Global and local alignments

Needleman-Wunsch algorithm, mathematical form

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

first row, p = gap penalty

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

first column

$$\mathbf{M}(i,j) = \max \begin{pmatrix} \mathbf{M}(i-1,j) + p \\ \mathbf{M}(i,j-1) + p \\ \mathbf{M}(i-1,j-1) + s(a_j,b_i) \end{pmatrix} \text{ diagonal}$$

 $s(a_j, b_i)$ = 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
                                       # diagonal
        match = score[i - 1][j - 1] + \setminus
                    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 <u>all</u> nucleotides
- Local: align subsequences with best score

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

```
global alignment: local alignment: GCAT ?
GC-T
```

Like Needleman-Wunsch, with 2 changes:

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

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	С	A	Т
_	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
С	-2	0	2	1	0
Т	-3	-1	1	1	2

GCAT GC-T

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

	ı	G	С	A	Т
_	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
С	-2	0	2	1	0
Т	-3	-1	1	1	2

GCAT GC-T

Smith-Waterman

	-	G	С	A	Т
_	0	0	0	0	0
G	0	1	0	0	0
С	0	0	2	1	0
Т	0	0	1	1	2

GC

GC

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

	ı	G	С	A	Т
-	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
С	-2	0	2	1	0
Т	-3	-1	1	1	2

Smith-Waterman

	ı	G	С	A	Т
_	0	0	0	0	0
G	0	1	0	0	0
С	0	0	2	1	0
Т	0	0	1	1	2

$$\frac{GC}{GC}$$
 or $\frac{GCAT}{GC-T}$

Smith-Waterman algorithm, mathematical form

$$M(0, j) = 0$$

first row

$$M(i,0) = 0$$

first column

$$\mathbf{M}(i,j) = \max \begin{pmatrix} \mathbf{0} \\ \mathbf{M}(i-1,j) + p \\ \mathbf{M}(i,j-1) + p \\ \mathbf{M}(i-1,j-1) + s(a_j,b_i) \end{pmatrix} \text{ top }$$
 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)

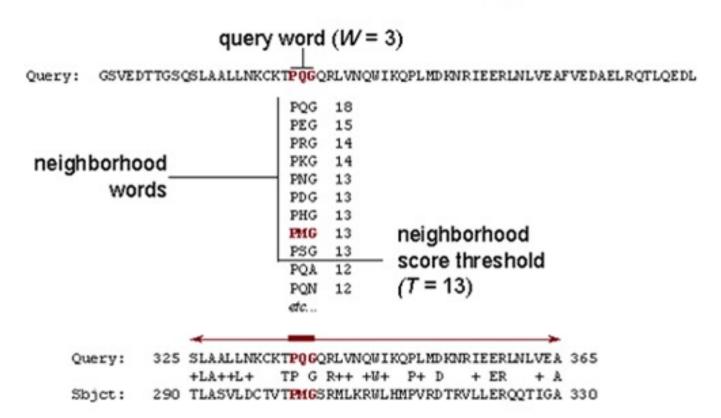
BLAST is the primary method to find sequences in modern sequence data bases



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, SE Altschul, MS Boguski, JS Liu, AE 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.

glycoprotein precursor [Junin virus]

Range 1: 1 to 241 GenPept Graphics

Sequence ID: gb|ABI51595.1| Length: 485 Number of Matches: 1

		241 denrept diapnics
Score		Expect Method Identities Positives Gaps
278 bits	s(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	MGQFISFMQEIPTFLQEALNIALVAVSLIAIIKGIVNLYKSGLFQFFVFLALAGRSCTEE 60
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYYMKGGANIFLIRVSDVSVLMK 120 FKIGLHTEFQ+V+F+M L +N+ ++LP LC LN S Y+KGG F+I D++VL+
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP 120
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSNIQFNIS 180
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPPFLCRNRTKTEGFIFQVNTS 180
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD 240 K + Y KK + GM HL+R + D C GK+C + P+S+ +C D
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSCLNGKLCLMKAQPTSWPLQCPLD 229
Query	241	HVNTLHFLVRSK 252 HVNTLHFL R K
Sbjct	230	HVNTLHFLTRGK 241

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	s(710)	1e-86	Composi	tional mat	trix adjust.	137/252(549	%) 168/252(66 ⁹	%) 11/252	(4%)
Query	1	~		~			SGLFQFIFFLFLA SGLFQF FL LA		60
Sbjct	1	_		~			SGLFQFFVFLAL		60
Query	61					SLCMLNNSFYY LC LN S Y	MKGGANIFLIRVS	SDVSVLMK D++VL+	120
Sbjct	61		~				IKGGNASFMISF	DDIAVLLP	120
Query	121	+ADA +	P D+	C D	W WF	A+GHDW +DP	PMLCRNKTKKEGS P LCRN+TK EG	QNS	180
Sbjct	121	QYDVVI	QHPADMS	WCSKSDDQ	IWLSQWFM	NAVGHDWHLDP	PFLCRNRTKTEGI	FIFQVNTS	180
Query	181	KADESR K +			GFYDPCEE	~	GDPSSFEYCGTNY P+S+	YLSKCQFD +C D	240
Sbjct	181	KTGVNE	NYAKKFK'	r F F F F F F F F F F F F F F F F F F F	EYPDSCLN	GKLCLMK	AQPTSWPL	QCPLD	229
Query	241		FLVRSK FL R K	252					
Sbjct	230	HVNTLH	FLTRGK	241					

glycoprotein precursor [Junin virus]

Range 1: 1 to 241 GenPept Graphics

Sequence ID: gb|ABI51595.1| Length: 485 Number of Matches: 1

itunge 1		<u> </u>	ii cpt oi	артнес	2					TICKLI III		LVIOUS
Score		Expect	E val	ΠΘ			Identit		Positi		Gaps	
278 bits	(710)	1e-86	Composil	tional	matrix a	adjust.	137/25	2(54%)	168/2	52(66%)	11/252(4%)
Query	1									FFLFLAGF FL LAGF		60
Sbjct	1	-	~	_					_	VFLALAGE		60
Query	61		HTEFQSVT							FLIRVSDV F+I D+	7SVLMK -+VL+	120
Sbjct	61		-							FMISFDDI		120
Query	121		YEPEDLGI P D+		SDSSWA D W		IALGHDW A+GHDW			TKKEGSNI TK EG	QFNIS O N S	180
Sbjct	121			_						TKTEGFIE	~	180
Query	181	KADESRY	VYGKKIRN Y KK +						PSSFE P+S+	YCGTNYLS	KCQFD +C D	240
Sbjct	181	KTGVNE								L		229
Query	241	HVNTLHI HVNTLHI		252								
Sbjct	230	HVNTLHI		241								

glycoprotein precursor [Junin virus]

Range 1: 1 to 241 GenPept Graphics

Sequence ID: gb|ABI51595.1| Length: 485 Number of Matches: 1

italige .		241 denrept drapnics		* NOXC ITO	iccii — Ficvio	us
Score		Expect Method	Identities	Positives	Gaps	٦
278 bits	s(710)	1e-86 Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIZ	av number	and % of	exact	
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLI	AI matches	s, near ma	atches,	
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELP; FKIGLHTEFQ+V+F+M L +N+ ++LP	and no	matches	7SVLMK 12 -+VL+	0
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLP				0
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFS: +YDV + P D+ C D W WF	IALGHDWLMDPPM A+GHDW +DPP		QFNIS 18 O N S	0
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFM			~	0
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEC			SKCQFD 24 +C D	0
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSCLNO		PTSWPL	-QCPLD 22	9
Query	241	HVNTLHFLVRSK 252 HVNTLHFL R K				
Sbjct	230	HVNTLHFLTRGK 241				

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	MGQLISFFQEIPVFLQEALNIALVAVSLIZ MGQ ISF QEIP FLQEALNIALVAVSLIZ	number and % of exact
Sbjct 1	MGQFISFMQEIPTFLQEALNIALVAVSLI	
Query 61	TFKIGLHTEFQSVTFTMQRLLANHSNELPS FKIGLHTEFQ+V+F+M L +N+ ++LP	and no matches
Sbjct 61		LLCTLNKSHLYIKGGNASFMISFDDIAVLLP 120
Query 121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFS: +YDV + P D+ C D W WF	IALGHDWLMDPPMLCRNKTKKEGSNIQFNIS 180 A+GHDW +DPP LCRN+TK EG ON S
Sbjct 121		NAVGHDWHLDPPFLCRNRTKTEGFIFQVNTS 180
Query 181		GKVCYVTINQCGDPSSFEYCGTNYLSKCQFD 240 GK+C + P+S+ +C D
Sbjct 181		GKLCLMKAQPTSWPLQCPLD 229
Query 241	HVNTLHFLVRSK 252 exact	match
Sbjct 230	HVNTLHFLTRGK 241	

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	MGQLISFFQEIPVFLQEALNIALVAVSLI	number and % of exact
Sbjct 1	MGQFISFMQEIPTFLQEALNIALVAVSLI	
Query 61	TFKIGLHTEFQSVTFTMQRLLANHSNELPS FKIGLHTEFQ+V+F+M L +N+ ++LP	and no matches +-v = 120
Sbjct 61		LLCTLNKSHLYIKGGNASFMISFDDIAVLLP 120
Query 121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFS: +YDV + P D+ C D W WF	IALGHDWLMDPPMLCRNKTKKEGSNIQFNIS 180 A+GHDW +DPP LCRN+TK EG O N S
Sbjct 121		NAVGHDWHLDPPFLCRNRTKTEGFIFQVNTS 180
Query 181		GKVCYVTINQCGDPSSFEYCGTNYLSKCQFD 240 GK+C + P+S+ +C D
Sbjct 181		GKLCLMKAQPTSWPLQCPLD 229
Query 241	HVNTLHFLVRSK 252 near	match (positive)
Sbjct 230	HVNTLHFLTRGK 241	

glycoprotein precursor [Junin virus]

Range 1: 1 to 241 GenPept Graphics

Sequence ID: gb|ABI51595.1| Length: 485 Number of Matches: 1

Range 1. 1 to 241 denrept draphics			
Score		Expect Method Identities Positives Gaps	
278 bit	s(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+	
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAI matches, near matches,	
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI and no matches 12 FKIGLHTEFQ+V+F+M L +N+ ++LP I and no matches	
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP 12	
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSNIQFNIS 18 +YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S	
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPPFLCRNRTKTEGFIFQVNTS 18	
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD 24 K + Y KK + GM HL+R + D C GK+C + P+S+ +C D	
Sbjct	181	KTGVNENYAKKFKTGMHHLYRAYPDSCLNGKLCLMKAQPTSWPLQCPLD 22	
Query	241	HVNTLHFLVRSK 252 no match	
Sbjct	230	HVNTLHFLTRGK 241	

glycoprotein precursor [Junin virus]

Range 1: 1 to 241 GenPept Graphics

Sequence ID: gb|ABI51595.1| Length: 485 Number of Matches: 1

Range 1. 1 to 241 denrept draphics				
	Expect Method Identities Positives Gaps			
s(710)	1e-86 Compositional matrix adjust. 137/252(54%) 168/252(66%) 11/252(4%)			
1	MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG 60 MGQ ISF QEIP FLQEALNIALVAVSLIA+IKGI+NLYKSGLFQF FL LAGRSC++			
1	MGQFISFMQEIPTFLQEALNIALVAVSLIAIIKGIVNLYKSGLFQFFVFLALAGRSCTEE 60			
61	FKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYYMKGGANIFLIRVSDVSVLMK 120 FKIGLHTEFQ+V+F+M L +N+ ++LP LC LN S Y+KGG F+I D++VL+			
61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP 120			
121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSNIQFNIS 180 +YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG O N S			
121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPPFLCRNRTKTEGFIFQVNTS 180			
181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD 240 K + Y KK + GM HL+R + D C GK+C + P+S+ +C D			
181	KTGVNENYAKKFKTGMHHLYREYPDSCLNGKLCLMKAQPTSWPLQCPLD 229			
241	HVNTLHFLVRSK 252 HVNTLHFL R K			
230	HVNTLHFLTRGK 241			
	1 N 1 N 61 7 61 H 121 H 121 G 181 H 181 H			