Local Alignment, BLAST and Scoring Matrices

Manu Madhavan

Lecture 6

Recap

- Sequence Alignment
- Global Alignment
- Needleman-Wunsch Algorithm

Outline

- Local Alignment
- Smith-Waterman algorithm
- BLAST - Basic Local Alignment Search Tool
- Scoring Matrices - PAM

Why not only Global Alignment

- Global alignment compares two sequences in their entirety; the gap penalty is assessed regardless of whether gaps are located internally within a sequence, or at the end of one or both sequences.
- Do you feel any issue with this?

Why not only Global Alignment

- Global alignment compares two sequences in their entirety; the gap penalty is assessed regardless of whether gaps are located internally within a sequence, or at the end of one or both sequences.
- Do you feel any issue with this?
- Suppose we wish to search for the short sequence `ACGT` within the longer sequence `AAACACGTGTCT`
- If two sequences have approximately the same length and are quite similar, they are suitable for Global alignment
- If sequences are divergent, global alignment not works well

Local Alignment

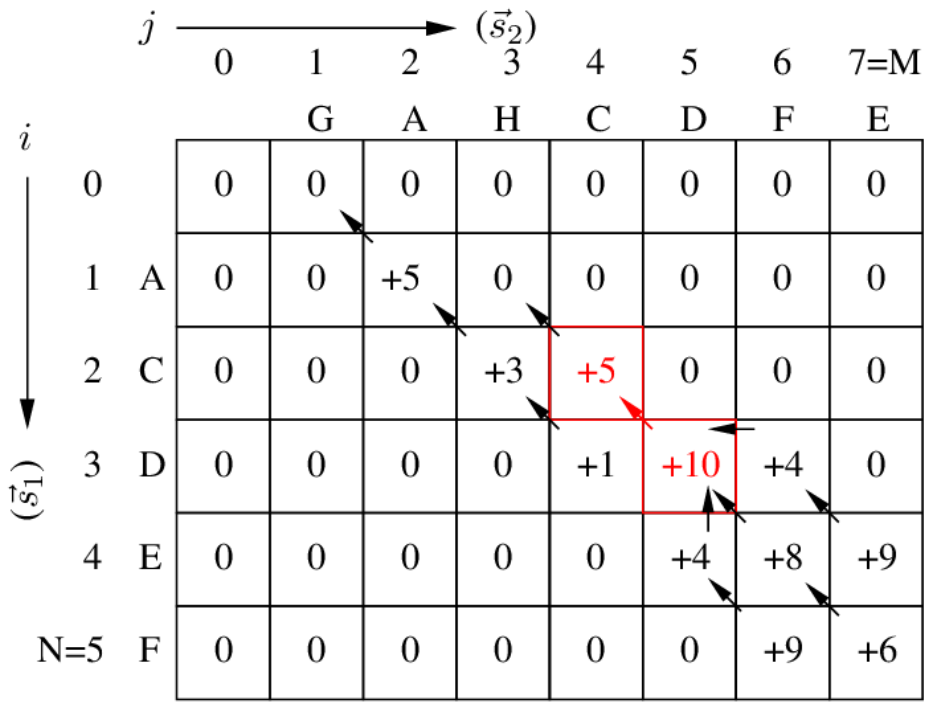
- Sequences which are suspected to have similarity or even dissimilar sequences can be compared with local alignment method.
- It finds the **local regions with high level of similarity**.
- Suitable for aligning more divergent sequences
- Can be used to find **conserved patterns** in DNA sequences

Smith-Waterman Algorithm

- Dynamic programming method for local alignment
- Objective is to find the the **optimal local alignment** with respect to the scoring system being used
- Extension of Needleman-Wunsch algorithm
- Changes:
  - Replace negative scoring matrix cells with zero
  - Traceback procedure starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment

Smith-Waterman Algorithm

- Extension of Needleman-Wunsch algorithm
- Changes:
  - Replace negative scoring matrix cells with zero
  - Traceback procedure starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment



Optimum alignment score: +10

Smith-Waterman Algorithm

Let $A = a_1 a_2 ... a_n$ and $B = b_1 b_2 ... b_m$ be the sequences to be aligned, where n and m A and B respectively.

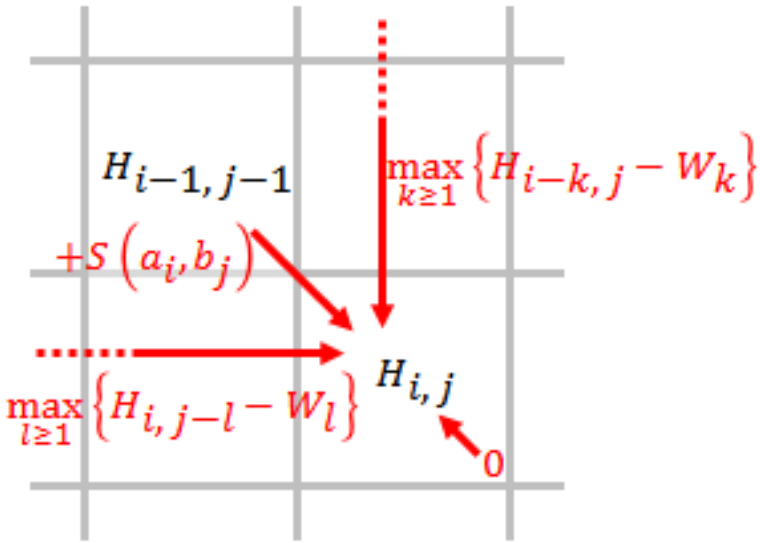
- Determine the substitution matrix and the gap penalty scheme
- Construct a scoring matrix H and initialize its first row and first column with $0s$.
- Fill the scoring matrix using the equation

$$H(i,j) = max \begin{cases} H(i-1, j-1) + s(a_i, b_j) \\ H(i, j-1) + g \\ H(i-1, j) + g \\ 0 \end{cases}$$

- Traceback. Starting at the highest score in the scoring matrix H and ending at a matrix cell that has a score of 0

Smith-Waterman Algorithm

$$H(i,j) = max \begin{cases} H(i-1,j-1) + s(a_i, b_j) \\ H(i,j-1) + g \\ H(i-1,j) + g \\ 0 \end{cases}$$



Try

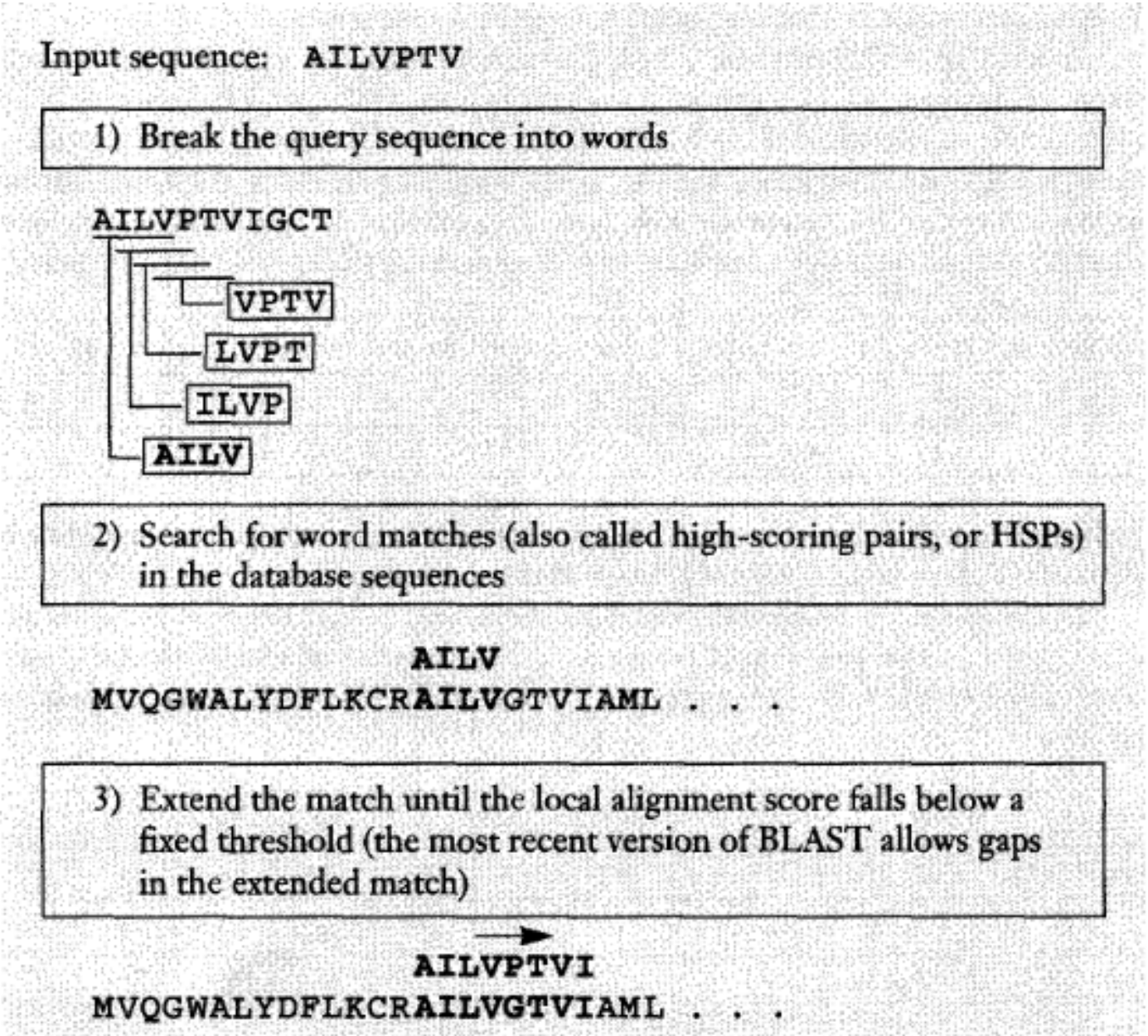
Use: match score = +1, mismatch=-1 and gap penalty as -1

		-	C	G	T	G	A	A	T	T	C	A	T
	-	0	0	0	0	0	0	0	0	0	0	0	0
G		0											
A		0											
C		0											
T		0											
T		0											
A		0											
C		0											

BLAST

- Basic Local Alignment Search Tool
- A searching method to retrieve similar sequences from databases (based on a query sequence)
- BLAST algorithm, introduced by S. Altschul et al.
- The original BLAST algorithm searches a sequence database for maximal ungapped local alignments
- Heuristic method

BLAST


Input sequence: AILVPTV
1) Break the query sequence into words
AILVPTVIGCT
VPTV
LVPT
ILVP
AILV
2) Search for word matches (also called high-scoring pairs, or HSPs) in the database sequences
AILV
MVQGWALYDFLKCRAILVGTVIAML . . .
3) Extend the match until the local alignment score falls below a fixed threshold (the most recent version of BLAST allows gaps in the extended match)
AILVPTVI
MVQGWALYDFLKCRAILVGTVIAML . . .

BLAST- e-value

- Alignment scores vary among the different database search algorithms, and are not a sufficient indicator that two sequences are related
- "Given a set of sequences not related to the query sequence (or even random sequences), what is the probability of finding a match with alignment score S simply by chance?"
- Given a database result with an alignment score S, the $E - score$ is the expected number of sequences of score $>= S$ that would be found by random chance.
- The $P - score$ is the probability that one or more sequences of score $>= S$ would have been found randomly
- Low value of E and P scores are desirable

BLAST- e-value

E-value significance

While E values of 10^{-3} and below are often considered indicative of statistically significant results, it is not uncommon for search algorithms to produce matches with E values on the order of 10^{50}, indicating a very strong likelihood of evolutionary relationship between the query sequence and the search results.

BLAST- variants

- **BLASTN** - for nucleotide sequence
- **BLASTP**- for protein sequence alignment
- **Translated Blast**: Translated BLAST searches use a genetic code to translate either the query, database subjects, or both, into protein sequences, which are then aligned as in blastp.
- **Genome Blast**: the application of any of the BLAST search programs to the complete genomic sequence of an organism or the transcript and protein sequences derived from its annotation.

BLAST- online

- Go through this online reference: `https://www.ncbi.nlm.nih.gov/books/NBK1734/`
- BLAST tool is available at `https://blast.ncbi.nlm.nih.gov/Blast.cgi`
- Try the exercises mentioned in the references
- Biopython-BLAST `https://www.tutorialspoint.com/biopython/biopython_overview_of_blast.htm`

Scoring Matrices

- A simple scheme:
  - A positive value or high score is given for a match
  - a negative value or low score for a mismatch and gaps.
  - This assignment is based on the assumption that the frequencies of mutation are equal for all bases.
- **Transitions**: substitutions between purines[1] and purines or between pyrimidines[2] and pyrimidines
- **Transversions**: substitutions between purines and pyrimidines
- Transitions occurs more frequently than Transversions

[1] A and G
[2] C and T

Scoring Matrices

- An amino-acid scoring matrix is a 20×20 table such that position indexed with amino-acids so that position X,Y in the table gives the score of aligning amino-acid X with amino-acid Y
- Identity matrix Exact matches receive one score and non-exact matches a different score (1 on the diagonal 0 everywhere else)
- Mutation data matrix a scoring matrix compiled based on observation of protein mutation rates: some mutations are observed more often then other (PAM, BLOSUM).
- Physical properties matrix amino acids with with similar biophysical properties receive high score.
- Genetic code matrix amino acids are scored based on similarities in the coding triple.

Next

- Scoring Matrices- Details
- MSA algorithms and Tools