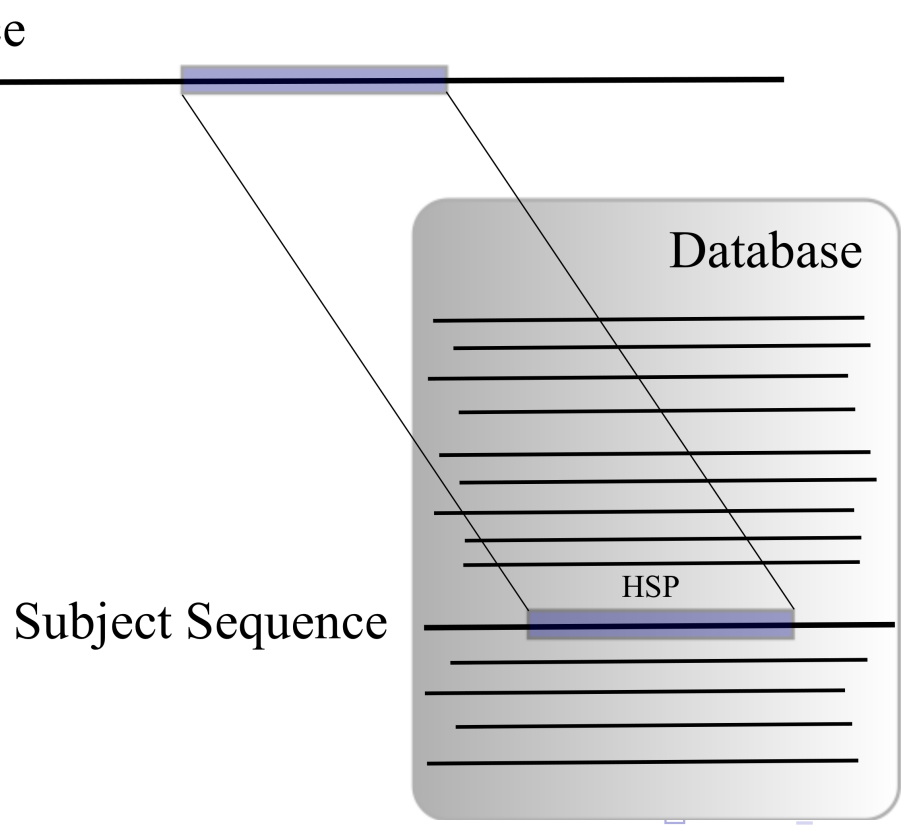
BLAST - Basic Local Alignment Search Tool

Sayyed Auwn Muhammad

Lecture for Algorithmic Bioinformatics (DD2450)

April 11, 2013

- Searching Algorithm
  1. Input: Query Sequence
  2. Database of sequences
  3. Subject Sequence(s)
  4. Output: High Scoring Segment Pairs (HSPs)

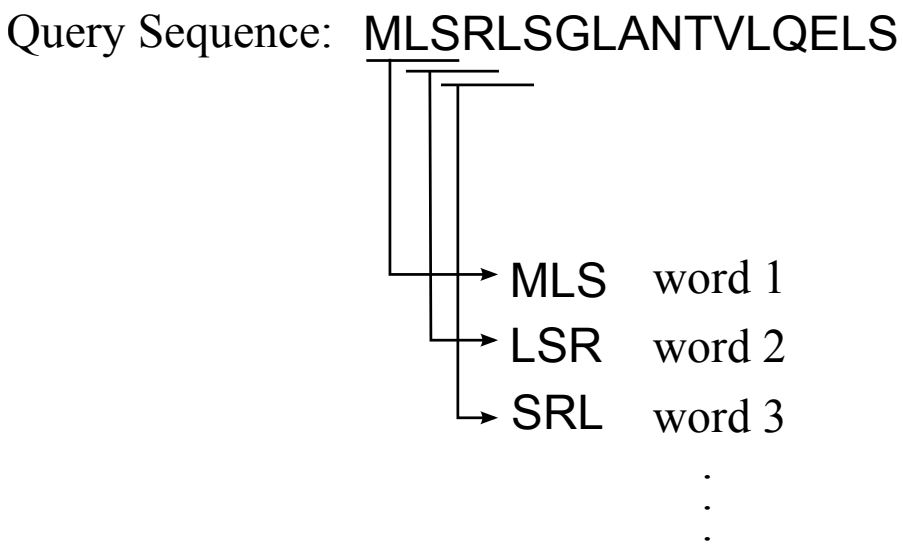


- Sequence Similarity Measures:
  1. **Global similarity algorithms** optimize the overall alignment of two sequences, which may include large stretches of low similarity.
  2. **Local similarity algorithms** seek only relatively conserved sub sequences, and a single comparison may yield several distinct subsequence alignments.
- Unconserved regions do not contribute to the measure of similarity.

" The main heuristic of BLAST is that there are often high-scoring segment pairs (HSPs) contained in a *statistically significant alignment*."

- High Scoring Segments Paris (HSPs)
  - Let a word pair be a segment pair of fixed length **W**.
  - The main strategy of BLAST is to search only those segment pairs that contain word with a score of at least threshold **T**.
  - Any such **hit** is extended to determine if it is contained within a segment pair whose score is greater than or equal to **S**.

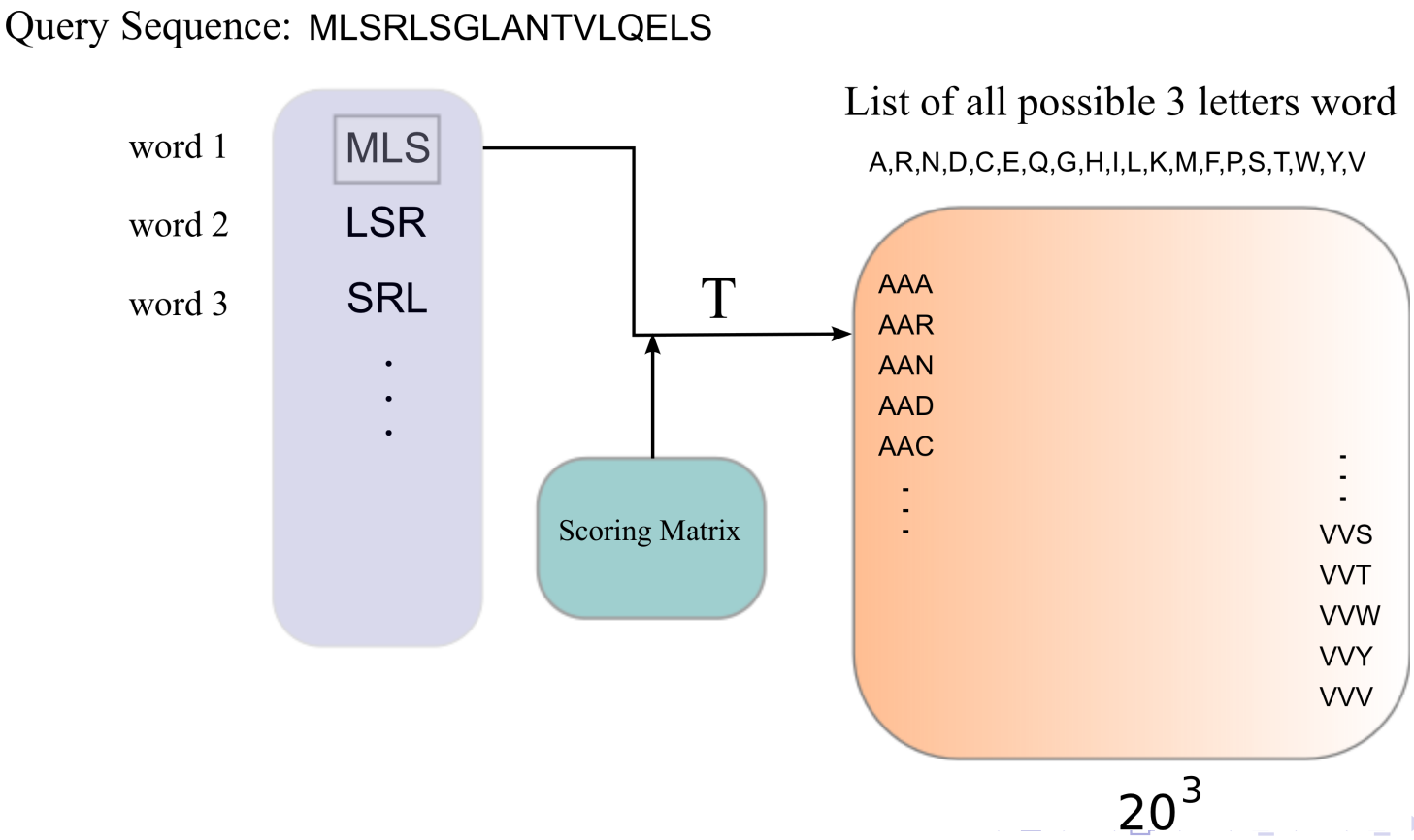
- **Step 1:**
  - Sequential Scanning of query sequence to construct the list of fixed length words.
  - For protein sequence $W = 3$ and for DNA sequence $W = 11$.



Derived from Wikipedia page on BLAST

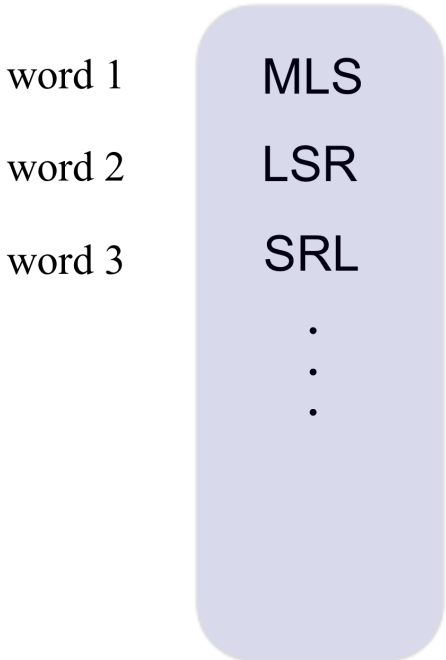
- **Step 0:**
  - Before going into Step 1...
  - Low-complexity Regions:
    - Def: Regions of protein sequences with biased amino acid composition are called Low-Complexity Regions (LCRs).
  - In Protein sequence: PPCDPPPPPKDKKKKDDGPP
  - In Nucleotide sequence: AAATAAAAAAAATAAAAAAT
  - We remove these Low Complexity Regions by **masking**.
    1. **Segmasker** masks low complexity regions of protein sequences.
    2. **Dustmasker** is for nucleotide sequences.

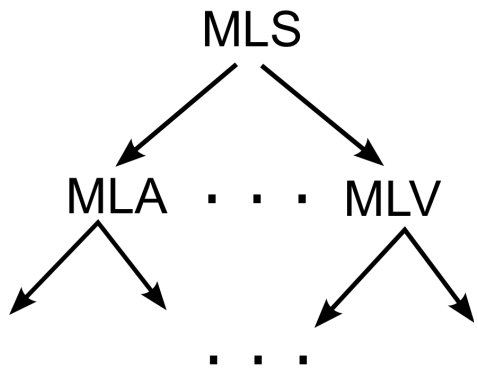
- **Step 2:**
  - Construct the lists of possible matching words by using the scoring matrix (substitution matrix) and filtering Threshold T.



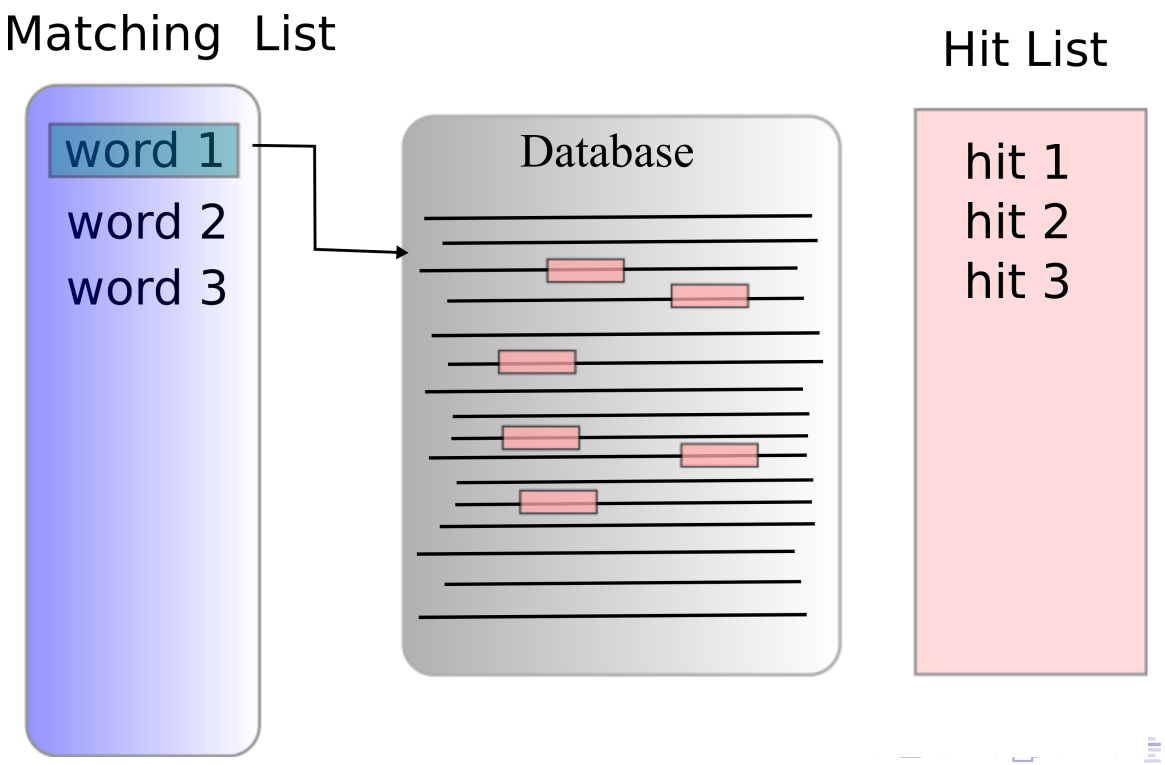
- **Step 3:**
  - Organize the lists of possible matching words in efficient tree format or finite automaton (finite state machine).

Query Sequence: MLSRLSGLANTVLQELS

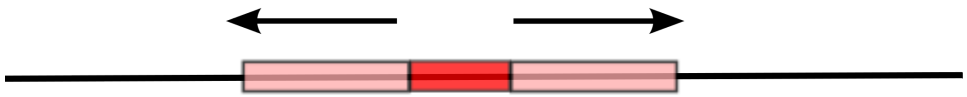




- **Step 4:**
  - Find the hits in the database by scanning the database sequences ...
  - This hit will be used to seed a possible alignment between the query and subject sequence.

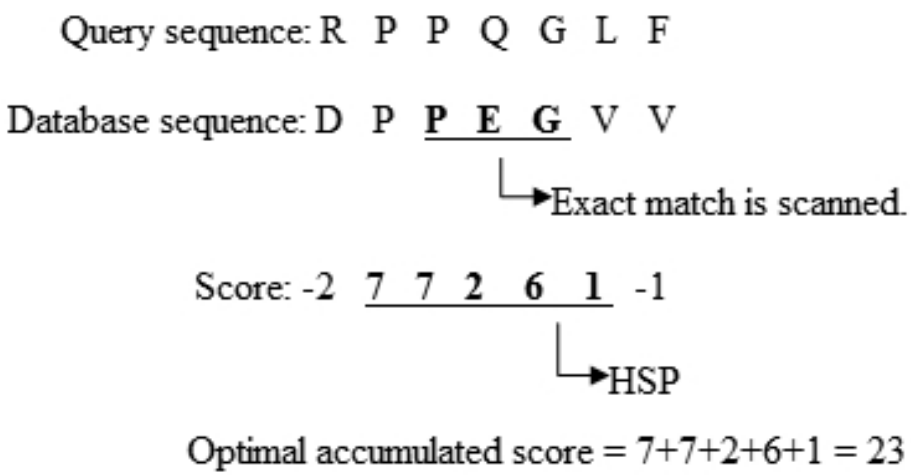


- **Step 5:**
  - Extend the Hits on both side of the subject sequence ...
  - The extension does not stop until the accumulated total score of the HSP begins to decrease with respect to threshold parameter S.



Extending hit

- **Step 6:**
  - Output the list the HSPs whose scores are greater than the empirically determined cutoff score S.



From Wikipedia page on BLAST

- How to compare sequence similarity ?
- Blast Scoring consists of three important components:
  1. Raw Score (S)
  2. Bit Score (S')
  3. E-Value (E)

- Raw Score:
- BLAST uses a substitution matrix, which specifies a score $s(i, j)$ for aligning each pair of amino acids i and j in an HSP.
- The aggregated score S in this way, is called raw Score i.e.

$$S = \sum_{i,j}^{L} s(i, j)$$

- This score is just a numerical value that will describe the overall quality of an alignment.

- Bit Score:
- Bit-score S' is a normalized score expressed in bits.
- This will estimate the size of the search space you would have to look through before you would expect to find a score as good as or better than this one by chance.
- According to definition by author:

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

- E value:
- The E-value (associated to a score S) is the number of distinct alignments, with a score equivalent to or better than S, that are expected to occur in a database search by chance.

Or

Expect value (E) parameter describes the number of hits one can "expect" to see by chance when searching a database of a particular size. Therefore it depends on the sizes i.e. $N = mn$, Where database size is m and Query size is n.

$$E = \frac{N}{2^{S'}}$$

- The statistical parameters λ and K are estimated by fitting the distribution of the un-gapped local alignment scores to the (Gumbel) extreme value distribution (EVD).
- The estimation of these parameters depends on the substitution matrix, gap penalties, and sequence composition (AA frequencies), and are the effective lengths of the query and database sequences, respectively.

- Applications:
  1. Homology Clustering
  2. Protein Domains
  3. Gapped BLAST improvements
  4. PSSM
  5. Neighbourhood Correlation