

A Practical Guide to NCBI BLAST

06/06/2016



1

NCBI Search Services and Tools

- **Entrez** [integrated literature and molecular databases](#)
 - Viewers
 - BLink [protein similarities](#)
 - Graphical Sequence Viewer [annotation viewer and analysis tool](#)
- **BLAST** [sequence similarity search service](#)
- **VAST** [structure similarity searches](#)
- **Tools, special services, standalone software**
 - Entrez Utilities [Entrez API](#) [<ncbi>/books/NBK25501/](#)
 - Standalone BLAST [BLAST programs + databases](#) [<ncbi>/books/NBK1762/](#)
 - Cn3D [3D structure viewer](#) [<ncbi>/Structure/CN3D/cn3d.shtml](#)
 - Genome Workbench [sequence analysis / annotation platform](#) [<ncbi>/tools/gbench/](#)
 - SRA Utilities [<ncbi>/Traces/sra/](#)
 - SRA Run Browser [web access](#)
 - SRA toolkit [standalone SRA manipulator and client](#)

06/06/2016

2

Topics

- Basics of using NCBI BLAST
 - Motivation, Statistics, Scoring, Search Programs
- Using the Web Interface
- Other Web services
 - COBALT – protein multiple alignment
 - Primer BLAST
 - MOLE-BLAST
- Live Searches

What is BLAST?

- Widely used sequence similarity search tool
- Finds high scoring local alignments between two sequences (protein or DNA)
- Includes a model of score distributions for random local alignments
- Provides statistical significance for alignments

BLAST Fundamentals

NCBI Public Services

- BLAST tells you about non-chance similarities between biological sequences.
- If similarities are not due chance then they must be due to something else!
 - Homology
 - Simple identification
- All BLAST searches begin with a sequence
 - protein or nucleotide
 - experimentally determined or one from database

06/06/2016

5

What BLAST tells you

NCBI Public Services

Here's my sequence.

1. What is it related to? (What does it do?)
 - Homology
 - Function
2. Is it already in the database? (Identification)
 - find the matching sequence in the database
 - organism of origin
3. Where is it located or how is it organized?
 - in a genome
 - other annotation problems
 - comparing sequences
 - looking for frame shifts

06/06/2016

6

BLAST Statistics

Score = 18.5 bits (36), **Expect = 47992**
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query 1 ELVIS 5
 ELVIS
 Sbjct 8 ELVIS 12

- Number of chance alignments = 48 thousand!
- Indistinguishable from chance

The most important statistic: **Expect value (e-value)**
 Expected number of random alignments with a particular score or better

Score = 89.7 bits (204), **Expect = 7e-18**
 Identities = 50/103 (49%), Positives = 54/103 (52%), Gaps = 18/103 (17%)

Query 1 MKLLAATVL---LLTICSLEGALVR...
 MK L VL LL +CSLEGA V
 Sbjct 1 MKVL---VLAMVLLCVCSLEGAVVM

- Number of chance alignments = 7×10^{-18}
- Not due to chance

Query 54 SPELQAEAKSYFEKSKEQLTPLIKKAGTELNVFLSYFVELGTQ 96
 E +AK Y E EQ P K TE F +L TQ
 Sbjct 57 AEEIKTOAKAYLEOANEQFSPIAKRLHTE-----FMDLLTQ 92

- The e-value depends directly on the size of the search space (database)
- Search the smallest database likely to contain the sequence of interest

Scoring: Nucleotide

Number of Chance Alignments = 2×10^{-73}

Score = 288 bits (318), Expect = $2e^{-73}$
 Identities = 262/325 (81%), Gaps = 8/325 (2%)
 Strand=Plus/Plus

```

Query 1923 TCAGCCTACCATGAGAATAAGAGAAAGA-AAATGAAGATCAAAGCTTATTCATCTGTTT 1981
Sbjct 33774 TCAGACTACCC TGAGAATAAGAGAAAGAAATGAAGACCTAGA-CTTATCCATCTCTTT 33832
Query 1982 TTTTTTTCGTTGGTGAAGCCAACCCCTGTCTAAAAACATAAAATTTCTTTAATCAT 2041
Sbjct LAAATTTCTTTAAATAT 33892
Query 2042 TTTGCCCTTTTTCT Mismatch=-3 SAATCTAATAGAGTGGT 2100
Sbjct 33893 TTTGCCCTTTTTCTCTGTCTACAAATTAATAAAAAAATGAAAAGAAATCTAATTTAATTGT 33952
Query 2101 ACAGCACTGTTA-TTTTTCALMGAGCCCTGCTGAGCCGCAAAAATGCTAGCTTCTCTGG 2159
Sbjct 33953 CTATGACTGTATTG Gap TTCTATGA 34012
Query 2160 AAGTTCAGTGTCT - (5 + 4(2)) = -13 GTGGGCTA 2219
Sbjct 34013 AAATTCACATATCTCTCTTCCCTATTTTCAATGGAGGACATCTAGTTCCTTCTGGATTA 34072
Query 2220 AT---TAAATAAATCATTAAATACT 2240
Sbjct 34073 ATTGCATAAAGAAACATTAAATACT 34097
    
```

Scoring: Protein

Number of Chance Alignments = 4×10^{-50}

Score = 176 bits (447), Expect = 4e-50, Method: Compositional matrix adjust.
Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

```

Query 30  MAKVLTLELYKKLRDKETPSGFTVDDVIQTGV--DNPGHPPIMTVGCVAGDEESYEVFKE 87
          + K LT +L+++ +D+  GF+  I +G  N G  VG AG +SY F
Sbjct 26  LQKCLTKDLWEQCKDRRDYGFSPKQAI FSGSKWTNSG-----VGVYAGSHDSYYAFAP 79

Query 88  LFDPIISDAHGGYKPTDKHKTDLNHEHLKGG---DDLDPNYVLSSRVRTGRSIRQYTLPP 144
          D  D  DKH  D D  +S+R+R
          FMD  SDKHIS  PADED-KMINSTRIRVA 137
          HCS  E +2  ALNSI  F -3  SMTEKEQQQLIDHFLF 204
          +  AL  +M++ E++QLI DHFLF
Sbjct 138  AVTRKERKEIEHLVTSALGFTGELKGYVCEPMDAPKVKLTAUHLR KQDPKMLQ 196

Query 205  SGMARDWPDARGIWHNDNKSFLVWVNEEI
          +G+ RDWP+ARGI+HND K+FLVWVNEEI
Sbjct 197  AGLERDWPEARGLFHNDAKTFLVWVNEEI

          Gap
          - (11 + 4(1)) = -14

```

Scores from [BLOSUM62](#), a position independent matrix

- Same substitution gets the same score at all positions
- All positions equally likely to change

06/06/2016

9

BLAST Search Programs

06/06/2016

10

Nucleotide Search Programs

NCBI Public Services

- **blastn**
 - traditional BLAST algorithm
 - most sensitive nucleotide search
- **megablast**
 - larger word size than blastn
 - different gapping model
 - **Contiguous megablast**
 - Nearly identical sequences
 - **Discontiguous megablast**
 - Cross-species comparisons

- Default nucleotide search program
- Best for
 - Identification
 - Same-species annotation

06/06/2016

11

Protein Search Programs

NCBI Public Services

Position Independent scoring

- **blastp**
- **translating searches**
 - useful for unannotated protein coding regions
 - six frame translations of query, database or both
 - **blastx** – translated query
 - **tblastn** – translated database
 - **tblastx** – translated query and database

06/06/2016

12

Protein Domains and Position Specific Scoring

Position-specific scoring model

- Multiple alignment based
- Substitution scores depend on the position in the protein.
- Some positions are more important (less likely to change)
- More accurate alignments
- More sensitive at identifying distant homologies
- Better at identifying structural / functional domain

06/06/2016

13

Position-specific Programs (protein only)

- **Position Specific Iterative BLAST (PSI-BLAST)**
Automatically generates a position specific score matrix (PSSM) from initial set of BLAST alignments
- **Position-Hit Initiated BLAST (PHI-BLAST)**
Focuses search around pattern (motif)
- **Domain Enhanced Lookup Time Accelerated (DELTA) BLAST**
Uses conserved domain PSSM in first round of search
- **Reverse PSI-BLAST (RPS-BLAST)**
Searches a database of PSI-BLAST PSSMs
Conserved Domain Database Search
Quickly identifies type of protein and potential function
 - Runs with all blastp searches at the NCBI
 - Identifies conserved domains in query

06/06/2016

14

Web Access

www.ncbi.nlm.nih.gov

The screenshot shows the NCBI homepage with a Google search bar containing 'ncbi blast'. The search results show 'Basic Local Alignment Search Tool (BLAST) - NCBI' as the top result. A blue arrow points to the 'BLAST' link in the search results. The NCBI logo and navigation menu are visible at the top.

NCBI Public Services

06/06/2016 15

BLAST Homepage

blast.ncbi.nlm.nih.gov

The screenshot shows the BLAST homepage with several overlapping callout boxes. The main page has a search bar and a 'GO' button. The callout boxes highlight different search options: 'BLAST Assembled Genomes', 'Basic BLAST', and 'Specialized BLAST'. The 'Basic BLAST' box shows options for 'nucleotide blast' and 'protein blast'. The 'Specialized BLAST' box shows various search types like 'Primer-BLAST', 'MOLE-BLAST', etc.

NCBI Public Services

06/06/2016 16

The New BLAST Homepage

BLAST®
Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

New BLAST home page preview.

The new design provides improved navigation, a cleaner look, and easier access to specialized BLAST services.

Thu, 12 May 2016 08:00:00 EST [More BLAST news...](#)

Standalone and API BLAST

[Download BLAST](#) Get BLAST databases and executables
 [Use BLAST API](#) Call
 [Use BLAST in the cloud](#)

Specialized searches

<p>SmartBLAST</p> <p>Find proteins highly similar to your query</p>	<p>Primer-BLAST</p> <p>Design primers specific to your PCR template</p>	<p>Global Align</p> <p>Compare two sequences across their entire span</p>	<p>CD-search</p> <p>Find conserved domains in your sequence</p>
<p>GEO</p> <p>Find matches to gene expression profiles</p>	<p>IgBLAST</p> <p>Search immunoglobulins and T cell receptor sequences</p>	<p>Vecscreen</p> <p>Search sequences for vector contamination</p>	<p>CDART</p> <p>Find sequences with similar conserved domain architecture</p>
<p>Targeted Loci</p> <p>Search markers for phylogenetic analysis</p>	<p>Multiple Alignment</p> <p>Align sequences using domain and protein constraints</p>	<p>BioAssay</p> <p>Search protein or nucleotide targets in PubChem BioAssay</p>	<p>MOLE-BLAST</p> <p>Establish taxonomy for uncultured or environmental sequences</p>

Nucleotide BLAST

nucleotide → nucleotide

tblastn

protein → translated nucleotide

BLAST Genomes

Enter organism common name, scientific name, or taxon

Human Mouse Rat Micro

06/06/2016
17

Query Sequences

06/06/2016
18

Queries

```

>seq1
CGAATTTGATCCTGGCTCAGAAATCAAGCTGGCGGCTGGCTCAGCCATGCAAGTCGAA
CGATTAATCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTACCCCTCTATTGGGATTACCTCCGGACACTTCTCTCTCTACCCCTATCCACCTCT
CCCTATGGCCCTTTTCTCTATTATCCCTTGTTCAGTTCTGGCCGCTTAGCTA
TTTGTGGTAAATCGCTACCAAGCTCGCTCTGTAACCGCTTAGAGCCGCTC
>seq2
CGAATTTGATCCTGGCTCAGAGCGAAGCTGGCGGCTGGCTAACACATGCCAGTCGAC
CGCCCTATCCTCCGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTTCAGAACCTGGTACCTCTGGCTCTACCTCTACCTCGCTGGCTCTCTCTCTCTCTCT
TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
>seq3
AGAATTTGATCCTGGCTCAGATTGAACCTGGCGGCTGGCTCACACATGCAAGTCGAA
GCCCCACCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT

```

Enter accession number(s), gi(s), or FASTA sequence(s)

NP_000032.1
 NP_033826.2
 NP_776416.1
 NP_001018401.1
 NP_001076112.1

Job Title: 6 sequences (seq1)

Analyze these sequences

- Run BLAST
- Align sequences with COBALT
- Identify Conserved Domains with CD-Search

1. **apolipoprotein E isoform b precursor**
 317 aa protein
 Accession: NP_000032.1 Gi: 4557329
 GenPept FASTA Graphics Related Sequences Identical Proteins

2. **apolipoprotein E precursor [Mus musculus]**
 311 aa protein
 Accession: NP_033826.2 Gi: 163644329
 GenPept FASTA Graphics Related Sequences Identical Proteins

3. **apolipoprotein E precursor [Bos taurus]**
 316 aa protein
 Accession: NP_776416.1 Gi: 27806739
 GenPept FASTA Graphics Related Sequences Identical Proteins

4. **apolipoprotein E precursor [Danio rerio]**
 269 aa protein
 Accession: NP_001018401.1 Gi: 68472620
 GenPept FASTA Graphics Related Sequences Identical Proteins

- FASTA format, single or multiple
- Accessions, single or multiple
 - Directly from the sequence dbs

06/06/2016

19

BLAST 2 (or more) Sequences

Standard Protein BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

>NP_000468
 MKWVTFISLFLFSSAYSRCVFRDRAHKSEVAHRFKDLGEENFKALVLIQFAQYL
 VNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQ
 KDDNPNLRLVRPEVDVMCTAFHDNEETFLKYLVEIARRHPFYAPPELLFFAK
 AADKAACLPLKLDLDEGKASSAKQRLKCSLQKQGERAFKAWAVARLSQR

Job Title: NP_000468

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Align Sequences Protein BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

>NP_000468
 MKWVTFISLFLFSSAYSRCVFRDRAHKSEVAHRFKDLGEENFKALVLIQFAQYLQCCPFEDHVKL
 VNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLOH
 KDDNPNLRLVRPEVDVMCTAFHDNEETFLKYLVEIARRHPFYAPPELLFFAKRYKAFFTECCQ
 AADKAACLPLKLDLDEGKASSAKQRLKCSLQKQGERAFKAWAVARLSQRFPKAEFAEVSKL

Job Title:

Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

QLSEDKLLACGEGAAADIIGHLCIRHEMTPVPGVGCCTSSYANRRPCFSSLVVDETYVPPAFSD
 DKFIFHKDLCAQGVALQTMKQEFINLVKQKQITEQLEAVIADFSGLLEKCCGGQEQVEVCF
 EEQQLKISKTRALGV
 >NP_000574
 MKLLKLTGFIFLFLTESLTPTRDIENFNSTQKFIENIEYIIAFAQYVQEQATFEEMKLVKD

BLAST Search protein sequence using Blastp (protein-protein BLAST)

- Any search page convertible to BLAST 2 (or more) Seqs
- Can search small custom database
- Also available under "Specialized BLAST"
- Many who use this really want a global alignment

06/06/2016

20

Global Alignment Tool

NCBI Public Services

Specialized BLAST

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

- Make specific primers with [Primer-BLAST](#)
- Cluster multiple sequences together with their database neighbors
- 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 and T cell receptor sequences](#) (igb)
- Screen sequence for [vector contamination](#) (vecscreen)
- Align two (or more) sequences using BLAST (tbl2seq)**
- Search [protein or nucleotide](#) targets in PubChem BioAssay
- Search [SRA by experiment](#)
- Consistent-based Protein Multiple Alignment Tool**
- Needleman-Wunsch Global Sequence Alignment Tool**
- Search [refseqs](#)
- Search [trace archives](#)

	NW Score	Identities	Positives	Gaps	
	267	127/613(21%)	228/613(37%)	143/613(23%)	
Query 1					60
Sbjct 1					57
Query 61					120
Sbjct 58					117
Query 121					180
Sbjct 118					177
Query 181					240
Sbjct 178					236
Query 241					299
Sbjct 237					296
Query 300					358
Sbjct 297					350
Query 359					418
Sbjct 351					394
Query 419					478
Sbjct 395					408
Query 479					538
Sbjct 409					416
Query 539					596
Sbjct 417					467

NP_000468 (ALB) vs. NP_000574 (GC)

06/06/2016 21

Restrictions on Web BLAST

NCBI Public Services

⚠ There was a problem with the search. Please, contact [Help Desk](#) and include RID 85ZEZVEW01R

ℹ Informational Message: [blastsrv4.REAL]: Error: CPU usage limit was exceeded, resulting in SIGXCPU (24).

- No fixed limit on size / number of queries
- Web BLAST limited by processing time (one hour) for single search
 - Difficult to estimate processing time
 - Factors
 - Size and number of queries
 - Size of database
 - Depends on size and complexity of results too

06/06/2016 22

BLAST Databases

06/06/2016

23

Protein Databases

Choose Search Database

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- UniProtKB/Swiss-Prot (swissprot)
- Patented protein sequences (pat)
- Protein Data Bank proteins (pdb)
- Metagenomic proteins (env_nr)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

Exclude Models (XMI/XP) Uncultured/environmental sample sequences

Services
blastp
blastx

- Default database (nr)
 - Most comprehensive
 - Useful subsets: RefSeq, Swiss-Prot, PDB
- What's not in nr?
 - US, European and Asian Patents
 - Proteins from metagenomes
 - Proteins from Next-Gen assemblies

06/06/2016

24

Nucleotide Databases

NCBI Public Services

Choose Search Set

<p>Database</p> <p>Organism Optional</p> <p>Exclude Optional</p> <p>Limit to Optional</p> <p>Entrez Query Optional</p>	<p>Genomic plus Transcript</p> <p>Human genomic plus transcript (Human G+T)</p> <p>Mouse genomic plus transcript (Mouse G+T)</p> <p>Other Databases</p> <p>✓ Nucleotide collection (nr/nt)</p> <p>Reference RNA sequences (refseq_rna)</p> <p>Reference genomic sequences (refseq_genomic)</p> <p>RefSeq Representative genomes (refseq_representative_genomes)</p> <p>NCBI Genomes (chromosome)</p> <p>Expressed sequence tags (est)</p> <p>Genomic survey sequences (gss)</p> <p>High throughput genomic sequences (HTGS)</p> <p>Patent sequences(pat)</p> <p>Protein Data Bank (pdb)</p> <p>Human ALU repeat elements (alu_repeats)</p> <p>Sequence tagged sites (dbsts)</p> <p>Whole-genome shotgun contigs (wgs)</p> <p>Transcriptome Shotgun Assembly (TSA)</p> <p>16S ribosomal RNA sequences (Bacteria and Archaea)</p> <p>Sequence Read Archive (SRA)</p>	<p>Others (nr etc.):</p> <p>de +</p> <p>be shown</p> <p>ces</p> <p>Create custom database</p>
---	--	---

Services

megablast

blastn

tblastn

tblastx

06/06/2016 25

Nucleotide Databases

NCBI Public Services

- Default database ([nr/nt](#)) is not comprehensive
 - Contains traditional GenBank and RefSeq RNA
 - Useful subsets: [RefSeq RNA](#), [16S rRNA](#) reference sequences
- What is not in nr/nt? The majority of nucleotide data
 - Bulk sequences ([EST](#), [GSS](#), [HTGS](#), [STS](#))
 - RefSeq Genomic Sequences ([Chromosome](#), [RefSeq Genomic](#), [RefSeq Representative Genomes](#))
 - US, European and Asian Patents ([pat](#))
 - Whole Genome Shotgun Contigs (WGS) (Second Largest)
 - Transcriptome Shotgun Assemblies (TSA)
 - Next-Gen RNA-Seq, DNA-Seq Reads (SRA) (Largest set of data)

06/06/2016 26

Limiting Databases

Search the smallest database likely to contain the sequence of interest.

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional
 bacteria (taxid:2) Exclude +
 Enterobacteriales (taxid:91347) Exclude
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional
 Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional
 25000:30000[Molecular Weight]
 Enter an Entrez query to limit search

Exclude predicted and uncultured

Limit with Entrez query

06/06/2016

27

NCBI Public Services

Genome Databases

BLAST Assembled Genomes

Find Genomic BLAST pages:

chimpanzee (taxid:9597)
 pygmy chimpanzee (taxid:9597)
 chimpanzee (taxid:9598)
 plownose chimaera (taxid:7868)
 chimney swift (taxid:8897)
 Chimarra obscura (taxid:178329)
 Plasmodium sp. chimpanzee clade C2 (taxid:8792...)
 chimney bellflower (taxid:239419)
 Vibrio mar...
 Chimarra socia (taxid:692083)
 Chimarrigale himalayica (taxid:227887)
 Mycobacterium chimaera (taxid:222805)
 Plasmodium sp. chimpanzee clade C1 (taxid:8805...)
 Plasmodium sp. chimpanzee clade C3 (taxid:8805...)
 Chimarra sp. AMI 1 (taxid:888128)
 Tropheus sp. 'Chimba' (taxid:1038501)
 Chimonomalampus pallens (taxid:145982)
 chimpanzee louse (taxid:240286)

GO

Human
 Mouse
 Rat
 Cow
 Pig
 Dog
 Rabbit
 Chimp
 Guinea pig
 Fruit fly
 Honey bee
 Chicken
 Zebrafish
 Clawed frog
 Arabidopsis
 Rice
 Yeast
 Microbes

Shortcuts to popular organisms

- Comprehensive search for genomic data
- Finds the best set (most assembled) of genomic sequences

...ing a protein query
 ...ast, phi-blast, delta-blast
 ...ing a translated nucleotide query
 ...e database using a protein query
 ...e database using a translated nucleotide query
 ...name in parentheses)

06/06/2016

28

NCBI Public Services

Web Program Selection

06/06/2016

29

Nucleotide Programs

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear Query subrange

From To

Or, upload file

Browse...

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Sensitivity

Speed

More

Less

BLAST Search database Human G+T using Megablast (Optimize for highly similar sequences)

Show results in a new window

06/06/2016

30

Protein Programs

NCBI Public Services

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

PHI-BLAST (Pattern Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

Show results in a new window

06/06/2016

31

Algorithm Parameters: General

NCBI Public Services

General Parameters

Max target sequences Select the number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Set to more stringent value:

- 1e-6
- 0.001

Increase Max target sequences

Decrease Expect threshold

Let Expect threshold govern output not Max target sequences

06/06/2016

32

Nucleotide Repeat Filters

Filters and Masking

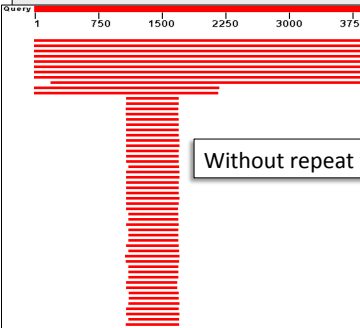
Filter

- Low complexity regions
- Species-specific repeats for: Homo sapiens (Human)

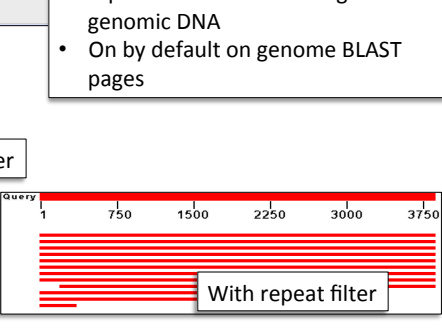
Mask

- Mask for lookup table only
- Mask lower case letters

- Select the matching interspersed repeat filter when working with genomic DNA
- On by default on genome BLAST pages



Without repeat filter



With repeat filter

06/06/2016

33

Formatting options

Pairwise

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

- Dots for identities
- Coding Sequence

Alignment View Pairwise with dots for identities

CDS feature

Color: Grey

Physical overview: 100 Line length: 60

Entrez query:

```

Query 7  AAGTCA-GGGCCAGTGGCGCAGCCAGCTAGCACCCAGTACCGCGCGTGTGGGCACCAT 65
Sbjct 975 AAGTCAGGGGCCAGTGGCGCTACAGCCAGCGCCAGTACCGCGGTGTGATGGGCACCAT 1034

CDS: Putative 1 1 S Q G Q C A R O L A P S T A A C N A P
Query 7 AAGTCA-GGGCCAGTGGCGCAGCCAGCTAGCACCCAGTACCGCGCGTGTGGGCACCAT 65
Sbjct 975 .....G.....G.TA....C..G.....T..A..... 1034
CDS:PREDICTED: mitoc 47 S Q G P V R A T A S A Q Y R G V M G T I

CDS: Putative 1 20 S * P W C A P R A P A A A S T T G W S P A
Query 66 CCTTACAACGGGCTGGTCCGGCCT 125
Sbjct 1035 T.....T.....T..... 1094
CDS:PREDICTED: mitoc 67 L Y N G L V A G L

CDS: Putative 1 39 C S A S T T L S S S
Query 126 GC CGGCCCTACGACTCTGTCAAGCAGTT 185
Sbjct 1095 .. ..G..T..T.....A..... 1154
CDS:PREDICTED: mitoc 87 G L Y D S V K Q F

CDS: Putative 1 58 S T P R A L S M L A L G A A S W P A A P
Query 186 CTACACCAAGGCTCTGAGCATGTGGCATTTGGAGCCGCTCTGGCCGCGAGCAC 245
Sbjct 1155 .....CA.....A.A..... 1214
CDS:PREDICTED: mitoc 107 Y T K G S E H A S I G S R L L A G S T T
                    
```

Highlights

- frameshifts
- sequence changes
- Nuc and Prot

06/06/2016

34

Managing Your Results

06/06/2016

35

The Request ID (RID) is the key

The BLAST Request ID (RID) is the key

refNP_001246.2| (499 letters)

RID HKZG2PPT013 (Expires on 04-01 09:53 am)	Database Name refseq_protein
Query ID gilt12 gilt12 refNP_001246.2 	Description NCBI Protein Reference Sequences
Description cell division cycle protein 20 homolog [Homo sapiens]	Program BLASTP 2.2.31+ Citation
Molecule type amino acid	
Query Length 499	

<http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=HKZG2PPT013>

- Uniquely identifies search settings and results
- Persists at NCBI for 36 hours
- View through Recent Results, My NCBI
- Allows sharing results and reformatting
- Send the RID to blast-help@ncbi.nlm.nih.gov to ask about a search

06/06/2016

36

Download Options

Download

Alignment

Text XML ASN.1 JSON Seq-align Hit Table(text) Hit
 Table(csv) Multiple-file XML2 Single-file XML2 Multiple-file
 JSON Single-file JSON SAM

Search Strategies

ASN.1

```

19 sec
# Query: seq6
# Iteration: 0
# RID: HNS0K2AV013
# Database: rRNA_typestrains/prokaryotic_16S_r1bosom1_SNA
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, eval, bit score
# 100 hits found
seq6 gi1444383911|ref|NR_074324.1| 90.27 298 29 0 1 298 9 386 2e-188 398
seq6 gi13431902491|ref|NR_040936.1| 89.97 299 28 2 1 298 1 298 8e-187 385
seq6 gi16312589931|ref|NR_112190.1| 89.93 298 26 3 1 298 1 294 1e-185 381
seq6 gi14443837831|ref|NR_074205.1| 89.63 299 28 2 1 298 1 297 1e-184 377
seq6 gi14443838911|ref|NR_074313.1| 89.38 299 30 2 1 298 10 307 2e-183 374
seq6 gi13432061541|ref|NR_044744.1| 89.23 297 31 1 1 297 11 306 2e-182 370

```

- Downloads all data for multiple queries in a single file
- XML / XML2 easiest to parse with script and / or redisplay
- Hit table compatible with Excel and other spreadsheet programs
- Search strategies can be used again on the web or in standalone

06/06/2016
37

Specialized BLAST Services

06/06/2016
38

Nucleotide Services

NCBI Public Services

- **PrimerBlast**
 - primer designer / specificity checker
 - Primer3 primer design
 - Uses RefSeq annotation
 - exon boundaries
 - splice variants
 - SNPs
- **MOLE-BLAST**
 - Helps identify sources of 16S and other targeted sequences
 - BLAST followed by global multiple alignment
 - Clusters queries plus most similar database sequences
 - Identifies taxonomic units (neighbors)
 - Labels database sequences from type material for accurate ID

06/06/2016

39

Protein Services

NCBI Public Services

- **COBALT** – Constraint Based Alignment Tool
 - Protein global multiple alignment tool
 - Uses conserved domains to guide alignment
 - Extension to BLAST search
- **SmartBLAST** – Rapid protein identification tool
 - Uses fast k-mer search
 - Identifies closest match in reference organism database
 - Produces multiple alignment and protein tree
 - Prototype for on-the-fly protein similarity (BLink)

06/06/2016

40

BLAST Help

NCBI Public Services

BLAST® Home Recent Results Saved Strategies Help

BLAST documentation

<p>Getting Started</p> <ul style="list-style-type: none"> • Guide to BLAST home and search pages • BLAST interface description • Blast report description <p>About BLAST</p> <ul style="list-style-type: none"> • Frequently Asked Questions • NCBI Handbook: BLAST • The Statistics of Sequence Similarity Scores • NAR 2004 Web server issue • NAR 2006 Web server issue • NAR 2008 Web server issue • BLAST glossary • References • Blast+ Command Line Applications User Manual • BLAST News directory 	<p>Getting Help</p> <ul style="list-style-type: none"> • Email blast-help • Mailing list • YouTube BLAST tutorials <p>Other BLAST information</p> <ul style="list-style-type: none"> • Download BLAST Software and Databases • Developer information • BLAST Searches at a Cloud Provider
---	---

Help desk: blast-help@ncbi.nlm.nih.gov

06/06/2016 41

More Help Links

NCBI Public Services

- Help Manual: ncbi/books/NBK3831/
- Learn: ncbi/home/learn.shtml
- Factsheets: <ftp://pub/factsheets/>
- NCBI YouTube: youtube/ncbinlm
- NCBI Helpdesks
 - General: info@ncbi.nlm.nih.gov
 - BLAST: blast-help@ncbi.nlm.nih.gov

06/06/2016 42

Web Demonstrations

- Basic BLAST
 - blastp, MLH1
 - COBALT extension
- Genome BLAST
 - blastn, macaque CDC20
 - Formatting options
 - Genome context
- SRA BLAST
 - Melanoma gene expression
- Primer BLAST
 - BRCA1 Exon Primers
- SmartBLAST
 - Finding yeast MLH1