Mini Courses

NCBI Resources

one course at a time

NCBI Minicourses

BLAST Quick Start

NCBI Minicourses

Algorithm Basics

- Introduction
- Words & extensions

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Basic Local Alignment Search Tool

NCBI Minicourses

Compare protein and nucleic acid sequences:

- in NCBI databases
- in local databases (standalone BLAST)
- to a single protein or nucleotide sequence (BLAST 2 Sequences, or pairwise BLAST)

Basic Local Alignment Search Tool

NCBI Minicourses

- local alignments; isolated regions of similarity
- fast and sensitive
- breaks the query sequence into “words”
- word hits to database sequences extended in both directions

Global vs Local Alignment


```
Seq1:  WHEREISWALTERNOW      (16aa)
Seq2:  HEWASHEREBUTNOWISHERE (21aa)
```

Global

```
Seq1:  1                W--HEREISWALTERNOW 16
                        W  HERE
Seq2:  1  HEWASHEREBUTNOWISHERE            21
```

Local

```
Seq1: 1  W--HERE 5        Seq1:  1 W--HERE  5
         W  HERE                   W  HERE
Seq2: 3  WASHERE 9        Seq2: 15 WISHERE 21
```

Basic Local Alignment Search Tool

- local alignments; isolated regions of similarity
- fast and sensitive
- breaks the query sequence into “words”
- word hits to database sequences extended in both directions

Nucleotide Words

Make a lookup table based on the word size.

11-mer

ATGCTGCTAGTCGATGACGTAGCTA

ATGCTGCTAGT
TGCTGCTAGTC
GCTGCTAGTCG
…

Protein Words

AIEKCYTGCTLAQEADDTA

AIE
IEK
EKC
KCY
CYT
…

LEK, IDK, IQK, IER, IDR, etc

Neighborhood words

Lookup table, including neighborhood words, is based on word size, score matrix, and threshold.

Scoring Systems - Proteins (BLOSUM62)

NCBI Minicourses

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1

IEK: keep LEK?
(threshold = 11)

I / I — IEK = 14

L / I — LEK = 12

Word Hits & Extensions

NCBI Minicourses

Nucleotide: one exact match

⇦ GCTGCTAGTCG ⇨

Protein: two matches within 40 residues

⇦ IDK ⇨ ⇦ EAD ⇨

BLAST

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

NCBI Minicourses

BLAST Programs

What is your goal?

blastn	nucleotide X nucleotide
blastp	protein X protein

6 frame, translated nucleotide searches

blastx	nucleotide X protein
tblastx	nucleotide X nucleotide
tblastn	protein X nucleotide

NCBI Minicourses

More BLAST Programs

NCBI Minicourses

	Word Size
MegaBLAST - batch nucleotide queries - very similar sequences	28
Discontiguous MegaBLAST - batch nucleotide queries - divergent sequences	11 matches out of 18
BLAST 2 Seq. (pairwise)	

Templates for Discontiguous Words

NCBI Minicourses

```
W = 11, t = 16, coding:       1101101101101101
W = 11, t = 16, non-coding:   1110010110110111
W = 12, t = 16, coding:       1111101101101101
W = 12, t = 16, non-coding:   1110110110110111
W = 11, t = 18, coding:       101101100101101101
W = 11, t = 18, non-coding:   111010010110010111
W = 12, t = 18, coding:       101101101101101101
W = 12, t = 18, non-coding:   111010110010110111
W = 11, t = 21, coding:       100101100101100101101
W = 11, t = 21, non-coding:   111010010100010010111
W = 12, t = 21, coding:       100101101101100101101
W = 12, t = 21, non-coding:   111010010110010010111
```

W = word size; # matches in template

t = template length

Reference: Ma, B, Tromp, J, Li, M. PatternHunter: faster and more sensitive homology search. Bioinformatics March, 2002; 18(3):440-5

BLAST

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

NCBI Minicourses

BLAST Databases – Nucleotide

nr	GenBank+EMBL+DDBJ+PDB sequences (no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS)
refseq_rna	RefSeq RNA sequences (NM_, XM_)
refseq_genomic	RefSeq genomic sequences (NC_, NT_, NG_)
month	new or revised GenBank+EMBL+DDBJ+PDB sequences from the last 30 days
est	Expressed Sequence Tags (EST division)
dbsts	Sequence Tag Site sequences (STS division)
htgs	Unfinished High Throughput Genomic Sequences, phases 0, 1 and 2 (finished phase 3 go to nr)

NCBI Minicourses

HTGS – High Throughput Genomic Sequences

NCBI Minicourses

Phase 0	single-pass reads of a single clone
Phase 1	unfinished, unordered and contain gaps
Phase 2	unfinished, ordered and contain gaps
Phase 3	high quality finished sequences no gaps

More BLAST Databases – Nucleotide

NCBI Minicourses

gss	genome survey sequences
pdb	Protein Data Bank nucleotide sequences
chromosome	equivalent to refseq_genomic; complete genomes, chromosomes, contigs

BLAST Databases – Protein

NCBI Minicourses

nr ≠ nr

nr	non-redundant GenBank CDS translations+PDB +SwissProt+PIR+PRF
refseq	RefSeq proteins
month	new or revised GenBank CDS translation+PDB +SwissProt+PIR+PRF released in the last 30 days
Swissprot	last major release of the SWISS-PROT database
pdb	Protein Data Bank protein sequences
env_nr	sequences from environmental samples

BLAST Databases – Genome-specific

NCBI Minicourses

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

BLAST

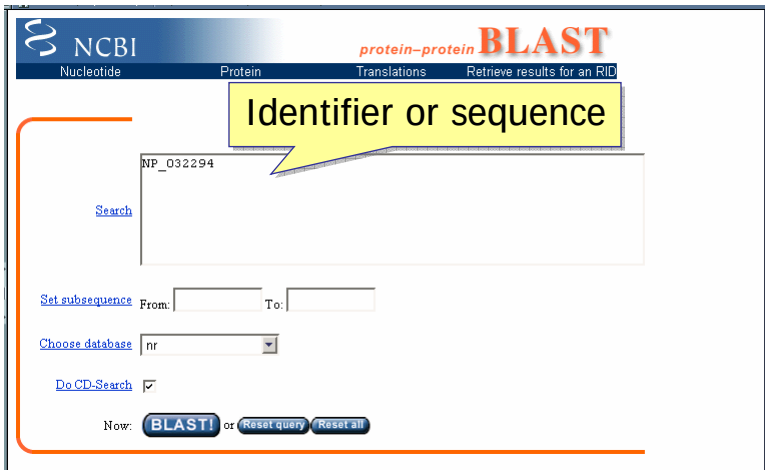
Search Basics

NCBI Minicourses

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Setting up a BLAST Search

NCBI Minicourses



NCBI

NCBI formatting BLAST

Nucleotide Protein Translations Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = gi|6680251 (269 letters)

Hit the button to See conserved domains from CDD

The request ID is 1170281941-31729-97914371633.BLASTQ4

Format! or Reset all

BLAST Output

NCBI Minicourses

NCBI results of BLAST

BLASTP 2.2.15 [Oct-15-2006]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference:
Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994-3005.

RID: 1170281941-31729-97914371633.BLASTQ4

Database: [illegible] non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples
4,500,144 sequences; 1,545,660,933 total letters

If [illegible] have any problems or questions with the results of this search pl[illegible] refer to the **BLAST FAQs**
Taxonomy reports

Tax BLAST Report

NCBI Minicourses

- Lineage Report
- Organism Report
- Taxonomy Report
- Help

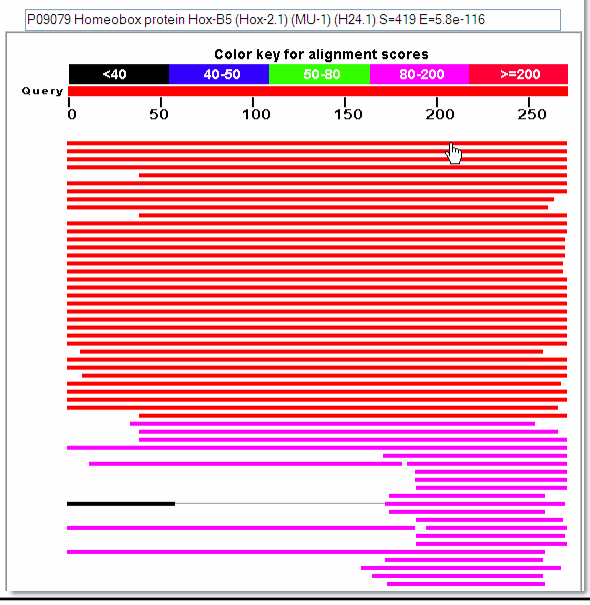
Lineage Report

```
root
. Bilateria          [animals]
. . Coelomata        [animals]
. . . Deuterostomia    [animals]
. . . . Chordata         [chordates]
. . . . . Vertebrata       [vertebrates]
. . . . . . Gnathostomata    [vertebrates]
. . . . . . . Euteleostomi     [vertebrates]
. . . . . . . . Sarcopterygii    [vertebrates]
. . . . . . . . . Tetrapoda        [vertebrates]
. . . . . . . . . . Amniota          [vertebrates]
. . . . . . . . . . . Eutheria         [mammals]
. . . . . . . . . . . . Euarchontoglires [mammals]
. . . . . . . . . . . . . Muroinae          [rodents]
. . . . . . . . . . . . . . Mus             [rodents]
. . . . . . . . . . . . . . . Mus musculus (mouse) ----------------------  420  41 hits  [rodents]   homeo box B5 [Mus musculus] >gi|5223335|g
. . . . . . . . . . . . . . . Mus sp. (mice) ..............................  126   2 hits  [rodents]   Hoxa-4 [Mus sp.] >gi|9488660|gb|AAB20705
. . . . . . . . . . . . . . . Rattus norvegicus (brown rat) ---------------  419  18 hits  [rodents]   PREDICTED: similar to Homeobox protein H
. . . . . . . . . . . . . . . Rattus sp. (rats) ............................  120   2 hits  [rodents]   Hox-B1Hox-2 [Rattus sp.]
. . . . . . . . . . . . . . Homo sapiens (man) ------------------------------  412  42 hits  [primates]   homeobox B5 [Homo sapiens] >gi|14666052
. . . . . . . . . . . . . . Lagothrix lagotricha (Humboldt's woolly monkey) .  248   1 hit   [primates]   HOXA5 [Lagothrix lagotricha]
. . . . . . . . . . . . . . Lemur catta ...................................  247   1 hit   [primates]   HOXA5 [Lemur catta]
. . . . . . . . . . . . . . Macaca mulatta (rhesus macaque) ...............  244  10 hits  [primates]   PREDICTED: similar to Homeobox protein H
. . . . . . . . . . . . . . Pan troglodytes ...............................  239  10 hits  [primates]   PREDICTED: hypothetical protein isoform.
. . . . . . . . . . . . . . Gorilla gorilla ...............................  165   2 hits  [primates]   HOXA5 [Gorilla gorilla]
. . . . . . . . . . . . . . Pan paniscus (bonobo) .........................  152   2 hits  [primates]   HOXB5 [Pan paniscus]
. . . . . . . . . . . . . . Saguinus labiatus .............................  152   3 hits  [primates]   HOXB5 [Saguinus labiatus]
. . . . . . . . . . . . . . Pongo pygmaeus (orang-utan) ...................  149   1 hit   [primates]   HOXB5 [Pongo pygmaeus]
. . . . . . . . . . . . . . Ateles geoffroyi (Central American spide...) ....  129   1 hit   [primates]   HOXD4 [Ateles geoffroyi] >gi|12293486|g
. . . . . . . . . . . . Canis familiaris (dogs) -----------------------------  416  19 hits  [mammals]   PREDICTED: similar to Homeobox protein H
. . . . . . . . . . . . Bos taurus (cow) ..................................  150  10 hits  [mammals]   PREDICTED: hypothetical protein [Bos tau
. . . . . . . . . . . . Ovis aries (domestic sheep) .......................  123   3 hits  [mammals]   Homeobox protein Hox-A7 >gi|14450868|gb|A
. . . . . . . . . . Gallus gallus (bantam) ------------------------------  347  22 hits  [birds]   homeo box B5 [Gallus gallus] >gi|5722548
. . . . . . . . . . Pelodiscus sinensis ...................................  320   1 hit   [turtles]   Hoxb-5 [Pelodiscus sinensis]
. . . . . . . . . . Coturnix japonica .....................................  115   1 hit   [birds]   Homeobox protein Hox-A7 (Qhox-1) >gi|213
. . . . . . . . . Xenopus tropicalis ------------------------------------  243   2 hits  [amphibians]   homeo box A5 [Xenopus tropicalis] >gi|156
. . . . . . . . . Xenopus laevis (clawed frog) ............................  229  17 hits  [amphibians]   Homeobox protein Hox-B5 (X1Hbox-4) (Xhox
. . . . . . . . . Ambystoma mexicanum .....................................  174   1 hit   [amphibians]   Homeobox protein Hox-A5 >gi|7464181|gb|AA
. . . . . . . . . Notophthalmus viridescens (red-spotted newt) ............  129   3 hits  [amphibians]   homeobox protein
. . . . . . . . . Pleurodeles waltl .......................................  116   1 hit   [amphibians]   transcription factor HoxA4 [Pleurodeles
. . . . . . . . Latimeria menadoensis (king of the sea) -----------------  292  11 hits  [coelacanths]   HoxB5 [Latimeria menadoensis]
. . . . . . . Danio rerio (zebra fish) ----------------------------------  327  21 hits  [bony fishes]   homeodomain protein [Danio rerio] >gi|62
. . . . . . . Megalobrama amblycephala ----------------------------------  310   5 hits  [bony fishes]   homeodomain protein [Megalobrama amblyce
. . . . . . . Oryzias latipes (Japanese rice fish) ----------------------  303  15 hits  [bony fishes]   hoxB5b [Oryzias latipes] >gi|8201695?|gb
. . . . . . . Takifugu rubripes (torafugu) ------------------------------  302  13 hits  [bony fishes]   Homeobox protein Hox-B5b >gi|9948280?|gb
. . . . . . . Sphoeroides nephelus ........................................  301   3 hits  [bony fishes]   HoxB5b [Sphoeroides nephelus]
```

BLAST Output – Graphical Overview

NCBI Minicourses

Distribution of 514 Blast Hits on the Query Sequence



BLAST Output - Descriptions

NCBI Minicourses

gi|6680251|ref|NP_032294.1|	homeo box B5 [Mus musculus] >gi|5...	420	4e-116	U G
gi|62656894|ref|XP_573183.1|	PREDICTED: similar to Homeobox p...	419	1e-115	U G
gi|73966238|ref|XP_548176.2|	PREDICTED: similar to Homeobox p...	416	6e-115	G
gi|4504469|ref|NP_002138.1|	homeobox B5 [Homo sapiens] >gi|11...	412	7e-114	U G
gi|12848710|dbj|BAB28059.1|	unnamed protein product [Mus musculu	364	3e-99	U G
gi|70778859|ref|NP_001020526.1|	homeo box B5 [Gallus gallus] ...	347	4e-94	U G
gi|62526079|dbj|BAD95556.1|	Hoxb-5 [Gallus gallus]	337	4e-91	U G
gi|62537|emb|CAA48320.1|	homeodomain protein [Danio rerio] >g...	327	5e-88	G
gi|54261740|ref|NP_571176.2|	homeo box B5a [Danio rerio] >gi|...	325	1e-87	U G
gi|62526089|dbj|BAD95561.1|	Hoxb-5 [Pelodiscus sinensis]	320	6e-86	
gi|60392405|sp|P09013|HXB5B_BRARE	Homeobox protein Hox-B5b (H...	315	1e-84	
gi|53749656|ref|NP_571612.2|	homeo box B5b [Danio rerio] >gi|...	315	1e-84	U G
gi|4322074|gb|AAD15946.1|	homeobox protein [Danio rerio]	312	1e-83	G
gi|123204462|gb|ABM73560.1|	homeodomain protein [Megalobrama amb	310	4e-83	
gi|74267553|dbj|BAE44275.1|	hoxB5b [Oryzias latipes] >gi|8301...	303	5e-81	
gi|119370786|sp|Q1KKX0|HXB5B_FUGRU	Homeobox protein Hox-B5b >...	302	1e-80	
gi|34484445|gb|AAQ72845.1|	Hoxb5b [Sphoeroides nephelus]	301	3e-80	
gi|28629649|gb|AAO43030.1|	HoxB5 [Latimeria menadoensis]	292	1e-77	
gi|62125375|gb|AAX63757.1|	HoxB5bi [Oncorhynchus mykiss]	284	3e-75	
gi|47216999|emb|CAG01627.1|	unnamed protein product [Tetraodon n	260	4e-68	
gi|119370785|sp|Q1KKX9|HXB5A_FUGRU	Homeobox protein Hox-B5a >...	259	1e-67	
gi|60392398|sp|Q6B3N0|HXA5_CHICK	Homeobox protein Hox-A5 >gi|...	253	5e-66	
gi|109472056|ref|XP_001064984.1|	PREDICTED: similar to Homeob...	248	2e-64	U G
gi|15029697|gb|AAH11063.1|	Homeo box A5 [Mus musculus]	248	2e-64	U G

BLAST Output - Statistics

NCBI Minicourses

Bit Score

[*E.g.*: Score = 419 bits (1077)]

- derived from raw alignment score
- normalized, so can be used to compare different searches

Expect value (E)

- <u>number</u> of alignments expected to occur by chance

E is dependent on search space

$$E = mn2^{-S'}$$

BLAST Output - Alignments

NCBI Minic

```
>gi|1708354|sp|P09079|HXB5_MOUSE  Homeobox protein Hox-B5 (Hox-2.1) (MU-1)
 gi|6680251|ref|NP_032294.1|  homeo box B5 [Mus musculus]
 gi|522335|gb|AAA37842.1|  homeobox protein
Length=269

 Score =  419 bits (1077),  Expect = 6e-116
 Identities = 269/269 (100%), Positives = 269/269 (100%), Gaps = 0/269 (0%)

Query  1    MSSYFVNSFSGRYPNGPDyqllnygsgsslsgsyRDPAAMHTGSYGYNYNGMDLSVNRSS  60
            MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAMHTGSYGYNYNGMDLSVNRSS
Sbjct  1    MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAMHTGSYGYNYNGMDLSVNRSS  60

Query  61   ASSSHFGAVGESSRAFPASAKEPRFRQAtsscslsspeslpctNGDSHGAKPSASSPSDQ  120
            ASSSHFGAVGESSRAFPASAKEPRFRQATSSCSLSSPESLPCTNGDSHGAKPSASSPSDQ
Sbjct  61   ASSSHFGAVGESSRAFPASAKEPRFRQATSSCSLSSPESLPCTNGDSHGAKPSASSPSDQ  120

Query  121  ATPASSSANFTEIDeasassepeeaasqlsspslaraqPEPMATSTAAPEGQTPQIFPWM  180
            ATPASSSANFTEIDEASASSEPEEAASQLSSPSLARAQPEPMATSTAAPEGQTPQIFPWM
Sbjct  121  ATPASSSANFTEIDEASASSEPEEAASQLSSPSLARAQPEPMATSTAAPEGQTPQIFPWM  180
```

low complexity

BLAST Output - Alignments

NCBI Minic

```
>gi|62526079|dbj|BAD95556.1|  Hoxb-5 [Gallus gallus]
Length=257

 Score =  294 bits (753),  Expect = 2e-78, Method: Composition-based stats.
 Identities = 204/262 (77%), Positives = 221/262 (84%), Gaps = 5/262 (1%)

Query  1    MSSYFVNSFSGRYPNGPDyqllnygsgsslsgsyRDPAAMHTGSYGYNYNGMDLSVNRSS  60
            MSSYFVNSFSGRYPNGPDYQLLNYG+ SS++GSYRD + MH+ SYGYNYNGMDLS+NRS
Sbjct  1    MSSYFVNSFSGRYPNGPDYQLLNYGTSSS VGSYRDSSTMHSSSYGYNYNGMDLSINRS-  59
```

positives, from score matrix

BLAST

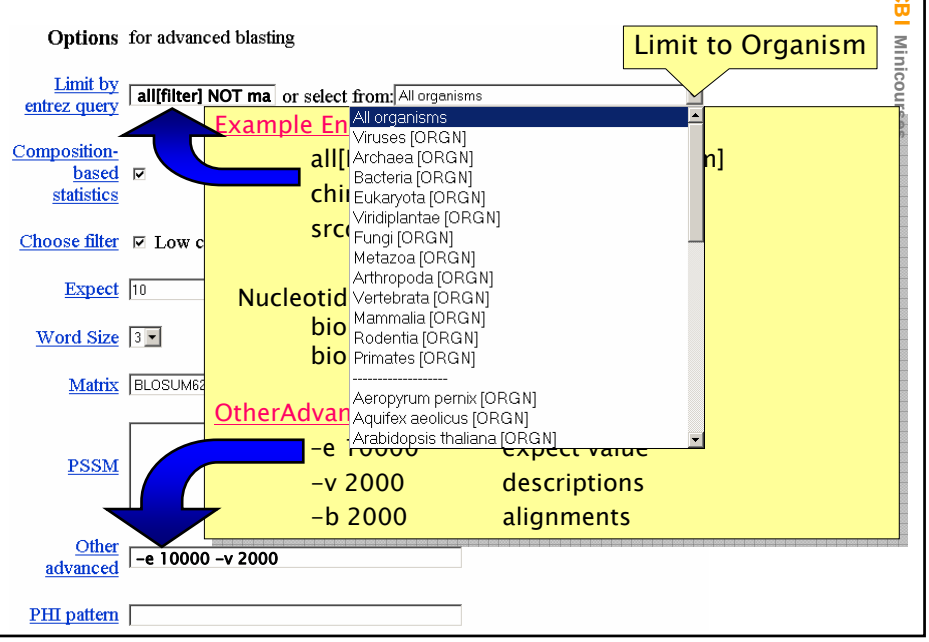
Search Basics

NCBI Minicourses

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Advanced Options

NCBI Minicourses



Other Advanced Options

NCBI Minicourses

-e expect value

-W word size

-v descriptions

-b alignments

For short queries: increase e and lower W

Other advanced | **– e 10000 –W 7 –v 2000 –b 1500**

BLASTP –
"...short, nearly exact matches"

NCBI Minico

Options for advanced blasting

Limit by entrez query | or select from: All organisms

Compositional adjustments | No adjustment

Choose filter | ☐ Low complexity ☐ Mask for lookup table only ☐ Mask lower case

Expect | 20000

Word Size | 2

Matrix | PAM30 | Gap Costs | Existence: 9 Extension: 1

Scoring Matrices

Nucleotide search: identity matrix

	A	T	C	G
A	1	-3	-3	-3
T	-3	1	-3	-3
C	-3	-3	1	-3
G	-3	-3	-3	1

Scoring Matrices

Protein search: substitution matrix

- Percent Accepted Mutation (PAM)
- Blocks Substitution Matrix (BLOSUM)

NCBI Minicourses

PAM Matrices

NCBI Minicourses

- global alignments of closely related proteins
- PAM1 calculated from sequences with no more than 1% divergence
- Other PAM matrices are extrapolated from PAM1
- Examples: PAM30 and PAM70

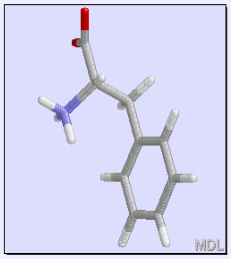
BLOSUM Matrices

NCBI Minicourses

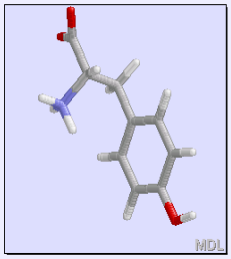
- local alignments
- all matrices based on observed alignments; not extrapolated
- BLOSUM 62 calculated from sequences with no more than 62% identity
- Examples: BLOSUM45, BLOSUM62 and BLOSUM80
- BLOSUM62: very good for detecting weak protein similarities
- BLOSUM62 is the default matrix in BLAST

Scoring Systems - Proteins (BLOSUM62)

L-phenylalanine (F)



L-tyrosine (Y)



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	-2	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	[illegible]	[illegible]	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	-1	-1	-2	-1	1	5				
W	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	[illegible]	[illegible]	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-1	1	-2	-2	0	-3	-1	4	
X	0	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	-2	0	0	-2	-1	-1	-1

Negative for less likely substitutions

Positive for more likely substitutions

Substitution Matrices

A Provisional Table

NCBI Minicourses

Query Length	Substitution Matrix
<35	PAM30
35-50	PAM70
50-85	BLOSUM80
>85	BLOSUM62

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

Less divergent ⟷ *More divergent*

BLAST

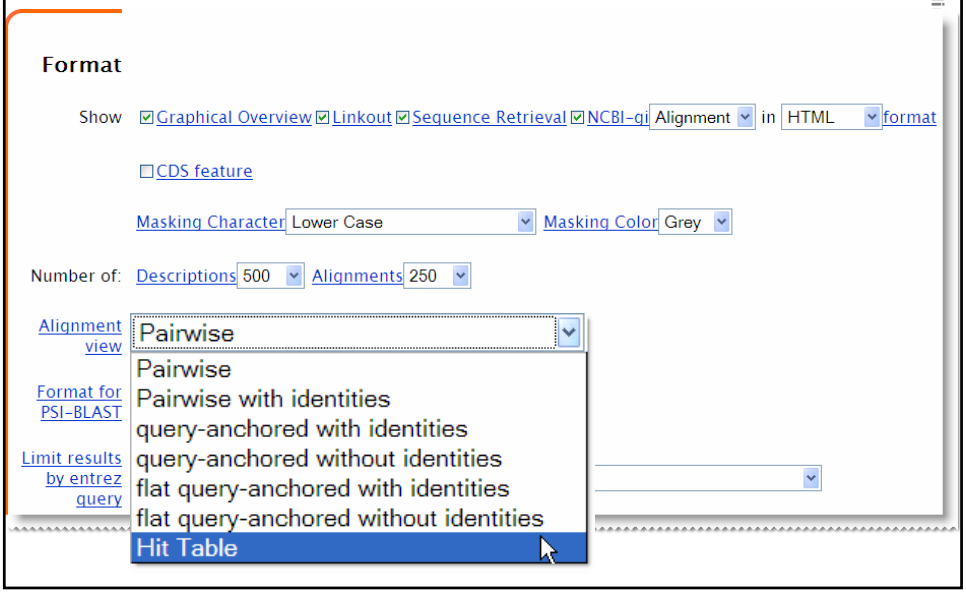
Search Basics

NCBI Minicourses

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Formatting Options

NCBI Mi



Flat Query-anchored with Identities

NCBI Mi

Get selected sequences | Select all | Deselect all | Tree View

```
Query      1     MSSYFVNSF-SGRYPN-GPDyqllnygsgsslsgsy--RDP--AAMH---TGSY-----G  46
1708354    1     .........-......-...................--...--....---....-----.  46
74225740   1     .........-......-...................--...--....---....-----.  46
84579185   1     .........-......-...................--...--....---....-----.  46
11138933   1     .........-......-...................--...--....---....-----.  46
12848710   1                                                  ..---....-----.  7
76644496   1     .........-......-...................--...--....---....-----.  46
57235483   1     .........-......-..........TS..MN...--..S--ST..---SS..-----.  46
62526079   1     .........-......-..........TS..MN...--..S--ST..---SS..-----.  46
62526089   1     .........-......-..E.......TS.AMN...--..S--GT..---SS..-----A  46
28629649   1                                                  ..---SS..-----.  7
54261740   1     .........-......-..........TS..AMNASY-..S--GT..---S...-----.  47
62537      1     .........-......-..........TS..AMNASY-..S--GT..---S...-----.  47
53749656   1     .....L...-......-.S........TNGAMNA..--..S--TS..---S..F-----.  46
22316136   1     .....L...-......-.S........TNGAMNA..--..S--TS..---S..F-----.  46
4322074    1     .....L...-......-.S........TNGAMNA..--..S--TS..---S..F-----.  46
74267553   1     .........-......-.S........TNGAV...F--...--GT..---S..F-----.  46
34484445   1     .........-......-.S........TNGAV...F--...--GT..---S..F-----.  46
50956652   1     .........-C.....-......H...DH..V.EQ.--..S--.S..---S.R.-----.  46
32879831   1     .........-C.....-......H...DH..V.EQF--..S--.S..---S.R.-----.  46
32879829   1     .........-C.....-......H...DH..V.EQF--..S--.S..---S.R.-----.  46
```

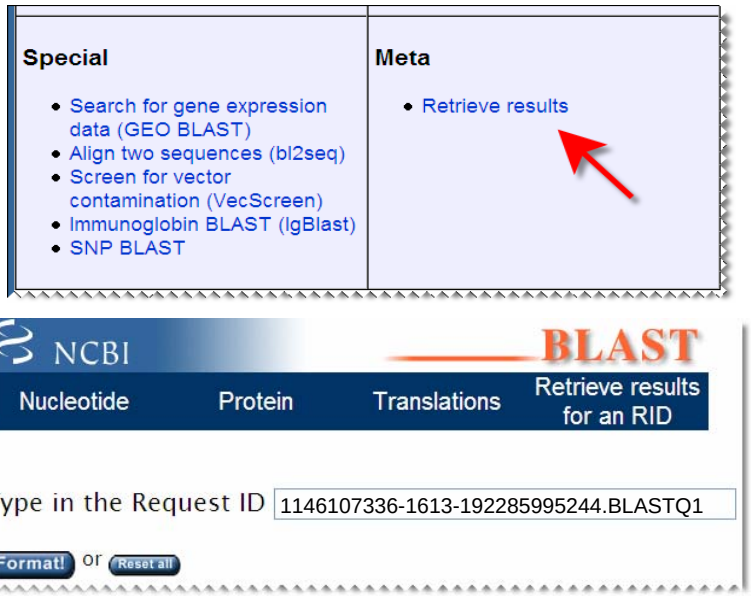
HitTable

NCBI Minicourses

```
# BLASTP 2.2.13 [Nov-27-2005]
# Query: gi|6680251|ref|NP_032294.1| homeo box B5 [Mus musculus]
# Database: nr
# Fields: query id, subject ids, % identity, % positives, alignment length, mismatches, gap
# 514 hits found
gi|6680251|ref|NP_032294.1|     gi|1708354|sp|P09079|HXB5_MOUSE;gi|6680251|ref|NP_032294.1|;
gi|6680251|ref|NP_032294.1|     gi|74225740|dbj|BAE21695.1|;gi|73695452|gb|AAI03605.1|;gi|73
gi|6680251|ref|NP_032294.1|     gi|84579185|dbj|BAE73026.1|;gi|73966238|ref|XP_548176.2|
gi|6680251|ref|NP_032294.1|     gi|11138933|gb|AAG31553.1|;gi|4504469|ref|NP_002138.1|;gi|40
gi|6680251|ref|NP_032294.1|     gi|12848710|dbj|BAB28059.1|     99.57   100.00  230     1
gi|6680251|ref|NP_032294.1|     gi|76644496|ref|XP_584774.2|     90.18   92.73   275     21
gi|6680251|ref|NP_032294.1|     gi|57235483|gb|AAW48484.1|;gi|70778859|ref|NP_001020526.1|
gi|6680251|ref|NP_032294.1|     gi|62526079|dbj|BAD95556.1|     77.86   84.35   262     53
gi|6680251|ref|NP_032294.1|     gi|62526089|dbj|BAD95561.1|     73.85   80.38   260     63
gi|6680251|ref|NP_032294.1|     gi|28629649|gb|AAO43030.1|     77.83   84.35   230     46
gi|6680251|ref|NP_032294.1|     gi|54261740|ref|NP_571176.2|;gi|47938028|gb|AAH71493.1|;gi|26
gi|6680251|ref|NP_032294.1|     gi|62537|emb|CAA48320.1|;gi|62555|emb|CAA31290.1|      70.14
gi|6680251|ref|NP_032294.1|     gi|15029697|gb|AAH11063.1|     58.93   68.21   280     94
gi|6680251|ref|NP_032294.1|     gi|47216999|emb|CAG01627.1|     55.35   61.64   318     92
```

Fields: query id, subject ids, % identity, % positives, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score

Retrieve Results for an RID



BLAST

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

NCBI Minicourses

Examples

NCBI Minicourses

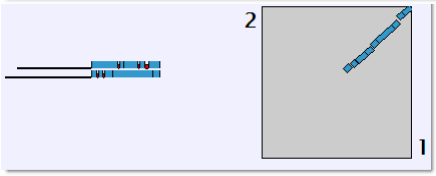
- Protein searches <u>more sensitive</u> than nucleotide searches
  - redundancy of the genetic code allows the same information to be encoded differently
- blastn best when searching within same organism

BLASTN vs BLASTP

NCBI Minicourses

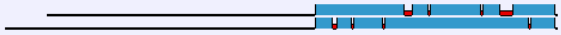
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.14 [Apr-09-2006]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 10.0000 wordsize: 3 Filter View option Standard
Masking character option X for protein, n for nucleotide Masking color option Black
Show CDS translation Align



BLASTP

NOTE: Bitscore and expect value are calculated based on the size of the nr database.

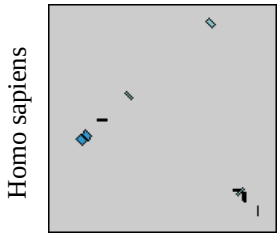


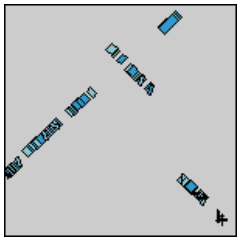
```
Score = 48.5 bits (114),  Expect = 7e-04
Identities = 50/181 (27%), Positives = 80/181 (44%), Gaps = 27/181 (14%)
```

BLASTN vs TBLASTX

Anopheles gambiae mitochondrion, complete genome.
ACCESSION NC_002084 GI:5834911

Homo sapiens mitochondrion, complete genome.
ACCESSION NC_001807 GI:17981852





Additional Questions?

blast-help@ncbi.nlm.nih.gov

NCBI Minicourses

Hands-On

NCBI Minicourses

www.ncbi.nlm.nih.gov/Class/minicourses/quickblast.html