Overview

- Introduction
- Heuristic approach
- Steps in the algorithm

Introduction

- The FASTA algorithm is a heuristic method for string comparison. It was developed by Lipman and Pearson in 1985 and further improved in 1988.
- It performs Local alignment of sequences.

Heuristic Approach

- Definition: A heuristic method is a method that uses hash coding in which the sequencce is broken into small “words or K-tuples”of specific sizes.
- Dynamic programming methods consume time to analyse and search the entire database.Heuristic approach overcomes this drawback.

FASTA - Idea -

- Problem of Dynamic Programming

  D.P. compute the score in a lot of useless area for optimal sequence

	G	A	A	T	T	C	A	G	T	T	A
G	1	1	1	1	1	1	1	1	1	1	1
G	1	1	1	1	1	1	1	2	2	2	2
A	1	2	2	2	2	2	2	2	2	2	2
T	1	2	2	3	3	3	3	3	3	3	3
C	1	2	2	3	3	4	4	4	4	4	4
G	1	2	2	3	3	4	4	5	5	5	5
A	1	2	3	3	3	4	5	5	5	5	6

- FASTA focuses on diagonal area

Steps in FASTA Algorithm

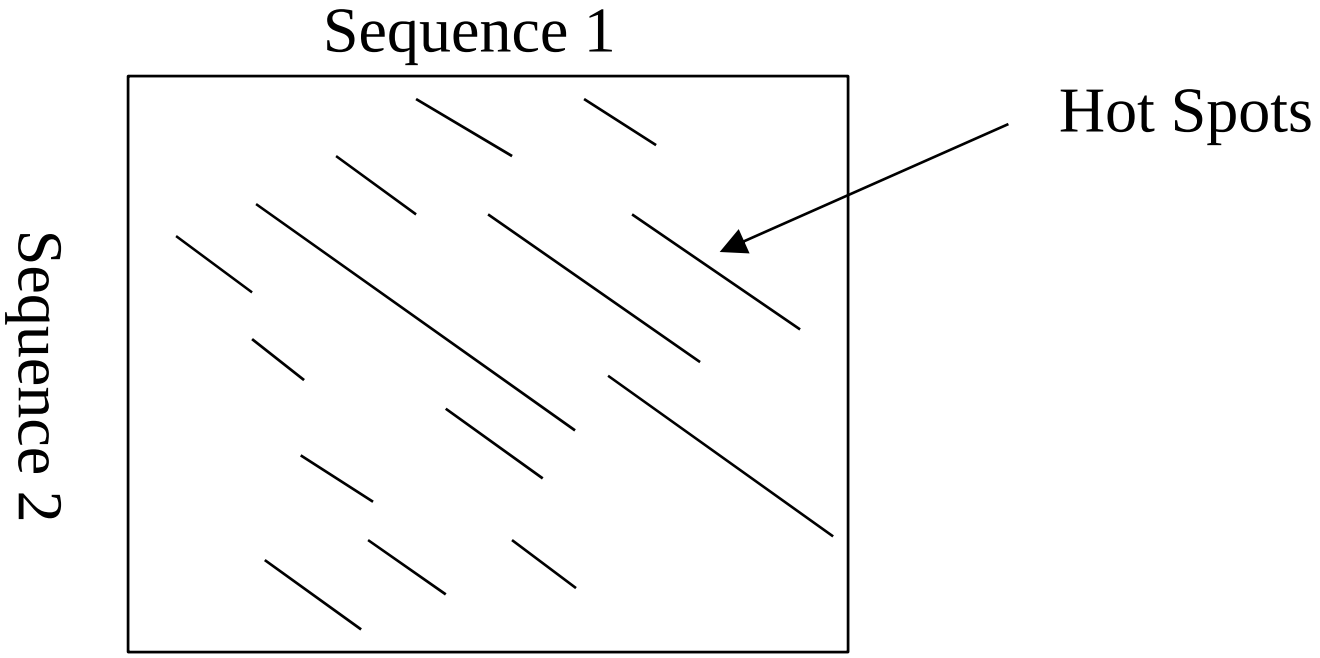
- Four steps:
  1) Identify regions of similarity:
     - Using the *ktup* parameter which specifies # consecutive identities required in a match
     - 10 best diagonal regions found based on #matches and distance between matches
  2) Rescore regions and identify best initial regions
     - PAM250 or 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

FASTA - Algorithm -

- Step 1

  Find all hot-spots

  // Hot spots is pairs of words of length k that exactly match



Steps in 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

OVERVIEW OF THE FASTA ALGORITHM

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

FASTA - Algorithm -

- Step 1 in detail

Use look-up Table

Query : G A A T T C A G T T A

Sequence: G G A T C G A

Look-up Table

Q	Location
A	2,3,7,11
C	6
G	1,8
T	4,5,9,10

Dot—Matrix

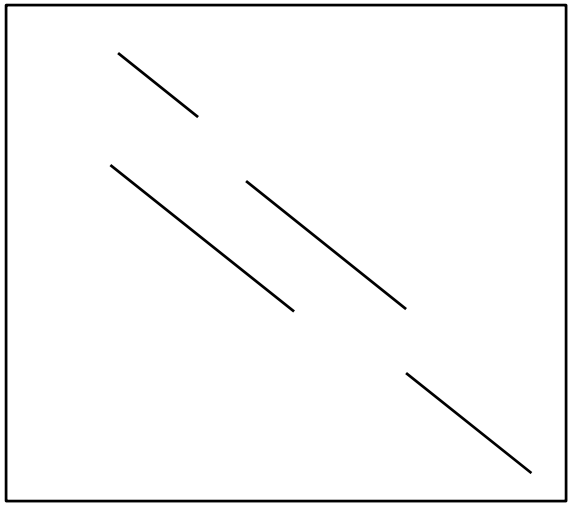
	G	A	A	T	T	C	A	G	T	T	A
G	*							*			
G	*							*			
A		*	*				*				*
T				*	*				*	*	
C						*					
G	*							*			
A		*	*				*				*

FASTA - Algorithm -

- Step 2

Score the Hot-spot and locate the ten best diagonal run.

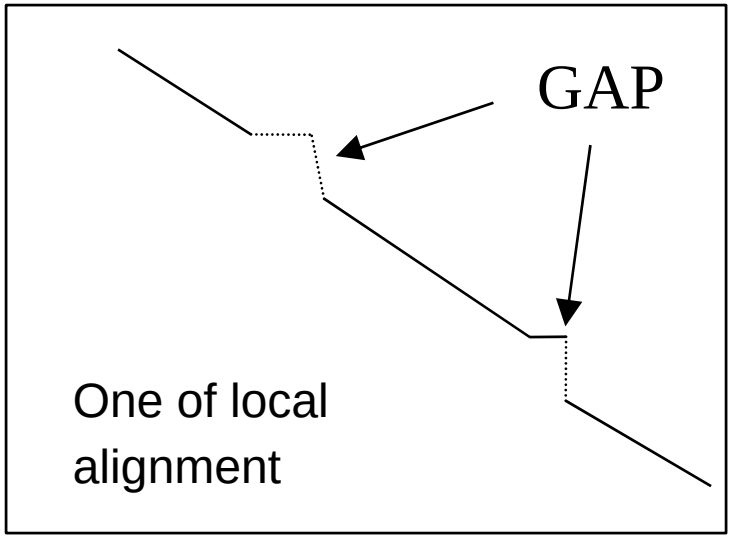
// There is some scoring system; ex. PAM250



FASTA - Algorithm -

- Step 3

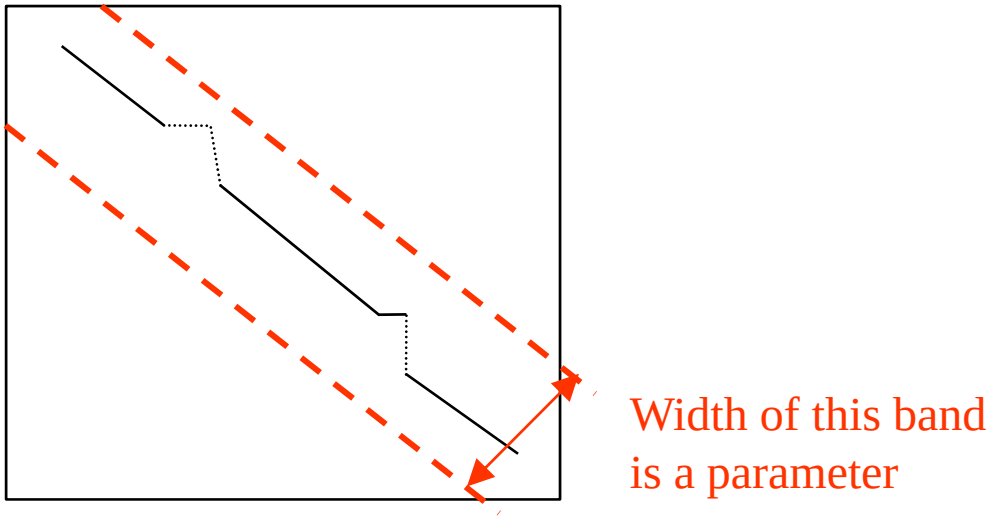
  Combine sub-alignments into one alignment with GAP



FASTA - Algorithm -

- Step 4

  Use the dynamic programming in restricted area around the best-score alignment to find out the higher-score alignment than the best-score alignment



Conclusion

Algorithm	Sensitivity	Running Time
D.P	1	3
FASTA	3	2
BLAST	2	1

Reference

- N Gautham, Introduction to bioinformatics databases and algorithms.
- T.K.Atwood,Parry Smith and Samiron Pukhan ,introduction to bioinformatics.

Thank you