

Global and local alignments

Needleman-Wunsch algorithm, mathematical form

$$\mathbf{M}(0, j) = j \times p \quad \text{first row, } p = \text{gap penalty}$$

$$\mathbf{M}(i, 0) = i \times p \quad \text{first column}$$

$$\mathbf{M}(i, j) = \max \left(\begin{array}{c} \mathbf{M}(i-1, j) + p \\ \mathbf{M}(i, j-1) + p \\ \mathbf{M}(i-1, j-1) + s(a_j, b_i) \end{array} \right) \begin{array}{l} \text{top} \\ \text{left} \\ \text{diagonal} \end{array}$$

$s(a_j, b_i)$ = match/mismatch score for sites j and i
in sequences a and b

Needleman-Wunsch in Python

```
# Fill in the first row
for j in range(0, n + 1):
    score[0][j] = gap_penalty * j

# Fill in the first column
for i in range(0, m + 1):
    score[i][0] = gap_penalty * i

# Fill in all other values in the score matrix
for i in range(1, m + 1):      # loop over all rows
    for j in range(1, n + 1):  # loop over all columns
        # Calculate the score by checking the top, left, and diagonal cells
        insert = score[i - 1][j] + gap_penalty      # top
        delete = score[i][j - 1] + gap_penalty      # left
        match = score[i - 1][j - 1] + \            # diagonal
            match_score(seq1[j-1], seq2[i-1])

        # Record the maximum score from the three possible ones
        score[i][j] = max(match, delete, insert)
```

Global vs. local alignments

- Global: align all nucleotides
- Local: align subsequences with best score

Align these sequences: GCAT, GCT
(match = 1, mismatch = -1, gap = -1)

global alignment:

GCAT

GC-T

local alignment:

?

We can make local alignments using the Smith-Waterman algorithm

Like Needleman-Wunsch, with 2 changes:

- Don't allow negative scores, set them to 0
- Backtrack from cell with highest score, stop at 0

We can make local alignments using the Smith-Waterman algorithm

Like Needleman-Wunsch, with 2 changes:

- Don't allow negative scores, set them to 0
- Backtrack from cell with highest score, stop at 0

Needleman-Wunsch

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
C	-2	0	2	1	0
T	-3	-1	1	1	2

GCAT

GC–T

We can make local alignments using the Smith-Waterman algorithm

Like Needleman-Wunsch, with 2 changes:

- Don't allow negative scores, set them to 0
- Backtrack from cell with highest score, stop at 0

Needleman-Wunsch

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
C	-2	0	2	1	0
T	-3	-1	1	1	2

GCAT

GC–T

Smith-Waterman

	–	G	C	A	T
–	0	0	0	0	0
G	0	1	0	0	0
C	0	0	2	1	0
T	0	0	1	1	2

GC

GC

Smith-Waterman algorithm, mathematical form

$$\mathbf{M}(0, j) = 0$$

first row

$$\mathbf{M}(i, 0) = 0$$

first column

$$\mathbf{M}(i, j) = \max \left(\begin{array}{c} 0 \\ \mathbf{M}(i-1, j) + p \\ \mathbf{M}(i, j-1) + p \\ \mathbf{M}(i-1, j-1) + s(a_j, b_i) \end{array} \right)$$

top
left
diagonal

$s(a_j, b_i)$ = match/mismatch score for sites j and i
in sequences a and b

BLAST

(Basic Local Alignment Search Tool)

BLAST is the primary method to find sequences in modern sequence data bases



Stephen Altschul

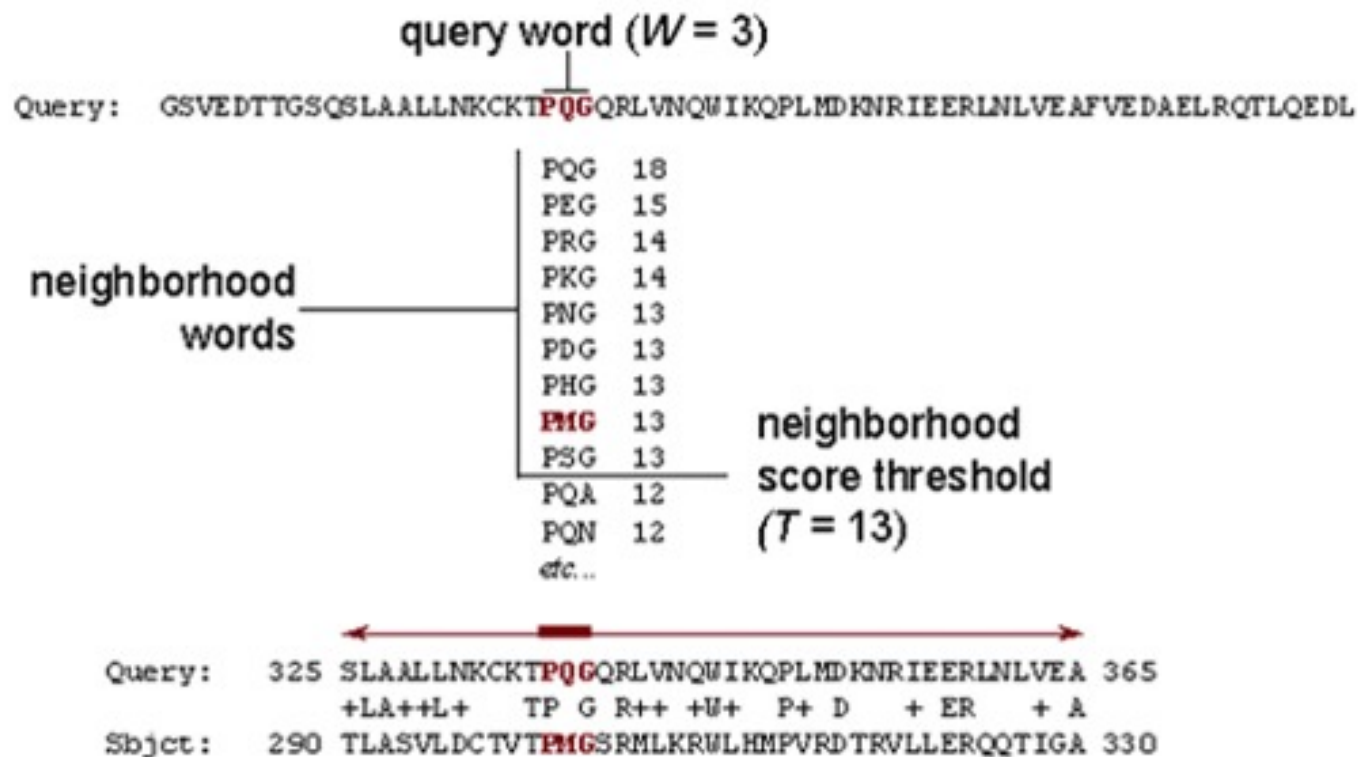
NCBI, NLM, NIH

[Bioinformatics](#)

Verified email at nih.gov

Title	1–20	Cited by	Year
Basic local alignment search tool	SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman Journal of molecular biology 215 (3), 403-410	54882	1990
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs	SF Altschul, TL Madden, AA Schäffer, J Zhang, Z Zhang, W Miller, ... Nucleic acids research 25 (17), 3389-3402	54781	1997
Identification of FAP locus genes from chromosome 5q21.	KW Kinzler, MC Nilbert, LK Su, B Vogelstein, TM Bryan, DB Levy, ... Science (New York, NY) 253 (5020), 661	2060	1991
Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment	CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, JC Wootton	1963	1993

The BLAST Search Algorithm



High-scoring Segment Pair (HSP)

Primary BLAST quality metric: E value

The **Expectation value** or **E value** represents the number of different alignments with scores equivalent to or better than the one observed that are expected to occur in a database search by chance.

The lower the E value, the more significant the score and the alignment.

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) Length: 485 Number of Matches: 1

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score		Expect Method		Identities	Positives	Gaps	
278 bits(710)		1e-86	Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)	
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG					60
		MGQ ISF QEIP FLQEALNIALVAVSLIA+IKGI+NLYKSGLFQF FL LAGRSC++					
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAIIKGIIVNLYKSGLFQFFVFLALAGRSCTEE					60
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYMKGGANIFLIRVSDVSVLMK					120
		FKIGLHTEFQ+V+F+M L +N+ ++LP LC LN S Y+KGG F+I D++VL+					
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP					120
Query	121	EYDVSVYEPEDLGNCNLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS					180
		+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S					
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPFPLCRNRTKTEGFIFQVNTS					180
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD					240
		K + Y KK + GM HL+R + D C GK+C + P+S+ +C D					
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSC LNGKLCLMK----AQPTSWPL-----QCPLD					229
Query	241	HVNTLHFLVRSK 252					
		HVNTLHFL R K					
Sbjct	230	HVNTLHFLTRGK 241					

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) **Length:** 485 **Number of Matches:** 1

sequence we found
(subject sequence)

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

Score		Expect	Method	Identities	Positives	Gaps
278 bits(710)		1e-86	Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG				60
Sbjct	1	MGQ ISF QEIP FLQEALNIALVAVSLIA+IKGI+NLYKSGLFQF FL LAGRSC++ MGQFISFMQEIPTFLQEALNIALVAVSLIAIIKGIIVNLYKSGLFQFFVFLALAGRSCTEE				60
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYMKGGANIFLIRVSDVSVLMK				120
Sbjct	61	FKIGLHTEFQ+V+F+M L +N+ ++LP LC LN S Y+KGG F+I D++VL+ AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP				120
Query	121	EYDVSVEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS				180
Sbjct	121	+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPFPLCRNRTKTEGFIFQVNTS				180
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFYECGTNYLSKCQFD				240
Sbjct	181	K + Y KK + GM HL+R + D C GK+C + P+S+ +C D KTGVNENYAKKFKTGMHHLYREYPDSC LNGKLCLMK----AQPTSWPL-----QCPLD				229
Query	241	HVNTLHFLVRSK 252				
Sbjct	230	HVNTLHFL R K HVNTLHFLTRGK 241				

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) Length: 485 Number of Matches: 1

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score		Expect	Method		Identities	Positives	Gaps
278 bits(710)		1e-86	E value		matrix adjust.	137/252(54%)	168/252(66%) 11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG					60
		MGQ ISF QEIP FLQEALNIALVAVSLIA+IKGI+NLYKSGLFQF FL LAGRSC++					
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAIIKGIIVNLYKSGLFQFFVFLALAGRSCTEE					60
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYMKGGANIFLIRVSDVSVLMK					120
		FKIGLHTEFQ+V+F+M L +N+ ++LP LC LN S Y+KGG F+I D++VL+					
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP					120
Query	121	EYDVSVYEPEDLGNCNLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS					180
		+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S					
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPFPLCRNRTKTEGFIFQVNTS					180
Query	181	KADES RVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSF EYCGTNYLSKCQFD					240
		K + Y KK + GM HL+R + D C GK+C + P+S+ +C D					
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSC LNGKLCLMK----AQPTSWPL-----QCPLD					229
Query	241	HVNTLHFLVRSK 252					
		HVNTLHFL R K					
Sbjct	230	HVNTLHFLTRGK 241					

Query	1	MGQLISFFQEIPVFLQ	EALNIALVAVSLIAV	number and % of exact	60
Sbjct	1	MGQ ISF QEIP FLQ	EALNIALVAVSLIA+	matches, near matches,	60
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI		and no matches	120
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLL	CTLNKSHLYIKGGNASFMISFDDIAVLLP		120
Query	121	EYDVSVEPEDLGNC	LNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS		180
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGH	DWHLDPFFLCRNRTKTEGFIFQVNTS		180
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEE	GKVVCYVTINQCGDPSSF		240
Sbjct	181	KTGVNENYAKKFKTG	MHHLYREYPDSC	LNKGLCLMK----	AQPTSWPL-----QCPLD 229
Query	241	HVNTLHFLVRSK	252		
Sbjct	230	HVNTLHFLTRGK	241		

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) Length: 485 Number of Matches: 1

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
278 bits(710)	1e-86	Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAV	number and % of exact matches, near matches, and no matches		
		MGQ ISF QEIP FLQEALNIALVAVSLIA+ IKGI+NLYKSGLFQF FL LAGRSC++			
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAI IKGI+NLYKSGLFQF FL LAGRSC++			
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI			
		FKIGLHTEFQ+V+F+M L +N+ ++LP ICGD++VL+			
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDI AVL LP			
Query	121	EYDVSVYEPEDLGNCNLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS			
		+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S			
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDP PFLCRNRTKTEGFIFQVNTS			
Query	181	KADES RVYGKKIRNGMRHLFRGFYDPCEE GKVCYVTINQCGDPSSF EYCGTNYLSKCQFD			
		K + Y KK + GM HL+R + D C GK+C + P+S+ +C D			
Sbjct	181	KTGVNENYAKKFKTGMMNLYREYPDSC L NGKLCLMK----AQPTSWPL-----QCPLD			
Query	241	HVNTLHFLVRSK 252	exact match		
		HVNTLHFL R K			
Sbjct	230	HVNTLHFLTRGK 241			

number and % of exact
matches, near matches,
and no matches

exact match

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) Length: 485 Number of Matches: 1

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
278 bits(710)	1e-86	Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAV	number and % of exact		
		MGQ ISF QEIP FLQEALNIALVAVSLIA+	matches, near matches,		
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAI	and no matches		
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI			
		FKIGLHTEFQ+V+F+M L +N+ ++LP I			
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLL			
Query	121	EYDVSVYEPEDLGNCNLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS			
		+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S			
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPFPLCRNRTKTEGFIFQVNTS			
Query	181	KADES RVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSF EYCGTNYLSKCQFD			
		K + Y KK + GM HL+R + D C GK+C + P+S+ +C D			
Sbjct	181	KTGVNENYAKKFKTGMHHLRYEYPDSCLNGKLCLMK----AQPTSWPL-----QCPLD			
Query	241	HVNTLHFLVRSK 252	near match (positive)		
		HVNTLHFL R K			
Sbjct	230	HVNTLHFLTRGK 241			

number and % of exact
matches, near matches,
and no matches

near match (positive)

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) Length: 485 Number of Matches: 1

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
278 bits(710)	1e-86	Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAV	number and % of exact matches, near matches, and no matches		
		MGQ ISF QEIP FLQEALNIALVAVSLIA+			
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAI			
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI			
		FKIGLHTEFQ+V+F+M L +N+ ++LP I			
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLL			
Query	121	EYDVSVYEPEDLGNCNLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS			
		+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S			
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDP PFLCRNRTKTEGFIFQVNTS			
Query	181	KADES RVYGKKIRNGMRHLFRGFYDPCEE GKVCYVTINQCGDPSSF EYCGTNYLSKCQFD			
		K + Y KK + GM HL+R + D C GK+C + P+S+ +C D			
Sbjct	181	KTGVNENYAKKFKTGMHHL YRYPD SCLNGKLCLMK----AQPTSWPL-----QCPLD			
Query	241	HVNTLHFLVRSK 252	no match (gap)		
		HVNTLHFL R K			
Sbjct	230	HVNTLHFLTRGK 241			

Anatomy of a BLAST result

glycoprotein precursor [Junin virus]

Sequence ID: [gb|ABI51595.1|](#) Length: 485 Number of Matches: 1

Range 1: 1 to 241 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score		Expect Method		Identities	Positives	Gaps	
278 bits(710)		1e-86	Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)	
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG					60
		MGQ ISF QEIP FLQEALNIALVAVSLIA+IKGI+NLYKSGLFQF FL LAGRSC++					
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAIIKGIIVNLYKSGLFQFFVFLALAGRSCTEE					60
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYMKGGANIFLIRVSDVSVLMK					120
		FKIGLHTEFQ+V+F+M L +N+ ++LP LC LN S Y+KGG F+I D++VL+					
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP					120
Query	121	EYDVSVYEPEDLGNCNLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKEGSNIQFNIS					180
		+YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S					
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPFPLCRNRTKTEGFIFQVNTS					180
Query	181	KADES RVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD					240
		K + Y KK + GM HL+R + D C GK+C + P+S+ +C D					
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSC LNGKLCLMK----AQPTSWPL-----QCPLD					229
Query	241	HVNTLHFLVRSK 252					
		HVNTLHFL R K					
Sbjct	230	HVNTLHFLTRGK 241					