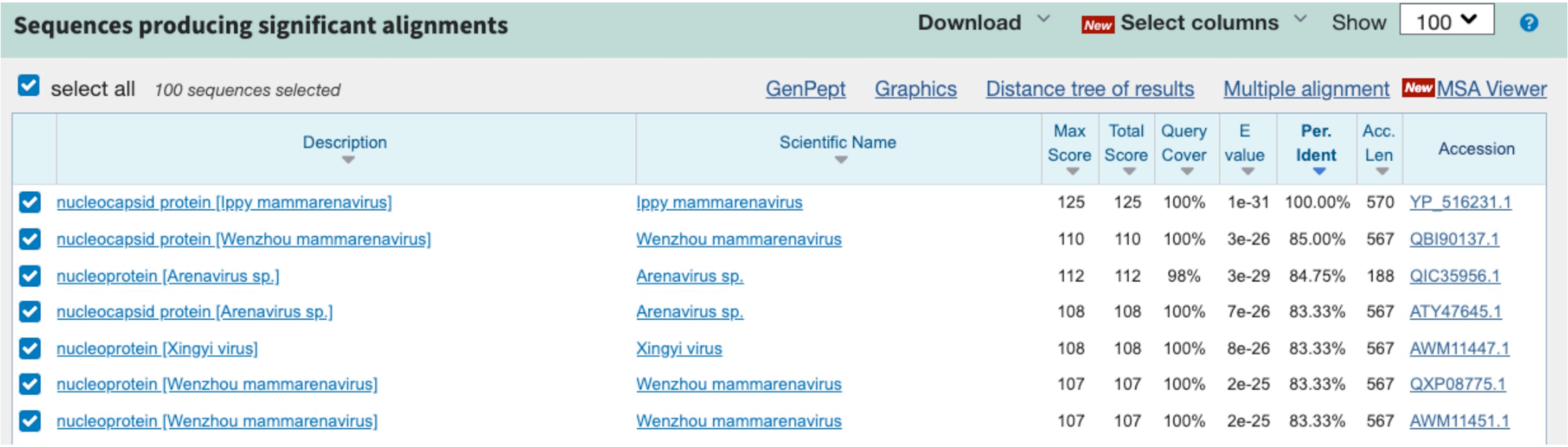
BLAST output interpretation

Expected value (E-value)

- Some amino acids are more common than others and so similarity among them can occur just by statistical chance. The significance of an alignment is given by the **Expected value (E-value).**
- The definition of the E – value is: the number of expected hits of similar quality (score) that could be found just by chance.

  e.g. E-value of 10 means that up to 10 hits can be expected to be found just by chance, given the same size of a random database.
- The typical threshold for a good E-value from a BLAST search is 10^{-5} or lower.
- Database size is taken into account during the E-value calculation. The same search done at different times may therefore give two E-values, if the size of the database has changed between the two searches.

BLAST output interpretation



Sequences producing significant alignments

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	nucleocapsid protein [Ippy mammarenavirus]	Ippy mammarenavirus	125	125	100%	1e-31	100.00%	570	YP_516231.1
☑	nucleocapsid protein [Wenzhou mammarenavirus]	Wenzhou mammarenavirus	110	110	100%	3e-26	85.00%	567	QBI90137.1
☑	nucleoprotein [Arenavirus sp.]	Arenavirus sp.	112	112	98%	3e-29	84.75%	188	QIC35956.1
☑	nucleocapsid protein [Arenavirus sp.]	Arenavirus sp.	108	108	100%	7e-26	83.33%	567	ATY47645.1
☑	nucleoprotein [Xingyi virus]	Xingyi virus	108	108	100%	8e-26	83.33%	567	AWM11447.1
☑	nucleoprotein [Wenzhou mammarenavirus]	Wenzhou mammarenavirus	107	107	100%	2e-25	83.33%	567	QXP08775.1
☑	nucleoprotein [Wenzhou mammarenavirus]	Wenzhou mammarenavirus	107	107	100%	2e-25	83.33%	567	AWM11451.1

Keep an eye on query coverage. A partial similarity may score better than a true protein homolog. Therefore:

- Do not trust the first hit alone.
- Be careful of homology between pathogens and host. For example, viruses and their host are very different organisms and often a protein can have acquired a very different function when moving from one to another.

Question 5

The higher the E-value, the more significant the alignment?

I. Yes
II. No

Reference & Useful Links

SIB e-learning resource (https://viralzone.expasy.org/e_learning/alignments/1/start.html)

Blast in NCBI tutorial (https://www.youtube.com/watch?v=RzC-V67z5LA)

Blast in Uniprot tutorial (https://www.youtube.com/watch?v=UPaConHNP7E)

Exercise

1. Use blast in NCBI to search the unknown nucleotide sequence

    - Which organism does this sequence belong to?
    - Pick one blast result. What is the accession number, max score, query cover and E value?
    - Which region does this sequence cover the subject sequence? (The answer could be different which depends on the accession that you choose)
    - Is it DNA or RNA sequence?
    - Does it encode a (part of) protein? If yes, which protein? (Hint: use different blast type)

Exercise

2. Use blast in Uniprot to search the unknown protein sequence

- Select the most possible one among manually reviewed entries. What is its Uniprot ID?
- What protein does this sequence come from?
- Which organism does this sequence belong to?
- What is the function of this protein?
- What is the variant associated with acute myeloid leukemia (AML) in this protein?

3. If you have more time, play around to feel the difference of blast service from different databases

For example,

- Use Blast in NCBI to query the protein sequence
- Use Blast in Uniprot to query the nucleotide sequence