BLAST MCDB 187

BLAST

- Basic Local Alignment Sequence Tool
- Uses shortcut to compute alignments of a sequence against a database very quickly
- Typically takes about a minute to align a sequence against a database of one million sequences
- Accurately computes the statistical significance of the alignment

One of most highly cited papers

000	blast - Goo	ale Sch	olar				
+ http://scholar.google.com/scholar?hl=en&g=blast&btnG=Search&as_sdt=0%20	C5&as vlo=19	900&as	vhi=2011&as	vis=0)		
Pellegrinilab PDF to Word UCLA/UCOP Reservation GATHER UniPROBE Database	Blackboard	IGET	Pellegrinilab	IDP	IPA	Precipitation	DAVID
Web Images Videos Maps News Shopping Gmail more V						matteope	@gmail
Google scholar blast Search Advanced Sch	oolar Search						
Scholar Articles and patents \$ 1900 - 2011 include citations \$ Created Scholar	ate email alert						
Gapped BLAST and PSI-BLAST: a new generation of protein database search p SF Altschul, TL Madden, AA Schaffer Nucleic acids, 1997 - Oxford Univ Press The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinement described here permits the execution time of the BLAST programs to be decreased Cited by 33636 - Related articles - BL Direct - All 200 versions	e s		[PDF] fro UC-eLii	<u>m nil</u> 1ks	<u>h.gov</u>		
BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequence TA Tatusova FEMS microbiology letters, 1999 - Wiley Online Library 'BLAST 2 Sequences', a new BLAST-based tool for aligning two protein or nucleotide sequences, is described. While the standard BLAST program is widely used to search for homolog sequences in nucleotide and protein databases, one often needs to compare only two <u>Cited by 1298</u> - <u>Related articles</u> - <u>All 12 versions</u>	e <mark>s</mark> jous		<u>[PDF] fro</u> <u>UC-eLii</u>	m cs 1ks	<u>hl.org</u>	1	
V-BLAST: An architecture for realizing very high data rates over the rich-scatterin PW Wolniansky, GJ Foschini, 1998. ISSSE 98, 2002 - ieeexplore.ieee.org V-BLAST: An Architecture for Realizing Very High P. W. Wolniansky, GJ Foschini, G. D. Golden, R. A. Valenzuela Bell Laboratories, Lucent Technologies, Crawford Hill Laboratory 791 Holmdel-Keyport Rd., Holmdel, NJ 07733 ABSTRACT Recent information theory Cited by 2149 - Related articles - UC-eLinks - All 23 versions	ig wireless	<u>chann</u>	iel [PDF] fro	maa	<u>u.dk</u>		
Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of	f chronic my	<u>eloid</u>	UC-eLi	<u>ıks</u>			

Global vs Local Alignments

- In 1981 Smith and Waterman introduced a modification to the Needleman-Wunsch algorithm to search for optimal alignment fragments
- Positions in the alignment matrix with negative scores are set to zero, and are considered new start positions

Definitions

- Maximum Segment Pair (MSP): highest scoring pair of identical length segments from 2 sequences
- Word pair: segment pair of fixed length w
- T: minimum word pair score

Algorithm

- Pre-compute all words that score above T against the query sequence
- Search for these words in a database
- For each match, extend the alignment to generate the maximum segment pair
- Compute statistics of the resulting score

Step I : break up sequence into words

- ATCGTCTATTCCCGG
- w=4 letter words
 - ATCG
 - TCGT
 - CGTC
 - etc.

Step 2: find high scoring matches

- Start with word I:ATCG
- score identities as +1, and mismatches a 0
- find all words that have a score of at least T=3
 - ATCAATCT
 - ATCC

etc.

Search for high scoring words in database

- Find all sequences in the database that contain a high scoring word
- For example:
- CCGGCTATCATCTATTCCCGGTTCG

Extend ALignments Around matching word

CCGGCTATCATCTATTCCCGGTTCG ATCGTCTATTCCCGG

 The Yellow sequence alignment corresponds to the MSP wth score S

NCBI BLAST - Netsca	ape			- P	×
<u>File Edit View Go B</u>	ookmarks <u>T</u> ools <u>W</u> indow <u>H</u> elp				
6, 0, 0	E http://www.ncbi.nlm.nih.gov/BLAST/		Search	2, N)
🔺 🗔 🖂 Mail 🚴 AIN	1 🐔 Home 😱 Radio 🕅 Netscape 🔍 Search 🖾 Boo	okmarks 🛇 Instant Message 🛇 WebMail 🛇 Radio 🛇 People	🛇 Yellow Pages 🛇 Download	🛇 Calendar	**
S NCBI BLAST				6	
S NCBI		BLAST			
PubMed	Entrez BLAST ON	IIM Taxonomy Structure			
Info	Nucleotide	Protein			
 FAQs News References Credits 	 Discontiguous megablast Megablast Nucleotide-nucleotide BLAST (blastn) Search for short, nearly exact matches Search trace archives with megablast or discontiguous megablast 	 Protein-protein BLAST (blastp) PHI- and PSI-BLAST Search for short, nearly exact matches Search the conserved domain database (rpsblast) Search by domain architecture (cdart) 			
 Program selection guide Tutorial URL API guide Download 	 Translated Translated query vs. protein database (blastk) Protein query vs. translated database (tblastn) Translated query vs. translated database (tblastk) 	Genomes Human, mouse, rat Fugu rubripes, zebrafish Insects, nematodes, plants, fungi, malaria Microbial genomes, other eukaryotic genomes 			111
Executables Databases	Special	Meta			
 Source code Support Helpdesk Mailing list 	 Align two sequences (bl2seq) Screen for vector contamination (VecScreen) Immunoglobin BLAST (IgBlast) 	 Retrieve results by RID Get this page with javascript-free links 			
	<u>Disc</u> <u>Privacy</u> <u>Acces</u> Valid <u>XHTI</u>	<u>laimer</u> <u>statement</u> <u>ssibility</u> ML 1.0, <u>CSS</u> .			1
	one			-I- ® 6	
🐉 start 🛛 🔞 т	-COFFEE server - N 🔕 NCBI BLAST - Netscape 🛽 🛽	postdoc - Inbox for d 📴 Bioinfo3.1	2 🛛 🕄 🏈 🖥 🖓 🧐	💭 🚛 10:11 AF	1
	bi	oinformatics.ca			

Different Flavours of BLAST

- BLASTP protein query against protein DB
- BLASTN DNA/RNA query against GenBank (DNA)
- **BLASTX** 6 frame trans. DNA query against proteinDB
- TBLASTN protein query against 6 frame GB transl.
- **TBLASTX** 6 frame DNA query to 6 frame GB transl.
- **PSI-BLAST** protein 'profile' query against protein DB
- PHI-BLAST protein pattern against protein DB



Running NCBI BLAST

👔 🚴 Instant Message 🛛 🖳 Secure	Web Shop 🖳 My Pre	esario 🖳 Compaq At H	ome 🖳 Compaq Support	🖳 Smart Update!	
S NCBI			BLAS	ST	
Nucleotide	Protein	Translations	Retrieve results for a	n RID	
Search			× ×		
Set subsequence From:	To:				
Choose database nr	•			-Click	BLAS
Do CD-Search					
Now: BL	AST! or Reset qu	ery Reset all			
					•
iff)http://www.n	cbi.nlm.nih.gov			🛞 🛀 🚽 🖓	11

MT0895

• MMKIQIYGTGCANCQMLEKNAREAVKELGID AEFEKIKEMDQILEAGLTALPGLAVDGELKI MGRVASKEEIKKILS

Lecture 3.1



Formatting Results

r r r						
RCBI Blast						
Open a new tab NCB	Prote	fc	ormatting B			
Nucleonide				ve resolts for all MD		
Your request has b	een successfully sub	mitted and put into	the Blast Queue.			
Query = (77 letter:)					
No putative conse	erved domains hav	ve been detected				
The request ID is 1	076347219-30071-1	7917880302.BLAST(23			
Format! or R	set all					
The results are estima	ted to be ready in 43 s	seconds but may be do	ne sooner.			
Please press "FORM/	T!" when you wish t	o check your results. Y	ou may change the forms	atting options for your res	ult via the form below and	press "FORMAT!" again. You may also
	ferent search by enter	nng any other valid rec	quest ID to see other rece	nt jobs.		
request results of a da						
request results of a da						
Format						
Format Show	Graphical Overvi	iew 🗹 Linkout 🗹 Sequ	uence Retrieval 🗹 NCBI-	gi Alignment 💌 in H	ГML <mark>⊻</mark> format	
Format Show Number of:	Graphical Overvi Descriptions 100	iew 🗹 Linkout 🗹 Sequences 50 💉	uence Retrieval 🔽 <u>NCBI-</u>	gi Alignment 💌 in H	ſML <mark>▼</mark> format	
Format Show Number of: <u>Alignment view</u>	Craphical Overvi Descriptions 100	iew 🗹 Linkout 🗹 Seq 🜱 Alignments 50 🛐	uence Retrieval 🗹 <u>NCBI-</u>	gi Alignment 💌 in H	ſML <mark>≰</mark> format	
Format Show Number of: <u>Alignment view</u> Format for	Caraphical Overvi Descriptions 100 Pairwise	iew 🗹 Linkout 🗹 Seq 🖌 Alignments 50 🔹	uence Retrieval 🗹 <u>NCBI-</u>	gi Alignment 💌 in H	ΓML ▼format	

BLAST Format Options

Show	🗹 <u>Graphical Overview</u> 🗹 <u>Linkout</u> 🗹 <u>Sequence Retrieval</u> 🗹 <u>NCBI-gi</u> Alignment 💌 in HTML 🛛 🚩 for
Number of:	Descriptions 100 💌 Alignments 50 💌
Alignment view	Pairwise 🔽
Format for PSI-BLAST	with inclusion threshold: 0.005
Limit results by entrez query	or select from: All organisms
Expect value range:	
Layout:	Two Windows 💌 Formatting options on page with results: None 💌
<u>Autoformat</u>	Semi-auto 💌

RID=1076347219-30071-1791788	0302.BLASTQ3, - Netscape	
0000	http://www.ncbi.nlm.nih.gov/BLAST/Blast.cg	🖸 🔍 Search 🖉 🔊
Mail & AIM 🛳 Home 😱 I	Radio 🕅 Netscape 🔍 Search 🛛 Bookmarks 🛇 Instant Message 🛇 WebMail 🛇 Radio	🛇 People 🛇 Yellow Pages 🛇 Download 🛇 Calendar
RID=1076347219-30071-1791788030	2.BL	(
S NCBI	results of BLAST	
BLASTP 2.2.8 [Jan-05-2004]		
Reference: Altschul, Stephen F., Thomas Jinghui Zhang, Zheng Zhang, "Gapped BLAST and PSI-BLAST: programs", Nucleic Acids Re:	L. Madden, Alejandro A. Schäffer, Webb Miller, and David J. Lipman (1997), a new generation of protein database search 3. 25:3389-3402.	
RID: 1076347219-30071-179178	30302.BLASTQ3	
Query= (77 letters)		
Database: All non-redundant of translations+PDB+SwissProt+P 1,624,011 sequence	GenBank CDS IR+PRF es; 534,067,077 total letters	
If you have any problems or please refer to the BLAST FA	questions with the results of this search	
Taxonomy reports		
S 🖂 🔏 💇 🔲 Done	Distribution of 43 Blast Hits on the Onery Sequence	

Distribution of 43 Blast Hits on the Query Sequence



bioinformatics.ca

RID=1076347219-30071-17917880302.BLASTQ3, - Netscape				- 7 🛛
Shttp://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi			Search	2. 🔊
🖌 🗔 Mail 🚴 AIM 🐔 Home 😱 Radio 🔤 Netscape 🔍 Search 🖾 Bookmarks 🛇 Inst	ant Message	e 🛇 WebMail 🛇 Radio	🖓 People 🛇 Yellow Pages 🛇 Download	🛇 Calendar 🛛 »
RID=1076347219-30071-17917880302.BL				×
	Score	Е		
Sequences producing significant alignments:	(bits)	Value		
gi 15678915 ref NP 276032.1 conserved protein [Methanother	124	5e-28 S		
gi 23111526 ref ZP 00097156.1 COG0526: Thiol-disulfide iso	71	4e-12		
gi 21226839 ref NP 632761.1 conserved protein [Methanosarc	70	9e-12		
gi 20092734 ref NP 618809.1 conserved hypothetical protein	69	2e-11		
gi 15643756 ref NP 228804.1 conserved hypothetical protein	65	4e-10		
gi 21674543 ref NP 662608.1 glutaredoxin family protein [C	64	5e-10		
gi 23054435 ref 2P 00080592.1 COG0526: Thiol-disulfide iso	62	3e-09		
gi[22971774 ref[ZP 00018701.1] hypothetical protein [Chloro		2e-08		
gi 39998047 ref NP 953998.1 redox-active disulfide protein		2e-08		
gi[34557156[ref]NP 906971.1] hypothetical protein WS0755 [W	_58	3e-08		
gi 17229003 ref NP 485551.1 hypothetical protein [Nostoc s	_57	7e-08		
gi 15668761 ref NP 247560.1 conserved hypothetical protein	_57	7e-08		
gi 23048165 ref 2P 00075893.1 COG0526: Thiol-disulfide iso	56	1e-07		
gi[21228351[ref]NP 634273.1] hypothetical protein [Methanos	_55	4e-07		
gi[29345530]ref[NP 809033.1] conserved hypothetical protein		1e-06		
g1 20092738 ref NP 618813.1 conserved hypothetical protein		1e-06		
gi 11499829 ref NP 071073.1 conserved hypothetical protein		1e-06		
g1[22299430[ref]NP 682677.1] ORF_ID:ts11887~hypothetical pr	52	Ze-06		
g1 29346208 ref NP 809711.1 conserved hypothetical protein	52	3e-06		
g1[23001521]ref[ZP 00045426.1] COG0526: Thiol-disulfide iso	46	Ze-04		
gi[23015000]ref[2P 00054791.1] COG0526: Thiol-disulfide iso	45	3e-04		
g1 22960655 ref 2P 00008294.1 COG0526: Thiol-disulfide iso	45	3e-04		
gi[24372130]ref[NP 716172.1] redox-active disulfide protein	42	0.002		
g1[39936619[ref]WP 948895.1] In101-disulfide isomerase and	39	0.016		
gillioggovispir42035/Into METIM Probable Intoredoxin (Giuta		0.17		
gi 15678829 ref NP 275946.1 thioredoxin [Methanothermobact	34	0.46 🔛		
gi 23060730 ref 2P 00085617.1 COG0438: Glycosyltransferase	34	0.77		
gi[23121123[ref]ZP 00103523.1] COG0526: Thiol-disulfide iso	32	1.9		
gi 34860800 ref XP 215715.2 similar to Alcohol dehydrogena	32	3.5		
gi 19684184 gb AAH26035.1 C4orf9 protein [Homo sapiens]	32	3.5 L		
gi 17933966 ref NP 530756.1 glutamine amidotransferase [Ag	32	3.8		
	~			
Start 🔊 5 Netscape 🚽 💷 Bioinfo3.1			2 2 3 4 4 4 4 4) 🕰 📕 10:27 AM

RID=1076347219-30071-17917880302.BLASTQ3, - Netscape	
Search Search	۷. 🔊
🖌 🖂 Mail 🚴 AIM 🐔 Home 🎧 Radio 🔤 Netscape 🔍 Search 🖾 Bookmarks 🛇 Instant Message 🛇 WebMail 🛇 Radio 🛇 People 🛇 Yellow Pages 🛇 Download 🛇	Calendar »
RID=1076347219-30071-17917880302.BL	×
Sequences producing significant alignments: Score E (bits) Value gi 15678915 ref NP 276032.1 conserved protein [Methanother 124 Se-22 S gi 23111526 ref ZP 00097156.1 COG0526: Thiol-disulfide iso 71 He-12	
gi 21226839 ref NP 632761.1 conserved protein [Methanosarc 70 9e-12 gi 20092734 ref NP 618809.1 conserved hypothetical protein 69 2e-11	
gi 15643756 ref NP 228804.1 conserved hypothetical pro NKBI Sequence Structure Alignment Visualization Service - Netscape gi 21674543 ref NP 662608.1 glutaredoxin family protei gi 23054435 ref ZP 00080592.1 COG0526: Thiol-disulfide. O O O O C Phtp://www.ndbi.nb.gov/Sructure/dist/dist.og?blat_RID=1076347219-30071-17917000302.BLASTOR hypothetical protein [Ch_, II, IMMa & Ann & Hone Q Rado M Netscape Q Search Bookmarks Vinstant Message Q WebMat Q Rado V People Q	Réfer Calendar *
gi 39998047/ref NP 953998.1 redox-active disulfide pro gi 34557156/ref NP 906971.1 hypothetical protein US075 gi 17229003/ref NP 485551.1 hypothetical protein [Nost gi 23048165/ref 2P 00075893.1 conserved hypothetical protein [Nost gi 21228351/ref NP 634273.1 conserved hypothetical protein [Meth gi 2092738/ref NP 634273.1 hypothetical protein [Meth gi 20092738/ref NP 618813.1 conserved hypothetical protein [Meth	
gi 11499829 ref NP 071073.1 conserved hypothetical pro gi 22299430 ref NP 682677.1 ORF_ID:ts11887~hypothetical pro gi 29346208 ref NP 809711.1 conserved hypothetical pro gi 2301521 ref 2P 00045426.1 COG0526: Thiol-disulfide gi 22960655 ref 2P 00008294.1 COG0526: Thiol-disulfide gi 24372130 ref NP 716172.1 redox-active disulfide pro	
<u>gi 39936619 ref NP 948895.1 </u> Thiol-disulfide isomerase <u>gi 1169967 sp P42035 THIO METTM</u> Probable Thioredoxin (G <u>gi 23060730 ref ZP 00085617.1 </u> COG0438: Glycosyltransfe <u>gi 23121123 ref ZP 00103523.1 </u> COG0526: Thiol-disulfide	
gi 34860800 ref XP 215715.2 similar to Alcohol dehydro gi 19684184 gb AAH26035.1 C4orf9 protein [Homo sapiens gi 17933966 ref NP 530756.1 glutamine amidotransferase	
S Click here to begin	* 日 1 () 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
🤣 start 🔊 🔊 5 Netscape 🚽 📴 Bioinfo3.1 🧷 😰 👶 🐁 🖗 🖉 🖉	🗊 💭 10:27 AM

RID=1076347219-30071-17917880302.BLASTQ3, - Netscape	
C C T C T C T C T C T C T C T C T C T C	🖸 🔍 Search
🖌 / 🖬, 🛛 🖼 Mail 🚴 AIM 🐔 Home 🎧 Radio 🔤 Netscape 🔍 Search 🖾 Bookmarks 🛇 Instant Message 🛇 WebMail 🛇 Radio 🛇 People	Yellow Pages 🛇 Download 🛇 Calendar 🛛 »
RID=1076347219-30071-17917880302.BL DecusLink - Netscape	
Sequences producing significant alignments:	2(pg) Search 🖧 🔊 ant Message 🛇 WebMal 🛇 Radio 🛇 People 🛇 Yellow Pages 🛇 Download 🛇 Calendar 🔅
gil 15678915 ref NP 276032.11conserved protein [Ngil 23111526 ref IP 00097156.11COG0526: Thiol-disgil 21226339 ref NP 632751.11conserved protein [Ngil 21226339 ref NP 632751.11conserved protein [Ngil 212092734 ref NP 618809.11conserved hypotheticgil 21674543 ref NP 228804.11gil acredoxin familygil 23054435 ref IP 00080592.11cogsical familygil 23054435 ref IP 00080592.11cogsical familygil 23054435 ref IP 953998.11cogsical familygil 23054435 ref NP 485551.11hypothetical proteingil 23048165 ref IP 963971.11hypothetical proteingil 23048165 ref IP 906971.11hypothetical proteingil 23048165 ref IP 906971.11hypothetical proteingil 23048165 ref IP 906971.11conserved hypotheticgil 23048165 ref IP 906971.11hypothetical proteingil 23048165 ref IP 906971.11hypothetical proteingil 228351 ref NP 638277.11conserved hypotheticgil 22946208 ref NP 63813.11conserved hypotheticgil 22946208 ref NP 682677.11ORF ID:tsl1887~hypotgil 22946208 ref NP 682677.11conserved hypotheticgil 23015001[ref IP 809731.11conserved hypotheticgil 23015001[ref IP 8097173.11conserved hypotheticgil 23015001[ref IP 8097173.11conserved hypotheticgil 23015001[ref IP 8097173.11conserved hypotheticgil 23015001[ref IP 8097173.11conserved hypotheticgil 23015001[ref IP 9005426.11conserved hypotheticgil 23015001[ref IP 9005426.11conserved hypothetic </td <td></td>	
gi 39936619 ref NP 948895.1 Thiol-disulfide ison	
gi 15678829 ref NP 275946.1 thioredoxin [Nethand Start Start Start Start Start	2 日 1 6 5 15 15 10 10 10 10 10
gi[23060730]ref[ZP_00085617.1] COG0438: Glycosyltransference 34 0.77 gi[23121123]ref[ZP_00103523.1] COG0526: Thiol-disulfide iso 32 1.9 gi[34860800]ref[XP_215715.2] similar to Alcohol dehydrogena 32 3.5 1 gi[19684184]gb[AAH26035.1] C4orf9 protein [Homo sapiens] 32 3.5 1 gi[17933966]ref[NP_530756.1] glutamine amidotransferase [Ag 32 3.8	
S Click here to begin	-11: 12:
start 🔊 5 Netscape - 💷 Bioinfo3.1	🧷 🛛 🕄 🔍 😼 🍓 🧐 ⁄ 🎵 💭 10:27 AM

gi|2621990|gb|AAB85393.1|

conserved protein [Methanothermobacter thermautotrophicus str.

Delta H] Length = 77

Score = 124 bits (310), Expect = 5e-28 Identities = 77/77 (100%), Positives = 77/77 (100%)

Query: 1 MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60 MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL

Sbjct: 1 MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60

- Query: 61 KIMGRVASKEEIKKILS 77 KIMGRVASKEEIKKILS
- Sbjct: 61 KIMGRVASKEEIKKILS 77

>gi|23111526|ref|ZP 00097156.1| COG0526: Thiol-disulfide isomerase and thioredoxins [Desulfitobacterium hafniense] Length = 76 Score = 71.2 bits (173), Expect = 4e-12 Identities = 40/76 (52%), Positives = 57/76 (75%) Query: 2 MKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGELK 61 M I+I GTGCANC+ LE NA+EA+KELG+DA EK++++ I+ G+ P L V+ ++K Sbjct: 1 MVIKILGTGCANCKKLEANAKEAIKELGLDAVVEKVEDLQAIMAYGVMKTPALVVNEQVK 60 Query: 62 IMGRVASKEEIKKILS 77 +MG+V S EEIKK L+ Sbjct: 61 VMGKVLSAEEIKKYLN 76



BLAST Parameters

- Identities No. & % exact residue matches
- Positives No. and % similar & ID matches
- Gaps No. & % gaps introduced
- Score Summed HSP score (S)
- Bit Score a normalized score (S')
- Expect (E) Expected # of chance HSP aligns
- P Probability of getting a score > X
- T Minimum word or k-tuple score (Threshold)

Lecture 3.1



BLAST - Rules of Thumb

- Expect (E-value) is equal to the number of BLAST alignments with a given Score that are expected to be seen simply due to chance
- Don't trust a BLAST alignment with an Expect score > 0.01 (Grey zone is between 0.01 1)
- Expect and Score are related, but Expect contains more information. Note that %Identies is more useful than the bit Score
- Recall Doolittle's Curve (%ID vs. Length, next slide) %ID > 30 - numres/50
- If uncertain about a hit, perform a PSI-BLAST search



Doolittle's Curve

Evolutionary Distance VS Percent Sequence Identity





Statistics

 Probability of observing a score greater than S

 Probability of observing multiple MSPs greater than S $P=1-e^{-y}$

 $y = Kmne^{-\lambda S}$

 $1 - e^{-y} \sum_{i=0}^{c-1} \frac{y^i}{i!}$

Extreme Value Distribution



Scoring Matrices

- BLOSUM Matrices
 - -Developed by Henikoff & Henikoff (1992)
 - -BLOcks SUbstitution Matrix
 - -Derived from the BLOCKS database
- PAM Matrices
 - -Developed by Schwarz and Dayhoff (1978)
 - -Point Accepted Mutation
 - Derived from manual alignments of closely related proteins

Lecture 3.1



Conclusions

- BLAST is the most important program in bioinformatics (maybe all of biology)
- BLAST is based on sound statistical principles (key to its speed and sensitivity)
- A basic understanding of its principles is key for using/interpreting BLAST output
- Use NBLAST or MEGABLAST for DNA
- Use PSI-BLAST for protein searches

