Lecture 11-12

Pairwise Sequence Alignment

```
Global FTFTALILLAVAV
       F--TAL-LLA-AV

Local  FTFTALILL-AVAV
       --FTAL-LLAAV--
```

Local Pairwise Sequence Alignment

- Smith-Waterman

Seq1: ACACACTA

Seq2: AGCACACA

S(match) = +2

S(mismatch) = *w(-,b)* = *w(a,-)* = -1

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

$$T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow \end{pmatrix}$$

Local pairwise alignment

- Align

S = TAATATATTTAT

T = AAGCGAATAATATATTTATACTCAGATTATTGCGCG

Local pairwise alignment

- Initial Seed

```
          TAT
          |||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

- Alignment by expansion of seed

```
       TAATATATTTAT
       ||||||||||||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

An example

- **Input:**

  S_q= AKLMAATCD

  S_i= ….ALPQRKLMMAKLPPRTLQ…..

  Window size $w = 4$

  Threshold $T = 3$ for determining seed points

- **Method:**

  1. Generating subsequences of length 4 ($w = 4$) from target string S_q.

- AKLMAATCD
- AKLM
- KLMA
- LMAA
- MAAT
- AATC
- ATCD

Considering threshold $T=3$, sequences that are considered valid for subsequence KLMA

Sequence	Score
KLMA	4
K*MA	3
KL*A	3
KLM*	3
* indicates a wild card character and can be any alphabet	

- Identifying valid entries in string S_i corresponding to seed KLMA

1100 = 2 X
KLMA
….ALPQ**RKLMMAKLP**PRTLQ…..
KLMA
1110 = 3 ✓

Perform all such string matching's using DP for such all such seed point contained extended regions and report the segment with the highest score in string S_i.

What is BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool

- Calculates similarity for biological sequences.

- Produces local alignments: only a portion of each sequence must be aligned.

- Uses statistical theory to determine if a match might have occurred by chance.

BLAST is a heuristic

• A lookup table is made of all the “words” (short subsequences) in the query sequence. In many types of searches “neighboring” words are included.

• The database is scanned for matching words (“hot spots”).

• Gapped and un-gapped extensions are initiated from these matches.

BLAST method

- It is an alignment heuristic that determines “local alignments” between a query and a database. It uses an approximation of the Smith-Waterman algorithm.

- BLAST consists of two components: a search algorithm and computation of the statistical significance of solutions.

- BLAST uses a heuristic method to find the highest scoring alignment between the query sequence and the search set sequence.

BLAST terminology

- Definition

Let *q* be the query and *d* the database. A segment is simply a substring *s* of *q* or *d*.

A segment-pair (*s*, *t*) (or hit) consists of two segments, one in *q* and one *d*, of the same length.

BLAST terminology

- Example

```
V A L L A R
P A M M A R
```

- We think of *s* and *t* as being aligned without gaps and score this alignment using a substitution score matrix, e.g. BLOSUM or PAM in the case of protein sequences.

- The alignment score for (*s*, *t*) is denoted by $\sigma(s, t)$.

BLAST terminology

- A locally maximal segment pair (LMSP) is any segment pair (*s*, *t*) whose score cannot be improved by shortening or extending the segment pair.

- A maximum segment pair (MSP) is any segment pair (*s*, *t*) of maximal alignment score $\sigma(s, t)$.

- Given a cutoff score *S*, a segment pair (*s*, *t*) is called a high-scoring segment pair (HSP), if it is locally maximal and $\sigma(s,t) \geq S$.

- Finally, a word is simply a short substring of fixed length *w*.

The BLAST algorithm

- **Goal:** Find all HSPs for a given cut-off score.

- Given three parameters, i.e. a word size *w*, a word similarity threshold *T* and a minimum cut-off score *S*. Then we are looking for a segment pair with a score of at least *S* that contains at least one word pair of length *w* with score at least *T*.

The BLAST algorithm

- **Preprocessing:** Of the query sequence *q* first all words of length *w* are generated. Then a list of all *w*-mers of length *w* over the alphabet $\sum$ that have similarity > *T* to some word in the query sequence q is generated.

Example

For the query sequence `RQCSAGW` the list of words of length $w = 2$ with a score $T > 8$ using the BLOSUM62 matrix are:

word	$2-$ *mer* with score > 8
RQ	RQ
QC	QC, RC, EC, NC, DC, KC, MC, SC
CS	CS,CA,CN,CD,CQ,CE,CG,CK,CT
SA	-
AG	AG
GW	GW,AW,RW,NW,DW,QW,EW,HW,KW,PW,SW,TW,WW

The BLAST algorithm

1. **Localization of the hits:** The database sequence d is scanned for all hits t of w-mers s in the list, and the position of the hit is saved.
2. **Detection of hits:** First all pairs of hits are searched that have a distance of at most A (think of them lying on the same diagonal in the matrix of the SW-algorithm).
3. **Extension to HSPs:** Each such *seed* (s, t) is extended in both directions until its score $\sigma(s, t)$ cannot be enlarged (LMSP). Then all best extensions are reported that have score $\geq S$, these are the HSPs. Originally the extension did not include gaps, the modern BLAST2 algorithm allows insertion of gaps.

The BLAST algorithm

- The list L of all words of length w that have similarity $> T$ to some word in the query sequence q can be produced in $O(|L|)$ time.
- These are placed in a "keyword tree" and then, for each word in the tree, all exact locations of the word in the database d are detected in time linear to the length of d.
- As an alternative to storing the words in a tree, a finite-state machine can be used, which Altschul et al. found to have the faster implementation.

The BLAST algorithm

Use of seeds of length *w* and the termination of extensions with fading scores (**score dropoff threshold X**) are both steps that speed up the algorithm.
Recent improvements (BLAST 2.0):

- Two word hits must be found within a window of *A* residues.
- Explicit treatment of gaps.
- Position-specific iterative BLAST (PSI-BLAST).

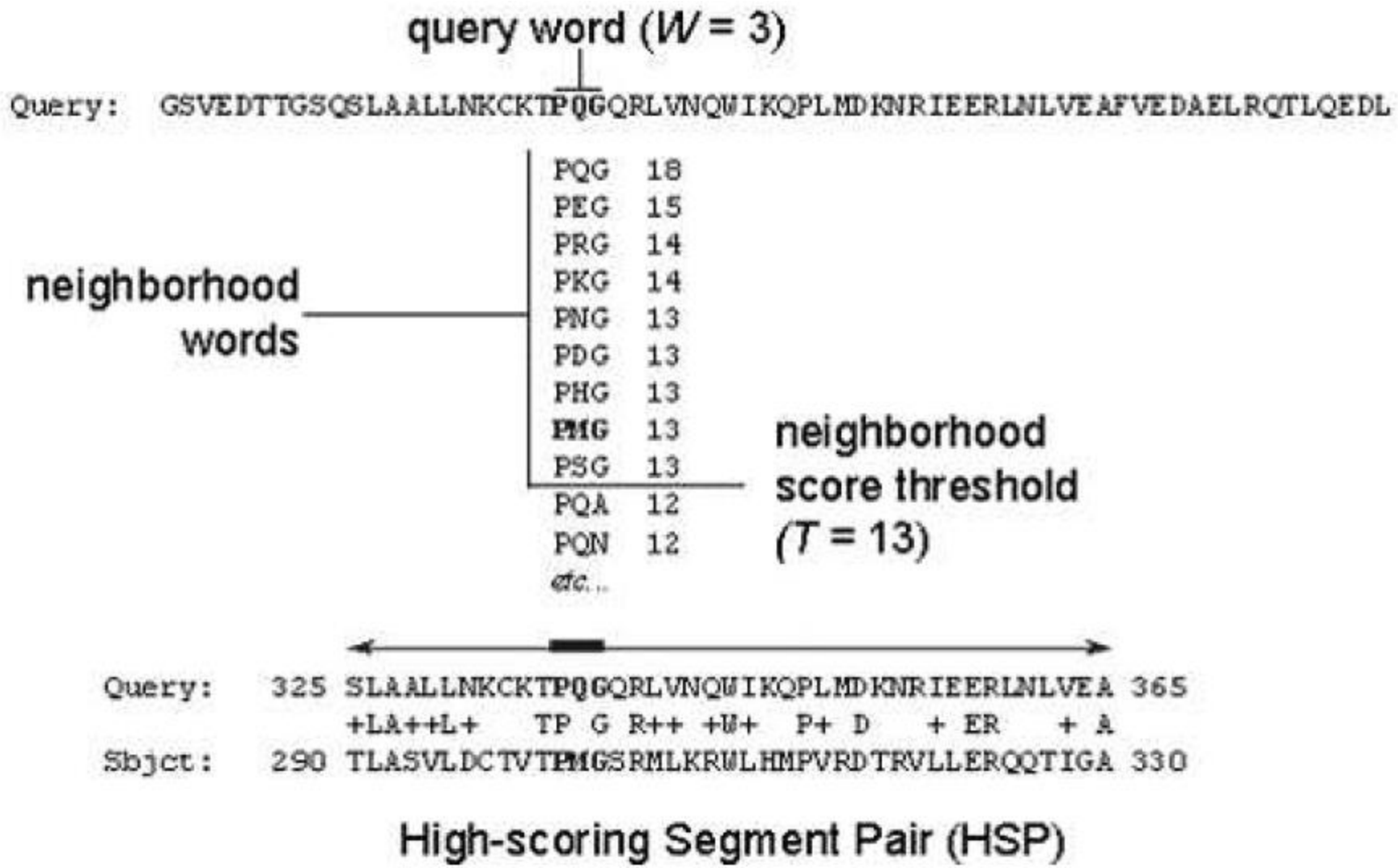
The BLAST algorithm for DNA

For *DNA* sequences, BLAST operates as follows:

- The list of all words of length w in the query sequence q is generated. In practice, $w = 12$ for DNA.
- The database d is scanned for all hits of words in this list. Blast uses a two-bit encoding for DNA. This saves space and also search time, as four bases are encoded per byte.

Note that the "T" parameter dictates the speed and sensitivity of the search.

The BLAST search algorithm



Statistical Significance of HSP

- **Problem:** ***Given an HSP (s,t) with score $\sigma(s,t)$. How significant is this match (i.e., local alignment)?***

  Given the scoring matrix $S(a, b)$, the expected score for aligning a random pair of amino acid is required to be negative:

$$E = \sum_{a,b \in \Sigma} p_a p_b S(a, b) < 0$$

  The sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution. The maximum of a large number of i.i.d. random variables tends to an extreme value distribution as we will see

Statistical Significance of HSP

HSP scores are characterized by two parameters, K and λ. The parameters K and λ depend on the background probabilities of the symbols and on the employed scoring matrix. λ is the unique value for y that satisfies the equation

$$\sum_{a,b \in \Sigma} p_a p_b e^{S(a,b)y} = 1$$

K and λ are scaling-factors for the search space and for the scoring scheme, respectively.

The number of random HSPs (s, t) with $\sigma(s, t) \geq S$ can be described by a Poisson distribution with parameter $v = Kmne^{-\lambda S}$. The number of HSPs with score $\geq S$ that we *expect* to see due to chance is then the parameter v, also called the *E-value*:

$$E(\texttt{HSPs with score} \geq S) = Kmne^{-\lambda S}$$

BLAST Statistics

- **Score:**
  - A statistical conversion of the score derived by summing using the substitution matrix.

- **Expect (e) value:**
  - Function of the S value and the database size
  - *An e value of 1:* One alignment using a query of this size will by chance produce a S score of this value in a database of this size
  - *e value of –10 (=1x10^{-10}):* Unlikely that random chance lead to this current alignment compared to an alignment with an e value of 1
  - Expect value is specific to a database of a certain size . Thus it may change later because of change in database size.

- **Rules of thumb:**
  - **E value of –30 or less:** Sequences are homologous
  - **E values of –5:** Often considered significant enough when annotating a genome

BLAST output

- Pair-wise report
- Query-anchored report
- Hit-table
- Tax BLAST
- Abstract Syntax Notation 1
- XML

BLAST family of programs

The BLAST family of programs allows all combinations of DNA or protein query sequences with searches against DNA or protein databases:

- Protein-protein (blastp): compares an amino acid sequence against a protein sequence database.
- Nucl.-nucl (blastn): compares a nucleotide query sequence against a nucleotide sequence database (in general optimized for speed, not sensitivity).
- Translated nucl.-protein (blastx): compares the six-frame conceptual translation products of a nucleotide query against a protein sequence database.
- Protein-translated nucl (tblastn): compares a protein query sequence against a sequence database dynamically translated in all six reading frames (useful for searching proteins against EST's).
- Translated nucl-translated nucl. (tblastx): compares the six frame translation of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

BLAST webserver

- BLAST
  - http://blast.ncbi.nlm.nih.gov/Blast.cgi
  - http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Work Out

- **Problem Statement**

  Given a query sequence S_q and a target sequence S_i (such that $|S_q| \ll |S_i|$) find an optimal alignment and alignment score of S_q with S_i.

  – *Additional information/input:* For a window size (*w*) 4 the minimum score (*T*) value is 3. First locate such window and expand on both the sides.