



Basic Local Alignment Search Tool



A blast from the past...

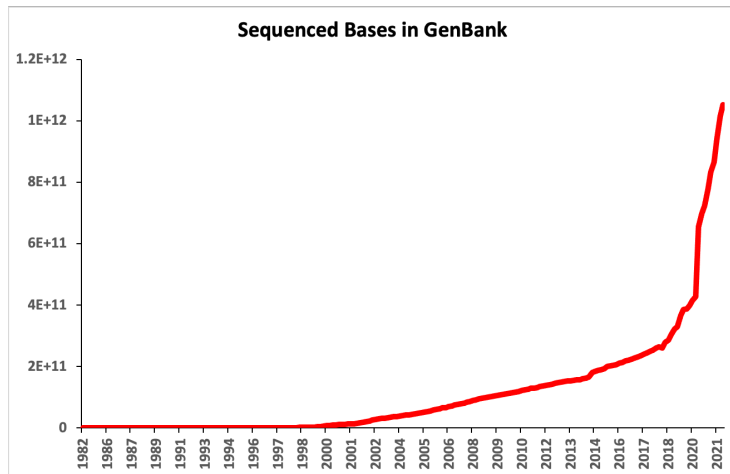
AGATCAC
CGACAG

	C	G	A	C	A	G
A	0	0	0	0	0	0
G	0	0	5	0	1	0
A	0	0	0	10	4	6
T	0	0	0	4	6	0
C	0	5	0	0	9	3
A	0	0	0	5	3	14
C	0	5	0	0	10	8

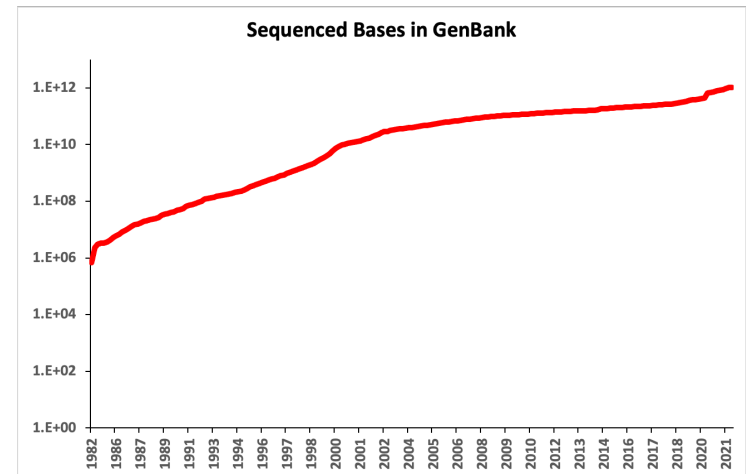
GATCA
|| ||
GA-CA



Why BLAST?



Why BLAST?

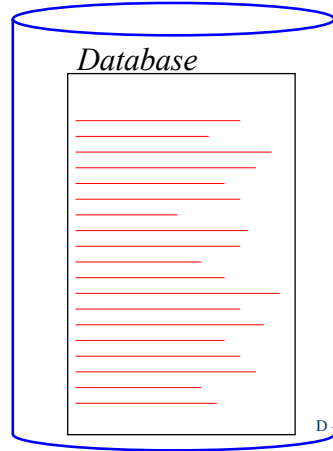




How Does BLAST Work?

Query sequence

MLVFAHAYHESKWAAHNQEILTPLV



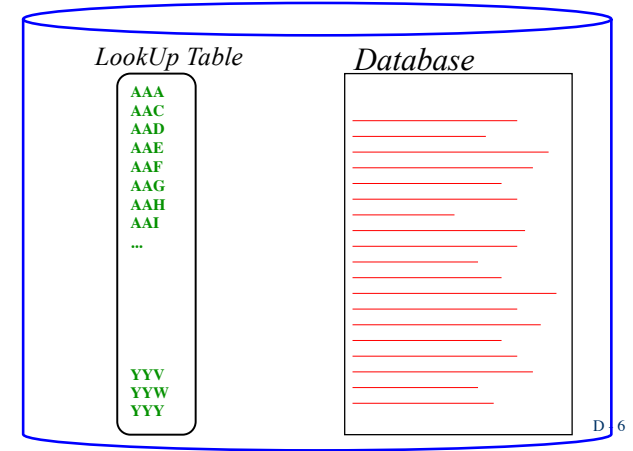
BLAST Example

Query sequence

MLVFAHAYHESKWAAHNQEILTPLV

Word List

MLV	AHN
LVF	HNQ
VFA	NQE
FAH	QEI
AHA	EIL
HAY	ILT
AYH	LTP
YHE	TPL
HES	PLV
ESK	
SKW	
KWA	
WAA	
AAH	



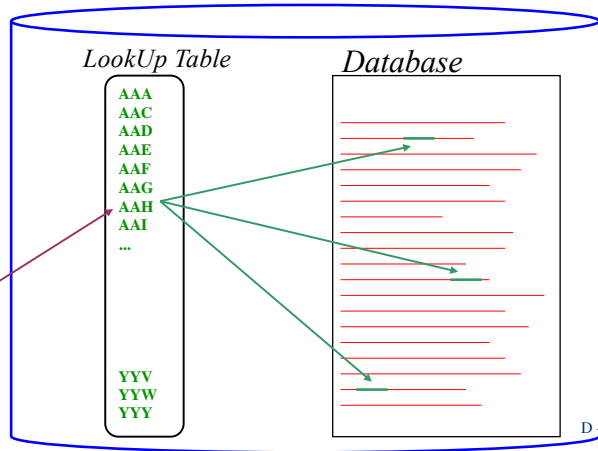
BLAST Example

Query sequence

MLVFAHAYHESKWAAHNQEILTPLV

Word List

MLV	AHN
LVF	HNQ
VFA	NQE
FAH	QEI
AHA	EIL
HAY	ILT
AYH	LTP
YHE	TPL
HES	PLV
ESK	
SKW	
KWA	
WAA	
AAH	



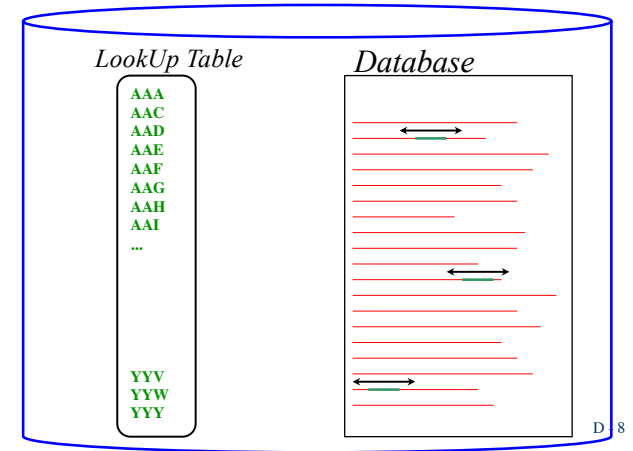
BLAST Example

Query sequence

MLVFAHAYHESKWAAHNQEILTPLV

Word List

MLV	AHN
LVF	HNQ
VFA	NQE
FAH	QEI
AHA	EIL
HAY	ILT
AYH	LTP
YHE	TPL
HES	PLV
ESK	
SKW	
KWA	
WAA	
AAH	



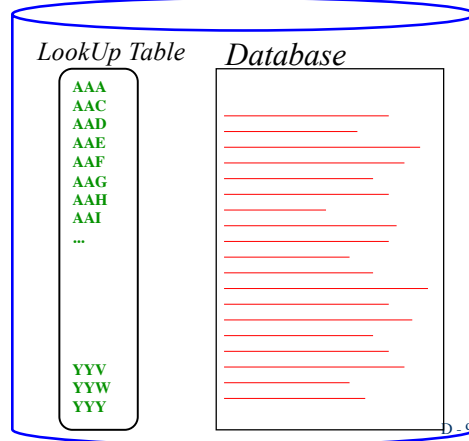


BLAST In a Nutshell

Query sequence

MLVFAHAYHESKWAAHNQEILTPLV

- Create "word list" from query sequence
- Locate words in database via "lookup table"
- Determine similarity of query sequence to each word-match sequence in database



BLAST Program

U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastp suite

blastn **blastp** blastx tblastn tblastx

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange From To

Or, upload file No file selected.

Job Title

Align two or more sequences

Choose Search Set
Database
Organism exclude
Exclude Models (XM/XP) Non-redundant RefSeq proteins (WPI) Uncultured/environmental sample sequences

Program Selection
Algorithm
 Quick BLASTP (Accelerated protein-protein BLAST)
 blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window



BLAST Output

universal stress protein [Pyrococcus horikoshii]

Sequence ID: [WP_048053230.1](#) Length: 167 Number of Matches: 3

Range 1: 1 to 152 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
96.7 bits(239)	2e-22	Compositional matrix adjust.	62/157(39%)	97/157(61%)	8/157(5%)

Query	4	MYKKILYPTDFSETAEIALKHVKAFKTLKAEVILLHVIDEREIKKRDI FSL LGVAGLNK	63
Sbjct	1	M++K+L+PTDFSE A A++ + ++ EVILLHVIDE +++ + G + MFRKVLFP TDFSE GAYRAVEVF EKRNKMEVGEVILLHVIDEGTLEE----LMDGYSFFYD	56
Query	64	SVEEFENELKNKLT EAKNKMENIKKELEDVGFVKD---IIVVGI PHEEIVKIAEDGV + E ++K KL EEA K++ +E++ F+ K+ II GIP +EIVK+AE+E V	120
Sbjct	57	NAEIELKDIKEKLKEEASRKLQEKAAEEVKR-AFRAKNVRTIIRFGIPWDEIVKVAEEENV	115
Query	121	DIIMGSHGKTNLKEILLGVS TENVIKSNKPVLVK 157 +II+ S GK +L LGS V++K+ KPVL++K	
Sbjct	116	SLIILPSRGKLSLSHEFLGSTVMRVL RRTKKPVLIIK 152	



BLAST Output

Search Parameters	
Program	blastn
Word size	28
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	1,-2
Gapcosts	3,1
Low Complexity Filter	Yes
Filter string	L,m;
Genetic Code	1

Database	
Posted date	Feb 14, 2022 11:51 AM
Number of letters	673,799,021,858
Number of sequences	79,695,929
Entrez query	None

Karlin-Altschul statistics		
Lambda	1.33271	1.32
K	0.620991	0.57
H	1.12409	1

Results Statistics	
Length adjustment	32
Effective length of query	1429
Effective length of database	671248752130
Effective search space	959214466793770
Effective search space used	959214466793770

Search Parameters	
Program	blastp
Word size	6
Expect value	0.05
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	Feb 12, 2022 2:31 AM
Number of letters	175,661,398,663
Number of sequences	460,231,190
Entrez query	None

Karlin-Altschul statistics		
Lambda	0.316534	0.267
K	0.135598	0.041
H	0.365158	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Results Statistics	
--------------------	--



BLAST Options

Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display ?

Short queries: Automatically adjust parameters for short input sequences ?

Expect threshold: 0.05 ?

Word size: 6 ?

Max matches in a query range: 0 ?

Scoring Parameters

Matrix: BLOSUM62 ?

Gap Costs: Existence: 11 Extension: 1 ?

Compositional adjustments: Conditional compositional score matrix adjustment ?

Filters and Masking

Filter: Low complexity regions ?

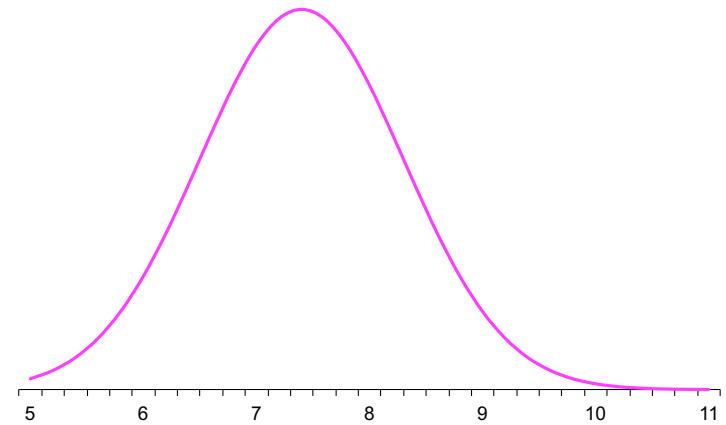
Mask: Mask for lookup table only ?
 Mask lower case letters ?

BLAST Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

D - 13



Normal Distributions

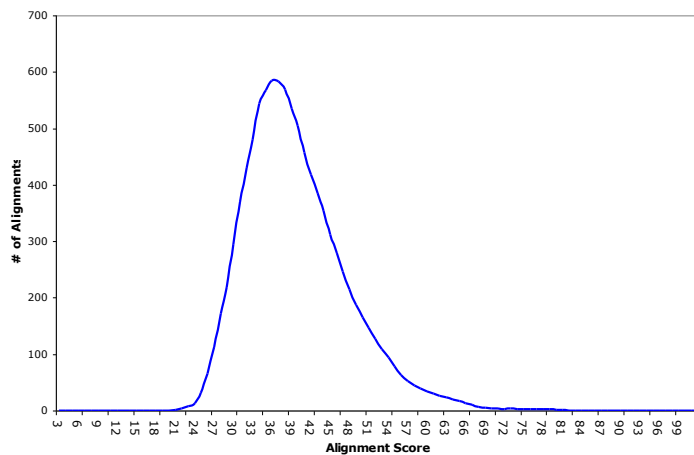


The widths of zebra stripes are normally distributed, with a mean of 7.3 centimeters and a standard deviation of 0.9 centimeters.

D - 14



Extreme Value Distributions



Scores of optimal local alignments correspond to extreme value distributions.

D - 15



Statistical Significance

Suppose we align two sequences, a query sequence and a target sequence, and we determine that their optimal local alignment score is $S = 60$.

Are the sequences similar? In other words, is a score of $S = 60$ significant? How likely is it that we would observe an alignment score of $S = 60$ by chance?

The *p-value* of an optimal local alignment score, S , is the likelihood that two random sequences* would have an optimal local alignment score greater than or equal to S .

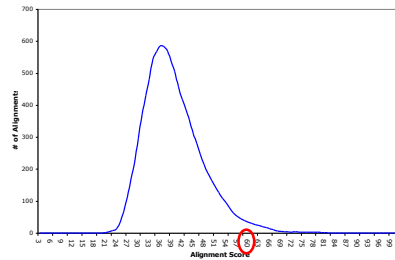
* of the same lengths and compositions as the query and target sequences

D - 16



p-values for pairs of sequences

What is the probability that the optimal local alignment score for two random sequences will be at least 60?

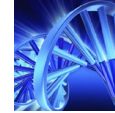


Solution 1: Count up all of the alignment scores greater than or equal to 60 and divide by the total number of alignment scores, i.e., 10,000.

Solution 2: Plug $x = 60$ into the the following expression, where $\mu = 34.2$ and $\beta = 6.1$

$$1.0 - e^{-e^{-\frac{x-\mu}{\beta}}}$$

D - 17



p-values for databases

When searching a large database with many target sequences, our previous definition of the p -value is problematic because we can expect some small p -values by chance. For example, if we align a query sequence to 6,000,000 target sequences in a database, we can expect 60,000 scores with a p -value less than 0.01.

When we BLAST a query sequence against a database of many target sequences, the p -value of one of the alignment scores, S , indicates the likelihood that we would see a score of at least S when BLASTing the query sequence against a comparable random database.

D - 18



E-values

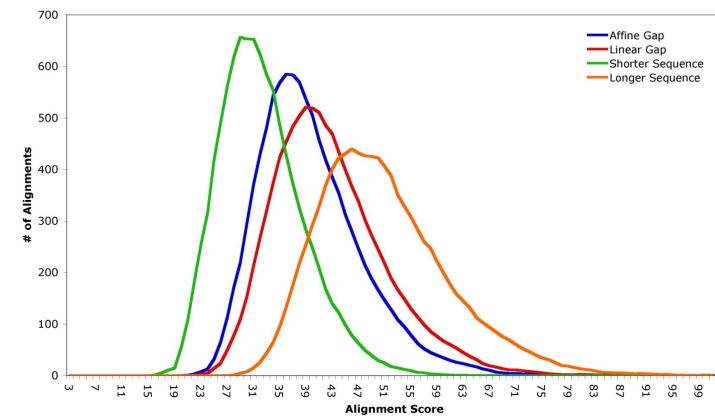
Instead of p -values, BLAST reports E-values. If the alignment score of a query sequence and some target sequence in the database is S , the E -value is the expected number of alignments with score S or higher in a random database.

Score	Expect	Method	Identities	Positives	Gaps
40.0 bits(92)	1.4	Compositional matrix adjust.	27/88(31%)	42/88(47%)	3/88(3%)
Query 9	MSKGAPWAKGRRGIIVLSRRLAGETSVAQSTPSYSD---KNLTQPPLNNINAYDSIILS	65			
	MSK P G RGI + + +A + Q+ + S +LT P+ + + ++S				
Sbjct 127	MSKAEPVESGERGIIINTASIAAFEGQIGQAAYAASKGAVHSLTLPAARELARHGIRVMS	186			
Query 66	MDQSNIGTPLTSGGSDDDVQEQAAGVKF	93			
	+ GTP+ SG DDVQ+ AA F				
Sbjct 187	IAPGLFGTPMLSGLPDDVQDSLAAANTPF	214			

D - 19



E-values depend on sequences and scoring



D - 20



Runtime of BLAST?

If n is the length of the query sequence and m is the length of the target database...

What is the runtime of computing a pairwise alignment?

What is the runtime of BLAST?

How can we improve this runtime?



Linear Alignment

