B.Sc. (Hons.) Biotechnology
Core Course 13:
Basics of Bioinformatics and Biostatistics (BIOT 3013)

Unit 5:
Sequence Alignment and database searching

Dr. Satarudra Prakash Singh
Department of Biotechnology
Mahatma Gandhi Central University, Motihari

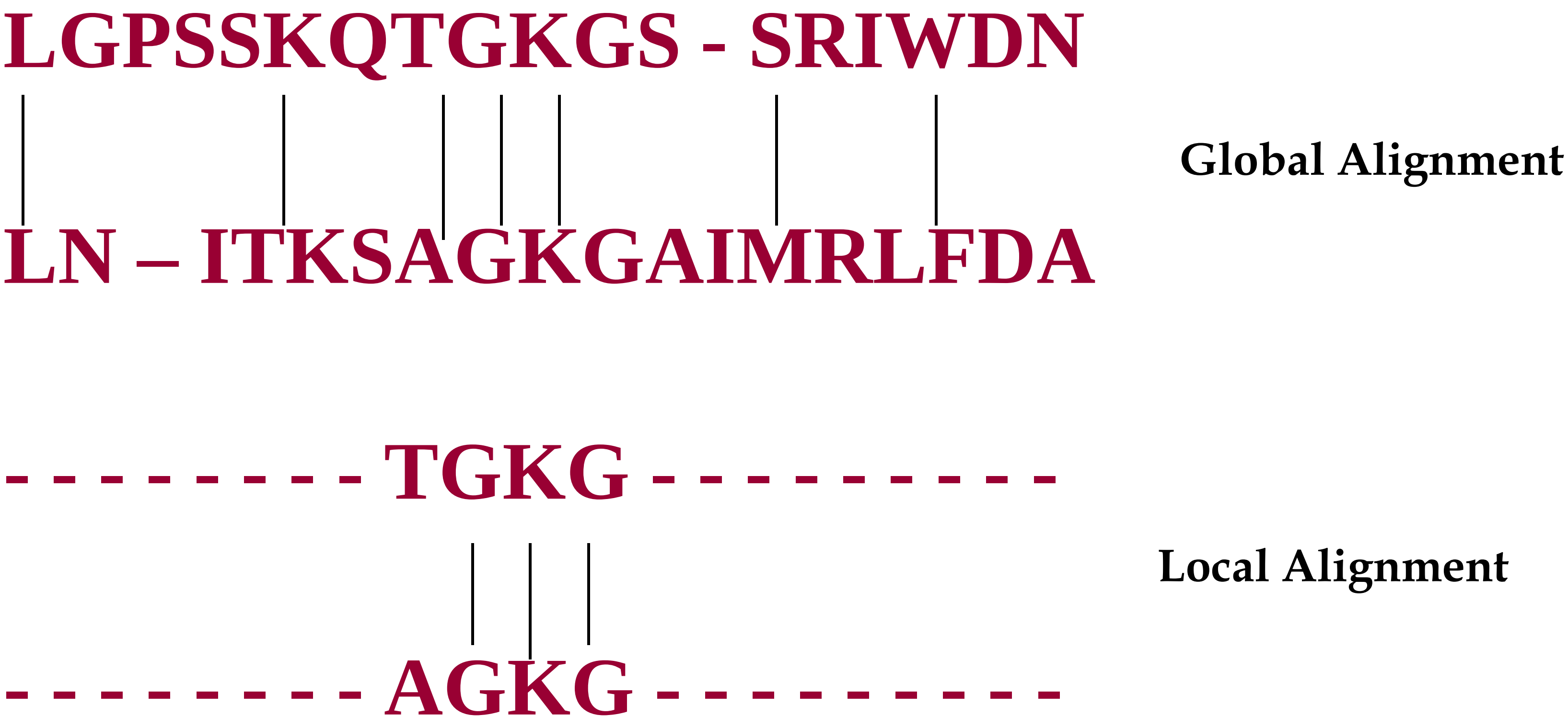
Challenges in bioinformatics

1. Obtain the genome of an organism.
2. Identify and annotate genes.
3. Find the sequences, three dimensional structures, and functions of proteins.
4. Find sequences of proteins that have desired three dimensional structures.
5. Compare DNA sequences and proteins sequences for similarity.
6. Study the evolution of sequences and species.

Sequence alignments lie at the heart of all bioinformatics

Definition of Sequence Alignment

- **Sequence alignment is the procedure of comparing two or more sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.**



LGPSSKQTGKGS - SRIWDN

LN – ITKSAGKGAIMRLFDA

Global Alignment

- - - - - - - - - - TGKG - - - - - - - - - -

- - - - - - - - - - AGKG - - - - - - - - - -

Local Alignment

- In global alignment, an attempt is made to align the entire sequences, as many characters as possible.

- In local alignment, stretches of sequence with the highest density of matches are given the highest priority, thus generating one or more islands of matches in the aligned sequences.

- Eg: problem of locating the famous *TATAAT*-box (a bacterial promoter) in a piece of DNA.

Method for pairwise sequence Alignment: Dynamic Programming

- **Global Alignment:** **Needleman-Wunsch Algorithm**
- **Local Alignment:** **Smith-Waterman Algorithm**

Needleman & Wunsch algorithm : Global alignment

- There are three major phases:

  1. initialization 2. Fill 3. Trace back.
- Initialization assign values for the first row and column.
- The score of each cell is set to the gap score multiplied by the distance from the origin.
- Each cell of the matrix contains two values: a score and an arrow that points up,left,or diagonally up.

Needleman-Wunsch: Global Alignments

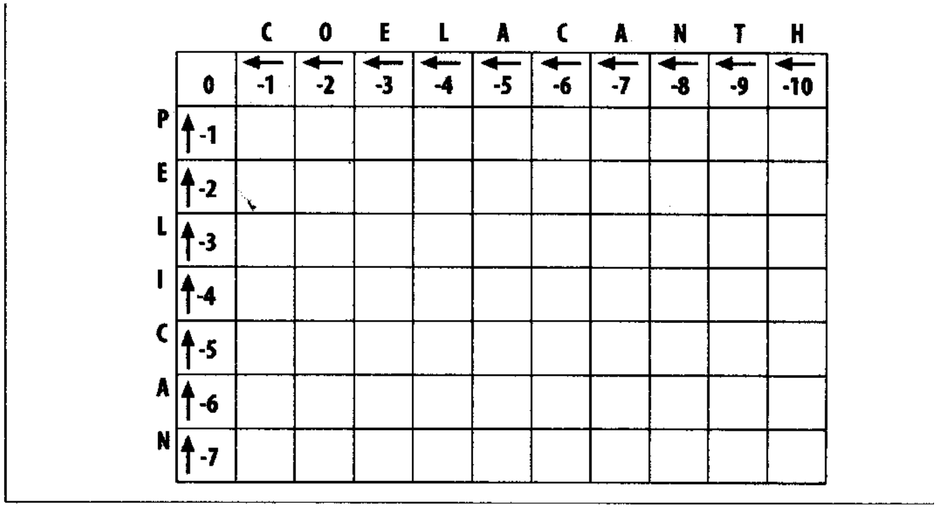
Two sequences

COELACANTH
PELICAN

Scoring scheme

- match = 1
- mismatch = -1
- gap penalty = -1

1. Initialization Phase



		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1										
E	↑ -2										
L	↑ -3										
I	↑ -4										
C	↑ -5										
A	↑ -6										
N	↑ -7										

Figure 3-2. Initialization of the alignment matrix

Scoring scheme

match = 1

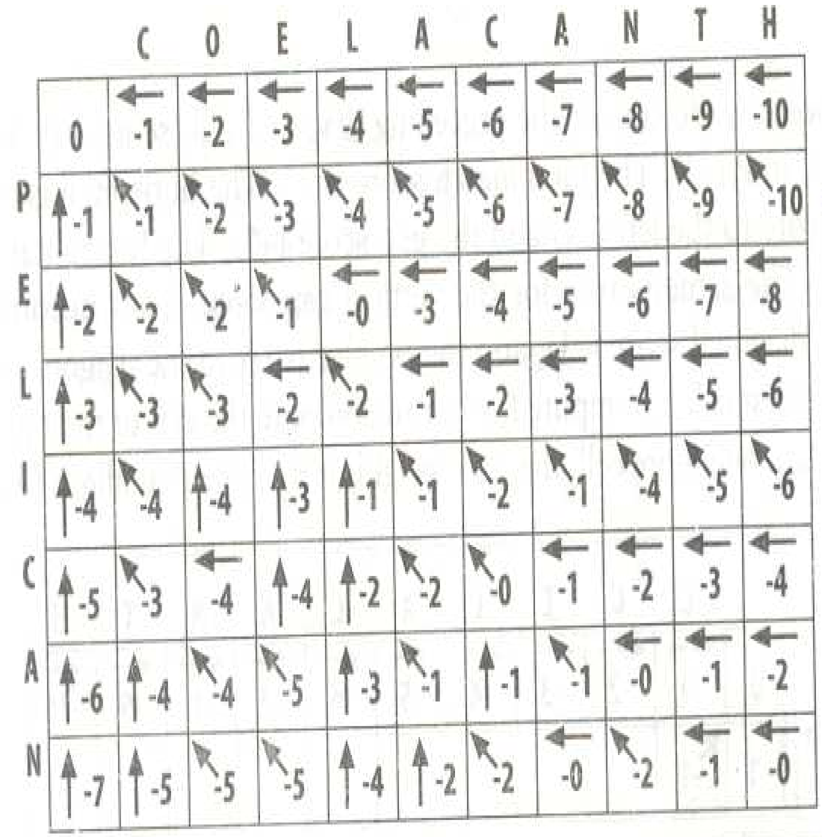
mismatch = -1

gap *g* = -1

- Compute three score for each matrix cell
- Assign max. value to the cell and point the arrow in the direction of the maximum score.
- Make the consistency when two scores are equal (always choose Diagonal vs gap).
- Continue operation until the entire matrix is filled.

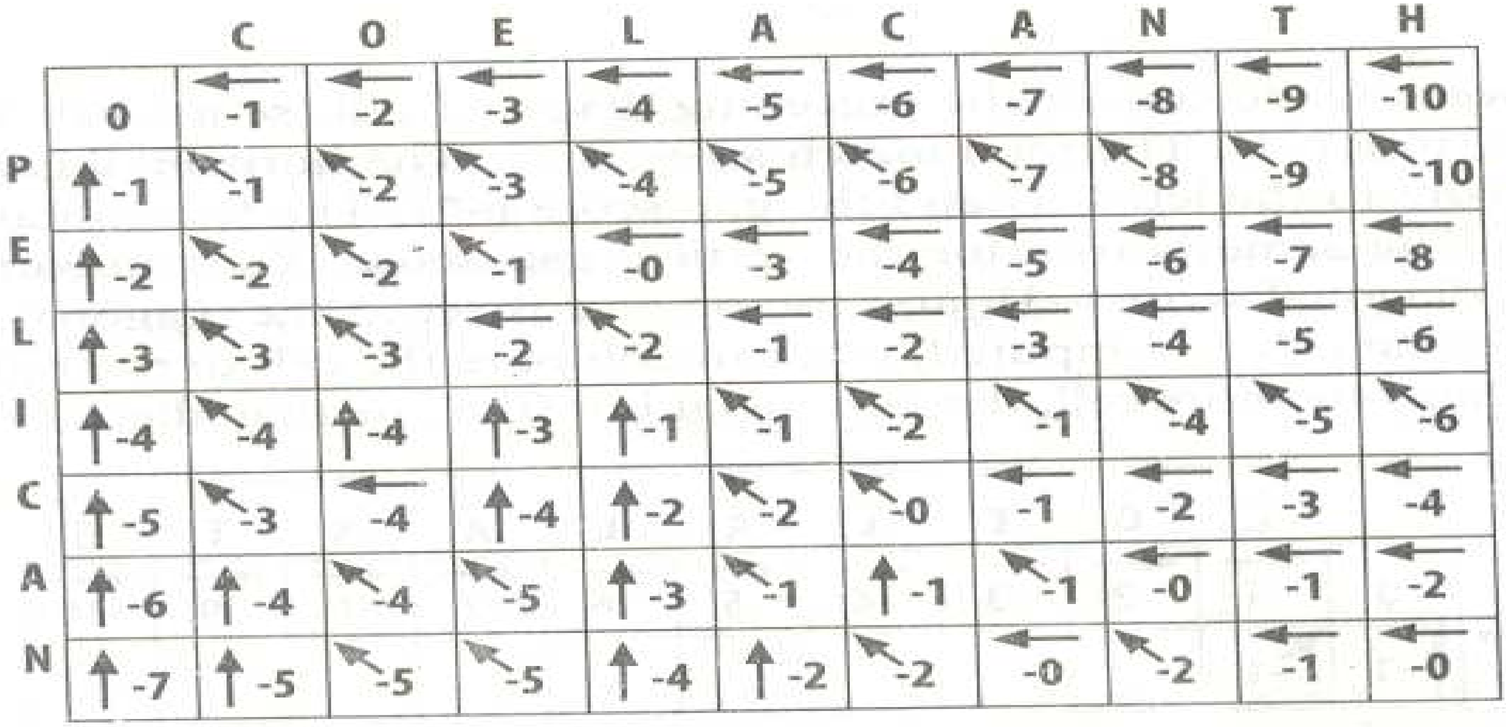
2. Fill Phase

$$M_{ij} = \max \begin{cases} M_{i-1,j-1} + S(c_i, c_j) \\ M_{i,j-1} + g \\ M_{i-1,j} + g \end{cases}$$



		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -0	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	← -2	↖ -2	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↑ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -1	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ -0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ -1	← -0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	← -0	↖ -2	← -1	← -0

3.Trace Back



		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -0	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	← -2	↖ -2	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↑ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -1	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ -0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ -1	← -0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	← -0	↖ -2	← -1	← -0

Globally Aligned Sequence

C O E L A C A N T H
- P E L I C A N - -

Smith-Waterman Algorithm :

Local alignment

- Simple modification of N-W algorithm (Only Four Changes)

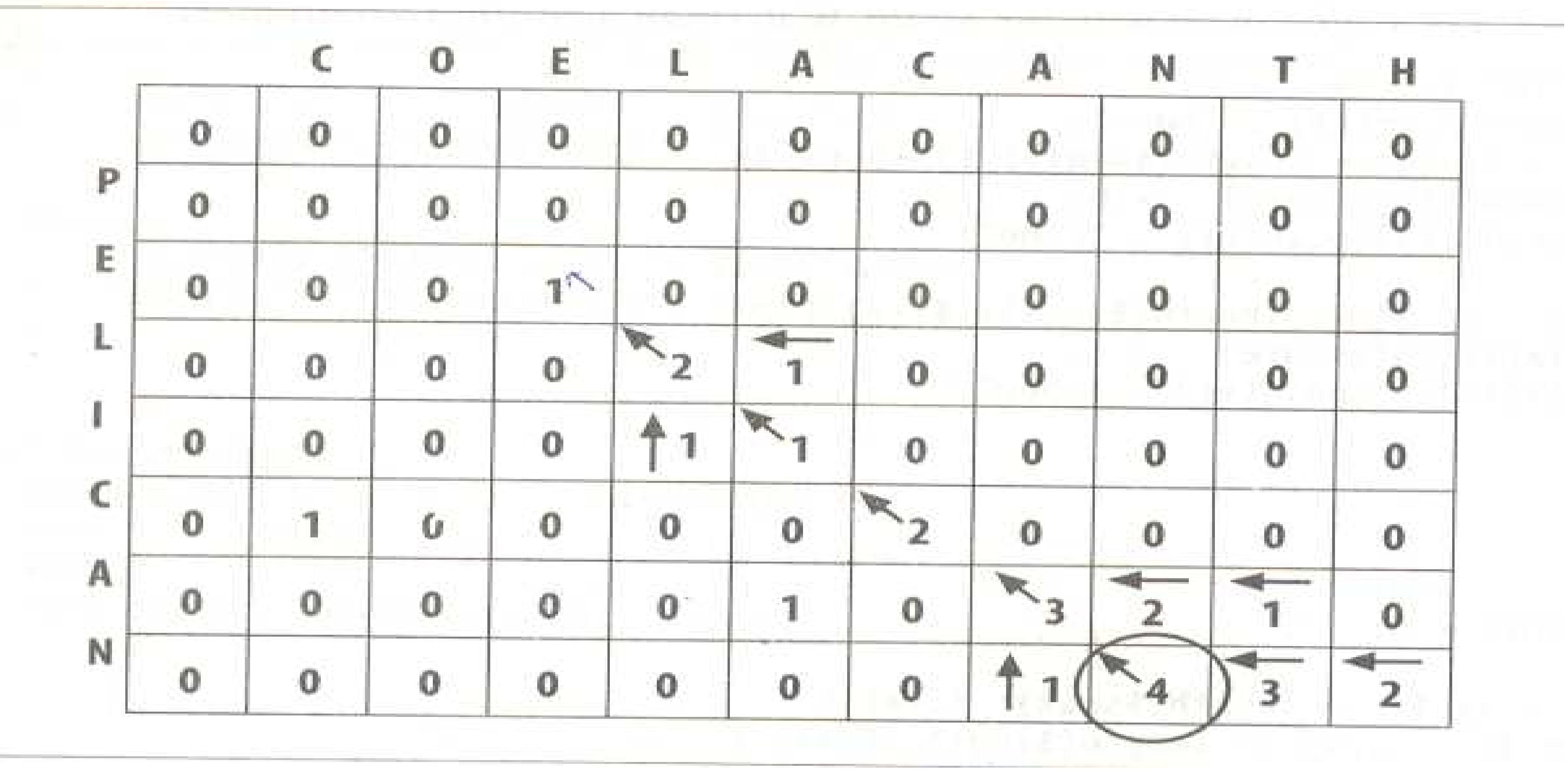
- The edges of the matrix are initialized to 0

- The maximum score is never less than 0 and no arrow is recorded unless the score is greater than 0

- Traceback is started at the highest values rather than at the lower right hand corner.

- Traceback is stopped as soon as a zero is encountered.

Trace Back



		C	O	E	L	A	C	A	N	T	H
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1↖	0	0	0	0	0	0	0
L	0	0	0	0	↖ 2	← 1	0	0	0	0	0
I	0	0	0	0	↑ 1	↖ 1	0	0	0	0	0
C	0	1	0	0	0	0	↖ 2	0	0	0	0
A	0	0	0	0	0	1	0	↖ 3	← 2	← 1	0
N	0	0	0	0	0	0	0	↑ 1	↖ 4	← 3	← 2

Locally Aligned Sequence

ELACAN
ELI CAN

Matrices: Measures of Similarity

- **Every sequence comparison method requires a set of scores.**
- **Thus, the similarity matrices are the basis of sequence analysis methods.**
- **Choice of matrix can influence outcome of analyse.**

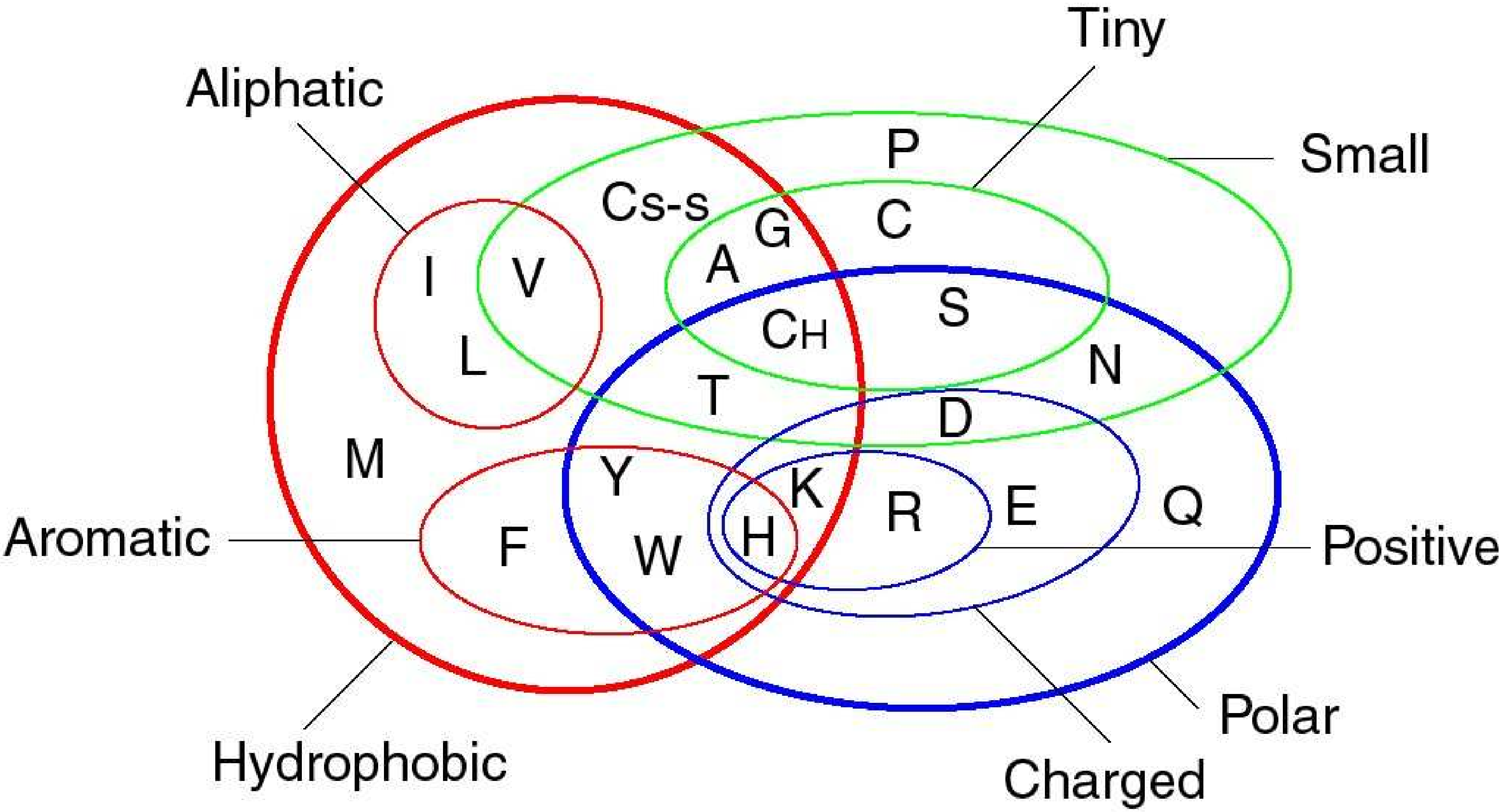
Amino acid substitution matrices

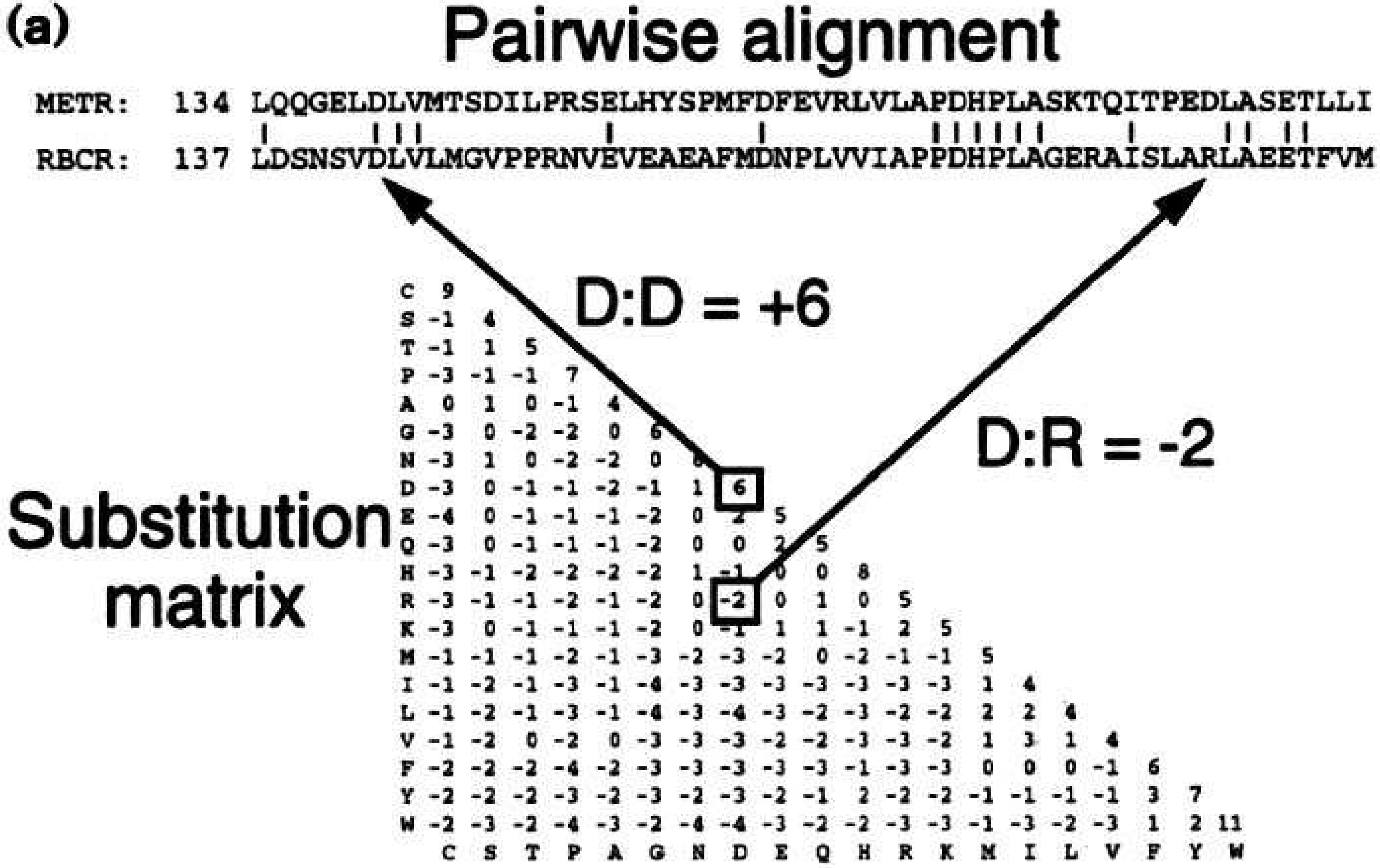
Amino acids are **not** equal:

1. Some are easily substituted because they have similar:
   - physico-chemical properties
   - structure
2. Some mutations between amino acids occur more often due to similar codons

The two above observations give us ways to define *substitution matrices*

Properties of Amino Acids





(a)

Pairwise alignment

```
METR:   134 LQQGELDLVMTSDILPRSELHYSPMFDFEVRLVLAPDHPLASKTQITPEDLASETLLI
            |     |||         |       |        ||||||    |    || ||
RBCR:   137 LDSNSVDLVLMGVPPRNVEVEAEAFMDNPLVVIAPPDHPLAGERAISLARLAEETFVM
```

D:D = +6

D:R = -2

Substitution matrix

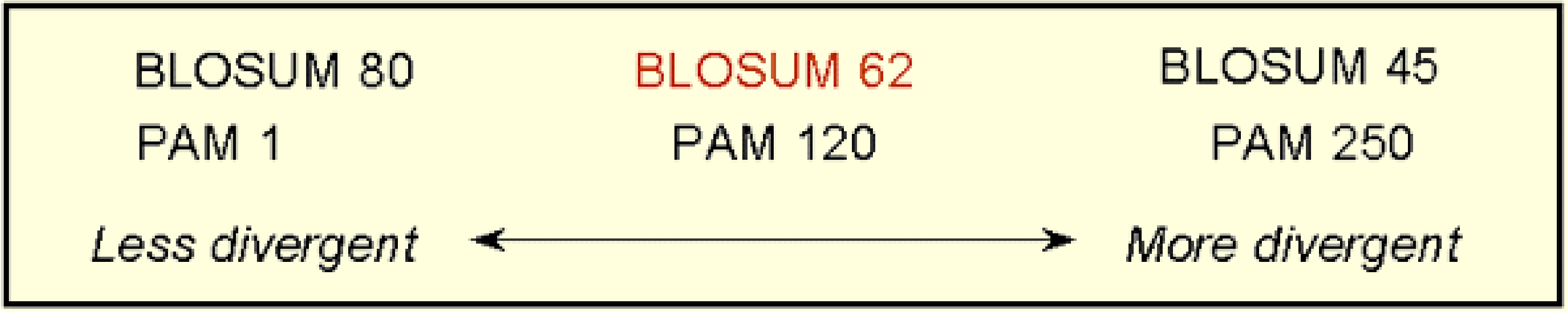
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5									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

Substitution Matrices

- PAM
  - Developed by Margaret Dayhoff and published in 1978
- BLOSUM
  - Developed by Henikoff and Henikoff and published in 1992

The relationship between BLOSUM and PAM substitution matrices

- BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related proteins.
- Overall BLOSUM62 is most effective for local alignment.



BLOSUM 80 | BLOSUM 62 | BLOSUM 45

PAM 1 | PAM 120 | PAM 250

Less divergent ← → *More divergent*

BLOSUM62

	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	

Example

- 1. FGKISESREFDNQNGPSTKDFGKIS

- 2. FGKINMRLEDALVQNQLERSFGKIN
- Matrix: EBLOSUM62
- Gap penalty: 10.0
- Extend penalty: 0.5
- Length: 25
- Identity: 9/25 (36.0%)
- Similarity: 12/25 (48.0%)
- Gaps: 0/25 (0.0%)
- Score: 32.0

Identity & similarity

- The %id is the percentage of identical matches between the two sequences over the reported aligned region.
- The %similarity is the percentage of matches between the two sequences over the reported aligned region where the scoring matrix value is greater or equal to 0.0.

Similarity versus Homology

- Similarity refers to the likeness or % identity between 2 sequences
- Similarity means sharing a statistically significant number of bases or amino acids
- Similarity does not imply homology

- Homology refers to shared ancestry
- Two sequences are homologous is they are derived from a common ancestral sequence
- Homology usually implies similarity

Similarity versus Homology

- Similarity can be quantified
- It is correct to say that two sequences are X% identical
- It is correct to say that two sequences have a similarity score of Z
- It is generally incorrect to say that two sequences are X% *similar*

Difference between Homology and Similarity

- **Since homology is a qualitative description of the relationship, the term “% homology” has no meaning.**
- **Supporting data for a homologous relationship may include sequence or structural similarities, which can be described in quantitative terms.**
  - **% identities, rmsd**

Some Simple Rules

- If two sequence are > 100 residues and > 25% identical, they are likely related
- If two sequences are 15-25% identical they may be related, but more tests are needed
- If two sequences are < 15% identical they are probably not related
- If you need more than 1 gap for every 20 residues the alignment is suspicious

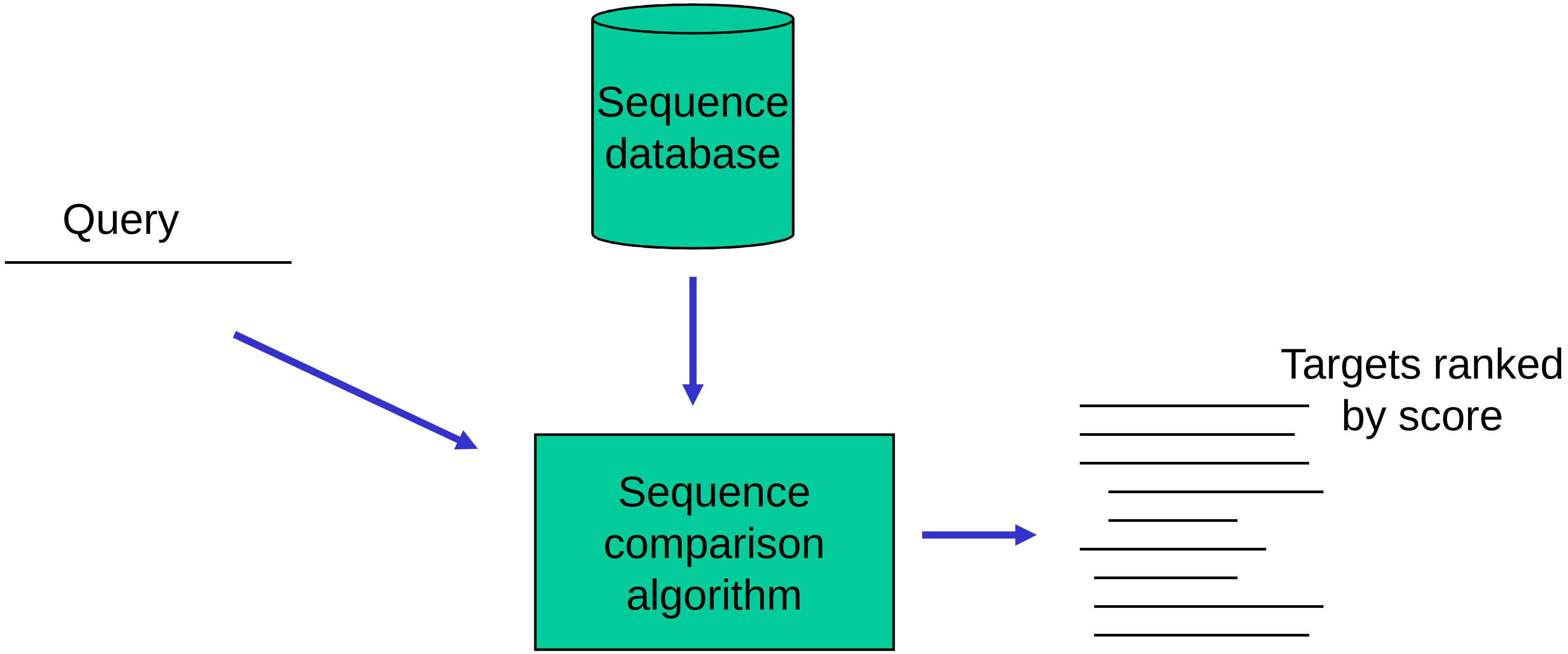
Dynamic Programming

- **Great for doing pairwise global alignments**
- **Produces a quantitative alignment “score”**
- **Problems if one tries to do alignments with very large sequences (memory requirement grows as N^2 or as N x M)**
- **Serious problems if one tries to align one sequence against a database (10’s of hours)**
- **Need an alternative ….. Like BLAST….**

Basic Local Sequence Alignment Tool

- Time complexity of dynamic programming algorithm lead to the development of BLAST algorithms which are significantly faster but do not guarantee to find the optimal alignment.
- BLAST does not explore the entire search space between two sequences.
- Minimizing the search space is the key to its speed but at the cost of a loss in sensitivity.

Database searching



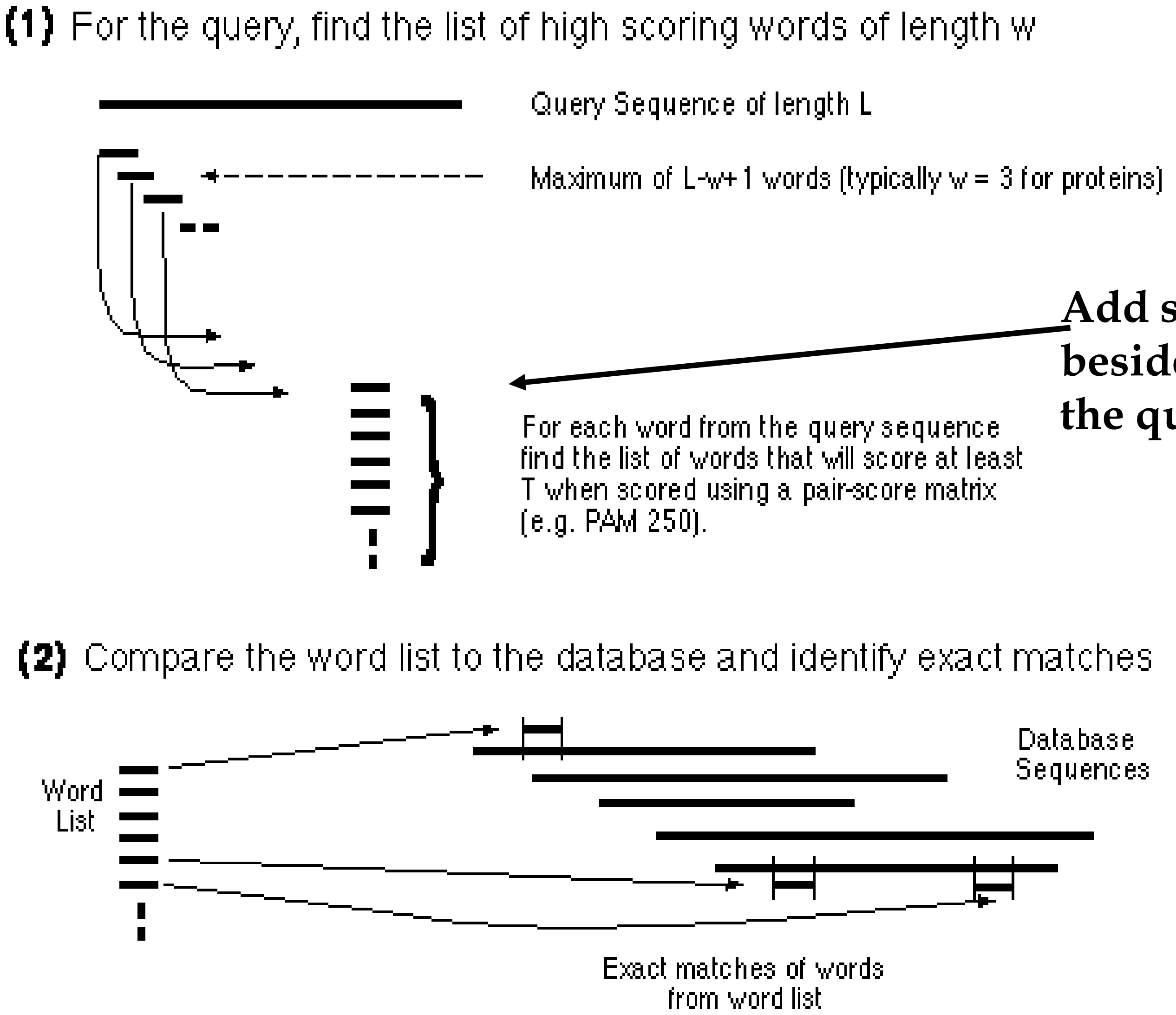
BLAST Algorithm

It uses three layers of rules to sequentially find potential high scoring pairs (HSPs)

1. **Seeding**
2. **Extension**
3. **Evaluation**

to sample the entire search space without wasting time on dissimilar regions.

BLAST Algorithm



(1) For the query, find the list of high scoring words of length w

Query Sequence of length L
Maximum of L-w+1 words (typically w = 3 for proteins)

For each word from the query sequence find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).

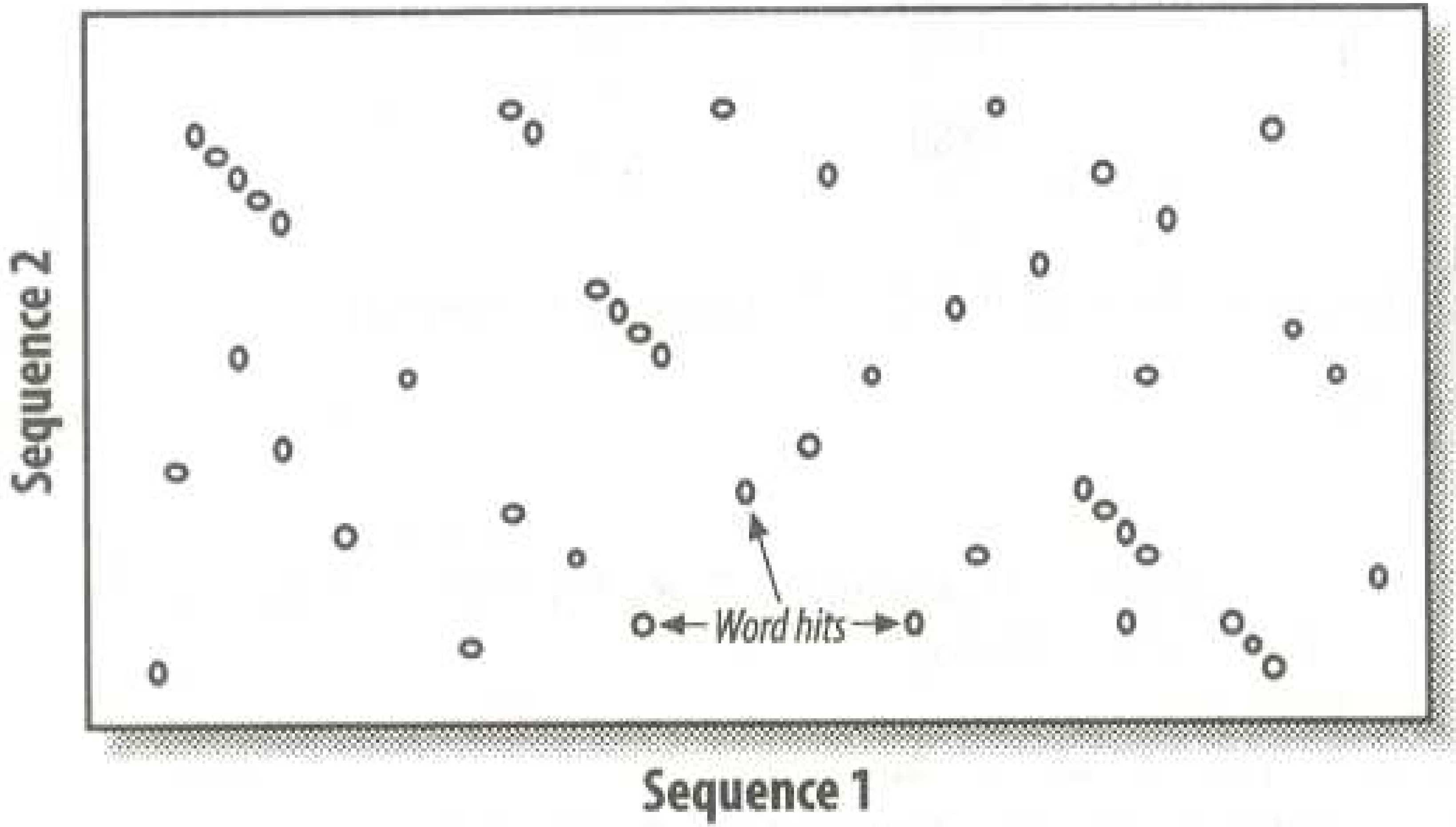
(2) Compare the word list to the database and identify exact matches

Word List
Database Sequences
Exact matches of words from word list

Add similar words besides those in the query.

BLOSUM62		PAM200	
Word	Score	Word	Score
RGD	17	RGD	18
KGD	14	RGE	17
QGD	13	RGN	16
RGE	13	KGD	15
EGD	12	RGQ	15
HGD	12	KGE	14
NGD	12	HGD	13
RGN	12	KGN	13
AGD	11	RAD	13
MGD	11	RGA	13
RAD	11	RGG	13
RGQ	11	RGH	13
RGS	11	RGK	13
RND	11	RGS	13
RSD	11	RGT	13
SGD	11	RSD	13
TGD	11	WGD	13

1.Seeding



Selection of T and W and Scoring matrix

- The proper value for T depends on both the values in the scoring matrix and the balance between speed and sensitivity.
- Higher values of T progressively remove more hits and reduces the search space (run faster) but increases the chance of missing an alignments
- Word size (w) also control the word hits.
- Smaller w increases sensitivity but decreases speed.
- So interplay between W,T and matrix is critical to control speed and sensitivity of BLAST.

Effect of threshold T

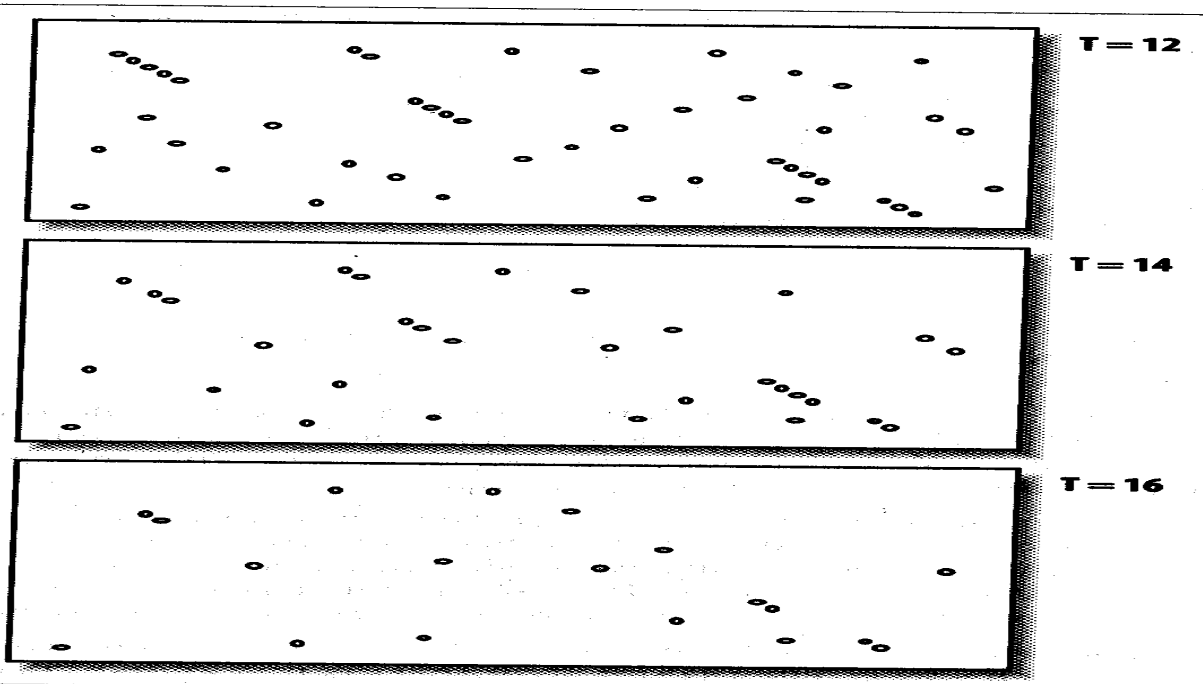


Figure 5-3. How T affects seeding

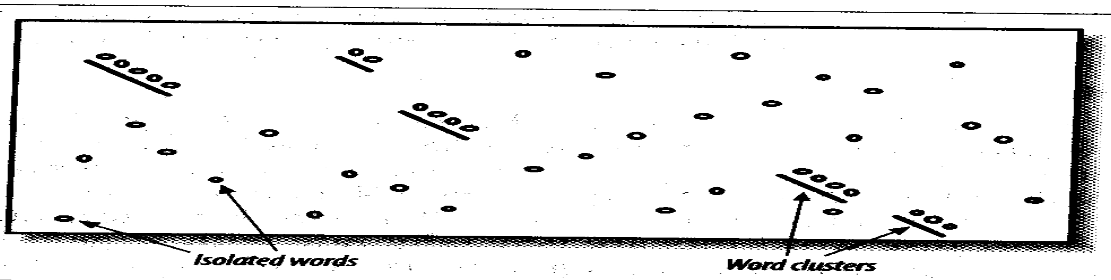
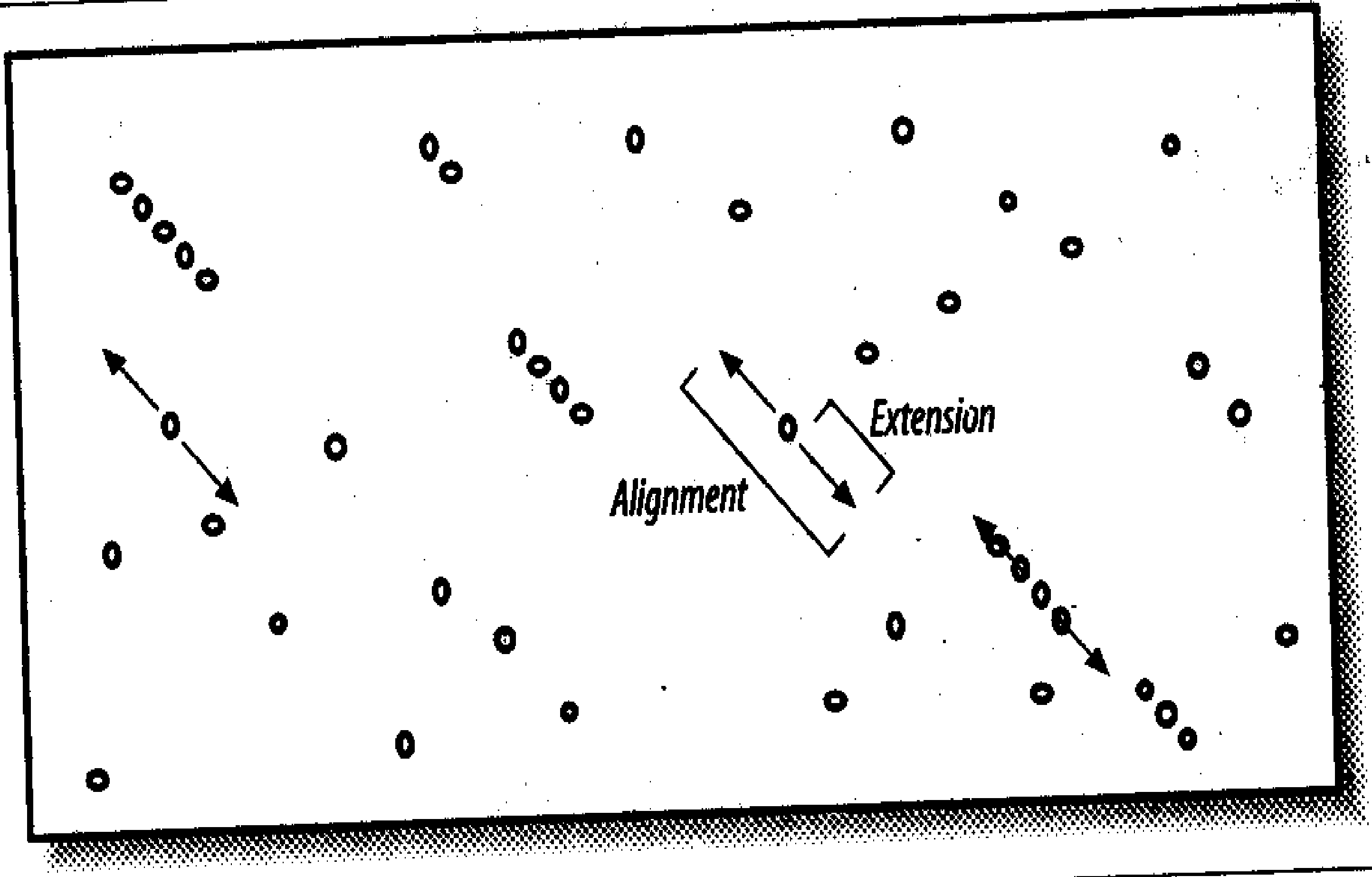


Figure 5-4. Isolated and clustered words

2. Extension

- **Once the search space is seeded, alignments can be generated from individual seeds in both direction.**

2. Extension



2: Extend matches

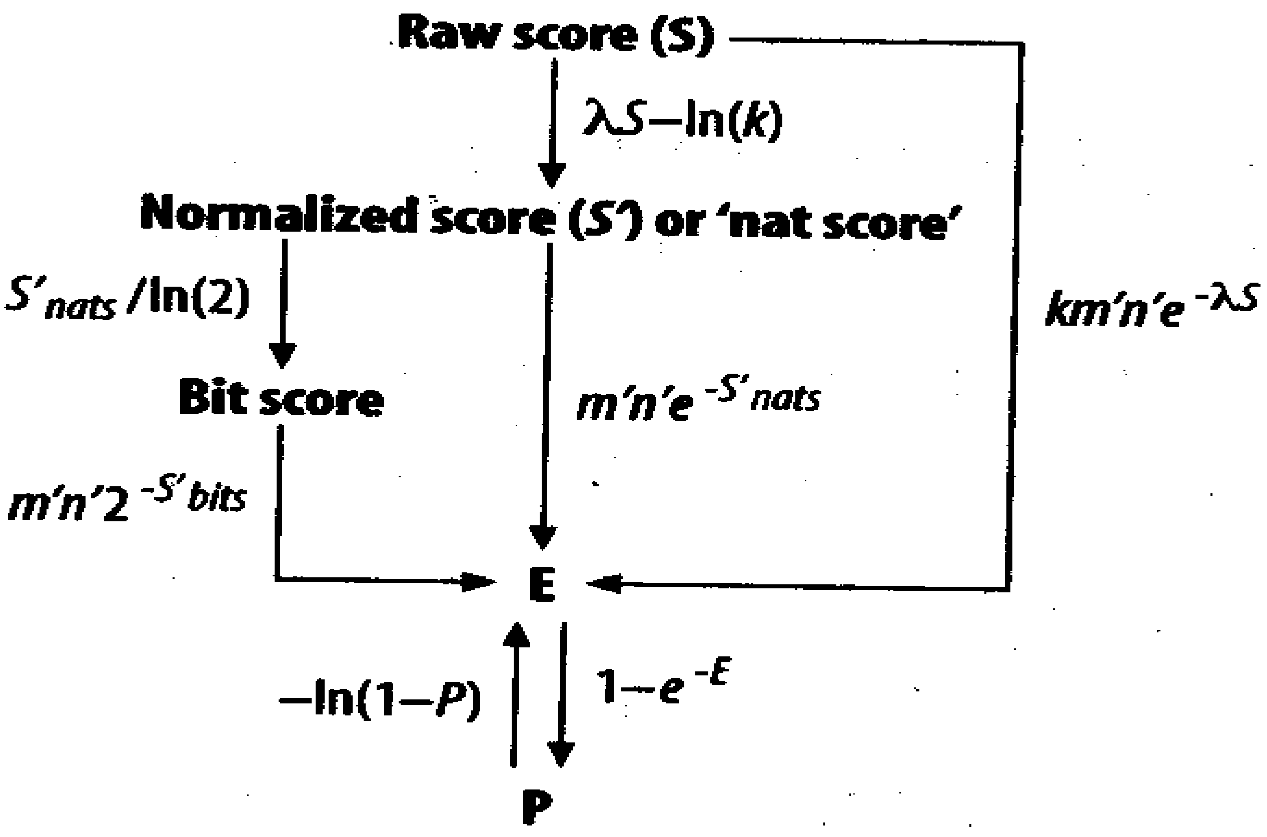
```
L P  P Q G  L L  Query sequence
M P  P E G  L L  Database sequence
     <word>
     7 2 6       BLOSUM62 scores
                 word score = 15
<---       --->
2 7  7 2 6  4 4  HSP SCORE = 32
```

- **Each match is extended to the left and right until a negative BLOSUM62 score is encountered.**

3. Evaluation

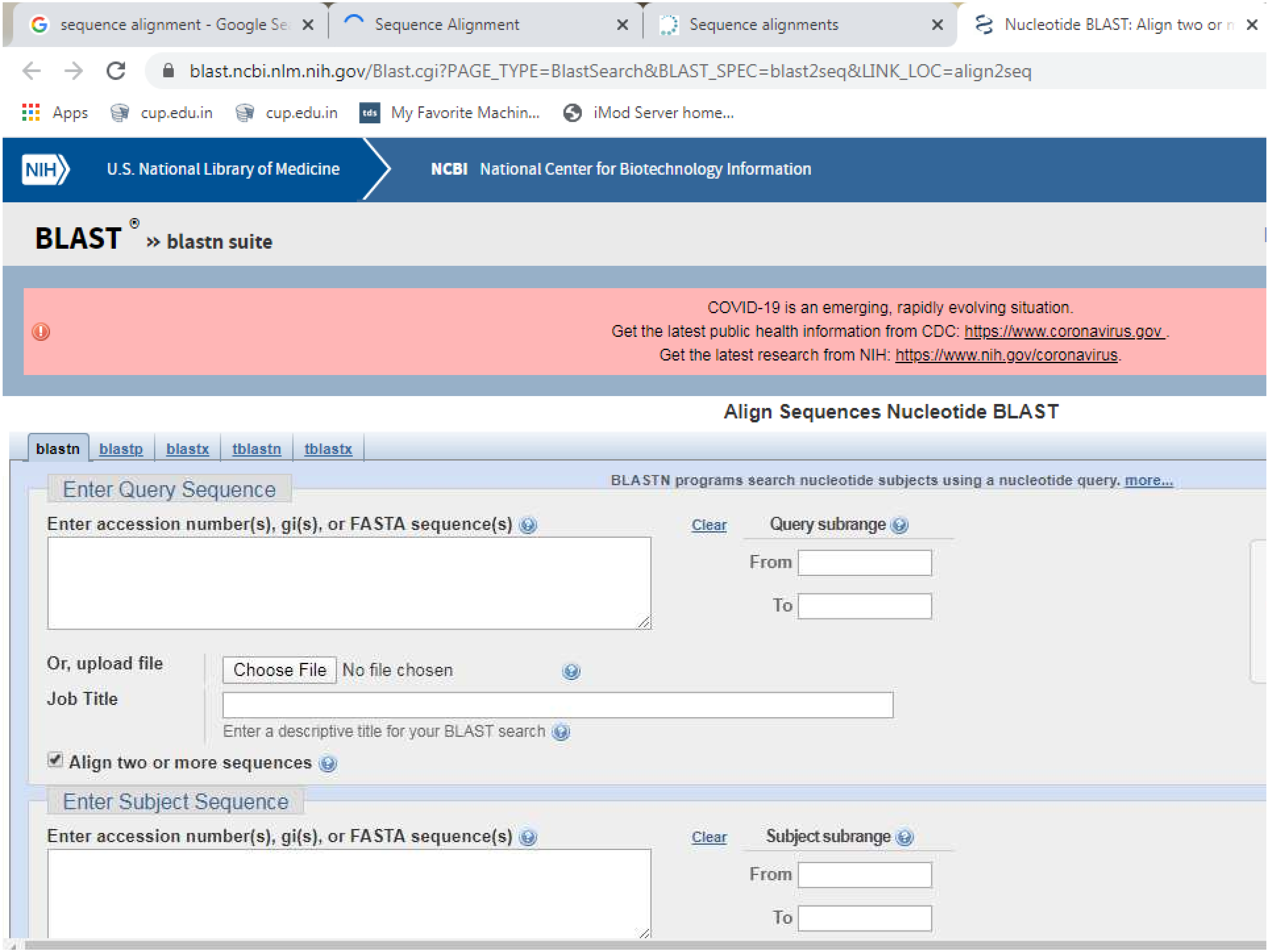
- Statistical significant of the alignments are evaluated and termed as HSPs.
- Because alignment score (S) and Expect(E) are directly related through Karlin-Altschul equation, so S is an synonymous with a statistical threshold.
- $E=k\ m\ n\ e^{-\lambda S}$
- m=no. of letters in query
- n= no. of letters in database
- K=minor constant
- λS= normalized score

Statistical parameter of BLAST



P and E-values

- A p-value is the probability of making a mistake.
- The E-value is the expected number of times that the given score would appear in a random database of the given size.
- The E-value is computed by multiplying the p-value times the size of the database.
- Thus, for a p-value of 0.001 and a database of 1,000,000 sequences, the corresponding E-value is 0.001 × 1,000,000 = 1,000.



sequence alignment - Google Se ✕ | Sequence Alignment ✕ | Sequence alignments ✕ | Nucleotide BLAST: Align two or n ✕

blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq

Apps | cup.edu.in | cup.edu.in | My Favorite Machin... | iMod Server home...

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information

BLAST ® » blastn suite

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.

Align Sequences Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide subjects using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) | Clear | Query subrange

From

To

Or, upload file | Choose File | No file chosen

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

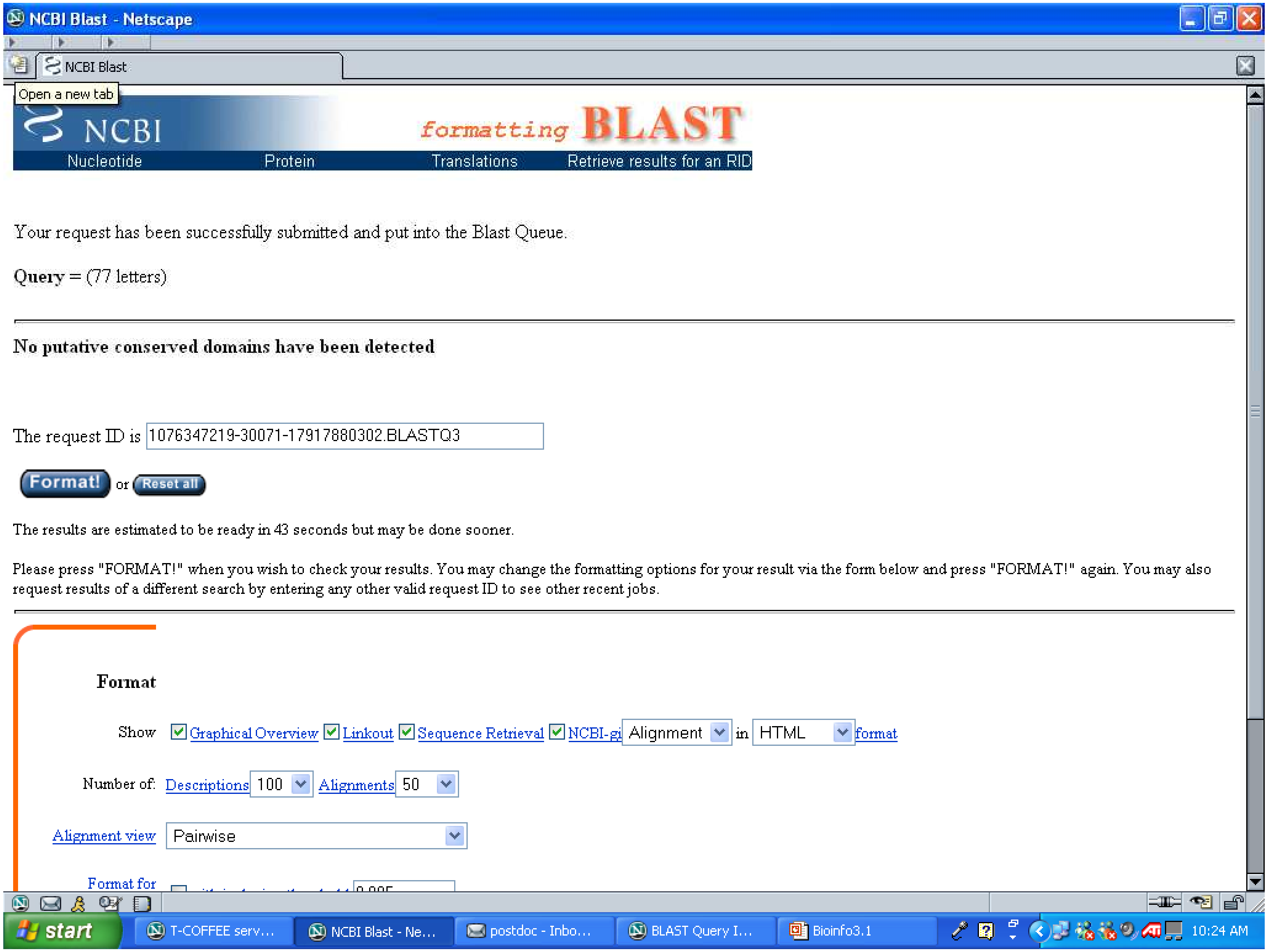
Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) | Clear | Subject subrange

From

To

Formatting Results



Your request has been successfully submitted and put into the Blast Queue.

Query = (77 letters)

No putative conserved domains have been detected

The request ID is 1076347219-30071-17917880302.BLASTQ3

Format! or Reset all

The results are estimated to be ready in 43 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

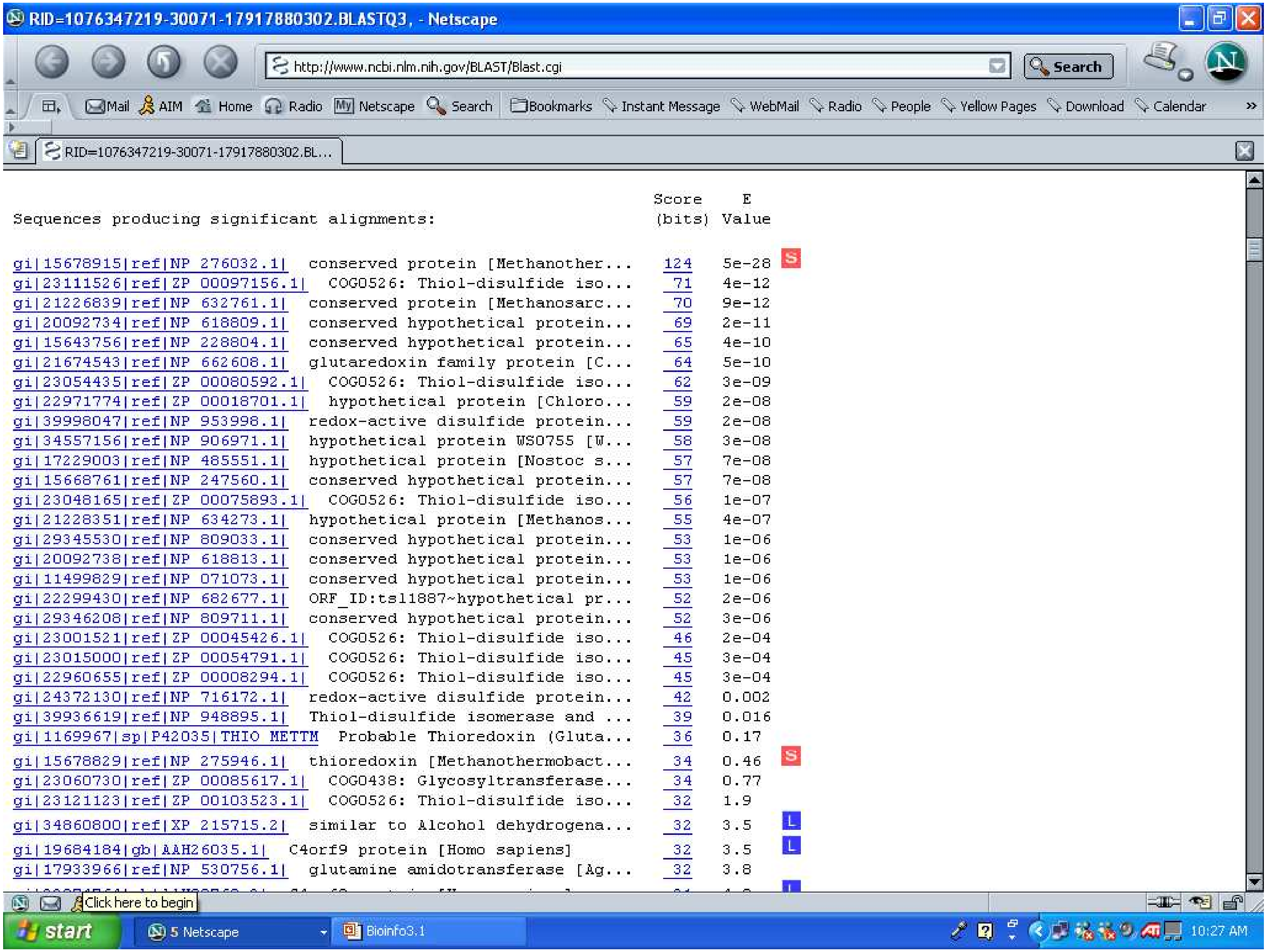
Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

Number of: Descriptions 100 Alignments 50

Alignment view Pairwise

Format for □ 0.005

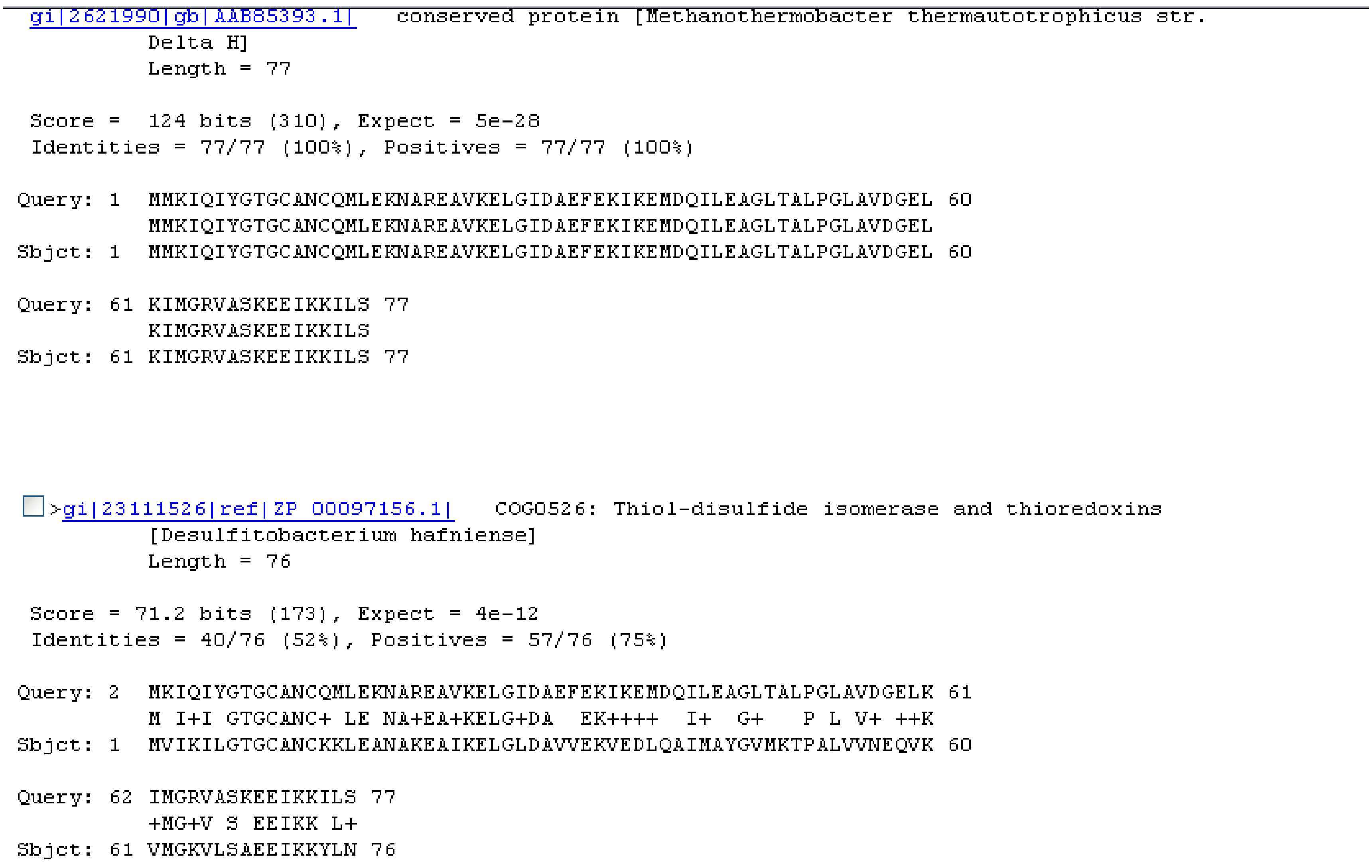
BLAST Output



```
                                                                 Score    E
Sequences producing significant alignments:                      (bits)  Value

gi|15678915|ref|NP 276032.1|  conserved protein [Methanother...   124    5e-28  S
gi|23111526|ref|ZP 00097156.1|  COG0526: Thiol-disulfide iso...    71    4e-12
gi|21226839|ref|NP 632761.1|  conserved protein [Methanosarc...    70    9e-12
gi|20092734|ref|NP 618809.1|  conserved hypothetical protein...    69    2e-11
gi|15643756|ref|NP 228804.1|  conserved hypothetical protein...    65    4e-10
gi|21674543|ref|NP 662608.1|  glutaredoxin family protein [C...    64    5e-10
gi|23054435|ref|ZP 00080592.1|  COG0526: Thiol-disulfide iso...    62    3e-09
gi|22971774|ref|ZP 00018701.1|  hypothetical protein [Chloro...    59    2e-08
gi|39998047|ref|NP 953998.1|  redox-active disulfide protein...    59    2e-08
gi|34557156|ref|NP 906971.1|  hypothetical protein WS0755 [W...    58    3e-08
gi|17229003|ref|NP 485551.1|  hypothetical protein [Nostoc s...    57    7e-08
gi|15668761|ref|NP 247560.1|  conserved hypothetical protein...    57    7e-08
gi|23048165|ref|ZP 00075893.1|  COG0526: Thiol-disulfide iso...    56    1e-07
gi|21228351|ref|NP 634273.1|  hypothetical protein [Methanos...    55    4e-07
gi|29345530|ref|NP 809033.1|  conserved hypothetical protein...    53    1e-06
gi|20092738|ref|NP 618813.1|  conserved hypothetical protein...    53    1e-06
gi|11499829|ref|NP 071073.1|  conserved hypothetical protein...    53    1e-06
gi|22299430|ref|NP 682677.1|  ORF_ID:tsl1887~hypothetical pr...    52    2e-06
gi|29346208|ref|NP 809711.1|  conserved hypothetical protein...    52    3e-06
gi|23001521|ref|ZP 00045426.1|  COG0526: Thiol-disulfide iso...    46    2e-04
gi|23015000|ref|ZP 00054791.1|  COG0526: Thiol-disulfide iso...    45    3e-04
gi|22960655|ref|ZP 00008294.1|  COG0526: Thiol-disulfide iso...    45    3e-04
gi|24372130|ref|NP 716172.1|  redox-active disulfide protein...    42    0.002
gi|39936619|ref|NP 948895.1|  Thiol-disulfide isomerase and ...    39    0.016
gi|1169967|sp|P42035|THIO METTM  Probable Thioredoxin (Gluta...    36    0.17
gi|15678829|ref|NP 275946.1|  thioredoxin [Methanothermobact...    34    0.46   S
gi|23060730|ref|ZP 00085617.1|  COG0438: Glycosyltransferase...    34    0.77
gi|23121123|ref|ZP 00103523.1|  COG0526: Thiol-disulfide iso...    32    1.9
gi|34860800|ref|XP 215715.2|  similar to Alcohol dehydrogena...    32    3.5    L
gi|19684184|gb|AAH26035.1|  C4orf9 protein [Homo sapiens]          32    3.5    L
gi|17933966|ref|NP 530756.1|  glutamine amidotransferase [Ag...    32    3.8
```

BLAST Output



```
 gi|2621990|gb|AAB85393.1|   conserved protein [Methanothermobacter thermautotrophicus str.
          Delta H]
          Length = 77

 Score =  124 bits (310), Expect = 5e-28
 Identities = 77/77 (100%), Positives = 77/77 (100%)

Query: 1   MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60
            MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL
Sbjct: 1   MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60

Query: 61  KIMGRVASKEEIKKILS 77
            KIMGRVASKEEIKKILS
Sbjct: 61  KIMGRVASKEEIKKILS 77
```

```
□ >gi|23111526|ref|ZP_00097156.1|   COG0526: Thiol-disulfide isomerase and thioredoxins
          [Desulfitobacterium hafniense]
          Length = 76

 Score = 71.2 bits (173), Expect = 4e-12
 Identities = 40/76 (52%), Positives = 57/76 (75%)

Query: 2   MKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGELK 61
            M I+I GTGCANC+ LE NA+EA+KELG+DA  EK++++  I+  G+   P L V+ ++K
Sbjct: 1   MVIKILGTGCANCKKLEANAKEAIKELGLDAVVEKVEDLQAIMAYGVMKTPALVVNEQVK 60

Query: 62  IMGRVASKEEIKKILS 77
            +MG+V S EEIKK L+
Sbjct: 61  VMGKVLSAEEIKKYLN 76
```

BLAST - Rules of Thumb

- Don’t trust a BLAST alignment with an Expect score > 0.01
- Expect and Score are related, but Expect contains more information. Note that %Identities is more useful than the bit Score
- If uncertain about a hit, perform a PSI-BLAST search

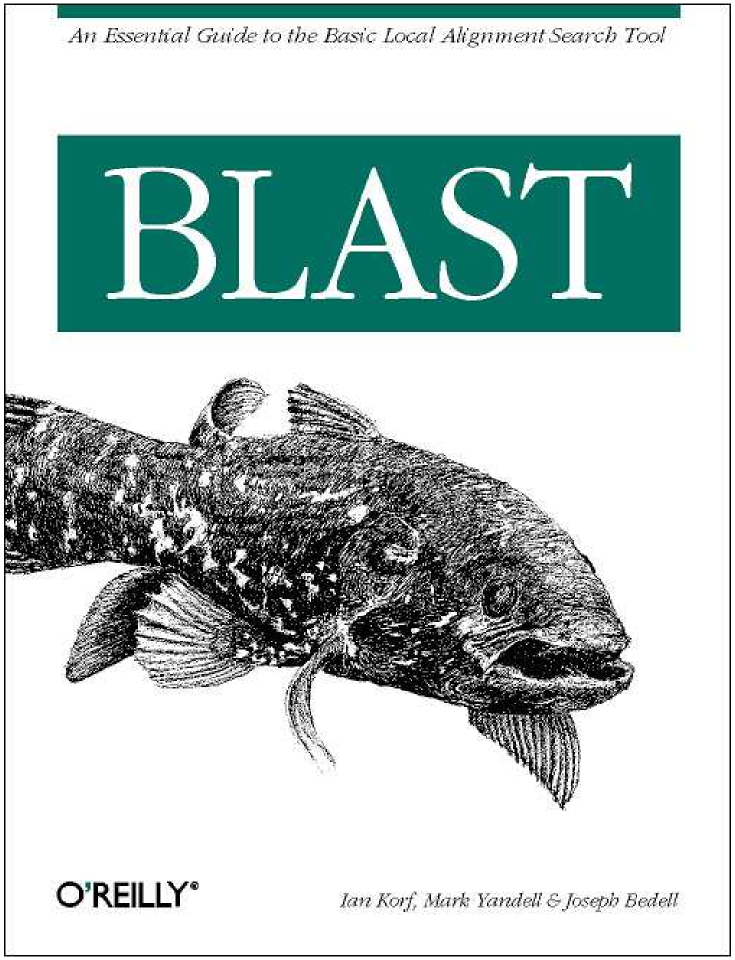
Different Flavours of BLAST

- BLASTP - protein query against protein DB
- BLASTN - DNA query against GenBank (DNA)
- BLASTX - 6 frame trans. DNA query against proteinDB
- TBLASTN - protein query against 6 frame GB transl.
- TBLASTX - 6 frame DNA query to 6 frame GB transl.
- PSI-BLAST - protein ‘profile’ query against protein DB

References

- https://bioinf.comav.upv.es/courses/biotech3/theory/sequence_alignment.html
- https://www.ebi.ac.uk/Tools/psa/
- https://www.ebi.ac.uk/Tools/msa/clustalw2/

O’Reilly Book


An Essential Guide to the Basic Local Alignment Search Tool
BLAST
O'REILLY®
Ian Korf, Mark Yandell & Joseph Bedell

Home Assignment

1. Discuss the importance of sequence alignment.
2. Differentiate between global and local sequence alignment with examples.
3. Describe the BLAST algorithm for database searching.

Last data of submission 20.04.2020

Thank you.

Email: sprakashsingh@mgcub.ac.in