Hot Topics

What's New with BLAST®?

Slides based on NCBI talk at American Society of Human Genetics
October 2005

Hot Topics Outline

I. New BLAST Algorithm: Discontiguous MegaBLAST

II. New Databases

III. New Formatting and Advanced Options

IV. Educational Resources

Search through NCBI BLAST Home Page

Web BLAST Help Doc

BLAST Statistics Doc

BLAST download

NCBI → BLAST

Latest news: 28 August 2005 : BLAST 2.2.12 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp NEW, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Environmental samples
- Malaria
- Insects, invertebrates, plants, fungi, microbial genomes, other eukaryotic genomes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST

Meta

- Retrieve results

Standard databases

Specialized databases

Retrieval Result with RID

New BLAST Algorithm

Discontiguous MegaBLAST

Why Do We Need Sequence Similarity Searching?

- To identify and annotate sequences
- To evaluate evolutionary relationships
- Other:
  - model genomic structure
  - check primer specificity *in silico*
  - Identify SNPs

BLAST: Sequence similarity search tool from NCBI

Basic *Local* Alignment Search Tool

- Is a widely used similarity search tool
- Uses Heuristic approach based on *Smith Waterman* algorithm
  - Sacrifices speed for sensitivity
- Finds best (biologically relevant) local alignments
- Provides statistical assessment on the significance
- Megablast – Similar to Blast, however sacrifices sensitivity for speed

Megablast: *contiguous* vs *discontiguous*

Contiguous megablast (NCBI Genome Annotator)

- *Long alignments* of *highly similar* sequences
- *Concatenation* of *query* sequences
- *Faster* and *less sensitive* than blastn

Discontiguous megablast (a more sensitive sibling)

- Uses *discontiguous* word matches
- Is more sensitive and better for *cross-species* comparisons
- Still maintains the speed edge over regular blastn

Whats a Discontiguous Word

Megablast
W = 11 — 11111111111

Discontiguous
W = 11, t = 16 — 110111111 0111

W = word size; # matches in template *[1=match; 0 = ignored, not evaluated]*

t = template length (window size within which the word match is evaluated)

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

Discontiguous Word Options

Options for advanced blasting

Limit by entrez query [ ] or select from: (none)

Choose filter ☑ Low complexity ☐ Human repeats ☑ Mask for lookup table only ☐ Mask lower case

Expect 10

Word Size 11

Percent Identity, match, mismatch scores None, 1, -2

Disontiguous Word options Template length 18 Template type Coding Require 2 word hits for extension ☑

Other advanced -G 5 -E 2

An Example . . .

Query: NM_078651

Drosophila melanogaster CG18582-PA (mbt) mRNA, (3244 bp)

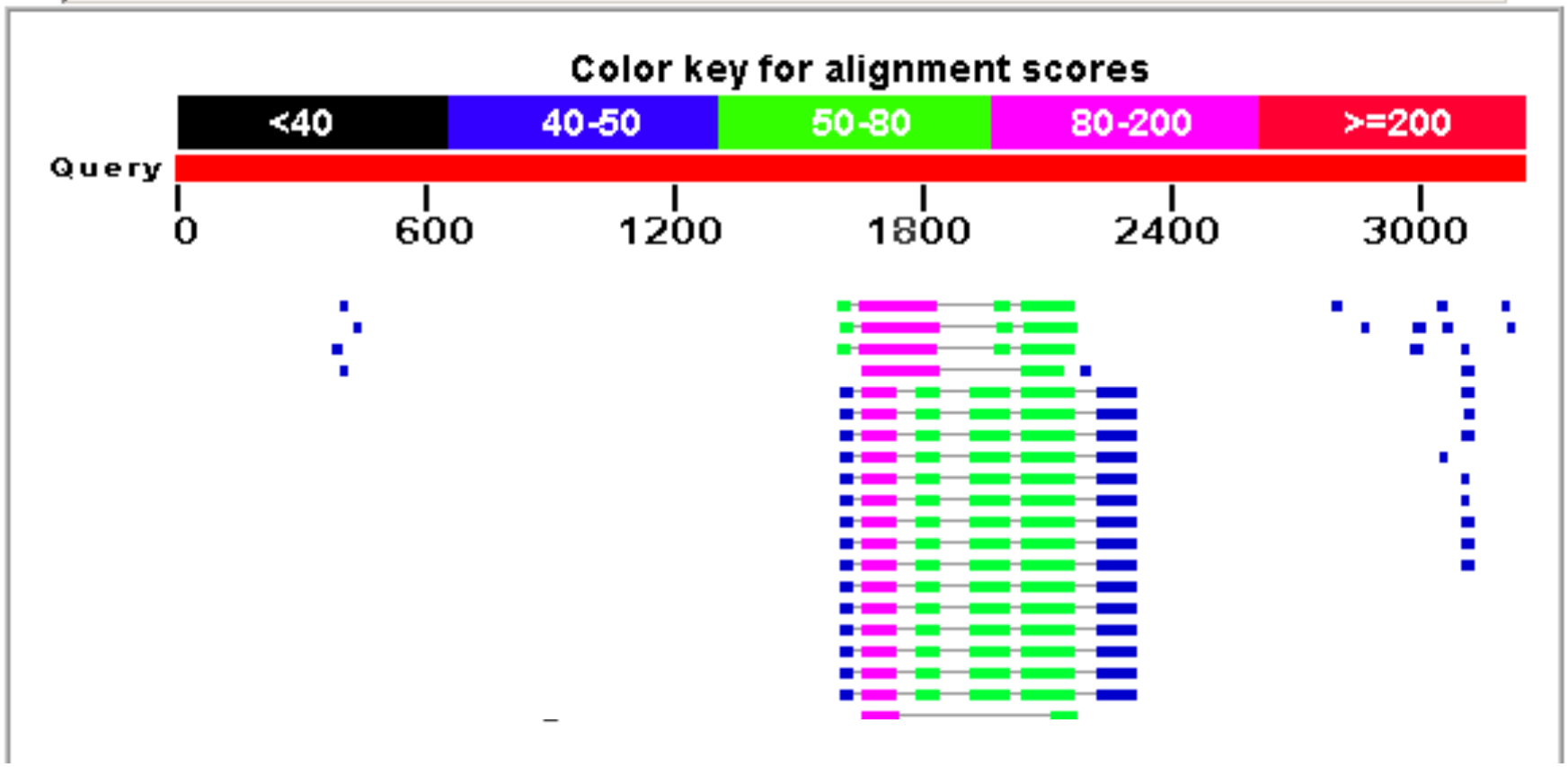
/note= mushroom bodies tiny; synonyms: Pak2, STE20, dPAK2

Database: nr (nt), Mammalia[orgn]

- MegaBLAST "No significant similarity found."

BLASTN Results

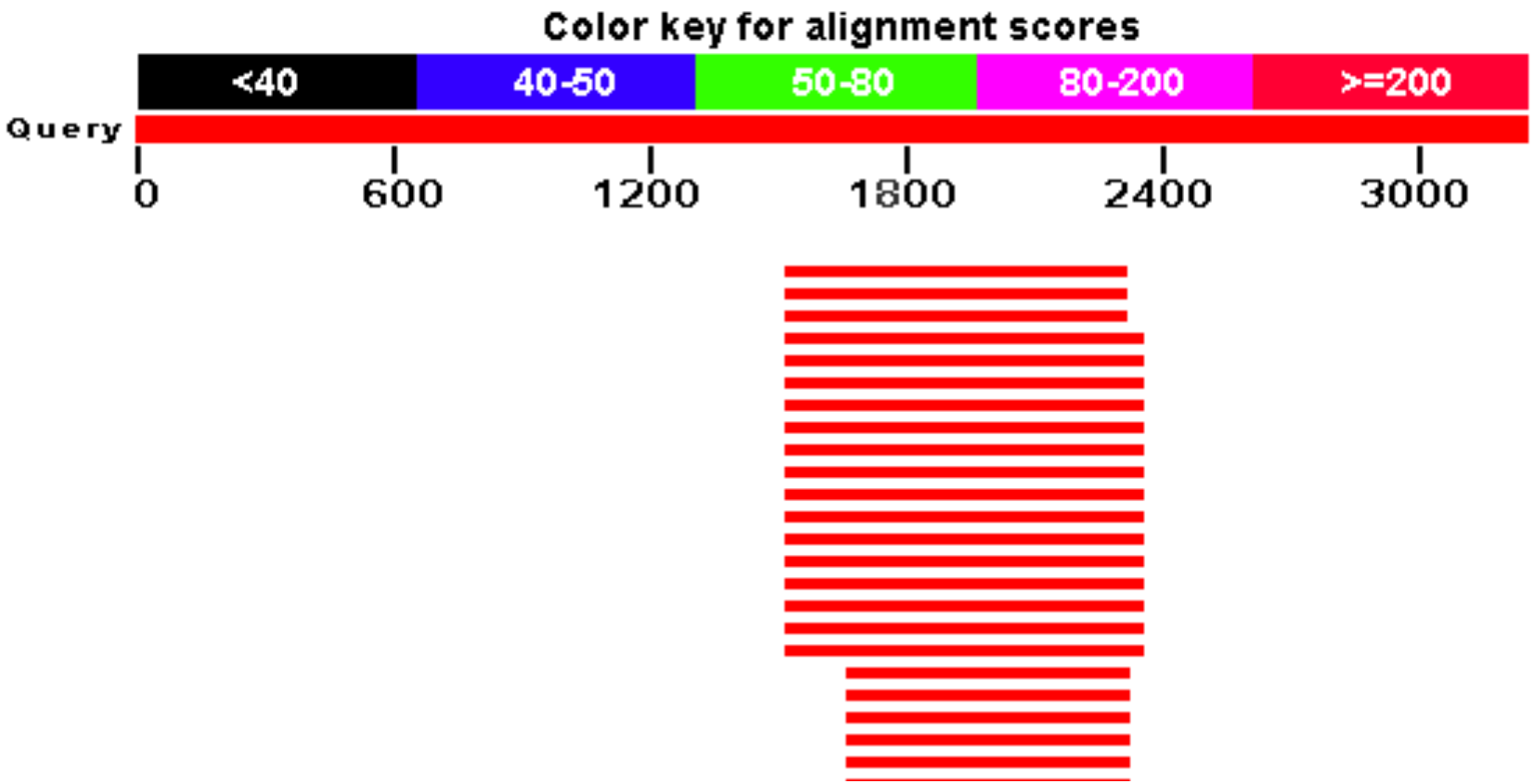
NM_005884 Homo sapiens p21(CDKN1A)-activated kinase 4 (PAK4), tra.. S=69 E=2.7e-08




```
Score = 87.7 bits (44),  Expect = 1e-13
Identities = 74/84 (88%), Gaps = 0/84 (0%)
Strand=Plus/Plus
```

Ex: Discontiguous MegaBLAST

NM_005884 Homo sapiens p21(CDKN1A)-activated kinase 4 (PAK4), tra.. S=365 E=3.4e-97




```
Score =  365 bits (190),  Expect = 3e-97
Identities = 624/836 (74%), Gaps = 2/836 (0%)
Strand=Plus/Plus
```

New BLAST Databases

Nucleotide and Protein BLAST Databases

Nucleotide

- refseq_rna = NM_*, XM_*
- refseq_genomic = NC_*, NG_*
- env_nt
  - environmental sample[filter], *e.g.*, 16S rRNA

Protein

- refseq = NP_*, XP_*
- env_nr

New Human Genome Databases

NCBI Home ▸ Genomic Biology ▸ Human Genome Resources ▸ BLAST

Search Map Viewer Go

BLAST
overview
FAQs
news
manual
references

Blast Human Sequences

Blast your sequence against Human specific sequences

Database: genome (all assemblies)

genome (reference only)
HTGS
RefSeq RNA
RefSeq protein
Non-RefSeq RNA
Non-RefSeq protein
Build RNA
Build protein
Ab initio RNA
Ab initio protein
ESTs
Clone end sequences
Traces- WGS
Traces- ESTs
Traces- other
Celera CSA
Celera cWGA
Celera WGSA
HSC_TCAG
SNPs

Program

Begin Sea

Enter an ac

FASTA format:

Optional pa

Expect 0.01 | default | 100 | 100 Alignments

Advanced options:

Begin Search | Clear Input

HTGS
RefSeq RNA
RefSeq protein
Non-RefSeq RNA
Non-RefSeq protein
Build RNA
Build protein
Ab initio RNA
Ab initio protein
ESTs
Clone end sequences
Traces- WGS
Traces- ESTs
Traces- other
Celera CSA
Celera cWGA
Celera WGSA
HSC_TCAG
SNPs

New Formatting Options

Masking Low Complexity Sequence

Format

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

Masking Character Default(X for protein, n for nucleotide) Masking Color Black

Number

Masking Character Default(X for protein, n for nucleotide) Masking Color Black

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

Alignment vi

Limit results by entrez query or select from: All organisms

Expect value range:

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

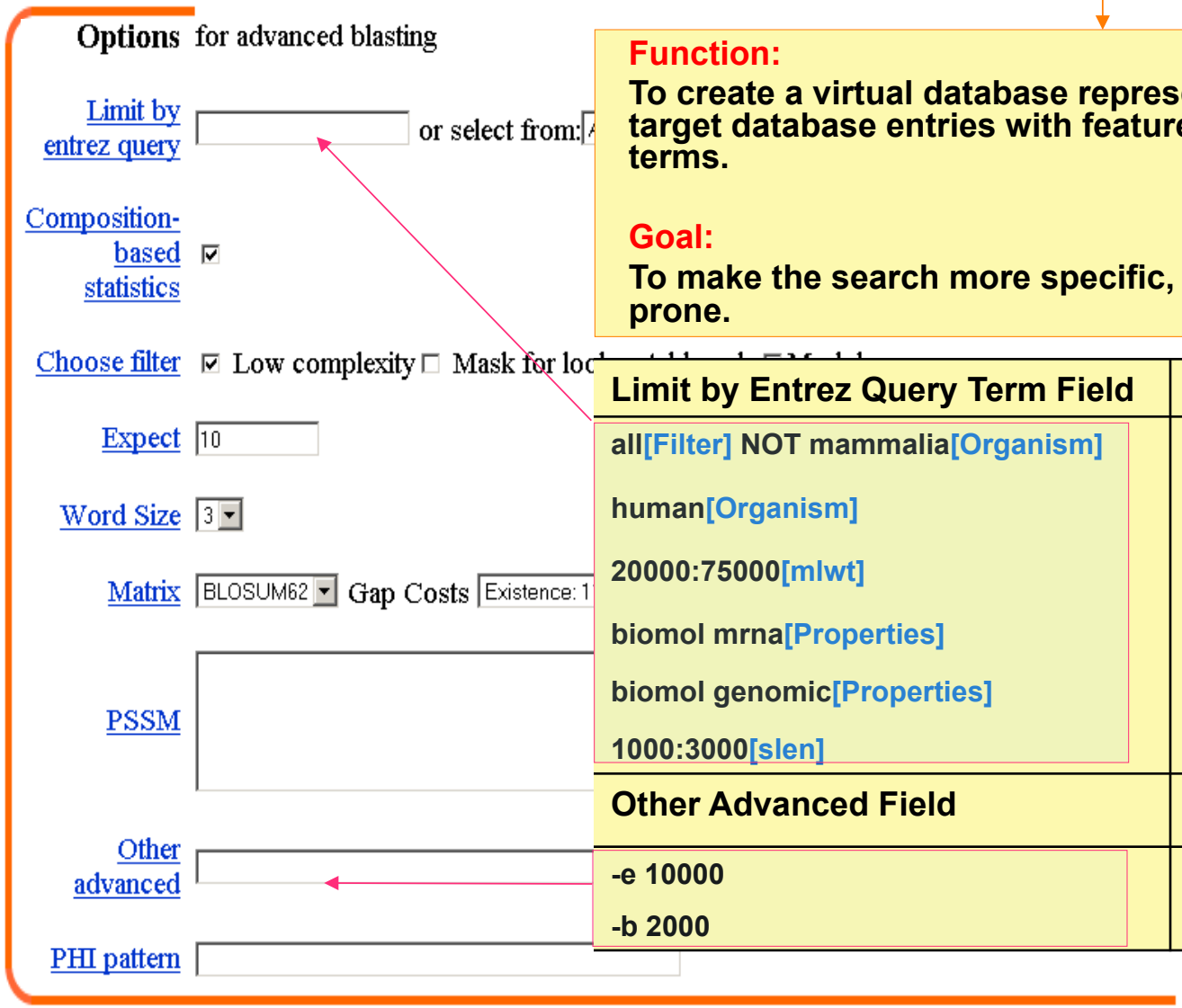
Autoformat Semi-auto

Select Black, Gray or Red

Why Filtering is Important

- When filtered, low-complexity sequences are treated as mismatches
  - "N" for nucleotide; "X" for proteins
  - Reduces the number of spurious database hits, thus improving E value
  - Caveat: Regions of percent identity not properly calculated
    - Altering the format, however, will report the correct percent identity

Customize the Search Using "Options"



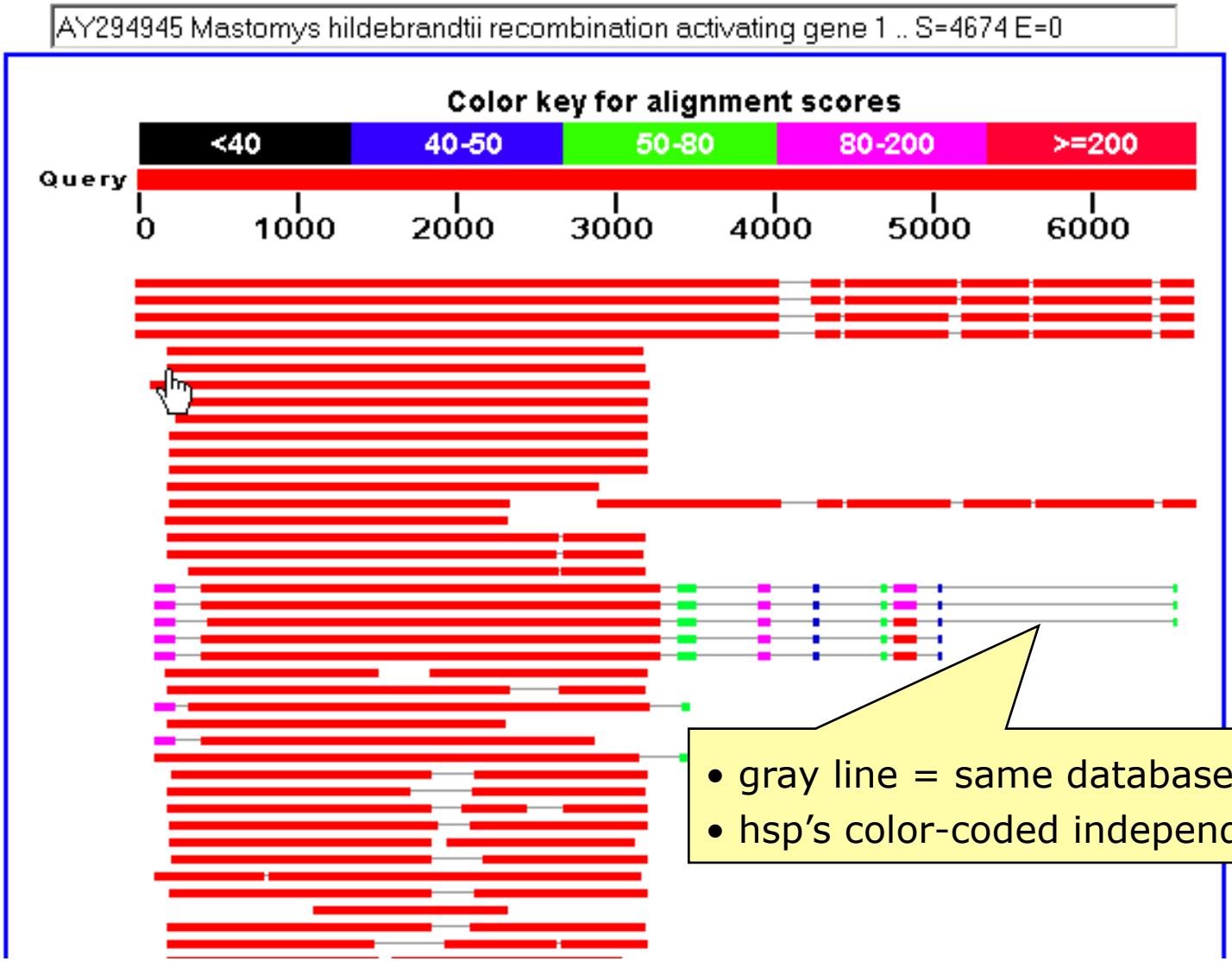
Function:
To create a virtual database representing a subset of the target database entries with features specified by the query terms.

Goal:
To make the search more specific, efficient, and less error-prone.

Limit by Entrez Query Term Field	Meaning
all[Filter] NOT mammalia[Organism]	non-mammalian entries
human[Organism]	human entries
20000:75000[mlwt]	Proteins 20 - 75 kD in size
biomol mrna[Properties]	mRNA entries
biomol genomic[Properties]	Genomic entries
1000:3000[slen]	Entries 1000 – 3000 k long

Other Advanced Field	Meaning
-e 10000	Set expect value to 10000
-b 2000	Set alignments to 2000

New BLAST Graphical Output



- gray line = same database hit
- hsp's color-coded independently

Predetermine The "Look" of Your Result: Format Section

Links to other dbs

Change to get XML

Display Retrieval Buttons

Summary Graphic

XML

Format

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

Use new formatter ☑ Masking Character Default(X for protein, n for nucleotide) Masking Color Black

Number of: Descriptions 500 Alignments 250

Alignment view Pairwise

Click to change alignment display, to "**Hit Table**" for example.

Format for PSI-BLAST ☐ with inclusion threshold: 0.005

Limit results by entrez query bacteria[orgn] AND All organisms

Expect value range:

Display only alignments with EXPCT values between the specified range

2e-40 **1e-2**

Example1

There are several ways that you can use BLAST to find SNPs

- BLAST2 Sequences
- SNP BLAST
  - Pairwise with Identities
    - Mismatches [SNP's] highlighted in red
- Traditional BLAST with altered alignment view
  - Flat query anchored with identities

SNP BLAST:
Finding coding SNPs in Cyp2C9

NCBI

dbSNP homepage

BLAST Home Page
BLAST overview
BLAST FAQs
BLAST news
BLAST manual

Single Nucleotide Polymorphism

Select the BLAST program

Program blastn Use Megablast Yes No

Choose a snp blast database

GenBank Division	snp blast database by organism					
Primate	chimpanzee	chimpanzee	chimpanzee	human		
Rodent	mouse	rat				
Other Mammal	bison	cow	pig	sheep		
Other Vertebrate	Collared_flycatcher	European_pied_flycat	bee	chicken	trout	zebrafish
Invertebrate	Nematodes	elegans	fruitfly	mosquito	plasmodium	
Plant	arabidopsis	corn	pine	rice	soybean	sugarcane

Click to blast human snp database by chromosome.

Query Sequence

Enter your sequence as: FASTA format Submit Query

```
>gi|13699817|ref|NM_000771.2| Homo sapiens
cytochrome P450, family 2, subfamily C,
polypeptide 9 (CYP2C9), mRNA
ATGGATTCTCTTGTGGTCCTTGTGCTCTGTCTCTCATGTTTGCTTCTCCT
CTGGGAGAGGAAAACTCCCTCCTGGCCCCACTCCTCTCCCAGTGATTGGA
TAAGGACATCAGCAAATCCTTAACCAATCTCTCAAAGGTCTATGGCCCGG
```

Reformatting of BLAST Search Results

The request ID is 1130418236-29933-176980377796.BLASTQ3

Format! or Reset all

The results are estimated to be ready in 20 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show ☑ Graphical Overview ☐ Linkout ☐ Sequence Retrieval ☐ NCBI-gi Alignment in HTML format

Masking Character Default(X for protein, n for nucleotide) Masking Color Black

Number of: Descriptions 100 Alignments 100

Alignment view Pairwise with identities

- Pairwise
- Pairwise with identities
- query-anchored with identities
- query-anchored without identities
- flat query-anchored with identities
- flat query-anchored without identities
- Hit Table

Start formatting from query #

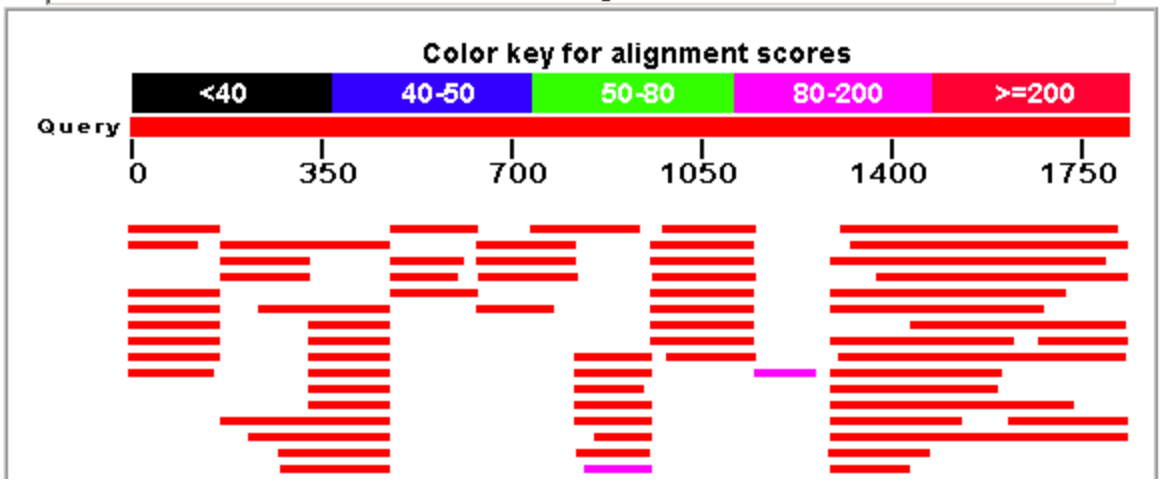
Limit results by entrez query organisms

Expect value range:

Results file ☐

Distribution of 107 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs9332241	allelePos=256totallen=511;taxid=9606;snpC...	958	0.0
gnl|dbSNP|rs9332242	allelePos=256totallen=511;taxid=9606;snpC...	956	0.0
gnl|dbSNP|rs9332240	allelePos=256totallen=511;taxid=9606;snpC...	956	0.0
gnl|dbSNP|rs9332243	allelePos=256totallen=511;taxid=9606;snpC...	865	0.0
gnl|dbSNP|rs9332239	allelePos=256totallen=511;taxid=9606;snpC...	823	0.0
gnl|dbSNP|rs1057911	allelePos=256totallen=511;taxid=9606;snpC...	746	0.0
gnl|dbSNP|rs9332244	allelePos=256totallen=476;taxid=9606;snpC...	740	0.0
gnl|dbSNP|rs17847030	allelePos=201totallen=401;taxid=9606;snp...	640	2e-180
gnl|dbSNP|rs17882796	allelePos=256totallen=511;taxid=9606;snp...	639	9e-180
gnl|dbSNP|rs2017319	allelePos=262totallen=542;taxid=9606;snpC...	598	1e-167
gnl|dbSNP|rs1934969	allelePos=122totallen=533;taxid=9606;snpC...	583	5e-163
gnl|dbSNP|rs5787121	allelePos=401totallen=801;taxid=9606;snpC...	519	7e-144
gnl|dbSNP|rs9332245	allelePos=256totallen=300;taxid=9606;snpC...	408	2e-110
gnl|dbSNP|rs17420162	allelePos=101totallen=201;taxid=9606;snp...	385	2e-103

SNP BLAST with Pairwise Alignment View

```
>gnl|dbSNP|rs9332241 allelePos=256totallen=511;taxid=9606;snpClass=1;alleles='C/T';mol=genomic;build=119
          Length=511

 Score =  958 bits (498),  Expect = 0.0
 Identities = 507/512 (99%), Gaps = 2/512 (0%)
 Strand=Plus/Plus

Query  1306  GTGGGAGAAGCCCTGGCCGGCATGGAGCTGTTTTTATTCCTGACCTCCATTTTACAGAAC  1365
Sbjct  1     ............................................................  60

Query  1366  TTTAACCTGAAATCTCTGGTTGACCCAAAGAACCTTGACACCACTCCAGTTGTCAATGGA  1425
Sbjct  61    ............................................................  120

Query  1426  TTTGCCTCTGTGCCGCCCTTCTACCAGCTGTGCTTCATTCCTGTCTGAAGAAGAGCAGAT  1485
Sbjct  121   ............................................................  180

Query  1486  GGCCTGGCTGCTGCTGTGCAGTCCCTGCAGCTCTCTTTCCTCTGGGGCATTATCCATCTT  1545
Sbjct  181   ............................................................  240

Query  1546  TGCACTATCTGTAATGCCTTTTCTCACCTGTCATCTCACATTTTCCCTTCCCTGAAGATC  1605
Sbjct  241   .-..............Y.........................................  299

Query  1606  TAGTGAACATTCGACCTCCATTACGGAGAGTTTCCTATGTTTCACTGTGCAAATATATCT  1665
Sbjct  300   ............................................................  359

Query  1666  GCTATTCTCCATACTCTGTAACAGTTGCATTGACTGTCACATAATGCTCATACTTATCTA  1725
Sbjct  360   ............................................................  419

Query  1726  ATGTAGAGT-ATTAATATGTTATTATTAAATAGAGAAATATGATTTGTGTATTATAATTC  1784
Sbjct  420   ....T....T.................................................  479

Query  1785  AAAGGCATTTCTTTTCTGCATGATCTAAATAA  1816
Sbjct  480   ......................T.........  511
```

Example 2

Mining Human EST Data for Biologically Significant Sequence Polymorphisms

1: P04156. Reports Major prion prote...[gi:130912]

BLink, Conserved Domains, Links

```
LOCUS       P04156                   253 aa            linear   PRI 01-MAY-2005
DEFINITION  Major prion protein precursor (PrP) (PrP27-30) (PrP33-35C) (ASCR)
            (CD230 antigen).
ACCESSION   P04156
VERSION     P04156  GI:130912
DBSOURCE    swissprot: locus PRIO HUMAN, accession P04156;
```

```
Region          129..130
                /gene="PRNP"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          129
                /gene="PRNP"
                /region_name="Variant"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="M -> V (polymorphism; determines the disease
                phenotype in patients who have a PrP mutation at position
                178. Patients with M-129 develop FFI, those with V-129
                develop CJD; dbSNP:1799990). /FTId=VAR_006467."
Region          131
                /gene="PRNP"
                /region_name="Variant"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="G -> V (in GSD). /FTId=VAR_014264."
```

TBLASTN Search of est_human

NCBI translating BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

Search

```
>gi|130912|sp|P04156|PRIO_HUMAN Major prion protein
precursor (PrP) (PrP27-30) (PrP33-35C) (ASCR) (CD230
antigen)
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH
GGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGGLGGYMLG
```

Choose a translation: PROTEIN query - TRANSLATED database [tblastn]

Set subsequence: From: To:

Choose database: est_human

Genetic codes: Disabled

Now: BLAST! or Reset query Reset all

Format the "Alignment View"

Format

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

Masking Character Default(X for protein, n for nucleotide) Masking Color Black

Number of: Descriptions 100 Alignments 50

Alignment view flat query-anchored with identities

- Pairwise
- Pairwise with identities
- query-anchored with identities
- query-anchored without identities
- flat query-anchored with identities
- flat query-anchored without identities
- Hit Table

Limit results by entrez query organisms

Expect value range:

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

Autoformat Semi-auto

BLAST! or Reset all

```
Query     121  XXXXXXXXXXXXXXXXRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV  180
46925984  410  ..........................................................  589
45857930  416  ..........................................................  595
45749230  403  M.........................................................  582
22697249  397  ..........................................................  576
22659989  401  ..........................................................  580
22271655  407  ..........................................................  586
20405989  374  ..........................................................  553
19370296  440  ..........................................................  619
18520023  386  ..........................................................  565
31446754  402  ..........................................................  581
45703565  375  ........V.................N.K...............................  554
46921643  412  ..........................Y...............................  591
34889317  410  ..........................................................  589
66264383  443  ..........................................................  622
22285044  409  ..........................................................  588
11002886  425  ..........................................................  604
21857720  429  ...........................D..A...........................  608
22662958  453  ..........................................................  632
15492735  429  ........V.................................................  608
15493048  425  ........V.................................................  604
45751517  408  ..........NKXXXXXX.......................................  587
14001854  408  ........V.................................................  587
13984551  430  ........V.......R..................R.....................  609
15440704  433  ........V.................................................  612
13987308  426  ........V.................................................  505
13976202  426  ........V.................................................  505
22705803  412  ..........................................................  591
13967872  433  ........V.................................................  512
15496140  426  ........V.................................................  505
```

Select EST sequences of interest to link from Entrez Nucleotide to UniGene to find the EST library.

Genomic BLAST

- Finding a Homolog in a Distant Organism
- Mapping Oligo's to the Genome
- Determining Gene Structure

Genome BLAST via Map Viewer

Click on the organism name to go to the genome view

Switch to Graphical View

Vertebrates

Mammals

- BLAST *Bos taurus* (cow)
- BLAST *Canis familiaris* (dog)
- BLAST *Felis catus* (cat)
- BLAST *Homo sapiens* (human)
- BLAST *Mus musculus* (mouse)
- BLAST *Ovis aries* (sheep)
- BLAST *Pan troglodytes* (chimpanzee)
- BLAST *Rattus norvegicus* (rat)
- BLAST *Sus scrofa* (pig)

Other Vertebrates

- BLAST *Danio rerio* (zebrafish)
- BLAST *Gallus gallus* (chicken)

Invertebrates

Insects BLAST

- BLAST *Anopheles gambiae* (mosquito)
- BLAST *Apis mellifera* (honey bee)
- BLAST *Drosophila melanogaster* (fruit fly)

Nematode BLAST

Caenorhabditis elegans (nematode)

Protozoa

Plants BLAST Search all plant maps

- BLAST *Arabidopsis thaliana* (thale cress)
- BLAST *Avena sativa* (oat)
- BLAST *Glycine max* (soybean)
- BLAST *Hordeum vulgare* (barley)
- BLAST *Lycopersicon esculentum* (tomato)
- BLAST *Oryza sativa* (rice)
- BLAST *Triticum aestivum* (wheat)
- BLAST *Zea mays* (corn)

Fungi BLAST

- BLAST *Candida glabrata*
- BLAST *Debaryomyces hansenii*
- BLAST *Encephalitozoon cuniculi*
- BLAST *Eremothecium gossypii*
- BLAST *Gibberella zeae*
- BLAST *Kluyveromyces lactis*
- BLAST *Magnaporthe grisea*
- BLAST *Neurospora crassa*
- BLAST *Saccharomyces cerevisiae* (baker's yeast)
- BLAST *Schizosaccharomyces pombe* (fission yeast)
- BLAST *Yarrowia lipolytica*

Mapping of PAH in Chicken Genome

NCBI Home ▸ Genomic Biology ▸ Chicken Genome Resources ▸ BLAST

Search Map Viewer Go Clear

BLAST
overview
FAQs
news
manual
references

Blast Chicken Sequences

Blast your sequence against Chicken specific sequences

Database: genome (reference only)

Program megaBLAST: Compare highly related nucleotide sequences

- megaBLAST: Compare highly related nucleotide sequences
- cross-species megaBLAST: Compare distantly related nucleotide sequences
- BLASTN: Compare nucleotide sequences
- BLASTP: Compare protein sequences
- BLASTX: Compare a nucleotide sequence against a protein database
- TBLASTN: Compare a protein sequence against a nucleotide database

Begin S

Enter an

```
>Human
CAGCTGGGGGTAAGGGGGGCGGATTATTCATATAATTGTTATACCAGACGGTCGCAGGCTTAGTCCAATT
GCAGAGAACTCGCTTCCCAGGCTTCTGAGAGTCCCGGAAGTGCCTAAACCTGTCTAATCGACGGGGCTTG
GGTGGCCCGTCGCTCCCTGGCTTCTTCCCTTTACCCAGGGCGGGCAGCGAAGTGGTGCCTCCTGCGTCCC
CCACACCCTCCCTCAGCCCCTCCCCTCCGGCCCGTCCTGGGCAGGTGACCTGGAGCATCCGGCAGGCTGC
CCTGGCCTCCTGCGTCAGGACAAGCCCACGAGGGGCGTTACTGTGCGGAGATGCACCACGCAAGAGACAC
CCTTTGTAACTCTCTTCTCCTCCCTAGTGCGAGGTTAAAACCTTCAGCCCCACGTGCTGTTTGCAAACCT
GCCTGTACCTGAGGCCCTAAAAAGCCAGAGACCTCACTCCCGGGGAGCCAGCATGTCCACTGCGGTCCTG
GAAAACCCAGGCTTGGGCAGGAAACTCTCTGACTTTGGACAGGAAACAAGCTATATTGAAGACAACTGCA
ATCAAAATGGTGCCATATCACTGATCTTCTCACTCAAAGAAGAAGTTGGTGCATTGGCCAAAGTATTGCG
```

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	low complexity	100	100

Advanced options:

Chicken Genome BLAST: Genome View

Genome View

Show positions of the BLAST hits in the chicken genome using the Entrez Genomes MapViewer

Query= Human phenylalanine hydroxylase (PAH), mRNA
(2680 letters)

Distribution of 13 Blast Hits on the

Mouse over to see the defline, click to show alignments

Color key for alignme

<40 | 40-50 | 50-80

Query

0 500 1000 1500

Sequences producing significant al

ref|NW_060209.1|Gga1_WGA18_1 Gall
ref|NW_060378.1|Gga5_WGA187_1 Gal

Gallus gallus (chicken) genome view

BLAST search the chicken genome

Build 1.1 statistics

1 2 3 4 5 6 7 8 9 10 11 12 13

Hit GIs: 1 1
Hits: 10 3

14 15 16 17 18 19 20 21 22 23 24 25 26

Hit GIs:
Hits:

27 28 29 30 31 32 33 34 35 36 37 38 W

Hit GIs:
Hits:

Z LGE22C19W28 LGE64 LGE26C13 LGE50C23 MT

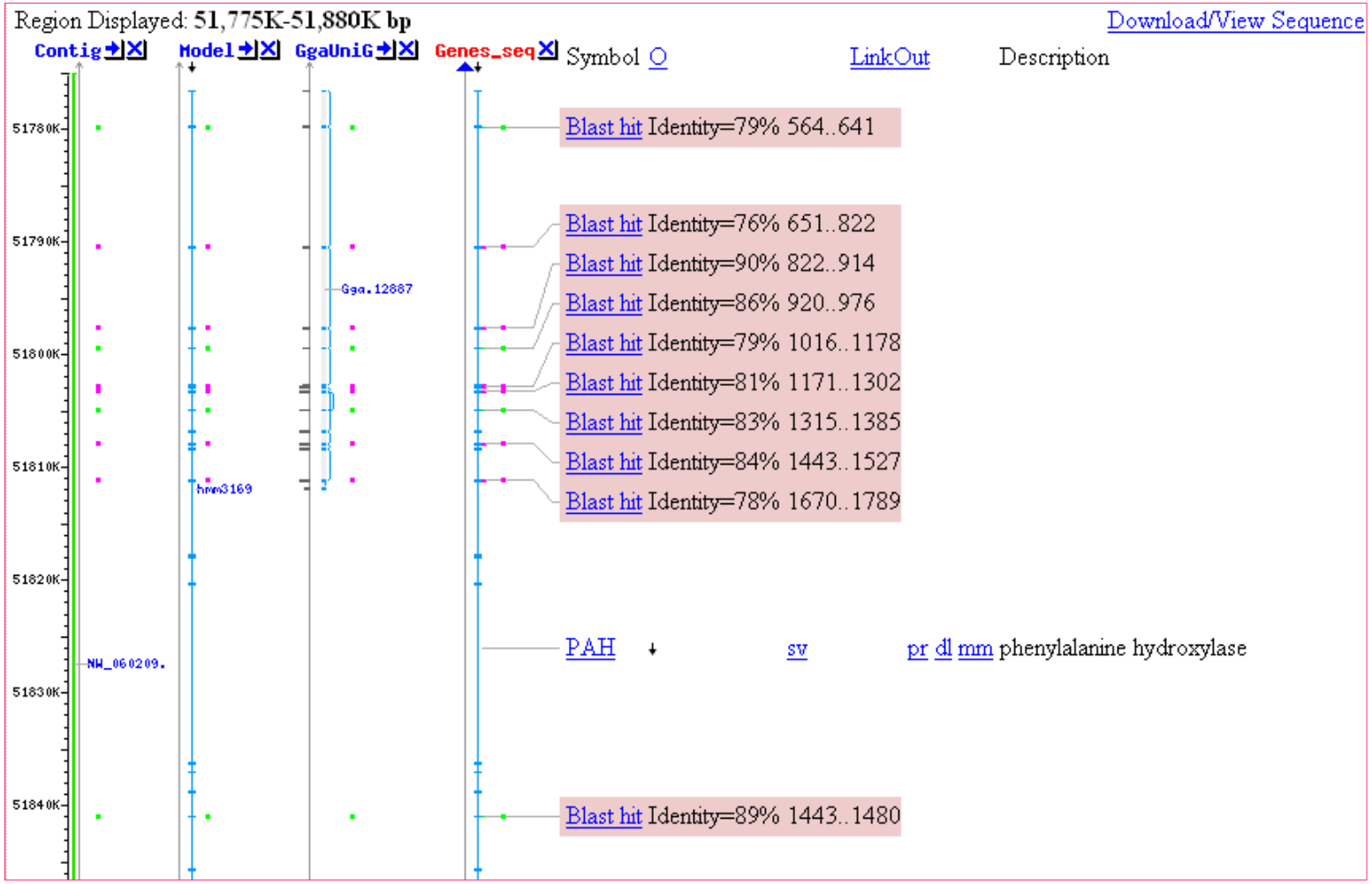
>ref|NW_060209.1|Gga1
sequence
Length=27618

Features in this part
phenylalanine hydro

Score = 127 bits (66
Identities = 84/93 (9
Strand=Plus/Plus

```
Query  822       CAGTGCCCTGGTTCCCAAGAACCATTCAAGAGCTGGACAGATTTGCCAATCAGATTCTCA  881
                 |||| |||||||||||||||  || || |||||||||||||||||||||||||| || |
Sbjct  11490574  CAGTTCCCTGGTTCCCAAGAAGTATCCAGGAGCTGGACAGATTTGCCAATCAGATCCTAA  11490633

Query  882       GCTATGGAGCGGAACTGGATGCTGACCACCCTG  914
                 ||||||||||||| |||||||||||||| ||||
Sbjct  11490634  GCTATGGAGCGGAGCTGGATGCTGACCATCCTG  11490666
```

BLAST "Genome View": Aligning BLAST Hits to the Genome



Map Oligos Onto Genome

NCBI Home ▸ Genomic Biology ▸ Human Genome Guide

Search LocusLink for Go

BLAST
- overview
- FAQs
- news
- manual
- references

Blast the Human Genome

Blast your sequence against Human specific sequences

Database: genome Program blastn

☐ use MegaBLAST

Begin Search

Enter an accession, gi, or a sequence in FASTA format:

>CCATGGCGACCCTGGAAAAGCNNNNNNNNNNCAGCAGCGGCTGTGCCTGCGG

forward primer reverse primer

Optional parameters

Expect Filter Descriptions Alignments

1 default 100 100

0.0001
0.001
0.01
1
10

ed options: -W 7 –e 1000

earch Clear Input

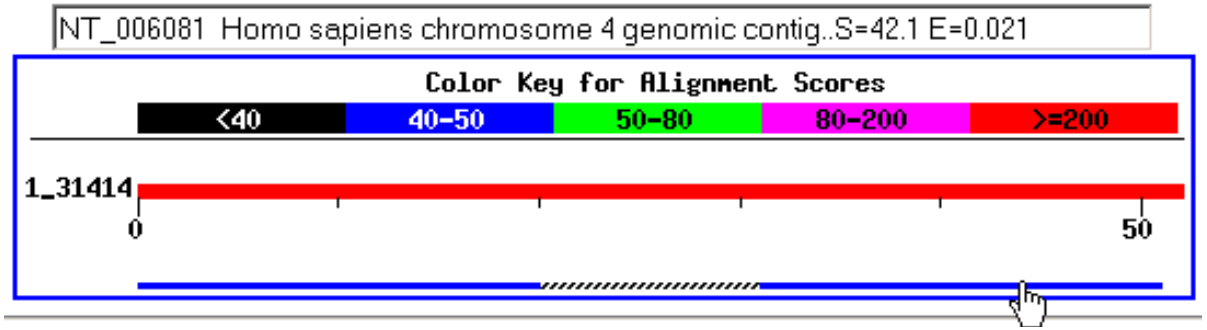
Genome BLAST Results

```
RID: 1076295772-31414-177358914251.BLASTQ3

Query= CCATGGCGACCCTGGAAAAGCNNNNNNNNNNCAGCAGCGGCTGTGCCTGCGG
         (52 letters)


Database: contig
           498 sequences; 3,020,300,271 total letters


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

Genome View

Show positions of the BLAST hits in the human genome using the Entrez Genomes MapViewer

Distribution of 2 Blast Hits on the Query Sequence



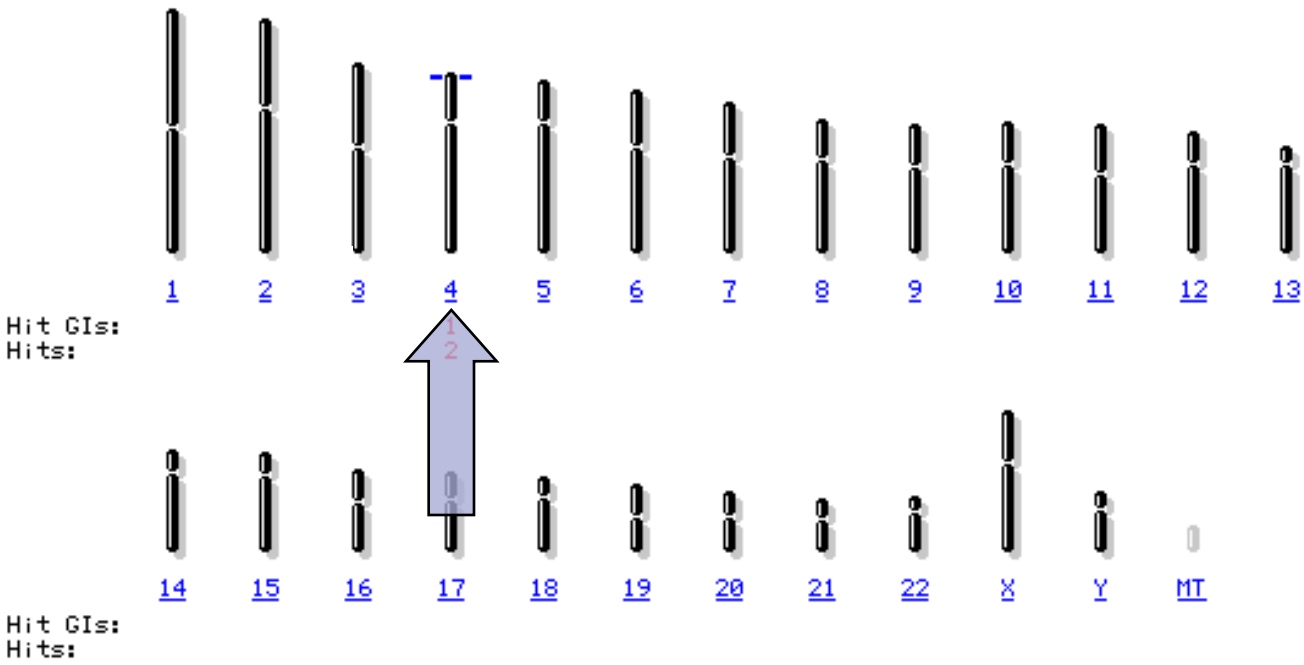
Primer Alignments

```
>ref|NT_006081.16|Hs4_6238 Homo sapiens chromosome 4 genomic contig
          Length = 1182262

 Score = 42.1 bits (21), Expect = 0.021
 Identities = 21/21 (100%)
 Strand = Plus / Minus


Query: 32      cagcagcggctgtgcctgcgg 52
               |||||||||||||||||||||
Sbjct: 463315 cagcagcggctgtgcctgcgg 463295




 Score = 42.1 bits (21), Expect = 0.021
 Identities = 21/21 (100%)
 Strand = Plus / Plus


Query: 1       ccatggcgaccctggaaaagc 21
               |||||||||||||||||||||
Sbjct: 463128 ccatggcgaccctggaaaagc 463148
```

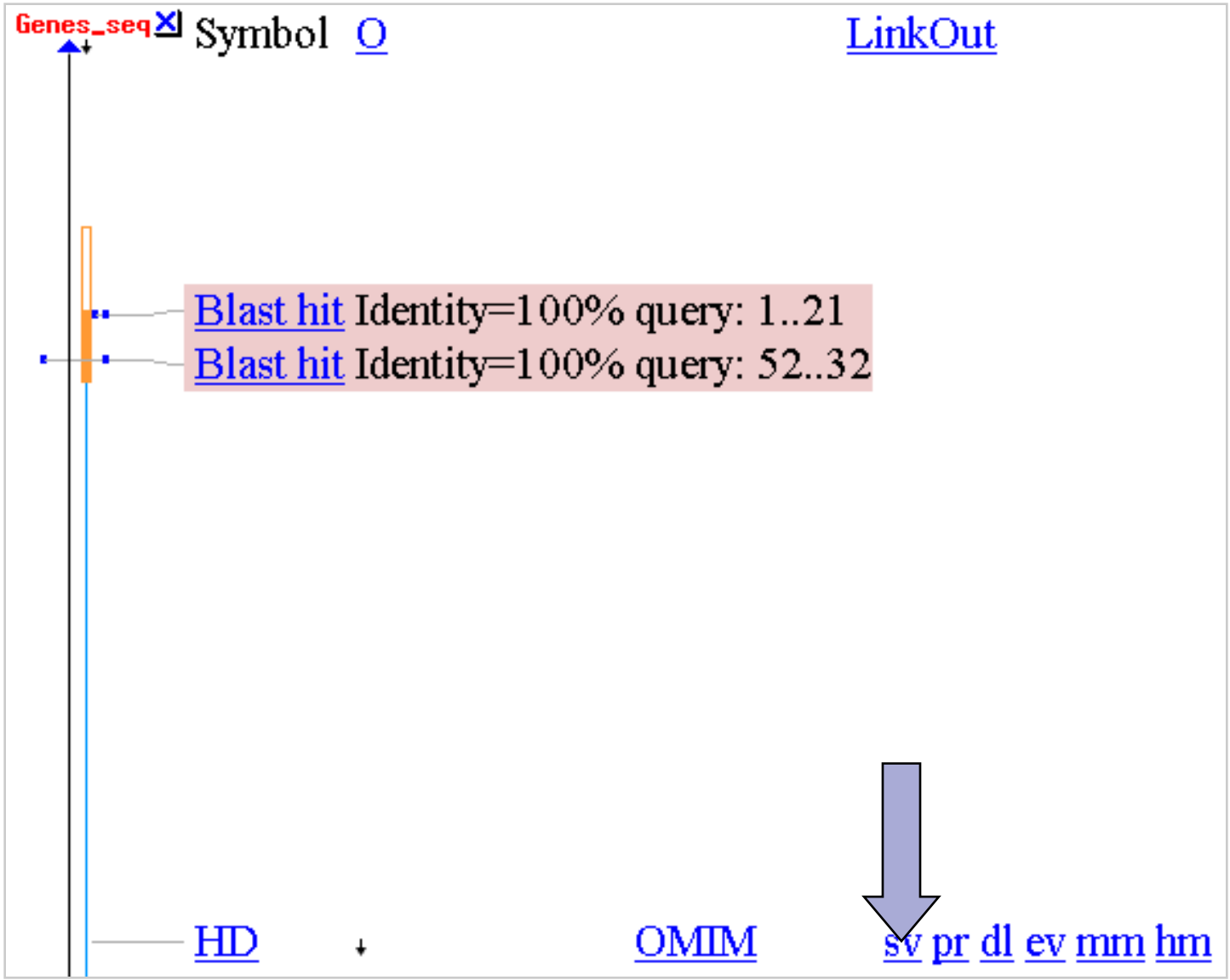
reverse primer

forward primer

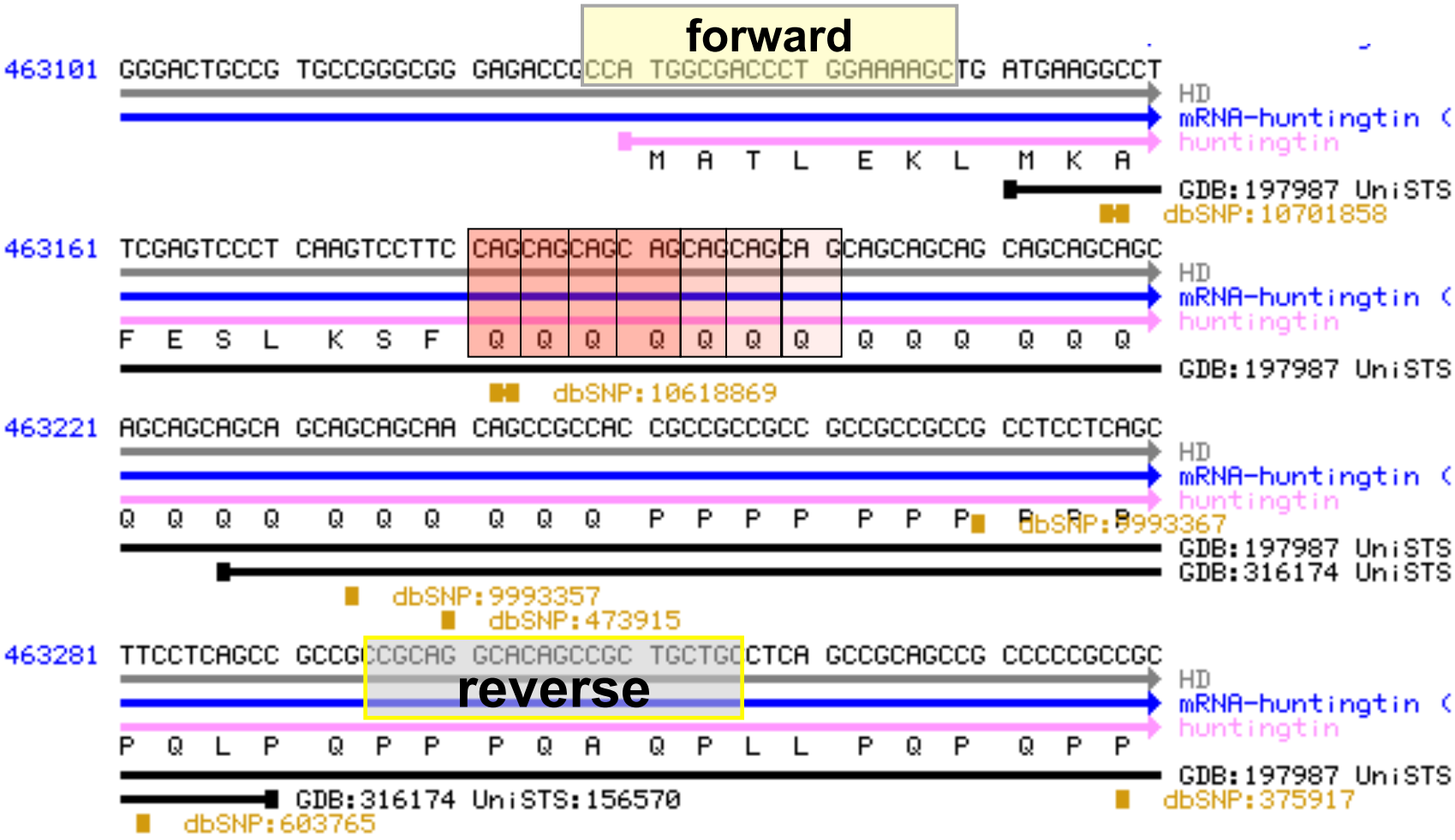
MapViewer


1
2
3
4
5
6
7
8
9
10
11
12
13
Hit GIs:
Hits:
14
15
16
17
18
19
20
21
22
X
Y
MT
Hit GIs:
Hits:

MapViewer


Genes_seq
Symbol
O
LinkOut
Blast hit Identity=100% query: 1..21
Blast hit Identity=100% query: 52..32
HD
OMIM
sv pr dl ev mm hm

Sequence View (sv)



BLAST Educational Resources

Literature

Training Courses/Workshops

The NCBI Newsletter: BLAST Labs

NCBI News

In this issue

BLAST Lab

```
caaatccgttcttgatcgtacatagcgcatgtcagncaaatc
||||||| |||||||||||||||| || |||||||| ||||||
caaatccattcttgatcgtacatggcacatgtcagtcaaatc
```

Using seedtop to find patterns in protein and nucleotide sequences

Seedtop is one of the programs included within the NCBI standalone BLAST package and is used to find matches to a pattern in a protein or nucleotide sequence database.

Using seedtop to locate a pattern in protein and nucleotide sequences

Seedtop, like blastall and formatdb, is a commandline program with parameters specified with a leading dash, followed by a one-letter parameter code. To find a pattern in a protein sequence, we may use:

```
seedtop -i input -k pat -p patmchp -o pat_out
```

The file "pat" contains the pattern for a serine protease motif:

```
ID Serine Protease Motif, cd00190
PA C-[AVLS]-X(3,9)-[DSNAR]-X-[CG]-X-[GSR]-[DE]-[SAPG]-G-[GS]-[PAG]-[LFMV]
```

The NCBI Handbook

The NCBI Handbook The National Library of Medicine

Short Contents | Full Contents Other books @ NCBI

Navigation

The NCBI Handbook → **Part 3. Querying and Linking the Data**

Created: October 9, 2002

Updated: August 13, 2003

16. The BLAST Sequence Analysis Tool

by Tom Madden

Summary

The comparison of nucleotide or protein sequences from the same or different organisms is a very powerful tool in molecular biology. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships. Now that whole genomes are being sequenced, sequence similarity searching can be used to predict the location and function of protein-coding and transcription-regulation regions in genomic DNA.

Basic Local Alignment Search Tool (BLAST) (1, 2) is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases. All BLAST applications, as well as information on which BLAST program to use and other help documentation, are listed on the BLAST homepage.

NCBI → BLAST

Latest news: 28 August 2005 : BLAST 2.2.12 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

BLAST Program Selection Guide

By blast-help group, NCBI User Service
NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894

Table of Content

1. Introduction

NCBI has provided BLAST sequence analysis services for over a decade. For many users, the first question they face is *"Which BLAST program should I use?"*

In order to help users arrive at an answer to this question, we have constructed this table called the "BLAST Program Selection Guide." It is divided into several categories according to the *nature* and *size* of the query and the primary goal of the search. Starting from the query sequence on the left and cross-referencing to the right, an user will arrive the specific BLAST program best suited for that search.

This document is also available in PDF (1056656 bytes).

Avoid the Lines

Precomputed BLAST Services

- Nucleotide or protein: Related Sequences
- BLAST link: BLink

- Transcript clusters: UniGene
- Protein homologs: HomoloGene

BLAST Technical Assistance

BLAST help at Whitehead:

wibr-bioinformatics@wi.mit.edu

NCBI contact information:

General questions other than BLAST:

info@ncbi.nlm.nih.gov

BLAST specific Questions:

blast-help@ncbi.nlm.nih.gov (preferred route)

NCBI "Hotline" (8.30 am-5:00 pm EST):

(301) 496-2475

Information needed for troubleshooting BLAST problems:

- RID
- Query, BLAST page used
- Database and search parameters
- Error messages encountered
- Computer platform and BLAST version
- Command line used for BLAST and formatdb

The END! Thank you.

Slides Taken from: NCBI talk at American Society of Human Genetics
October 2005