









# **Defining the Terms**

- The quantitative measure: Similarity
  - Always based on an observable
  - Usually expressed as percent identity
  - Quantify changes that occur as two sequences diverge (substitutions, insertions, or deletions)
  - Identify residues crucial for maintaining a protein's structure or function
- High degrees of sequence similarity *might* imply
  - a common evolutionary history

• possible commonality in biological function

# **Defining the Terms**

### The conclusion: *Homology*

- Homology: Implies an evolutionary relationship
- *Homologs*: Genes that have arisen from a common ancestor
- Genes either *are* or *are not* homologous (not measured in degrees)

It is worth repeating here that homology, like pregnancy, is indivisible<sup>8</sup>. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Fitch, Trends Genet. 16: 227-231, 2000

### **Defining the Terms**

Orthologs: Genes that diverged as a result of a speciation event

- Sequences are direct descendants of a sequence in a common ancestor
- Most likely have similar domain and three-dimensional structure
- Usually retain same biological function over evolutionary time
- Can be used to predict gene function in novel genomes

# **Defining the Terms**

*Paralogs*: Genes that arose by the duplication of a single gene in a particular lineage

- Perhaps less likely to perform similar functions
- Can take on new functions over evolutionary time
- Provides insight into "evolutionary innovation"





# **Global Sequence Alignments**

- Sequence comparison along the entire length of the two sequences being aligned
- Best for highly-similar sequences of similar length
- As the degree of sequence similarity declines, global alignment methods tend to miss important biological relationships

# **Local Sequence Alignments**

- Sequence comparison intended to find the most similar regions in the two sequences being aligned ("paired subsequences")
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated for any two sequences being compared
- Best for sequences that share some similarity, or for sequences of different lengths

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Scoring Matrices: Construction and Proper Selection

# **Scoring Matrices**

- Empirical weighting scheme representing physicochemical and biological characteristics of nucleotides and amino acids
  - Side chain structure and chemistry
  - Side chain function
- Amino acid-based examples of considerations:
  - Cys/Pro important for structure and function
  - Trp has bulky side chain

• Lys/Arg have positively charged side chains

### **Scoring Matrices**

• *Conservation:* What residues can substitute for another residue and not adversely affect the function of the protein?

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

- Ile/Val both small and hydrophobic
- Ser/Thr both polar

- Conserve charge, size, hydrophobicity, additional physicochemical factors
- *Frequency:* How often does a particular residue occur amongst the entire constellation of proteins?

# <section-header><section-header><section-header><section-header><list-item><list-item><list-item><list-item>

Matrix Structure: Nucleotides						
<ul> <li>Simple match/mismatch scoring scheme:</li> <li>Match +2 Mismatch -3</li> </ul>	A T G C	A 2 -3 -3 -3	T -3 2 -3 -3	G -3 -3 2 -3	C -3 -3 -3 2	
<ul> <li>Assumes each nucleotide occurs 25% of the time</li> </ul>						
<b>XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX</b>		NATIO	NAL HUMAN	N GENOME Div	RESEARCH ision of Intron	INSTITUTE nural Research

NHGRI Current Topics in Genome Analysis Week 2: Biological Sequence Analysis I March 5, 2014 Andy Baxevanis, Ph.D.

Matrix Structure: Proteins					
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$					
$\begin{array}{cccccccccccccccccccccccccccccccccccc$					
BLOSUM62 NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intromurol Research					

# **BLOSUM Matrices**

- Look only for differences in conserved, ungapped regions of a protein family ("blocks")
- Directly calculated based on local alignments
  - Substitution probabilities (*conservation*)
  - Overall *frequency* of amino acids

- Sensitive to detecting structural or functional substitutions
- Generally perform better than PAM matrices for local similarity searches (*Henikoff and Henikoff, 1993*)
- BLOSUM series can be used to identify both closely and distantly related sequences



# BLOSUM *n*

- Clustering reduces contribution of closely related sequences (less bias towards substitutions that occur in the most closely related members of a family)
- Reducing *n* yields more distantly related sequences
- Increasing *n* yields more closely related sequences

Which one to	choose?	
BLOSUM		% Similarity
90	Short alignments, highly similar	70-90
80	Best for detecting known members of a protein family	50-60
62	Most effective in finding all potential similarities	30-40
30	Longer, weaker local alignments	< 30
XXXXXXXX		L HUMAN GENOME RESEARCH INSTITUTE Division of Intromural Research



# Gaps

- Used to improve alignments between two sequences
  - Compensate for insertions and deletions
  - As such, gaps represent biological events
- Must be kept to a reasonable number, to not reflect a biologically implausible scenario (~1 gap per 20 residues good rule-of-thumb)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

• Cannot be scored simply as a "match" or a "mismatch"

# **Affine Gap Penalty**

Fixed deduction for introducing a gap *plus* an additional deduction proportional to the length of the gap





### **BLAST**

- Seeks high-scoring segment pairs (HSPs)
  - Pair of sequences that can be aligned with one another
  - When aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - Score must be above score threshold *S*
  - Gapped or ungapped
- Results not limited to the "best" high-scoring segment pair for the two sequences being aligned

Altschul et al., J. Mol. Biol. 215: 403-410, 1990

	BLAST Algorithms					
	Program	Query Sequence	Target Sequence			
	BLASTN	Nucleotide	Nucleotide			
	BLASTP	Protein	Protein			
	BLASTX	Nucleotide, six-frame translation	Protein			
	TBLASTN	Protein	Nucleotide, six-frame translation			
	TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation			
)	$\overline{\mathbf{x}}$	000000000000000000000000000000000000000	NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intromurol Research			















achi sin aib ser			
	How To 🖻	http://ncbi.r	lm.nih.gov
SNCBI National Center for Biotechnology Information	All Detabases 1	nup.meor.i	Search
NCBI Home	Welcome to NCBI		Popular Resources
Resource List (A-Z)	The National Center for Biotechnology Informa	tion advances science and health by providing access to biomedical	PubMed
All Resources	and genomic information.		Bookshelf
Chemicals & Bioassays	About the NCB   Mission   Organization   Re	search   NCBI News	PubMed Central
Data & Software	Construction of the New York Construction		D.hMad Health
DNA & RNA	0.100.001		BLAST
Domains & Structures	Get Started		Nucleotide
Genes & Expression	<ul> <li>Tools: Analyze data using NCBI software</li> </ul>		Genome
Genetics & Medicine	<ul> <li>Downloads: Get NCBI data or software</li> <li>How-To's: Learn how to accomplish spe</li> </ul>	cife tasks at NCBI	SNP
Genomes & Maps	<ul> <li>Submissions: Submit data to GenBank</li> </ul>	or other NCBI databases	Gene
Homology			Protein
Literature			PubChem
Proteins	3D Structures	All and a second se	
Sequence Analysis	And an Annual sector of the		NCBI Announcements
Taxonomy	teini, DNA, and ENA molecules. Examine		Human genome annotation release 105
Training & Tutorials	molecular interactions, biological activities of		now available Feb 11, 2014
Variation			The human (Homo sapiens) genome
	▶ 1		NCBI releases Entrez Direct, the Entrez utilities on the UNIX command line Feb.5.2014 NCBI has just released Entrez Direct, a
			Sequence Mewer updated to version 3.1 Feb 4, 2014 NCBI Sequence Viewer provides a graphical view of sequences and
			More
You are here: NCBI > National	Center for Biolechnology Information		Write to the Help Der
GETTING STARTED	RESOURCES POPUL	AR FEATURED	NCBI INFORMATION
NCBI Help Menual	Data & Software Bookst	elf PubMed Health	Research at NCBI
	DNA & RNA Puchle	1 Central Ger Bank	NCRI News

	RLAST. Basic Local Alignment Search Tool	
d Marzichinin nin gov, fr		
BLAST <sup>®</sup> Home Recent	Basic Local Alignment Search Teol http://ncbi	i.nlm.nih.gov/BLAST
CBI BLAST Home	nne of similarity babusan Nialoniani sanuanaar	Your Recent Results New!
and a more regin	DELTA-BLAST a more sensitive protein-protein search	ID All Recent resulta
BLAST Assemi	bled RefSeq Genomes	NEWS
Choose a species g	enome to search, or list all genomic BLAST databases.	BLAST 2.2,29+ relased
= <u>Human</u>	© Oryza sativa © Gallus gallus	A new version of the stand-alone BLAST+ applications is available.
© Rat	Danio rerio     Microbes	Mon, 05 Jan 2014 12:00:00 EBT
Arabidopsis the	aliana © Drosophila melanogaster © Apis melilfera	D More DLAST nows
Basic BLAST		
Chocse a BLAST pr	rogram to run.	Tip of the Day
nucleotide blast	Search a nucleotide database using a nucleotide query Algorithms: blach, megablast, disconfiguous megablast	is monotone
protein blast	Search protein database using a protein quary Algorithms: blastp, psi-blast, phi-blast, delta-blast	
blastx	Search protein database using a translated nucleotide query	
<u>tblastn</u>	Search translated nucleotide database using a protein query	
<u>tblastx</u>	Search translated nucleotide database using a translated nucleotide quary	
Specialized BL	AST	
Choose a type of so	vedalized search (or database name in parentheses.)	
- 11-1	AL AL 1950	
<ul> <li>Make sp</li> <li>Search t</li> </ul>	race archives	
= Find con	served domains in your sequence (cds)	
Find seq	uences with similar conserved domain architecture (cdart)	
Search s	equences that have gene expression profiles (GEO)	
<ul> <li>Search is</li> <li>Research is</li> </ul>	mmunogroouens and 1 cen receptor sequences (IgBLAST)	
- Screen s	equence for <u>vector contentinettori</u> (vececien)	
<ul> <li>Search g</li> </ul>	protein or nucleotide targets in PubChem BioAssay	
· Search §	RA by experiment	
Constrair	nt Based Protein Multiple Alignment Tool	



C C	ein a 🛛 🔲	Protein BLAST: search protein dat	abases using a protein quary		
B Macrichi alta alta gov	* Barco TROCIMI-Barco PACE TYR - BarGarchill NK LOC-barcome			- (a) + c (B) -	9.4.6
BLAST <sup>®</sup> Home Rece	Basic Local Alignm	ent Search Tool			My NCBI Welcome basevanis. (Sign C
NCBI/ BLAST/ blastp	p suite	Standard Prot	tein BLAST		
Enter Query S	setz thissin thisetz Sequence BLAS	P programs search protein data	bases using a protein query. <u>more</u>	-	Reset cage Bookmark
Enter scoession P Research Second Control Cont	NUMBER (S), (JS), OF ANS IAL SEQUENCE), SALAR SELECTION (S) DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION DESCRIPTION OF A DESCRIPTION OF A DESCR	From To			
Or, upload file	Browse No file selected.				
Align two or m	Enter a descriptive title for your BLAST search @				
Choose Searc	ch Set				_
Database	Non-redundant protein sequences (nr) 2	<del>~                                    </del>	Available p	rotein databases ir	iclude:
Organism Optional Exclude Optional Entrez Query	Erter ingenser nime of 16-consistions will be suggested Enter organism common name, binomial, or tax Id. Only 20 top Models (XMXP) Uncultured/environmental sample	clude 💽 i taxa wil be shown. 🚇 i sequences	nr refseq	Non-redundant Reference Sequ	ences
Optional	Enter an Entrez query to limit search 🕼		swissprot	SWISS-PROT	
Program Sele	ection		pat	Patents	
Algorithm	blastp (protein-protein BLAST)     PB-8LAST (Position-Specific Iterated BLAST)     PH-8LAST (Pattern Hit Initiated BLAST)     OELTA-BLAST (Domain Enhanced Lookup Time Acc     Choose a BLAST algorithm @	elerated BLAST)	pdb env_nr	Protein Data Be Environmental	ank samples
BLAST	Search database Non-redundant protein sequences	(nr) using Blastp (protein-pr	otein BLAST)		
Algorithm parame	etera Note: Parameter values	hat differ from the default a	re highlighted in yellow and r	narked with + sign	
	BLAST is a regis	ered trademark of the National Lib	rary of Medicine.		
looyright   Disclaimer   Pe	ivacy   Accessibility   Contact   Send Realback				NCB( ) N.M ( NH ) DH

# NCBI RefSeq Database

- *Goal:* Provide a single reference sequence for each molecule of the central dogma (DNA, mRNA, and protein)
- Distinguishing features
  - Non-redundancy
  - Updates to reflect the current knowledge of sequence data and biology
  - Includes biological attributes of the gene, gene transcript, or protein
  - Encompasses a wide taxonomic range, with primary focus on mammalian and human species
  - Ongoing updates and curation (both automated and manual review), with review status indicated on each record

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Pruitt et al., Nucleic Acids Res. 42: D756-D763, 2014

RefSeq Accession	Number Prefixes
From curation of Ge.	nBank entries:
NT_ NM_ NP_ NR_	Genomic contigs mRNAs Proteins Non-coding transcripts
From genome annote	ation:
XM_ XP_	Model mRNA Model proteins
Complete list of moleculo http://ncbi.nlm.nih.gov/bo	e types in Chapter 18 of the NCBI Handbook poks/NBK21091 NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intromural Research

0.0		Protein BLAST: search protein databases using a protein quar	r	
Protein BLAST search protei	ne [+]			0.000
() (5 Marcricht alls with gov.	Star.cp?ffx0GMM+blaca&PAGE_TYfE+3karSearch&UNE_LOC+5kathome		G ≠ C [1]	9 4 6 6
BLAST <sup>®</sup> Home Recen	Basic Local Align It Results Saved Strategies Help			My NCBI Welcome baxevenia. (Bign O
NCBI/ BLAST/ blastp	suite	Standard Protein BLAST		
blastn blastp bla	stx thiastn thiastx			
Enter Query S	equence BLA	STP programs search protein databases using a protein qu	ry. <u>more</u>	Reset page Bookmark
Enter accession n	umber(s), gi(s), or FASTA sequence(s)	Crear Query subrange G		
YOLATY REGISTICS MERAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	PUPONYTANE GEENNINGETPAKANIE PTEINE VYLASIANE SLUTANEO REDAOATSIL POLINIKEAANETYNTÄÄLLI DELINETRELLOODINNE REDAOATSIL POLINIKEÄÄNEN KEINELLI DELINETRELLOODINNE REDAATSI SEISTEN VALTANEN KEINELLI DELINETRELOODIN LUARITELIKKI LOOKKEINELLOILI ATMADOODOOLOKKEILII KLINININEN	From		
Or, upload file	Browse No file selected.			
Job Title	Query sequence			
	Enter a descriptive title for your BLAST search 🙀			
C Align two or mo	ore sequences 😡			
Choose Searc	h Set			
Database	Non-redundant protein sequences (nr) 2		Limit by organism or	
Organism	Enter presentate name or id-consistions will be accounted	ixilda 🗉 🔶	Limii by organism or	
Optional	Enter organism common name, binomial, or tax id. Only 20 t	op taxa will be shown. @	taxonomic group	
Exclude Optional	Models (XM/XP) Uncultured/environmental samp	Ne sequences	0 1	
Entrez Query Optional	Enter an Entrez query to limit search @			
Program Selec	tion			
Algorithm	Bibastp (protein-protein BLAST)     PH-BLAST (Position-Specific Iterated BLAST)     PHI-BLAST (Param Hit Initiated BLAST)     DELTA-BLAST (Domain Enhanced Lookup Time Ar Choose a BLAST algorithm      ⊕	coelerated BLAST)		
BLAST	Search database Non-redundant protein sequence	s (nr) using Blastp (protein-protein BLAST)		
Algorithm parame	Note: Parameter values	that differ from the default are highlighted in yello	w and marked with + sign	
	BLAST is a reg	eared addemark of the National Library of Medicine.		

Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign     General Parameters     Values	
Ver of the second	
Program Selection         Algorithm	9 4 6 0-
Algorithm	
PSI-BLAST (Position-Specific Iterated BLAST)     PH-BLAST (Pastion-Specific Iterated BLAST)     PH-BLAST (Pastion-Specific Iterated BLAST)     DELTALAST (Pastion-Specific Iterated BLAST)     CELTALAST (Pastion-Reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Coorse a BLAST agontion @      BLAST     Search databases Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Coorse a BLAST agontion @      BLAST     Search databases Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Coorse a BLAST agontion @      BLAST     Search databases Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search databases Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search databases Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search databases Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-reduced and protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-sequences (nr) using Blastp (protein-protein-BLAST)     Search databas	
OH-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)     OBELIA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)     Obcover a BLAST agentime @      BLAST     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences (nr) using Blastp (protein-protein BLAST)     Search database Non-redundent protein sequences to database Non-redundent protein blast differ from the default are highlighted in yellow and merked with + sign     General Paramoters     Max target     * 250 = 3     Botot the maximum number of aligned sequences to database Non-redundent of aligned sequences to database Non-redundent of aligned sequences to database Non-redundent of aligned sequences (Protein-protein BLAST)     Search database Non-redundent of aligned sequences to database Non-redundent of aligned sequences (Protein-protein BLAST)     Search database Non-reducted Sec.extended Non-reducted Non-reducted Non-reducted Sec.extended Non-reducted	
DELTA-BLAST (gondan Enhanced Lookup Time Accelerated BLAST) Choose a BLAST algorithm @      DELTA-BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)     default are highlighted in yellow and marked with + sign     General Parameters     Max target     + 230 2     Boot the maximum number of aligned sequences to disjon @     Default = 100      Short queries     Max target     + 230 2     Boot the maximum number of aligned sequences (or input sequences (o	
Choose a BLAST agention agent for the sequences (nr) using Blastp (protein-protein BLAST)  BLAST Second database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)  Concert Parameter  Concert Parame	
BLAST       Search database Non-redundant protein sequences (n) using Blastp (protein-protein BLAST)         Characteristic is a new stocker       Stoor results is a new stocker         Concruit Paramolers       Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign General Paramolers         Max target       (250 ±) Botect for maxmum number of aligned sequences to display (a)         Short queries       Max database Non-redundant protein sequences (b)         Word size       1 (a)         Watch size       1 (a)         Max matches in a       0 (a)	
Show results is a new statew     Source of the answering is a new stateweight in a new stateweight is a new stateweight in a new stateweight in a new stateweight is a new stateweight in a new stateweight in a new stateweight is a new stateweight in a new	
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign     General Parameters     Section to examine an unmore of algoed sequences to deplay      Short queries     Macromatically adjust parameters for short input sequences      Expect threshold     O     More size     X = 0     Max matches in a     O	
General Parameters       Max target       \$250     2       Short quarties     Matomatically adjust parameters for short input sequences @       Expect threshold     0       Word size     \$ 2       Max matches in a     0	
Max target sequences     • 250 _ 2 Bioloci the maximum number of aligned sequences to display @     Default = 100       Shert queries     Ø Automatically adjust parameters for short input sequences @       Expect threshold     @       Word size     • _ @       Max matches in a     0	
Interpretation     Balact the maximum number of aligned sequences to diskipy @     Despective     100       Short queries     M Automatically adjust parameters for short input sequences @       Expect threshold     10     @       Word size     10     @       Max matches in a     0     @	
Short queries          Modulat parameters for short input sequences @        Expect threshold     10       Word size     3       Max matches in a     0	
Expect threshold 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
Word size 3 2 00 Max matches in a 0 00	
Max matches in a	
query range	
Scoring Parameters	
Matrix (8.050M62 :)	
Gap Costs Existence 11 Exercise: 1 = 1	
Compositional Conditional compositional score matrix adjustment 5) @	
Filters and Masking	
Filter + Ef Low complexity regions @	
Mask Dispersion ratio only de	
Mask lower case letters @	
BLAST Search database Non-redundant protein sequences (nr) uting Blastp (protein-protein BLAST)	

10.0	Protein #LAST	T: search protein databases using a protein query	
Protein BLAST search protein d			
() Bilastrachi nim nih gov, Sta	In coPRECEAM-Macca@PACE_TYR -RainSearch&UNE_LOC-Statitione	( <u>i</u> + c)( <u>i</u> -	9 4 1 12-
1.1.1	Crime an crimer crimination intersearch rate		
Program Selectio	on .		
Algorithm	Childrede (anabele anabele D/ ADT)		
	State (Street in Street)     State (Street in Street)     State (Street in Street)		
	O PHI-BLAST (Pattern Hit Initiated BLAST)		
	O DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAS	ST)	
	Choose a BLAST algorithm @		
( NIACT )	Percela de la barra Non codo adoral ante la comunación de la cina Oli	and Anatola anatola BLAST	1
BLAST	Search database Non-redundant protein sequences (nr) using dia Showresults in a new window	Istb (brown-broken BEAG)	
	n Note: December values that differ fro	on the default are highlighted in unline and marked with a size	
General Paran	nelers	In the detault are regregated at yearder and marked with 7 sign	
Max tarnet			
sequences	* 250		
	and the maximum memory of any not and reasons to extrate the		
Short queries	Automatically adjust parameters for short input sequences @		
Expect threshold	10 😡		
Word size	(1)0	<i>E-value threshold</i>	
		<b>Reports all hits with <math>E &lt; 10</math></b>	
Max matches in a query range	0 9		
Seoring Darse	notore		
aconny Paran	REATS		
Matrix	BLOSUM62 : W		
Costs	Existence 11 Extension 1 1 1		
Compositional	Conditional compositional score matrix adjustment 🔹 😡		
adjustments			
Filters and Mar	sking		
Filtor	+ E Low complexity regions 🚇		
Mask	Mask for lookup table only @		
	Mask lower case letters		
DIACT	Search database Non-redundant protein sequences (ed. using (	Blaste (protein protein BLAST)	
BLAST	Show results in a new window	meeth (hi stambiotenii Orbol)	
			Y

000	Protein BLAST: search ;	protein databases using a protein query	
🗧 Protein BLAST: search protein d	*		
(4) B Mascricht nin mit gov, Sta	et.cp/TROCIMM-blacca&PAGE.TYTE-BlackeetsALINE.LOC-blacknone	(2 ≠ c) [B].	9 4 6 0-
	Criter an Criter Construction sea on the		
Program Selection	on		
Algorithm	Biblish (protein-protein BLAST)     PRI-BLAST (Position-Specific Iterated BLAST)     PHI-BLAST (Pattern Hit Initiated BLAST)     DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)     Choose a BLAST algorithm @		
BLAST	Search database Non-redundant protein sequences (nr) using Blastp (pr of Showresults in a new wholew	rotein-protein BLAST)	
General Param	ra Note: Parameter values that differ from the d motors	sefault are highlighted in yellow and marked with + sign	
Max target sequences	♦ (250 2) Solect the maximum number of aligned sequences to display @		
Short queries	Automatically adjust parameters for short input sequences @		
Expect threshold	10 0		
Wardalas	(200)	PAM30	
HOID SIZE	( <u>3</u> ,)	PAM70	
Max matches in a query range	0 9		
Seoring Darag	notore	BLOSUM80	
Matrix		BLOSUM62	
	BLUSUNDZ + UP	DLOGUN 45	
Gap Costs	Dillarge 11 Dranslow 1 1 1 1	BLOSUM45	
Company Wangel		BLOSUM50	
adjustments	Conditional compositional score matrix adjustment	PLOSUMOO	
Filters and Ma	isking	BL030M90	
Filtor	+ ≝ Low complexity regions @		
Mask	☐ Mask for tookup table only @ ☐ Mask lower case letters @		
BLAST	Search database Non-redundant protein sequences (nr) using Blastp i Show results in a new window	(protein-protein BLAST)	
			4

900	Protein BLAST: search protein	fatabases vsing a protein query	
🗧 Protein BLAST: search protein d	+		
() 🗄 Mastricht nim wit gov, Ste	er op/TROCIAM-bland&MCE.TYR-Bardeath&URE.LOC-bastnone	G * C (N-	9 🚯 🏚 🔟 -
	Crise ar chief don't o mini sea or no		
Program Selecti	on		
Algorithm	Bidaste (protein-protein BLAST)     ORIELAST (Peation-Specific Iterated BLAST)     OPHIGLAST (Peation-Kit Initiated BLAST)     DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)     Choose a BLAST algorithm @		
BLAST	Search database Non-redundant protein sequences (nr) using Blastp (protein- S Show results is a new window	protein BLAST)	
General Parameter	ITS Note: Parameter values that differ from the default motors	are highlighted in yellow and marked with + sign	
Max target sequences	♦ 250 ± Solect the maximum number of aligned sequences to display ₩		
Short queries	Automatically adjust parameters for short input sequences		
Expect threshold	10 😡		
Word size	( <u>3-1</u> ) @		
Max matches in a query range	0		
Scoring Paran	neters		
Metrix	( BLOSUM62		
Gap Costs	Existence 11 Extension: 1 1 1		
Compositional adjustments	Conditional compositional score matrix adjustment		
Filters and Ma	isking		
Filter	+ S Low complexity regions @		
Mask	☐ Mask for lookup table only @ ☐ Mask lower case letters @		
BLAST	Search database Non-redundant protein sequences (nr) using Blastp (prote	n-protein BLAST)	

# **Low-Complexity Regions**

- Defined as regions of "biased composition"
  - Homopolymeric runs
  - Short-period repeats
  - Subtle over-representation of several residues
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

• Often lead to false positives

• Filtering is advised (but *not* enabled by default)

000	Protein RAST: search protein databases using a protein	a dravA	
(d) et blan schi alta nët pov the	n cu/WiOAAH-SancaAMCZ TYR - European AB, NK UC-Santhome	Sec M.	9.5.0 12-
	Criser an Crisez Query to test search tag		
Program Selection	n		
Algorithm	blastip (protein-protein BLAST)     OP8ELAST (Peaklon-Speaklic Iterated BLAST)     OPHELAST (Paaklon-Speaklic Iterated BLAST)     DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)     Croces a BLAST algorithm		
BLAST	Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST) Stow results is a new vectore		
Algorithm parameter General Parameter	Note: Parameter values that differ from the default are highlighted in neters	yellow and marked with + sign	
Max target sequences Short queries Expect threshold Word size Max matches in a query range			
Scoring Paran	neters		
Matrix	8.0500452 ;) 🖗		
Gap Costs	Existence 11 Extension 1 2 @		
Compositional adjustments	Conditional compositional score matrix adjustment 💷 😡		
Filters and Ma	sking		
Filtor	+ S Low complexity regions @		
Mask	Mask for lookup table only @     Mask lower case letters @		
BLAST	Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)		

000	NCBI Blast-Query s	equence	2
🗧 Protein BLAST: search procein d . 🛛 🗧 NCBI BlastCluery sequence	× (+)		
(d) @ blastacbi.nim.mik.gov/Blast.cgi		습 = C (BF	9 🖶 🏦 🛄-
⊊_ BLAST <sup>®</sup>			My NCBI E2
Home Recent Results Saved Strategies H	elp		Welcome baxevanis. [Sign Out]
NCBI/ BLAST/ blastp suite/ Formatting Results - G4Y4D4	ZS014		
Edit and Resubmit Save Search Strategies P Fo	matting options > Download	Yes How to read this page	Blast report description
Query sequence			
RID G4Y40425014 (Expires on 02-1 Query ID kcij56462 Description Query sequence Molecule Sype amino odd Query Length 1403	9 04:01 am)	Database Name nr Decorption Al non-redundant GenBank CDS translations+PD8+Swir acululing environmental samples from WGS projects Program BLASTP 2.2.29+ P Citation	ssProt+PIR+PRF
Other reports: ▷Search Summary [Taxonomy re;	orts] [Distance tree of results] [Related Structures] [New DELTA-BLAST, a more sensiti	kultiple alignment] ve protein-protein search (a)	
Show Conserved Domains			
Query seq. Specific hito Superfamilies	, , , , , , , , , , , , , , , , , , ,	The state st	
	Distribution of 225 Blast Hits on th	he Query Sequence @	
	Mouse over to see the define, dick to show align	menis	
	Color key for al	lignment scores	
	Query		



ncti nin win gan, Taur, cui	· + c	81-		_		9
scriptions		Г		0	0.m	2020
Sequences producing significant alignments:				0.	10	1000
Select All None Selected:0				<	10-	1000
I Alignments Download - GenPapt Graphics Distance trie of results Multiple alignment					_	0
Description	Max score	Total score	Quary cover	E value	Ident	Accession
prospero, soform M [Drosophila melanogaster] rsp(P28617.3]PROS_DROME RecName: Full-Homeobox protein prospero repl/AFH06384.1] prospero, is	994	1938	93%	0.0	100%	NP 001247046.1
prospero, isoform J [Drosophila melanogastar] >emb(CAA77602.1] prospero [Drosophila melanogastar] >gb(AAF64628.4] prospero, isoform J [Drosophila	993	1936	93%	0.0	100%	NP 524317.4
prospero (Drosophila melanopaster)	993	1932	93%	0.0	100%	BAA01464 1
homeodomain transcription factor Prospero (Drospohila melanopaster)	990	1821	93%	0.0	100%	AAF05703.1
GG18085 [Drosophia erecta] >pbiEDV/4931.11 GG18089 [Drosophia erecta]	989	1885	93%	0.0	99%	XP_001580573.1
Prosistation (Drosophila melenopaster)	982	1811	93%	0.0	97%	AAA28541.1
GE28090 [Drosophila yakuba] >gb)EDW98913.1] GE28090 [Drosophila yakuba]	981	1885	93%	0.0	97%	XP_002097201.1
prospero, isoform H [Drosophila melanogaster] >pbiAFH06362.1] prospero, isoform H [Drosophila melanogaster]	944	1862	93%	0.0	100%	NP 001247044.1
prospero, isoform L. [Drosophila melanogoster] >gb]AAT94492.1] LD37627p [Grosophila melanogoster] >gb[AAN13600.4] prospero, isoform L. [Drosophila	943	1858	93%	0.0	100%	NP 788636.3
prospero, isoform i [Drosophila melanogaster] -gbiAFH06363.1[ prospero, isoform i [Drosophila melanogaster]	942	1864	93%	0.0	100%	NP 001247045 1
prospero, isoform K [Drosophila melanogester] >gb(AAN13501.4] prospero, isoform K [Drosophila melanogester]	942	1863	93%	0.0	100%	NP 731565.4
GM23939 [Drosophia sechelia] >gb(EDW42617,1) GM23939 [Drosophia sechelia]	935	1987	93%	0.0	98%	XP 002031631.1
GE 10657 (Droacohila ananasael >gt:(EDV42775.1) GE 10657 (Droscohila ananasaan)	904	1673	93%	0.0	82%	XP_001954214.1
GK11290 [Drosophila willston] >gb[EDW80945.1] GK11250 [Drosophila willston]	845	1532	85%	0.0	83%	XP 002069959.1
GA 14403, isoform B [Drosophile pseudoobscura] pseudoobscura] -pg/(EAL29137.3); GA 14403, isoform B [Drosophile pseudoobscura] pseudoobscura]	845	1476	90%	0.0	81%	XP_001359965.3
pros (Drosophia virila) > gb(EDW66804.1) pros (Drosophia virila)	821	1429	85%	0.0	79%	XP_002063284.1
GH21437 [Drosophila grimshaw] >gb(EDV95086.1] GH21437 [Drosophila grimshawi]	809	1374	84%	0.0	80%	XP 001994360 1
GI22695 (Drosophila moisvensia) >sb/EDW15591.1] GI22695 (Drosophila moisvensia)	804	1392	84%	0.0	80%	XP_002000130.1
PREDICTED: homoobox protein prospero-like isoform X2 (Constitis capitata)	691	11	30.	-17	9	$= 3 \times 10^{-1}$
GA14403, isoform C IDrosophile pseudoobscural pseudoobscural >cb/EM52398.11 GA14403, isoform C IDrosophile pseudoobscural pseudoobscural	611	14	52-	± /	<i>.</i>	= 5 ×10
homecbex protein prospero/prox-1 [Aedea aegys6] >gb/EAT46002.1[ AAEL.002769-PA, partial [Aedea aegys6]	571	770	62%	30-179	59%	XP_001655942.1
PRED/CTED: homeobox protein prospero-ike isoform X1 (Cerattis capitata)	491	1093	84%	5e-144	62%	XP 004529242 1
Rocheme: Full=Homeobox protein prospere >eb(AAF06880.1)AF190405_1 homeodomain transcription factor Prospere [Drosophila virilis]	440	1309	75%	30-126	80%	Q9U6A1.1
prospero [Drosophia simulans] >gb/EDX13377.1] prospero [Drosophia simulans]	378	606	28%	3e-112	100%	XP_002103874.1
PREDICTED: LOW QUALITY PROTEIN: homeobox protein prospera-like [Musca domestica]	397	1080	72%	1e-110	82%	XP_005188256.1
Chain A, Structural Basis Of Prospero-Dna Interaction; Implications For Transcription Regulation In Developing Cells	347	347	11%	20-107	100%	1XPX_A

4

thin with good Theoremail	· · c	81-				9
prospero homeotox protein 2 [Rattus nervegicus]	214	214	11%	38-55	59%	NP 001178700 1
PREDICTED, prospero homeolos: protein 2.8 osodorita africana)	214	214	10%	3e-55	60%	XP_003408594.1
PREDICTED: prospero homeobox protein 2 (Tursiops truncatus)	214	214	11%	40-55	58%	XP_004324368.1
<ul> <li>PREDICTED: prospero homeobox protein 2 (Orcinus orca)</li> </ul>	214	214	11%	48-55	58%	XP 004262290 1
PREDICTED: prospero homeobox protein 2-like [Monode/phia domentica]	214	214	10%	46-55	62%	XP 001376160.2
PREDICTED: prospero homeobox protein 2 (Aligator mississippiensis)	214	214	11%	60-55	57%	XP_006268418.1
PREDICTED: prospero homeolos profein 2.[Vougne pecos]	213	213	11%	6e-55	58%	XP 006206136.1
PREDICTED: prospero homeobox protein 2 (Sonox ananeus)	213	213	12%	60-55	54%	XP_004612447.1
PREDICTED: prospero homeobox protein 2 (Ceratotherium simum)	213	213	11%	6e-55	59%	XP 004426417.1
🗇 transcription factor protein JCiona intestinalia	210	210	11%	80-55	58%	BAE06658.1
prospero homeobox 2 [Danio reno] >gb/ABS30417.2[ prox2 protein [Danio reno]	212	212	10%	99-55	59%	NP_001121764.2
PREDICTED: prospero homeobox protein 2 (Aligator silvensis)	214	214	11%	96-55	57%	XP. 006015400.1
PREDICTED: prospero homeobox protein 2 isoform X1 [Equus cabalus]	213	213	10%	10-54	50%	XP 005605386.1
PREDICTED: prospero homeoloo protein 1-like [Tek/Ligu rubripes]	214	214	10%	1e-54	¢ ,	Accont
PREDICTED: prospero homeobox protein 1-like [Oreesthomia nilotous]	214	214	10%	10-54		necepi
PREDICTED: prospero homeobox protein 2 (Leptonychotes weddelli)	212	212	11%	1e-54	: (	for now)
PREDICTED: LOW QUALITY PROTEIN: prospero homeobox protein 2 (Califithis Jacobus)	211	211	11%	2e-54	57%	XP_003734094.1
PREDICTED: prospero homeobox protein 1-like isoform X1 (Pundamilia riverarel) het(XP_005742789.1) PREDICTED: prospero homeobox protein 1-like	213	213	10%	29-54	59%	XP_005742768.1
unverred protein product [Tetrandon nigroviridia]	198	198	10%	2e-54	59%	CA004605.1
PREDICTED: prospero homeobox protein 1-lika isoform X1 [Hoploshromia burton]; het/XP_005918895.1] PREDICTED: prospero homeobox protein 1-lika	213	213	10%	20-54	59%	XP_005016804.1
PREDICTED: prospero homeobox protein 1-like [Neolamprologue brichend]	213	213	10%	2e-54	59%	XP_006786125.1
PREDICTED: prospero homesica protein 1-like [May/andia.zabre]	213	213	10%	20-54	59%	XP_004540774.1
PREDICTED: prospero homeobox protein 2 [Cavia porcelus]	211	211	11%	28-54	59%	XP_003472362 1
PREDICTED: prospero homeobox protein 2 (Octodon degua)	211	211	11%	28-54	57%	XP 004625022.1
PRED/CTED; prospero homeobox protein 2 (Carnelus ferus)	211	211	11%	30-54	57%	XP 006184539.1
PREDICTED: prospero homeobox protein 2 isoform 2 [Gorilla gorilla]	210	210	10%	7e-54	58%	XP_004055501.1
PREDICTED: LOW QUALITY PROTEIN: prospero homeobox protein 2 (Ocholone princepa)	212	212	13%	80-54	51%	XP_004584644.1
PREDICTED: prospero homeobox protein 2 isoform 2 (Pan troglodytes)	209	209	10%	9e-54	58%	XP 522907.2
PRED-CTED: uncharacterized protein LDC779983 (Gone Intestinelia)	214	214	11%	16-53	58%	XP_002122508.1
Reject above desired threshold	(F	<	10-	3)		
Reject above destred intestiona	(12		10	/		

000		NCBI Blast:Query sequence	
S Protein BLAST: Sear	ch protein d 🚬 🗧	NCBI BlastQuery sequence × +	
( ) & blattechista	prospero, is Sequence ID:	soform L [Drosophila melanogaster] : <u>cffINP 786536.3</u> Length: 1374 Number of Matches: 2	$\geq$ 25% for proteins $\geq$ 70% for nucleotides
	P See 2 mor	ne uue(s)	Related Information
	Range 1: 17 t	to 704 GenPept Graphics Vext Match - Fr	revious Match Gene - associated gene details
	Score 943 bits(243	Expect         Method         Identities         Positives         G           137)         0.0         Compositional matrix adjust.         688/688(100%)         688/688(100%)         688/688(100%)	Apps UniGene - clustered expressed
	Query 17	LPOPOSVSTAnssssnnnssTPAALATHsptsnspvsgassasslltaaPGNLPGGSSA 7 LPOPOSVSTANSSSSNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNLPGGSSA	76 Map Viewer - aligned genomic
	Sbjct 17	LFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNLFGGSSA	76 context
	Query 77 Sbjct 77	KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN 1	136 identical to the subject
	Query 137	NSIAPANSTPHSNOTNasispgsahssshshqqvspKGSRRVSACSDRSLEANAADVAGG NSIAPANSTPHSNGTNASISPGSAHSSSHSHQVSPKGSRRVSACSDRSLEANAADVAGG	- Gap
	Sbjet 137	7 NSIAPANSTPMSNGTNASISPGSABSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG 1	Job Low complexity
	Query 197	SPPRASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMqqqq 2 SPPRASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQ	a Low complexity
	50jet 197	SPPARSVSDDSUGADSUEQHQQQQQADLYABBALAATLQAKELAQLQQLATAAQQQQ 2	
	Sbjct 257	qq1qekeq1BSklnnnnnnaatannnnttkEsINLIDSEMADIKIKSEPQTAPQPQ QQLQEKEQLBSKLNNNNNNAATANNNNTTMESINLIDSEMADIKIKSEPQTAPQPQ QQLQEKEQLBSKLNNNNNNIAATANNNNTTMESINLIDSEMADIKIKSEPQTAPQPQ 3	
	Query 317	QsphqsshssrsqsqsgshssmasdqslrrkssdsldsHGaqddaqdeedaaPTGQRSES 3 QSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDDAQDEEDAAPTGQRSFS	376
	Sbjet 317	7 QSPHGSSHSSRSGSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDDAQDEEDAAPTGODSES 3	376
	Query 377	7 RAPEEPOLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSdmmlldkddvldedddd 4 RAPEEPOLPTKKESVDDMLDEVELGLHSRGSDMDSLASPSHSDMMLDXDDVLDEDDDD 9 SPEFPOLPTKKESVDDMLDEVELGLHSRGSDMDSLASPSHSDMMLDXDDVLDEDDDDD	136
	Oueru 437		-30 
	Chick 437	DCVEQKTSGSGCLKKPGMDLKRARVERTVSGRRCSPSSGLAQAGQLQVNGCKKRKLYQPQ	196
	abjec 437		
	Query 497	QHAMERYVAAAAGLNFGLNLQSMALDQEDSESNELESVQIQQKRVEKNALKSQLRSNQEQ QHAMERYVAAAAGLNFGLNLQSMALDQEDSESNELESVQIQQKRVEKNALKSQLRSNQEQ	
	abjet 497	QRAMER I VARAMULNIGLNIGSMILDQEDSESNELESPQIQQKNVEKNALKSQLRSNQEQ     5	120
	Query 557	LARENQUATVOLCSREQESECGEIGGGGGVEGEGEPDNGSSDHIELSPSPTLTGGGDVSP LAENQQKYVQLCSREGESECGELDQDQDVEGEGEPDNGSSDHIELSPSPTLTGGGDVSP	210
	abjec 557	LARAQQAIVQLCSAREQESECQELDQDQDVEQEQEPDXGSSDHIELSPSPTLTGDGDVSP     6	20
	Sbjct 617	nn.klivqlergissspspspl.kpkrs.ig.s5DSGANNLSQNMSKMNSGKLHNPLVGVGHP NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANNLSQNMSKMNSGKLHNPLVGVGHP NHKEETGQERPGSSSPSPLKPKTSLGESSDSGANNLSQNMSKMNSGKLHNPLVGVGHP	576
	Query 677	ALPQGFPPLLQHMGDNSHAAANYQQFFF 704	
	Sbjct 677	ALPQGFPPLLQHMGDNSHAAANYQQFFF ALPQGFPPLLQHMGDNSHAAANYQQFFF 704	







## **BLAST 2 Sequences**

- Finds local alignments between two protein or nucleotide sequences of interest
- All BLAST programs available
- Select BLOSUM and PAM matrices available for protein comparisons
- Same affine gap costs (adjustable)
- Input sequences can be masked

900	RLAST: Basic Local Alignment Search Tool	
S BLAST: Basic Local Alignment S		
() Bilascricht nim nih gov, R		
BLAST <sup>®</sup> Home Recent	Results Boved Strategies Help	m.nin.gov/BLASI
NCBI BLAST Home     BLAST finds reak	ons of similarity between biological sequences. more	Your Recent Results Newl
	DELTA-BLAST, a more sensitive protein-protein search	III. All Record results
BLAST Assem	bled RefSeq Genomes	News
Choose a species g	enome to search, or list all genomic BLAST databases.	BLAST 2.2.29+ released
= <u>Human</u> = <u>Mouae</u> = <u>Rat</u> = <u>Arabidopsis th</u>	Oryza sativa      Odliss pailus     Odliss pailus     Oddiss pailus     Oddiss pailus     Oddiss     Oddis     Oddiss     Oddiss     Oddiss     Oddiss     Oddis     Oddis	A new venion of the stand-store BLAST- applications is available. Mon. 05 Jan 2014 12:00:00 EST
Basic BLAST		
Chocse a BLAST p	rogram to run.	Tip of the Day
nucleotide blast	Search a nucleotide database using a nucleotide query Algorithms: blastn, mogablast, discontiguous mogablast	
protein blast	Search protein database using a protein quary Algorithms: blast, psi-blast, phi-blast, delta-blast	
blastx	Search protein database using a translated nucleotide query	
tblastn	Search translated nucleotide database using a protein query	
<u>tblastx</u>	Search translated nucleotide database using a translated nucleotide query	
Specialized BL	AST	
Choose a type of sp	ecialized search (or database name in parentheses.)	
<ul> <li>Make sp</li> <li>Search t</li> <li>Find cor</li> <li>Find seq</li> <li>Search t</li> <li>Search t</li> </ul>	eclic primers with <u>Primer-BLAST</u> rate as tohins served domains in your sequence (cds) percent with aniMar conserved domain architecture (cdart) express that have gene as parasets on profiles (CBC) mmunoglobulitins and T cell receptor sequences (igBLAST)	
<ul> <li>Screen s</li> <li><u>Align</u> tw</li> <li>Search g</li> <li>Search g</li> <li>Constrait</li> </ul>	or more sequences using BLAST (2020c)  Votelini or nucleotide targets in PubChem BicAssay BRA by experiment Based Process Multiple Alignment Tool	

O O Pristein BLAST Align two or mer.		Proteir	BLAST Align	two or more sequences using BLAST		
) 🖲 blacechi nin nin gov, Sta	n.co/PROCIAM-Necol/PACE_TYPE-Reidershill	AST SPEC-blant2xxxxXXXXX.LOC-bla	middLAST.PAG	C+ElemAdit.AST_NIT+Sem25ep4Q423t+>>>P,006872.150: ☆ ± C 【日・		9.4.6.10
BLAST <sup>®</sup> Home Recen	nt Resulta Saved Strategies	Basic Local Alignm Help			My NCBI Welcome be	xevanis. <u>(Sign O</u> u
NCBV BLAST/ blastp	suite	Alig	n Seque	nces Protein BLAST		
blastn blastp bla	stx tblastn tblastx				Queret enne	Backmark
Enter Query S	equence	BLAS IP programs	search prot	sein subjects using a protein query. more	00000.49928	MONTH SET
Enter accession n >NP_008872.1 SOX- MAEEQOLSEVELSPVGSI VCTREAVSOVLSGTOWTI GELWALLNESDKRPPTEI KSAHLDEREPGSOSPHSI	umber(s), gl(s), or FASTA seque 10 (Вово варіеля) верксіля озанісця рокососодо валя пор даля и компански порактика порактика валя компански порактика порактика одаля валя компански порактика одаля валя порактика по по по по по по по по по по по по по	Ince(s)	Clear	Ouery subrange 😡 From To		
Or, upload file Job Title	Browse No file selected.	θ				
Align two or mo	Enter a descriptive tibe for your Bi ore sequences 😡	AST search 👹				
Enter Subject 3 Enter accession n NP 003331.1 sex 4 957ASAMLSVINSCOM WORDORRHALINFERR	Sequence sumber(s), gi(s), or FASTA seque determining region Y   Bore sapir swymin Harksserc-reachistroctr HBIERGLOYOWORLTBARMOPTCEACHO EVOLUMELYBECCTEATHSSMEMCLORLPPE	INCE(S)	Clear	Subject subrange 😡 From To		
Or, upload file	Browse No file selected.					
Program Selec	ction					
Algorithm	blastp (protein-protein BLA: Choose a BLAST sigorithm @	ST)				
BLAST	Search protein sequence us	ng Blastp (protein-prote	in BLAST	)		
Algorithm parame	ators					
opyright   Disclaimer   Pri	vacy   Accessibility   Contact   Send fee	BLAST is a regis Stack	lered tradem	hark of the National Library of Medicine.	NCBI	

9 () () 12 Printelin BLAST: Align Two or mor	Protein BLAST: A	lign two or more sequences using ILAST	
()) () Hert of a sin of gov the region com	egenerative seasoned from an development of the seasonable of the	UT, MAG-biometer AGT, Jult-SeendberskippONt-SAF, Older 12, 1 🗇 * C 📗 🖬 *	9 (\$) @ D-
BLAST	Search protein sequence using Blastp (protein-protein BLA 영 Show results in a new window	ST)	
Algorithm paramete	rs Note: Parameter values that diff	fer from the default are highlighted in yellow and marked with $st$ sign	
General Paran	neters		
Max target sequences	Select the maximum number of aligned sequences to display @		
Short queries	Automatically adjust parameters for short input sequences		I
Expect threshold	10		
Word size	3.: 0		
Max matches in a query range	0		
Scoring Paran	neters		
Matrix	BLOSUM62 1 🚱 🔶	PAM30 PAM70	
Gap Costs	Existence: 11 Extension: 1 💠 😡	PAM250	
Compositional adjustments	Conditional compositional score matrix adjustment	BLOSUM80	
Filters and Ma	sking	BLOSUM62	
Filter	+ S Low complexity regions 😥	BLOSUM45	
Mask	Mask for lookup table only      Mask lower case letters	BLOSUM50 BLOSUM90	
BLAST	Search protein sequence using Blastp (protein-protein B	LAST)	
Copyright   Disclatimer   Privac	BLAST is a registered tre y   Accessibility   Contact   Send feedback	demark of the National Library of Medicine.	NCBI ( NLM   NH   DHHS

			NC81 8last:NP_008872.1 SO	X-10 Biomo sapiens] (466				
NCN NaseN*,008872.1 SOR-1.	+			and the second second			_	
) 🗄 Martinchi nin min gov, Martin	a)				1940 C	N-		9 4 4
NCBI/ BLAST/ blastp st	uite-2sequences/ Formatting	Results - G4U6505P11	4					
Edit and Resubmit	Save Search Strategies	Formatting options.	▷ Download		You	How to read this	s page Bla	est report description
			Blast 2 se	quences				
NP_008872.1 SO	X-10 [Homo sapiens]	(466						
RID	G4U6505P114 (Expires or	02-19 02:53 am)						
Query ID	ld]19239			Subject ID	Icl 19241			
Description	NP_008872.1 SOX-10 [Ho	omo sapiens]		Description	NP_003131.1 sex d	etermining region	n Y [Homo s	apiens]
Molecule type	amino acid			Molecule type	amino acid			
Query Length	400			Program	BLASTP 2.2.29+ >0	Citation		
				1.1				
Other reports: >	Search Summary [Taxonor	my reports] [Multiple al	ignment]					
Graphic Sum	mary							
		Distrib	ution of 2 Blast Hits or	the Query Sequence	9			
		Mouse over to see	e the defline, click to show a	signments				
			Color key fo	r alignment scores				
		<40	Color key fo 40-50	r alignment scores	:00 >=200			
		<40 Duery	Color key fo 40-50	r alignment scores	:00 >=200			
		<40 Avery 1 9	Color key fo 40-50 0 180	r alignment scores	1 00 ≻=200 360 45			
		-40 Duery 1 9	Color key fo 40-50 0 180	r alignment scores 50-80 80-9 270	00 ≻=200 360 45	0		
		-40 1 9	Color key fo 40-50 0 180	r alignment scores 50.50 80.5 270	00 >=200 360 45	0		
	•	Suery 1 9 -	Color key fo 40-50 0 180	r alignment scores 50.50 80-5 270	00 >≈200 360 45	0		
⊕Dot Matrix Vie			Color key fo 40-50 0 180	r alignment scores 50.00 80-0 270	.00 ≻≈200 360 45	0		
Dot Matrix Vie     Descriptions	•	3uery <mark>- 40</mark> 1 91 -	Color key fo 40-50	ralignment scores	00 >≈200 360 45	0		
⊕ <u>Dot Matrix Vie</u> ⊖ <u>Descriptions</u>		-40 1 9	Color key fo 40.50	ralignment scores	360 →=200 360 45	o		
⊕ <u>Dot Matrix Vie</u> ⊖ <u>Descriptions</u>		-40 1 91 -	Color key fo 40.50	r alignment score 1 50-50 270	360 →≈200 360 45	0		
Dot Matrix Vie     Descriptions     Sequences p	w Z	Suery 40 1 9 ments:	Color key fo 40-50	r alignment scores 50.50 80.4 270	360 ×=200	0		
Dot Matrix Vie     Descriptions     Sequences p     Select: All N	wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww	overy <a href="https://www.selfacture.com">&lt; 40</a>	Color key to 40-50	r alignment scores couronal 80-2 270	00 >≈200 360 45	o		
Dot Matrix Vie     Descriptions     Sequences p     Select: Al N     Algnmen	wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww	nents:	Color key fo 40-50 0 180	r alignment scores consoliniti 80 d 270	00 >⊭200 360 45	o		0
Dot Matrix Vie     Descriptions     Sequences p     Select: All N     Algomen	www. croducing significant alignm tone Selected 0 ta @Download = Graphi	nents:	Color Key to 40-50 0 180	r alignment scores colocidade 270	00 >=200 360 45	Total Query	Eigen	0
Dot Matrix Vie     Descriptions     Sequences p     Select: All N     Select: All y	w Z producing significant alignm kone Selected 0 ts @Download = Graphi	eents: Multicle_slightneet Desc	Color key to 40-50 0 180	r alignment scores concord 80 d 270	360 →=200 360 45	Total Query score cover	E Ident	O Accession

Average .004872.1 503-1 +	NCBI Blast:NP_008872.1	SCX-10 Biomo sapier	43 M66							
at Achi ala Air gay, Blan, ca					∀¢ R	•		_	٩	4
Dot Matrix View /										
Sequences producing significant alignments:										
ii Alignments El Download - Graphics Multiple al	lanment.								0	
Descrip	tion			Max score	Total score	Query cover	E value	Ident	Accession	
NP_003131.1 sex determining region Y [Homo sapiens]				94.0	109	19%	1e-26	46%	19241	
Alignments	:)				₩ Ne	ixt a P	revious	A Det	scriptions	
Dialignments Download ~ <u>Graphics</u> sortby: <u>Evalue</u> NP_003131.1 sex determining region Y [Homo Sequence ID: Icl]19241 Length: 204 Number of Match	sapiens] ses: 2				₩ Nc	Rt A P	revious	₫ Det	scriptions	
DAlignments Download ~ Graphics sert by: Evalue NP_003131.1 sex determining region Y [Homo Sequence ID: Id](19241 Length: 204 Number of Match Range 1: 51 to 134 Graphics	t] sapiens] ues: 2	♥ Next Mat	ch 🛓 Previous I	fatch	₩ Ne	Relat	revious ied Info	∳ Det	scriptions	
Developments  Download ~ Graphics sert by: Evalue  NP_003131.1 sex determining region Y [Homo Sequence ID: Id][19241 Length: 204 Number of Match Range 1: 51 to 134 Graphics Score Expect Method 94.0 bits(22) 1:=-26 Compositional metrix adjus	: sapiens] tes: 2 Identities st. 39/84(46%)	▼ Next Matt Positives 62/84(73%)	ch Previous Gaps 0/84(0%)	fatch	₩ Ne	Relat	revious red Info	∦ De ormati	scriptions on	
Alignments     Download ~ <u>Graphics</u> sort by: Evalue     NP_003131.1 sex determining region Y [Homo     Sequence ID: Id]19241 Length: 204 Number of Match     Range 1: 51 to 134 <u>Graphics</u> Score Expect Method     94.0 bits(232) 1e-26 Compositional metrix adjus     Query 95 NOARSERIPTIKETPREATIVENT REFLATS     Sbjec 51 NSARKERIPTIKETPREATIVENT REFLATS     Sbjec 51 NSARKERIPTIKETPREATIVENT REFLATS     Duery 155 EDERELSPORTKETPREATIVENTORED     Duery 155 EDERELSPORTKETPREATIVENTORED	: sapiens] mes: 2 Identities t. 39/84(46%) YYPHLHNALEKSTCK P + 39-85K LG NIRPARENSEISKOLGY	V Next Matt Positives 62/84(73%) IMRLINESDRR R++1 E++X QMRMLTEAERW	ch A Previous Gaps 0/84(0%) PPI 154 PP PPF 110	fatch	₩ Ne	Relat	ravious ced Info	A De	on	
Alignments     Download ~ <u>Graphics</u> Sert by: Evalue     NP_003131.1 sex determining region Y [Homo     Sequence ID: Id]19241 Length: 204 Number of Match     Range 1: 51 to 134 <u>Graphics</u> Score Expect Method     94.0 bits(232) 1e-26 Compositional matrix adjust     Ouery 55 NOARESERIFYREDERS/PVMACDARRELAD     Socie 51 NSCREWORKEDIDGYKTOPERGAR 178     ethol 11 OEAOKLOADIRERKTPMYKTERREX 134     Dence 2.0 54 to 121 Compo	2 sapiensj wes: 2 Identities st. 39/84(46%) YYPHIMANALSKITCK P + %+2+5K LC NNPARHSEISKOLCY	▼ Next Mat Positives 62/84(73%) LVRLLNESDKRUNESDKR WKNLTEAEKW	ch & Previous Gaps 0/84(0%) PPI 154 PPF 110	tatch	₩ Ne	Relat	ravious ed Info	≩ De: ormati	on	
Alignments  Download ~ Graphics sort by: Evalue  NP_003131.1 sex determining region Y [Homo Sequence ID: bi[19241 Length: 204 Number of Match Range 1: 51 to 134 Graphics Score Expect Method 94.0 bits(232) 1e-26 Compositional matrix adjus Ouery 95 NOARSERINGKONDENCHAPPENDALMERLA Ouery 155 EBABBLENOIRKONDENCHAPPENDALMERLA Ouery 155 EBABBLENOIRKONDENCHAPPENDAL OUERY 155 EBABBLENOIRKONDENCHAPPENDENCHAPPENDAL OUERY 155 EBABBLENOIRKONDENCHAPPENDEN	2) sapiens] mes: 2 Identities st. 39/84(46%) NYPHILMALSKILC NYPHIL	▼ Next Mat Positives 62/84(73%) UH7LLNESDK4(73%) UH7LLNESDK4 W+5L E++X (WK0L17EAEKW (MK0L17EAEKW (MK0L17EAEKW (MK0L17EAEKW) (MK0L17EAEKW) (MK0L17EAEKW) (MK0L17S) (MK	ch A Previous Gaps O/B4(0%) PPF 154 PP PFF 110 Match & First Gaps	4arch	₩ Ne	Relat	ravious ced Info	<b>▲</b> De	on	
Alignments     Bownload ~ Graphics sertey: Evalue     NP_003131.1 sex determining region Y [Homo     Sequence ID: [d] 19241 Length: 204 Number of Match     Range 1: 51 to 134 Gappics     Score Expect Method     94.0 bits(212) 1=-26 Compositional matrix adju     0uery 95 NGARKERSHWEINDRAPHYMELOD     8 VERDBRACHWEIN REAL     0uery 155 EEAPERCHMENDRAPHYMELORERSMAR     0uery 155 EEAPERCHMENDRAPHYMERTERS     51 SISSERVORKWEINDERSMAPTINGERSMAR     0uery 155 EEAPERCHMENDRAPHYMERTERS     5bject 111 OBAOKLOANIIRERTERVERTERS     5core Expect Method     15.4 bits(28) 1.3 Compositional matrix adju	2) sapiens] ses: 2 Identities st. 39/84(46%) YYPHIEMALESTICS 19 + 8454554 NRPARESTICS NRPARESTICS 1 dentities st. 3/7(43%)	V Rest Mat Positives 62/84(73%) UH7LINES0K4(73%) W+L E++K GWKMLTELAEKN Positives 5/7(71%)	ch ▲ Previous Gaps 0/84(0%) PFI 154 PFFI 110 Match ▲ First Gaps 0/7(0%)	faich faich	♥ No	Relat	revious red Info	🛓 De	on	
Alignments     Download ~ Graphics sort by: Evalue     NP_003131.1 sex determining region Y [Homo     Sequence ID: Id][19241 Length: 204 Number of Match     Range 1: 51 to 134 Graphics     Score Expect Method     94.0 bits(22) 1=-26 Compositional matrix adju     0uery 95 NCARESCHIVENEDRAFYWARDARSHLAD     0uery 95 NCARESCHIVENEDRAFYWARDARSHLAD     0uery 155 BEARLIADOURKCHIPDYRYGORERS 1/8     85ject 111 0EAGKLADOURKCHIPDYRYGORERS 1/8     85ject 111 0EAGKLADOURKCHIPDYRYGORERS 1/8     Score Expect Method     15.4 bits(28) 1.9 Compositional matrix adju     Query 82 OXYMPLY 88	2 sapiens] mes: 2 LdentHise st. 39/84(45%) YPHLMNSLEKTLG IDENTIFIES IDE	Next Mat     Society     Next Mat     Society     Society     Society     Society     Society     Next Mathematical     Society     S	ch & Provideos Gape 0/84(0%) PPI 154 PPF 110 Match & First I Gaps 0/7(0%)	fatch fatch	▼ Ne	Relat	revious red Info	â De:	on	





0	2A.B	T. Basic Local Algement Search Tool
LAST: Basic Local Alignment 5		
P DELOCIO PERMISSIONI D		http://nchi.nlm.nih.gov/BLAST
BLAST*	Basic Local Alignment Search To	mup.//ncol.num.nun.gov/DLASI
Home Recent	Results Saved Strategies Help	
381/ BLAST Home		Your Resent Results New
BLAST finds regio	one of similarity between biological sequences. more-	
	DELTA-BLAST, a more sensitive pr	rotein-protein search Gal
BLAST Assem	bled RefSeg Genomes	News
		BLAST 2.3.29+ released
Unocse a species g	enome to search, or list all genomic BLAST databases.	NOTE AND A DESCRIPTION
= Human	Oryza sativa	<ul> <li>Galfus galfus</li> <li>A new vention of the stand-atone BLAST+</li> </ul>
Mouse	<ul> <li>Bos taurus</li> </ul>	Pan troglodytes     Ann 00 lar 3614 (2 00 m EPT
· Rat	a Danio rario	Microbes
Arabidopsis the Arabidopsis	Drosophile melanogaster	Apis melilifera
Basic BLAST		
Chocse a BLAST p	rogram to run.	Tip of the Day
		10 More Spelin
nucleotide blast	Search a nucleotide database using a nucleotide query Algorithms: blastn, megablast, discontiguous megablast	
protein blast	Search protein database using a protein query Algonithms: blastp, psi-blast, phi-blast, delta-blast	
blastx	Search protein database using a translated nucleotide query	
tblastn	Search translated nucleotide database using a protein query	
tblastx	Search translated nucleotide database using a translated nucleotide	s dneuλ
Specialized BL	AST	
Choose a type of st	verialized search /or database name in narentheses )	
choose a type of a		
Make sp	ecific primers with Primer-BLAST	
· Search t	race archives	
Find con	served domains in your sequence (ods)	
Find seq	uences with similar conserved domain architecture (cdart)	
Search s	sequences that have gene expression profiles (GEO)	
Search i	mmunogroousins and T cell receptor sequences (IgBLAST)	
<ul> <li>acreens</li> <li>Allers treet</li> </ul>	c (or more) and warder up a RI AST (bland)	
- Align tw	o (or more) sequences carry ouncer (0(2500) rentals or nucleating tomate in D-it/Cham BioAceau	
<ul> <li>Search 5</li> </ul>	SRA by experiment	
Constrai	nt Based Protein Multiple Alignment Tool	

0.0	Nucleot de BLAST: Search nucleot	de databases using a nucleatide query	
() () Marchelinin all gov	Bar, cp/WOCKM-blacs&WCE, TVIC-BarCearth&, NE LOC-barrowe	(1 × c) (B-	9400
DI AOT			Mu NCBI
Home Racer	Eastic Local Alignment Search 1001		Welcome baxevania. (Sign.O
NCBI BLAST/ blastr	suita Standard Nu	cleotide BLAST	
blastn blastp	blasts tblastn tblasts		
Enter Query S	BLASTN programs search nucleotide	artabases using a nucleotide query. more	Reset page Bookmark
Enter accession r	number(s), gi(s), or FASTA sequence(s) @ Gen Query subz	noe G	
	From		
	d To		
Or, upload file	Browse No file selected.		
Job Title			
	Enter a descriptive title for your BLAST search @		
Align two or m	ore sequences 🚇		
Choose Searc	th Set		
Database	GHuman genomic + transcript GMouse genomic + transcript GOthers (nr etc.)	¢	
	Nucleotide collection (nr/nt) 0		
Organism Optional	Triterorganism name or id-completions will be suggested		
	Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown 🔛		
Exclude Optional	Models (XM/XP)      Uncultured/environmental sample sequences		
Entrez Query Optional			
	Enter an Entrez query to limit search 🥩		
Program Sele	ction		
Optimize for	Highly similar sequences (megablast)		
	<ul> <li>More dissimilar sequences (discontiguous megablast)</li> </ul>		
	Somewhat similar sequences (biastn)		
	Cione a post adoram de		
BLAST	Search database Nucleotide collection (nrint) using Megablast (Optimize for	highly similar sequences)	
	Show results in a new window		
Algorithm param	oters		
General Pa	amelers		
sequences	Select the maximum number of aligned sequences to display @		
Sheet must be			
Short queries	M Automatically adjust parameters for short input sequences 42		

Nucleotide-Based BLAST Algorithms							
		+/	Gaps				
Optimized for aligning very long and/or high	ly similar	sequences (>	<b>&gt; 95%</b> )				
MegaBLAST (default)	28	1, -2	Linear				
Better for diverged sequences and/or cross-s	pecies com	parisons (<	80%)				
Discontiguous MegaBLAST	11	2, -3	Affine				
BLASTN	11	2, –3	Affine				
Finding short, nearly exact matches (< 20 bases)							
BLASTN	7	2, -3	Affine				
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	000 <u>0</u>	IATIONAL HUMAN GEN	DME RESEARCH INSTITUTE Division of Intramural Research				

# **BLAT**

- "BLAST-Like Alignment Tool"
- Designed to rapidly align longer nucleotide sequences (L ≥ 40) having ≥ 95% sequence similarity
- Can find exact matches reliably down to L = 33
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster than BLAST for mRNA/DNA searches
- May miss divergent or shorter sequence alignments
- Can be used on protein sequences, but BLASTP is more efficient

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

### When to Use BLAT

- · To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates

- Determine gene structure (the presence and position of exons)
- Identify markers of interest in the vicinity of a sequence
- To find highly similar (or identical) sequences
  - Alignment of mRNA sequences onto a genome assembly
  - Identification of gene family members

- · Cross-species alignment to identify putative homologs
- To display a specific sequence as a separate track within the UCSC Genome Browser

000	UCSC Cenome Browser Home	
(d) @ genome JCSC Add	ar water and a second and a s	http://genome ucsc edu
11000	De De la companya de	mip.mgenome.nese.eau
UCSC	, Genome Bioinformatics	
Genomes -	Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help	
Genome	About the UCSC Genome Bioinformatics Site	
Browser	Welcome to the UCSC Genome Browser website. This site contains the reference sequence also provides portals to the ENCODE and Neandertal projects.	e and working draft assemblies for a large collection of genomes. It
Neandertal	We encourage you to explore these sequences with our tools. The Genome Browser zooms worldwide. The Gene Sorter shows expression, homology and other information on groups encluses to the genome. The Table Reuser provides groups access to the underlands.	and scrolls over chromosomes, showing the work of annotators of genes that can be related in many ways. Blat quickly maps your database. VielGage lets you browse through a large collection of in
Table	situ mouse and frog images to examine expression patterns. Genome Graphs allows you to	upload and display genome-wide data sets.
Browser	The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics G	Froup, a cross-departmental team within the Center for Biomolecular
Gene Sorter	Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you h website, feel free to contact us on our public mailing list.	have feedback or questions concerning the tools or data on this
In Silico PCR	· · ·	
Genome	News 🖸	News Archives ►
Galaxy	To receive announcements of new genome assembly releases, new software features, upda announce mailing list.	ates and training seminars by email, subscribe to the genome-
VisiGene	22 January 2014 - All the DNA on the internet now at your fingertips	1
Utilities Downloads	We're pleased to announce the release of the Web Sequences track on the UCSC Genome Research, contains the results of a 30-day scan for DNA sequences from over 40 billion diff human genome (hg19) and numerous other species including mouse (mm9), rat (m4), and squares including naterats online favitholosis, help forams and any other webaresets that conto	Browser. This track, produced in collaboration with Microsoft erent webpages. The sequences were then mapped with Blat to the zebrafish (danRer7). The data were extracted from a variety of an DNA sequence. In service, this track displays the Blat alignments
Release Log	of nearly every DNA sequence on the internet! The Web Sequences track description page	contains more details on how the track was generated.
Custom Tracks	We would like to acknowledge Max Haeussler and Matt Speir from the UCSC Genome Brow work in creating this track.	vser staff and Bob Davidson from Microsoft Research for their hard
Cancer Browser	14 January 2014 - Two E. coli comparative assembly hubs now available	le
Microbial Genomes	We're pleased to add two new assembly hubs produced by the UCSC David Haussler lab to 60 bacterial assemblies, including more than 55 different E. coli strains. The assembly anno repetitive elements and much more.	o our collection of publicly available hubs. The new hubs feature over tations include genes, pathogenic genes, conservation, GC percent,
Mirrors	These hubs focus on comparative genomics and showness the new "enake" track tupe. Sha	kes which visualize alignments from Higrarchical Alignment (H&I.)
Training	These much source of comparison generating and/case and and/case in the window of acc type. The files, provide a way to view sets of pairwise gapless alignments that may everage on both the various types of genomic variations such as insertions, substitutions, and duplications. More options can be found on our Genome Browser help page.	e chosen genome (reference) and the query genome, and show o details about the new snake track display and its configuration
Publications	To access and view these hubs, navigate to the Track Hub gateway page and select one of	the two E. coli comparative assembly hubs from the Public Hubs list.
Cite Us	We would like to acknowledge Ngan Nguyen, Glenn Hickey, Brian Raney, Joel Armstrong, E	Benedict Paten, Matt Speir, and Luvina Guruvadoo for their hard work

						Rat BLAT Search			
B	Rat BLAT Search	(*						() = () (B)	
100	Genomes	Genome Browser	Tools	Mirrors	Downloads	s My Dat	a About Us	Help	
let I	DI AT Search								
(at I	BLAI Search								
BL	AT Search	I Genome							
	Genome:	Assembly:	Qu	uery type:	Sort output:	Output type			
Rat	;	Nov. 2004 (Baylor 3.4/r	n4) : DNA		query,score 🖻	hyperlink	-		
>CB3 GGGG	12815 NICHD_Rr	Pit1 Rattus norveg	JICUS CDNA	clone	ATTTCCTABACTC	TC			
TACO	TCTGGTTCATGTTC	GCTCTTCTGGATAGTCTG	TGTGCAATGA	GCCCTTAAA	GGAATATTGCAAT	GA			
GCTO	CTANGTCACAGGTT	ATCCAGAGCCCACTTTAC	CCARGAGAG	AGCCTCTCC	CCCATCCCTAGGA	22			
CAGI	AGAGCTTAGGAAAA	TGAATGACTCCACCACAT	ICAAGAGGCT	TCAAATTGT	ATACTTGGCATTT ACACCTTGTCATT	CT TA			
222	ACCATTGAATTAAG	AGAAATGGAAAATCATGC	CACATAAAA	CATGTATGG	AAGTGTTCATGTT	17			
GATO	CATGGCGGGGGGATAT GCACCGAAAAAAGAG	AGCTCAGTCATGGAGTGC	TIGCATAGCA CATTCACAGO	ATGTGCATA	ATCCGAGGTTCAA AGTATAGGCGCAA	GC AG			
GGG7	AGGAGTTTAAACAC	CTACTGAGGGAATGGATAJ	GCGGAGTGC	CCTTGTCTA	TACTCGGGGATGG	CT			
AGIC	arcatornaonaan	OTTIGGRANI OTTIG	PAGGANTONS	AIGGAIGGG	CITIAAACCATCO				
		7 7	,				la la		
		submit	m feeling luck	clear					
Pas	te in a query see	quence to ind its loca	ation in the	the genor	me. Multiple se	equences ma	У		
be s	earched if sepa	rated by lines starting	g with '>' fo	blowed by	the sequence	name.			
File	Upload: Rathe	r than pasting a secu	ence. vou	can choos	se to upload a	text file			
cont	aining the sequ	ence.	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,						
Uplo	ad sequence:	Browse No file selected.	. sub	mit file					
Only		e of 25 000 or forwar	hacee an	d protein e	r translatori en	nuence of			
1000	00 or fewer lette	rs will be processed	Un to 25	sequences	can be submi	itted at the			
sam	e time. The tota	I limit for multiple sec	uence sul	bmissions	is 50.000 base	s or 25.000			
lette	rs.					,			
Fee		imore use la Cilica I		at reaults 1	note of DLA	-			
For	locating PCR pr	imers, use in-Silico P	CR for be	IST RESULTS I	nstead of BLA	1.			
Abo	ut BLAT								
BLA	T on DNA is des	igned to guickly find	sequence	s of 95% a	nd greater sim	nilarity of leng	th 25 bases or	more. It may miss	s more divergent or shorter sequence
align	ments. It will fin	d perfect sequence r	matches of	20 bases.	BLAT on prot	eins finds se	quences of 80°	% and greater simi	larity of length 20 amino acids or more.
In pr	actice DNA BLA	T works well on prim	ates, and	protein bla	t on land verte	ebrates.			

0 0 0	+				Rat BLAT Results			
() @ genome.ucsc.edu/cgi-b	bin/hgliat						合호 @ (원*	9.4.6
🏫 Genor	nes Genome I	Browser To	ols Mirror	s Download	s My Data	About Us	Help	
Rat BLAT Resu	ults							
BLAT Sear	rch Result	s						
ACTIONS	QUERY	SCORE STAR	T END QSIZE	IDENTITY CHRO	STRAND START	END	SPAN	
browser details browser details	CB312815 CB312815	710 1 29 501	733 768 537 768	98.18 5 89.28 2	+ 101455599 + 38736251	101456323 38736287	725 37	
bro ser details	CB312815 CB312815	25 501 22 341	529 768 363 768	93.28 3 100.08 1	+ 22960346 + 122930956	22960374 122930979	29 24	
bro ser details bro ser details	CB312815 CB312815 CB312815	21 202 21 706 21 552	222 768 727 768 574 769	100.08 17 100.08 3	- 33248146 + 46857920	33248166 46857942	21 23	
bro ser details bro ser details	CB312815 CB312815 CB312815	20 277 20 442	298 768 461 768	95.5% 2 100.0% 1	- 240446870 - 216323127	240446891 216323146	22 20	
bro ser details bro ser details	CB312815 CB312815	20 508 20 453	527 768 474 768	100.0% 1 95.5% 2	- 56102029 + 186587336	56102048 186587357	20 22	



0 0 0			Rat RLAT Results		
(d) @ genome.ucsc.edu/cgi-bin/hgliat				☆ * C (B*	9
A Genomes Genome B	Browser Tools	Mirrors Download	ds My Data About L	Js Help	
Rat BLAT Results					
BLAT Search Results	5				
ACTIONS QUERY	SCORE START END	QSIZE IDENTITY CHRO	0 STRAND START END	SPAN	
browser details CB312815	710 1 733	768 98.16 5	+ 101455599 101456323	725	
browser det ls CB312815	25 501 537	768 93.26 3	+ 22960346 22960374	29	
browser det ils CB312815	21 202 222	768 100.08 17	- 33248146 33248166 + 46857920 46857942	21	
browser det ils CB312815	21 552 574	768 95.78 1	+ 157973111 157973133	23	
browser det ils CB312815	20 442 461	768 100.08 1	- 216323127 216323146 - 56102029 56102048	20	
browser det ils CB312815	20 453 474	768 95.5% 2	+ 186587336 186587357	22	

000	User Sequence vs Genomic al							
User Sequence vs Genomic								
(d) @ genome ucsc.edu/cgi-bin/1	hg/tw=1014555566g=hcdbiesAkl4=httph://git./hgfs.genceme_967_270e800 gale+_362/trash62fhg562fhg							
Alignment of	Alignment of CB312815 and chr5:101455599-101456323							
CB312815	inginient of oboligite and enterior levers for levere							
	Click on Victoria the forme to the 160 to environt the element. Matching bases in cDNA and experies converses on colored							
CB312815	blue on any field Lick to the fert to having the hourd grine of energy in attenting bases in cDNA and genomic sequences are colored							
Rat.chr5	blue and capitalized. Light blue bases mark the boundaries of gaps in entire sequence (often spice sites).							
block1								
together	cDNA CB312815							
	Generating consecuting representation accounter 50							
	TOTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT 100							
	AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG 150							
	TIGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT 200							
	AARACARCC CCCCCCCCAR CCCCARGETAR CCRARGCCC CTTACCARARC 300							
	AATGACTICCA CCACATTCAA GAGCCTTCAA ATTGTATACT TGGCATTTCT 350							
	GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG 400							
	GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA 450							
	TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG 500							
	GOGGATATAG CTCAGTCATG GAGAGCATCA ATAGCAATGT GCATAATCG 550							
	CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGAGT TTAAACACCT 550							
	ACTGAGGgAA TGGATAAGCG GAGTGCCCTT GTCTATACTC GGGgatgGCT 700							
	AGTCATCACG taAGAAAAGT TTGgaAAATG ATAaaatacc aatgggatgg 750							
	atcccttta aaccatcc							
	Genomic chr5 :							
	cttggaagaa ggtaactata cattaatata gagcootott tttotttgca 101455548							
	ggcccaggac acacaggacg gatgtttcca agtcactcca gggacagcat 101455598							
	GROCIETE ETGEVENTE ALCONTROL TETTETE ACTIVE ACTIVITET 10145569							
	AGTCTGTGTGT CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG 101455748							
	TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT 101455798							
	CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA 101455848							
	ALAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAATG 101455898							
	GATTCAGT CREANTICE COCCTAGT CREGGARA ATAGAAAG 101453998							
	GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA 101456048							
	TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG 101456098							
	GEGEATATAG CTCAGTCATG GAGTCCTTCC ATAGCAATGT CCATAATCCG 101455148							
	AUSTICARSC CULASCAUCS ARAAGAGAGA GCGGGAGGAG TGGAGGCATT 101455198							
	ACTGAGGAAT GGATAACCG ACTCCTTGT CTATACTCGG GALGCTAGTC 101455298							
	ATCAgtaGAA AAGTTTGAAA TGATAgatac gatggatgat cccttaaaca 101456348							
	torcoctaag taacgaggto agostaggaa ggostatgtt coatattet 101456398							

0 0 0 User Sequence vs Genomic	User Sequence vs Cenomic	
() @ genome JESE edu/cgi-bin/hg	gch=1014555988g=hcdlueAl&i=_thubhdps/hcfs.genome_967.270x80 psixe=32fhubHcfhgfsk2fhgfs.genome_997.270x80.fs+C03128158c=chr54i=1014 🗇 🕫 🛛 🔯 •	* * 0-
Alignment of CB312815	Side by Side Alignment	
CB312815 Rat.chr5	00000001 ggggctctcgctggctctggtgtctcagaagstgcttctccacctttcct 00000005 >>>>>>> [	
block1 together	00000051 tgtgaatttoctaaactototaoototggttoatgttogotottotogat 000000100 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000101 agtctgtgtgcaatgagcccttaaaggaatattgcaatgagctataagag 000000150 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000151 ttgtgagoctgcggtaggcaaggcctgcactgggacagcaaaggaaatt 000000200 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000201 cattgctctctaagtcacaggttatccagagcccactttacccca 000000250 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000251 agagacagcctctccccctacccctaggaacagtagagcttaggaaaatg 000000300 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000301 aatgactocaccacattcaagaggottcaaattgtatacttggcatttot 000000350 >>>>>>>> atgactocaccacattcaagaggottcaaattgtatacttggcatttot 101455948	
	000000351 gatttcagtctqaaattctgtcccttagtcgtgggaaaataagaaatg 000000400 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000401 gagttacaccttgtcatttaaaaaaccattgaattaagagaaatggaaaa 000000450 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000451 toatgoccacataaaacatgtatggaagtgttcatgttttgatcatggcg 000000500 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000501 ggggatatagctoagtoatggagtgcttgcatagcaatgtgcataatccg 000000550 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>	
	000000551 aggttcaagccccagcaccgaaaaagagaagaggagtggaggagtggaggagt >>>>>>>>>>	
	000000601 cacagcagcgttttcagtataggcgcaaaggggaaggagtttaaacacct 000000650	

# FASTA

- Identifies regions of local alignment
- Employs an approximation of the Smith-Waterman algorithm to determine the best alignment between two sequences
- Method is significantly different from that used by BLAST
- Online implementations at:

http://fasta.bioch.virginia.edu http://www.ebi.ac.uk/fasta33

