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

NCBI Minicourses

Basic Local Alignment Search Tool

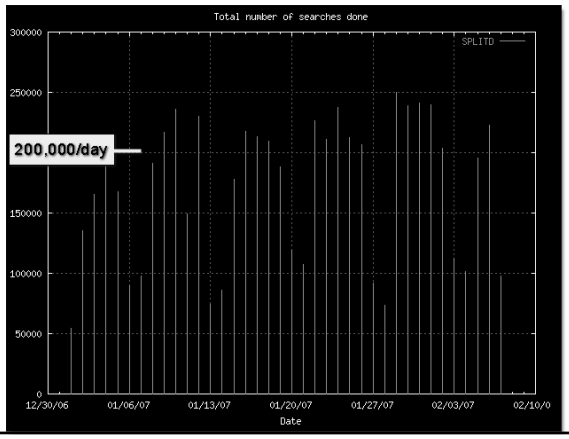
Compare protein or nucleic acid sequences to protein or nucleic acid databases

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

(BLAST 2 Sequences, or pairwise BLAST)

NCBI Minicourses

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

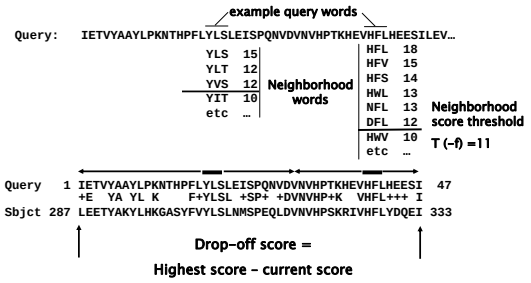
Nucleotide: one exact match

ATGCTGCTAGTCGATGACGTAGCTA
⇦ GCTGCTAGTCG ⇨

Protein: two matches within 40 residues

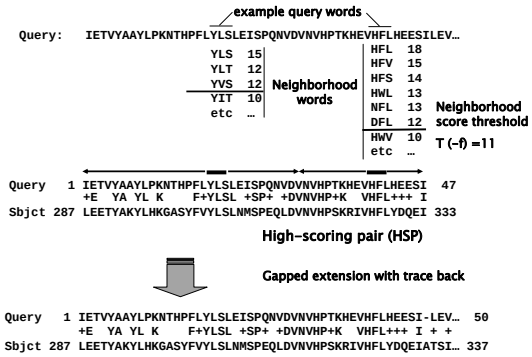
AIEKCYTGCTLAQEADDTA
⇦IDK⇨ ⇦ EAD ⇨

BLASTP Summary



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

BLASTP Summary

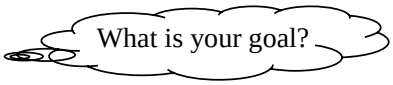


BLAST

Search Basics

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

BLAST Programs



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

Word Size

Trade-off: sensitivity vs speed

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

Templates for Discontiguous Words

```
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

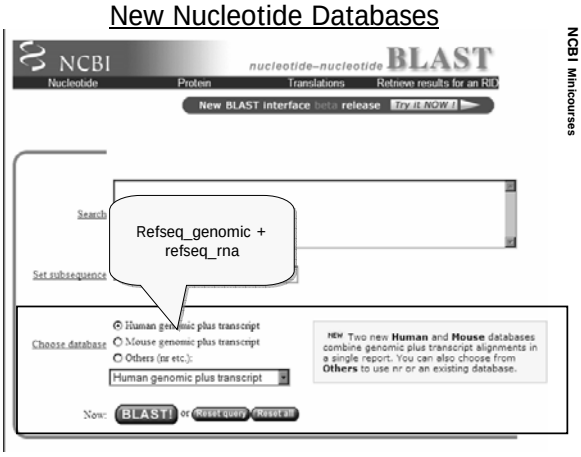
- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Nucleotide BLAST Databases

- **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

NCBI Minicourses

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

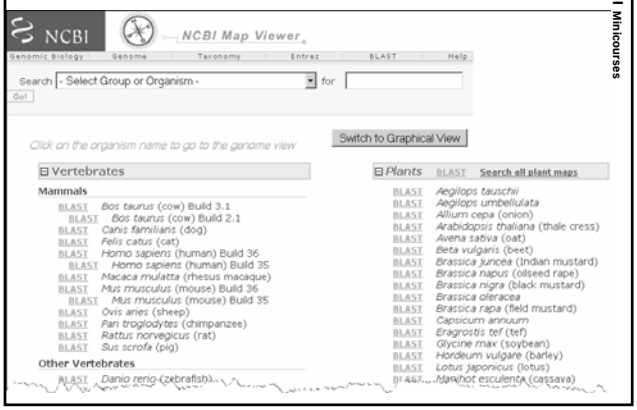
NCBI Minicourses

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

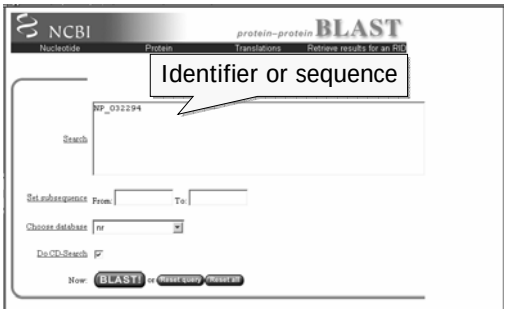


Search Basics

NCBI Minicourses

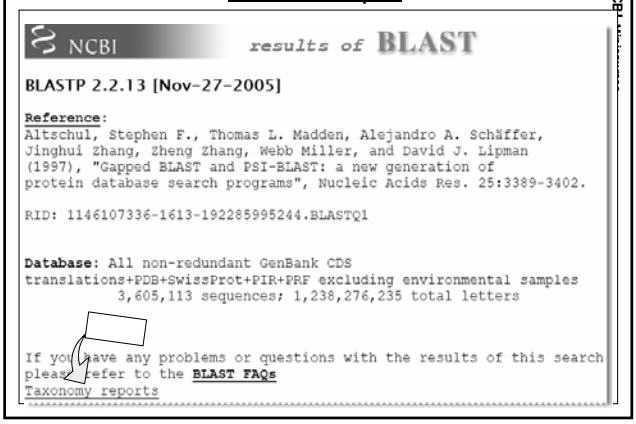
- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Setting up a BLAST Search



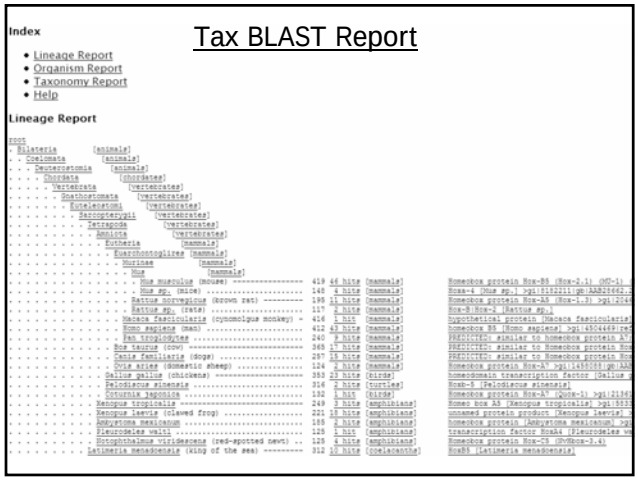
NCBI Minicourses

BLAST Output



NCBI Minicourses

Tax BLAST Report

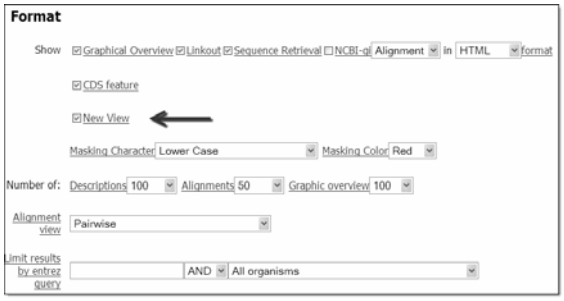


Tax BLAST Report

```
. . . . . 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]
```

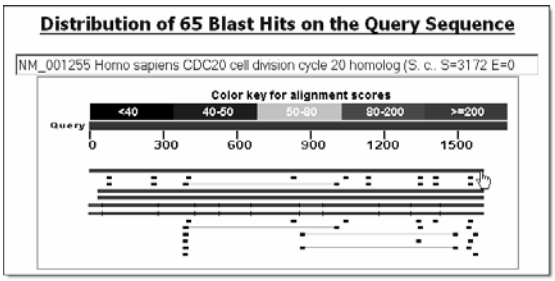
New Output View

NCBI Minicourses



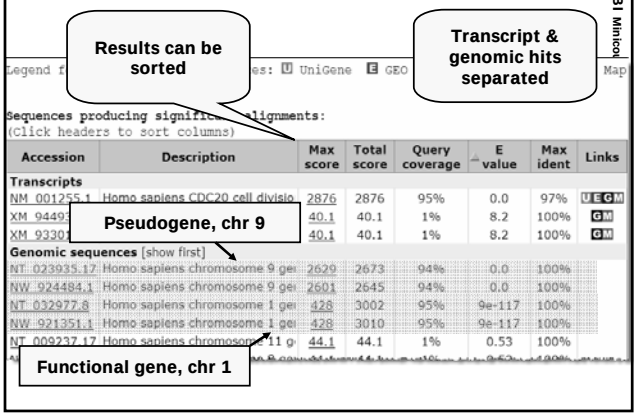
Graphic Overview

NCBI Minicourses



New Output View

NCBI Minicou



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 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 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 chrc	40.1	40.1	1%	8.2	100%	G M
XM_944935.2	PREDICTED: Homo sapiens chrc	40.1	40.1	1%	8.2	100%	G M
Genomic sequences [show first]							
NW_921351.1	Homo sapiens chromosome 1 g	428	3010	95%	9e-117	100%	
NT_032977.8	Homo sapiens chromosome 1 g	428	3002	95%	9e-117	100%	
NT_023935.17	Homo sapiens chromosome 9 g	2629	2673	94%	0.0	100%	
NW_924484.1	Homo sapiens chromosome 9 g	2601	2645	94%	0.0	100%	
NW_921807.1	Homo sapiens chromosome 3 g	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

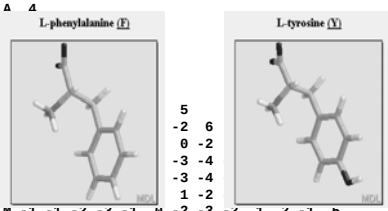
- global alignments of closely related proteins
- PAM1 calculated from sequences with no more than 1% divergence
- Other PAM matrices are extrapolated from PAM1
- Examples: PAM30 and PAM70

BLOSUM Matrices

- 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

NCBI Minicourses

BLOSUM62




```
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
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  X
```

Negative for less likely substitutions

Positive for more likely substitutions

NCBI Minicourses

BLOSUM62

```
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
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  X
```

J (leucine or isoleucine) and O (pyrrolysine)

NCBI Minicourses

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

NCBI Minicourses

Serine/Threonine protein kinases catalytic loop

PSSM scores 1 5 ####7 4 ## 4

```
        K----------GIIHRDIKPENILLDE----
DAF-1   QiggskesnkpamahrdikskniMVKN----
        A----------GIIHRDIKPSNIVVKs----
        P----------AIAHRDIKSKNILVKk----
        P----------AMAHRDIKSKNIMVKn----
        Q----------GVIHRDIKPENILINe----
        P----------AIAHRDIKSKNILVKk----
        T----------PLIHGDIKPANILLDq----
        P----------PIVHRDIKSPNLLVDk----
        V----------PVIHRDIKSNNILLLqpies
```

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 chrc	40.1	40.1	1%	8.2	100%	G M
XM_944935.2	PREDICTED: Homo sapiens chrc	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

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

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

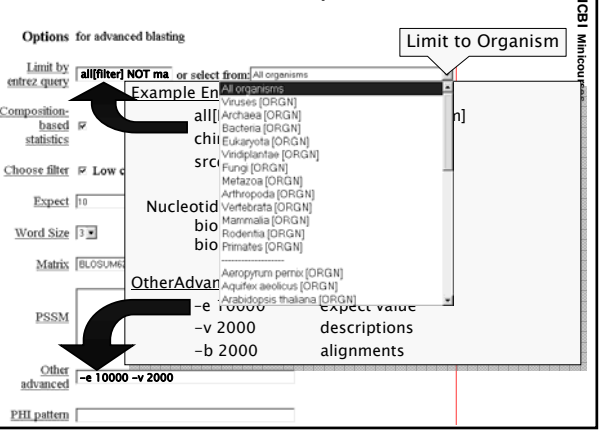
Search Basics

NCBI Minicourses

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

Advanced Options

NCBI Minicourses



Other Advanced Options

NCBI Minicourses

For short queries: increase e and lower W

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

-e expect value
-W word size
-v descriptions
-b alignments

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

Search Basics

NCBI Minicourses

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

'New' Formatter

NCBI Minicourses

Format

Show: Graphical Overview, Linkout, Sequence Retrieval, NCBI-gi Alignment in HTML format

Default(X for protein, n for nucleotide) | Maskin: Black

Default(X for protein, n for nucleotide)
Lower Case

Black
Grey
Red

Select lower case

Select red

Alignment view

Format for PSI-BLAST: with inclusion threshold

Limit results by entrez query | or select from: All organisms

Expect value range:

Layout: Two Windows | Formatting options on page with results: None

Autoformat: Semi-auto

BLAST Output: Alignments & Filter

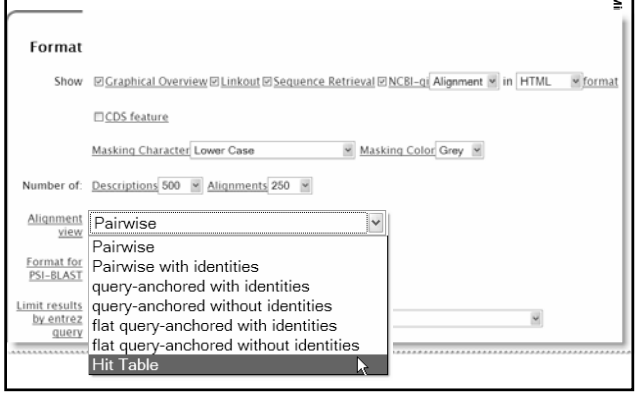
NCBI

```
>gi|4664621|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-sttsltsss   gsDKVYAHQMVRTDSREQKLDA  118
            GSNSSRMYFTQTLLPGLAGPSGEMVK +T+  +SS++  SDKVYAHQMVRTDSREQKLDA
Sbjct  336  GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA  395

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

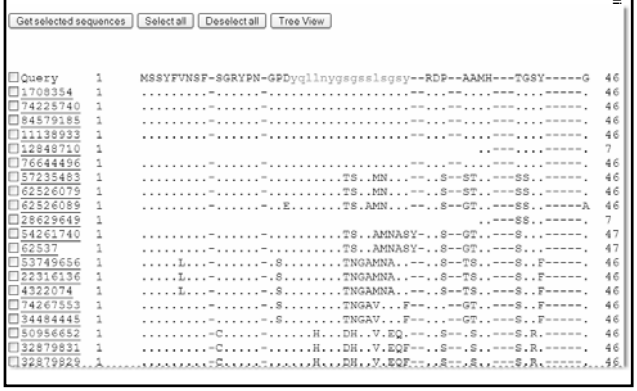
positives, from score matrix

low complexity sequence filtered

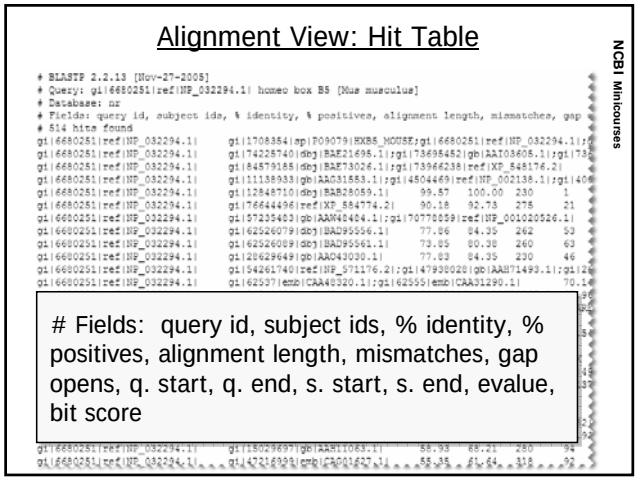
Alignment View Options



Alignment View: Flat Query-Anchored with Identities



Alignment View: Hit Table



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

Search Basics

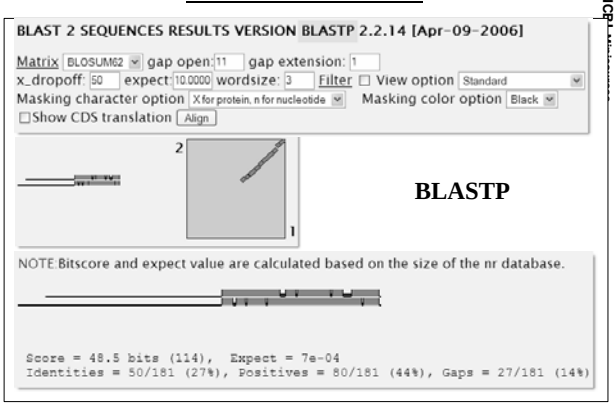
- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Examples

- Protein searches <u>more sensitive</u> than nucleotide searches
  - redundancy of the genetic code
- blastn, megablast best when searching within same organism

NCBI Minicourses

BLASTN vs BLASTP



NCBI Minicourses

BLAST is a shortcut . . .

An alignment BLAST cannot make:

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

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

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

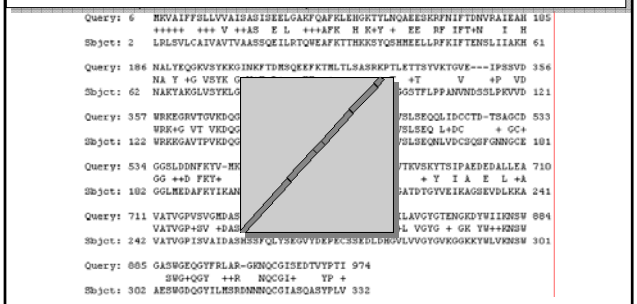
Reason:

no contiguous exact match of 7 bp.

NCBI Minicourses

An Alignment BLAST Can Make

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



NCBI

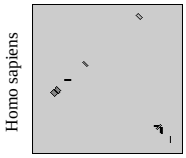
BLASTN vs TBLASTX

NCBI Minicourses

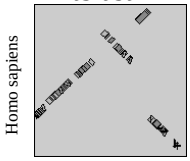
Anopheles gambiae mitochondrion, complete genome.
ACCESSION NC_002084 GI:5834911

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