Mini Courses NCBI Resources one course at a time

BLAST Quick Start

Wayne Matten
NCBI

Basic Local Alignment Search Tool

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

- 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

ATGCTGCTAGTCGATGACGTAGCTA

ATGCTGCTAGT
  TGCTGCTAGTC
    GCTGCTAGTCG
      …

Make a lookup table based on the word size.

Protein Words

AIEKCYTGCTLAQEADDTA

AIE
IEK LEK, IDK, IQK, IER, IDR, etc
EKC
KCY
CYT

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

Finding the Word Hits

Nucleotide: one exact match

ATGCTGCTAGTCGATGACGTAGCTA
⇦ GCTGCTAGTCG ⇨

Protein: two matches within 40 residues

AIEKCYTGCTLAQEADDTA
⇦IDK ⇨ ⇦ EAD ⇨

Search Basics

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

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

More BLAST Programs

	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)	

Search Basics

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

Select 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)

HTGS

Phase 0	single-few 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

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

Select BLAST Databases – Protein

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

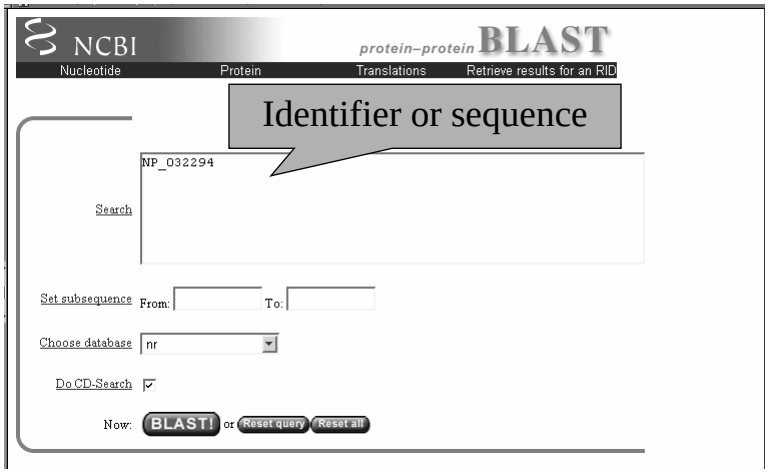
Select BLAST Databases – Genome-specific

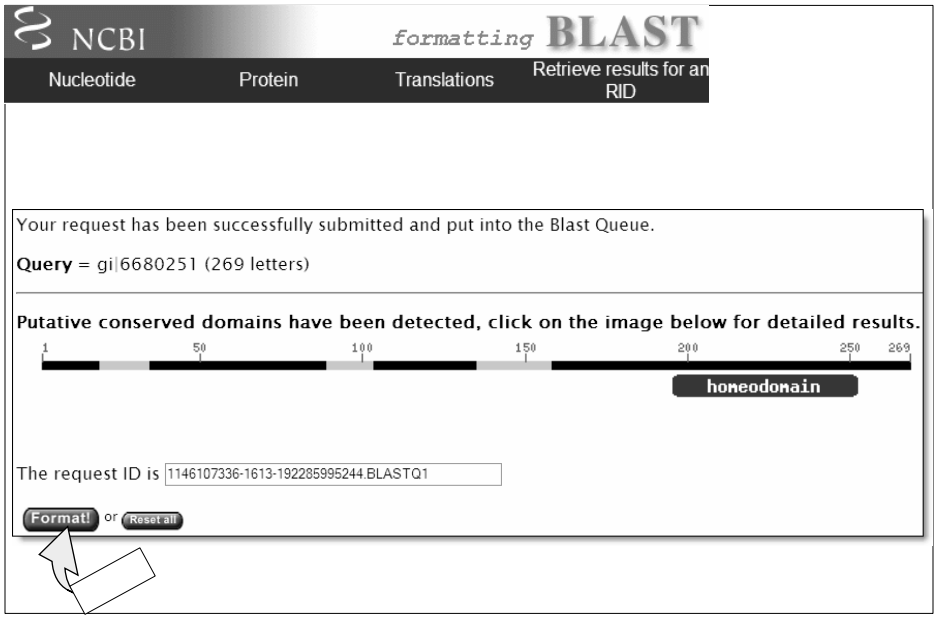
- Human Genome
- Mouse Genome
- Rat Genome
- Other eukaryotes
- Microbial Genomes
- Arabidopsis thaliana
- Insects
- Plants

Search Basics

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

Setting up a BLAST Search





Search Basics

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

BLAST Output

NCBI *results of* **BLAST**

BLASTP 2.2.13 [Nov-27-2005]

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.

RID: 1146107336-1613-192285995244.BLASTQ1

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples
3,605,113 sequences; 1,238,276,235 total letters

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**
Taxonomy reports

Index

Tax BLAST Report

- 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]
. . . . . . . . . . . . . Murinae          [mammals]
. . . . . . . . . . . . . . Mus              [mammals]
. . . . . . . . . . . . . . . Mus musculus (mouse) ----------------  419 46 hits [mammals]   Homeobox protein Hox-B5 (Hox-2.1) (MU-1) (
. . . . . . . . . . . . . . . Mus sp. (mice) ......................  148  4 hits [mammals]   Hoxa-4 [Mus sp.] >gi|8182211|gb|AAB28662.2
. . . . . . . . . . . . . . . Rattus norvegicus (brown rat) ---------  195 11 hits [mammals]   Homeobox protein Hox-A5 (Hox-1.3) >gi|2046
. . . . . . . . . . . . . . . Rattus sp. (rats) ....................  117  2 hits [mammals]   Hox-B|Hox-2 [Rattus sp.]
. . . . . . . . . . . . . . . Macaca fascicularis (cynomolgus monkey) -  416  1 hit  [mammals]   hypothetical protein [Macaca fascicularis]
. . . . . . . . . . . . . . . Homo sapiens (man) ......................  412 43 hits [mammals]   homeobox B5 [Homo sapiens] >gi|4504469|ref
. . . . . . . . . . . . . . . Pan troglodytes .........................  240  9 hits [mammals]   PREDICTED: similar to homeobox protein A7;
. . . . . . . . . . . . . . Bos taurus (cow) -------------------------  365 17 hits [mammals]   PREDICTED: similar to Homeobox protein Hox
. . . . . . . . . . . . . . Canis familiaris (dogs) ..................  257 15 hits [mammals]   PREDICTED: similar to Homeobox protein Hox
. . . . . . . . . . . . . . Ovis aries (domestic sheep) ..............  124  2 hits [mammals]   Homeobox protein Hox-A7 >gi|1458088|gb|AAB
. . . . . . . . . . . . . Gallus gallus (chickens) -------------------  353 23 hits [birds]     homeodomain transcription factor [Gallus g
. . . . . . . . . . . . . Pelodiscus sinensis .........................  316  2 hits [turtles]   Hoxb-5 [Pelodiscus sinensis]
. . . . . . . . . . . . . Coturnix japonica ...........................  132  1 hit  [birds]     Homeobox protein Hox-A7 (Quox-1) >gi|21361
. . . . . . . . . . . . Xenopus tropicalis ----------------------------  249  3 hits [amphibians]  Homeo box A5 [Xenopus tropicalis] >gi|5833
. . . . . . . . . . . . Xenopus laevis (clawed frog) ..................  221 18 hits [amphibians]  unnamed protein product [Xenopus laevis] >
. . . . . . . . . . . . Ambystoma mexicanum ...........................  185  2 hits [amphibians]  homeobox protein [Ambystoma mexicanum] >gi
. . . . . . . . . . . . Pleurodeles waltl ..............................  125  1 hit  [amphibians]  transcription factor HoxA4 [Pleurodeles wa
. . . . . . . . . . . . Notophthalmus viridescens (red-spotted newt) ..  125  4 hits [amphibians]  Homeobox protein Hox-C5 (NvHbox-3.4)
. . . . . . . . . . . Latimeria menadoensis (king of the sea) ---------  312 10 hits [coelacanths]  HoxB5 [Latimeria menadoensis]
```

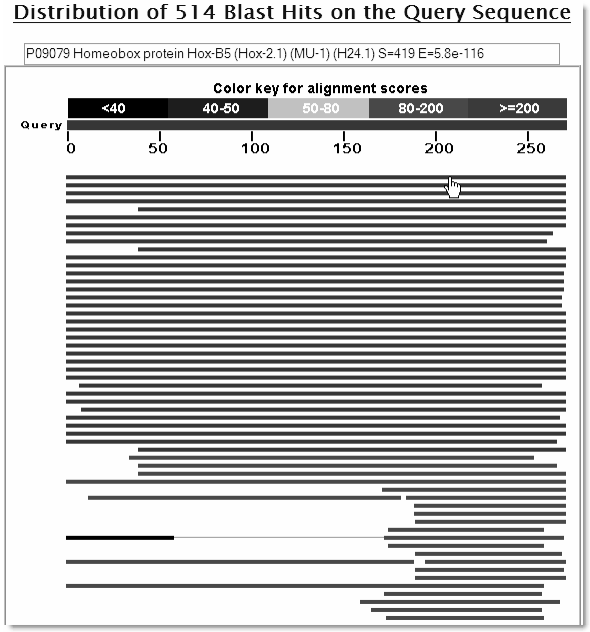
Index

Tax BLAST Report

- Lineage Report
- Organism Report

```
. . . . . . Mus musculus (mouse) ----------------  419 46 hits [mammals]
. . . . . . Mus sp. (mice) ......................  148  4 hits [mammals]
. . . . . Rattus norvegicus (brown rat) ---------  195 11 hits [mammals]
. . . . . Rattus sp. (rats) .....................  117  2 hits [mammals]
. . . . Macaca fascicularis (cynomolgus monkey) -  416  1 hit  [mammals]
. . . . Homo sapiens (man) ......................  412 43 hits [mammals]
. . . . Pan troglodytes .........................  240  9 hits [mammals]
. . . Bos taurus (cow) -------------------------  365 17 hits [mammals]
. . . Canis familiaris (dogs) ...................  257 15 hits [mammals]
. . . Ovis aries (domestic sheep) ...............  124  2 hits [mammals]
. . Gallus gallus (chickens) -------------------  353 23 hits [birds]
. . Pelodiscus sinensis .........................  316  2 hits [turtles]
. . Coturnix japonica ...........................  132  1 hit  [birds]
. Xenopus tropicalis ----------------------------  249  3 hits [amphibians]
. Xenopus laevis (clawed frog) ..................  221 18 hits [amphibians]
. Ambystoma mexicanum ...........................  185  2 hits [amphibians]
. Pleurodeles waltl .............................  125  1 hit  [amphibians]
. Notophthalmus viridescens (red-spotted newt) ..  125  4 hits [amphibians]
Latimeria menadoensis (king of the sea) ---------  312 10 hits [coelacanths]
```

BLAST Output – Graphical Overview



BLAST Output - Descriptions

Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
gi|1708354|sp|P09079|HXB5_MOUSE Homeobox protein Hox-B5 (Hox-...	419	6e-116	G
gi|74225740|dbj|BAE21695.1| unnamed protein product [Mus musc...	417	2e-115	G
gi|84579185|dbj|BAE73026.1| hypothetical protein [Macaca fasc...	416	6e-115	
gi|11138933|gb|AAG31553.1| homeobox B5 [Homo sapiens] >gi|450...	412	5e-114	G
gi|12848710|dbj|BAB28059.1| unnamed protein product [Mus musculu	375	1e-102	G
gi|76644496|ref|XP_584774.2| PREDICTED: similar to Homeobox p...	365	8e-100	G
gi|57235483|gb|AAW48484.1| homeodomain transcription factor [...	353	5e-96	G
gi|62526079|dbj|BAD95556.1| Hoxb-5 [Gallus gallus]	339	8e-92	G
gi|62526089|dbj|BAD95561.1| Hoxb-5 [Pelodiscus sinensis]	316	7e-85	
gi|28629649|gb|AAO43030.1| HoxB5 [Latimeria menadoensis]	312	1e-83	
gi|54261740|ref|NP_571176.2| homeo box B5a [Danio rerio] >gi|...	310	4e-83	G
gi|62537|emb|CAA48320.1| homeodomain protein [Danio rerio] >g...	306	4e-82	G
gi|53749656|ref|NP_571612.2| homeo box B5b [Danio rerio] >gi|...	297	3e-79	G
gi|22316136|emb|CAD44458.1| homeo box protein B5b [Danio reri...	297	3e-79	G
gi|4322074|gb|AAD15946.1| homeobox protein [Danio rerio]	294	3e-78	G
gi|74267553|dbj|BAE44275.1| hoxB5b [Oryzias latipes] >gi|8301...	279	7e-74	
gi|34484445|gb|AAQ72845.1| Hoxb5b [Sphoeroides nephelus]	277	3e-73	
gi|50956652|gb|AAT90845.1| homeodomain transcription factor [...	261	2e-68	
gi|32879831|gb|AAP88746.1| homeo box A5 [Homo sapiens] >gi|24...	260	3e-68	G
gi|32879829|gb|AAP88745.1| homeo box A5 [synthetic construct]...	260	3e-68	
gi|387668|gb|AAA58663.1| HOX 1.3	259	8e-68	G

BLAST Output - Statistics

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) [*E.g.*: Expect = 6e-116]

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

E = mn2-S’

BLAST Output - Alignments

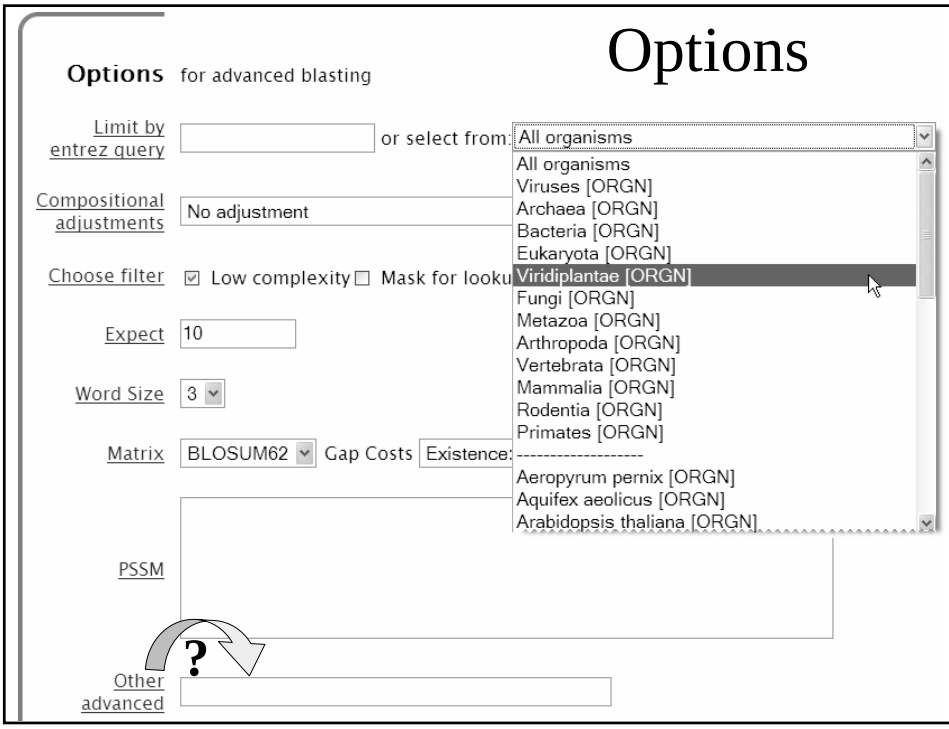
```
>gi|1708354|sp|P09079|HXB5_MOUSE  Homeobox protein Hox-B5 (Hox-2.1) (MU-1)
 gi|6680251|ref|NP_032294.1|  homeo box B5 [Mus musculus]
 gi|522335|gb|AAA37842.1|  homeobox protein
Length=269

 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

Search Basics

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

Options



Other Advanced Options

-e expect value

-W word size

-v descriptions

-b alignments

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

For short queries: increase e and lower W

BLASTP –
"…short, nearly exact matches"

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)

PAM Matrices

- 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

- local alignments
- all matrices based on observed alignments; not extrapolated
- BLOSUM 62 calculated from sequences with no more than 62% identity
- Examples: BOLSUM45, BLOSUM62 and BLOSUM80
- BLOSUM62: very good for detecting weak protein similarities
- BLOSUM62 is the default matrix in BLAST

BLOSUM62

	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

frequently observed substitution

rare residue

Substitution Matrices

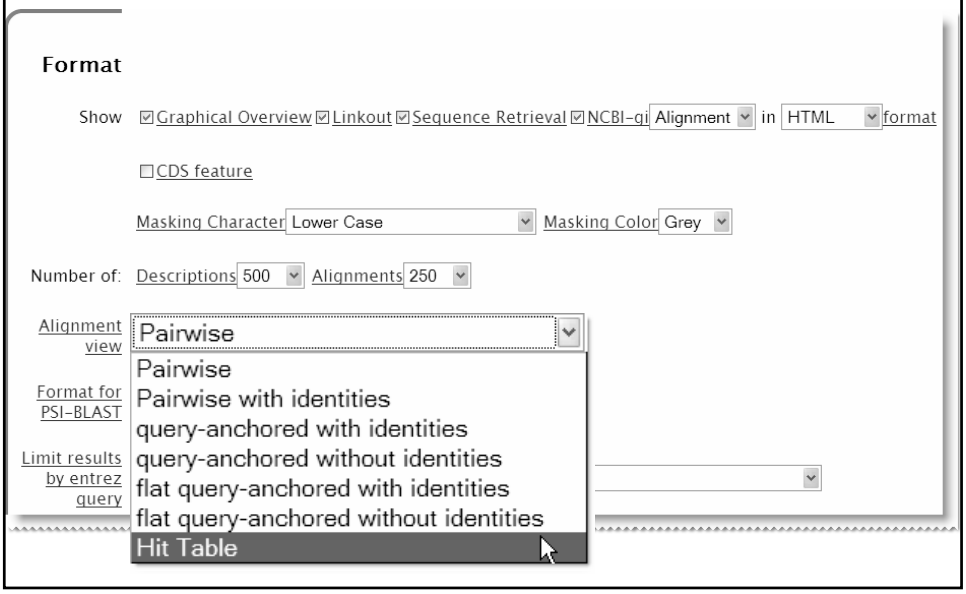
A Provisional Table

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

Search Basics

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

Formatting Options



Flat Query-anchored with Identities

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
```

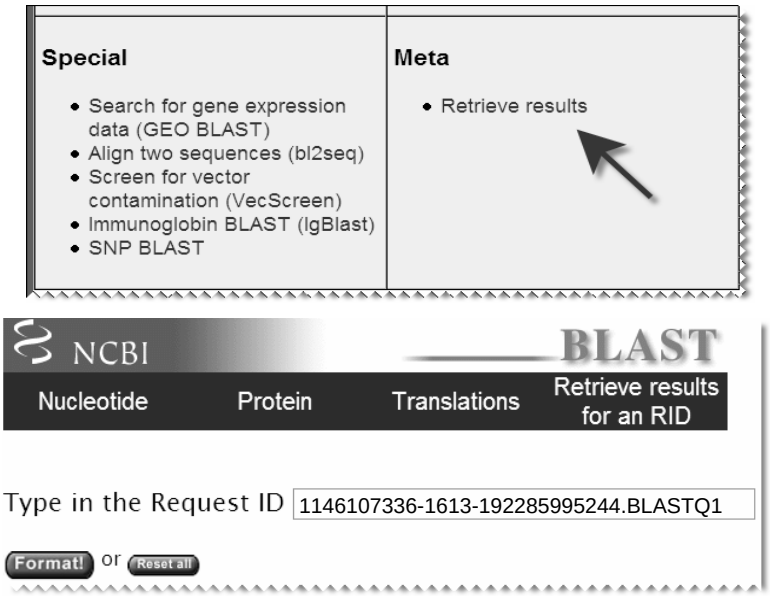
HitTable

```
# 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 o
# 514 hits found
gi|6680251|ref|NP_032294.1|     gi|1708354|sp|P09079|HXB5_MOUSE;gi|6680251|ref|NP_032294.1|;g
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
```

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

```
gi|6680251|ref|NP_032294.1|     gi|32879831|gb|AAP88746.1|;gi|24497517|ref|NP_061975.2|;gi|49
gi|6680251|ref|NP_032294.1|     gi|32879829|gb|AAP88745.1|;gi|61370212|gb|AAX43455.1|;gi|6137
gi|6680251|ref|NP_032294.1|     gi|387668|gb|AAA58663.1|        59.29   69.29   280     93
gi|6680251|ref|NP_032294.1|     gi|49457548|emb|CAG47073.1|     59.29   68.93   280     93
gi|6680251|ref|NP_032294.1|     gi|51372|emb|CAA68364.1|;gi|51471|emb|CAA34738.1|;gi|6754232
gi|6680251|ref|NP_032294.1|     gi|73976525|ref|XP_539487.2|;gi|87578311|gb|AAI13285.1| 58.92
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
```

Retrieve Results for an RID



Search Basics

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

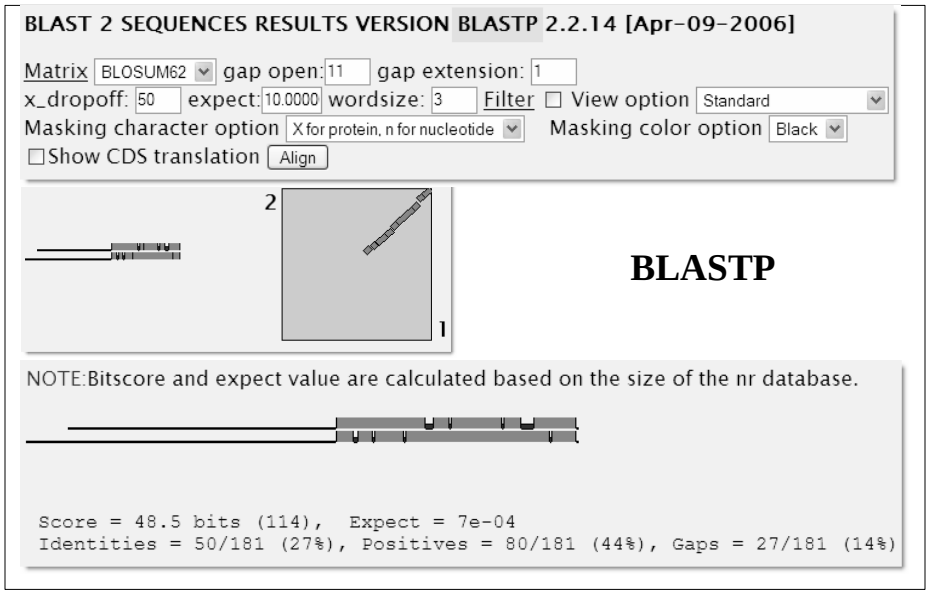
Examples

- 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



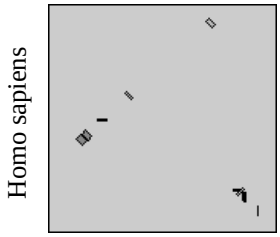
BLASTP

BLASTN vs TBLASTX

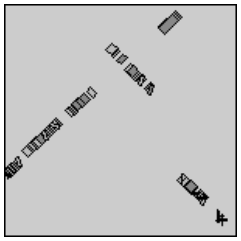
Anopheles gambiae mitochondrion, complete genome.
ACCESSION NC_002084 GI:5834911

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

blastn



tblastx



Hands-On

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