NCBI Public Services

An Update on NCBI BLAST

Slides: ftp.ncbi.nih.gov/pub/cooper/ASCB2009

NCBI Exhibit Booth: 633

American Society for Cell Biology
December 2009

Peter Cooper
The National Center for Biotechnology Information

NCBI Public Services

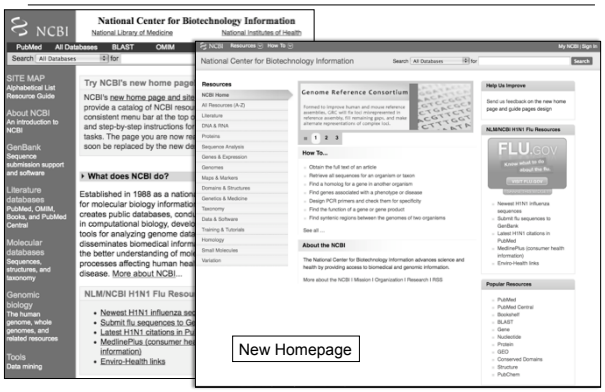
The National Center for Biotechnology Information

Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

NCBI Public Services

Web Access: www.ncbi.nlm.nih.gov



NCBI Public Services

What we will cover

- Basics of using NCBI BLAST
- Using the new Interface
  - Improved organism and filter options
- New Services
  - Primer BLAST
  - Align 2 Sequences Integration
  - COBALT – protein multiple alignment
  - SRA BLAST Service
- C++ BLAST binaries

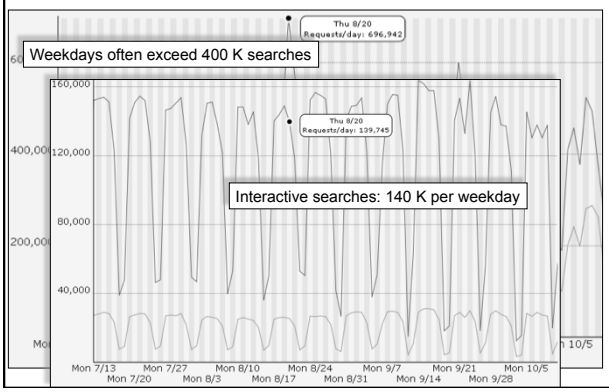
Basic Local Alignment Search Tool

NCBI Public Services

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network client

BLAST Activity

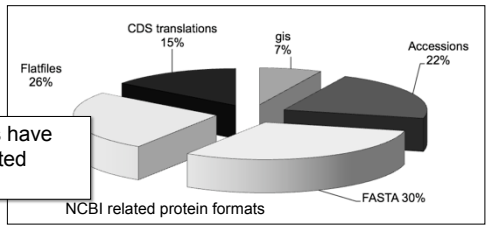
NCBI Public Services



BLAST usage observations

NCBI Public Services

- 20% of interactive nucleotide searches are short queries, under 50 bases (primers)
- Up to 20% of protein queries seem to be copy – pasted NCBI formats.



All proteins have pre-computed similarities!

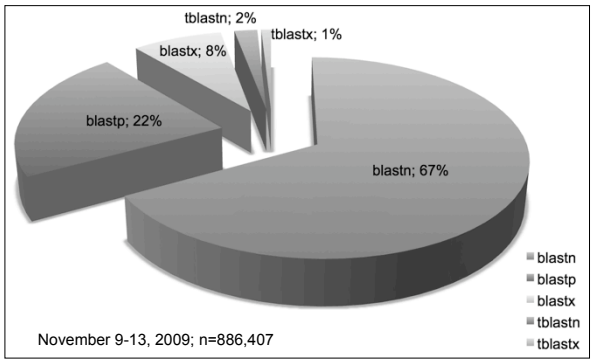
BLAST and BLAST-like programs

NCBI Public Services

- **Traditional BLAST (formerly blastall) nucleotide, protein, translations**
  - **blastn nucleotide query vs. nucleotide database**
  - **blastp protein query vs. protein database**
  - **blastx nucleotide query vs. protein database**
  - **tblastn protein query vs. translated nucleotide database**
  - **tblastx translated query vs. translated database**
- **Megablast nucleotide only**
  - **Contiguous megablast**
    - **Nearly identical sequences**
  - **Discontiguous megablast**
    - **Cross-species comparison**
- **Position Specific BLAST Programs protein only**
  - **Position Specific Iterative BLAST (PSI-BLAST)**
    - **Automatically generates a position specific score matrix (PSSM)**
  - **Reverse PSI-BLAST (RPS-BLAST)**
    - **Searches a database of PSI-BLAST PSSMs**

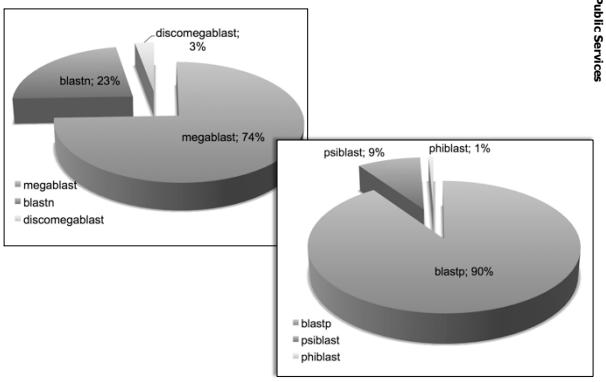
BLAST Searches by Program

NCBI Public Services



Nucleotide and Protein BLAST Programs

NCBI Public Services



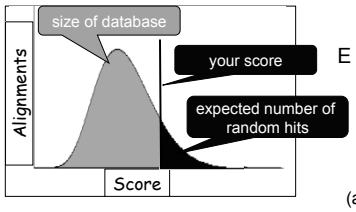
Local Alignment Statistics

NCBI Public Services

High scores of local alignments between two random sequences follow the Extreme Value Distribution

Expect Value

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



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

K = scale for search space
λ = scale for scoring system
$S' = \text{bitscore} = (\lambda S - \ln K)/\ln 2$

(applies to ungapped alignments)

The BLAST homepage

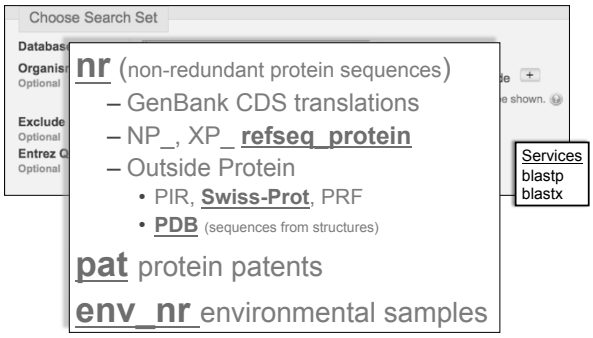
http://blast.ncbi.nlm.nih.gov/

NCBI Public Services

Basic BLAST: Databases

Non-redundant protein

NCBI Public Services



nr (non-redundant protein sequences)
- GenBank CDS translations
- NP_, XP_ **refseq_protein**
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
  - **PDB** (sequences from structures)

pat protein patents

env_nr environmental samples

Protein Database Sizes

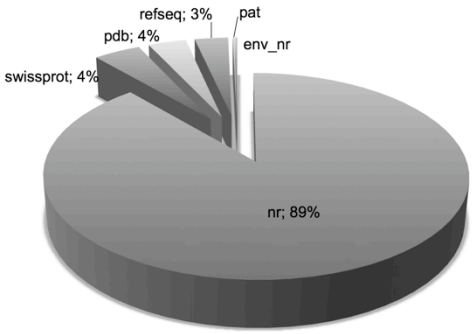
NCBI Public Services

Database	Sequences	Residues
nr	10,133,783	3,456,922,644
refseq_protein	7,413,069	2,589,005,568
swissprot	430,511	159,291,105
pat	817,680	166,184,433
pdb	44,202	10,171,945

12/04/2009

Protein Database Selection

NCBI Public Services

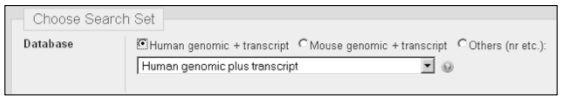


November 9-13, 2009; n=222,791

Nucleotide Databases: Human and Mouse

NCBI Public Services

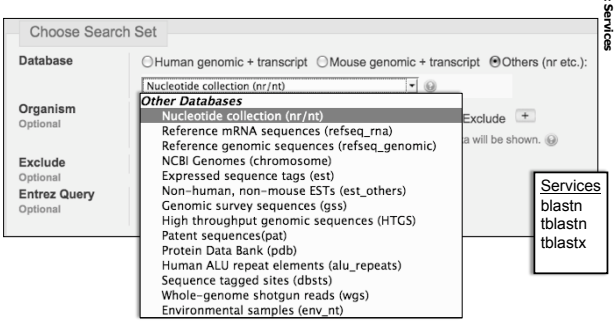
Megablast, blastn service



- Human and mouse genomic and transcript now default
- Separate sections in output for mRNA and genomic
- Direct links to Map Viewer for genomic sequences

Nucleotide Databases: Traditional

NCBI Public Services



Nucleotide Databases: Traditional

NCBI Public Services

Databases are mostly non-overlapping

- **nr (nt)**
  - Traditional GenBank
  - NM_ and XM_ RefSeqs
    - **refseq_rna**
- **NCBI Genomes**
  - NC_ RefSeqs
  - GenBank Chromosomes
- **dbest**
  - EST Division
    - **non-human, non-mouse ests**

- **htgs**
  - HTG division
- **gss**
  - GSS division
- **wgs**
  - whole genome shotgun contigs
- **env_nt**
  - environmental samples

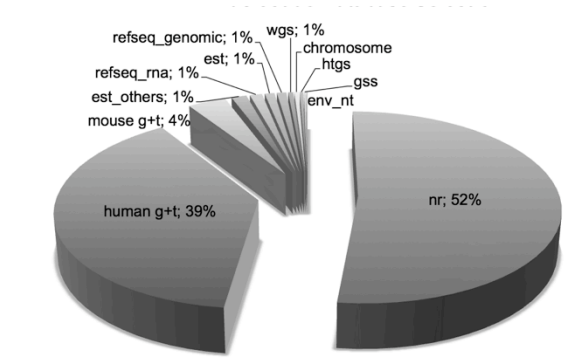
Nucleotide Database Sizes

NCBI Public Services

Database	Sequences	Residues
nr/nt	10,362,162	29,617,088,643
refseq_rna	2,042,538	3,240,301,155
NCBI genomes	10,047	49,094,451,709
est	63,832,451	35,136,825,005
htgs	143,742	24,082,224,044
gss	27,198,629	17,658,377,015
wgs	31,377,631	149,309,157,200
env_nt	17,708,548	7,218,208,433

12/04/2009

Nucleotide Database Selection



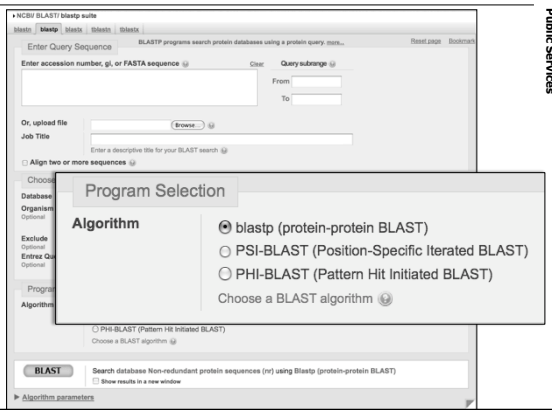
November 9-13, 2009; n=535,836

NCBI Public Services

Using Basic BLAST

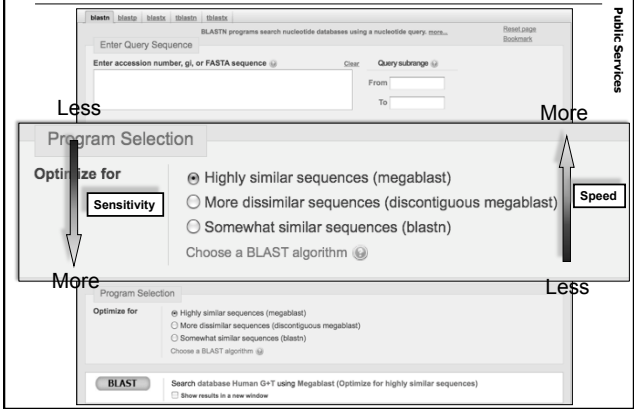
NCBI Public Services

Universal Form: Protein



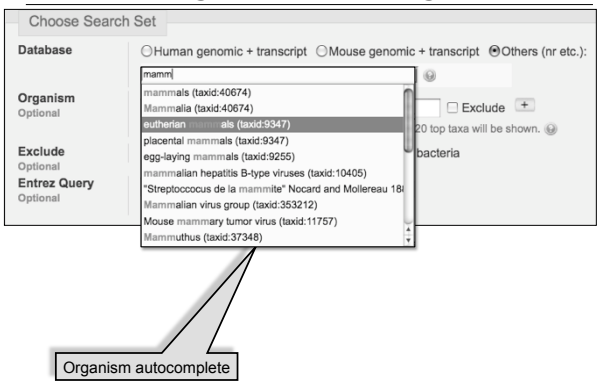
NCBI Public Services

Universal Form: Nucleotide



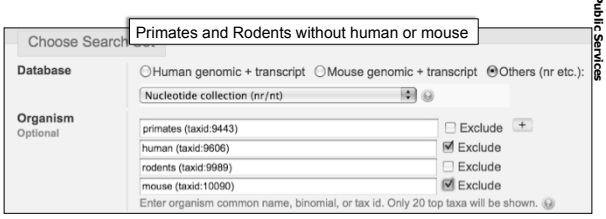
NCBI Public Services

Limiting Database: Organism



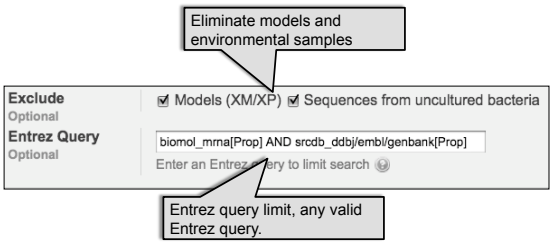
NCBI Public Services

Combining Organisms



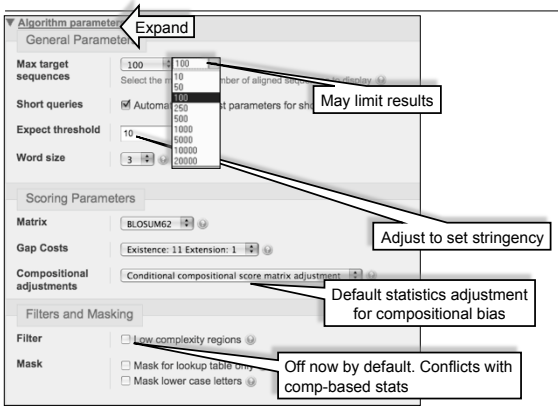
NCBI Public Services

More Limits



NCBI Public Services

Algorithm parameters: Protein



NCBI Public Services

Automatic Short Sequence Adjustment

NCBI Public Services

Job Title: Elvis Lives!

No putative conserved domains have been detected

Your search parameters were adjusted to search for a short input sequence.

Request ID	HGC9100G
Status	Searching
Submitted at	Thu Dec 3
Current time	Thu Dec 3
Time since submission	00:00:08

This page will be automatically updated in 5 seconds

Protein	
e-value	20000
Word Size	2
Matrix	PAM30
Comp Stats	Off
Low Comp Filter	Off

Nucleotide	
e-value	1000
Word Size	7
Matrix	1,-3
Low Comp Filter	Off

New Designing or Testing PCR Primers? Try your search in Primer-BLAST. Go

Algorithm parameters: Nucleotide

NCBI Public Services

General Parameters

Max target sequences: 100

Short queries: Automatically adjust parameters for

Expect threshold: 10

Word size: 11

blastn

•Masks species-specific interspersed repeats
•Essential for genomic query sequences

Human
Human
Rodents
Arabidopsis
Rice
Mammals
Fungi
C. elegans
A. gambiae
Zebrafish
Fruit fly

Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: Low complexity regions; Species-specific repeats for: Human

Mask: Mask for lookup table only; Mask lower case letters

Masks LC sequence (simple repeats)

•Prevents starting alignment in masked region
•Allows extensions through masked regions

NCBI Public Services

Basic BLAST: Protein

The hard way to run a BLAST Search

NCBI Public Services

1. Search protein with "Human Muscle Creatine Kinase"
2. Click on summary for NP_001815
3. Change format to FASTA
4. Select sequence
5. Copy sequence
6. Google search "BLAST"
7. Link to NCBI BLAST Homepage
8. Link to Protein BLAST form
9. Paste FASTA sequence into form
10. Click BLAST button

muscle creatine kinase [Homo sapiens]
1. 381 aa protein
NP_001815.2 GI:21536288

muscle creatine kinase [Homo sapiens]

NCBI B

ncbi blast — Suggestions
ncbi books
ncbi blast download
ncbi blast2
ncbi blast tutorial
ncbi biosystems

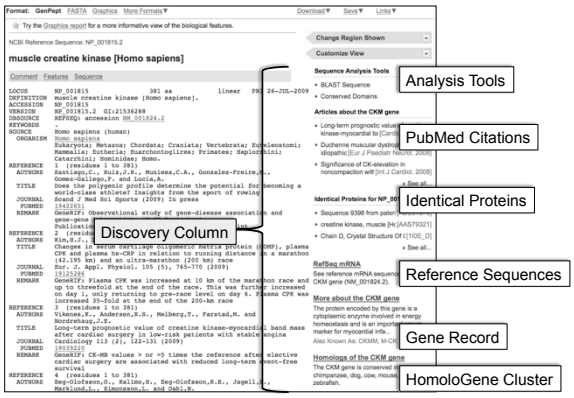
protein blast — Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast

Enter accession number, gi, or FASTA sequence — Clear

```
>gi|21536288|ref|NP_001815.2| muscle creatine kinase [Homo sapiens]
MPFGNTHNKFKLNYKPEEEYPDLSKHNNHMAKVLTLELYKKLRDKETPSGFTVDDVIQTGVDNPGHP
TVGCVAGDEESYEVFKELFDPIISDRHGGYKPTDKHKTDLNHENLKGGDDLDPNYVLSSRVRTGRSI
TLPPHCSRGERRAVEKLSVEALNSLTGEFKGKYYPLKSMTEKEQQQLIDDHFLFDKPVSPLLLASGM
WPDARGIWHNDNKSFLVWVNEEDHLRVISMEKGGNMKEVFRRFCVGLQKIEEIFKKAGHPFMWNQHL
GYV
```

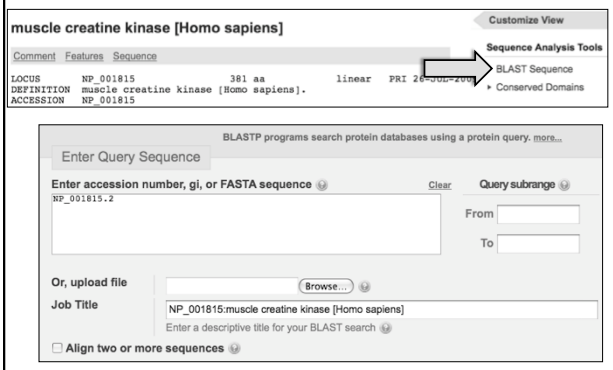
BLAST

An easier way: Entrez protein record



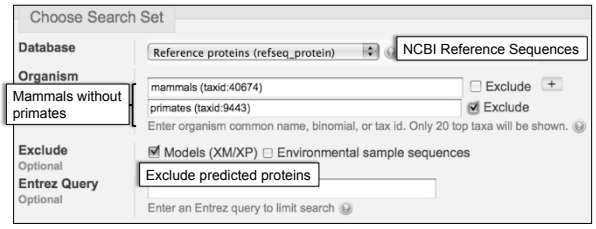
NCBI Public Services

BLAST Ad to BLAST form



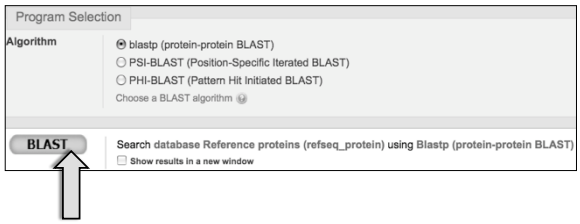
NCBI Public Services

Database and limits



NCBI Public Services

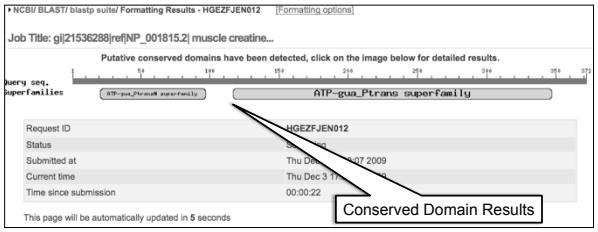
Run Search



NCBI Public Services

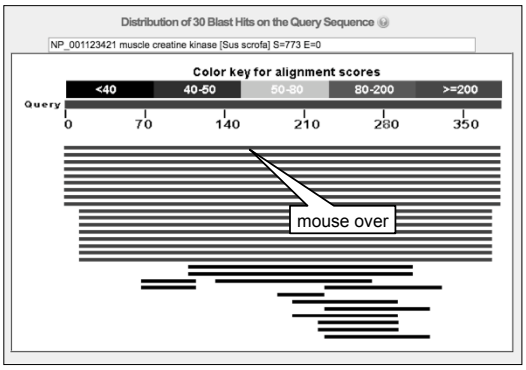
BLAST Formatting Page

NCBI Public Services



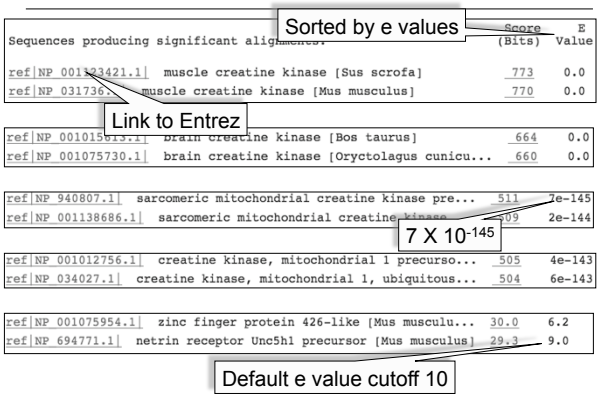
BLAST Output: Graphical Overview

NCBI Public Services



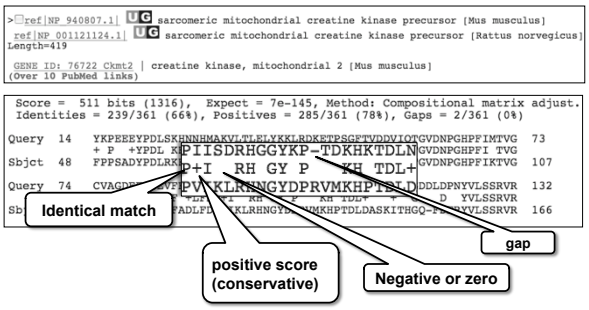
BLAST Output: Descriptions

NCBI Public Services



BLAST Output: Alignments

NCBI Public Services



What happens without XP_ filter?

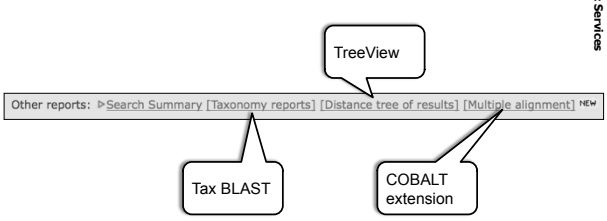
NCBI Public Services

```
Sequences producing significant alignments:                          Score    E
                                                                     (Bits)  Value

ref|XP_533641.2|  PREDICTED: similar to Creatine kinase M-type...     772    0.0
ref|XP_537561.2|  PREDICTED: similar to Creatine kinase B-type...     653    0.0
ref|XP_868659.1|  PREDICTED: similar to Creatine kinase B-type...     551    1e-156
ref|NP_001138686.1|  sarcomeric mitochondrial creatine kinase ...     509    6e-144
ref|XP_535451.1|  PREDICTED: similar to creatine kinase, mitoc...     503    4e-142
ref|XP_860111.1|  PREDICTED: similar to creatine kinase, mitoc...     501    2e-141
ref|XP_859467.1|  PREDICTED: similar to creatine kinase, mitoc...     498    2e-140
ref|XP_858430.1|  PREDICTED: similar to creatine kinase, mitoc...     497    3e-140
ref|XP_854438.1|  PREDICTED: similar to Creatine kinase B-type...     433    8e-121
ref|XP_853207.1|  PREDICTED: similar to creatine kinase, brain...     412    2e-114
ref|XP_851000.1|  PREDICTED: similar to creatine kinase, brain...     358    3e-98
ref|XP_853614.1|  PREDICTED: similar to Creatine kinase B-type...     318    2e-86
ref|XP_542948.2|  PREDICTED: similar to creatine kinase, brain...     239    2e-62
ref|XP_860041.1|  PREDICTED: similar to creatine kinase, mitoc...     224    6e-58
ref|XP_546543.2|  PREDICTED: similar to Creatine kinase B-type...     202    1e-51
ref|XP_853217.1|  PREDICTED: similar to Creatine kinase B-type...     199    1e-50
ref|XP_852237.1|  PREDICTED: similar to low-density lipoprotei...     197    8e-50
ref|XP_859393.1|  PREDICTED: similar to Creatine kinase, sarco...     144    8e-34
ref|XP_849487.1|  PREDICTED: similar to creatine kinase, brain...     140    1e-32
ref|XP_544323.2|  PREDICTED: similar to creatine kinase, brain...     103    9e-22
ref|XP_851013.1|  PREDICTED: similar to Creatine kinase B-type...     98.6   4e-20
ref|XP_849758.1|  PREDICTED: similar to otoferlin isoform a is...     97.8   6e-20
ref|XP_849487.1|  PREDICTED: similar to creatine kinase, brain...     95.1   4e-19
ref|XP_544323.2|  PREDICTED: similar to creatine kinase, brain...     74.7   6e-13
ref|XP_851013.1|  PREDICTED: similar to Creatine kinase B-type...     72.8   2e-12
ref|XP_849758.1|  PREDICTED: similar to otoferlin isoform a is...     31.6   6.1
```

Results filtered for domestic dog proteins.
26 additional gene predictions from Dog alone.
Many are extra splice variants predicted by Gnomon.

Other Reports

NCBI Public Services



TaxBLAST: Taxonomy Reports

NCBI Public Services

Lineage Report

```
Eutheria          [placentals]
. Laurasiatheria  [placentals]
. . Cetartiodactyla [even-toed ungulates]
. . . Sus scrofa (wild boar) ---------------  773 2 hits [even-toed ungulates]
. . . Bos taurus (COW) ....................  768 4 hits [even-toed ungulates]
. . Canis lupus familiaris (dogs) ---------  509 1 hit  [carnivores]
. Mus musculus (mouse) ------------------  770 4 hits [rodents]
. Oryctolagus cuniculus (domestic rabbit) .  768 3 hits [rabbits & hares]
. Rattus norvegicus (brown rat) ...........  766 4 hits [rodents]
```

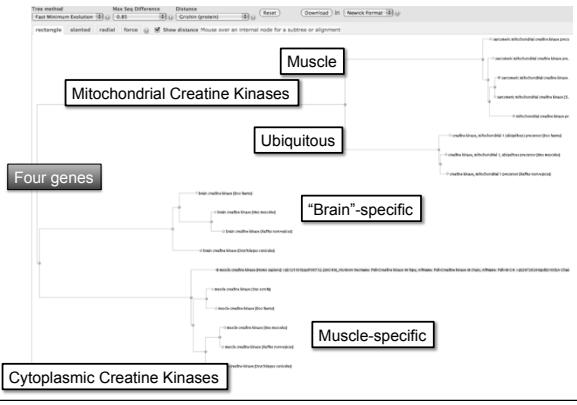
Organism Report

```
 Sus scrofa (wild boar, ...) [even-toed ungulates] taxid 9823
ref|NP_001123421.1| muscle creatine kinase [Sus scrofa]          773  0.0
ref|NP_001038016.1| sarcomeric mitochondrial creatine kina...    509  2e-144

 Mus musculus (mouse) [rodents] taxid 10090
ref|NP_031736.1| muscle creatine kinase [Mus musculus]           770  0.0
ref|NP_067248.1| brain creatine kinase [Mus musculus]            652  0.0
ref|NP_940807.1| sarcomeric mitochondrial creatine kinase ...    511  8e-145
ref|NP_034027.1| creatine kinase, mitochondrial 1, ubiquit...    504  6e-143

 Bos taurus (cow, ...) [even-toed ungulates] taxid 9913
ref|NP_777198.2| muscle creatine kinase [Bos taurus]             768  0.0
ref|NP_001015613.1| brain creatine kinase [Bos taurus]           664  0.0
ref|NP_001029828.1| sarcomeric mitochondrial creatine kina...    508  5e-144
ref|NP_776700.1| creatine kinase, mitochondrial 1 (ubiquit...    504  7e-143

 Oryctolagus cuniculus (domestic rabbit, ...) [rabbits & hares] taxid 9986
ref|NP_001075708.1| muscle creatine kinase [Oryctolagus cu...    768  0.0
ref|NP_001075730.1| brain creatine kinase [Oryctolagus cun...    660  0.0
ref|NP_001156542.1| mitochondrial creatine kinase precurso...    500  1e-141

 Rattus norvegicus (brown rat, ...) [rodents] taxid 10116
ref|NP_036662.1| muscle creatine kinase [Rattus norvegicus]      766  0.0
ref|NP_036661.2| brain creatine kinase [Rattus norvegicus]       650  0.0
ref|NP_001121124.1| sarcomeric mitochondrial creatine kina...    511  8e-145
ref|NP_001012756.1| creatine kinase, mitochondrial 1 precu...    505  4e-143

 Canis lupus familiaris (dogs) [carnivores] taxid 9615
ref|NP_001138686.1| sarcomeric mitochondrial creatine kina...    509  2e-144
```

Four genes in each mammal.

TreeView: Distance Tree

NCBI Public Services



Position Specific Iterative BLAST

NCBI Public Services

Program Selection

Algorithm

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

Choose a BLAST algorithm

MLH1 and ETR1

NCBI Public Services

```
>gi|4557757|ref|NP_000240.1| MutL protein homolog 1 [Homo sapiens]
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIVKEGGLKLIQIQDNGTGIRK
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPK
PCAGNQGTQITVEDLFYNIATRRKALKNPSEEYGKILE
STVDNIRSIFGNAVSRELIEIGCEDKTLAFKMNGYISN
AAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLLGSNSSRMYFTQTLLP
GLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDAFLQFLSKPLSSQPQAIVTEDKTDIS
SGRARQQDEEMLELPAPAEVAAKNQSLEGDTTKGTSEMSEKRGPTSSNPRKRHREDSDVEMVEDDSRKEM
TAACTPRRRIINLTSVLSLQEEINEQGHEVLREMLHNHSFVGCVNPQWALAQHQTKLYLLNTTKLSEELF
YQILIYDFANFGVLRLSEPAPLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEMLADYFSLEI
DEEGNLIGLPLLIDNYVPPLEGLPIFILRLATEVNWDEEKECFESLSKECAMFYSIRKQYISEESTLSGQ
QSEVPGSIPNSWKWTVEHIVYKALRSHILPPKHFTEDGNILQLANLPDLYKVFERC
```

Human Mismatch Repair Protein

```
>gi|22095656|sp|O81122.1|ETR1_MALDO Ethylene receptor
MLACNCIEPQWPADELLMKYQYISDFFIALAYFSIPLELIYF
INLWTFSIHSRTVAMVMTTAKVLTAVVSCATALMLVHIIPDL
ETGRHVRMLTHEIRSTLDRHTILKTTLVELGRTLALEECALWMPTFTGLELQLSYTLRQQNFVGYTVFIH
LPVINQVFSSNRAVKISANSPVAKLRQLAGRHIPGEVVAVRVPLLHLSNFQINDWPELSTKRYALMVLML
PSDSARQWHVHELELVEVVADQVAVALSHAAILEESMRARDLLMEQNIALDLARREAETAIRARNDFLAV
MNHEMRTPMHAIIALSSLLQETELTAEQRLMVETILRSSNLLATLINDVLDLSRLEDGSLQLEIATFNLH
SVFREVHNMIKPVASIKRLSVTLNIAADLPMYAIGDEKRLMQTILNVVGNAVKFSKEGSISITAFVAKSE
SLRDFRAPDFFPVQSDNHFYLRVQVKDSGSGINPQDIPKLFTKFAQTQALATRNSGGSGLGLAICKRFVN
LMEGHIWIESEGLGKGCTATFIVKLGFPERSNESKLPFAPKLQANHVQTNFPGLKVLVMDDNGVSRSVTK
GLLAHLGCDVTAVSLIDELLHVISQEHKVVFMDVSMPGIDGYELAVRIHEKFTKRHERPVLVALTGSIDK
ITKENCMRVGVDGVILKPVSVDKMRSVLSELLEHRVLFEAM
```

Apple ethylene receptor

PSI-BLAST: Iteration 1

NCBI Public Services

Sequences producing significant alignments: Scores E

Run PSI-BLAST Iteration 2

Sequences with E-value worse than threshold

PSI-BLAST:Iteration 4

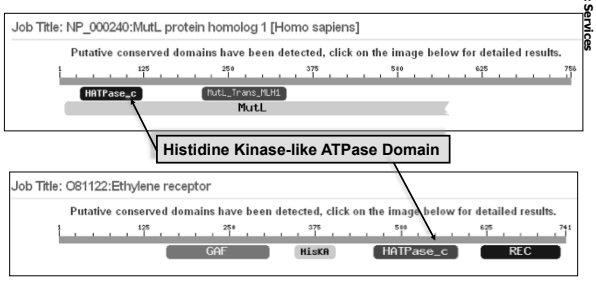
NCBI Public Services

	sp|Q827H3|PHOQ_SALTI	Virulence sensor histidine kinase phoQ	48.0	9e-06	
	sp|P14147|PHOQ_SALTY	Virulence sensor histidine kinase phoQ >...	48.0	9e-06	
NEW	sp|Q9M7M1|ETR1_PRUPE	Ethylene receptor	47.6	1e-05	
	sp|Q55168|PHY1_SYNY3	Phytochrome-like protein cphl (Light-reg...	47.6	1e-05	G
NEW	sp|P40330|BVGS_BORPA	Virulence sensor protein bvgS precursor	47.2	1e-05	
	sp|Q03069|DEGM_BACB2	Sensor protein degM	47.2	1e-05	
	sp|P09431|NTRB_RHOCA	Nitrogen regulation protein ntrB	47.2	1e-05	
NEW	sp|P26762|BVGS_BORBR	Virulence sensor protein bvgS precursor	47.2	1e-05	
	sp|P10955|FIXL_RHIME	Sensor protein fixL	47.2	1e-05	
	sp|P21865.2|KDPD_ECOLI	Sensor protein kdpD	47.2	2e-05	G
NEW	sp|O48929|ETR1_TOBAC	Ethylene receptor (NT-ETR1)	47.2	2e-05	
NEW	sp|Q8FK37|CUSS_ECOL6	Sensor kinase cusS	46.9	2e-05	
	sp|P23222|FIXL_BRAJA	Sensor protein fixL	46.9	2e-05	
NEW	sp|Q4A009|LYTS_STAS1	Sensor protein lytS (Autolysin sensor kinas	46.5	2e-05	G
NEW	sp|O49230|ETR1_BRAOL	Ethylene receptor	46.5	3e-05	
NEW	sp|P08982|ENVZ_SALTY	Osmolarity sensor protein envZ	46.1	3e-05	
NEW	sp|P16575|BVGS_BORPE	Virulence sensor protein bvgS precursor	46.1	4e-05	
NEW	sp|O32193|CSSS_BACSU	Sensor protein cssS	45.7	5e-05	
NEW	sp|P48027|GACS_PSESY	Sensor protein gacS	45.3	5e-05	
NEW	sp|O81122|ETR1_MALDO	Ethylene receptor	45.3	5e-05	
NEW	sp|O82436|ETR1_CUCMX	Ethylene receptor (MELETR1) (CM-ETR1)	45.3	5e-05	

Plant ethylene receptors, bacterial two-component regulatory system kinases

RPS-BLAST: Conserved Domains

NCBI Public Services

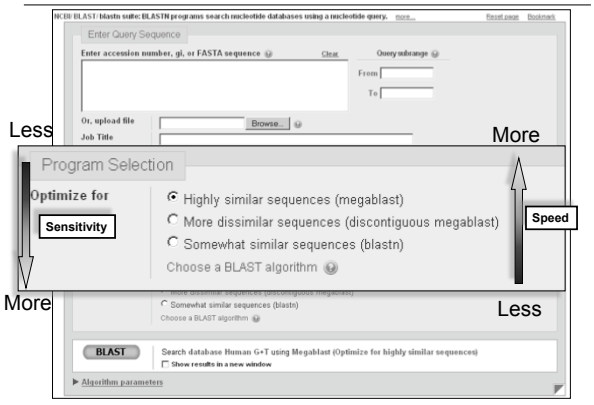


NCBI Public Services

Basic BLAST: Nucleotide

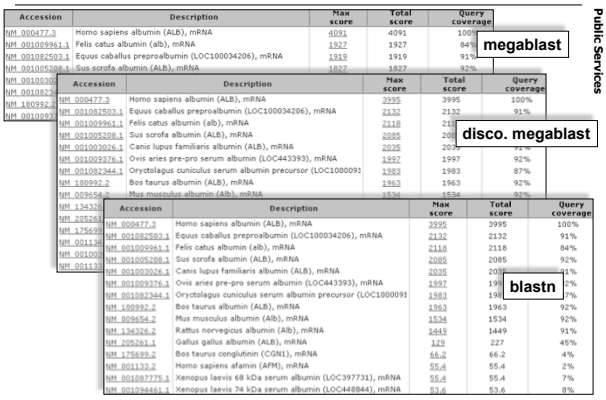
Universal Form: Nucleotide

NCBI Public Services



Nucleotide Results: ALB mRNA

NCBI Public Services



Nucleotide BLAST: Human Genome

NCBI Public Services

NCBI/BLAST/blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence | Clear

Query subrange: From / To

Or, upload file: Browse...

Job Title: Crab eating macaque CDC20 mRNA

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.): Human genomic plus transcript

Entrez Query (Optional): Enter an Entrez query to limit search

Sortable Results

NCBI Public Services

Separate Sections for Transcript and Genome

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

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

Accession	Description	Max score	Tot score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell division cycle 20 ho	2876	2876	95%	0.0	97%	U E G M
Genomic sequences [show first]							
NT_023935.17	Homo sapiens chromosome 9 genomic contig	2629	2629	94%	0.0	95%	
NW_924484.1	Homo sapiens chromosome 9 genomic contig	2601	2601	94%	0.0	95%	
NT_032977.8	Homo sapiens chromosome 1 genomic contig	428	3002	95%	9e-117	100%	
NW_921351.1	Homo sapiens chromosome 1 genomic contig	428	3010	95%	9e-117	100	

Pseudogene on Chromosome 9

Direct links to Entrez Databases

Functional Gene on Chromosome 1

Total Score: All Segments

NCBI Public Services

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

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

Accession	Description	Max score	Tot score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell division cycle 20 hc	2876	2876	95%	0.0	97%	U E G M
Genomic sequences [show first]							
NW_921351.1	Homo sapiens chromosome 1 genomic contig	428	3010	95%	9e-117	100%	
NT_032977.8	Homo sapiens chromosome 1 genomic contig	428	3002	95%	9e-117	100%	
NT_023935.17	Homo sapiens chromosome 9 genomic contig	2629	2629	94%	0.0	95%	
NW_924484.1	Homo sapiens chromosome 9 genomic contig	2601	2601	94%	0.0	95%	

Functional Gene Now First

Alignments: Sorting in Exon Order

NCBI Public Services

>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

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

Query start position
Exon order

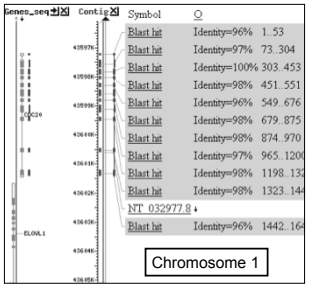
Default
Longest

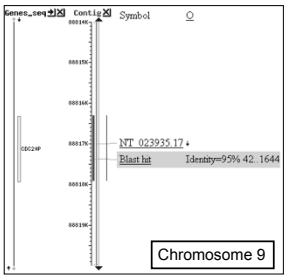
Score = 412 bits (208), Expect = 5e-112
Identities = 226/232 (97%), Gaps = 0/232 (0%)
Strand=Plus/Plus

Query 73 GGGCTCCGCAGGCACCAACTGCAAGGACCCCTCCCGCTGCGGGCGTTCCCATGGCACAAT 132
Sbjct 13796755 GGGCTCCGTAGGCACCAACTGCAAGGACCCCTCCCCCTGCGGGCGCTCCCATGGCACAGT 13796814

Query 133 TCGCGTTCGAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCCAATGCAC 192
Sbjct 13796815 TCGCGTTCGAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCCAATGCAC 13796874

Links to Map Viewer



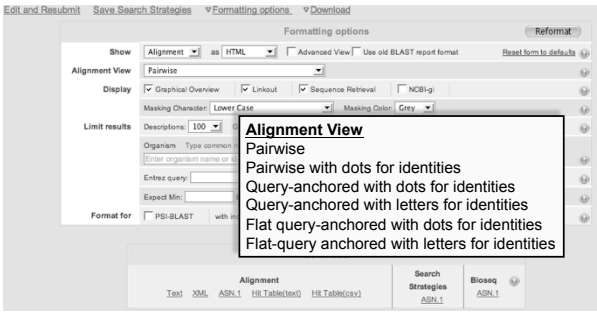


NCBI Public Services

BLAST Formatting Options

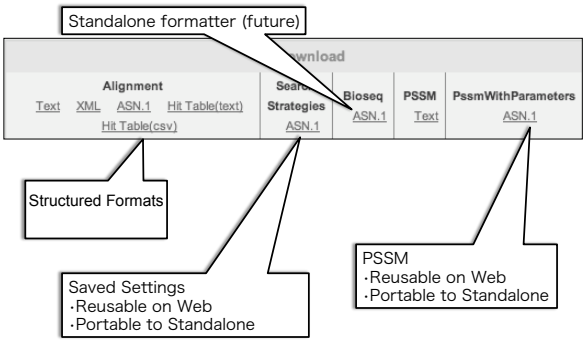
NCBI Public Services

Formatting Page (Now on Results)



NCBI Public Services

Download Options (Now on Results)



NCBI Public Services

The Hit Table

NCBI Public Services

```
# BLASTP 2.2.17 (Aug-26-2007)
# Query: gi|4557757|ref|NP_000240.1| MutL protein homolog 1 [Homo sapiens]
# Database: swissprot
# Fields: query id, subject ids, % identity, % positives, alignment length, mismatches,
gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 80 hits found
ref|NP_000240.1||gi|4557757 gi|1709056|sp|P38920|MLH1_YEAST  36.68 56.91 796 426 18 8 756 5 769 7e-138 481
ref|NP_000240.1||gi|4557757 gi|4847886|sp|Q9P7N6|MUH1_SCHPO 37.24 54.04 768 371 16 8 756 9 684 8e-122 437
ref|NP_000240.1||gi|4557757 gi|2509
ref|NP_000240.1||gi|4557757 gi|2509
ref|NP_000240.1||gi|4557757 gi|127552|sp|P23367.2|MUTL_ECOLI 35.99 58.11 339 202 7 8 334 3 338 8e-55  214
ref|NP_000240.1||gi|4557757 gi|29427778|sp|Q8FAK9|MUTL_ECOL6 35.99 58.11 339 202 7 8 334 3 338 1e-54  214
ref|NP_000240.1||gi|4557757 gi|20455084|sp|Q8XDN4|MUTL_ECO57 35.99 58.11 339 202 7 8 334 3 338 1e-54  214
ref|NP_000240.1||gi|4557757 gi|59798328|sp|Q72PF7|MUTL_LEPIC 36.27 55.20 375 221 8 6 375 2 363 3e-54  213
ref|NP_000240.1||gi|4557757 gi|13431695|sp|P57886|MUTL_PASMU 35.48 58.94 341 213 6 8 345 3 339 4e-54  212
ref|NP_000240.1||gi|4557757 gi|1171080|sp|P44494|MUTL_HAEIN  35.74 59.87 319 198 6 8 323 3 317 5e-54  212
ref|NP_000240.1||gi|4557757 gi|20455102|sp|Q8ZIW4|MUTL_YERPE 36.01 58.63 336 207 6 8 339 3 334 6e-54  212
ref|NP_000240.1||gi|4557757 gi|20455152|sp|Q9JYT2|MUTL_NEIMB 33.96 55.35 374 224 8 8 376 4 359 2e-53  210
ref|NP_000240.1||gi|4557757 gi|20139217|sp|Q9KAC1|MUTL_BACHD 35.39 55.90 356 214 6 8 362 4 344 2e-53  209
ref|NP_000240.1||gi|4557757 gi|31076794|sp|Q87LG5|MUTL_VIBPA 35.33 58.38 334 210 5 8 338 3 333 3e-53  209
ref|NP_000240.1||gi|4557757 gi|20455150|sp|Q9JTS2|MUTL_NEIMA 36.94 58.28 314 183 5 8 316 4 307 5e-53  209
ref|NP_000240.1||gi|4557757 gi|56749233|sp|Q6GHD9|MUTL_STAAR 38.28 58.46 337 193 7 6 335 2 330 1e-52  207
ref|NP_000240.1||gi|4557757 gi|25090739|sp|Q8NWX9|MUTL_STAAW 38.28 58.46 337 193 7 6 335 2 330 1e-52  207
ref|NP_000240.1||gi|4557757 gi|71151979|sp|Q5HGD5|MUTL_STAAC 38.28 58.46 337 193 7 6 335 2 330 1e-52  207
ref|NP_000240.1||gi|4557757 gi|54037875|sp|P65492|MUTL_STAAN 38.28 58.46 337 193 7 6 335 2 330 2e-52  207
ref|NP_000240.1||gi|4557757 gi|20043258|sp|Q9KV13|MUTL_VIBCH 35.74 58.56 333 204 6 8 335 3 330 2e-52  207
ref|NP_000240.1||gi|4557757 gi|127553|sp|P14161|MUTL_SALTY   35.10 56.93 339 205 7 8 334 3 338 3e-52  206
ref|NP_000240.1||gi|4557757 gi|20455140|sp|Q9CDL1|MUTL_LACLA 36.31 56.55 336 196 5 6 334 2 326 4e-52  206
ref|NP_000240.1||gi|4557757 gi|61214242|sp|Q7MH01|MUTL_VIBVY 34.63 58.51 335 213 5 8 339 3 334 4e-52  206
ref|NP_000240.1||gi|4557757 gi|20455099|sp|Q8Z187|MUTL_SALTI 35.10 56.93 339 205 7 8 334 3 338 4e-52  206
ref|NP_000240.1||gi|4557757 gi|31076809|sp|Q8DCV0|MUTL_VIBVU 34.63 58.51 335 213 5 8 339 3 334 6e-52  205
ref|NP_000240.1||gi|4557757 gi|71648717|sp|Q5E2C6|MUTL_VIBF1 36.71 59.81 316 186 6 8 316 3 311 1e-51  204
ref|NP_000240.1||gi|4557757 gi|37999611|sp|Q88DD1|MUTL_PSEPK 30.34 48.97 435 278 7 8 419 7 439 2e-51  203
```

Also available in comma separated format for Excel

Structured formats: XML and ASN.1

NCBI Public Services

XML

```
<Iteration_hits>
—<Hit>
<Hit_num>1</Hit_num>
<Hit_id>gi|730028|sp|P40692|MLH1_HUMAN<
—<Hit_def>
DNA mismatch repair protein Mlh1 (MutL
homolog 1)
</Hit_def>
<Hit_accession>P40692</Hit_accession>
<Hit_len>756</Hit_len>
—<Hit_hsps>
—<Hsp>
<Hsp_num>1</Hsp_num>
<Hsp_bit-score>1568.9</Hsp_bit-score>
<Hsp_score>4061</Hsp_score>
<Hsp_evalue>0</Hsp_evalue>
<Hsp_query-from>1</Hsp_query-from>
<Hsp_query-to>756</Hsp_query-to>
<Hsp_hit-from>1</Hsp_hit-from>
<Hsp_hit-to>756</Hsp_hit-to>
<Hsp_query-frame>0</Hsp_query-frame>
<Hsp_hit-frame>0</Hsp_hit-frame>
<Hsp_identity>0</Hsp_identity>
<Hsp_positive>0</Hsp_positive>
<Hsp_gaps>0</Hsp_gaps>
<Hsp_align-len>756</Hsp_align-len>
```

ASN.1

```
Seq-annot ::= {
  desc {
    user {
      type
        str "Hist Seqalign" ,
      data {
        {
          label
            str "Hist Seqalign" ,
          data
            bool TRUE } } } ,
    user {
      type
        str "Blast Type" ,
      data {
        {
          label
            id 0 ,
          data
            int 0 } } } ,
    user {
      type
        str "BLAST database title" ,
      data {
        {
          label
            str "Non-redundant SwissProt
```

PSSMs: Restart PSI-BLAST

NCBI P

```
PSSM:2
425A68313141592653591152764D0031A0
85AC6ADB26353660A64A5ACD6B550F9E84
EB5935D6DBBB42AD12495AE8AEB97674E3
AC9B65354C55B6AA954D130C6A32D6B59E
36A191B507A9A0353C2052AA993000098O
E00E79BAFAFEFF5CFEC9E404543E947EEE
A1294A1680A4A028428880 29A5A0429481
28AA40A0A44A029588295625A0A281290A
643452D0940342942148555214 0D09434D
554A990192449 4A6406488508D214ACC94
1840D0C43300528D0525225252506439 2D
34214252AD2508E40A6425356485014D28
DFAEF3BF10F93EBDE2B476022A19C046B6
BCA4B317968BDADB75A23926609BB8CA02
7BDE03081372D56D17565D3F000903D86C
DB1777325999496E3BA4E5675F26BA60D4E
A24CB6E3DAE1738F3D93D1575AA1254445
33C56551601D3A5B2DDE5594F16006B095
C706E734F83A2B730AAEA309BE828DACCA
35B9435BE6DE4C8385BEAAAB9AAF1FDC73
7C9C8C0A42E94743755E385 9DD8EDE6E2
1 26E834D4EB
C F605DDDB2D
```

ASCII encoded, Web only

```
PssmWithParameters ::= {
  pssm {
    numRows 28,
    numColumns 157,
    query seq {
      id {
        other {
          accession "YP_385691"
        }
      },
      inst {
        repr raw,
        mol aa,
        length 616
      }
    },
    intermediateData {
      freqRatios {
        { 0, 10, 0 },
        { 427396, 10, -7 },
        { 0, 10, 0 },
        { 593523, 10, -8 },
        { 18841, 10, -6 },
        { 231628, 10, -7 },
        { 729652, 10, -8 },
        { 245444, 10, -7 },
        { 622078, 10, -8 },
        { 118663, 10, -7 },
```

ASN.1 ScoreMat, Portable

NCBI Public Services

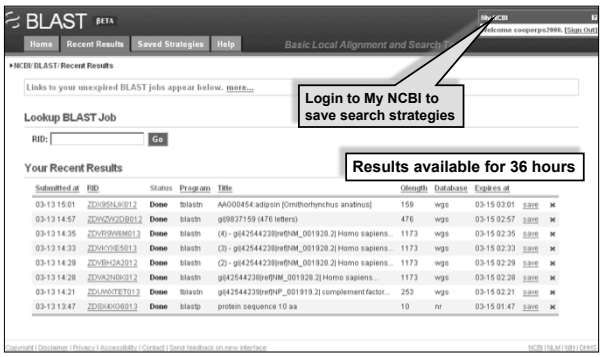
Managing Searches

Recent Results

Saved Strategies

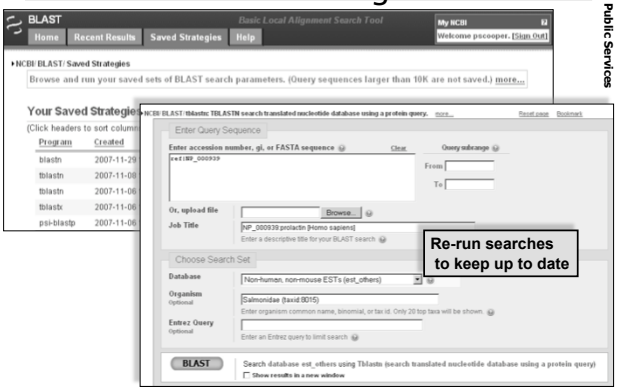
Recent Results

NCBI Public Services



Saved Strategies

NCBI Public Services



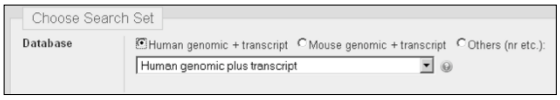
NCBI Public Services

Genome and Specialized BLAST

Nucleotide Databases: Human and Mouse

NCBI Public Services

Megablast, blastn service



- Human and mouse genomic and transcript now default
- Separate sections in output for mRNA and genomic
- Direct links to Map Viewer for genomic sequences

Genome BLAST pages

NCBI Public Services

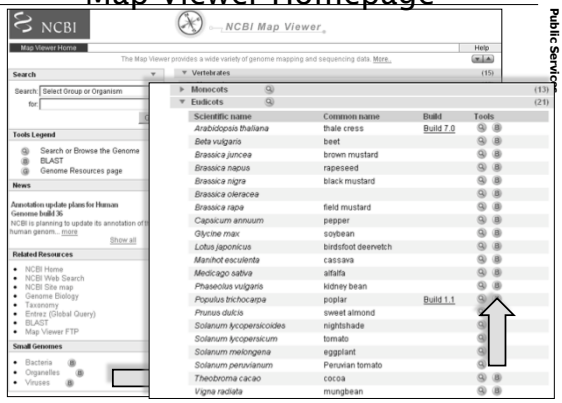
BLAST Assembled Genomes

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

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

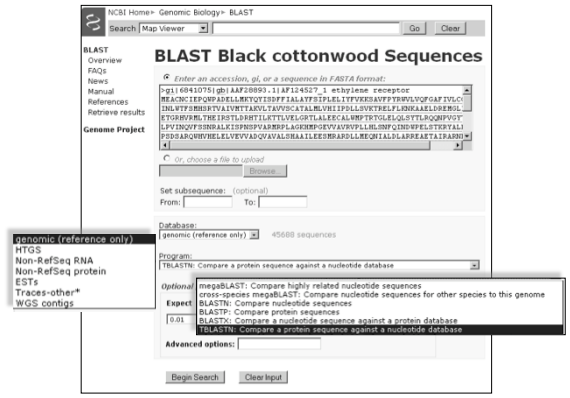
Map Viewer Homepage

NCBI Public Services



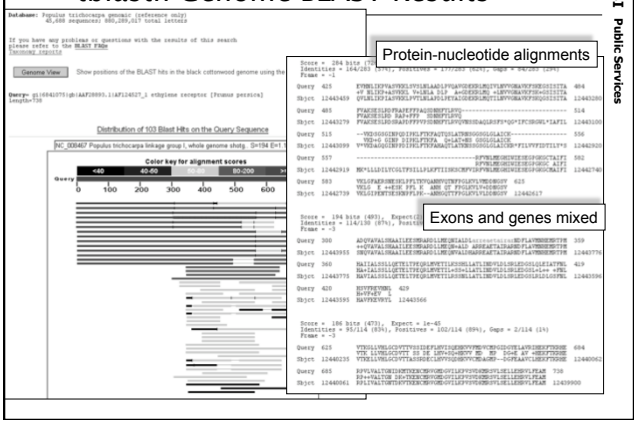
Poplar Genome BLAST

NCBI Public Services



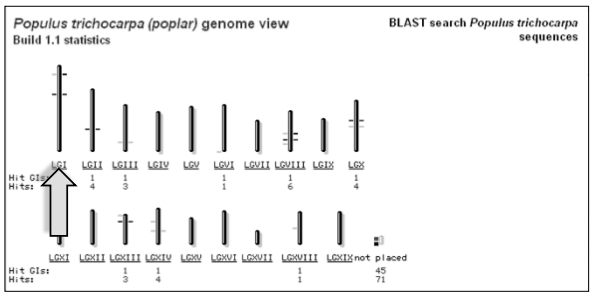
tblastn Genome BLAST Results

NCBI Public Services



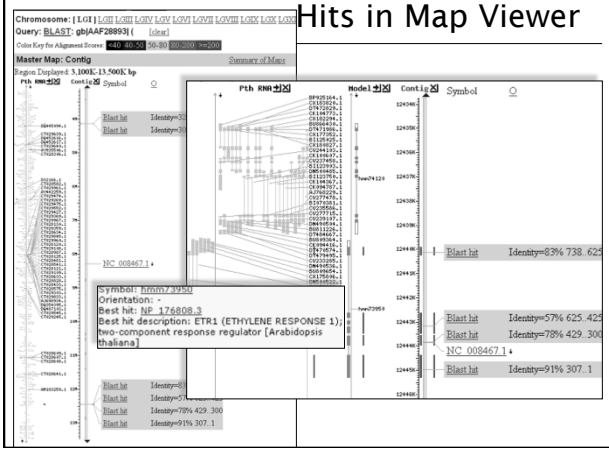
Genomic Context of BLAST Hits

NCBI Public Services



Hits in Map Viewer

NCBI Public Services



Specialized BLAST Pages

NCBI Public Services

Specialized BLAST

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

- Make specific primers with **Primer-BLAST**
- Search **trace archives**
- Find **conserved domains** in your sequence (cds)
- Find sequences with similar **conserved domain architecture** (cdart)
- Search sequences that have **gene expression profiles** (GEO)
- Search **immunoglobulins** (IgBLAST)
- Search for **SNPs** (snp)
- Screen sequence for **vector contamination** (vecscreen)
- **Align** two (or more) sequences using BLAST (bl2seq)
- Search **protein** or **nucleotide** targets in PubChem BioAssay
- Search SRA **transcript libraries**
- Constraint Based Protein **Multiple Alignment Tool**

BLAST extensions and improvements

NCBI Public Services

- PrimerBlast – primer designer / specificity checker
- COBALT – Protein Multiple Alignment tool
- Integration / expansion of BLAST 2 Sequences

Primer BLAST: Template and Primers

NCBI Public Services

PCR Template

Reset page | Save search parameters

Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

NM_001077.2

Range

From | To

Forward primer

Reverse primer

Or, upload FASTA file Browse...

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Clear

Use my own reverse primer (5'->3' on minus strand) Clear

PCR product size Min 100 Max 1000

# of primers to return 10

Primer melting temperatures (T_m) Min 57.0 Opt 60.0 Max 63.0 Max T_m difference 3

Please note the recent change in default Tm calculation

Primer BLAST: specificity params

NCBI Public Services

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span: No preference

Exon junction match: Exon at 5' side 7 4

Minimal number of

No preference

Primer must span an exon-exon junction

Primer may not span an exon-exon junction

Intron inclusion: Primer must be separated by at least one intron on the corresponding genomic DNA

Intron length range: Min 1000 Max 1000000

Get Primers

Primer Pair Specificity Checking Parameters

Specificity check: Enable search for primer pairs specific to the intended PCR template

Organism: 9606

Organism-specific search

Enter an organism name, taxonomy id or select from the suggestion list as you type.

Add more organisms

Database: Refseq RNA (refseq_rna)

Primer specificity stringency: At least 2 total

at least 2 misma

Refseq RNA (refseq_rna)

Genome (reference assembly from selected organisms)

Genome (chromosomes from all organisms)

nr

Misprimed product size deviation: 1000

Splice variant handling: Allow primer to ampli

Primer Results

NCBI Public Services

0 400 800 1200 1600 2000

Conserved region

Exon boundary

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	AAGCACCTTCGGGTCGCAGC	Plus	20	1466	1485	59.97	65.00%
Reverse primer	GGAAGCTCAGTAACTTTTGTGTGGGGT	Minus	27	1928	1902	59.17	48.15%
Product length	463						

Specific for this family member

Products on intended target

>NM_001077.2 Homo sapiens UDP glucuronosyltransferase 2 family, polypeptide B17 (UGT2B17), mRNA

```
product length = 463
Forward primer  1     AAGCACCTTCGGGTCGCAGC  20
Template        1466  ....................  1485

Reverse primer  1     GGAAGCTCAGTAACTTTTGTGTGGGGT  27
Template        1928  ............................  1902
```

BLAST 2 Sequences

NCBI Public Services

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide subjects using a nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

Query subrange From To

>ref|NC_000004.11|:c69538106-69310907 Homo sapiens chromosome 4,GRCh37 primary reference assembly

TTGCATCACTGTTGCATTTTTGCACTGCAGAAACCAAGAGAGGCCATGTAGATTTTCCCCTGTTAACCGG

GCTCCAAGGCAAACTTACCAGTAGGCGAGATCAGTGGCTGATGCGCATGCACAGAGAGGCGACTGCAGGC

TCAGCAGCTTCCTCTCTCTAATTGCTCTGGCCTGCTCTCTGGCGTGAGGGATAGGTCCACAAGCGACAC

AGACCTGAGCCCCTCCTAGGATTTCGCACCAGATGTAAGGTTCTTGAATCAGTTTGAACCCCAACAGCGC

Or, upload file Choose File no file selected

Job Title ref|NC_000004.11|c69538106-69310907 Homo...

Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Clear

>ref|NT_167250.1|:c249882-139868 Homo sapiens chromosome 4 genomic c

TCAGGAACTTCTCCTGCCAACACAACTTCAGCTTCTCCAGTGGGCATCTGTATTCTC

CCCTTTCCCACTTCCATTGTTCGGGCACTCACAGTGTTTGGGATGTCTCCCAGCTCC

CGCTTCCTGCACAGCGTGTGTGCATCCTCTCAGAATTGCTGCTCTGTTCTTGCAGTTGATCTGCAGCTAA

AATTCACAATGCAAGCCTCTGCATGCTGCTCTGTCTGGAGCTGCAATCTAGTCCTGCATCCCATCTGCCA

Or, upload file Choose File no file selected

Filters and Masking

Filter: Low complexity regions; Species-specific repeats for: Human

Mask: Mask for lookup table only; Mask lower case letters

Region between UGT2B15 and TMRSS11E in primary reference and alternate locus

Program Selection

Optimize for: Highly similar sequences (megablast); More dissimilar sequences (discontiguous megab; Somewhat similar sequences (blastn)

Choose a BLAST algorithm

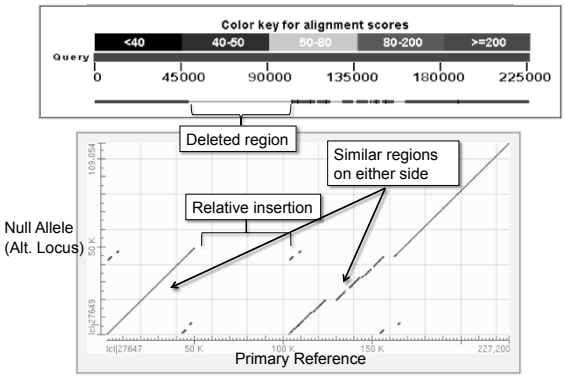
BLAST — Search nucleotide sequence using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

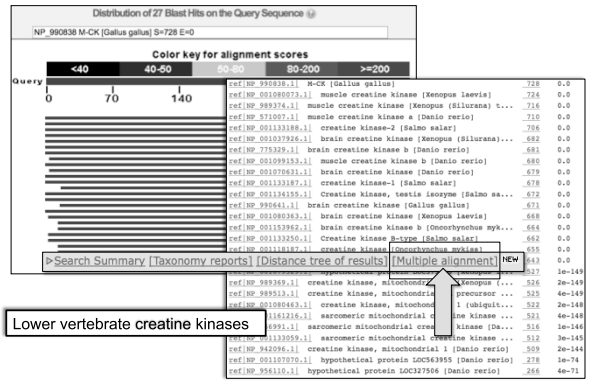
Alignment of reference and null-allele

NCBI Public Services



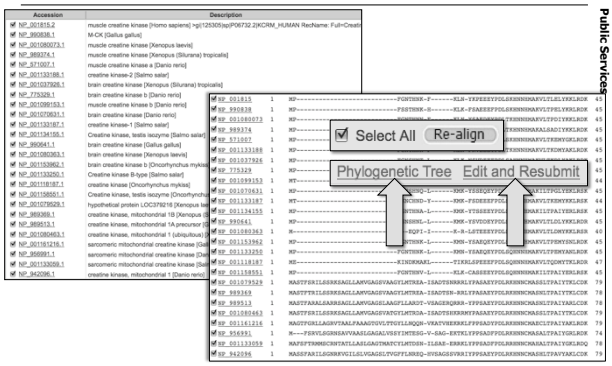
COBALT Extension of BLAST

NCBI Public Services



COBALT (Constraint Based Alignment Tool)

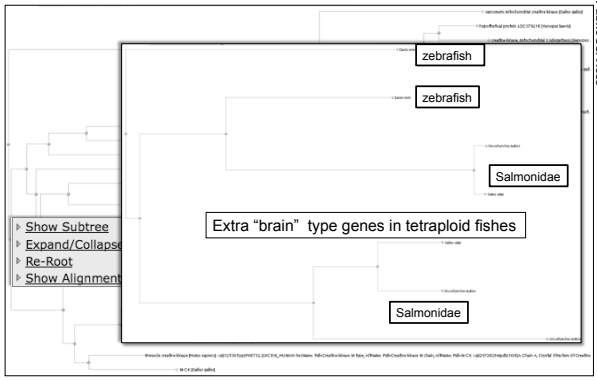
NCBI Public Services



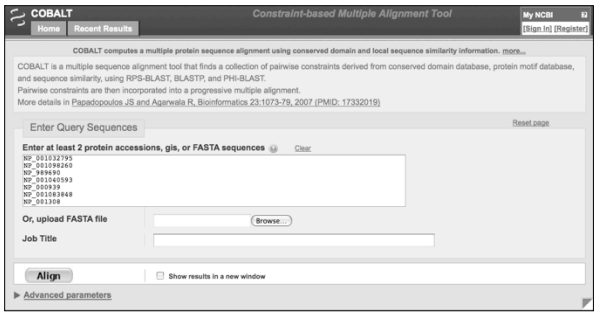
True multiple sequence alignment that uses conserved domain information

COBALT Tree

NCBI Public Services



COBALT Interface



NCBI Public Services

Short Read Archive BLAST

NCBI Public Services

NCBI's Sequence Read Archive

- Repository for next generation (massively parallel) sequencing data
- Platforms: 454, Illumina, SOLid, Heliscope, Complete Genomics
- Genome (re)-sequencing, transcriptome sequencing, Environmental sequencing
- Thousand times larger than GenBank

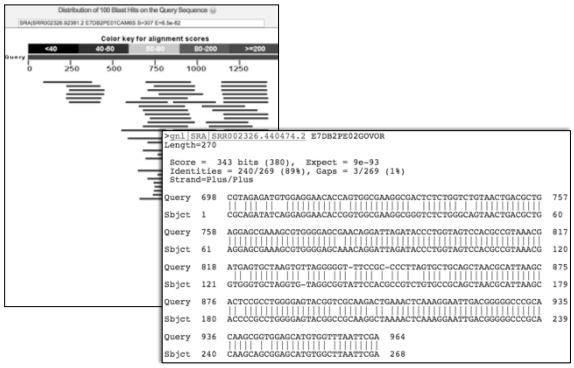
NCBI Public Services

SRA BLAST against Ice metagenome

NCBI Public Services

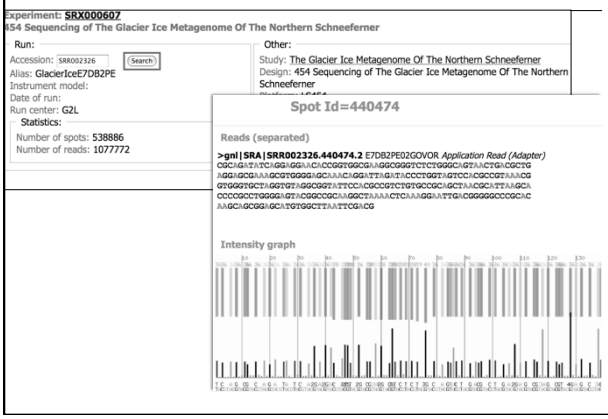
SRA Results

NCBI Public Services



SRA Run

NCBI Public Services

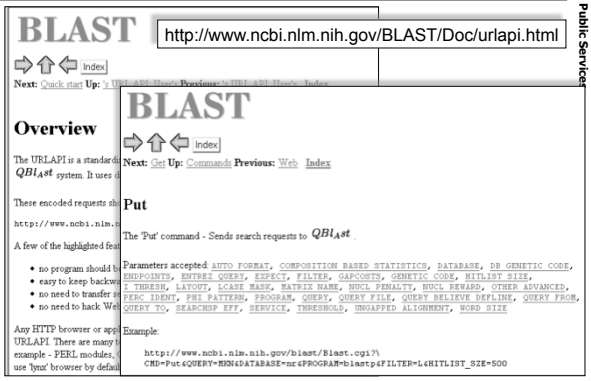


NCBI Public Services

Running BLAST without the Web

BLAST URL API

NCBI Public Services



BLAST: standalone, clients, databases

NCBI Public Services

```
ftp> open ftp.ncbi.nih.gov
.
.
ftp> cd blast
```

ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

C++ Toolkit BLAST (blast+) the focus of future development.

Name↓	
ChangeLog	
ncbi-blast-2.2.22+-1.i686.r	
ncbi-blast-2.2.22+-1.src.rp	
ncbi-blast-2.2.22+-1.x86_6	
ncbi-blast-2.2.22+-ia32-lin	
ncbi-blast-2.2.22+-ia32-win32.tar.gz	30095 KB
ncbi-blast-2.2.22+-spa	
ncbi-blast-2.2.22+-src.tar.gz	9770 KB
ncbi-blast-2.2.22+-src.zip	
ncbi-blast-2.2.22+-universal-maco	
ncbi-blast-2.2.22+-win32.exe	
ncbi-blast-2.2.22+-win64.exe	47649 KB
ncbi-blast-2.2.22+-x64-linux.tar.gz	221565 KB
ncbi-blast-2.2.22+-x64-solaris.tar.gz	115616 KB
ncbi-blast-2.2.22+-x64-win64.tar.gz	47517 KB
ncbi-blast-2.2.22+.dmg	168258 KB
user_manual.pdf	374 KB

Standalone BLAST

NCBI Public Services

C toolkit BLAST

C++ toolkit BLAST

```
BLAST
E:\Blast\bin>blastall -i purf.fsa -d swissprot -p blastp
BLAST+
E:\blast+\bin>blastp -query purf.fsa -db swissprot
```

Keeping up with what's new

NCBI Public Services

NCBI News

National Center for Biotechnology Information | National Library of Medicine | National Institutes of Health
Department of Health and Human Services

Search for Within This book All books PubMed Go

NCBI » Bookshelf » NCBI News »

NCBI News on Bookshelf

www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=newsncbi

- NCBI News, August 2009
- NCBI News, July 2009
  - Featured Resource: The BioSystems Database of Biological Pathways
  - New Databases and Tools
  - GenBank News
  - Updates and Enhancements
  - Announce Lists and RSS Feeds
- NCBI News, June 2009
- NCBI News, May 2009
- NCBI News, April 2009
- NCBI News, March 2009
- NCBI News, February 2009
- NCBI News, January 2009
- NCBI News, December 2008
- NCBI News, November 2008
- NCBI News, August 2008

NCBI News

NCBI News available online
August issue

- Featured Resource: The NCBI Short Read Archive (SRA) of Next-Generation Sequencing Data
- New Databases and Tools
- GenBank News
- Updates and Enhancements
- Announce Lists and RSS Feeds

Service Addresses

NCBI Public Services

- ***General Help*** `info@ncbi.nlm.nih.gov`
- ***BLAST*** `blast-help@ncbi.nlm.nih.gov`

Telephone support: 301- 496- 2475