Techniques for Protein Sequence Alignment and Database Searching

G P S Raghava

Scientist & Head Bioinformatics Centre,

Institute of Microbial Technology,

Chandigarh, India

Email: raghava@imtech.res.in

Web: http://imtech.res.in/raghava/

Importance of Sequence Comparison

- Protein Structure Prediction
  - Similar sequence have similar structure & function
  - Phylogenetic Tree
  - Homology based protein structure prediction
- Genome Annotation
  - Homology based gene prediction
  - Function assignment & evolutionary studies
- Searching drug targets
  - Searching sequence present or absent across genomes

Protein Sequence Alignment and Database Searching

- **Alignment of Two Sequences (Pair-wise Alignment)**
  - The Scoring Schemes or Weight Matrices
  - Techniques of Alignments
  - DOTPLOT
- **Multiple Sequence Alignment (Alignment of > 2 Sequences)**
  - Extending Dynamic Programming to more sequences
  - Progressive Alignment (Tree or Hierarchical Methods)
  - Iterative Techniques
    - Stochastic Algorithms (SA, GA, HMM)
    - Non Stochastic Algorithms
- **Database Scanning**
  - FASTA, BLAST, PSIBLAST, ISS
- **Alignment of Whole Genomes**
  - MUMmer (Maximal Unique Match)

Pair-Wise Sequence Alignment

Scoring Schemes or Weight Matrices

- ➢ Identity Scoring
- ➢ Genetic Code Scoring
- ➢ Chemical Similarity Scoring
- ➢ Observed Substitution or PAM Matrices
- ➢ PEP91: An Update Dayhoff Matrix
- ➢ BLOSUM: Matrix Derived from Ungapped Alignment
- ➢ Matrices Derived from Structure

Techniques of Alignment

- ➢ Simple Alignment, Alignment with Gaps
- ➢ Application of DOTPLOT (Repeats, Inverse Repeats, Alignment)
- ➢ Dynamic Programming (DP) for Global Alignment
- ➢ Local Alignment (Smith-Waterman algorithm)

Important Terms

- ➢ Gap Penalty (Opening, Extended)
- ➢ PID, Similarity/Dissimilarity Score
- ➢ Significance Score (e.g. Z & E)

The Scoring Schemes or Weight Matrices

For any alignment one need scoring scheme and weight matrix

Important Point

- All algorithms to compare protein sequences rely on some scheme to score the equivalencing of each 210 possible pairs.
- 190 different pairs + 20 identical pairs
- Higher scores for identical/similar amino acids (e.g. A,A or I, L)
- Lower scores to different character (e.g. I, D)

Identity Scoring

- Simplest Scoring scheme
- Score 1 for Identical pairs
- Score 0 for Non-Identical pairs
- Unable to detect similarity
- Percent Identity

Genetic Code Scoring

- Fitch 1966 based on Nucleotide Base change required (0,1,2,3)
- Required to interconvert the codons for the two amino acids
- Rarely used nowadays

The Scoring Schemes or Weight Matrices

Chemical Similarity Scoring

- Similarity based on Physio-chemical properties
- MacLachlan 1972, Based on size, shape, charge and polar
- Score 0 for opposite (e.g. E & F) and 6 for identical character

Observed Substitutions or PAM matrices

- Based on Observed Substitutions
- Chicken and Egg problem
- Dayhoff group in 1977 align sequence manually
- Observed Substitutions or point mutation frequency
- MATRICES are PAM30, PAM250, PAM100 etc

```
AILDCTGRTG......
ALLDCTGR--......
SLIDCSAR-G......
AILNCTL-RG......
```

PET91: An update Dayhoff matrix

BLOSUM- Matrix derived from Ungapped Alignment

- Derived from Local Alignment instead of Global
- Henikoff and Henikoff derived matric from conserved blocks
- BLOSUM80, BLOSUM62, BLOSUM35

The Scoring Schemes or Weight Matrices

Matrices Derived from Structure

- ➢Structure alignment is true/reference alignment
- ➢Allow to compare distant proteins
- ➢Risler 1988, derived from 32 protein structures

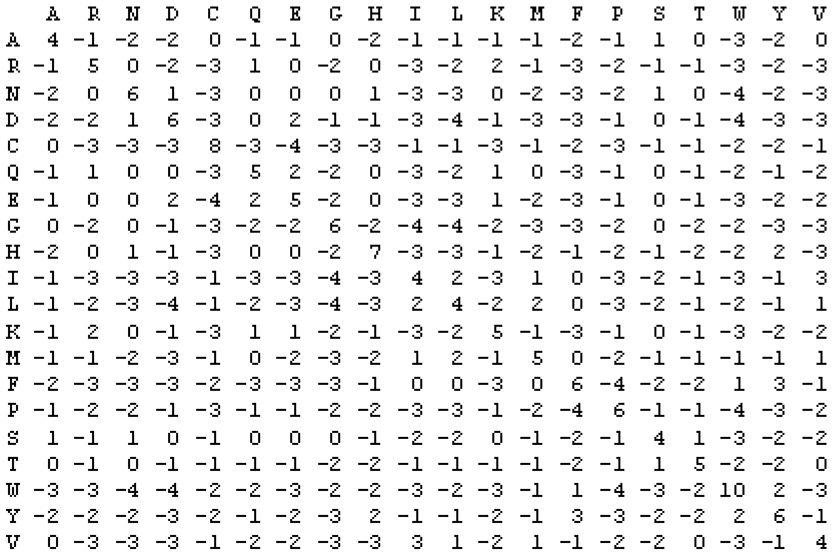
Which Matrix one should use

- ➢Matrices derived from Observed substitutions are better
- ➢BLOSUM and Dayhoff (PAM)
- ➢BLOSUM62 or PAM250

●

Similarity (Substitution) Matrix

- Identity Matrix
  - ◊ Match L with L => 1
    Match L with D => 0
    Match L with V => 0??
- S(aa-1,aa-2)
  - ◊ Match L with L => 1
    Match L with D => 0
    Match L with V => .5
- Number of Common Ones
  - ◊ PAM
  - ◊ Blossum
  - ◊ Gonnet



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	8	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	7	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	6	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	10	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	6	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Alignment of Two Sequences

<u>Dealing Gaps in Pair-wise Alignment</u>

Sequence Comparison without Gaps

Slide Windos method to got maximum score

ALGAWDE

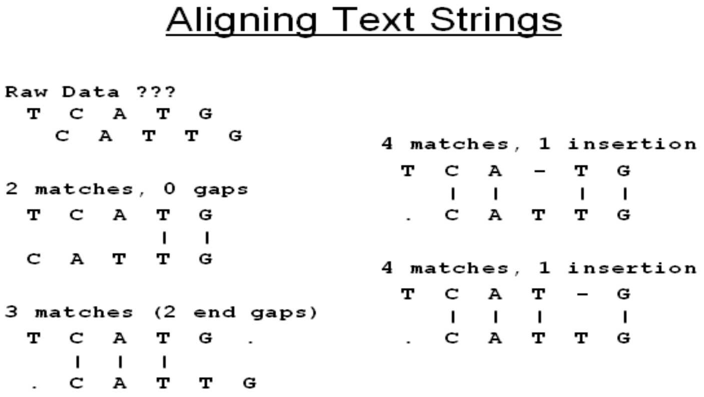
ALATWDE

Total score= 1+1+0+0+1+1+1=5 ; (PID) = (5*100)/7

Sequence with variable length should use dynamic programming

Sequence Comparison with Gaps

- Insertion and deletion is common
- Slide Window method fails
- Generate all possible alignment
- 100 residue alignment require > 10^{75}

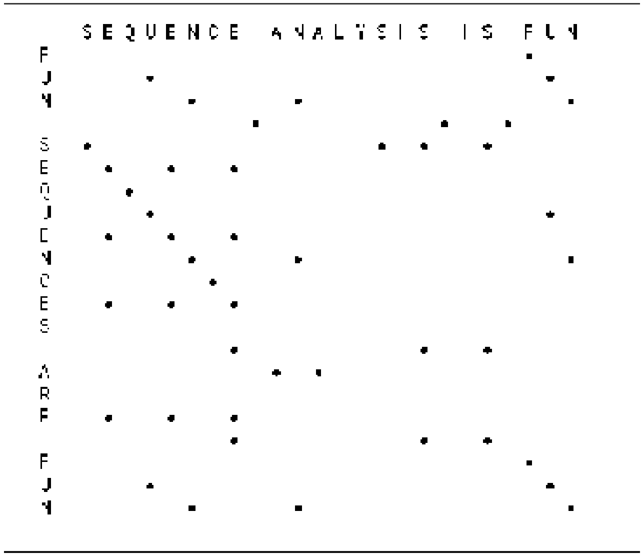


Aligning Text Strings

```
Raw Data ???
 T   C   A   T   G
   C   A   T   T   G

2 matches, 0 gaps
 T   C   A   T   G
             |   |
 C   A   T   T   G

3 matches (2 end gaps)
 T   C   A   T   G   .
     |   |   |
 .   C   A   T   T   G
```

```
4 matches, 1 insertion
 T   C   A   -   T   G
     |   |       |   |
 .   C   A   T   T   G

4 matches, 1 insertion
 T   C   A   T   -   G
     |   |   |       |
 .   C   A   T   T   G
```

Alternate Dot Matrix Plot
Diagnoal * shows align/identical regions



Dynamic Programming

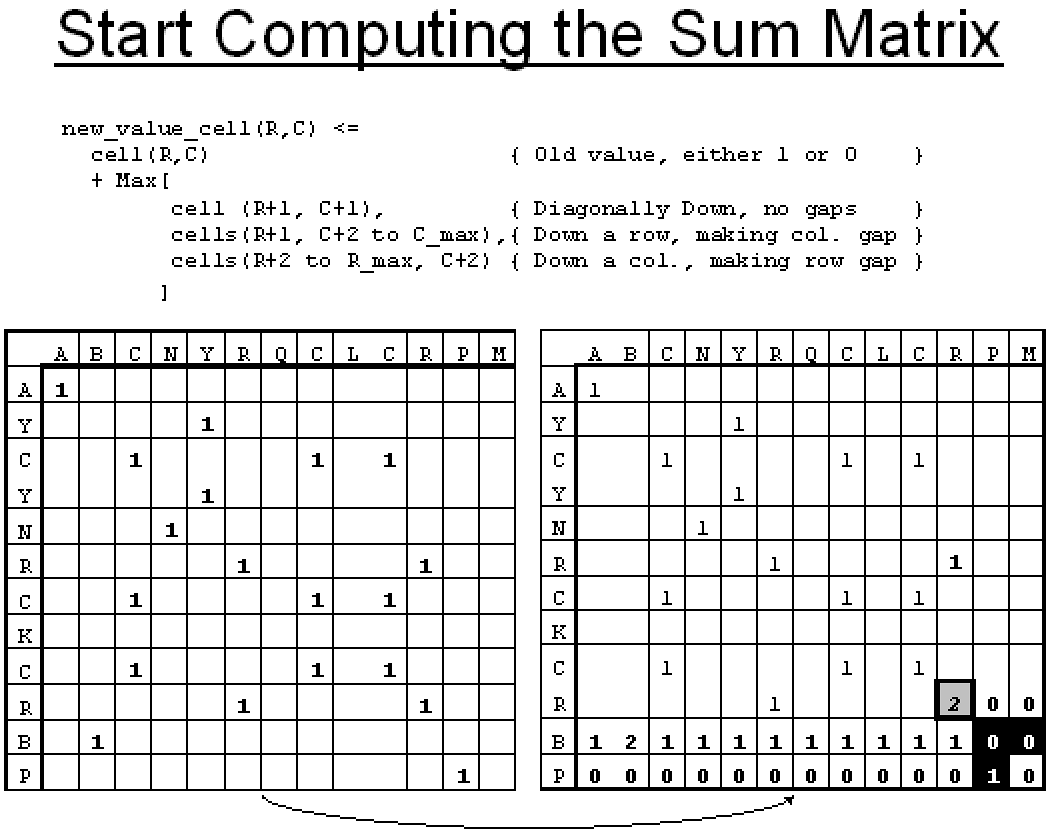
- Dynamic Programming allow Optimal Alignment between two sequences
- Allow Insertion and Deletion or Alignment with gaps
- Needlman and Wunsh Algorithm (1970) for global alignment
- Smith & Waterman Algorithm (1981) for local alignment
- Important Steps
  - Create DOTPLOT between two sequences
  - Compute SUM matrix
  - Trace Optimal Path

Step 1 -- Make a Dot Plot (Similarity Matrix)

Put 1's where characters are identical.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

Steps for Dynamic Programming



Start Computing the Sum Matrix

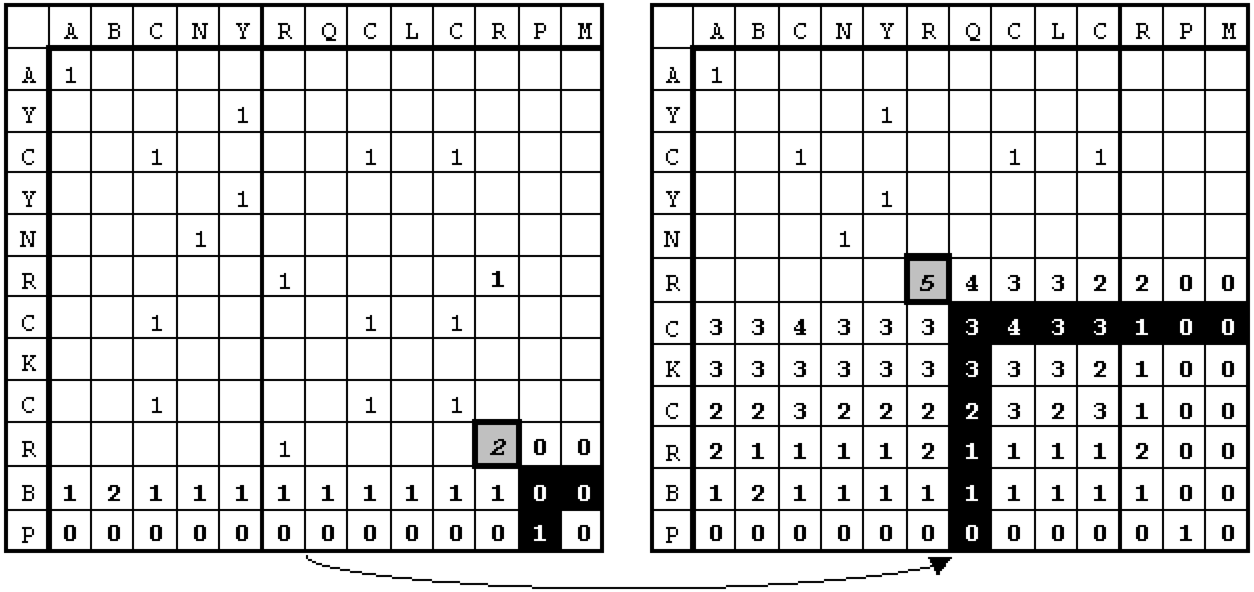
```
new_value_cell(R,C) <=
  cell(R,C)                        { Old value, either 1 or 0    }
  + Max[
         cell (R+1, C+1),          { Diagonally Down, no gaps    }
         cells(R+1, C+2 to C_max),{ Down a row, making col. gap }
         cells(R+2 to R_max, C+2) { Down a col., making row gap }
        ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Steps for Dynamic Programming

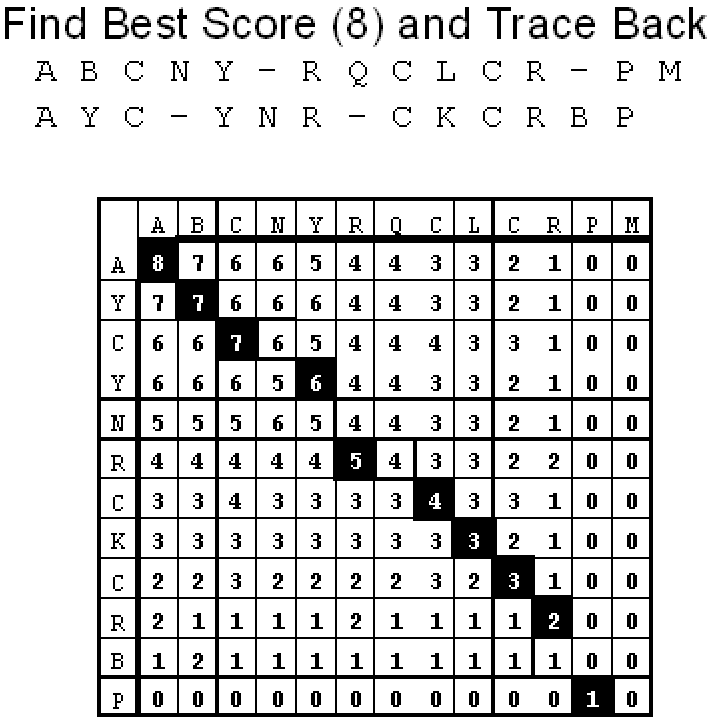
Keep Going



	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Steps for Dynamic Programming

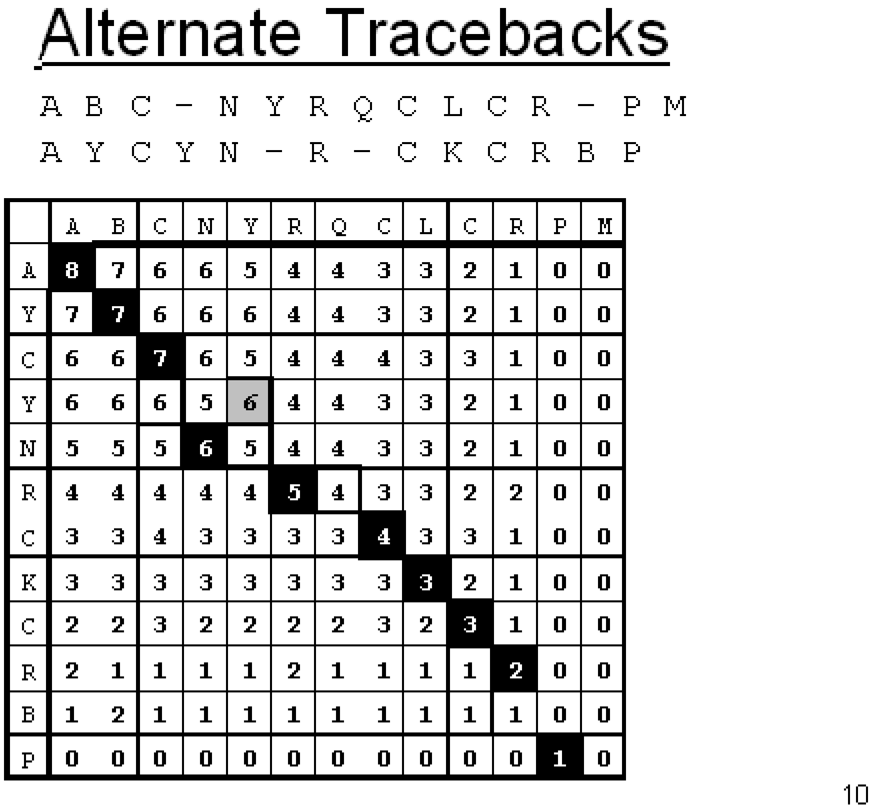


Find Best Score (8) and Trace Back

```
A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Steps for Dynamic Programming



Alternate Tracebacks

```
A B C - N Y R Q C L C R - P M
A Y C Y N - R - C K C R B P
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

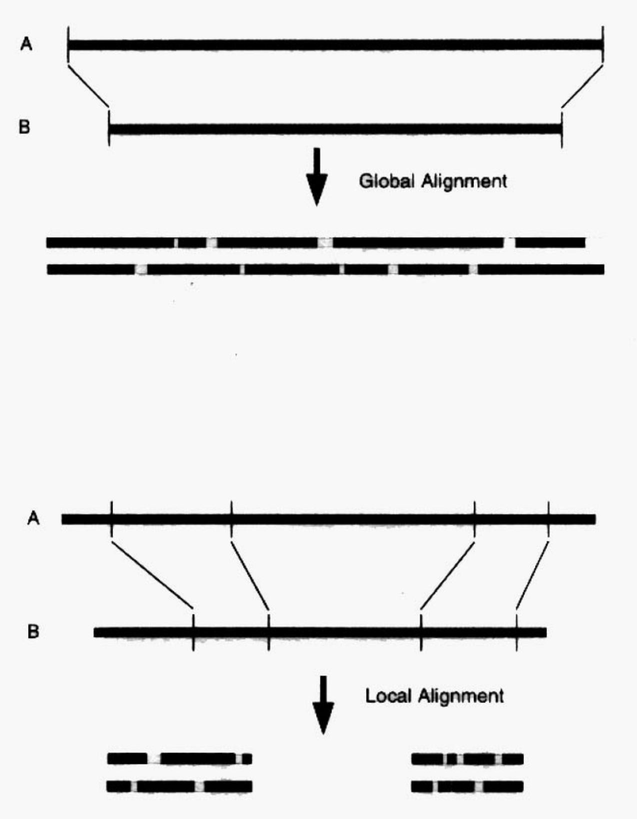
Important Terms in Pairwise Sequence Alignment

Global Alignment

- –Suite for similar sequences
- –Nearly equal legnth
- – Overall similarity is detected

Local Alignment

- –Isolate regions in sequences
- –Suitable for database searching
- –Easy to detect repeats



•Gap Penalty (Opening + Extended)

```
ALTGTRTG...CALGR …
AL.GTRTGTGPCALGR …
```

Important Points in Pairwise Sequence Alignment

Significance of Similarity

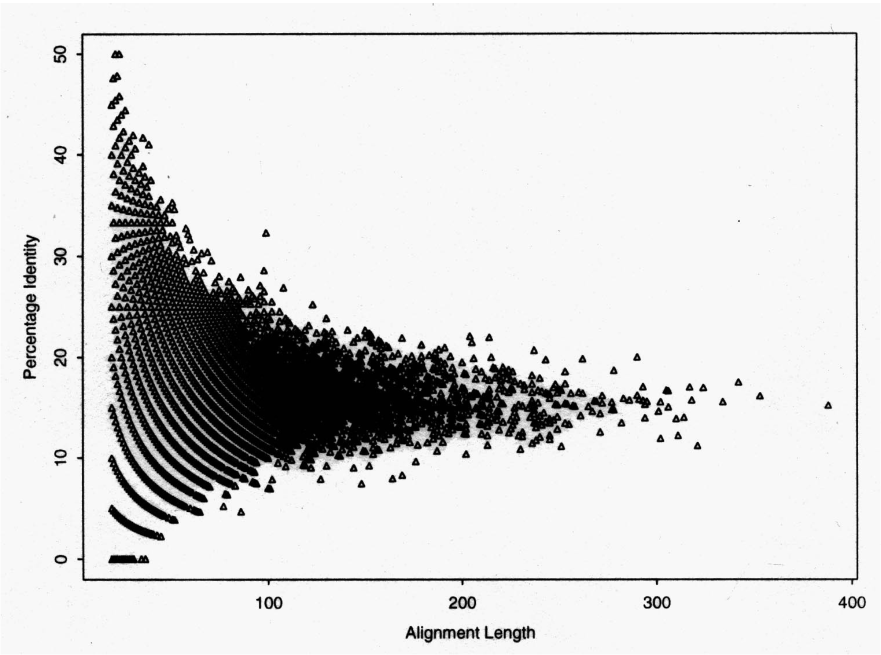
– Dependent on PID (Percent Identical Positions in Alignment)

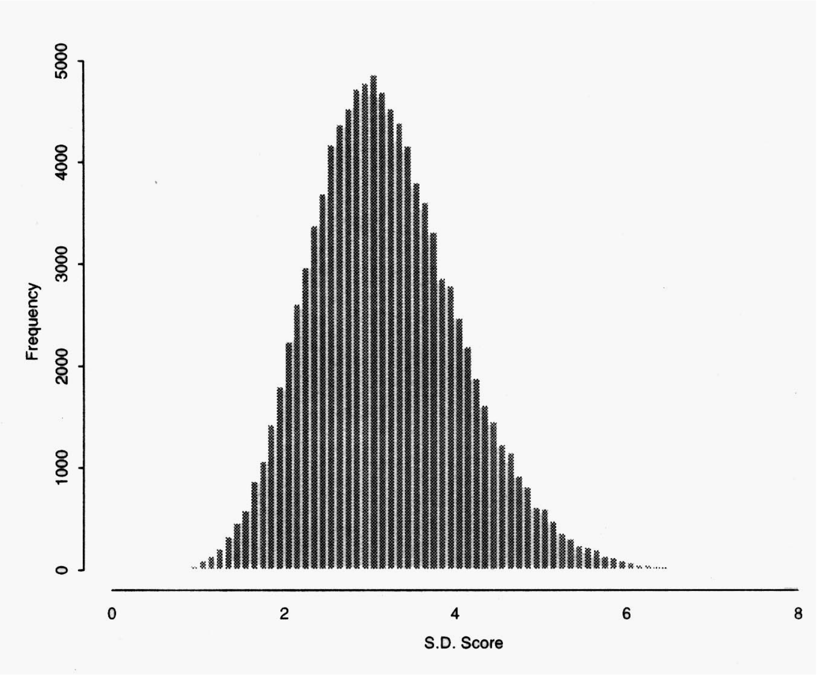
–Similarity/Disimilarity score

– Significance of score depend on length of alignment

–Significance Score (Z) whether score significant

–Expected Value (E), Chances that non-related sequence may have that score





Alignment of Multiple Sequences

Extending Dynamic Programming to more sequences

- –Dynamic programming can be extended for more than two
- –In practice it requires CPU and Memory (Murata et al 1985)
- – MSA, Limited only up to 8-10 sequences (1989)
- –DCA (Divide and Conquer; Stoye et al., 1997), 20-25 sequences
- –OMA (Optimal Multiple Alignment; Reinert et al., 2000)
- –COSA (Althaus et al., 2002)

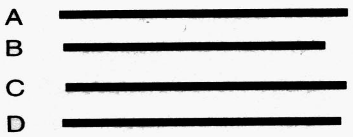
Progressive or Tree or Hierarchical Methods (CLUSTAL-W)

- –Practical approach for multiple alignment
- –Compare all sequences pair wise
- –Perform cluster analysis
- –Generate a hierarchy for alignment
- –first aligning the most similar pair of sequences
- –Align alignment with next similar alignment or sequence

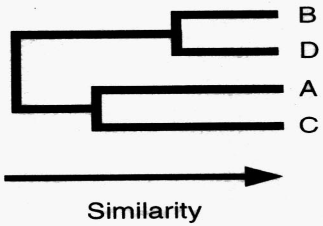
Steps in Multiple Alignment

(A) Pairwise Alignment

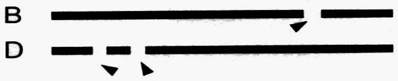
Example - 4 Sequences, A, B, C, D.



6 Pairwise Comparisons
then Cluster analysis

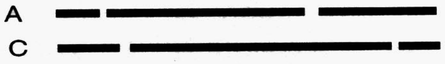


(B) Multiple alignment following the tree from A.



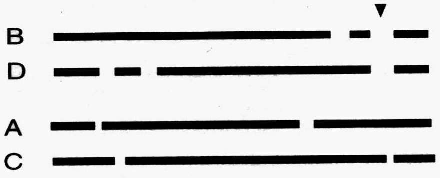
Gaps to optimise alignment

Align most similar pair.



Align next most similar pair.

New gap to optimise
alignment of (BD) with
(AC).



Align alignments - preserve gaps.

Alignment of Multiple Sequences

Iterative Alignment Techniques

- Deterministic (Non Stochastic) methods
  - They are similar to Progressive alignment
  - Rectify the mistake in alignment by iteration
  - Iterations are performed till no further improvement
  - AMPS (Barton & Sternberg; 1987)
  - PRRP (Gotoh, 1996), Most successful
  - Praline, IterAlign
- Stochastic Methods
  - SA (Simulated Annealing; 1994), alignment is randomly modified only acceptable alignment kept for further process. Process goes until converged
  - Genetic Algorithm alternate to SA (SAGA, Notredame & Higgins, 1996)
  - COFFEE extension of SAGA
  - Gibbs Sampler
  - Bayesian Based Algorithm (HMM; HMMER; SAM)
  - They are only suitable for refinement not for producing *ab initio* alignment. Good for profile generation. Very slow.

Alignment of Multiple Sequences

Progress in Commonly used Techniques (Progressive)

Clustal-W (1.8) (Thompson et al., 1994)

- Automatic substitution matrix
- Automatic gap penalty adjustment
- Delaying of distantly related sequences
- Portability and interface excellent

T-COFFEE (Notredame et al., 2000)

- Improvement in Clustal-W by iteration
- Pair-Wise alignment (Global + Local)
- Most accurate method but slow

MAFFT (Katoh et al., 2002)

- Utilize the FFT for pair-wise alignment
- Fastest method
- Accuracy nearly equal to T-COFFEE

Database scanning

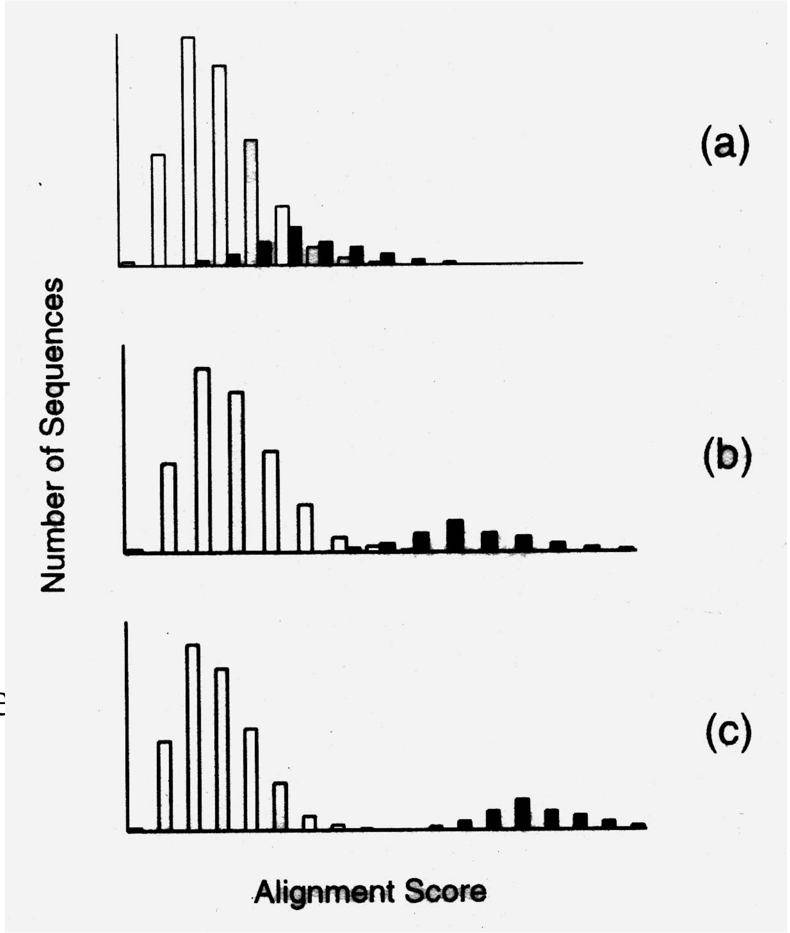
Basic principles of Database searching

- – Search query sequence against all sequence in database
- – Calculate score and select top sequences
- – Dynamic programming is best

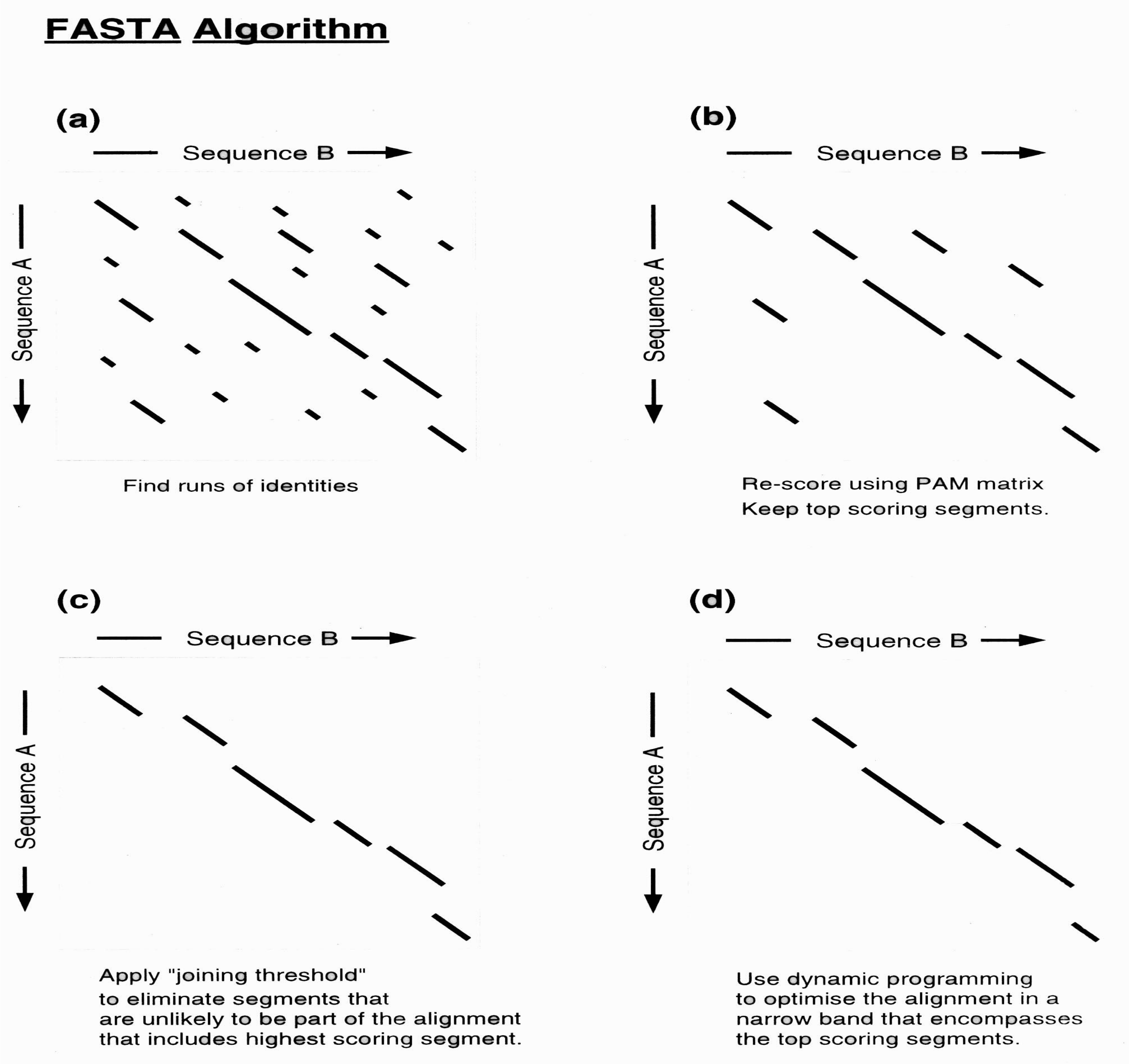
Approximation Algorithms

FASTA

- ✓Fast sequence search
- ✓Based on dotplot
- ✓Identify identical words (k-tuples)
- ✓Search significant diagonals
- ✓Use PAM 250 for further refinement
- ✓Dynamic programming for narrow re



Principles of FASTA Algorithms


FASTA Algorithm

(a) Find runs of identities

(b) Re-score using PAM matrix
Keep top scoring segments.

(c) Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.

(d) Use dynamic programming
to optimise the alignment in a
narrow band that encompasses
the top scoring segments.

Database scanning

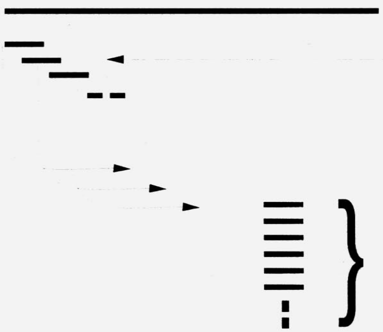
Approximation Algorithms

BLAST

- ✓Heuristic method to find the highest scoring
- ✓Locally optimal alignments
- ✓Allow multiple hits to the same sequence
- ✓Based on statistics of ungapped sequence alignments
- ✓The statistics allow the probability of obtaining an ungapped alignment
- ✓MSP - Maximal Segment Pair above cut-off
- ✓All world (k > 3) score grater than T
- ✓Extend the score both side
- ✓Use dynamic programming for narrow region

BLAST Algorithm

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

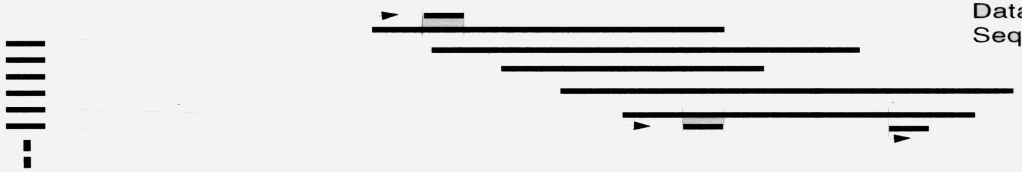


Query Sequence of length L

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 pairscore matrix (e.g. PAM 250). For typical parameters there are around 50 words per residue of the query..

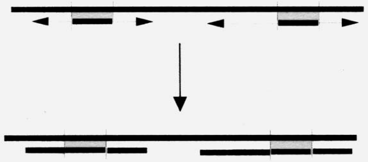
(2) Compare the word list to the database and identify exact matches.



Database Sequences

Exact matches of words from word list

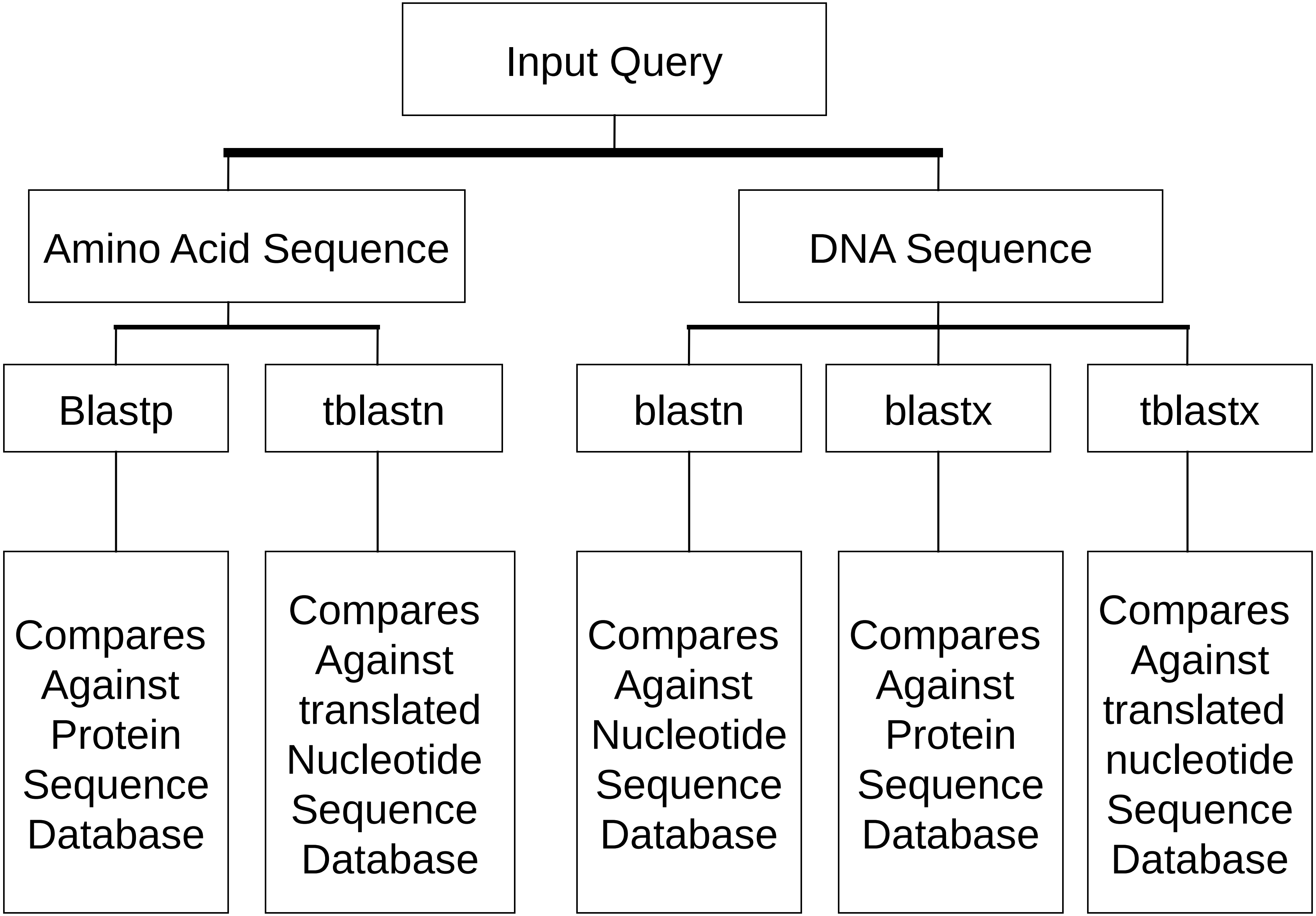
(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.



Maximal Segment Pairs (MSPs)

BLAST-Basic Local Alignment Search Tool

- Capable of searching all the available major sequence databases
- Run on nr database at NCBI web site
- Developed by Samuel Karlin and Stevan Altschul
- Method uses substitution scoring matrices
- A substitution scoring matrix is a scoring method used in the alignment of one residue or nucleotide against another
- First scoring matrix was used in the comparison of protein sequences in evolutionary terms by Late Margret Dayhoff and coworkers
- Matrices –Dayhoff, MDM, or PAM, BLOSUM etc.
- Basic BLAST program does not allow gaps in its alignments
- Gapped BLAST and PSI-BLAST



Input Query

Amino Acid Sequence

DNA Sequence

Blastp

tblastn

blastn

blastx

tblastx

Compares Against Protein Sequence Database

Compares Against translated Nucleotide Sequence Database

Compares Against Nucleotide Sequence Database

Compares Against Protein Sequence Database

Compares Against translated nucleotide Sequence Database

An Overview of BLAST

NCBI

Overview

Frequently Asked Questions

New/Noteworthy

NEW PSI-BLAST 2.1

NEW Search Conserved Domains with CD-Search

Receive e-mail about BLAST changes

BLAST course

BLAST Information/Tutorial

References

FTP Site

BLAST

NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein databases.

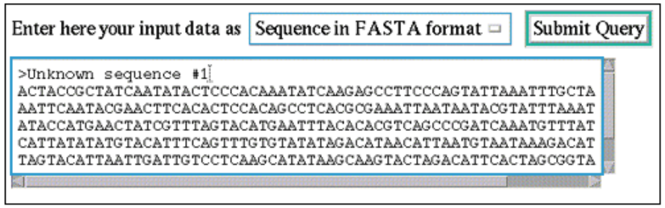
Please see the **BLAST Frequently Asked Questions** for tips on running BLAST searches.

NEW PSI-BLAST 2.1 is here! This version offers improved sensitivity with Composition based statistics

NEW Search the Conserved Domain Database with reverse position-specific Blast! See the link to **CD-Search** below.

BLAST 2.1

- Basic BLAST search
- Advanced BLAST search



Enter here your input data as Sequence in FASTA format

```
>Unknown sequence #1
ACTACCGCTATCAATATACTCCCACAAATATCAAGAGCCTTCCCAGTATTAAATTTGCTA
AATTCAATACGAACTTCACACTCCACAGCCTCACGCGAAATTAATAATACGTATTTAAAT
ATACCATGAACTATCGTTTAGTACATGAATTTACACACGTCAGCCCGATCAAATGTTTAT
CATTATATATGTACATTTCAGTTTGTGTATATAGACATAACATTAATGTAATAAAGACAT
TAGTACATTAATTGATTGTCCTCAAGCATATAAGCAAGTACTAGACATTCACTAGCGGTA
```

Searches against Profile Databases

- CD-Search: search the Conserved Domain Database using RPS-Blast

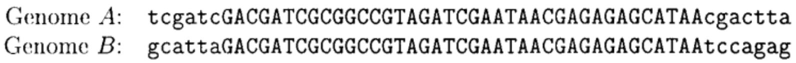
Position Specific Iterated BLAST

- PSI-BLAST search

Database Scanning or Fold Recognition

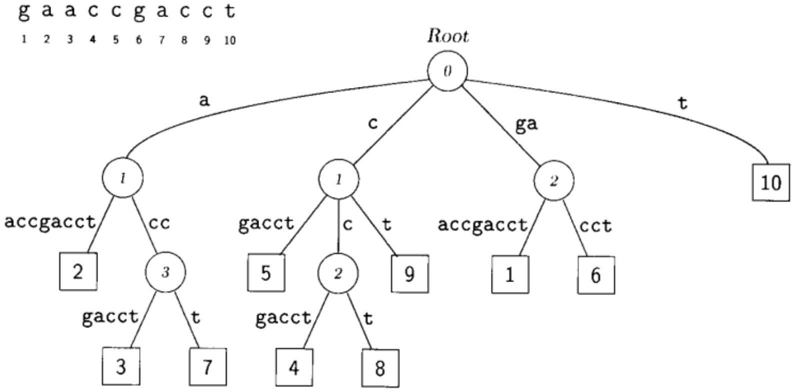
- **Concept of PSIBLAST**
  - Perform the BLAST search (gap handling)
  - GeneImprove the sensivity of BLAST
  - rate the position-specific score matrix
  - Use PSSM for next round of search
- **Intermediate Sequence Search**
  - Search query against protein database
  - Generate multiple alignment or profile
  - Use profile to search against PDB

Comparison of Whole Genomes



Genome *A*: tcgatcGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAcgactta
Genome *B*: gcattaGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAtccagag

- **MUMmer (Salzberg group, 1999, 2002)**
  - Pair-wise sequence alignment of genomes
  - Assume that sequences are closely related
  - Allow to detect repeats, inverse repeats, SNP
  - Domain inserted/deleted
  - Identify the exact matches
- **How it works**
  - Identify the maximal unique match (MUM) in two genomes
  - As two genome are similar so larger MUM will be there
  - Sort the matches found in MUM and extract longest set of possible matches that occurs in same order (Ordered MUM)
  - Suffix tree was used to identify MUM
  - Close the gaps by SNPs, large inserts
  - Align region between MUMs by Smith-Waterman



Thanks