Mini Courses NCBI Resources one course at a time

NCBI Minicourses

BLAST Quick Start

blast-help@ncbi.nlm.nih.gov

BLAST Quick Start

NCBI Minicourses

Introduction

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 or nucleic acid sequences to protein or nucleic acid databases

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

  (BLAST 2 Sequences, or pairwise BLAST)

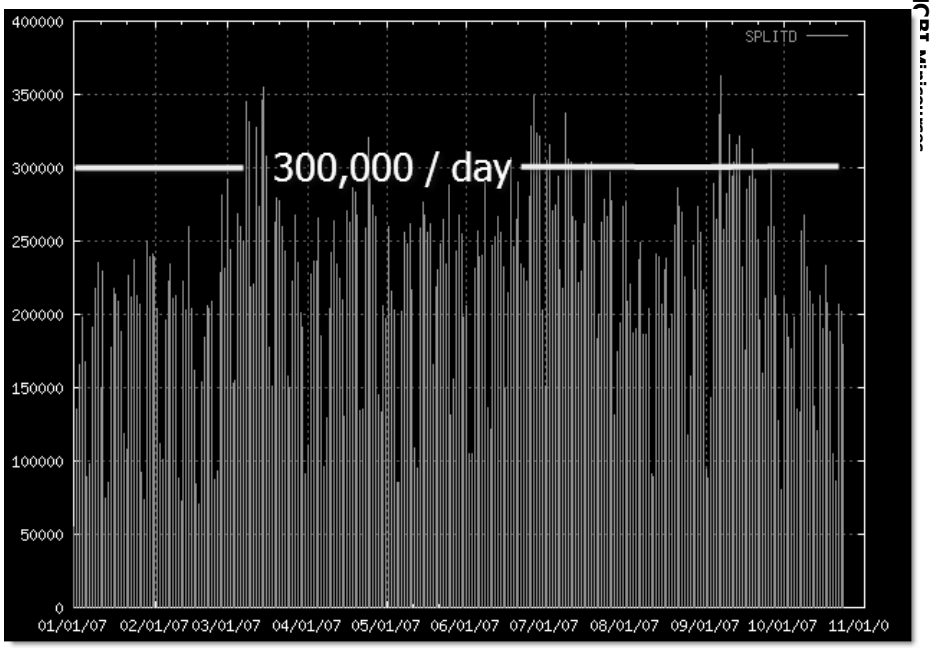
Why do we need similarity searching?

NCBI Minicourses

- To identify and annotate sequences with…
  - incomplete, incorrect, or absent annotations
- To assemble genomes
- To explore evolutionary relationships by…
  - finding homologous molecules
  - developing phylogenetic trees

NOTE: Similar sequences may NOT have similar function!

BLAST Web Searches, 2007



NCBI Minicourses

Basic Local Alignment Search Tool

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

NCBI Minicourses

Global vs Local Alignment

NCBI Minicourses

How BLAST Works

NCBI Minicourses

1. Make lookup table of “words” for query
2. Scan database for hits
3. Extend alignment both directions

Nucleotide Words

NCBI Minicourses

Make a lookup table based on the word size.

11-mer

ATGCTGCTAGTCGATGACGTAGCTA

ATGCTGCTAGT
TGCTGCTAGTC
GCTGCTAGTCG
. . .

Protein Words

NCBI Minicourses

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.

Neighborhood Words (BLOSUM62)

NCBI Minicourses

IEK: keep LEK?
(threshold = 11)

LEK = 12

E / E

L / I

K / K

	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

Word Hits & Extensions

NCBI Minicourses

Nucleotide: one exact match

ATGCTGCTAGTCGATGACGTAGCTA

⇦ TAGTCGATGA ⇨

Protein: two matches within 40 residues

PHAIEKCYTGCTLAQEADDTA

⇦ IDK ⇨ ⇦ EAD ⇨

BLASTP Summary

NCBI Minicourses

example query words

```
Query:   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILEV…
```

YLS	15		HFL	18
YLT	12		HFV	15
YVS	12	Neighborhood words	HFS	14
YIT	10		HWL	13
etc	…		NFL	13
			DFL	12
			HWV	10
			etc	…

Neighborhood score threshold

T (-f) =11

```
Query   1 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESI  47
          +E  YA YL K    F+YLSL +SP+ +DVNVHP+K  VHFL+++ I
Sbjct 287 LEETYAKYLHKGASYFVYLSLNMSPEQLDVNVHPSKRIVHFLYDQEI 333
```

Drop-off score =

Highest score – current score

-X X dropoff value for gapped alignment (in bits)
blastn 30, megablast 20, tblastx 0, all others 15

BLASTP Summary (gapped extension)

NCBI Minicourses

```
Query   1 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESI  47
          +E  YA YL K    F+YLSL +SP+ +DVNVHP+K  VHFL+++ I
Sbjct 287 LEETYAKYLHKGASYFVYLSLNMSPEQLDVNVHPSKRIVHFLYDQEI 333
```

High-scoring pair (HSP)

Gapped extension with trace back

```
Query   1 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESI-LEV…  50
          +E  YA YL K    F+YLSL +SP+ +DVNVHP+K  VHFL+++ I + +
Sbjct 287 LEETYAKYLHKGASYFVYLSLNMSPEQLDVNVHPSKRIVHFLYDQEIATSI… 337
```

Final HSP

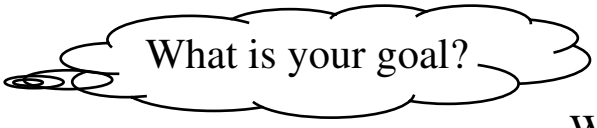
Search Basics

NCBI Minicourses

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

BLAST Programs

NCBI Minicourses



		Word Size
blastn	nucleotide X nucleotide	11
blastp	protein X protein	3

6 frame, translated nucleotide searches

blastx	nucleotide X protein	3
tblastx	nucleotide X nucleotide	3
tblastn	protein X nucleotide	3

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

Search Basics

NCBI Minicourses

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

Nucleotide BLAST Databases

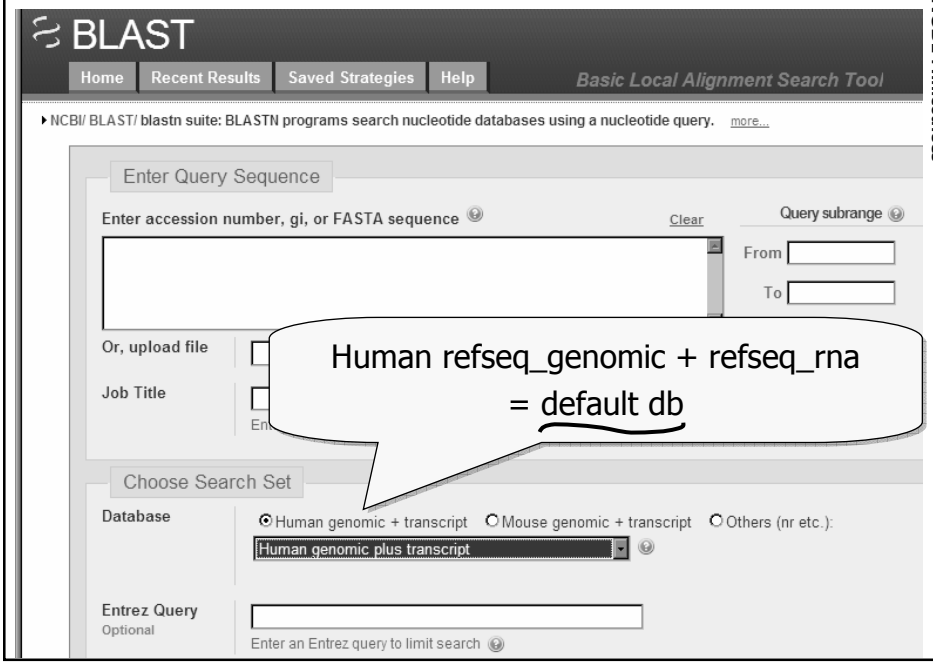
NCBI Minicourses

- **nr (nt)**
  - Traditional GenBank Divisions
  - NM_ and XM_ RefSeqs
- **refseq_rna**
  - NM_ , XM_ , NR_
- **refseq_genomic**
  - NC_ , NT_ , NG_
- **est**
  - EST Division
- **htgs**
  - HTG division
- **dbsts**
  - STS Division
- **chromosome**
  - NC genomic records
- **gss**
  - GSS division
- **wgs**
  - wgs entries from traditional divisions
- **pdb**
  - Nucleotide sequences from structures
- **env_nt**
  - environmental samples

combined, for human and mouse

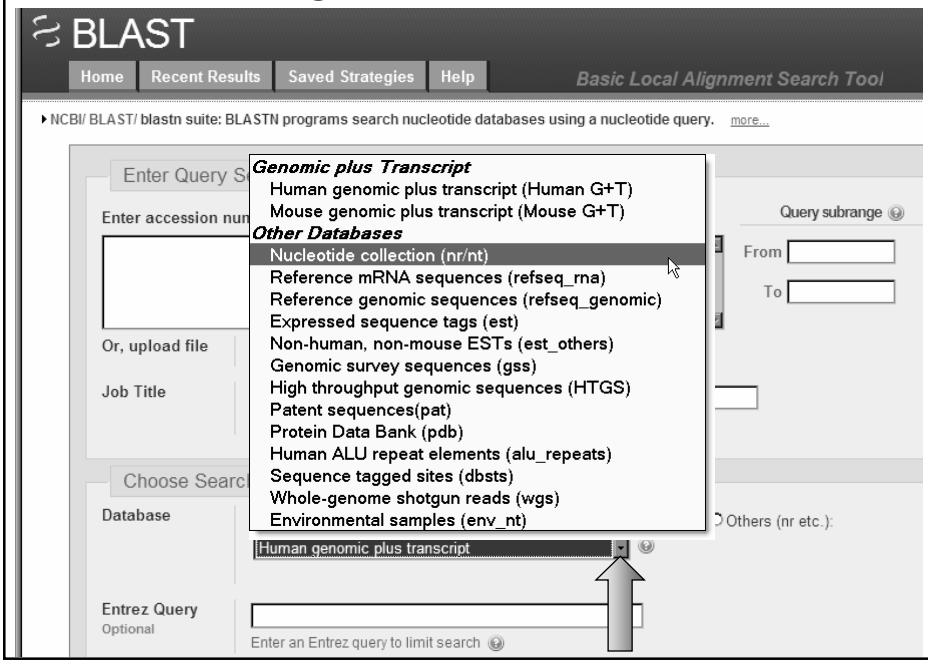
New Nucleotide Databases

NCBI Minicourses



Human refseq_genomic + refseq_rna
= default db

Change Nucleotide Database



NCBI Minicourses

Protein BLAST Databases

Protein

- nr

  traditional GenBank records
- refseq = NP_, XP_
- swissprot
- pdb
- pat
- env_nr

nr ≠ nr

NCBI Minicourses

BLAST Databases: Genome-specific

NCBI Minicourses

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast

BLAST

My NCBI [Sign In] [Register]

Home | Recent Results | Saved Strategies | Help

Basic Local Alignment and Search Tool

Save your searches

What else is new?

NCBI/ BLAST Hom

BLAST fi etween biological sequences. more...

BLAST

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabadopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Escherichia coli*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-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** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search trace archives with within species or among species

News

BLAST 2.2.13 now available
The new release includes a new engine for blastall, changes to statistical parameters, and bug fixes.
2005-12-06 11:55:00

BLAST 2.2.14 now available
BLAST release 2.2.14 offers a universal binary for Mac OS X, and improved performance on some platforms.
2006-06-07 12:05:00

More BLAST news...

Tip of the Day

Checking PCR Primers

A frequent use of nucleotide-nucleotide BLAST is to check the specificity of oligonucleotides for hybridization in PCR. The goal is usually to make sure that the primers will give a unique product from the target genome or cDNA population. Because BLAST

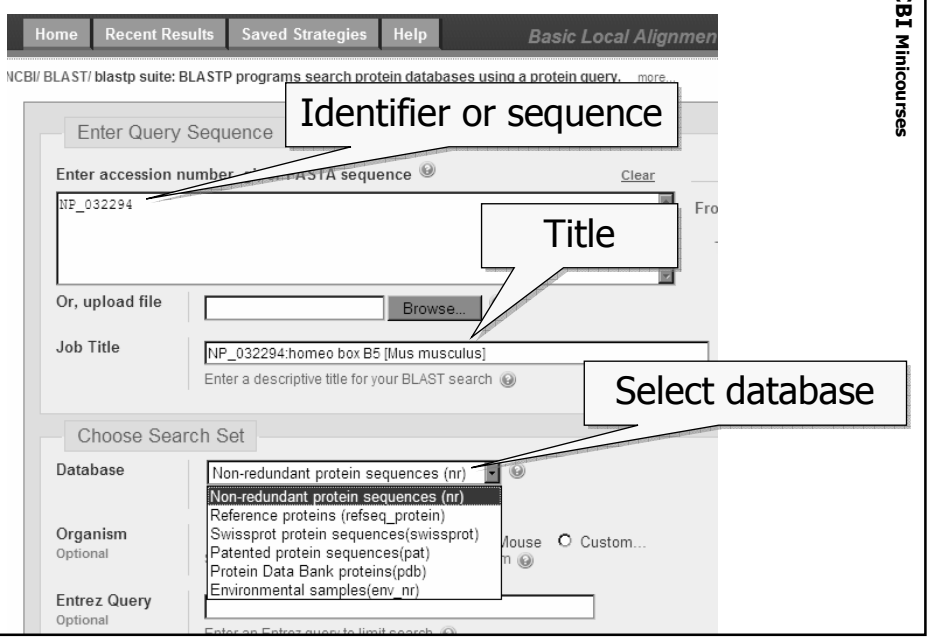
Search Basics

NCBI Minicourses

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

Setting up a BLAST Search

NCBI Minicourses



BLAST

Home | Recent Results | Saved Strategies | Help

Basic Local Alignment Search Tool

NCBI/ BLAST/ Formatting Results - 41Z6F6C012 [Formatting options]

Job Title: NP_032294:homeo box B5 [Mus musculus]

Putative conserved domains have been detected, click on the image below for detailed results.

1 50 100 150 200 250 269

homeodomain

WAITING

Request ID	41Z6F6C012
Status	Searching
Submitted at	Thu Mar 22 00:36:54 2007
Current time	Tue May 8 18:42:35 2007
Time since submission	00:00:11

Conserved Domain search run for protein queries

Default Output

NCBI Minicourses

BLAST

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp/ Formatting Results - 6B1GYW57014 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your searc

Job Title: NP_032294:homeo box B5 [Mus musculus] Show Co

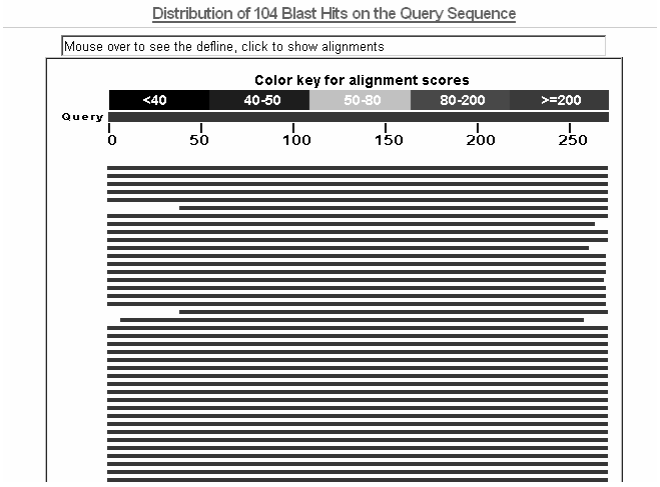
```
BLASTP 2.2.16 (Mar-25-2007)

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: 6B1GYW57014


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
          5,002,253 sequences; 1,729,200,953 total letters


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


Query= gi|6680251|ref|NP_032294.1| homeo box B5 [Mus musculus]
Length=269
```

header

Default Output con't

NCBI Minicourses

Graphical Overview



Default Output con't

NCBI Minicourses

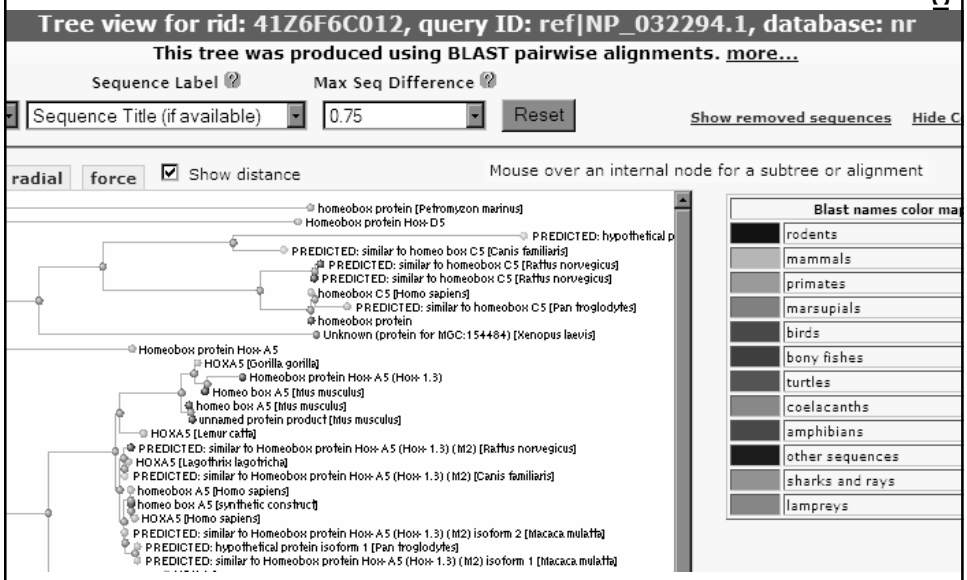
One line descriptions (default view)

Distance tree of results NEW

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|NP 032294.2| homeo box B5 [Mus musculus] >gi|62656894|ref...	553	4e-156	G
gb|AAA37842.1| homeobox protein	551	1e-155	U G
ref|XP 548176.2| PREDICTED: similar to Homeobox protein Hox-B...	548	1e-154	U G
ref|XP 001502124.1| PREDICTED: hypothetical protein [Equus cabal	546	4e-154	G
ref|NP 002138.1| homeobox B5 [Homo sapiens] >ref|XP_001173004...	546	7e-154	U G
dbj|BAB28059.1| unnamed protein product [Mus musculus]	474	3e-132	U G
ref|XP 001367145.1| PREDICTED: similar to homeobox protein [Mono	462	8e-129	G
ref|NP 001020526.1| homeo box B5 [Gallus gallus] >gb|AAW48484...	431	3e-119	U G
dbj|BAD95556.1| Hoxb-5 [Gallus gallus]	417	4e-115	U G
ref|NP 571176.2| homeo box B5a [Danio rerio] >sp|P09014|HXB5A...	394	3e-108	U G
emb|CAA48320.1| homeodomain protein [Danio rerio] >emb|CAA312...	392	1e-107	G
dbj|BAD95561.1| Hoxb-5 [Pelodiscus sinensis]	385	2e-105	

BLAST Output – tree view

NCI



Alignment View - pairwise

NCBI Minicourses

```
> emb|CAA48320.1| G homeodomain protein [Danio rerio]
 emb|CAA31290.1| G ZF-21 gene product [Danio rerio]
Length=275

 Score =  327 bits (837),  Expect = 5e-88, Method: Composition-based stats.
 Identities = 200/278 (71%), Positives = 227/278 (81%), Gaps = 12/278 (4%)

Query  1    MSSYFVNSFSGRYPNGPDYQLLNYGSGSS-LSGSYRDPAAMHTGSYGYNYNGMDLSVNRS  59
            MSSYFVNSFSGRYPNGPDYQLLNYG+ SS ++ SYRD   MH+GSYGYNYNGMDLSVNRS
Sbjct  1    MSSYFVNSFSGRYPNGPDYQLLNYGTSSSAMNASYRDSGTMHSGSYGYNYNGMDLSVNRS  60

Query  60   SASSSHFGAVGESSRAFPASAKEPRFRQATSSCSLSSPESLPCTNGDSHGAKPSASSPSD  119
            + S+ HFGAVG++SR F + A E RFRQ  SSCSL+SPE LPC+N +S G +    +PSD
Sbjct  61   T-STGHFGAVGDNSRVFQSPAPETRFRQP-SSCSLASPEPLPCSNSESFGTQ-RLFAPSD  117

Query  120  QATPAS-----SSANFTEIDEASASSEPEEAASQLSSPSLARAQPEPMATSTAAP---EG  171
            Q+T  +     S+ +FTEIDEASASSE EEA+ + ++ +    Q +    ++      +G
Sbjct  118  QSTTTAGNNLNSNTHFTEIDEASASSETEEASHRANNSAPRTQQKQETTATSTTSATSDG  177

Query  172  QTPQIFPWMRKLHISHDMTGPDGKRARTAYTRYQTLELEKEFHFNRYLTRRRRIEIAHAL  231
            Q PQIFPWMRKLHISHDMTGPDGKRARTAYTRYQTLELEKEFHFNRYLTRRRRIEIAHAL
Sbjct  178  QAPQIFPWMRKLHISHDMTGPDGKRARTAYTRYQTLELEKEFHFNRYLTRRRRIEIAHAL  237

Query  232  CLSERQIKIWFQNRRMKWKKDNKLKSMSLATAGSAFQP  269
            CLSERQIKIWFQNRRMKWKKDNKLKSMSLATAGSAFQP
Sbjct  238  CLSERQIKIWFQNRRMKWKKDNKLKSMSLATAGSAFQP  275
```

Where do Scores Come From?

NCBI Minicourses

One line descriptions
(Advanced view)

Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query average	E value	Links
NP_032294.2	homeo box B5 [Mus musculus] >ref|XP_573183.1| PREDICTED: similar to	553	553	.00%	4e-156	G
AAA37842.1	homeobox protein	551	551	.00%	1e-155	U G
XP_548176.2	PREDICTED: similar to Homeobox protein Hox-B5 (Hox-2A) (HHO.C10) (H	548	548	.00%	1e-154	U G
XP_001502124.1	PREDICTED: hypothetical protein [Equus caballus]	546	546	.00%	4e-154	G
NP_002138.1	homeobox B5 [Homo sapiens] >ref|XP_001173004.1| PREDICTED: home	546	546	.00%	7e-154	U G
BAB28059.1	unnamed protein product [Mus musculus]	474	474	85%	3e-132	U G
XP_001367145.1	PREDICTED: similar to homeobox protein [Monodelphis domestica]	462	462	.00%	8e-129	G
NP_001020526.1	homeo box B5 [Gallus gallus] >gb|AAW48484.1| homeodomain transcrip	431	431	.00%	3e-119	U G
BAD95556.1	Hoxb-5 [Gallus gallus]	417	417	97%	4e-115	U G
NP_571176.2	homeo box B5a [Danio rerio] >sp|P09014|HXB5A_BRARE Homeobox prot	394	394	.00%	3e-108	U G
CAA48320.1	homeodomain protein [Danio rerio] >emb|CAA31290.1| ZF-21 gene prod	392	392	.00%	1e-107	G
BAD95561.1	Hoxb-5 [Pelodiscus sinensis]	385	385	96%	2e-105	

Scoring Matrices

NCBI Minicourses

Nucleotide = 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

```
CAGGTAGCAAGCTTGCATGTCA
|| |||||||||||  |||||    raw score = 19-9* = 10*
CACGTAGCAAGCTTG-GTGTCA
```

* ignores gap costs

Scoring Matrices

NCBI Minicourses

Protein = substitution matrix

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

BLOSUM Matrices

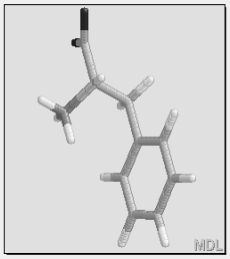
NCBI Minicourses

- 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

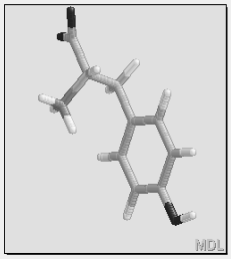
BLOSUM62

NCBI Minicourses

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	-3	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-2	-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

Negative for less likely substitutions

Positive for more likely substitutions

Where do Expect Values Come From?

NCBI Minicourses

Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value
NP_032294.2	homeo box B5 [Mus musculus] >ref|XP_573183.1| PREDICTED: similar to H	553	553	100%	4e-156
AAA37842.1	homeobox protein	551	551	100%	1e-155
XP_548176.2	PREDICTED: similar to Homeobox protein Hox-B5 (Hox-2A) (HHO.C10) (HU-	548	548	100%	1e-154
XP_001502124.1	PREDICTED: hypothetical protein [Equus caballus]	546	546	100%	4e-154
NP_002138.1	homeobox B5 [Homo sapiens] >ref|XP_001173004.1| PREDICTED: homeob	546	546	100%	7e-154
BAB28059.1	unnamed protein product [Mus musculus]	474	474	85%	3e-132
XP_001367145.1	PREDICTED: similar to homeobox protein [Monodelphis domestica]	462	462	100%	8e-129
NP_001020526.1	homeo box B5 [Gallus gallus] >gb|AAW48484.1| homeodomain transcriptio	431	431	100%	3e-119
BAD95556.1	Hoxb-5 [Gallus gallus]	417	417	97%	4e-115
NP_571176.2	homeo box B5a [Danio rerio] >sp|P09014|HXB5A_BRARE Homeobox protein	394	394	100%	3e-108
CAA48320.1	homeodomain protein [Danio rerio] >emb|CAA31290.1| ZF-21 gene produc	392	392	100%	1e-107
BAD95561.1	Hoxb-5 [Pelodiscus sinensis]	385	385	96%	2e-105

NCBI Minicourses

Expect Value

E = number of database hits you expect to find by chance, ≥ S

$$E = Kmne^{-\lambda S} \quad \text{or} \quad E = mn2^{-S'}$$

K = scale for search space
λ = scale for scoring system
λS' = bitscore = (λS – lnK)/ln2
m = query length
n = database length

E is dependent on m x n (search space)

More info: **The Statistics of Sequence Similarity Scores**

NCBI Minicourses

Search Basics

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

Initial Options – Search Set

NCBI Minicourses

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: mamm

- mammals (taxid:40674)
- Mammalia (taxid:40674)
- placental mammals (taxid:9347)
- eutherian mammals (taxid:9347)
- egg-laying mammals (taxid:9255)
- mammalian hepatitis B-type viruses (taxid:10405)
- Mammalian virus group (taxid:353212)
- Mammalian orthoreovirus (taxid:351073)
- Mouse mammary tumor virus (taxid:11757)
- Mammillaria (taxid:130139)

be shown.

Entrez Query Optional

Program Selectio

Algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Initial Options – Search Set

NCBI Minicourses

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional

Enter an Entrez query to limit search

all[filter] NOT mammalia[organism]

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)

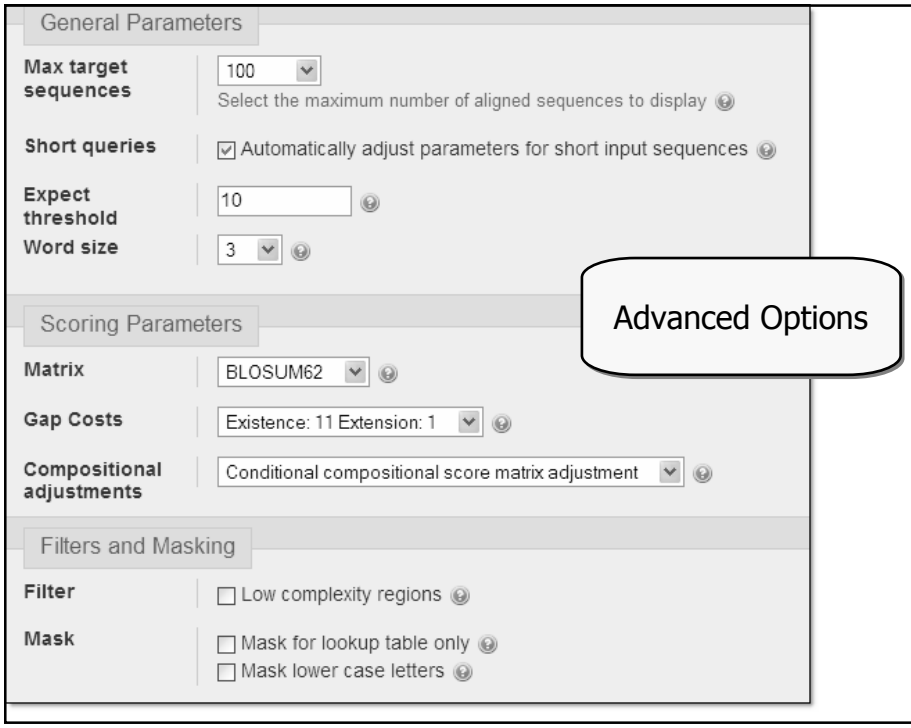
Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters



NCBI Minicourses

Search Basics

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

NCBI Minicourses

Reformat

NCBI Minicourses

NCBI/ BLAST/ blastp/ Formatting Results - N5W6JP2F013 [Reformat these Results] [Edit and Resubmit] [Sign in

Job Title: NP_032294:homeo box B5 [Mus musculus]

BLASTP 2.2.17 (Aug-26-2007)

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:
Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo Yu "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

RID: N5W6JP2F013

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
5,678,482 sequences; 1,961,803,296 total letters

Reformat

NCBI

Query	ref|np_032294| (269 letters)
Database	nr
Job title	NP_032294:homeo box B5 [Mus musculus]
Request ID	N5W6JP2F013 View report ☐ Show results in a new window
Format	
Show	Alignment as HTML ☑ Advanced View Reset form to defaults
Alignment View	Pairwise
Display	☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☐ NCBI-gi
Limit results	Masking Character: Lower Case / X for protein, n for nucleotide / Lower Case; Descriptions: 100; Organism; Entrez query; Expect Min; Expect Max
Format for	☐ PSI-BLAST with inclusion threshold

BLAST Output: Low Complexity Filter

NCBI

```
>gi|466462|gb|AAA17374.1| L human homolog of E. coli mutL gene product, Swiss-Prot
          Number P23367
        Length=756

 Score = 219.935 bits (559),  Expect = 8.71148e-57
 Identities = 120/131 (91%), Positives = 125/131 (95%), Gaps = 2/131 (1%)

Query  1    IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILE-VQQHIESKLL  59
            IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILE VQQHIESKLL
Sbjct  276  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL  335

Query  60   GSNSSRMYFTQTLLPGLAGPSGEMVK-sttsltssstsgsDKVYAHQMVRTDSREQKLDA  118
            GSNSSRMYFTQTLLPGLAGPSGEMVK +T+  +SS++  SDKVYAHQMVRTDSREQKLDA
Sbjct  336  GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA  395

Query  119  FLQPLSKPLSS  129
            FLQPLSKPLSS
Sbjct  396  FLQPLSKPLSS  406
```

low complexity sequence filtered

Alignment View Options

NCBI Minicourses

Alignment View: Pairwise

- Pairwise
- Pairwise with dots for identities
- Query-anchored with dots for identities
- Query-anchored with letters for identities
- Flat query-anchored with dots for identities
- Flat query-anchored with letters for identities
- Hit Table

Alignment View: Pairwise

- Pairwise
- Pairwise with dots for identities
- Query-anchored with dots for identities
- Query-anchored with letters for identities
- Flat query-anchored with dots for identities
- Flat query-anchored with letters for identities
- Hit Table

Alignment View: Flat Query-Anchored with Dots for 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
```

Alignment View: Hit Table

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 o
# 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

BLAST

Search Basics

NCBI Minicourses

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

Examples

NCBI Minicourses

- Protein searches more sensitive than nucleotide searches
  - redundancy of the genetic code
- megablast best when searching within same organism

BLAST is a shortcut . . .

NCBI Minicourses

An alignment BLAST cannot make:

```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    | || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |     |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
    |||| || ||||| ||  ||    |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

Reason:

no contiguous exact match of 7 bp.

Nucleotide vs. Protein BLAST

NCBI Minicourses

Comparing ADSS from *H. sapiens* and *A. thaliana*

```
         aaccggtgacggtggtgctcggtgcgcagtggggcgacgaaggc
Human:    N  R  V  T  V  V  L  G  A  Q  W  G  D  E  G
          +  +  V  +     V  L  G     Q  W  G  D  E  G
A.th.:    S  Q  V  S  G  V  L  G  C  Q  W  G  D  E  G
         agtcaagtatctggtgtactcggttgccaatggggagatgaaggt
```

BLASTp finds three matching words

BLASTn finds no match, because there are no 7 bp words

Protein searches are generally more sensitive than nucleotide searches.

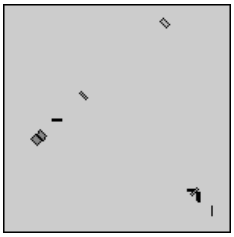
BLASTN vs TBLASTX

NCBI Minicourses

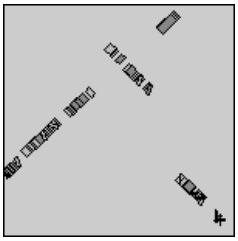
Anopheles gambiae mitochondrion, complete genome.
ACCESSION NC_002084 GI:5834911

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

blastn



tblastx



Hands-On

NCBI Minicourses

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