Mini Courses

NCBI Resources

one course at a time

NCBI Minicourses

BLAST Quick Start

blast-help@ncbi.nlm.nih.gov

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

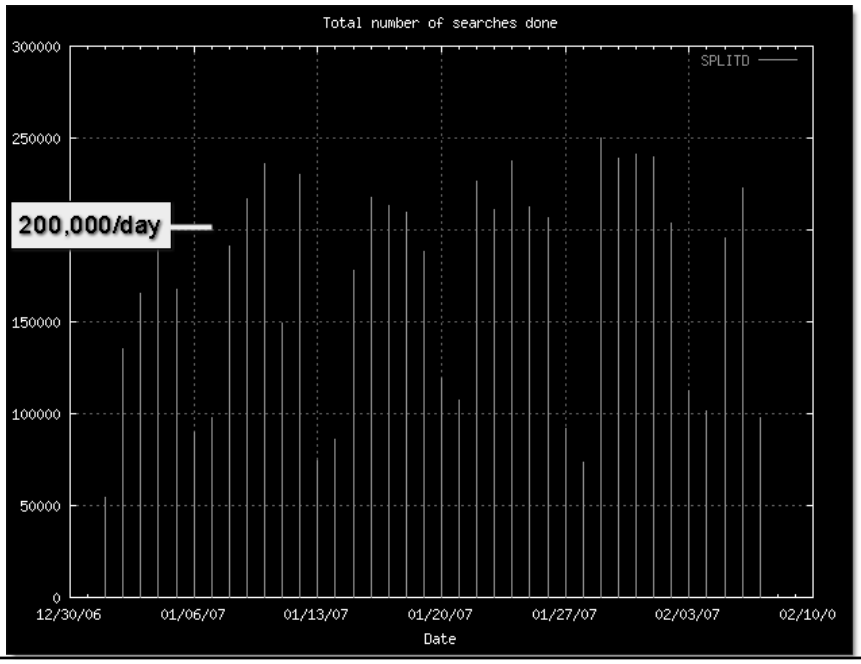
NCBI Minicourses

Compare protein 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)

BLAST Web Searches, 2007



NCBI Minicourses

Basic Local Alignment Search Tool

NCBI Minicourses

- 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

Global vs Local Alignment

NCBI Minicourses

Global vs Local Alignment

NCBI Minicourses

```
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
```

How BLAST Works

NCBI Minicourses

1. Make lookup table of “words” for query
2. Scan database for hits
3. Extend alignment both directions
   - Ungapped extensions of hits (initial HSPs)
   - Gapped extensions (no traceback)
   - Gapped extensions (traceback – alignment details)

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.

Scoring Systems - Proteins (BLOSUM62)

NCBI Minicourses

IEK: keep LEK?
(threshold = 11)

E / E

I / I

IEK = 14

L / I

LEK = 12

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

⇦ GCTGCTAGTCG ⇨

Protein: two matches within 40 residues

AIEKCYTGCTLAQEADDTA

⇦IDK ⇨ ⇦ EAD ⇨

BLASTP Summary

NCBI Minicourses

example query words

```
Query:   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILEV…
```

YLS	15	Neighborhood words	HFL	18
YLT	12		HFV	15
YVS	12		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

NCBI Minicourses

example query words

```
Query:   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILEV…
```

YLS	15	Neighborhood words	HFL	18
YLT	12		HFV	15
YVS	12		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
```

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

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

Word Size

Trade-off: sensitivity vs speed

NCBI Minicourses

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

NCBI Minicourses

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

NCBI Minicourses

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

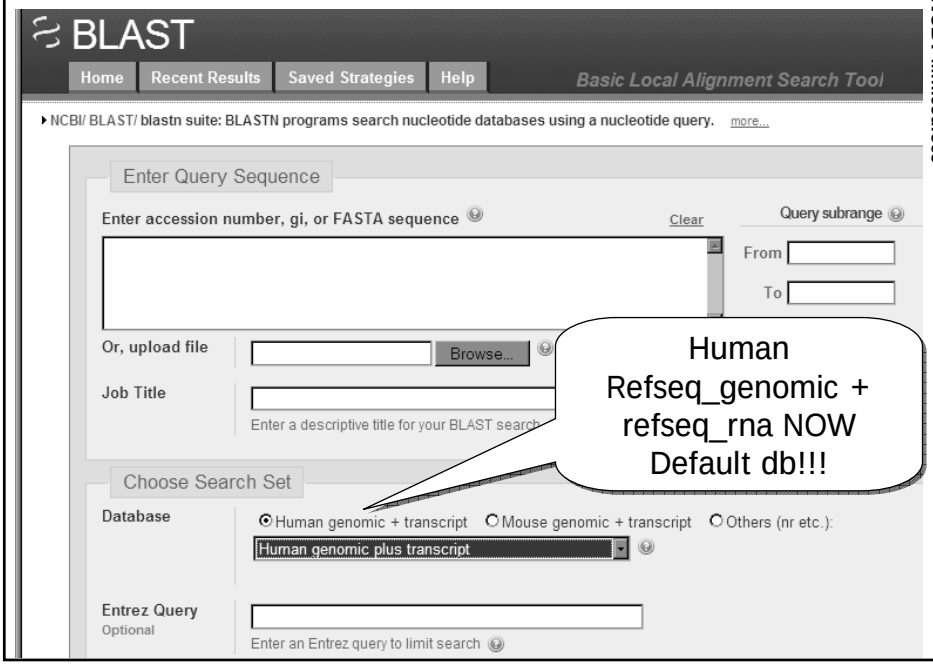
Nucleotide BLAST Databases

NCBI Minicourses

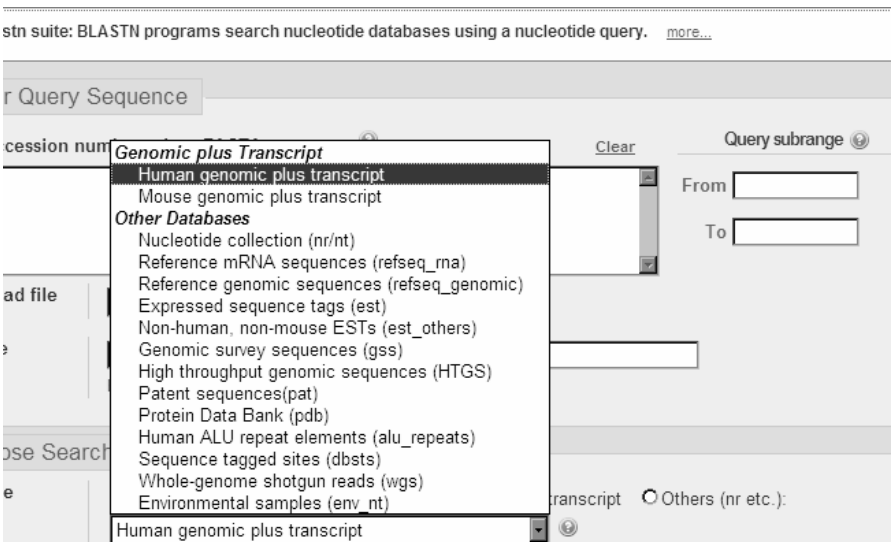
- **nr (nt)**
  - Traditional GenBank Divisions
  - NM_ and XM_ RefSeqs
- **refseq_rna**
  - NM_ , XM_ , NR_
- **refseq_genomic**
  - NC_ , NT_ , NG_
- **est**
  - EST Division
- **htgs**
  - HTG division
- **dbsts**
  - STS Division
- **chromosome**
  - NC genomic records
- **gss**
  - GSS division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

New Nucleotide Databases

NCBI Minicourses



New Nucleotide Databases



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

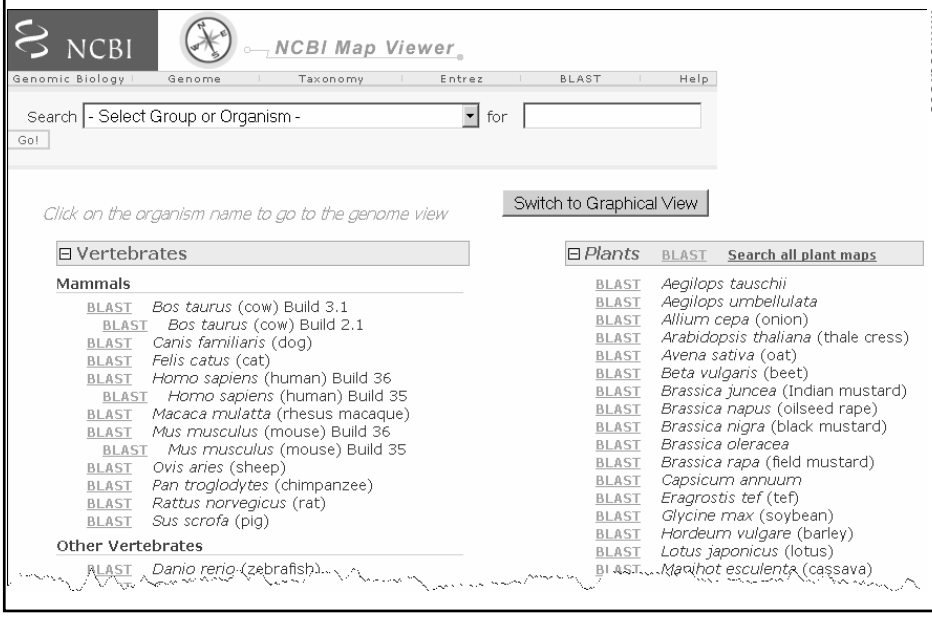
NCBI Minicourses

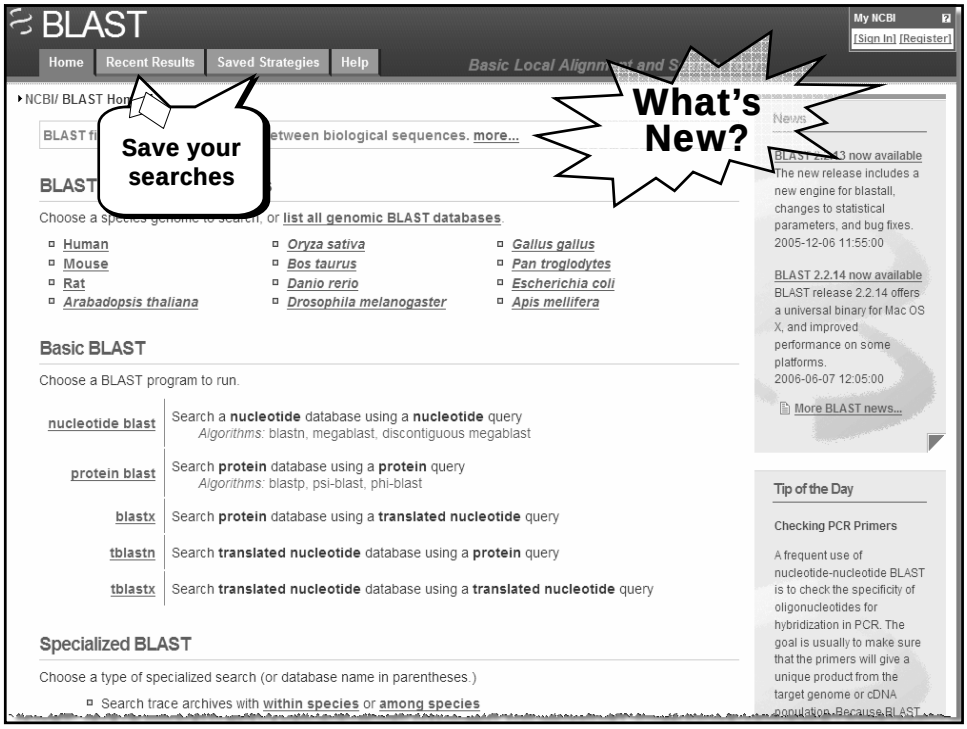
Genomes

- 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

Genome-specific: Map Viewer

NCBI Minicourses





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

Home | Recent Results | Saved Strategies | Help | Basic Local Alignmen

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more

Enter Query Sequence

Enter accession number, GI, or FASTA sequence | Clear

NP_032294

Or, upload file | Browse...

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Non-redundant protein sequences (nr)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Swissprot protein sequences(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Environmental samples(env_nr)

Organism Optional | Mouse | Custom...

Entrez Query Optional

Identifier or sequence

Title

NP_032294:homeo box B5 [Mus musculus]

Select database

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

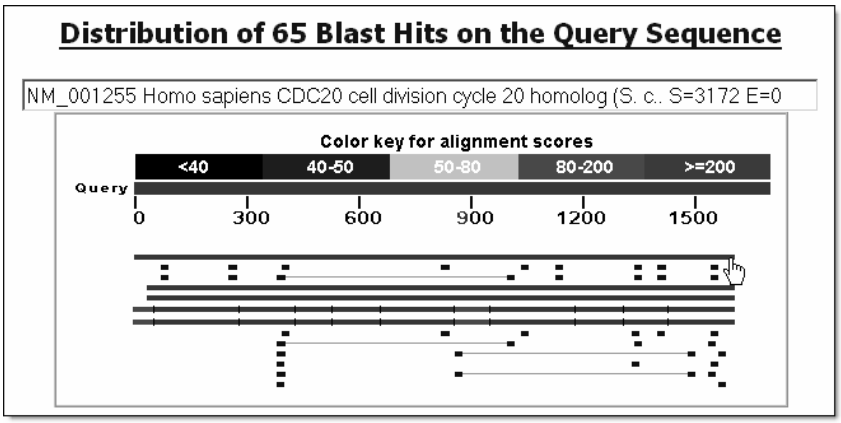
Request ID	41Z6F6C012
Status	Searching
Submitted at	Thu Mar 22 00:36:54 2007
Current time	Tue May 8 18:42:35 2007
Time since submission	00:00:11

Conserved Domain search run for protein queries

Default Output

Graphic Overview

NCBI Minicourses



Summary Table

NCBI Minicourses

Distance tree of results NEW

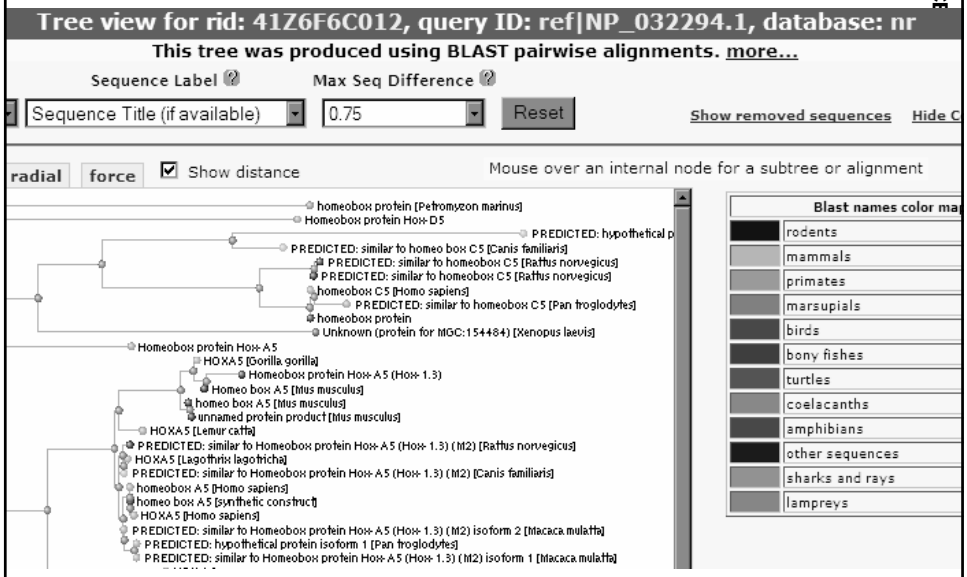
Sequences producing significant alignments:	Score (Bits)	E Value	
ref|NP_032294.1| homeo box B5 [Mus musculus] >gb|AAA37842.1| hom	420	4e-116	U G
ref|XP_573183.1| PREDICTED: similar to Homeobox protein Hox-B...	419	1e-115	U G
ref|XP_548176.2| PREDICTED: similar to Homeobox protein Hox-B...	416	7e-115	U G
ref|NP_002138.1| homeobox B5 [Homo sapiens] >ref|XP_001173004...	412	8e-114	U G
ref|XP_001367145.1| PREDICTED: similar to homeobox protein [Mono	378	2e-103	G
dbj|BAB28059.1| unnamed protein product [Mus musculus]	364	3e-99	U G
ref|NP_001020526.1| homeo box B5 [Gallus gallus] >gb|AAW48484...	347	4e-94	U G
dbj|BAD95556.1| Hoxb-5 [Gallus gallus]	337	4e-91	U G
emb|CAA48320.1| homeodomain protein [Danio rerio] >emb|CAA312...	327	5e-88	G
ref|NP_571176.2| homeo box B5a [Danio rerio] >sp|P09014|HXB5A...	325	1e-87	U G
dbj|BAD95561.1| Hoxb-5 [Pelodiscus sinensis]	320	6e-86	
ref|XP_001336387.1| PREDICTED: similar to homeo box protein B...	315	1e-84	U G
ref|NP_571612.2| homeo box B5b [Danio rerio] >gb|AAH78285.1| ...	315	1e-84	U G
gb|AAD15946.1| homeobox protein [Danio rerio]	312	1e-83	G
gb|ABM73560.1| homeodomain protein [Megalobrama amblycephala]	310	5e-83	

Alignment View - pairwise

NCBI Minicourses

```
> emb|CAA48320.1| homeodomain protein [Danio rerio]
 emb|CAA31290.1| 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
```

BLAST Output – tree view



New Output View

NCBI Minicourses

43CZKJFR014 | View report | Show results in a new window

Show: Alignment as HTML | Advanced View | Reset form to defaults

Alignment View: Pairwise

Display: Graphical Overview | Linkout | Sequence Retrieval | NCBI-gi

Masking Character: Lower Case | Masking Color: Grey

Limit results: Descriptions: 100 | Graphical overview: 100 | Alignments: 100

Entrez query:

Organism: Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Expect Min: | Expect Max:

Format for: PSI-BLAST | with inclusion threshold:

New Output View

NCBI Minicou

Results can be sorted

Transcript & genomic hits separated

Legend f ... es: U UniGene E GEO ... Map

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

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell divisio	2876	2876	95%	0.0	97%	U E G M
XM_94493		40.1	40.1	1%	8.2	100%	G M
XM_93301		40.1	40.1	1%	8.2	100%	G M
Genomic sequences [show first]							
NT_023935.17	Homo sapiens chromosome 9 ge	2629	2673	94%	0.0	100%	
NW_924484.1	Homo sapiens chromosome 9 ge	2601	2645	94%	0.0	100%	
NT_032977.8	Homo sapiens chromosome 1 ge	428	3002	95%	9e-117	100%	
NW_921351.1	Homo sapiens chromosome 1 ge	428	3010	95%	9e-117	100%	
NT_009237.17	Homo sapiens chromosome 11 g	44.1	44.1	1%	0.53	100%	

Pseudogene, chr 9

Functional gene, chr 1

Sorting Results

NCBI Minicourses

Resorted by Total score

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

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell divisi(	2876	2876	95%	0.0	97%	U E G M
XM_933013.2	PREDICTED: Homo sapiens chro	40.1	40.1	1%	8.2	100%	G M
XM_944935.2	PREDICTED: Homo sapiens chro	40.1	40.1	1%	8.2	100%	G M
Genomic sequences [show first]							
NW_921351.1	Homo sapiens chromosome 1 ge	428	3010	95%	9e-117	100%	
NT_032977.8	Homo sapiens chromosome 1 ge	428	3002	95%	9e-117	100%	
NT_023935.17	Homo sapiens chromosome 9 ge	2629	2673	94%	0.0	100%	
NW_924484.1	Homo sapiens chromosome 9 ge	2601	2645	94%	0.0	100%	
NW_921807.1	Homo sapiens chromosome 3 ge	44.1	44.1	1%	0.53	100%	

Functional gene now first

Sorting Hits: by Score

NCBI Minic

```
> ref|NT_032977.8|Hs1_33153  D  Homo sapiens chromosome 1 genomic contig, reference assembly
Length=73835825

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position

 Features in this part of subject sequence:
   cell division cycle 20

 Score =  428 bits (216),  Expect = 9e-117
 Identities = 231/236 (97%), Gaps = 0/236 (0%)
 Strand=Plus/Plus

Query  965       CTCCAGTGGTTCACGTTCTGGCCACATCCACCACCATGATGTTCGGGTAGCAGAACACCA  1024
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13798316  CTCCAGTGGTTCACGTTCTGGCCACATCCACCACCATGATGTTCGGGTAGCAGAACACCA  13798375

Query  1025      TGTGGCTACACTGAGTGGCCACAGCCAGGAAGTGTGTGGGCTGCGCTGGGCCCCAGATGG  1084
```

Sort by Score: longest exon usually first

Sorting Hits: by Query Start

NCBI Minicou

```
>ref|NT_032977.8|Hs1_33153  Homo sapiens chromosome 1 genomic contig, reference assembly
Length=73835825

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position

 Features flanking this part of subject sequence:
   6169 bp at 5' side: myeloproliferative leukemia virus oncogene
   223 bp at 3' side: cell division cycle 20


 Score = 89.7 bits (45),  Expect = 1e-14
 Identities = 51/53 (96%), Gaps = 0/53 (0%)
 Strand=Plus/Plus

Query  1         AGCGGAGAGTTTAAGAGGCGTAAGCGAGGCGTGTTAAACCCGGTCGGAACTGC  53
                 ||||||||||||||||||||||||||| |||||||||||| |||||||||||||
Sbjct  13796530  AGCGGAGAGTTTAAGAGGCGTAAGCCAGGCGTGTTAAAGCCGGTCGGAACTGC  13796582
```

Sort by Query start:
Proper exon order

BLAST

Search Basics

NCBI Minicourses

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

...con't

Where do Scores Come From?

NCBI Minicourses

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

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell divisi(	2876	2876	95%	0.0	97%	U E G M
XM_933013.2	PREDICTED: Homo sapiens chro	40.1	40.1	1%	8.2	100%	G M
XM_944935.2	PREDICTED: Homo sapiens chro	40.1	40.1	1%	8.2	100%	G M
Genomic sequences [show first]							
NW_921351.1	Homo sapiens chromosome 1 ge	428	3010	95%	9e-117	100%	
NT_032977.8	Homo sapiens chromosome 1 ge	428	3002	95%	9e-117	100%	
NT_023935.17	Homo sapiens chromosome 9 ge	2629	2673	94%	0.0	100%	
NW_924484.1	Homo sapiens chromosome 9 ge	2601	2645	94%	0.0	100%	
NW_921807.1	Homo sapiens chromosome 3 ge	44.1	44.1	1%	0.53	100%	

Scoring Matrices

NCBI Minicourses

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

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

* ignores gap costs

Scoring Matrices

NCBI Minicourses

Protein search: substitution matrix

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

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

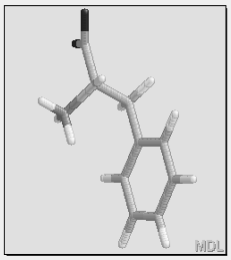
NCBI Minicourses

- 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

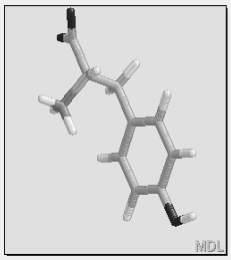
BLOSUM62

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	-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]	[illegible]	-1	-1	-2	-1	1	5				
W	[illegible]	[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

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*

NCBI Minicourses

Position-Specific Score Matrix

Serine/Threonine protein kinases catalytic loop

```
PSSM scores    1    5     ####7 4 ## 4
        K----------GIIHRDLKPENILLDE----
DAF-1   QiggskesnkpAMAHRDLKSKNIMVKN----
        A----------GIIHRDLKPSNIVVKs----
        P----------AIAHRDLKSKNILVKk----
        P----------AMAHRDLKSKNIMVKn----
        Q----------GVIHRDLKPENILINe----
        P----------AIAHRDLKSKNILVKk----
        T----------PLIHGDIKPANILLDq----
        P----------PIVHRDLKSPNLLVDk----
        V----------PVIHRDLKSNNILLLqpies
```

NCBI Minicourses

Position-Specific Score Matrix

NCBI Minicourses

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
435	K	-1	0	0	-1	-2	3	0	3	0	-2	-2	1	-1	-1	-1	-1	-1	-1	-1	-2
436	E	0	1	0	2	-1	0	2	-1	0	-1	-1	0	0	0	-1	0	0	-1	-1	-1
437	S	0	0	-1	0	1	1	0	1	1	0	-1	0	0	0	2	0	-1	-1	0	-1
438	N	-1	0	-1	-1	1	0	-1	3	3	-1	-1	1	-1	0	0	-1	-1	1	1	-1
439	K	-2	1	1	-1	-2	0	-1	-2	-2	-1	-2	5	1	-2	-2	-1	-1	-2	-2	-1
440	P	-2	-2	-2	-2	-3	-2	-2	-2	-2	-1	-2	-1	0	-3	7	-1	-2	-3	-1	-1
441	A	3	-2	1	-2	0	-1	0	1	-2	-2	-2	0	-1	-2	3	1	0	-3	-3	0
442	M	-3	-4	-4	-4	-3	-4	-4	-5	-4	7	0	-4	1	0	-4	-4	-2	-4	-1	2
443	A	4	-4	-4	-4	0	-4	-4	-3	-4	4	-1	-4	-2	-3	-4	-1	-2	-4	-3	4
444	H	-4	-2	-1	-3	-5	-2	-2	-4	10	-6	-5	-3	-4	-3	-2	-3	-4	-5	0	-5
445	R	-4	8	-3	-4	0	-1	-2	-3	-2	-5	-4	0	-3	-2	-4	-3	-3	0	-4	-5
446	D	-4	-4	-1	8	-6	-2	0	-3	-3	-5	-6	-3	-5	-6	-4	-2	-3	-7	-5	-5
447	I	-4	-5	-6	-6	-3	-4	-5	-6	-5	3	5	-5	1	1	-5	-5	-3	-4	-3	1
448	K	0	0	1	-3	-5	-1	-1	-3	-3	-5	-5	7	-4	-5	-3	-1	-2	-5	-4	-4
449	S	0	-3	-2	-3	0	-2	-2	-3	-3	-4	-4	-2	-4	-5	2	6	2	-5	-4	-4
450	K	0	3	0	1	-5	0	0	-4	-1	-4	-3	4	-3	-2	2	1	-1	-5	-4	-4
451	N	-4	-3	8	-1	-5	-2	-2	-3	-1	-6	-6	-2	-4	-5	-4	-1	-2	-6	-4	-5
452	I	-3	-5	-5	-6	0	-5	-5	-6	-5	6	2	-5	2	-2	-5	-4	-3	-5	-3	3
453	M	-4	-4	-6	-6	-3	-4	-5	-6	-5	0	6	-5	1	0	-5	-4	-3	-4	-3	0
454	V	-3	-3	-5	-6	-3	-4	-5	-6	-5	3	3	-4	2	-2	-5	-4	-3	-5	-3	5
455	K	-2	1	1	4	-5	0	-1	-2	1	-4	-2	4	-3	-2	-3	0	-1	-5	-2	-3
456	N	1	1	3	0	-4	-1	1	0	-3	-4	-4	3	-2	-5	-2	2	-2	-5	-4	-4
457	D	-3	-2	5	5	-1	-1	1	-1	0	-5	-4	0	-2	-5	-1	0	-2	-6	-4	-5
458	L	-3	-1	0	-3	0	-3	-2	3	-4	-2	3	0	1	1	-2	-2	-3	5	-1	-3

catalytic loop

Where do Expect Values Come From?

NCBI Minicourses

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

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell divisi	2876	2876	95%	0.0	97%	U E G M
XM_933013.2	PREDICTED: Homo sapiens chro	40.1	40.1	1%	8.2	100%	G M
XM_944935.2	PREDICTED: Homo sapiens chro	40.1	40.1	1%	8.2	100%	G M
Genomic sequences [show first]							
NW_921351.1	Homo sapiens chromosome 1 ge	428	3010	95%	9e-117	100%	
NT_032977.8	Homo sapiens chromosome 1 ge	428	3002	95%	9e-117	100%	
NT_023935.17	Homo sapiens chromosome 9 ge	2629	2673	94%	0.0	100%	
NW_924484.1	Homo sapiens chromosome 9 ge	2601	2645	94%	0.0	100%	
NW_921807.1	Homo sapiens chromosome 3 ge	44.1	44.1	1%	0.53	100%	

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**

E is dependent on m x n (search space)

NCBI Minicourses

Short query =

low score

high Expect

```
 Score = 30.2 bits (15),  Expect =    19
 Identities = 15/15 (100%), Gaps = 0/15 (0%)
 Strand=Plus/Plus

Query  4         GTAGCAAGCTTGCAT  18
                 |||||||||||||||
Sbjct  40347983  GTAGCAAGCTTGCAT  40347997
```

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

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

BLAST

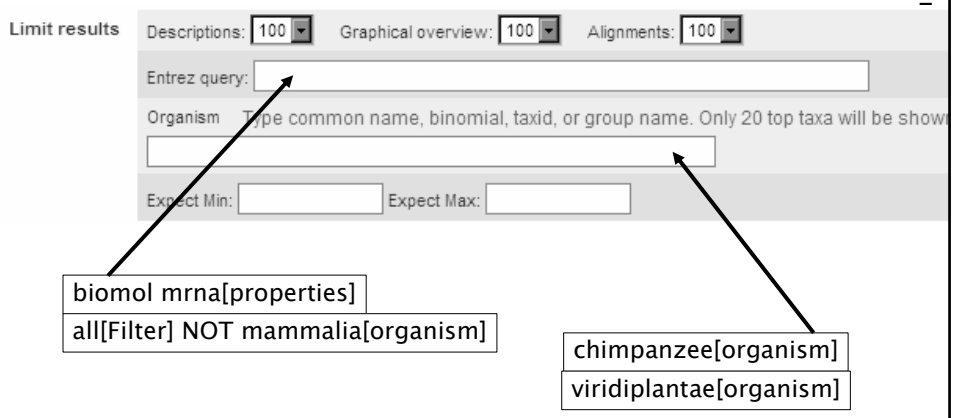
Search Basics

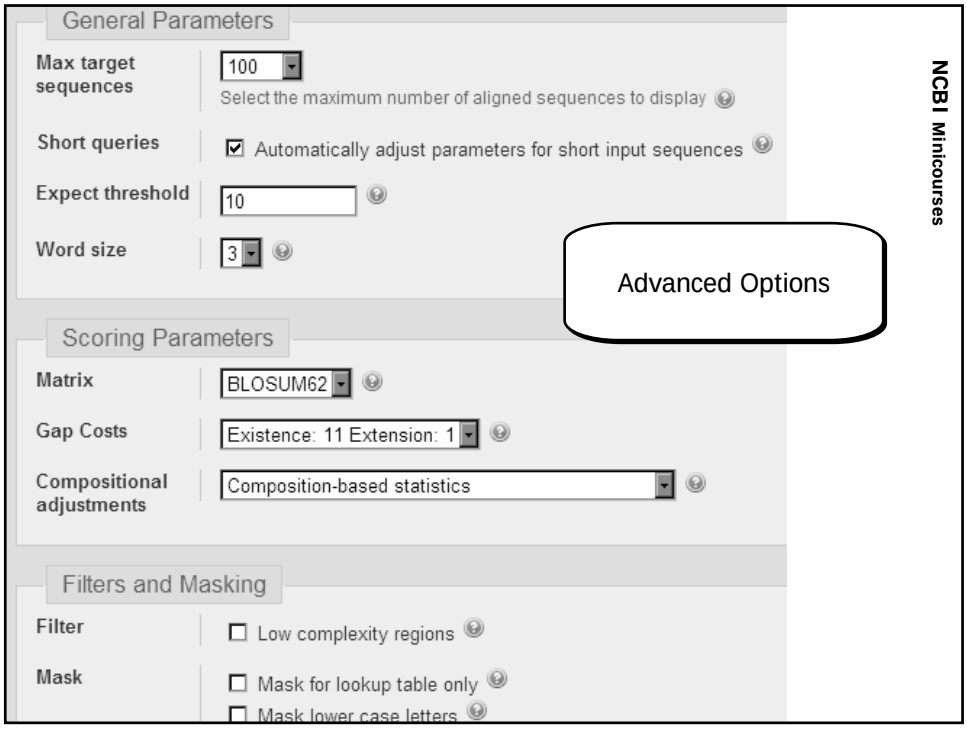
NCBI Minicourses

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

Initial Options

NCBI





BLAST

Search Basics

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

NCBI Minicourses

BLAST Output: Alignments & Filter

NCBI

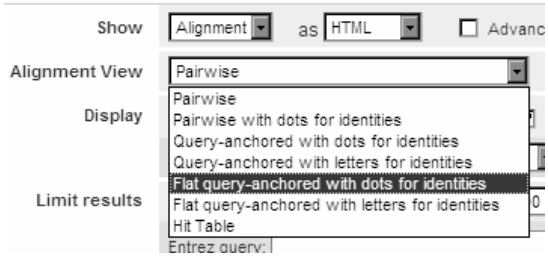
```
>gi|466462|gb|AAA17374.1| human homolog of E. coli mutL gene product, Swiss-Prot
          Number P23367
         Length=756

 Score = 219.935 bits (559),  Expect = 8.711
 Identities = 120/131 (91%), Positives = 125

Query  1    IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKH      EESILE-VQQHIESKLL  59
            IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKH      HEESILE VQQHIESKLL
Sbjct  276  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKH      LHEESILERVQQHIESKLL  335

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

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

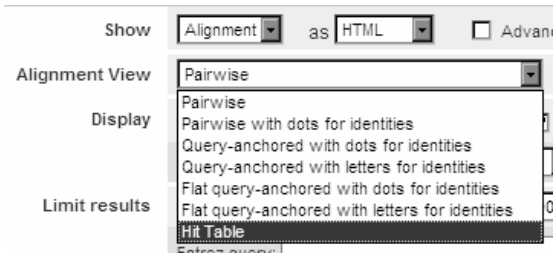
positives, from score matrix

low complexity sequence filtered

Alignment View Options

NCBI Minicourses





Alignment View: 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
```

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|;2
gi|6680251|ref|NP_032294.1|     gi|74225740|dbj|BAE21695.1|;gi|73695452|gb|AAI03605.1|;gi|730
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|400
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|15029897|gb|AAH11083.1|      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

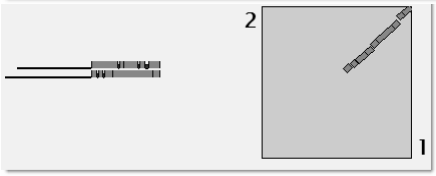
- blastn, megablast best when searching within same organism

BLASTN vs BLASTP

NCBI Minicourses

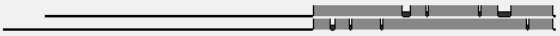
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.14 [Apr-09-2006]

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%)
```

BLAST is a shortcut . . .

NCBI Minicourses

An alignment BLAST cannot make:

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

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

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

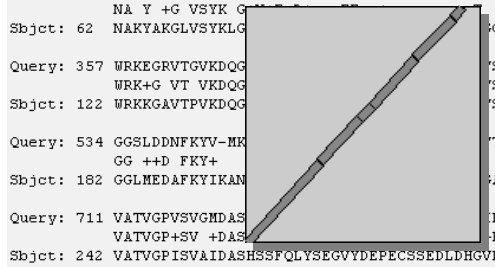
Reason:

no contiguous exact match of 7 bp.

An Alignment BLAST Can Make

NCBI

```
Score = 290 bits (741), Expect = 7e-77
Identities = 147/331 (44%), Positives = 206/331 (61%), Gaps = 8/331 (2%)
Frame = +3
```



```
Query: 6   MKVAIFFSLLVVAISASISEELGAKFQAFKLEHGKTYLNQAEESKRFNIFTDNVRAIEAH 185
           +++++  +++ V ++AS  E L  +++AFK  H K+Y +  EE  RF IFT+N   I  H
Sbjct: 2   LRLSVLCAIVAVTVAASSQEILRTQWEAFKTTHKKSYQSHMEELLRFKIFTENSLIIAKH 61

Query: 186 NALYEQGKVSYKKGINKFTDMSQEEFKTMLTLSASRKPTLETTSYVKTGVE---IPSSVD 356
           NA Y +G VSYK G                                +T      V    +P  VD
Sbjct: 62  NAKYAKGLVSYKLG                                GSTFLPPANVNDSSLPKVVD 121

Query: 357 WRKEGRVTGVKDQG                                YSLSEQQLIDCCTD-TSAGCD 533
           WRK+G VT VKDQG                                YSLSEQ L+DC     + GC+
Sbjct: 122 WRKKGAVTPVKDQG                                YSLSEQNLVDCSQSFGNNGCE 181

Query: 534 GGSLDDNFKYV-MK                                TKVSKYTSIPAEDEDALLEA 710
           GG ++D FKY+                                    + Y  I A  E  L +A
Sbjct: 182 GGLMEDAFKYIKAN                                ATDTGYVEIKAGSEVDLKKA 241

Query: 711 VATVGPVSVGMDAS                                LAVGYGTENGKDYWIIKNSW 884
           VATVGP+SV +DAS                                L VGYG + GK YW++KNSW
Sbjct: 242 VATVGPISVAIDASHSSFQLYSEGVYDEPECSSEDLDHGVLVVGYGVKGGKKYWLVKNSW 301

Query: 885 GASWGEQGYFRLAR-GKNQCGISEDTVYPTI 974
             SWG+QGY  ++R   NQCGI+    YP +
Sbjct: 302 AESWGDQGYILMSRDNNNQCGIASQASYPLV 332
```

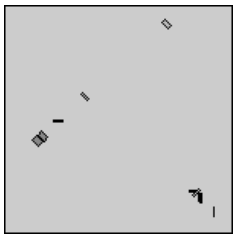
BLASTN vs TBLASTX

NCBI Minicourses

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

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

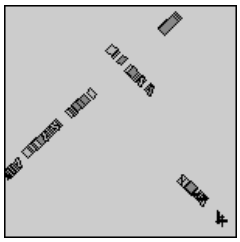
blastn



Homo sapiens

Anopheles gambiae

tblastx



Homo sapiens

Anopheles gambiae

Hands-On

NCBI Minicourses

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