Database Scanning/Searching FASTA/BLAST/PSIBLAST

G P S Raghava

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)

Database scanning

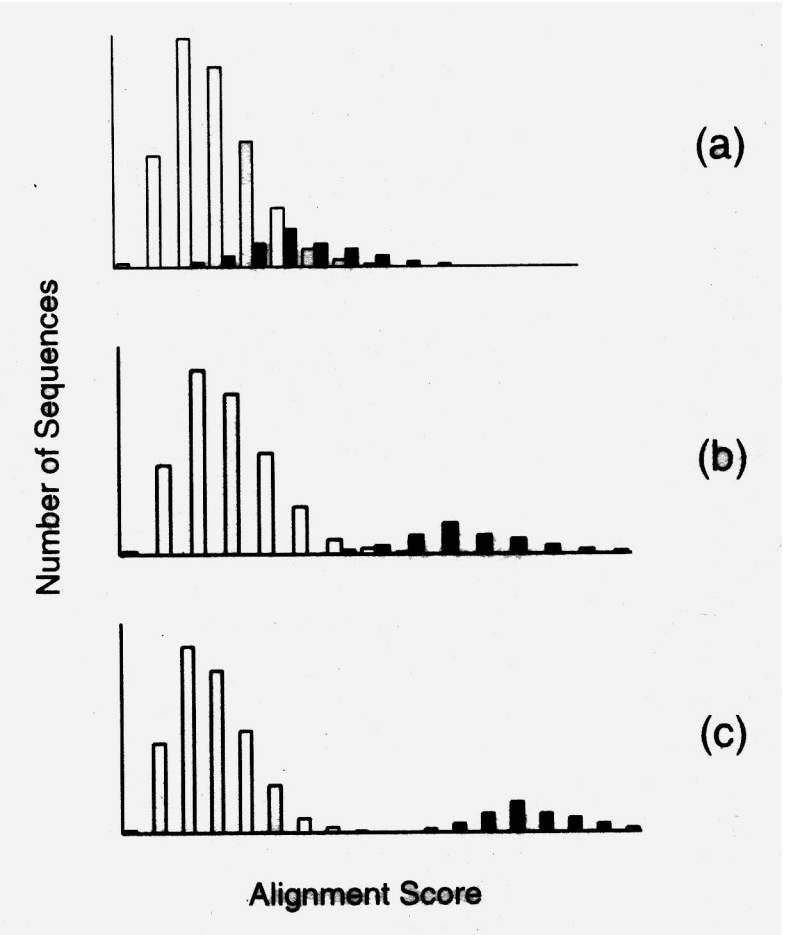
Basic principles of Database searching

- 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



WHAT IS A HEURISTIC METHOD?

Definition: A *heuristic method* is a method that uses "rules of thumb" to give an answer that, while it is not necessarily exact, is fast and is based on sound reasoning.

When comparing two sequences, exact methods such as dynamic programming is possible. When comparing a sequence against a database, dynamic programming consumes far too much computer time and space. Heuristic methods are needed.

The two most widely used heuristic methods in sequencing are FASTA and BLAST.

The time is close to linear in time and space (rather than quadratic as with dynamic programming.)

The FASTA algorithm

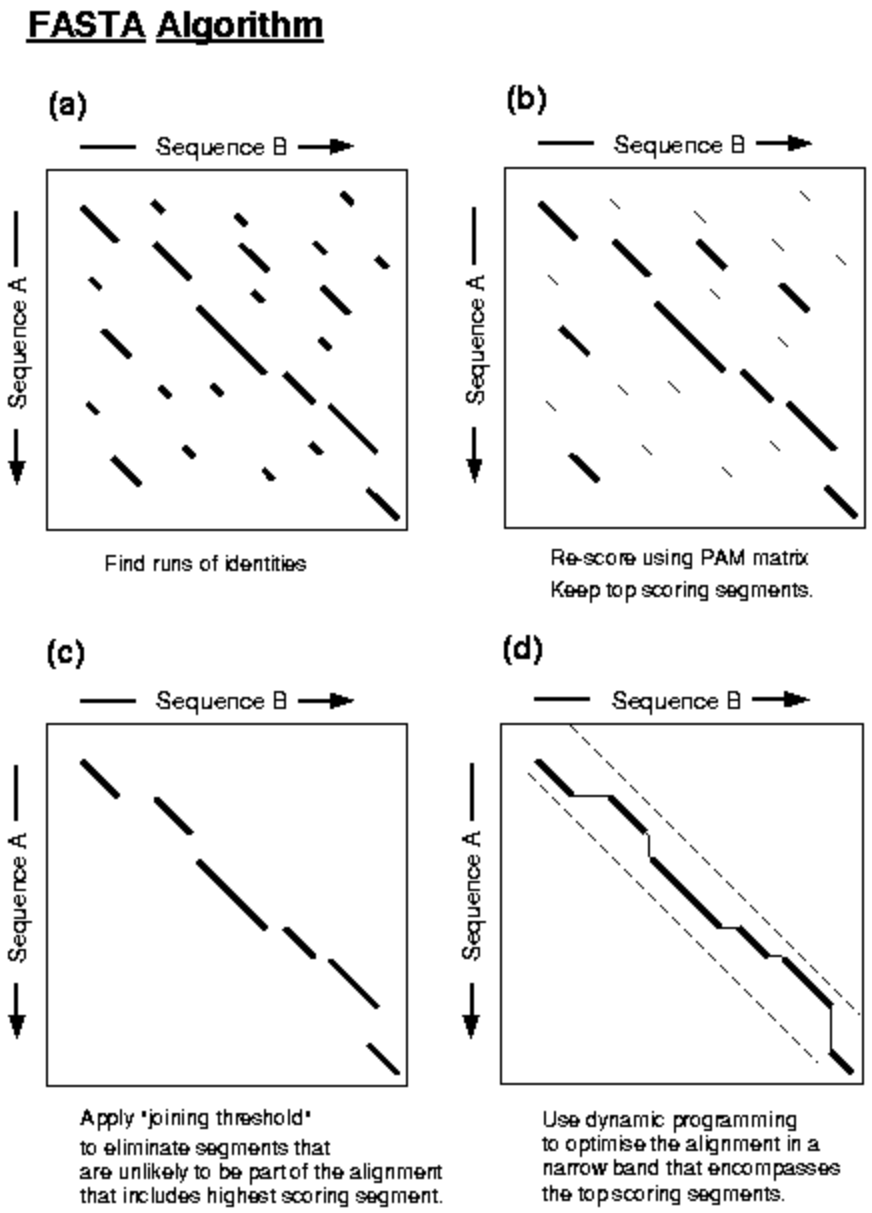
- Four steps:
  1) Identify regions of similarity:
     - Using the *ktup* parameter which specifies # consecutive identities required in a match
     - 10 best diagonal regions found based on #matches and distance between matches
  2) Rescore regions and identify best initial regions
     - PAM250 or other scoring matrix used for rescoring the 10 diagonal regions identified in step 1 to allow for conservative replacements and runs of identities shorter than *ktup*
     - For each the best diagonal regions, identify “initial region” that is best scoring subregion

The FASTA algorithm

3) Optimally join initial regions with scores > T
   - Given: location of initial regions, scores, gap penalty
   - Calculate an optimal alignment of initial regions as a combination of compatible regions with maximal score
   - Use resulting score to rank the library sequences
   - Selectivity degradation limited by using initial regions that score greater than some threshold T

4) Align the highest scoring library sequences using modification of global and local alignment algorithms
   - Considers all possible alignments of the query and library sequence that falls within a band centered around the highest scoring initial region

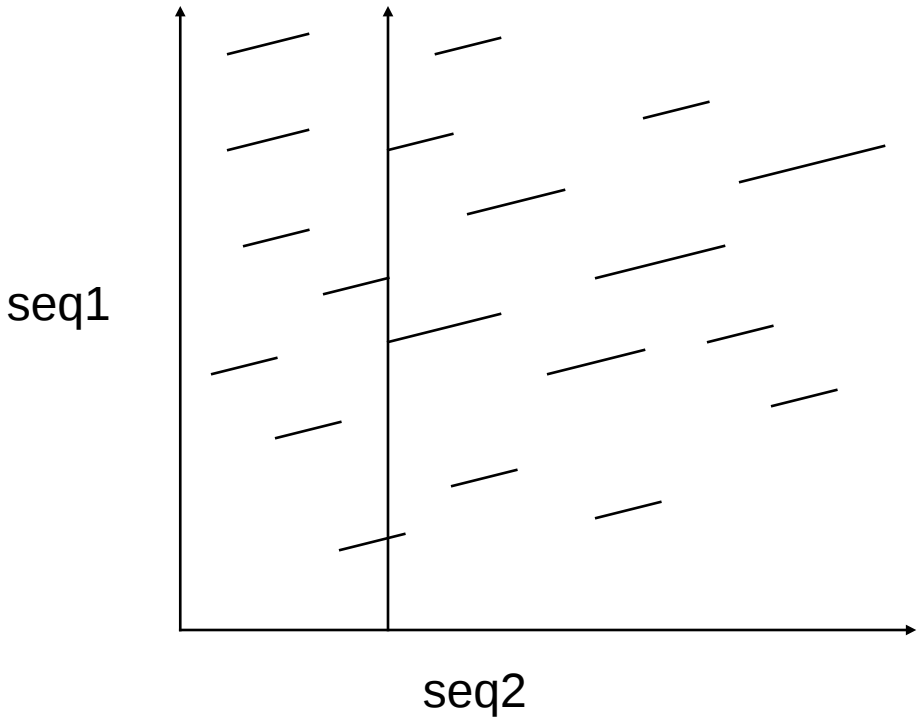
The Four-Step FASTA Process



Overview of the fasta algorithm

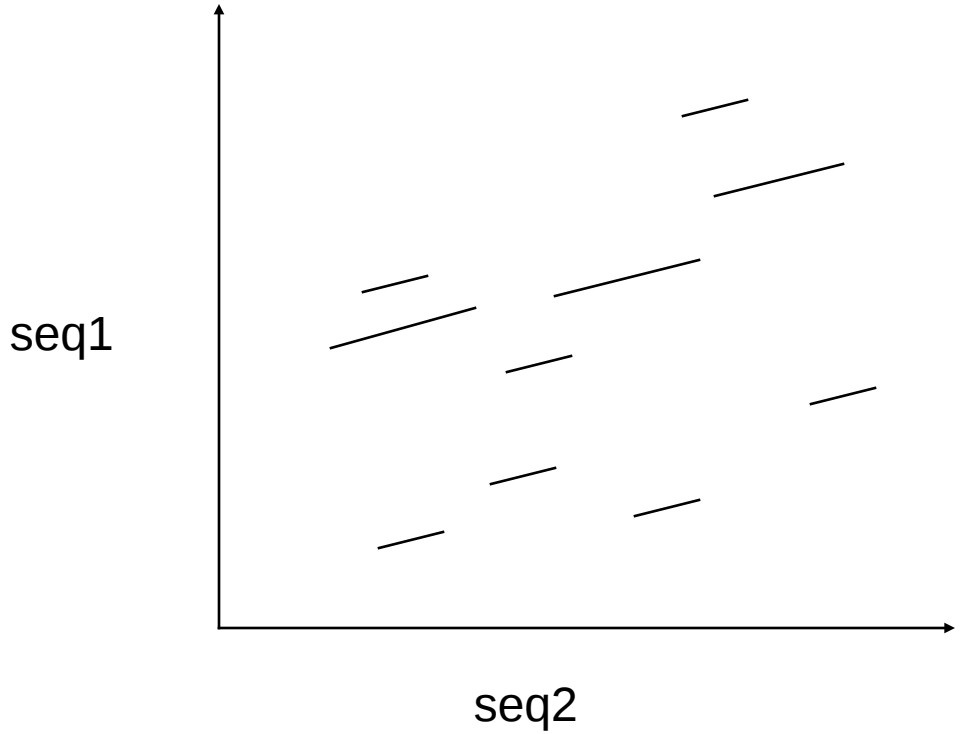
1.

FastA locates regions of the query sequence and the search set sequence that have high densities of exact word matches.
For DNA sequences the word length usually used is 6.



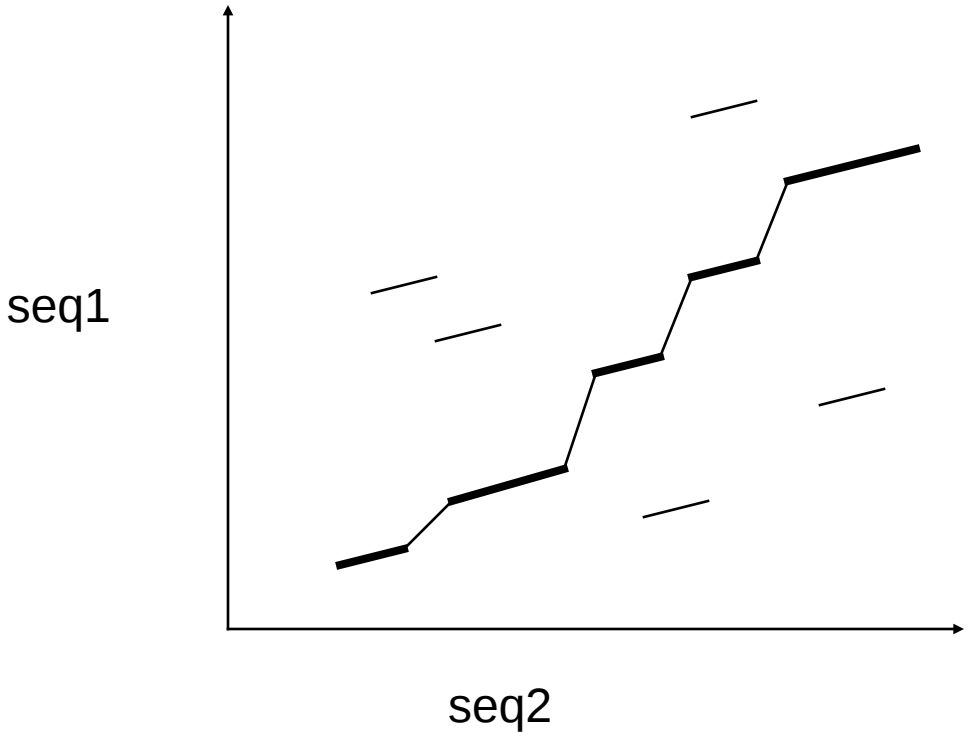
2.

The 10 highest-scoring sequence regions are saved and re-scored using a scoring matrix. These scores are the **init1 scores**



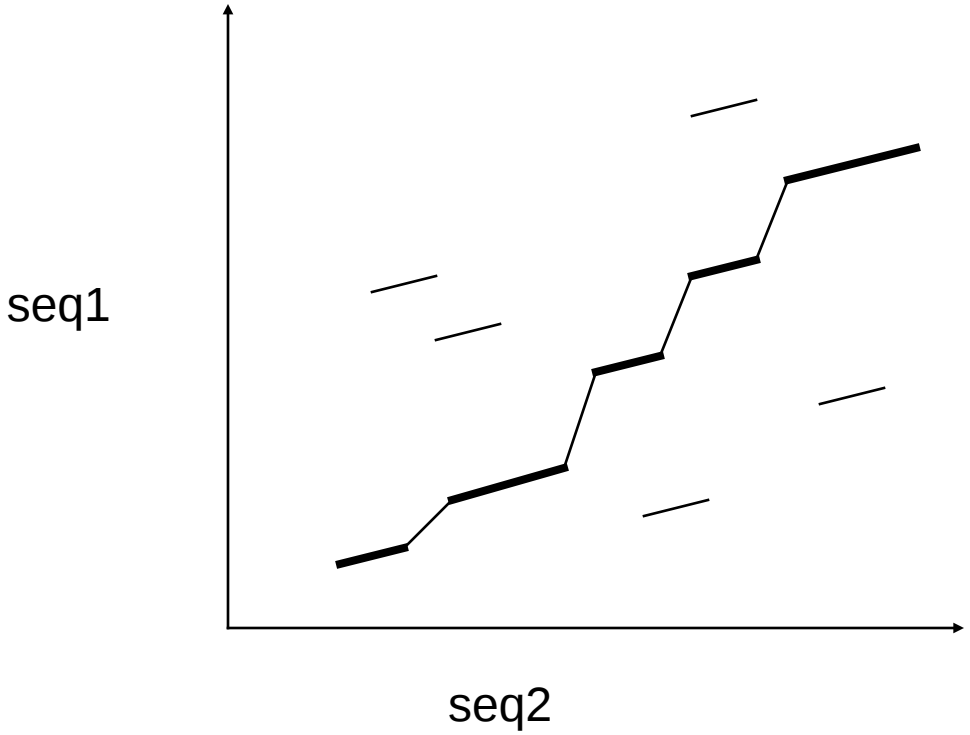
3

FastA determines if any of the initial regions from different diagonals may be joined together to form an approximate alignment with gaps. Only non-overlapping regions may be joined.



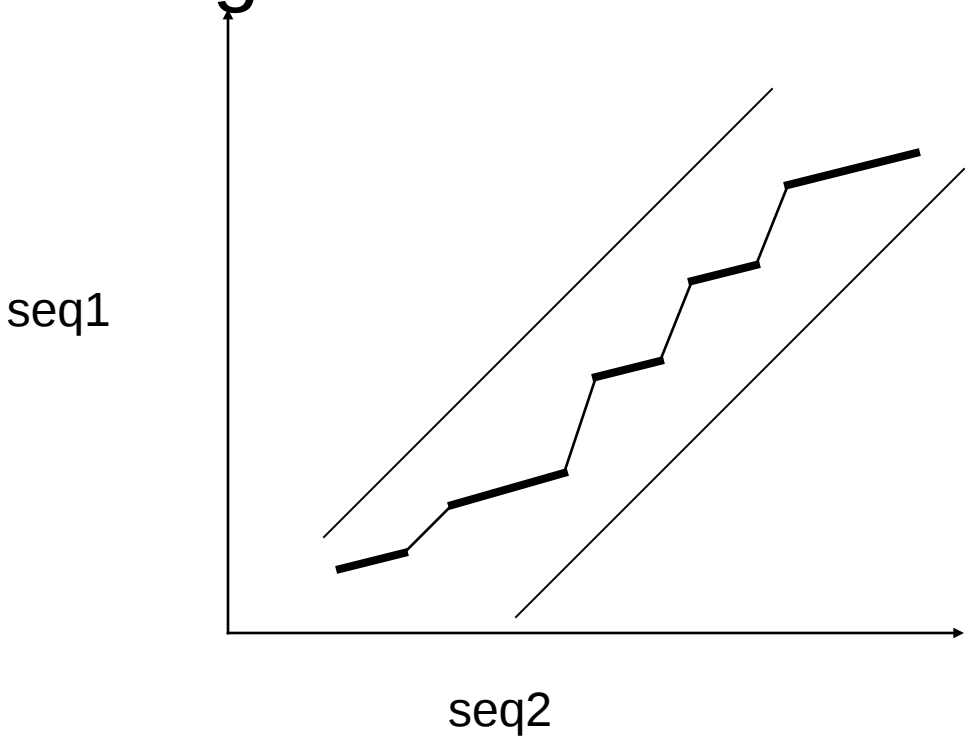
3 - cont

The score for the joined regions is the sum of the scores of the initial regions minus a joining penalty for each gap. The score of the highest scoring region, at the end of this step, is saved as the **initn score**.



4

FastA uses dynamic programming (Smith-Waterman algorithm) over a narrow band of high scoring diagonals between the query sequence and the search set sequence, to produce an alignment with a new score.



Opt_score

Overview of the fasta algorithm

- Search for regions with exact word matches
- keep 10 highest scoring regions and re-score them using a scoring matrix : Init1
- Join diagonals by introducing gaps : Initn
- Apply Smith-Waterman algorithm to achieve best alignment : Opt
- Correct for sequence lengths : Z-score
- Evaluate significance of Z_scores: E values

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 – Basic Local Alignment Search Tool

- Find the highest scoring locally optimal alignments between a query sequence and a database.

- Very fast algorithm

- Can be used to search extremely large databases
(uses a pre-indexed database which contributes to its great speed)

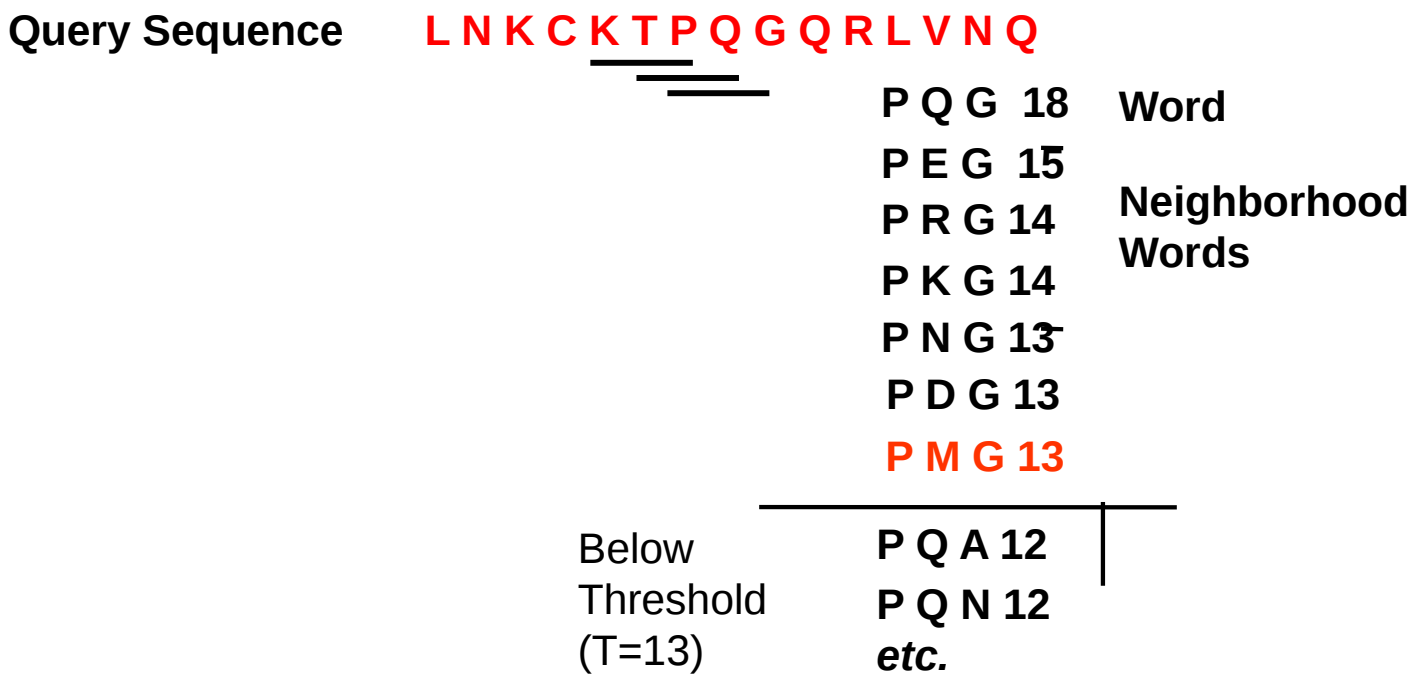
- Sufficiently sensitive and selective for most purposes

- Robust – the default parameters can usually be used

BLAST Algorithm, Step 1

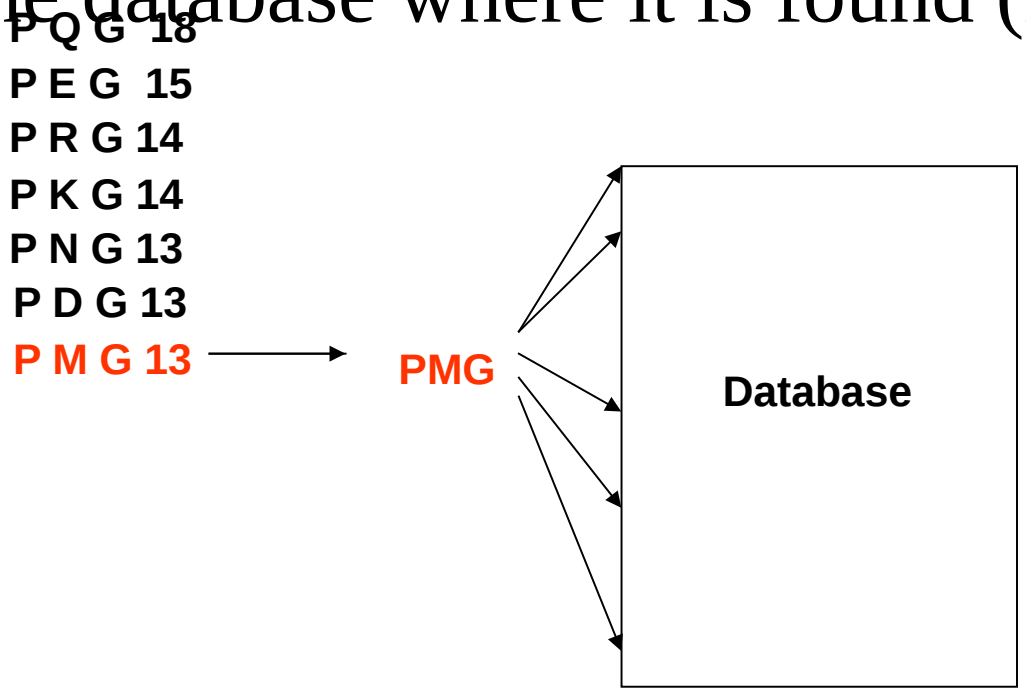
- For a given word length *w* (usually 3 for proteins) and a given score matrix:

  Create a list of all words (*w*-mers) that can can score $>T$ when compared to *w*-mers from the query.



BLAST Algorithm, Step 2

- Each neighborhood word gives all positions in the database where it is found (hit list).



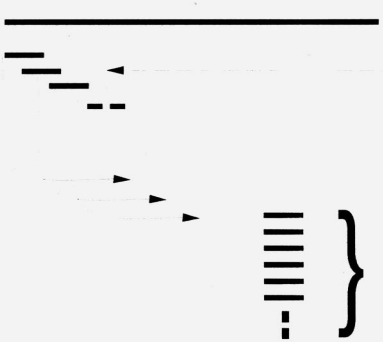
BLAST Algorithm, Step 3

- The program tries to extend matching segments (seeds) out in both directions by adding pairs of residues. Residues will be a
b

High-scoring Segment Pair (HSP)

BLAST Algorithm

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

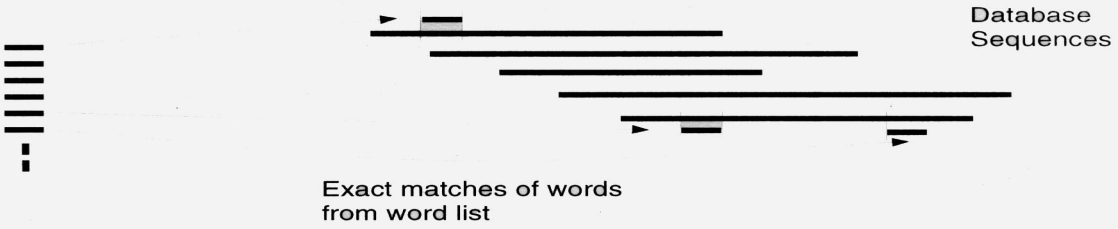


Query Sequence of length L

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..

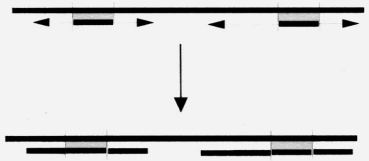
(2) Compare the word list to the database and identify exact matches.



Database Sequences

Exact matches of words from word list

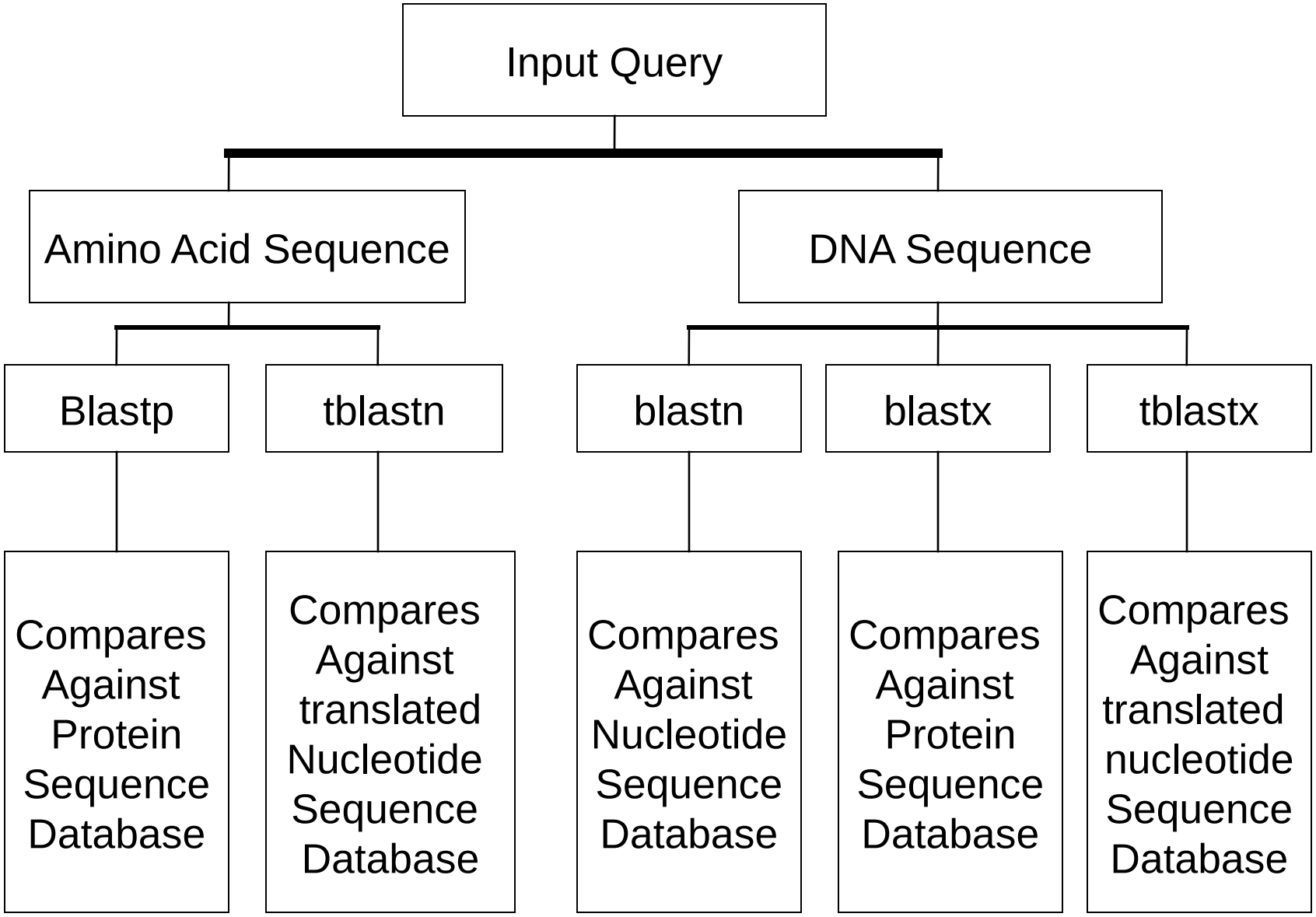
(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.



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



An Overview of BLAST

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

Searches against Profile Databases

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

Position Specific Iterated BLAST

- PSI-BLAST search

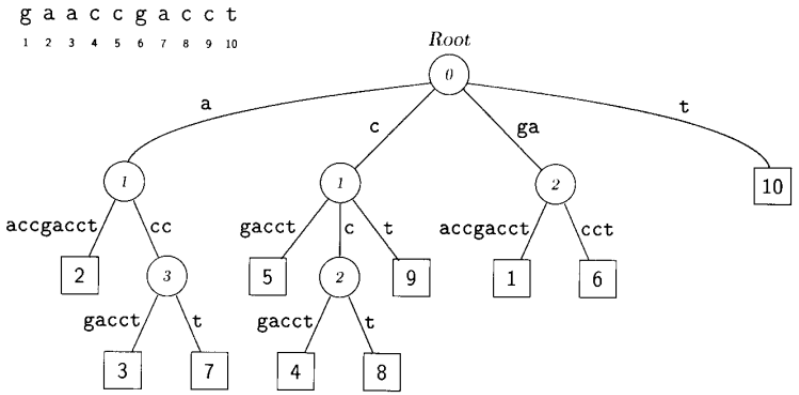
Database Scanning or Fold Recognition

- **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