Sequence Alignment: III

Dr Safdar Ali
Department of Biological Sciences
Aliah University, Kolkata
Date: 04/04/2020

Topics to be covered

- Sequence formats
- Dot Plot
- Global and Local alignment
- Scoring Matrices
- Needleman Wunsch Algorithm
- Smith Waterman Algorithm
- BLAST and FASTA
- HMM
- Multiple sequence alignment

Need for BLAST and FASTA

- The Needleman–Wunsch algorithm requires **O(mn)** steps, while the Smith–Waterman algorithm requires **O(m^{2}n)** steps.

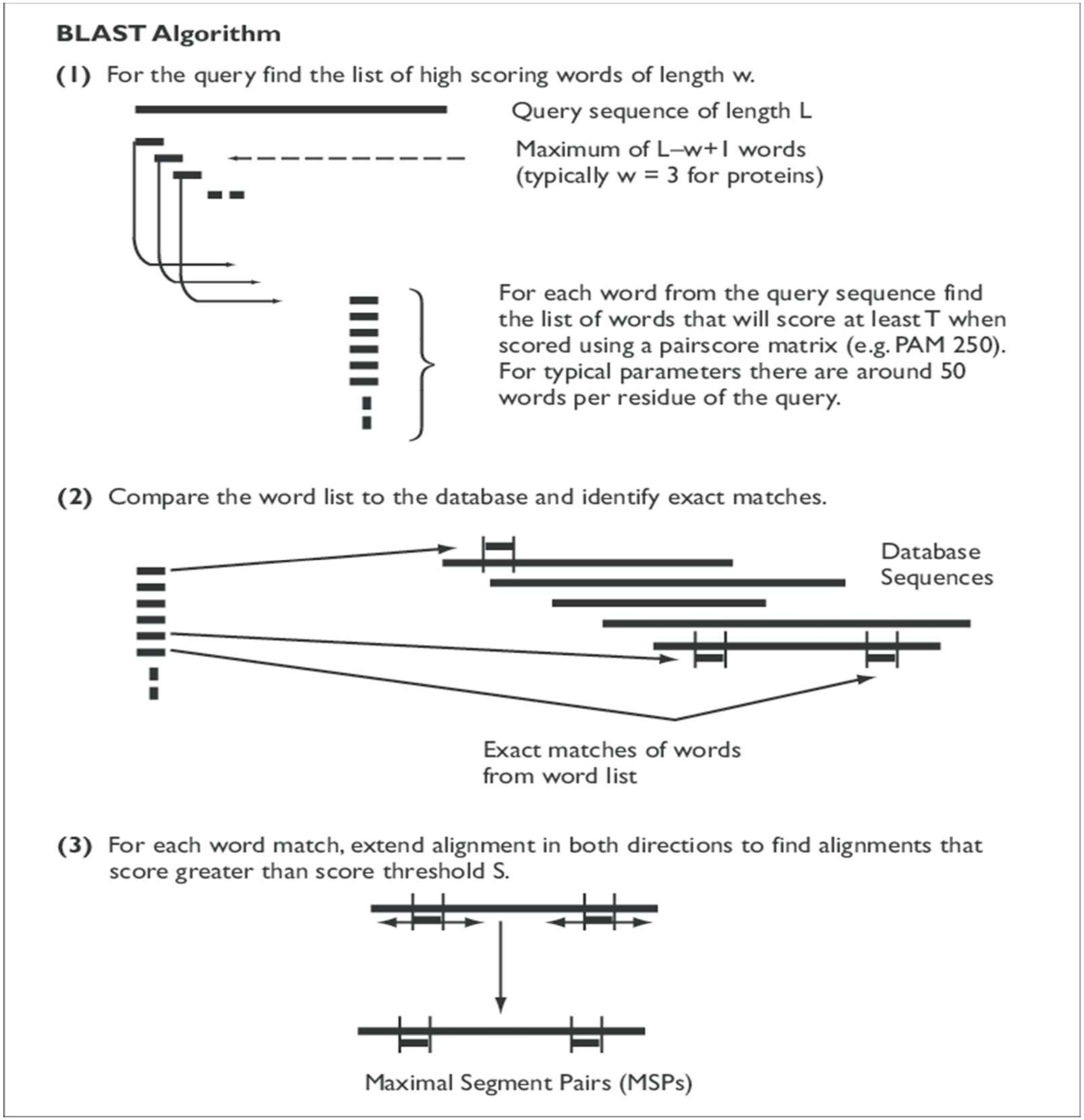
- Two popular local alignment algorithms have been developed that provide rapid alternatives are
  - FASTA (Pearson and Lipman, 1988) and
  - BLAST(Basic Local Alignment Search Tool) (Altschul et al., 1990).

BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- Set of alignment algorithms to
  - Find a short fragment of a query sequence
  - Aligns with a fragment of a subject sequence in a database

BLAST Algorithm

- Sequence (query) is broken into words of length W
- Align all words with sequences in the database
- Calculate score T for each word that aligns with a sequence in the database using a substitution matrix (PAM or BLOSUM)
- Discard words whose T value is below a threshold
- Extend words in both directions until score falls by dropoff value X when compared to previous best score



BLAST Algorithm

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

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.

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

https://www.researchgate.net/figure/Schematic-illustration-of-the-BLAST-algorithm-Originally-published-in_fig1_11875325

Types of BLAST

	Query	Database	Word size
BLASTN	Nucleotide	Nucleotide	11
BLASTX	Translated nucleotide in all six frames	Protein	3
BLASTP	Protein	Protein	3
TBLASTN	Protein	Translated nucleotide in all six frames	3
TBLASTX	Translated nucleotide in all six frames	Translated nucleotide in all six frames	3

E Value

- The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.
- It decreases exponentially as the Score (S) of the match increases.
- **E value describes the random background noise.**
- The lower the E-value, or the closer it is to zero, the more "significant" the match is.
  - **$<10^{-4}$: Suggest significant homology**
  - **10^{-4} to 10^{-2} : Similar domains, non homologous**
  - **10^{-4} to 1: No significant homology**

- Identical short alignments have relatively high E values because the calculation of the E value takes into account the length of the query sequence.
- The gapped BLAST algorithm allows for gaps to be introduced in the alignments and is often more accurate. This also introduces additional parameters k and λ for the calculation of E value for a particular score S.
- Mathematically,

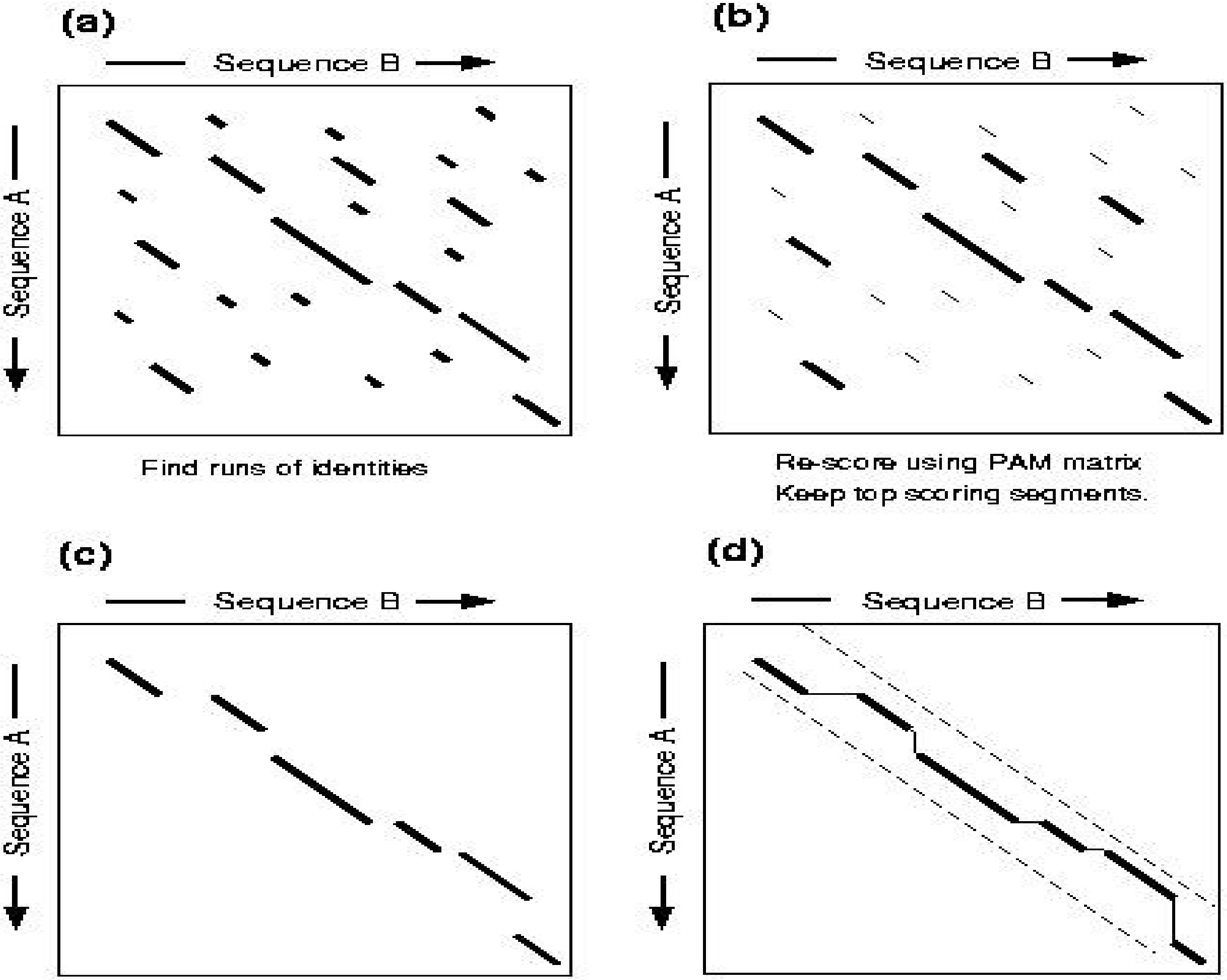
$$E = Kmne^{\lambda S}$$

m: length of database

n: length of query sequence

FASTA Algorithm

1. A lookup table is generated consisting of short stretches of amino acids or nucleotides from a database. The size of these stretches is determined from the ***ktup*** parameter. If ***ktup = 3*** for a protein search, then the query sequence is examined in blocks of three amino acids against matches of three amino acids found in the lookup table. The FASTA program identifies the 10 highest scoring segments that align for a given ktup.
2. These 10 aligned regions are rescored, allowing for conservative replacements, using a scoring matrix such as PAM250.
3. High-scoring regions are joined together if they are part of the same proteins.
4. FASTA then performs a global (Needleman–Wunsch) or local (Smith–Waterman) alignment on the highest scoring sequences, thus optimizing the alignments of the query sequence with the best database matches.



(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.

Thus, dynamic programming is applied to the database search in a limited fashion, allowing FASTA to return its results very rapidly because it evaluates only a portion of the potential alignments.

BLAST vs FASTA

- FASTA begins the search by looking for exact matches of words, whereas BLAST allows for conservative substitutions in the first step.

- FASTA will return only one alignment for sequence in one list, BLAST can return multiple results.

- FASTA better for finding distantly related sequences than BLAST.

- For highly similar sequences FASTA and BLAST perform similarly.

Thank You