Dept. of Computer Science and Engineering
University of Rajshahi
www.ru.ac.bd

Dr. Shamim Ahmad

BLAST:
Basic local alignment search tool

B L A S T !

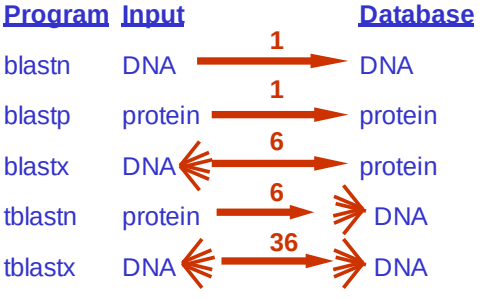
Courtesy of Jonathan Pevsner
Johns Hopkins U.

BLAST

BLAST (Basic Local Alignment Search Tool) allows rapid sequence comparison of a query sequence against a database.

The BLAST algorithm is fast, accurate, and web-accessible.

Choose the BLAST program



BLAST: background on sequence alignment

There are two main approaches to sequence alignment:

[1] Global alignment (Needleman & Wunsch 1970) using dynamic programming to find optimal alignments between two sequences. (Although the alignments are optimal, the search is not exhaustive.) Gaps are permitted in the alignments, and the total lengths of both sequences are aligned (hence "global").

BLAST: background on sequence alignment

[2] The second approach is local sequence alignment (Smith & Waterman, 1980). The alignment may contain just a portion of either sequence, and is appropriate for finding matched domains between sequences. S-W is guaranteed to find optimal alignments, but it is computationally expensive (requires $(O)n^2$ time).

BLAST and FASTA are heuristic approximations to local alignment. Each requires only $(O)n^2/k$ time; they examine only part of the search space.

How a BLAST search works

"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)

How the original BLAST algorithm works: three phases

Phase 1: compile a list of word pairs (w=3) above threshold T

Example: for a human RBP query
…FSGTWYA… (query word is in yellow)

A list of words (w=3) is:
FSG SGT GTW TWY WYA

Phase 1: compile a list of words (w=3)

neighborhood
word hits
> threshold

(T=11)

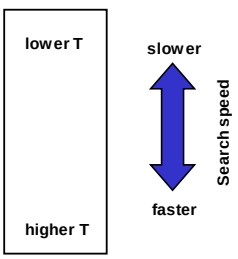
GTW	6,5,11	22
ASW	6,1,11	18
ATW	0,5,11	16
NTW	0,5,11	16
GTY	6,5,2	13
GNW		10
GAW		9

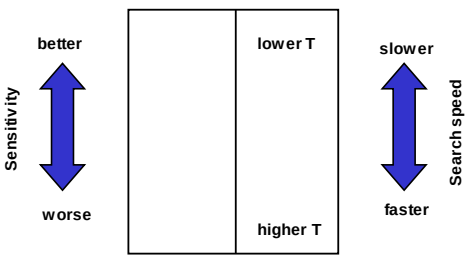
neighborhood
word hits
< below threshold

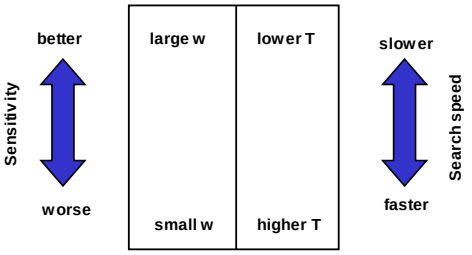
Fig. 4.13
page 101

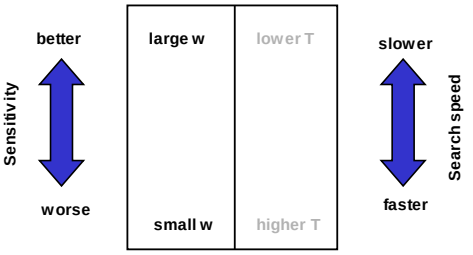
Pairwise alignment scores are determined using a scoring matrix such as Blosum62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4



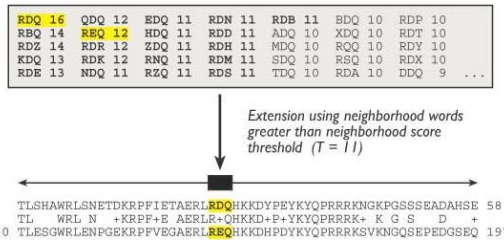


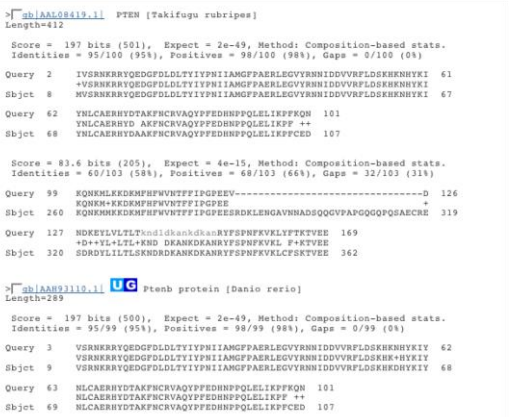




For proteins, default word size is 3.
(This yields a more accurate result than 2.)

BLAST Algorithm





STEP 1

Remove low-complexity region or sequence repeats in the query sequence

- "Low-complexity region" means a region of a sequence composed of few kinds of elements
- The regions will be marked with an X (protein sequences) or N (nucleic acid sequences) and then be ignored by the BLAST program

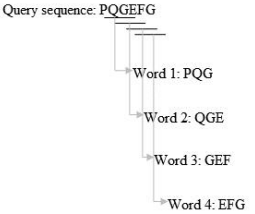
Low-complexity sequence can often be recognized by visual inspection.
For example,
Protein sequence
PPCDPPPPPKDKKKKDDGPP
Nucleotide sequence
AAATAAAAAAAATAAAAAAT.

To filter out the low-complexity regions,
SEG program is used for protein sequences
DUST program is used for DNA sequences

STEP 2

Make a k-letter word list of the query sequence.

Take k=3 for example, we list the words of length 3 in the query protein sequence
(k is usually 11 for a DNA sequence)



STEP 3
List the possible matching words.

- BLAST only cares about the high-scoring words.
- Example: the score obtained by comparing PQG with PEG and PQA is respectively 15 and 12
- For DNA words, a match is scored as +5 and a mismatch as -4, or as +2 and -3

STEP 4
Organize the remaining high-scoring words into an efficient search tree.

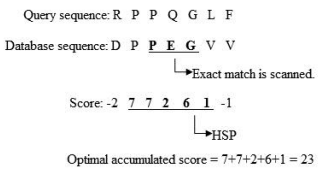
- This allows the program to rapidly compare the high-scoring words to the database sequences.

STEP 5
Repeat step 3 to 4 for each k-letter word in the query sequence

STEP 6
Scan the database sequences for exact matches with the remaining high-scoring words.
The BLAST program scans the database sequences for the remaining high-scoring word, such as PEG, of each position

STEP 7
Extend the exact matches to high-scoring segment pair (HSP).

- The original version of BLAST stretches a longer alignment between the query and the database sequence in the left and right directions

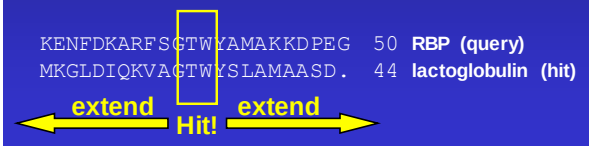


How a BLAST search works: 3 phases

Phase 3: when you manage to find a hit (i.e. a match between a "word" and a database entry), extend the hit in either direction.

Keep track of the score (use a scoring matrix)

Stop when the score drops below some cutoff.



STEP 8

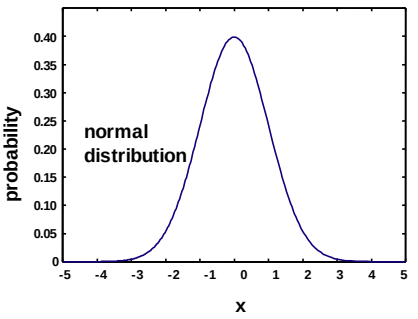
List all of the HSPs in the database whose score is high enough to be considered.

- List the HSPs whose scores are greater than the empirically determined **cutoff score S.**
- By examining the distribution of the alignment scores modeled by comparing random sequences, a cutoff score S can be determined such that its value is large enough to guarantee the significance of the remaining HSPs

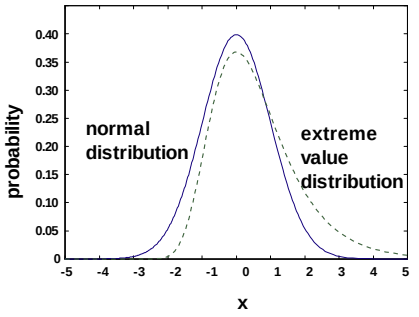
STEP 9

Evaluate the significance of the HSP score.

- BLAST next assesses the statistical significance of each HSP score by exploiting the Gumbel extreme value distribution (EVD).
- It is proved that the distribution of Smith-Waterman local alignment scores between two random sequences follows the Gumbel EVD



The probability density function of the extreme value distribution (characteristic value u=0 and decay constant λ=1)



particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

The lower the E-value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query

STEP 10

Make two or more HSP regions into a longer alignment.

Sometimes, two or more HSP regions in one database sequence that can be made into a longer alignment.