FASTA: SIMILARITY SEARCHING AND ITS APPLICATIONS

WHAT IS FASTA?

FASTA is a database similarity search tool which uses a standard format for sequence data of DNA and proteins. First developed by Lipman and Pearson, it was used to compare protein sequences against protein databases. But now it is used to compare both DNA and protein sequences against various databases.

FASTA uses a “hashing” strategy to find matches for a short stretch of identical residues with a length of k. Typically, a k-tuple is composed of two residues for protein sequences and six residues for DNA sequences.

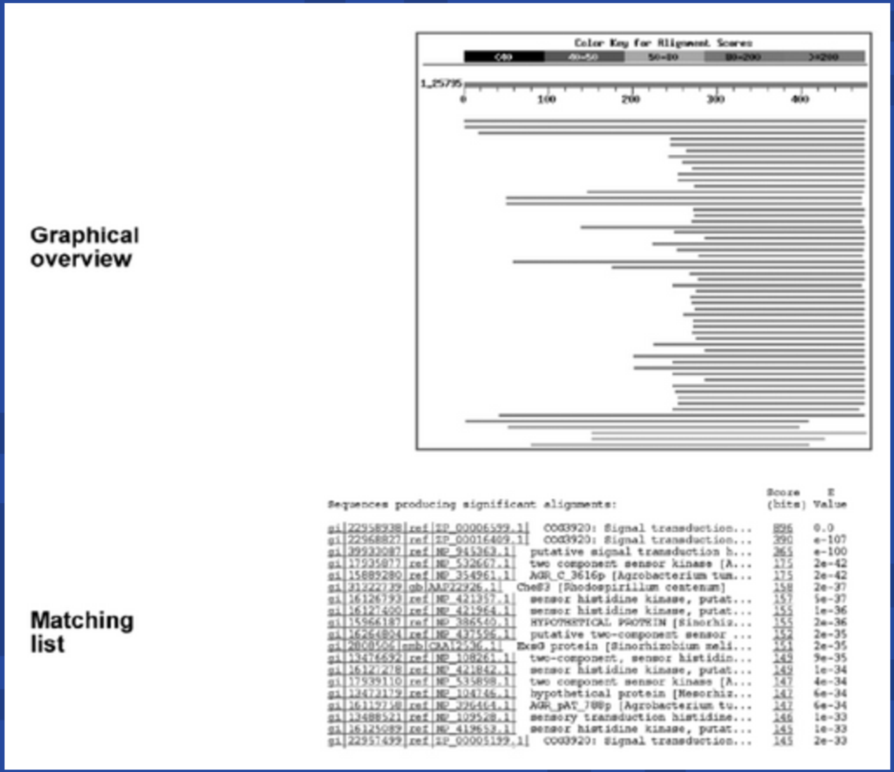
TYPES OF FASTA

- **FASTA** – compares DNA/protein sequence against DNA/protein database respectively.
- **SSEARCH** – performs protein-protein or DNA-DNA comparisons using local alignment algorithm.
- **GGSEARCH/GLSEARCH** – works using global alignment (GGSEARCH) or a combination of global-local alignment (GLSEARCH) to compare protein and nucleotide sequence.
- **FASTX/FASTY** – compares a DNA sequences to protein database by translating the DNA sequence into 3 frames and allowing gaps and frameshifts.
- **TFASTX/TFASTY** – compares a protein sequence to a DNA database by translating the DNA sequence into 6 frames, 3 in the forward direction and 3 in the reverse direction.
- **FASTF/TFASTF** – compares a mixed peptide sequence against a protein database (FASTF) or translated DNA database (TFASTF).

HOW IT WORKS?

FASTA works by comparing the query sequence to a database of sequences to identify similar matches. It uses heuristic algorithms to perform the searches and identify significant matches based on statistical parameters.

The mechanism involves 4 steps: identifying high similarity regions, re-scoring of the best aligned sequences, joining threshold to remove unlikely segments and final alignment of the new sequence.

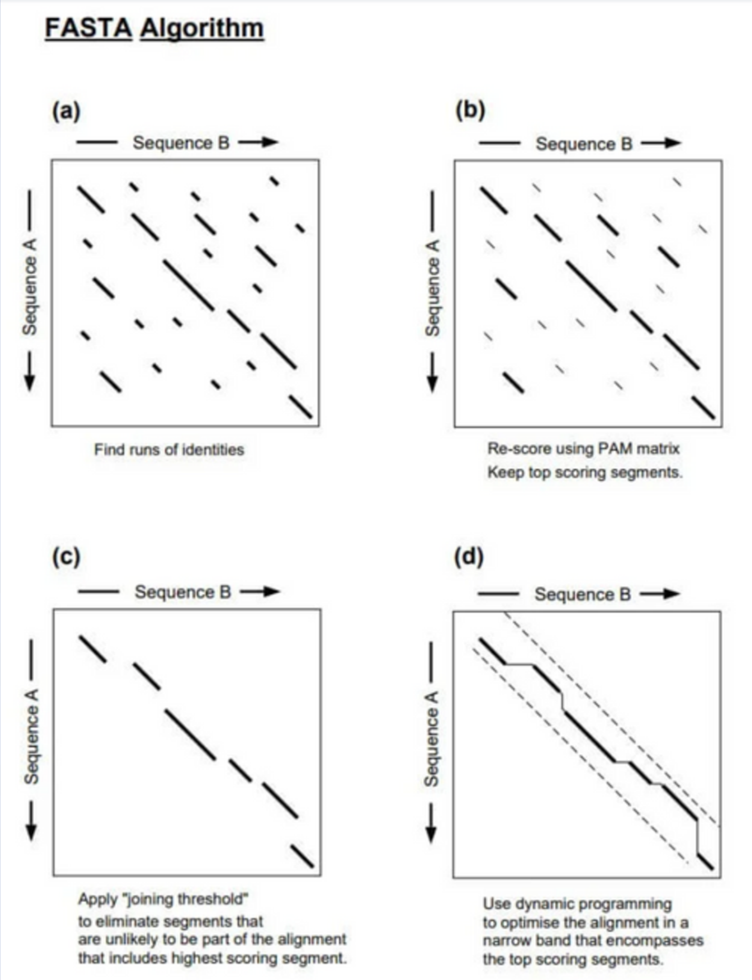


```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|22958938|ref|ZP_00006599.1|  COG3920: Signal transduction...   896    0.0
gi|22968827|ref|ZP_00016409.1|  COG3920: Signal transduction...   390    e-107
gi|39933087|ref|NP_945363.1|  putative signal transduction h...   365    e-100
gi|17935877|ref|NP_532667.1|  two component sensor kinase [A...   175    2e-42
gi|15889280|ref|NP_354961.1|  AGR_C_3616p [Agrobacterium tum...   175    2e-42
gi|31222739|gb|AAP22926.1|  CheS3 [Rhodospirillum centenum]       158    2e-37
gi|16126793|ref|NP_421357.1|  sensor histidine kinase, putat...   157    5e-37
gi|16127400|ref|NP_421964.1|  sensor histidine kinase, putat...   155    1e-36
gi|15966187|ref|NP_386540.1|  HYPOTHETICAL PROTEIN [Sinorhiz...   155    2e-36
gi|16264804|ref|NP_437596.1|  putative two-component sensor ...   152    2e-35
gi|2808506|emb|CAA12536.1|  ExsG protein [Sinorhizobium meli...   151    2e-35
gi|13476692|ref|NP_108261.1|  two-component, sensor histidin...   149    9e-35
gi|16127278|ref|NP_421842.1|  sensor histidine kinase, putat...   149    1e-34
gi|17939110|ref|NP_535898.1|  two component sensor kinase [A...   147    4e-34
gi|13473179|ref|NP_104746.1|  hypothetical protein [Mesorhiz...   147    6e-34
gi|16119758|ref|NP_396464.1|  AGR_pAT_788p [Agrobacterium tu...   147    6e-34
gi|13488521|ref|NP_109528.1|  sensory transduction histidine...   146    1e-33
gi|16125089|ref|NP_419653.1|  sensor histidine kinase, putat...   145    1e-33
gi|22957499|ref|ZP_00005199.1|  COG3920: Signal transduction...   145    2e-33
```

1. IDENTIFYING SIMILAR REGIONS

- FASTA identifies regions with high similarity by creating a lookup table by hashing method. The query is broken down to k-tuples.
- K-tuple values are increased to reduce the number of background hits, so it focuses more on the significant hits. K-tuple is 2 for proteins and 6 for nucleotides.
- The similar regions are plotted in a 2D matrix as diagonals and the top scoring diagonals are saved which are having the highest similarity.



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.

2. RE-SCORING

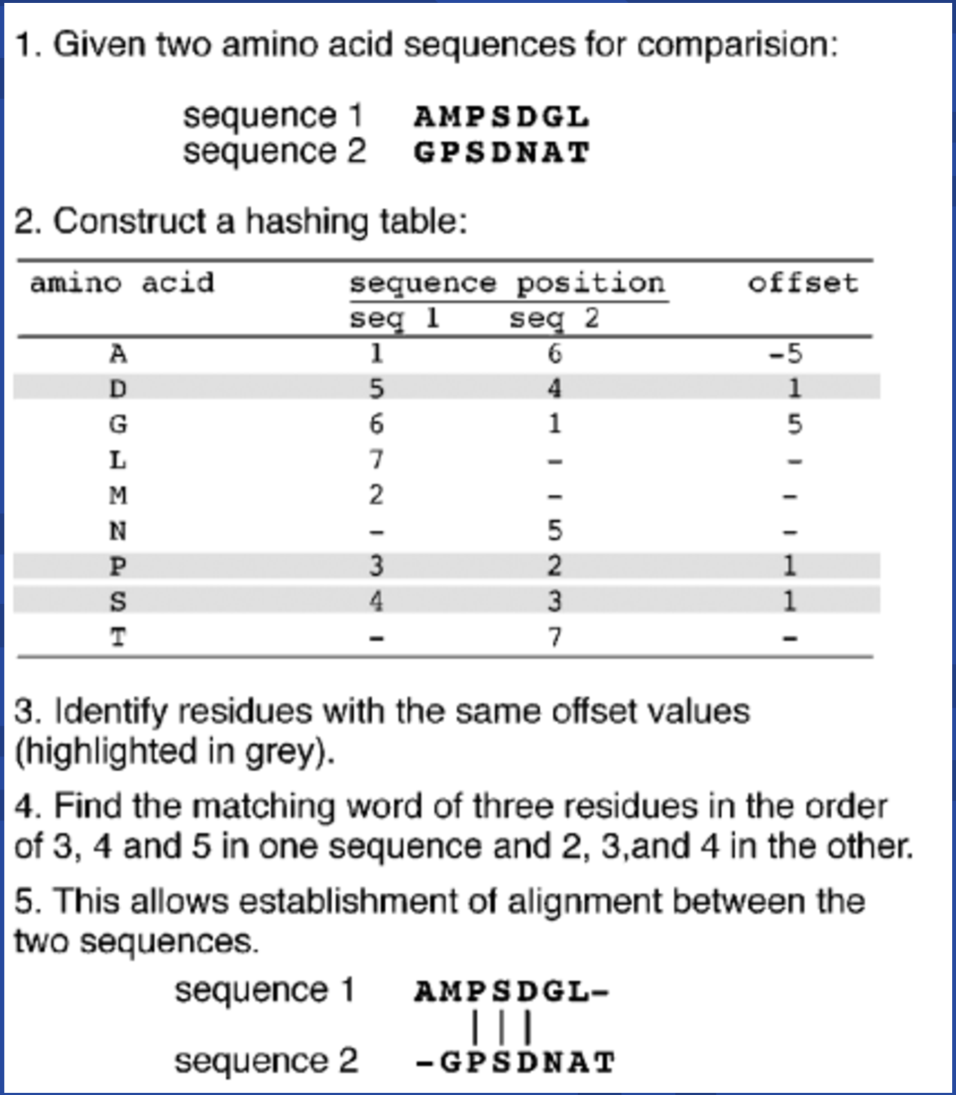
- The 10 best diagonals are re-scored using scoring matrices – BLOSUM50 for proteins and identity matrices for DNA. A sub region with the highest score is identified for each of the diagonals which are called initial regions.

3. JOINING THRESHOLD

- A joining threshold is applied that excludes segments unlikely to be part of the final alignment. The selected regions with initial scores above the pre-set threshold are joined. This introduces gaps between diagonals while applying gap penalties. The score of the gapped alignment is calculated by subtracting a penalty for each gap, which is used to rank the database sequence by similarity.

4. FINAL ALIGNMENT

Finally, the alignment is refined to produce the final alignment. This is done by using the banded Smith-Waterman algorithm, which is a dynamic programming algorithm that calculates the optimal score (opt) for alignment. This score is used for statistical calculations when measuring e-value, bit scores and finally, functional domains are studied and a phylogenetic analysis is performed for the given species.



1. Given two amino acid sequences for comparision:

```
sequence 1   AMPSDGL
sequence 2   GPSDNAT
```

2. Construct a hashing table:

amino acid	sequence position		offset
	seq 1	seq 2	
A	1	6	-5
D	5	4	1
G	6	1	5
L	7	-	-
M	2	-	-
N	-	5	-
P	3	2	1
S	4	3	1
T	-	7	-

3. Identify residues with the same offset values (highlighted in grey).

4. Find the matching word of three residues in the order of 3, 4 and 5 in one sequence and 2, 3,and 4 in the other.

5. This allows establishment of alignment between the two sequences.

```
sequence 1   AMPSDGL-
               |||
sequence 2   -GPSDNAT
```

STATISTICAL SIGNIFICANCE OF FASTA

- FASTA provides an estimate of statistical significance of each alignment found, which is evaluated using E-value (the likelihood of obtaining a sequence alignment score by chance). Smaller the E-value, more significant is the alignment.
- FASTA also uses bit scores and similarity scores based on the scoring matrix and gap penalties, to evaluate the significance of sequence alignments.
- Z-score is another parameter that represents the number of standard deviations from the mean score of the database search. A higher z-score indicates a higher similarity match.

APPLICATIONS OF FASTA

SIMILARITY SEARCHING

Used to identify similar regions in protein and DNA sequences to understand conserved domains or motifs.

FUNCTIONAL ANNONATION

Used to search database of sequences to identify homologous sequences to predict the function of a newly identified sequence.

PHYLOGENETIC ANALYSIS

Multiple sequence alignment can be done to plot phylogenetic trees by identifying evolutionary relationships between species.