## Global and local alignments

## 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 PGC-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

	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

#### 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**

	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

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

	-	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

	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

GC GCAT
GC GC-T

## Smith-Waterman algorithm, mathematical form

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

first row

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

first column

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

$$\mathbf{M}(i-1,j-1) + s(a_j,b_i)$$
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)

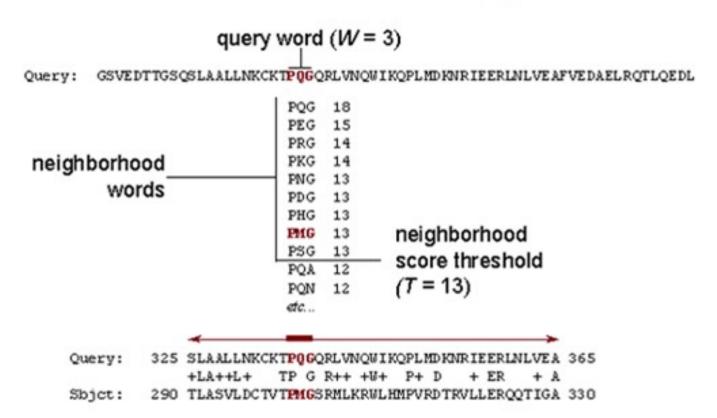
## BLAST is the primary method to find sequences in modern sequence data bases



# 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, 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.

glycoprotein precursor [Junin virus]

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

Range 1:	1 to	241 GenPept Graphics		▼ Next Ma	itch 🔺 Pr	evious 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	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAV				60
Sbjct 1	1	MGQFISFMQEIPTFLQEALNIALVAVSLIA		~		60
Query 6	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI FKIGLHTEFQ+V+F+M L +N+ ++LP I			/SVLMK ++VL+	120
Sbjct 6	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLI				120
Query 1	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIA +YDV + P D+ C D W WF A	ALGHDWLMDPPM: A+GHDW +DPP :		IQFNIS O N S	180
Sbjct 1	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNA			~	180
Query 1	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGI K + Y KK + GM HL+R + D C GI		PSSFEYCGTNYLS P+S+	SKCQFD +C D	240
Sbjct 1	181	KTGVNENYAKKFKTGMHHLYREYPDSCLNG		PTSWPL	-QCPLD	229
Query 2	241	HVNTLHFLVRSK 252 HVNTLHFL R K				
Sbjct 2	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 N	1ethod				Identiti	es	Positiv	res	Gaps	
278 bit	s(710)	1e-86 C	Composit	ional ma	atrix adju	ıst.	137/252	(54%)	168/25	2(66%)	11/252(	4%)
Query	1	MGQLISFI	~	~					-			60
Sbjct	1	MGQFISF	~	-					-			60
Query	61	TFKIGLHT			LANHSNE						7SVLMK -+VL+	120
Sbjct	61	AFKIGLH	~								AVLLP	120
Query	121	EYDVSVYI +YDV +					ALGHDWI A+GHDW				QFNIS Q N S	180
Sbjct	121	QYDVVIQ	HPADMSW	CSKSDD	QIWLSQW	/FMN	AVGHDWI	HLDPPF	LCRNRT	KTEGFIE	TOVNTS	180
Query	181	KADESRV			RGFYDPO			INQCGD		CGTNYLS	KCQFD +C D	240
Sbjct	181	KTGVNEN	YAKKFKI	GMHHLY	REYPDSC	LNG	KLCLMK-	AQ	PTSWPL		-QCPLD	229
Query	241	HVNTLHF1 HVNTLHF1		252								
Sbjct	230	HVNTLHFI	LTRGK	241								

glycoprotein precursor [Junin virus]

Range 1: 1 to 241 GenPept Graphics

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

		ZTZ OCIII	CPC CI	артнос	2							011001
Score		Expect	E val	ПР			Identit		Positi		Gaps	
278 bits	s(710)	1e-86	Composit	ional	matrix	adjust.	137/25	2(54%)	168/2	52(66%)	11/252	(4%)
Query	1	MGQLISF										60
Sbjct	1	MGQFISF	~	_					_			60
Query	61	TFKIGLH FKIGLH			-		SLCMLNN LC LN				/SVLMK ++VL+	120
Sbjct	61	AFKIGLH	~								IAVLLP	120
Query	121	EYDVSVY +YDV +			SDSSWA D W		IALGHDW A+GHDW				QFNIS O N S	180
Sbjct	121	QYDVVIQ	HPADMSV	CSKS							~	180
Query	181	KADESRV K +	YGKKIRN Y KK +						PSSFEY	YCGTNYLS	SKCQFD +C D	240
Sbjct	181	KTGVNEN								L		229
Query	241	HVNTLHF:		252								
Sbjct	230	HVNTLHF		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 Ochi ept Oraphics		THEXE FIE	
Score		<b>Expect Method</b>	Identities	Positives	Gaps
278 bits	s(710)	1e-86 Compositional matrix ad	just. 137/252(54%)	168/252(66%)	11/252(4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAV	sliav number	and % of	exact
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAV			atches,
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSN FKIGLHTEFQ+V+F+M L +N+ +	ELPSI and no	matches	7SVLMK 120 -+VL+
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPH			
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIF +YDV + P D+ C D W	WFSIALGHDWLMDPPM WF A+GHDW +DPP		QFNIS 180 O N S
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQ			~
Query	181	KADESRVYGKKIRNGMRHLFRGFYDE K + Y KK + GM HL+R + D		PSSFEYCGTNYLS	SKCQFD 240 +C D
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDS			
Query	241	HVNTLHFLVRSK 252 HVNTLHFL R K			
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

ixange		2-12 <u>Genrepe</u> Graphics
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+ IXCHNIXKS LFOR FL LAGRSCH
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	KTGVNENYAKKFKTGMHKLYREYPDSCLNGKLCLMKAQPTSWPLQCPLD 22
Query	241	HVNTLHFLVRSK 252 exact 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 60	241 denrept diapnics		V NEXT MA	ccii = ric	vious
Score		Expect Method Ide	entities	Positives	Gaps	
278 bits	s(710)	1e-86 Compositional matrix adjust. 13	7/252(54%)	168/252(66%)	11/252(4	4%)
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAVI MGQ ISF QEIP FLQEALNIALVAVSLIA+I	number	and % of	exact	60
Sbjct	1	~ ~ ~	_	s, near ma	atches	s <sup>6,0</sup>
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSL FKIGLHTEFQ+V+F+M L +N+ ++LP I	and no r	matches	SVLMK +VL+	120
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLC				120
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIAL	GHDWLMDPPMI GHDW +DPP 1		QFNIS O N S	180
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAV			~	180
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKV K + Y KK + GM HL+R + D C GK+		PSSFEYCGTNYLS P+S+	KCQFD +C D	240
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSCLNGKL				229
Query	241	HVNTLHFLVRSK 252 near m	atch (pc	sitive)		
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

Name 1. 1 to 2-11 denirept oraphics								
Score		Expect Method Identities Positives Gaps	٦					
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+ INC. THE PROPERTY OF T						
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIAI matches, near matches,						
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPSI and no matches 12 FKIGLHTEFQ+V+F+M L +N+ ++LP I and no matches	0					
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPLLCTLNKSHLYIKGGNASFMISFDDIAVLLP 12	0					
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSNIQFNIS 18 +YDV + P D+ C D W WF A+GHDW +DPP LCRN+TK EG Q N S	0					
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMNAVGHDWHLDPPFLCRNRTKTEGFIFQVNTS 18	0					
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD 24 K + Y KK + GM HL+R + D C GK+C + P+S+ +C D	0					
Sbjct	181	KTGVNENYAKKFKTGMHHLYRAYPDSCLNGKLCLMKAQPTSWPLQCPLD 22	9					
Query	241	HVNTLHFLVRSK 252 no 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		<b>Expect Method</b>	Identities	Positives	Gaps				
278 bits	s(710)	1e-86 Compositional matrix adjust.	137/252(54%)	168/252(66%)	11/252(	4%)			
Query	1	MGQLISFFQEIPVFLQEALNIALVAVSLIAMGQ ISF QEIP FLQEALNIALVAVSLIA				60			
Sbjct	1	MGQFISFMQEIPTFLQEALNIALVAVSLIA				60			
Query	61	TFKIGLHTEFQSVTFTMQRLLANHSNELPS FKIGLHTEFQ+V+F+M L +N+ ++LP			7SVLMK -+VL+	120			
Sbjct	61	AFKIGLHTEFQTVSFSMVGLFSNNPHDLPI				120			
Query	121	EYDVSVYEPEDLGNCLNKSDSSWAIHWFS1 +YDV + P D+ C D W WF	ALGHDWLMDPPM A+GHDW +DPP		QFNIS O N S	180			
Sbjct	121	QYDVVIQHPADMSWCSKSDDQIWLSQWFMN			~	180			
Query	181	KADESRVYGKKIRNGMRHLFRGFYDPCEEG K + Y KK + GM HL+R + D C		PSSFEYCGTNYLS P+S+	KCQFD +C D	240			
Sbjct	181	KTGVNENYAKKFKTGMHHLYREYPDSCLNG		PTSWPL	-QCPLD	229			
Query	241	HVNTLHFLVRSK 252 HVNTLHFL R K							
Sbjct	230	HVNTLHFLTRGK 241							