joanne@msl.ubc.ca

wireless login:
mslguest
4myguest

Laboratory Bioinformatics

Common tools, useful databases, and tricks of the trade for practical use in the laboratory.

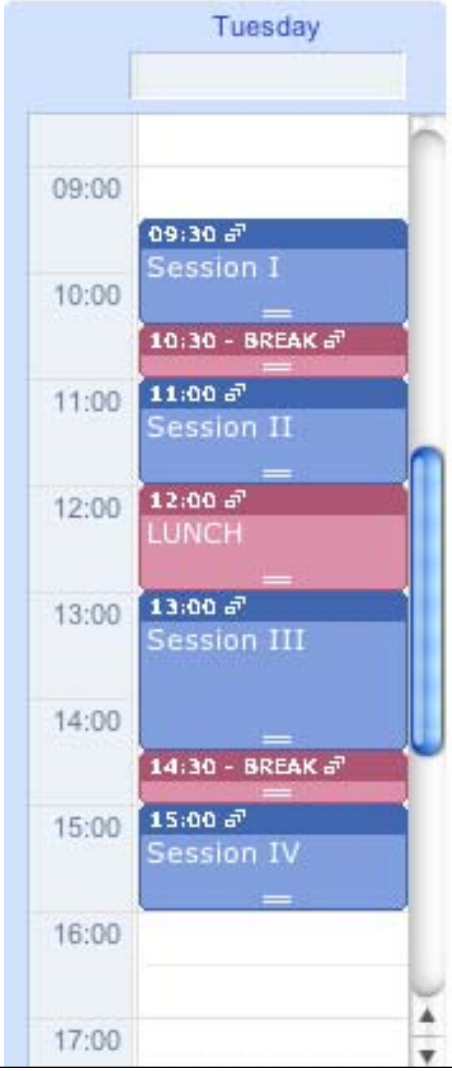
bioteach.ubc.ca/bioinfo2009

Workshop Schedule

- Laptops, available here for your use 9am - 4:30pm
- wireless login

  mslguest

  4myguest
- Vancouver guide books available



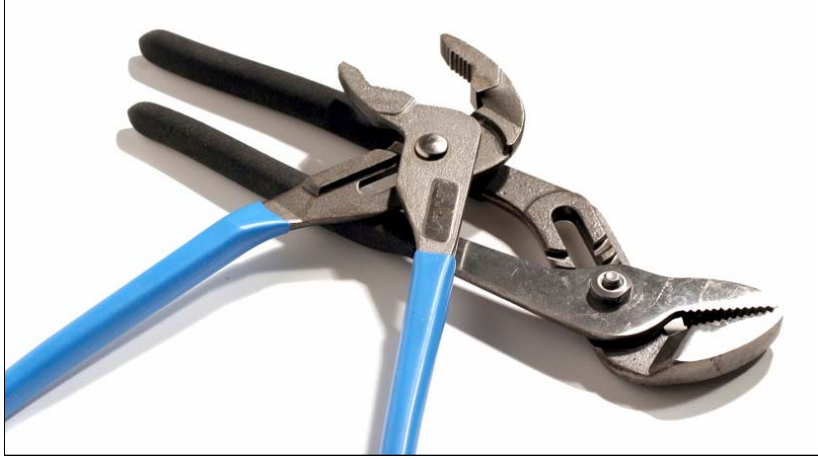

Thank You
Enjoy
Thank
You

Today's Topics

- **BLAST -** Finding Function by Sequence Similarity
- **GUIDED TOUR** - Advanced Tips & Tricks for Using BLAST
- **PRACTICAL EXERCISES** - The Jurassic Park Detective Story
- **COMMON TASKS -** Basic Search; Searching Sets of Sequences (multiple inputs; small custom databases); Primer Design

BLAST

Finding Function By Sequence Similarity



What do the Score and the e-value really mean?

- The quality of the alignment is represented by the Score (S).

  The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .

- The significance of each alignment is computed as an E value (E).

  Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".

BLAST Algorithm

Query Word (W = 3)

TLSHAWRLSNETDKRPFIETAERL**RDQ**HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE

Determine neighborhood

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".

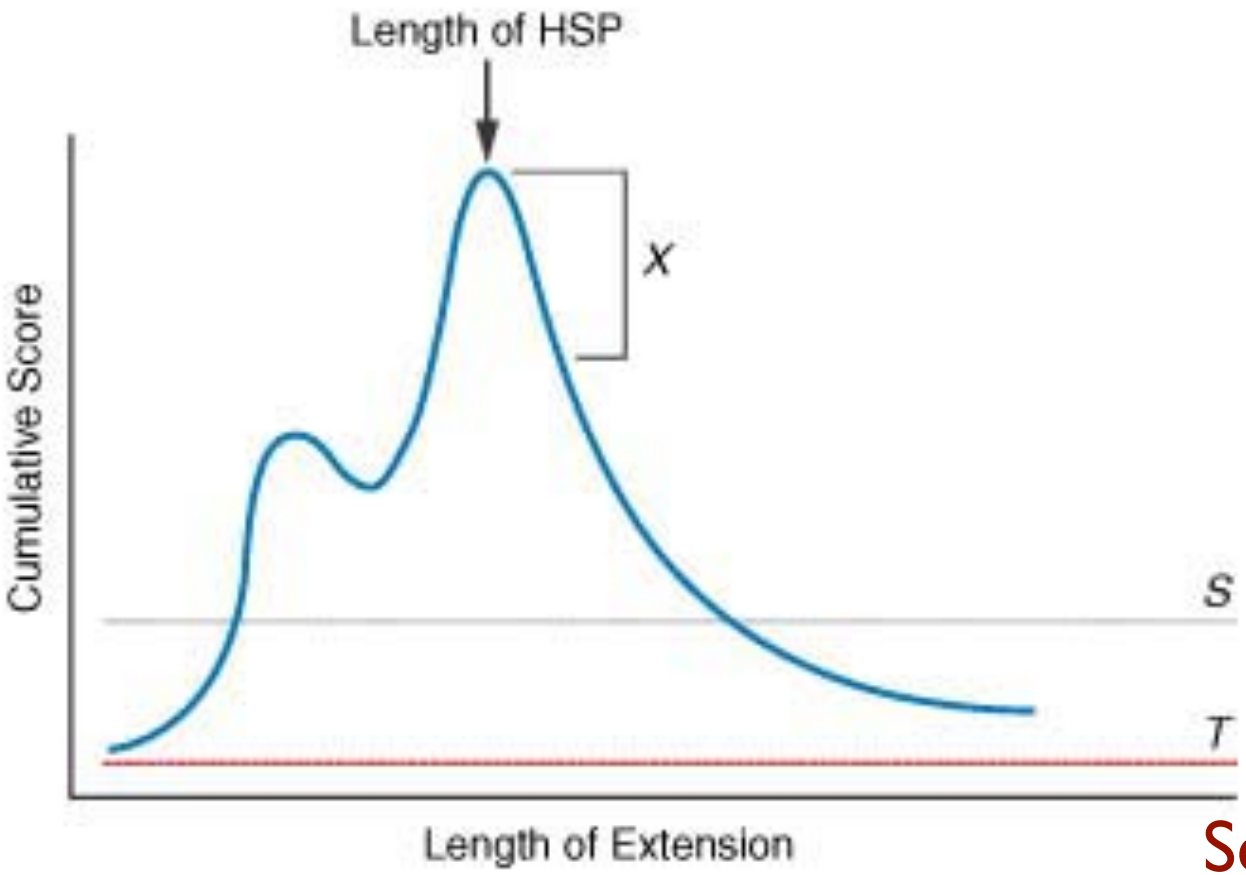
BLAST Algorithm

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

Extension using neighborhood words greater than neighborhood score threshold (T = 11)

```
Query: 1   TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
           TL   WRL N  +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S   D   +
Sbjct: 140 TLESGWRLENPGEKRPFVEGAERLREQHKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197
```

Extending the High Scoring Segment Pair (HSP)



Minimum Score (S)

Neighborhood Score Threshold (T)

```
>gb|AAL08419.1|  PTEN [Takifugu rubripes]
Length=412

 Score =  197 bits (501),  Expect = 2e-49, Method: Composition-based stats.
 Identities = 95/100 (95%), Positives = 98/100 (98%), Gaps = 0/100 (0%)

Query  2    IVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  61
            +VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI
Sbjct  8    MVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  67

Query  62   YNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            YNLCAERHYD AKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  68   YNLCAERHYDAAKFNCRVAQYPFEDHNPPQLELIKPFCED  107


 Score = 83.6 bits (205),  Expect = 4e-15, Method: Composition-based stats.
 Identities = 60/103 (58%), Positives = 68/103 (66%), Gaps = 32/103 (31%)

Query  99   KQNKMLKKDKMFHFWVNTFFIPGPEEV--------------------------------D  126
            KQNKM+KKDKMFHFWVNTFFIPGPEE                                  +
Sbjct  260  KQNKMMKKDKMFHFWVNTFFIPGPEESRDKLENGAVNNADSQQGVPAPGQGQPQSAECRE  319

Query  127  NDKEYLVLTLTkndldkankdkanRYFSPNFKVKLYFTKTVEE  169
            +D++YL+LTL+KND DKANKDKANRYFSPNFKVKL F+KTVEE
Sbjct  320  SDRDYLILTLSKNDRDKANKDKANRYFSPNFKVKLCFSKTVEE  362


>gb|AAH93110.1|  Ptenb protein [Danio rerio]
Length=289

 Score =  197 bits (500),  Expect = 2e-49, Method: Composition-based stats.
 Identities = 95/99 (95%), Positives = 98/99 (98%), Gaps = 0/99 (0%)

Query  3    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIY  62
            VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct  9    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY  68

Query  63   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  69   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCED  107
```

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

Credits

- Materials for this presentation have been adapted from the following sources:

  NCBI HelpDesk - Field Guide Course Materials

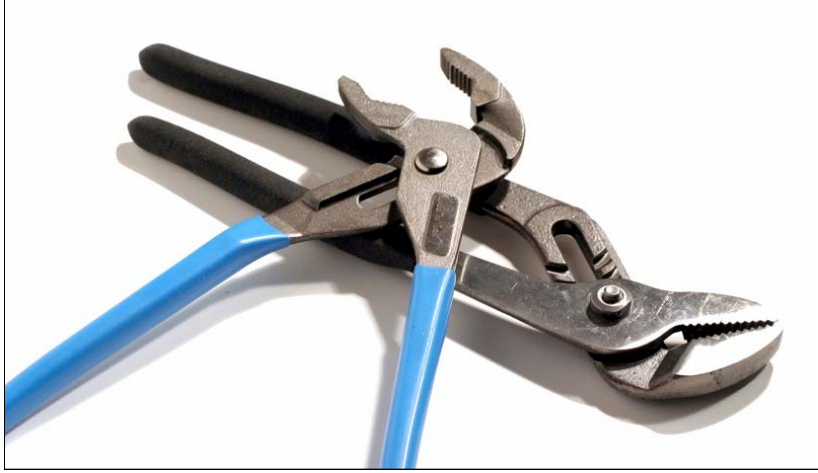
  Bioinformatics: A practical guide to the analysis of genes and proteins

- Questions? Please contact:

Dr. Joanne Fox
Michael Smith Laboratories
joanne@msl.ubc.ca

BLAST

GUIDED TOUR: Advanced Tips & Tricks for Using BLAST



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

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

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

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*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

Align two sequences form.

The Align two sequences link on the BLAST home page now uses the standard BLAST submission form.
Tue, 03 Feb 2009 16:00:00 EST

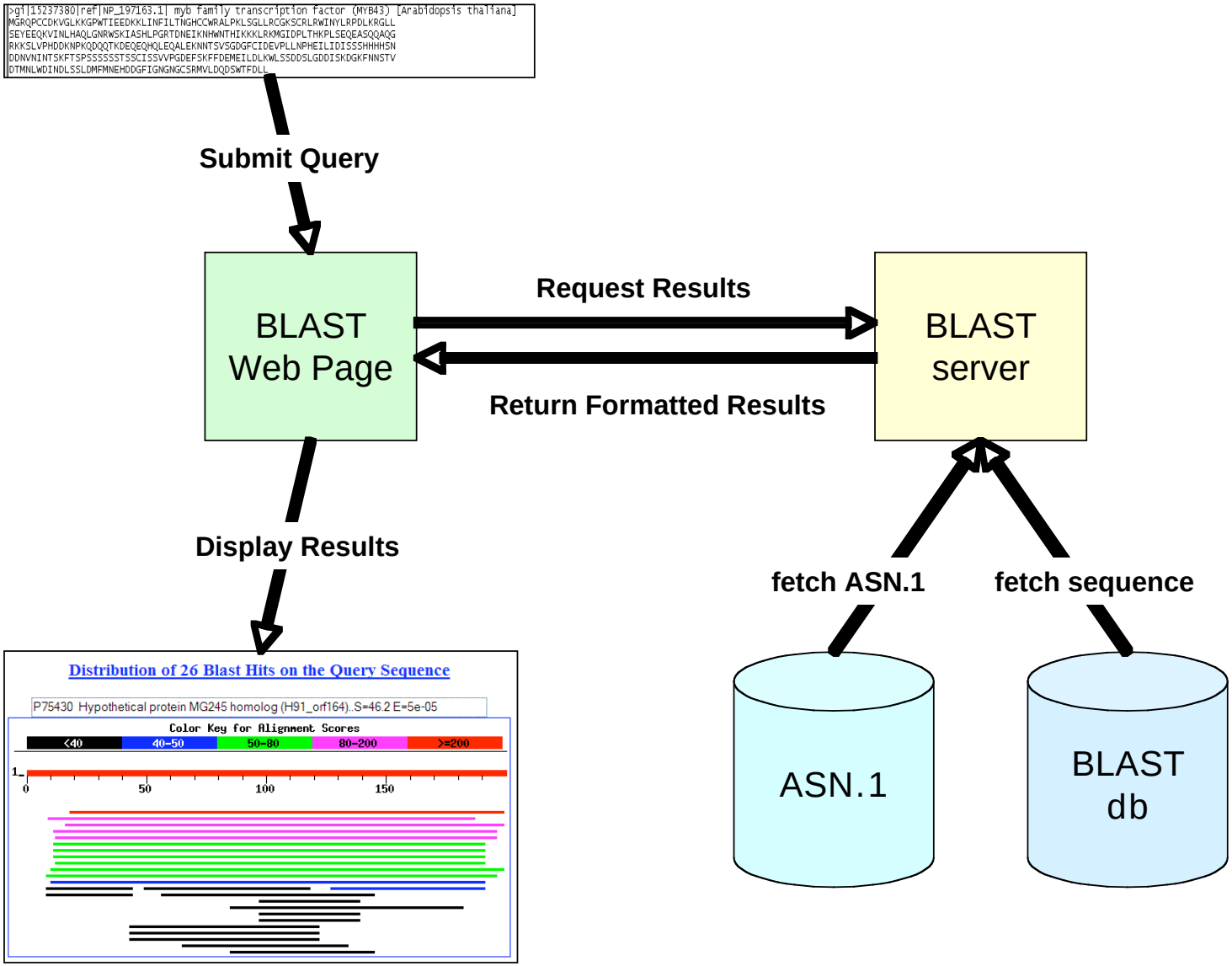
More BLAST news...

Tip of the Day

How to do Batch BLAST jobs.

BLAST makes it easy to examine a large group of potential gene candidates.

More tips...


>gi|15237380|ref|NP_197163.1| myb family transcription factor (MYB43) [Arabidopsis thaliana]
MGRQPCCDKVGLKKGPWTIEEDKKLINFILTNGHCCWRALPKLSGLLRCGKSCRLRWINYLRPDLKRGLL
SEYEEQKVINLHAQLGNRWSKIASHLPGRTDNEIKNHWNTHIKKKLRKMGIDPLTHKPLSEQEASQQAQG
RKKSLVPHDDKNPKQDQQTKDEQEQHQLEQALEKNNTSVSGDGFCIDEVPLLNPHEILIDISSSHHHHSN
DDNVNINTSKFTSPSSSSSSTSSCISSVVPGDEFSKFFDEMEILDLKWLSSDDSLGDDISKDGKFNNSTV
DTMNLWDINDLSSLDMFMNEHDDGFIGNGNGCSRMVLDQDSWTFDLL
Submit Query
BLAST Web Page
Request Results
Return Formatted Results
BLAST server
Display Results
Distribution of 26 Blast Hits on the Query Sequence
P75430 Hypothetical protein MG245 homolog (H91_orf164)..S=46.2 E=5e-05
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1
0
50
100
150
fetch ASN.1
fetch sequence
ASN.1
BLAST db

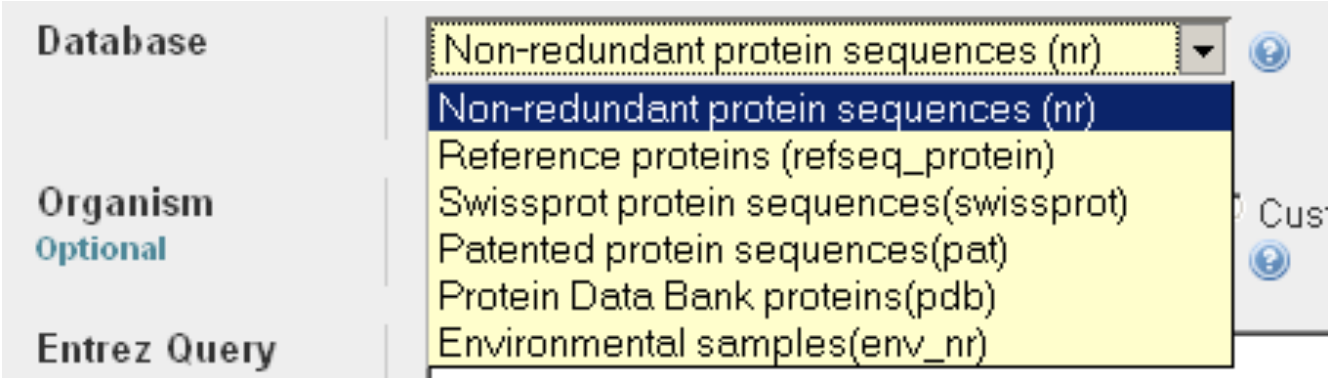
Consider your research question ...

- Are you looking for an particular gene in a particular species?
- Are you looking for additional members of a protein family across all species?
- Are you looking to annotate genes in your species of interest?

Know your reagents

- Changing your choice of database is changing your search space
- Database size affects the BLAST statistics
- Databases change rapidly and are updated frequently

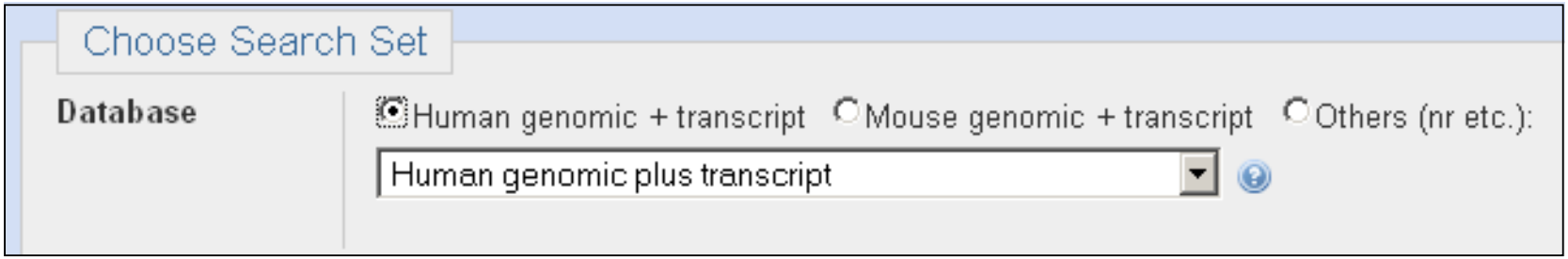
Protein Databases: nr



- nr (non-redundant protein sequences) default
  - GenBank CDS translations
  - NP_ RefSeqs
  - Outside Protein
    - PIR, Swiss-Prot, PRF
    - PDB (sequences from structures)
- pat protein patents
- env_nr environmental samples

<u>Services</u>
blastp
blastx

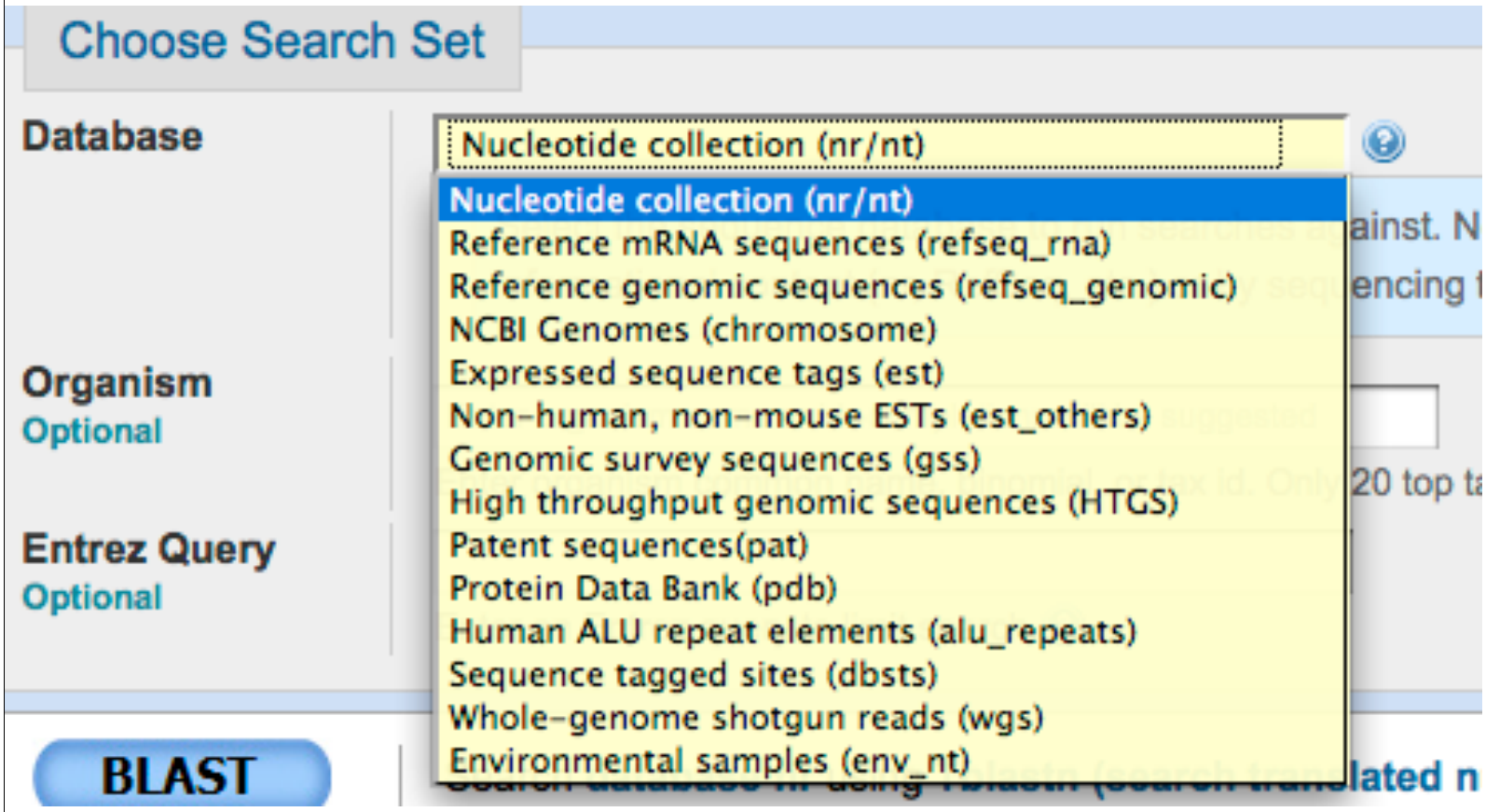
Nucleotide Databases: Human and Mouse



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

Megablast, blastn service

Nucleotide Databases: Traditional



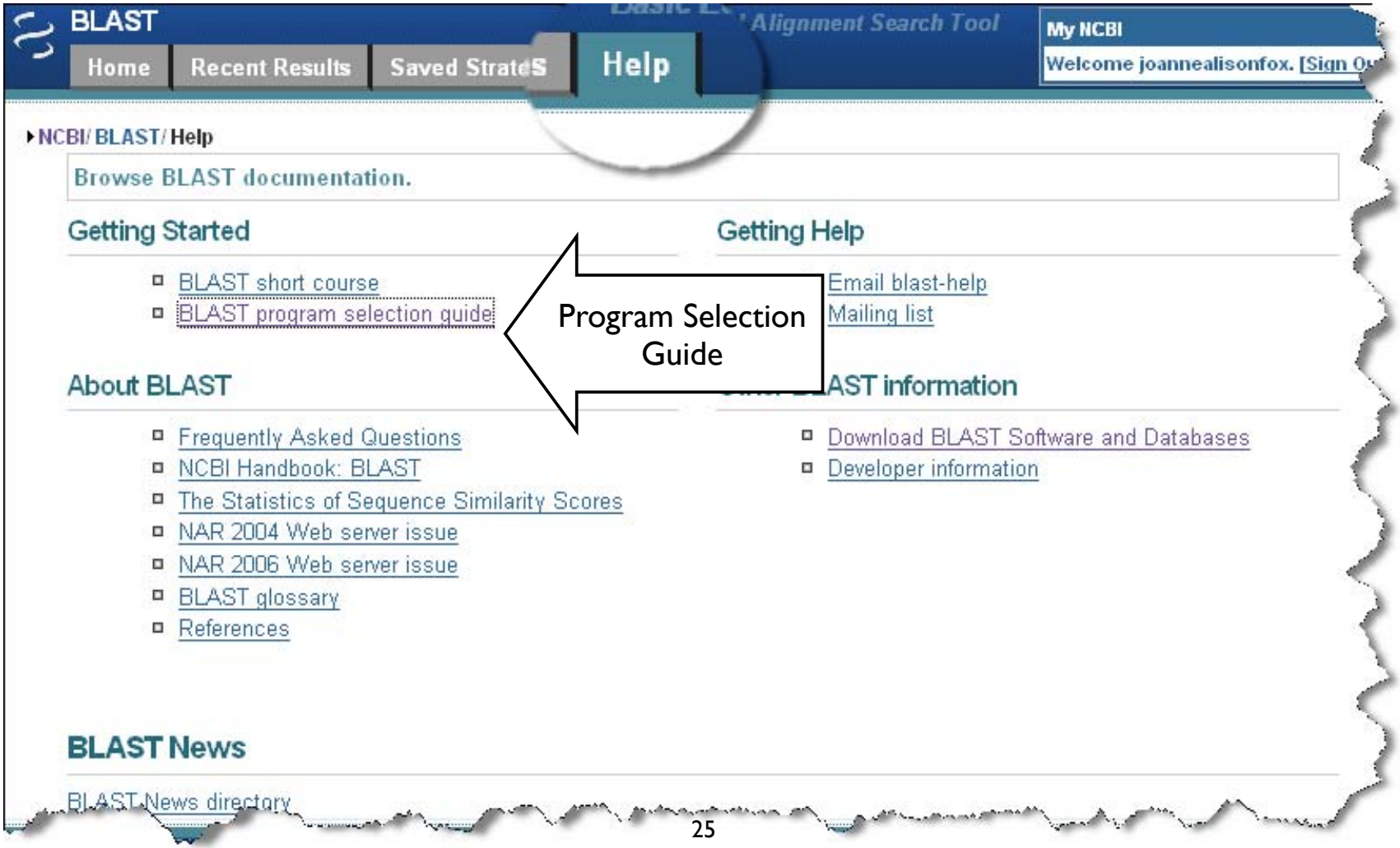
<u>Services</u>
blastn
tblastn
tblastx

Nucleotide Databases:

- **nr (nt)** Traditional GenBank
  + RefSeq nucleotides
  + PDB sequences
- **refseq_rna**
- **refseq_genomic** NC_
- **NCBI genomes**
  complete genomes
  + chromosomes from RefSeq
- **est** expressed sequence tags
  - human + mouse, others
- **htgs** high throughput genomic
  - unfinished
- **gss** genome survey sequence
  - single-pass genomic data
- **pdb** protein data bank
  - derived from 3D structures
- **wgs**
  - whole genome shotgun
- **env_nt**
  - environmental samples

Databases are mostly non-overlapping

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



3. Program Selection Tables

The appropriate selection of a BLAST program for a given search is influenced by the following three factors **1)** the nature of the query, **2)** the purpose of the search, and **3)** the database intended as the target of the search and its availability. The following tables provide recommendations on how to make this selection.

Table 3.1 Program Selection for Nucleotide Queries

Length [1]	Database	Purpose	Program	Explanation
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	discontiguous megablast, megablast, or blastn	Learn more ...
		Find sequences similar to query sequence	discontiguous megablast or blastn	Learn more ...
		Find similar sequence from the Trace archive	Trace megablast, or Trace discontiguous megablast	Learn more ...
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)	Learn more ...
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)	Learn more ...
7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches	Learn more ...

NOTE:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in the Section 4 below. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for MEGABLAST.

Table 3.2 Program Selection for Protein Queries				
Length [1]	Database	Purpose	Program	Explanation
15 residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)	Learn more ...
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST	Learn more ...
		Find proteins similar to the query around a given pattern	PHI-BLAST	Learn more ...
		Find conserved domains in the query	CD-search (RPS-BLAST)	Learn more ...
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)	Learn more ...
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)	Learn more ...
5-15 residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches	Learn more ...

Note:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in Section 4 below.

As genomic and other specialized sequence information is made available to the public, NCBI creates specialized BLAST pages for those sequences. The table below provides a general guide on how to select and use those special BLAST databases.

Table 3.3 Search against Organism Specific or Genome Databases [1]

Query [2]	Database	Purpose	BLAST Pages to Use [3]	Explanation
Nucleotide: 20 or 28 bp and above Protein: 15 residues and above	Human Genome	Map the query sequence Determine the genomic structure Identify novel genes Find homologs Other data mining	Human	Learn more ...
	Mouse Genome		Mouse	Learn more ...
	Rat Genome		Rat	Learn more ...
	Chimp, Cow, Dog, or Chicken Genome		Chimp, or Cow, Dog, Chicken	Learn more ...
	Cat, Sheep, or Pig Genome		Cat, Sheep, or Pig	Learn more ...
	Zebrafish or Fugu (Pufferfish)		Zebrafish or Fugu rubripes	Learn more ...
	Insects (flies and honeybees)		Insects	Learn more ...
	Nematodes (worms)		Nematodes	Learn more ...
	Plants		Plants	Learn more ...
	Fungi Genomes (including yeasts)		Fungi	Learn more ...
	Protozoa		Protozoa	Learn more ...
	Environmental Samples		Environmental Samples	Learn more ...
	Other Lower Eukaryotic Genomes		Other eukaryotes genomes	Learn more ...
	Microbial Genomes		Microbial genomes	Learn more ...

NOTE:
[1] Those pages access the genome database consisting of contig assemblies and other sequences specific to the organisms. Not all organisms listed here have genome assemblies available.
[2] Sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable searches with a short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, -F F -e 20000 -W 2 should be used. For nucleotide comparison, use -F F -e 1000 -W 7. This also requires the uncheck of the megablast checkbox.
[3] Available databases and their contents are described in Section 5.

BLAST pages for special purposes are listed under Special and Meta sections. Their functions are described in Table 3.4 below.

Table 3.4 Function of Special BLAST Pages under Special/Meta Sections				
Query [1]	Database	Purpose	BLAST Page to Use	Explanation
Nucleotide: 11 bp or above Protein: 15 or above	- [2]	Compare two sequences directly	Align two sequences	Learn more ...
	Immunoglobulin sequences	Find matches to curated immunoglobulin sequences	igBLAST	Learn more ...
Nucleotide: 20 or 28 bp and above	UniVec	Screen for vector contamination	VecScreen	Learn more ...
	GEO	Find matches to sequences with MicroArray information	GEO BLAST	Learn more ...
	SNP	Find matches to human reference SNPs	SNP BLAST	Learn more ...
-	- [3]	To retrieve results for a search with its RID	Retrieve result for an RID	Learn more ...

Note:
[1] The query sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable better handling of short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, -F F -e 20000 -W 2 should be used. For nucleotide comparison, use -F F -e 2000 -W 7.
[2] "Align two sequences" treats the second sequence as the database.
[3] Requires valid RIDs that are assigned within the past 24 hours.

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI

Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

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*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

Old BLAST Web Pages to be deleted June 11th 2007

As previously announced access to the old pages will be removed on June 11, 2007.

2007-06-01 12:15:00

More BLAST news...

Tip of the Day

How to use BLAST to find human sequences in a database that can be amplified with a particular primer pair.

A frequent use of nucleotide-nucleotide BLAST is to check the specificity of oligonucleotides for hybridization in PCR. The goal is usually to make sure that the primers will give a unique product from the target genome or cDNA

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

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

231571

231571

Query subrange

From

To

Or, upload file Browse...

Job Title Q02067:Achaete-scute homolog 1 (Mash-1)

Enter a descriptive title for your BLAST search

Choose Search Set

Database Swissprot protein sequences(swissprot)

Organism Optional Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

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

Choose a BLAST algorithm

Let's look at some of the options!

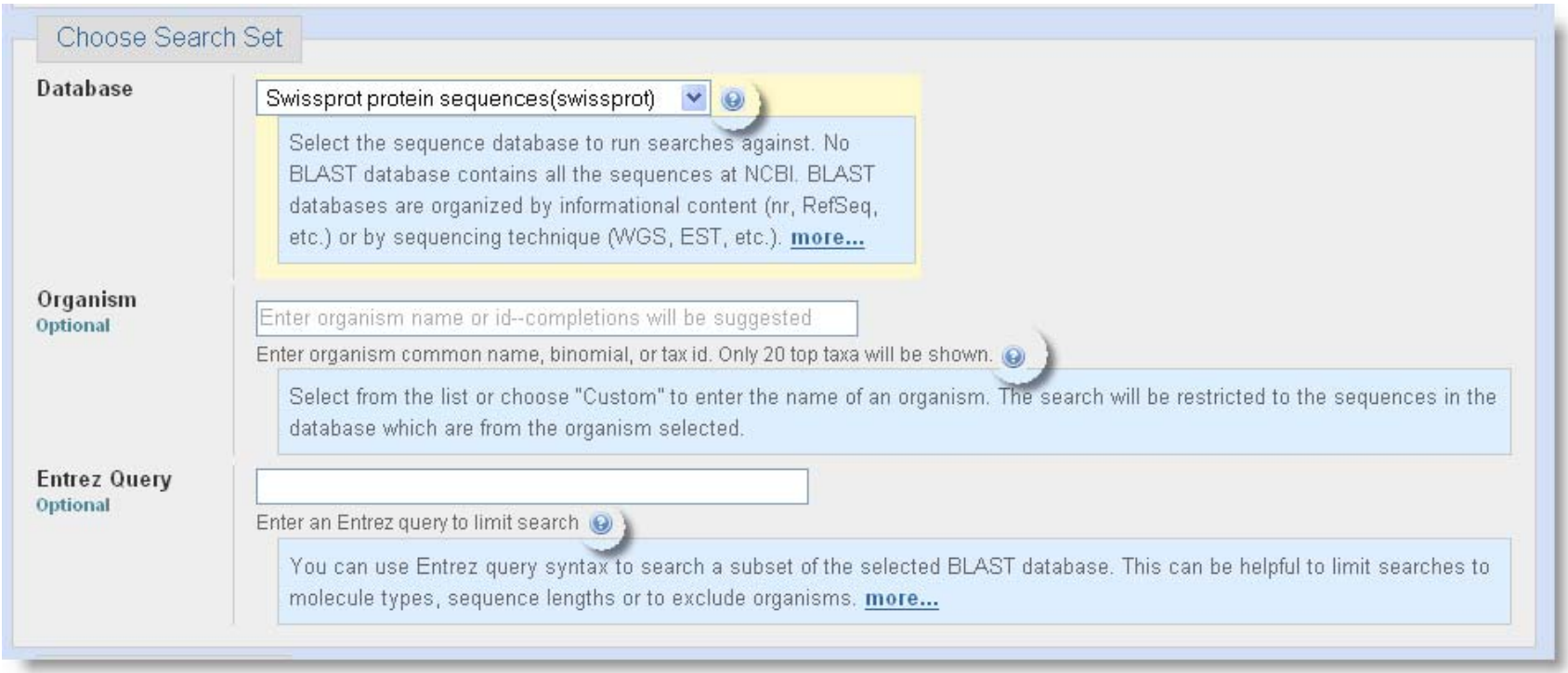
BLAST Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window

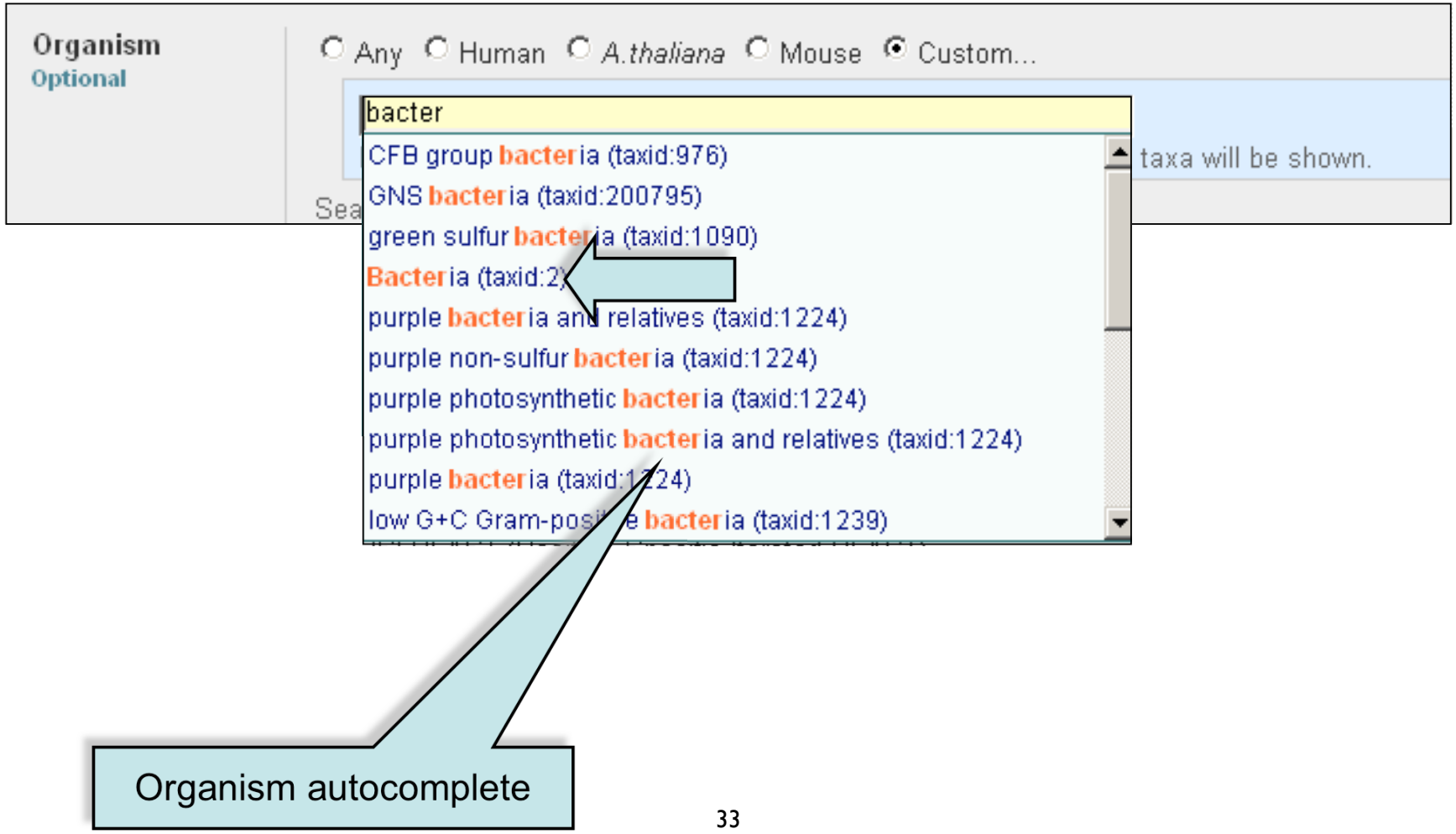
Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

Context Specific Help



Limiting Database: Organism



Limiting Database: Entrez Query

all[filter] NOT mammals[organism]

gene_in_mitochondrion[Properties]
2006:2007 [Modification Date]

Nucleotide
biomol_mrna[Properties]
biomol_genomic[Properties]

BLAST

Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the defau

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

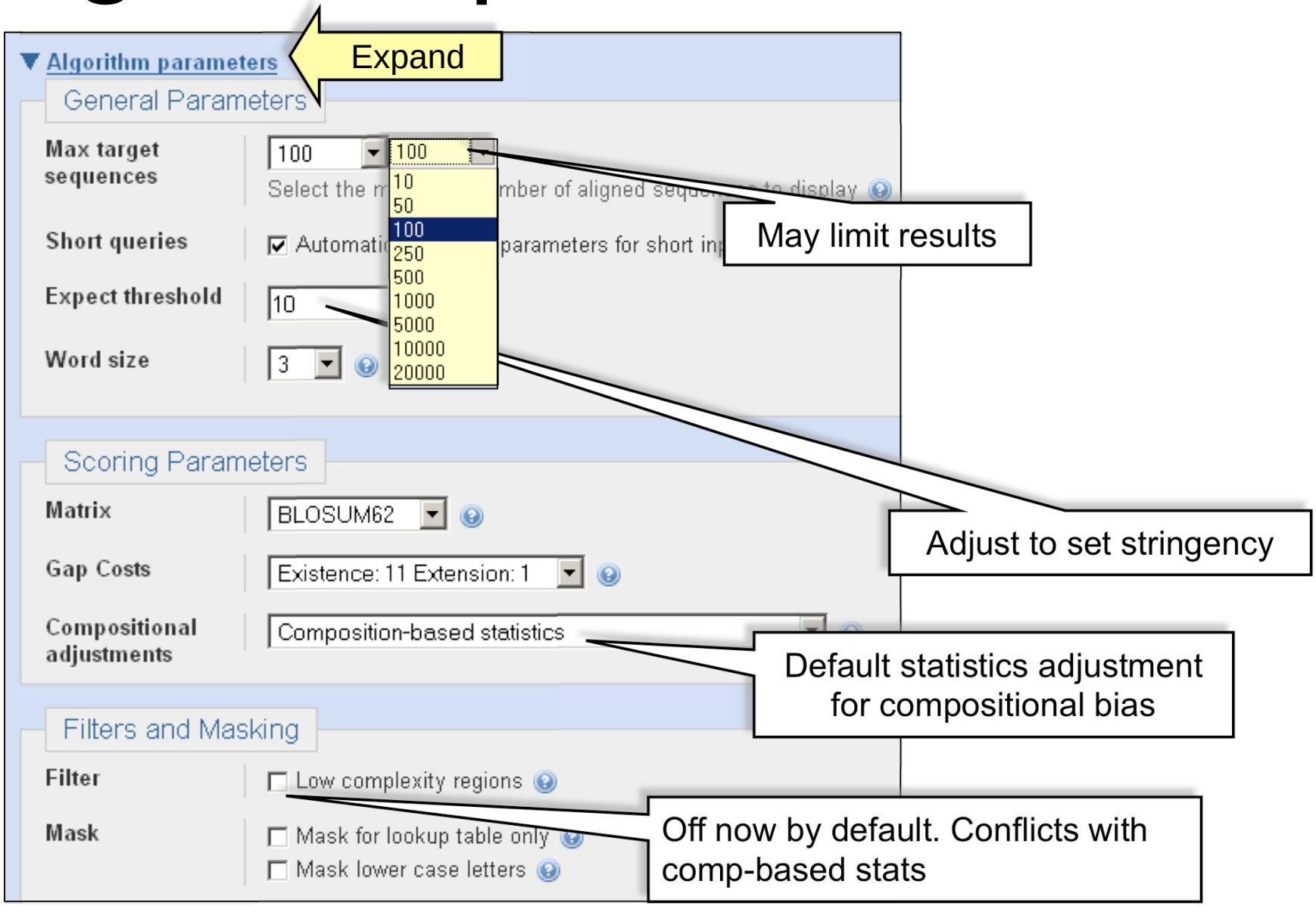
Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Algorithm parameters: Protein



Automatic Short Sequence Adjustment

Job Title: Elvis Lives!

No putative conserved domains have been detected

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

WAITING

Request ID	1WSB0FX012
Status	Searching

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

```
>ref|ZP_01712014.1|  conserved hypothetical protein [Pseudomonas put
Length=245

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1    ELVIS  5
            ELVIS
Sbjct  126  ELVIS  130


>ref|ZP_01712512.1|  Substrate-binding region of ABC-type glycine be
system [Pseudomonas putida GB-1]
Length=342

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1    ELVIS  5
            ELVIS
Sbjct  172  ELVIS  176


>ref|XP_001366374.1|  PREDICTED: similar to R7 binding protein [Mo
Length=257

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1   ELVIS  5
           ELVIS
Sbjct  69  ELVIS  73


>ref|ZP_01711731.1|  GCN5-related N-acetyltransferase [Caldivirga ma
Length=166

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1   ELVIS  5
           ELVIS
Sbjct  20  ELVIS  24
```

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

```
>gi|231571|sp|Q02067|ASCL1_MOUSE Achaete-scute homolog 1
(Mash-1)
MESSGKMESGAGQQPQPPQPFLPPAACFFATAAAAAAAAAAAAQSAQQQQPQAPPQQAPQLS
GGGHKSAAKQDKRQRSSSPELMRCKRRLNFSGFGYSLPQQQPAAVARRNERERNRVKLVNLG
PNGAANKKMSKVETLRSAVQYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGS
```

Query subrange

From

To

Or, upload file Browse...

Job Title MASH1 BLAST for CBW

Enter a descriptive title for your BLAST search

Choose Search Set

Database Swissprot protein sequences(swissprot)

Organism Optional

Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

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

Choose a BLAST algorithm

BLAST

Search database **swissprot** using **Blastp (protein-protein BLAST)**

☐ **Show results in a new window**

BLAST — Basic Local Alignment Search Tool

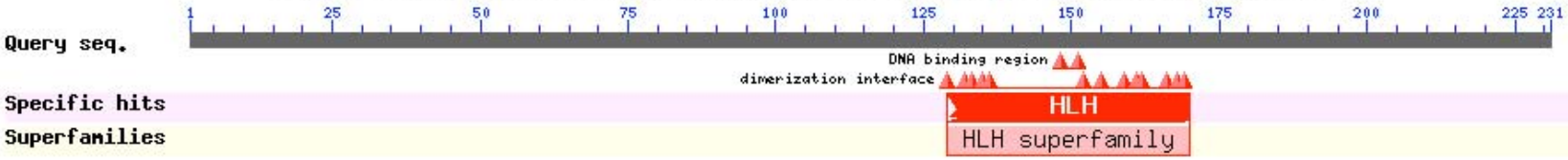
Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - T9U0ZFN4011 [Formatting options]

Job Title: Q02067:RecName: Full=Achaete-scute homolog...

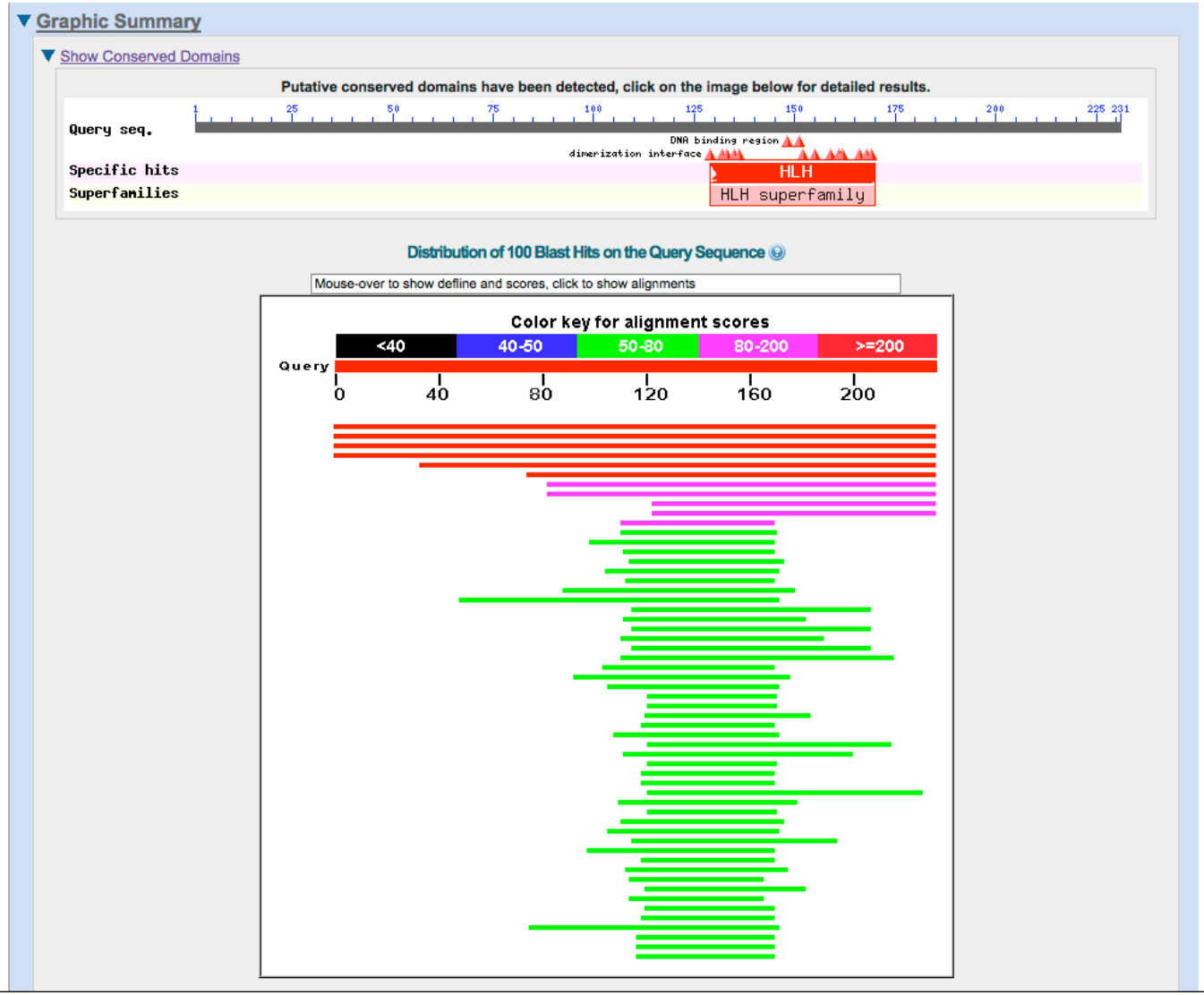
Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	**T9U0ZFN4011**
Status	Searching
Submitted at	Thu Feb 12 22:25:19 2009
Current time	Thu Feb 12 22:25:26 2009
Time since submission	00:00:06

This page will be automatically updated in **78** seconds

A graphical view



The BLAST hit list

BLAST Home

My NCBI [Sign In] [Register]

NCBI/ BLAST

Edit and

Q02067:

Des

Full=Mash-1

Program BLASTP 2.2.19+ ▸Citation

Molecule type amino acid
Query Length 231

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results]

▶ Graphic Summary

▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
sp|Q02067.1|ASCL1_MOUSE RecName: Full=Achaete-scute homolog 1...	466	4e-131	G
sp|P19359.1|ASCL1_RAT RecName: Full=Achaete-scute homolog 1	347	4e-95	G
sp|P50553.2|ASCL1_HUMAN RecName: Full=Achaete-scute homolog 1...	332	1e-90	G
sp|Q90259.1|ASL1A_DANRE RecName: Full=Achaete-scute homolog 1...	298	1e-80	G
sp|Q06234.1|ASCL1_XENLA RecName: Full=Achaete-scute homolog 1	289	9e-78	G
sp|Q90260.1|ASL1B_DANRE RecName: Full=Achaete-scute homolog 1...	217	3e-56	G
sp|Q2EGB9.1|ASCL2_BOVIN RecName: Full=Achaete-scute homolog 2...	135	1e-31	G
sp|Q99929.2|ASCL2_HUMAN RecName: Full=Achaete-scute homolog 2...	124	3e-28	G
sp|P19360.1|ASCL2_RAT RecName: Full=Achaete-scute homolog 2; ...	106	8e-23	G
sp|O35885.2|ASCL2_MOUSE RecName: Full=Achaete-scute homolog 2...	103	1e-21	G
sp|Q7RTU5.2|ASCL5_HUMAN RecName: Full=Achaete-scute homolog 5	80.5	6e-15	G
sp|Q6XD76.1|ASCL4_HUMAN RecName: Full=Achaete-scute homolog 4...	78.2	4e-14	G
sp|Q9NQ33.2|ASCL3_HUMAN RecName: Full=Achaete-scute homolog 3...	75.9	2e-13	G
sp|Q9JJR7.1|ASCL3_MOUSE RecName: Full=Achaete-scute homolog 3...	75.1	3e-13	G
sp|P10083.1|AST5_DROME RecName: Full=Achaete-scute complex pr...	74.7	3e-13	
sp|P10084.2|AST4_DROME RecName: Full=Achaete-scute complex pr...	71.6	3e-12	
sp|Q10007.1|HLH6_CAEEL RecName: Full=Helix-loop-helix protein 6	64.3	5e-10	G

BLAST Alignments

```
>sp|P20389|MYC2_MARMO  N-myc 2 proto-oncogene protein
Length=454

 Score = 35.8 bits (81),  Expect = 0.14, Method: Composition-based stats.
 Identities = 22/52 (42%), Positives = 30/52 (57%), Gaps = 4/52 (7%)

Query  133  FATLREHVPNGAANKKMSKVETLRSAVQYIRALQ----QLLDEHDAVSAAFQ  180
            F TLR+HVP    N+K +KV  L+ A +Y+  LQ    QLL E + + A  Q
Sbjct  391  FTTLRDHVPELVKNEKAAKVVILKKACEYVHYLQAKEHQLLMEKEKLQARQQ  442
```

Identical match

positive score (conservative)

Negative or zero

gap

BLAST Alignments

```
>sp|P04198|MYCN_HUMAN  N-myc proto-oncogene protein
Length=464

 Score = 35.4 bits (80),  Expect = 0.025, Method: Composition-based stats.
 Identities = 22/52 (42%), Positives = 31/52 (59%), Gaps = 4/52 (7%)

Query  133  FATLREHVPNGAANKKMSKVETLRSAVQYIRALQ----QLLDEHDAVSAAFQ  180
            F TLR+HVP    N+K +KV  L+ A +Y+ +LQ    QLL E + + A  Q
Sbjct  401  FLTLRDHVPELVKNEKAAKVVILKKATEYVHSLQAEEHQLLLEKEKLQARQQ  452


>sp|Q02363|ID2_HUMAN  DNA-binding protein inhibitor ID-2 (Inhibitor of DNA binding
2)
Length=134

 Score = 35.4 bits (80),  Expect = 0.025, Method: Composition-based stats.
 Identities = 19/47 (40%), Positives = 29/47 (61%), Gaps = 0/47 (0%)

Query  129  VNLGFATLREHVPNGAANKKMSKVETLRSAVQYIRALQQLLDEHDAV  175
            +N  ++ L+E VP+   NKK+SK+E L+  + YI  LQ  LD H  +
Sbjct  39   MNDCYSKLKELVPSIPQNKKVSKMEILQHVIDYILDLQIALDSHPTI  85


>sp|P12980|LYL1_HUMAN  Protein lyl-1 (Lymphoblastic leukemia-derived sequence 1)
Length=267

 Score = 35.4 bits (80),  Expect = 0.025, Method: Composition-based stats.
 Identities = 22/50 (44%), Positives = 31/50 (62%), Gaps = 0/50 (0%)

Query  129  VNLGFATLREHVPNGAANKKMSKVETLRSAVQYIRALQQLLDEHDAVSAA  178
            VN  FA LR+ +P    ++K+SK E LR A++YI  L +LL +  A  AA
Sbjct  153  VNGAFAELRKLLPTHPPDRKLSKNEVLRLAMKYIGFLVRLLRDQAAALAA  202
```

- ## Similarity

  The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score.

- ## Identity

  The extent to which two (nucleotide or amino acid) sequences are invariant.

- ## Homology

  Similarity attributed to descent from a common ancestor.

  It is your responsibility as an informed bioinformatician to use these terms correctly: A sequence is either homologous or not. Don't use % with this term!

Re-Format and/or Download your BLAST results

Edit and Resubmit | Save Search Strategies | Formatting options | Download

Formatting options — Reformat

Show Alignment as HTML ☐ Advanced View ☐ Use old BLAST report format — Reset form to defaults

Alignment View Pairwise

Display ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☐ NCBI-gi

Masking Character: Lower Case Masking Color: Grey

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

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

Entrez query:

Expect Min: Expect Max:

Format for ☐ PSI-BLAST with inclusion threshold:

Download

Alignment					Search Strategies	Bioseq
Text	XML	ASN.1	Hit Table(text)	Hit Table(csv)	ASN.1	ASN.1

Sorting BLAST by Taxonomy

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Regis

NCBI/ BLAST/ blastp suite/ Formatting Results - T9U0ZFN4011

Edit and Resubmit Save Search Strategies ▷Formatting options ▷Download

Q02067:RecName: Full=Achaete-scute homolog...

Query ID gi|231571|sp|Q02067.1|ASCL1_MOUSE
Description RecName: Full=Achaete-scute homolog 1; AltName: Full=Mash-1
Molecule type amino acid
Query Length 231

Database Name swissprot
Description Non-redundant SwissProt sequences
Program BLASTP 2.2.19+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results]

▶ Graphic Summary

▼ Descriptions

Sequences producing significant alignments:		Score (Bits)	E Value	
sp|Q02067.1|ASCL1_MOUSE	RecName: Full=Achaete-scute homolog 1...	466	4e-131	G
sp|P19359.1|ASCL1_RAT	RecName: Full=Achaete-scute homolog 1	347	4e-95	G
sp|P50553.2|ASCL1_HUMAN	RecName: Full=Achaete-scute homolog 1...	332	1e-90	G
sp|Q90259.1|ASL1A_DANRE	RecName: Full=Achaete-scute homolog 1...	298	1e-80	G
sp|Q06234.1|ASCL1_XENLA	RecName: Full=Achaete-scute homolog 1	289	9e-78	G

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp/ Formatting Results - VVH6PBD3011 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

Job Title: gi|231571 (231 letters)

Show Conserved Domains

Tax BLAST Report

Index

- Lineage Report
- Organism Report
- Taxonomy Report
- Help

Lineage Report

```
Bilateria          [animals]
. Coelomata          [animals]
. . Euteleostomi       [vertebrates]
. . . Tetrapoda          [vertebrates]
. . . . Amniota            [vertebrates]
. . . . . Eutheria           [placentals]
. . . . . . Euarchontoglires [placentals]
. . . . . . . Glires           [placentals]
. . . . . . . . Muroidea         [rodents]
. . . . . . . . . Murinae          [rodents]
. . . . . . . . . Mus musculus (mouse) ----------------  466 22 hits [rodents]              Achaete-scute homolog 1 (Mash-1)
. . . . . . . . . Rattus norvegicus (brown rat) .......  347 10 hits [rodents]              Achaete-scute homolog 1
. . . . . . . . . Mesocricetus auratus (Syrian hamster) -   50  2 hits [rodents]              Neurogenic differentiation factor 1 (NeuroD1)
. . . . . . . . . Oryctolagus cuniculus (domestic rabbit) -  49  1 hit  [rabbits & hares]      Heart- and neural crest derivatives-expressed
. . . . . . . . Homo sapiens (man) ------------------------  332 25 hits [primates]            Achaete-scute homolog 1 (HASH1)
. . . . . . . . Macaca fascicularis (cynomolgus monkey) ...   48  1 hit  [primates]            Neurogenic differentiation factor 6 (NeuroD6)
. . . . . . Bos taurus (cow) ----------------------------  135  4 hits [even-toed ungulates]  Achaete-scute homolog 2 (Mash2)
. . . . . . Ovis aries (domestic sheep) ................   50  1 hit  [even-toed ungulates]  Heart- and neural crest derivatives-expressed
. . . . . Gallus gallus (bantam) ------------------------   60  8 hits [birds]               Heart- and neural crest derivatives-expressed
. . . . . Coturnix japonica ............................   50  1 hit  [birds]               Myogenic factor 5 (Myf-5) (Myogenic factor 3)
. . . . Xenopus laevis (common platanna) ----------------  289 10 hits [frogs & toads]        Achaete-scute homolog 1
. . . . Notophthalmus viridescens (red-spotted newt) ....   49  1 hit  [salamanders]           Myogenic factor 5 (Myf-5)
. . . Danio rerio (leopard danio) ------------------------  298  8 hits [bony fishes]          Achaete-scute homolog 1a (Zash-1a) (Pituitary
. . Drosophila melanogaster ------------------------------   74  5 hits [flies]               Achaete-scute complex protein T5 (Achaete)
. Caenorhabditis elegans (nematode) ----------------------   64  4 hits [nematodes]           Helix-loop-helix protein 6
```

Organism Report

```
 Mus musculus (mouse) [rodents] taxid 10090
sp|Q02067|ASCL1_MOUSE Achaete-scute homolog 1 (Mash-1)          466  4e-131
sp|O35885|ASCL2_MOUSE Achaete-scute homolog 2 (Mash-2)          103  9e-22
sp|Q9JJR7|ASCL3_MOUSE Achaete-scute homolog 3 (bHLH transc...    75  2e-13
sp|Q61039|HAND2_MOUSE Heart- and neural crest derivatives-...   60  7e-09
sp|P27792|LYL1_MOUSE Protein lyl-1 (Lymphoblastic leukemia...    53  8e-07
```

Distance Tree of Results

Tree view for rid: T9U0ZFN4011, query ID: sp|Q02067.1, database: swissprot

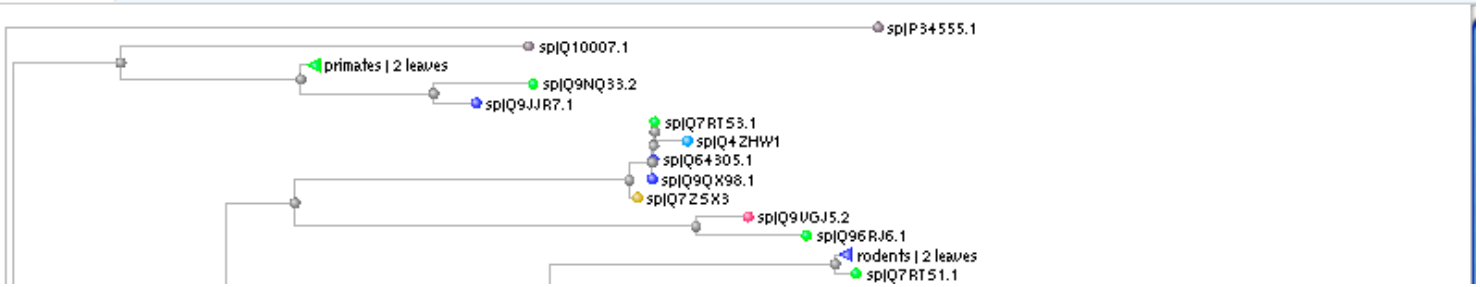
This tree was produced using BLAST pairwise alignments. more...

BLAST computes a pairwise alignment between a query and the database sequences searched. It does not explicitly compute an alignment between the different database sequences (i.e., does not perform a multiple alignment). For purposes of this sequence tree presentation an implicit alignment between the database sequences is constructed, based upon the alignment of those (database) sequences to the query. It may often occur that two database sequences align to different parts of the query, so that they barely overlap each other or do not overlap at all. In that case it is not possible to calculate a distance between these two sequences and only the higher scoring sequence is included in the tree.

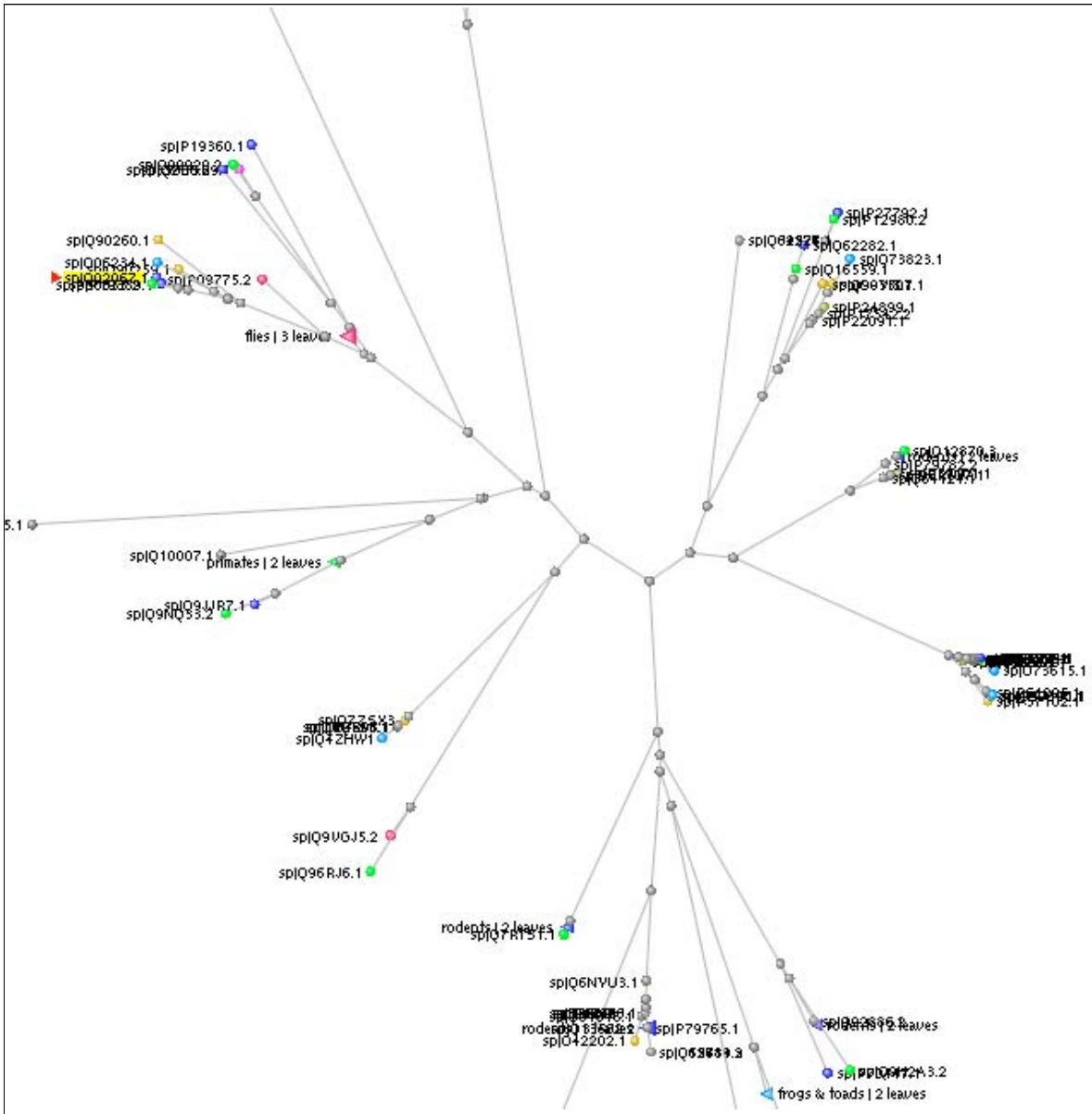
Tree method: Fast Minimum Evolution | Max Seq Difference: 0.85 | Distance: Grishin (protein) | Reset | Download

Tree Method:
Algorithm used to produce a tree from given distances (or dissimilarities) between sequences. Available options:
1) Fast Minimum Evolution (*Desper R and Gascuel O, Mol Biol Evol 21:587-98, 2004*)
2) Neighbor Joining (*Saitou N and Nei M, Mol Biol Evol, 4:406-25, 2004*)
Note: Both algorithms produce un-rooted tree such as ones shown as *radial* or *force* in the tabs below. The rooted trees are created by placing a root in the middle of the longest edge.

rectangle | slanted | radial | force | Show distance | Mouse over an internal node for a subtree or alignment



read more in context specific help menus



- ✓ Rectangle: rectangular shaped rooted tree, where root is places in the longest edge
- ✓ Slanted: similar to rectangle, but with triangular tree shape
- ✓ Radial: un-rooted tree
- ✓ Force: similar to radial, where nodes are pushed away from one another for better presentation.

Nucleotide BLAST

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI

Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

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*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

New Human and Mouse pre-indexed databases

Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.

2007-09-04 10:55:00

More BLAST news...

Tip of the Day

Using Genomic BLAST

Genomic BLAST pages are helpful because they allow the genomic context of a BLAST search to be displayed in the Map Viewer. For example, discontiguous (cross-species) MegaBLAST against the human RefSeq transcript for albumin (NM_000477) can be used to identify the homolog in the rat genome.

More tips...

nt BLAST: New Output

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 FAST AB168636 Clear

```
>Crab eating macaque CDC20 mRNA
AGCGGAGAGTTTAAGAGGCGTAAGCGAGGCGTGTTAAACCCGGTCGGAACTGCAACTTGCTC
ACGGGCTCCGCAGGCACCAACTGCAAGGACCCCTCCCGCTGCGGGCGTTCCCATGGCACAAT
GAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCCAATGCACCCCCTGCGCG
GCAAAGCCAAGGAAGCCTCAGGCCCGGCCCCCTCACCCATGCGGGCCGCCAACCGATCCCAC
```

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

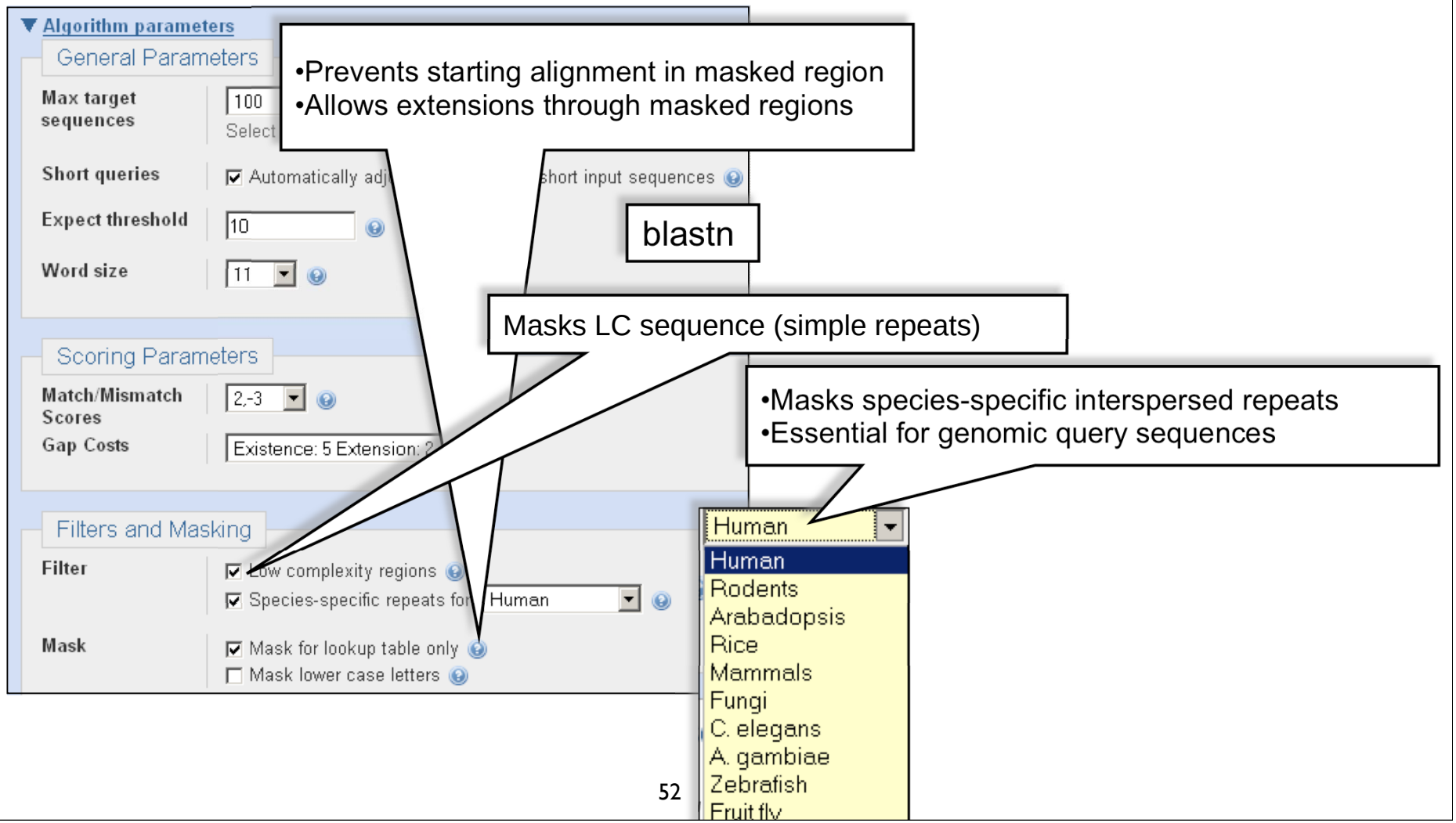
Database ◉ Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript

Entrez Query
Optional

Enter an Entrez query to limit search

Algorithm parameters: Nucleotide



Sortable Results

Legend for links to other resources: U UniGene E GEO G ene S Structure

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%	

Separate Sections for Transcript and Genome

Pseudogene on Chromosome 9

Functional Gene on Chromosome 1

Total Score: All Segments

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%	

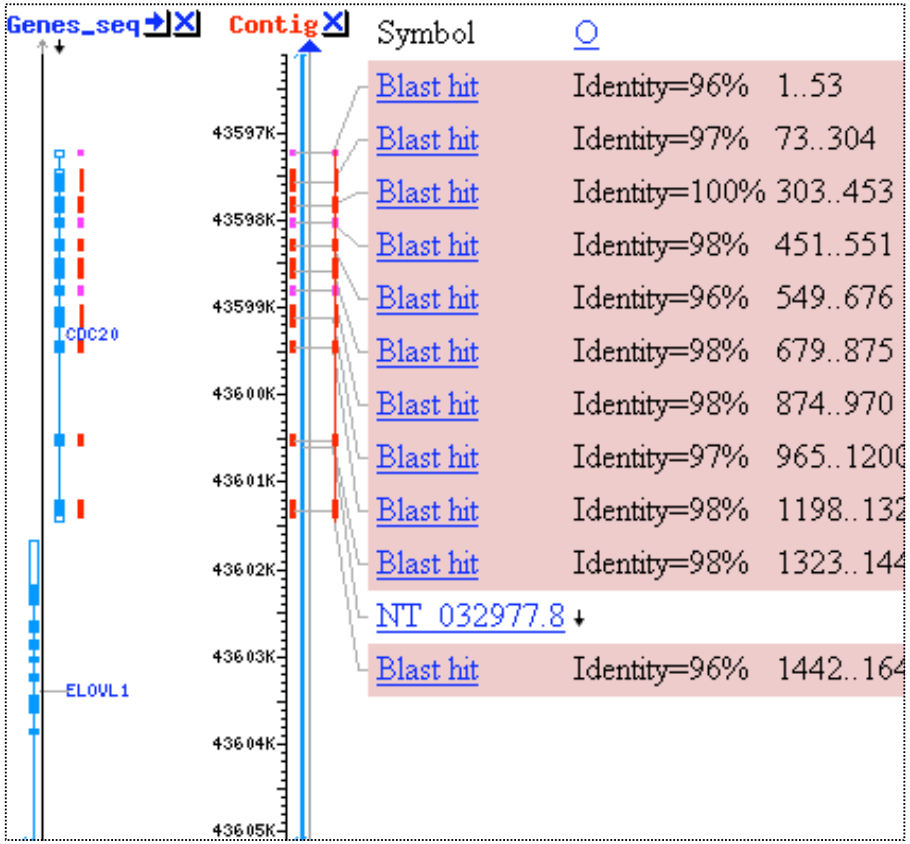
Functional Gene Now First

Sorting in Exon Order

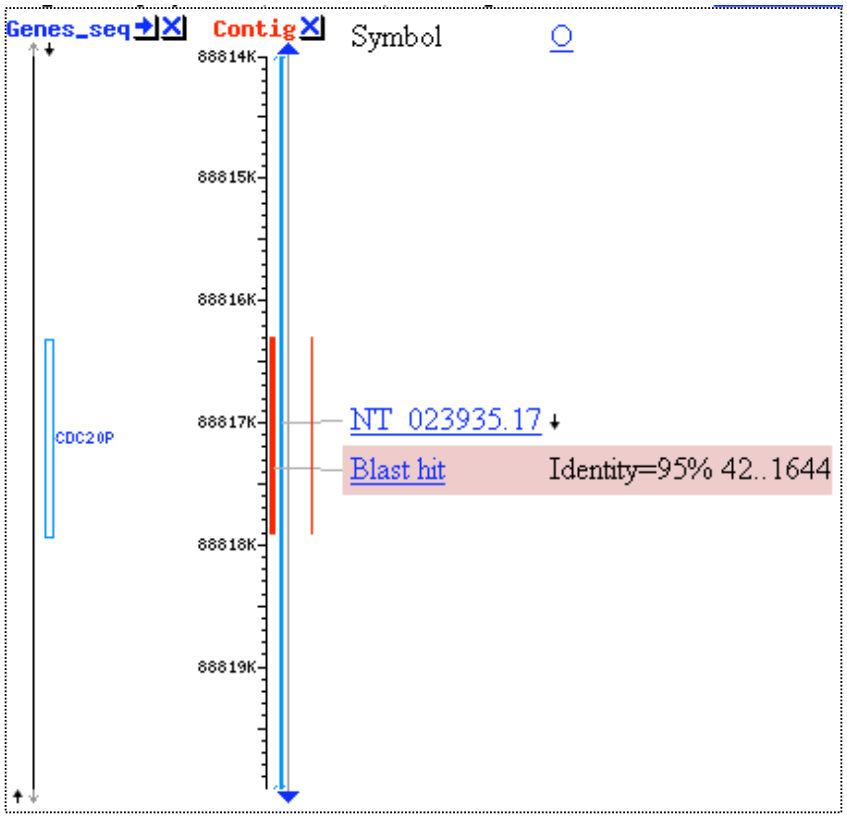
```
> ref|NT_032977.8|Hs1_33153 D Homo sapiens chromosome 1 genomic contig, reference assembly
Length=73835825

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

 Features i
   cell div

 Score =  4
 Identities
 Strand=Plu

Query  965
Sbjct  1379

Query  1025
Sbjct  1379

Defau
Lon

 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

 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
```

Query start position
Exon order

Links to Map Viewer

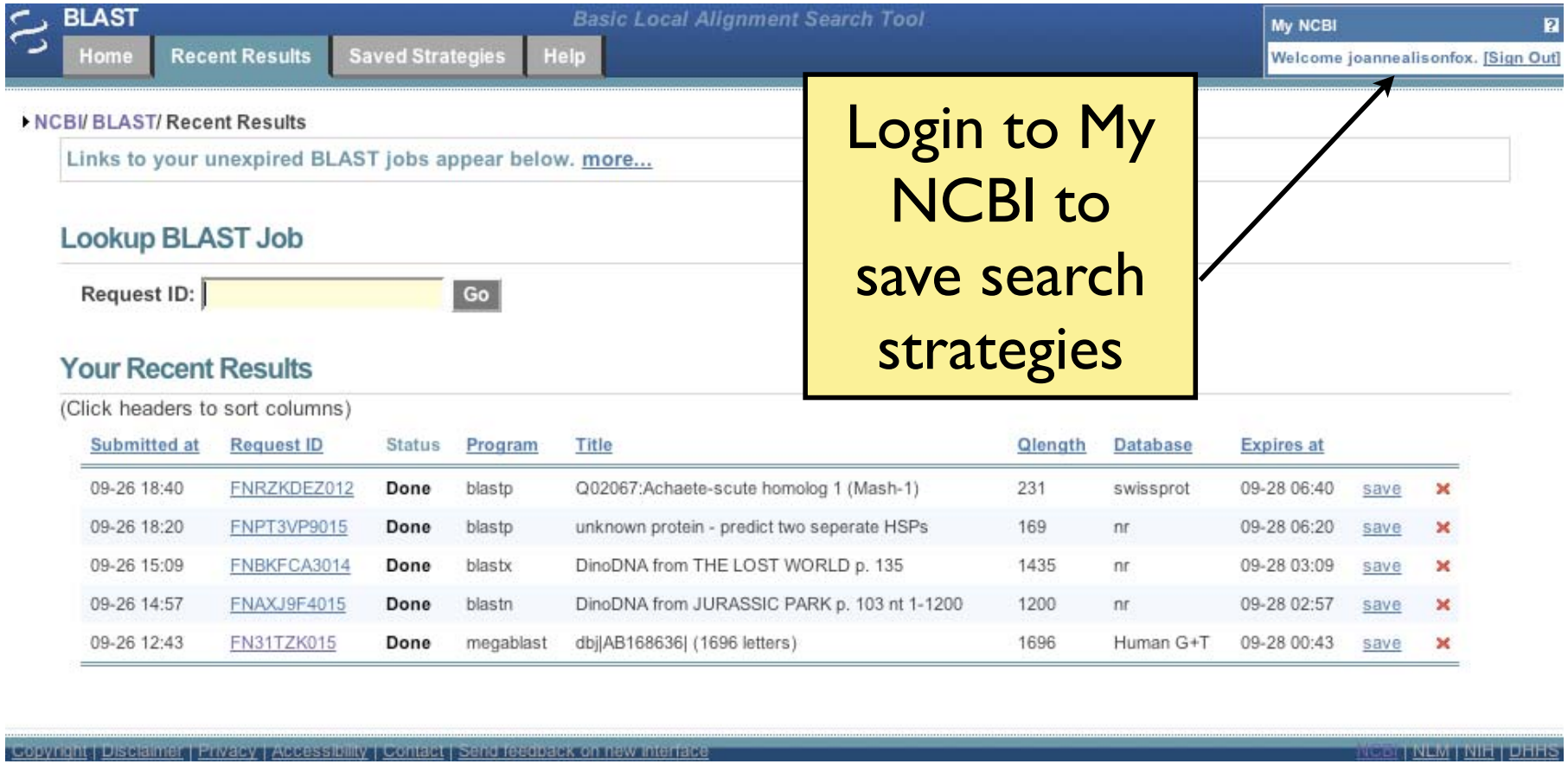


Chromosome 1



Chromosome 9

Recent and Saved Strategies



Login to My NCBI to save search strategies

Genomic and Specialized BLAST pages

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*

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 sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay

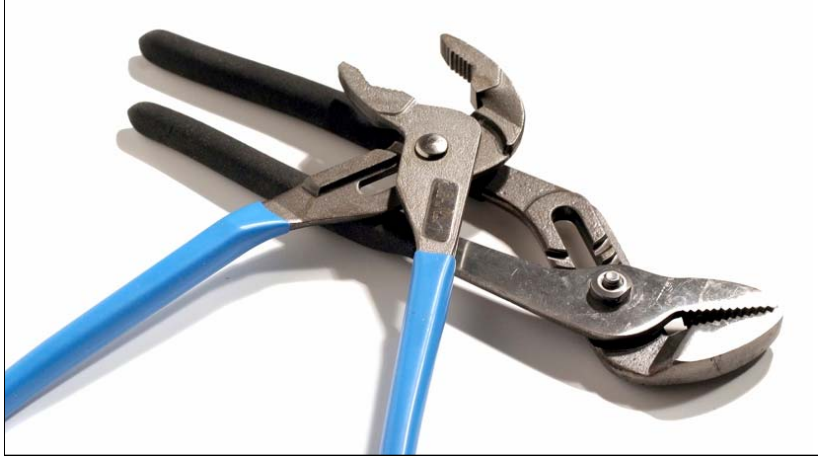
Service Addresses

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

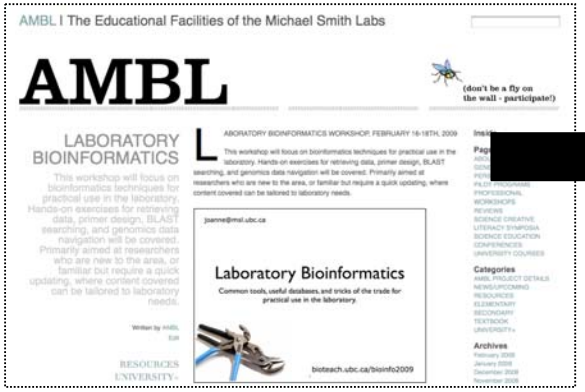
Telephone support: 301- 496- 2475

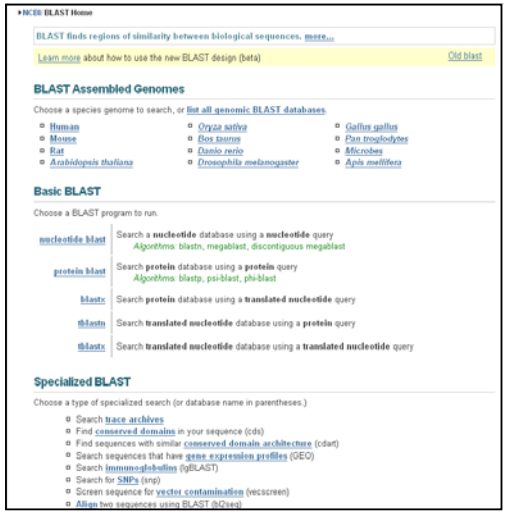
BLAST

PRACTICAL EXERCISE: The Jurassic Park Detective Story



navigate to:
bioteach.ubc.ca/bioinfo2009





Let's compare our results

Get the sequences from the webpage and carry out BLAST searches

Can you identify the Dinosaur sequences?

Search #1:
Jurassic Park sequence

use blastn

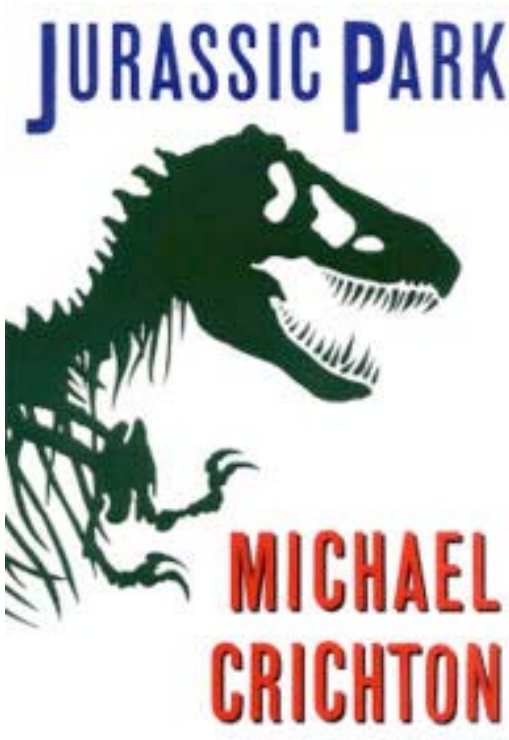
Search #2:
The Lost World sequence

use blastx

Try some BLAST searches with your own sequence of interest…

Explore what happens when you change advanced parameters…

Search #1 - blastn against nr

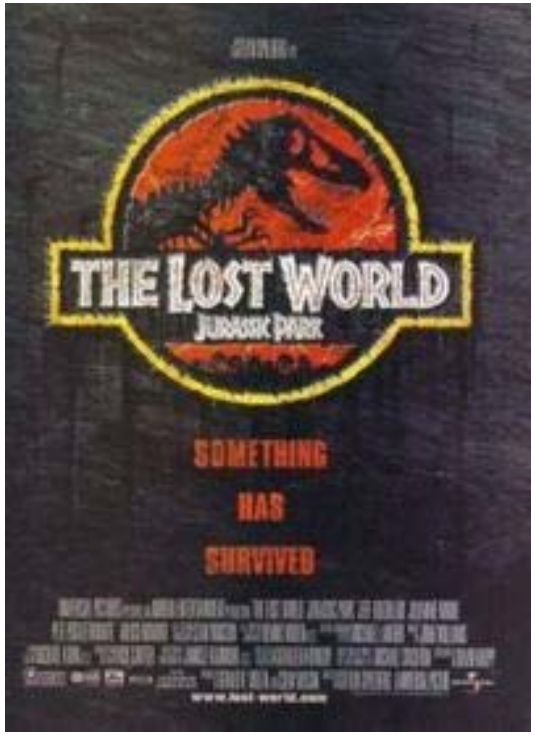


- Most common use of blastn
  - ✓ Sequence identification
  - ✓ Establish whether an exact match for a sequence is already present in the database

```
>gi|157064989|gb|EU118176.1|  Cloning vector pCM433, complete sequence
Length=8081

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

 Score =  437 bits (484),  Expect = 4e-119
 Identities = 297/340 (87%), Gaps = 40/340 (11%)
 Strand=Plus/Plus

Query  1     GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7309  GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGC  7368

Query  61    ----------GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGA  110
                       ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7369  TCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGA  7428

Query  111   AGCTCCCTCG----------TGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTT  160
             ||||||||||          ||||||||||||||||||||||||||||||||||||||||
Sbjct  7429  AGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTT  7488

Query  161   CTCCCTTCGGGAAGCGTGGC----------TGCTCACGCTGTACCTATCTCAGTTCGGTG  210
             ||||||||||||||||||||           ||||||||||||  ||||||||||||||||
Sbjct  7489  CTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTG  7548

Query  211   TAGGTCGTTCGCTCCAAGCTGGGCTGTGTG----------CCGTTCAGCCCGACCGCTGC  260
             ||||||||||||||||||||||||||||||          ||||||||||||||||||||
Sbjct  7549  TAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGC  7608

Query  261   GCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA  300
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  7609  GCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA  7648


 Score =  536 bits (594),  Expect = 6e-149
 Identities = 360/410 (87%), Gaps = 50/410 (12%)
 Strand=Plus/Plus

Query  302   GTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAG-  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3591  GTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGC  3650

Query  361   ---------ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCC  411
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3651  GCGACGATGATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCC  3710

Query  412   TTCGTCACT----------CCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATG  461
             |||||||||          |||||||||||||||||||||||||||||||||||||||||
Sbjct  3711  TTCGTCACTGGTCCCGCCACCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATG  3770

Query  462   GCGGCCGACGCGCTGGGCT----------GGCGTTCGCGACGCGAGGCTGGATGGCCTTC  511
             |||||||||||||||||||          ||||||||||||||||||||||||||||||||
Sbjct  3771  GCGGCCGACGCGCTGGGCTACGTCTTGCTGGCGTTCGCGACGCGAGGCTGGATGGCCTTC  3830

Query  512   CCCATTATGATTCTTCTCGCTTCCGGCG----------GCCCGCGTTGCAGGCCATGCTG  561
             ||||||||||||||||||||||||||||          ||||||||||||||||||||||
Sbjct  3831  CCCATTATGATTCTTCTCGCTTCCGGCGGCATCGGGATGCCCGCGTTGCAGGCCATGCTG  3890

Query  562   TCCAGGCAGGTAGATGACGACCATCAGGGACAGCTTCAA----------CGGCTCTTACC  611
             ||||||||||||||||||||||||||||||||||||||||          ||||||||||||
Sbjct  3891  TCCAGGCAGGTAGATGACGACCATCAGGGACAGCTTCAAGGATCGCTCGCGGCTCTTACC  3950

Query  612   AGCCTAACTTCGATCACTGGACCGCTGATCGTCACGGCGATTTATGCCGC  661
             |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3951  AGCCTAACTTCGATCACTGGACCGCTGATCGTCACGGCGATTTATGCCGC  4000
```

Search #2 - blastx against nr

- Translating BLAST programs (blastx, tblastn, tblastx)
  - ✓ Look for similar proteins
  - ✓ Identify potential homologs in other species



```
>gi|45382623|ref|NP_990795.1|  erythroid-specific transcription factor eryf1 [Gallus gallus]
 gi|120955|sp|P17678|GATA1_CHICK  Erythroid transcription factor (GATA-binding factor 1) (GATA-1)
(Eryf1) (NF-E1 DNA-binding protein) (NF-E1A)
 gi|212629|gb|AAA49055.1|  Eryf1 protein
Length=304

 Score =  366 bits (940),  Expect = 2e-99
 Identities = 304/318 (95%), Positives = 304/318 (95%), Gaps = 14/318 (4%)
 Frame = +1

Query  121   MEFVALGGPDAGSPTPFPDeagaflglgggerteaggllaSYPPSGRVSLVPWADTGTLG  300
             MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG
Sbjct  1     MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG  60

Query  301   TPQWVPPATQMEPPHYLEllqpprgspphpssgpllplssgpppCEARECVMARKNCGAT  480
             TPQWVPPATQMEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECV    NCGAT
Sbjct  61    TPQWVPPATQMEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECV----NCGAT  116

Query  481   ATPLWRRDGTGHYLCNWASACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCSHERENCQT  660
             ATPLWRRDGTGHYLCN   ACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCS    NCQT
Sbjct  117   ATPLWRRDGTGHYLCN---ACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCS----NCQT  169

Query  661   STTTLWRRSPMGDPVCNNIHACGLYYKLHQVNRPLTMRKDGIQTRNRKVsskgkkrrppg  840
             STTTLWRRSPMGDPVCN   ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPG
Sbjct  170   STTTLWRRSPMGDPVCN---ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPG  226

Query  841   ggnpsatagggapmggggdpsmpppppppaaappQSDALYALGPVVLSGHFLPfgnsggf  1020
             GGNPSATAGGGAPMGGGGDPSMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGF
Sbjct  227   GGNPSATAGGGAPMGGGGDPSMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGF  286

Query  1021  fgggaggYTAPPGLSPQI  1074
             FGGGAGGYTAPPGLSPQI
Sbjct  287   FGGGAGGYTAPPGLSPQI  304
```

Mark was here, NIH

BLAST

COMMON TASKS - Basic Search; Searching Sets of Sequences (multiple inputs; small custom databases); Primer Design

Research article

Open Access

A salmonid EST genomic study: genes, duplications, phylogeny and microarrays

Ben F Koop*[1,6], Kristian R von Schalburg[1], Jong Leong[1], Neil Walker[1], Ryan Lieph[1], Glenn A Cooper[1], Adrienne Robb[1], Marianne Beetz-Sargent[1], Robert A Holt[2], Richard Moore[2], Sonal Brahmbhatt[3], Jamie Rosner[3], Caird E Rexroad III[4], Colin R McGowan[5] and William S Davidson[5]

Address: [1]Centre for Biomedical Research, University of Victoria, Victoria, British Columbia, V8W 3N5, Canada, [2]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, V5Z 4S6, Canada, [3]Prostate Centre, Vancouver, British Columbia, V6H 3Z6, Canada, [4]ARS, USDA, Natl Ctr Cool & Cold Water Aquaculture, Kearneysville, WV 25430, USA, [5]Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada and [6]Department of Biology, University of Victoria, P.O. Box 3020, Victoria, British Columbia, V8W 3N5, Canada

Email: Ben F Koop* - bkoop@uvic.ca; Kristian R von Schalburg - krvs@uvic.ca; Jong Leong - jong@uvic.ca; Neil Walker - nwalker@uvic.ca; Ryan Lieph - handsomryan@gmail.com; Glenn A Cooper - gac@uvic.ca; Adrienne Robb - arobb@uvic.ca; Marianne Beetz-Sargent - marianbs@uvic.ca; Robert A Holt - rholt@bcgsc.ca; Richard Moore - rmoore@bcgsc.ca; Sonal Brahmbhatt - Sonal.Brahmbhatt@vch.ca; Jamie Rosner - Jamie.Rosner@vch.ca; Caird E Rexroad - caird.rexroadIII@ARS.USDA.GOV; Colin R McGowan - cmcgowan@icywaters.com; William S Davidson - wdavidso@sfu.ca

* Corresponding author

Published: 17 November 2008

BMC Genomics 2008, **9**:545 doi:10.1186/1471-2164-9-545

Received: 13 June 2008
Accepted: 17 November 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/545

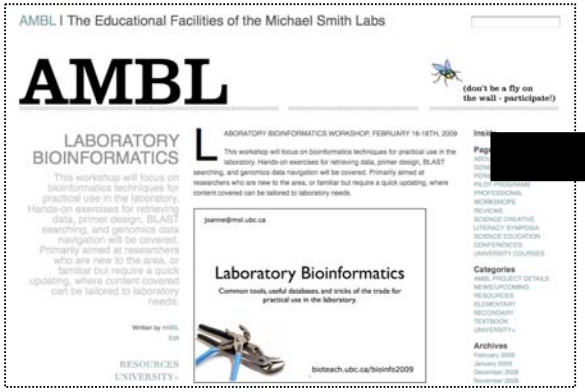
Abstract

Background: Salmonids are of interest because of their relatively recent genome duplication, and their extensive use in wild fisheries and aquaculture. A comprehensive gene list and a comparison of genes in some of the different species provide valuable genomic information for one of the most widely studied groups of fish.

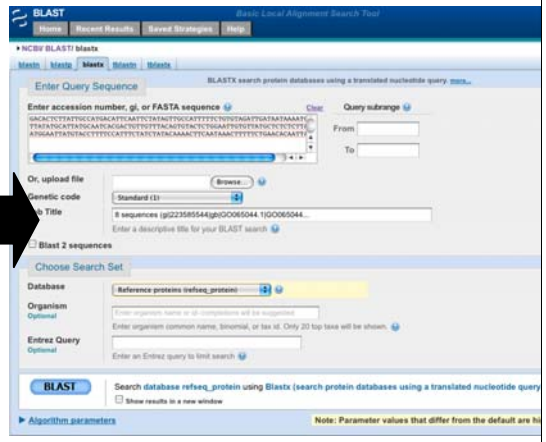
Results: 298,304 expressed sequence tags (ESTs) from Atlantic salmon (69% of the total), 11,664 chinook, 10,813 sockeye, 10,051 brook trout, 10,975 grayling, 8,630 lake whitefish, and 3,624 northern pike ESTs were obtained in this study and have been deposited into the public databases. Contigs were built and putative full-length Atlantic salmon clones have been identified. A database containing ESTs, assemblies, consensus sequences, open reading frames, gene predictions and putative annotation is available. The overall similarity between Atlantic salmon ESTs and those of rainbow trout, chinook, sockeye, brook trout, grayling, lake whitefish, northern pike and rainbow smelt is 93.4, 94.2, 94.6, 94.4, 92.5, 91.7, 89.6, and 86.2% respectively. An analysis of 78 transcript sets show *Salmo* as a sister group to *Oncorhynchus* and *Salvelinus* within Salmoninae, and Thymallinae as a sister group to Salmoninae and Coregoninae within Salmonidae. Extensive gene duplication is consistent with a genome duplication in the common ancestor of salmonids. Using all of the available EST data, a new expanded salmonid cDNA microarray of 32,000 features was created. Cross-species hybridizations to this cDNA microarray indicate that this resource will be useful for studies of all 68 salmonid species.

Conclusion: An extensive collection and analysis of salmonid RNA putative transcripts indicate that Pacific salmon, Atlantic salmon and charr are 94–96% similar while the more distant whitefish, grayling, pike and smelt are 93, 92, 89 and 86% similar to salmon. The salmonid transcriptome reveals a complex history of gene duplication that is consistent with an ancestral salmonid genome duplication hypothesis. Genome resources, including a new 32 K microarray, provide valuable new tools to study salmonids.

navigate to:
bioteach.ubc.ca/bioinfo2009



We'll walk through this example together



Get the Salmon sequences and carry out the BLAST searches

Can you identify the ESTs?

Search #1: Use multiple EST sequences as input query

use blastx

Is the hbaa1 gene present?

Search #2: Use the hbaa1 sequence as input, search against Salmon EST custom database

use blast2seq option with tblastn

Searching with Multiple Sequences as Input

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI — Welcome joannealisonfox. [Sign Ou

NCBI/ BLAST/ blastx

blastn | blastp | blastx | tblastn | tblastx

BLASTX search protein databases using a translated 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...

Genetic code — Standard (1)

Job Title — 8 sequences (gi|223585544|gb|GO065044.1|GO065044...

Enter a descriptive title for your BLAST search

Blast 2 sequences

Choose Search Set

Database — Reference proteins (refseq_protein)

Organism (Optional) — Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional)

Enter an Entrez query to limit search

BLAST — Search database refseq_protein using Blastx (search protein databases using a translated nucleotide query)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

11 sequences (gi|223585544|gb|GO065044.1|GO065044...

Results for: 7:lcl|5773 gi|223585538|gb|GO065038.1|GO065038 EST_ssal_rgh_1084502 ssalrgh mixed_tissue full-length Salmo sal...(614bp)

What's this?

Query ID
Description

1:lcl|5767 gi|223585544|gb|GO065044.1|GO065044 EST_ssal_rgh_1084509 ssalrgh mixed_tissue full-length Salmo sal...(725bp)
2:lcl|5768 gi|223585543|gb|GO065043.1|GO065043 EST_ssal_rgh_1079901 ssalrgh mixed_tissue full-length Salmo sal...(897bp)
3:lcl|5769 gi|223585542|gb|GO065042.1|GO065042 EST_ssal_rgh_1079900 ssalrgh mixed_tissue full-length Salmo sal...(266bp)
*4:lcl|5770 gi|223585541|gb|GO065041.1|GO065041 EST_ssal_rgh_1084506 ssalrgh mixed_tissue full-length Salmo sal...(290bp)
*5:lcl|5771 gi|223585540|gb|GO065040.1|GO065040 EST_ssal_rgh_1079898 ssalrgh mixed_tissue full-length Salmo sal...(310bp)
6:lcl|5772 gi|223585539|gb|GO065039.1|GO065039 EST_ssal_rgh_1084505 ssalrgh mixed_tissue full-length Salmo sal...(432bp)
7:lcl|5773 gi|223585538|gb|GO065038.1|GO065038 EST_ssal_rgh_1084502 ssalrgh mixed_tissue full-length Salmo sal...(614bp)
8:lcl|5774 gi|223585537|gb|GO065037.1|GO065037 EST_ssal_rgh_1079894 ssalrgh mixed_tissue full-length Salmo sal...(629bp)
9:lcl|5775 gi|223585536|gb|GO065036.1|GO065036 EST_ssal_rgh_1079893 ssalrgh mixed_tissue full-length Salmo sal...(884bp)
10:lcl|5776 gi|223585535|gb|GO065035.1|GO065035 EST_ssal_rgh_1084500 ssalrgh mixed_tissue full-length Salmo sal...(821bp)
11:lcl|5777 gi|223585534|gb|GO065034.1|GO065034 EST_ssal_rgh_1079892 ssalrgh mixed_tissue full-length Salmo sal...(791bp)

Molecule type
Query Length

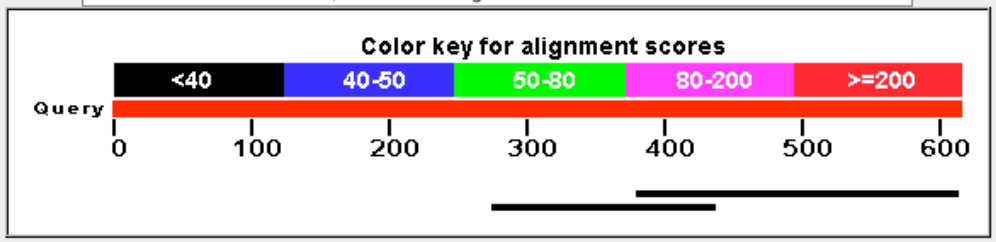
Other reports: ▷Search Summary | Taxonomy Reports

Results for: pull down list

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

Alignments

Select All Get selected sequences

```
>ref|YP_934206.1|  hypothetical protein azo2703 [Azoarcus sp. BH72]
Length=774

 GENE ID: 4607585 azo2703 | hypothetical protein [Azoarcus sp. BH72]
(10 or fewer PubMed links)

 Score = 35.0 bits (79),  Expect = 4.3
 Identities = 20/80 (25%), Positives = 36/80 (45%), Gaps = 2/80 (2%)
 Frame = -2

Query  613  GEKPPQYPCNAAYSKL--DILILNGCQRHFKDIPAFYVNFFCVFHGEHETTHWALTSIPR  440
            G++PP  P + A + L  D L+L    +H+K   A +  +  +  G  +   W L   P
Sbjct  557  GQRPPVTPLSRAEAGLPDDALVLAAFHQHYKITRASFALWMRLLRGLPDALLWLLEGAPS  616

Query  439  WFKVISLK*HGNNIDPTSVC  380
            +S +   + +DP  +C
Sbjct  617  AMARLSQEARAHGVDPARLC  636
```

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI — Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST/ tblastn

blastn | blastp | blastx | tblastn | tblastx

TBLASTN search translated nucleotide subjects using a protein query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>gi|47271417|ref|NP_571332.2| hemoglobin alpha ad
MSLSDTDKAVVKAIWAKISPKADEIGAEALARMLTVYPQTKTYFSHWAD
AVSKIDDLVGGLAALSELHAFKLRVDPANFKILSHNVIVVIAMLFPADFTPEVHVSVDKFFNNLALALSE
KYR
```

paste hbaa1 sequence

From

To

Or, upload file — Browse...

Job Title — gi|47271417|ref|NP_571332.2| hemoglobin alpha...

Enter a descriptive title for your BLAST search

Blast 2 sequences

Use BLAST 2 Sequences for Searching against small custom databases

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence — Clear — Subject subrange

```
GACACTCTTATTGCCATGACATTCAATTCTATAGTTGCCATTTTTCTGTGTAGATTGATAATAA
TTATATGCATTATGCAATCACGACTGTTGTTTACAGTGTACTCTGGAATTGTGTTATGCTCTCT
ATGGAATTATGTACCTTTTCCATTTCTATCTATACAAAACTTCAATAAACTTTTTCTGAACACA
```

From

paste Salmon ESTs

Or, upload file — Browse...

BLAST — Search nucleotide sequence using Tblastn (search translated nucleotide subjects using a protein query)

Show results in a new window

Algorithm parameters

Search against small custom database

Blast 2 sequences

gi|47271417|ref|NP_571332.2| hemoglobin alpha...

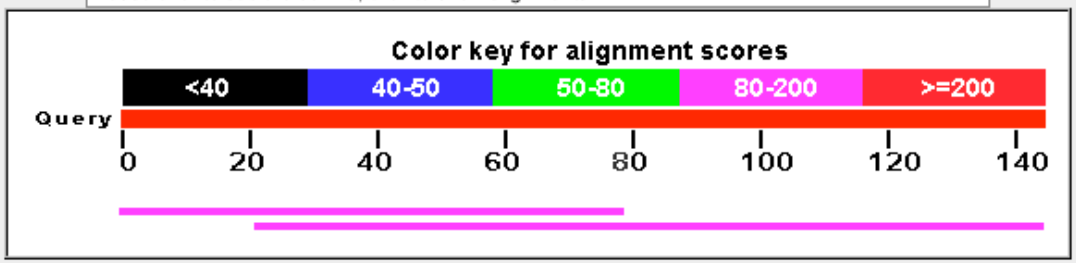
Query ID	lcl|20148	**Subject ID**	8 subjects
Description	gi|47271417|ref|NP_571332.2| hemoglobin alpha adult-1 [Danio rerio]	**Description**	See details
Molecule type	amino acid	**Molecule type**	nucleic acid
Query Length	143	**Subject Length**	n/a
		Program	TBLASTN 2.2.19+ Citation

Other reports: Search Summary [Taxonomy reports]

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

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

lcl|20152  gi|223585542|gb|GO065042.1|GO065042 EST_ssal_rgh_10...  116    3e-31
lcl|20155  gi|223585539|gb|GO065039.1|GO065039 EST_ssal_rgh_10...  178    4e-50
```

BLAST tasks

Basic BLAST

- ✓ Jurassic Park examples

Batch BLAST searching

- ✓ Use Salmon ESTs as input

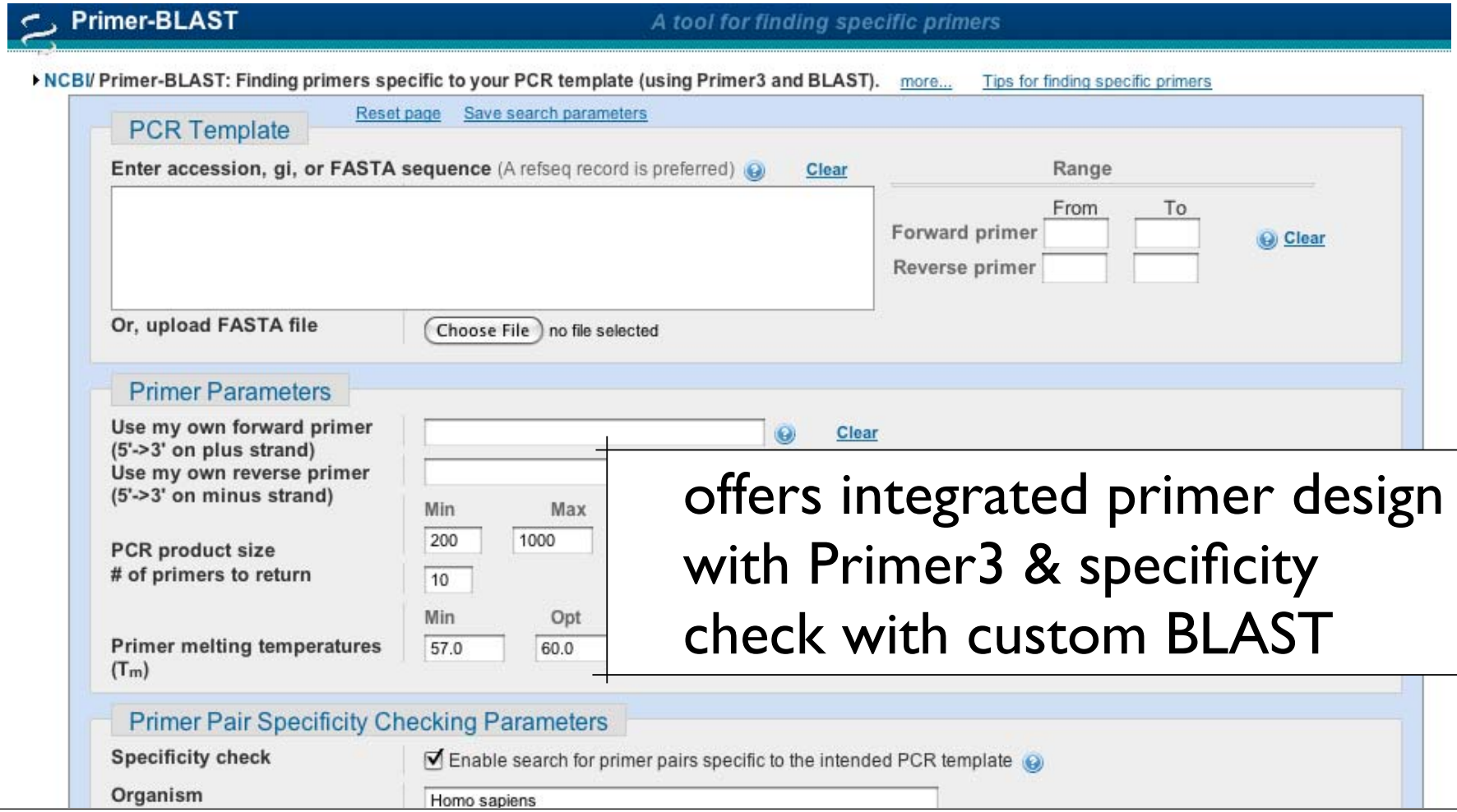
Search against a small custom database

- ✓ Use BLAST 2 Sequences utility

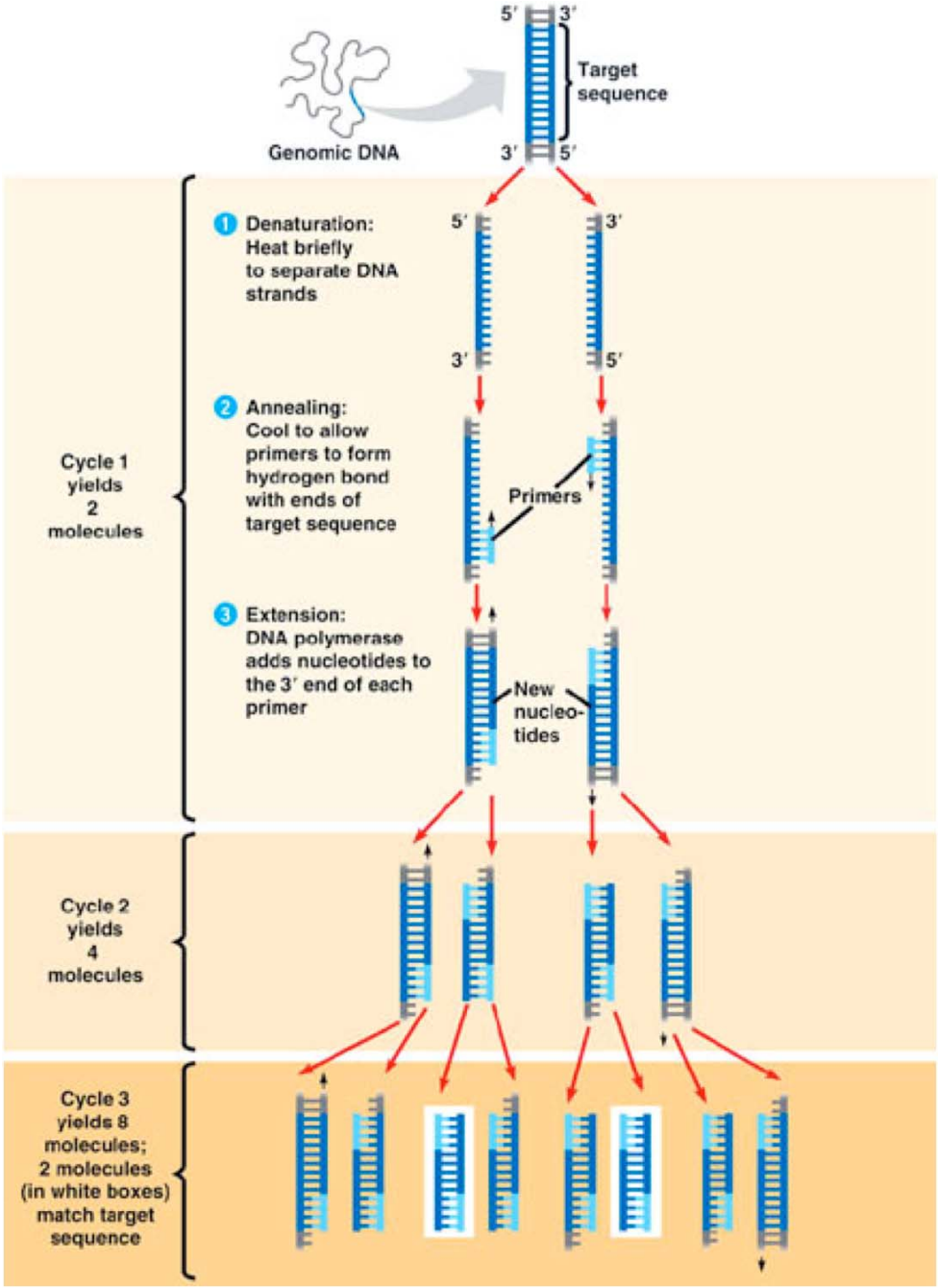
Primer-BLAST

NCBI's Primer Designer and Specificity Checker

http://www.ncbi.nlm.nih.gov/tools/primer-blast/



offers integrated primer design with Primer3 & specificity check with custom BLAST


Genomic DNA
5′
3′
Target sequence
3′
5′
1 Denaturation: Heat briefly to separate DNA strands
5′
3′
3′
5′
2 Annealing: Cool to allow primers to form hydrogen bond with ends of target sequence
Primers
3 Extension: DNA polymerase adds nucleotides to the 3′ end of each primer
New nucleo-tides
Cycle 1 yields 2 molecules
Cycle 2 yields 4 molecules
Cycle 3 yields 8 molecules; 2 molecules (in white boxes) match target sequence

Primer Design

Balance:

✓ Specificity - frequency of mispriming

✓ Efficiency of Amplification - 2X increase

Consider:

- primer length (18-24nt)
- primer Tm (>54°C)
- 3' end (G or C)
- GC content (45-55%)
- primer dimers
- for cDNA - coding region; across intron/ exon boundary

General Concepts for PCR Primer Design.
Dieffenback CW, Lowe TMJ, Dveksler GS Genome Research 3 (1993) S30-37 [PMID:8118394]

Primer-BLAST input

designs primers specific to target template and unique in the target database

NCBI/ Primer-BLAST: Finding primers

PCR Template

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

Range

From To

Forward primer

Reverse primer

Clear

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

Min Max

PCR product size 200 1000

# of primers to return 10

Min Opt Max Max Tm difference

Primer melting temperatures (Tm) 57.0 60.0 63.0 3

can specify primer sequence(s), desired product size, Tm ranges, Tm difference (can be used with or without template)

Primer-BLAST Specificity

By default human sequences are searched in specificity check

Primer Pair Specificity Checking Parameters

Specificity check

☑ Enable search for primer pairs specific to the intended PCR template

With this option on, the program will search the primers against the selected database and determine whether a primer pair can generate a PCR product on any targets in the database based on their matches to the targets and their orientations. The program will return, if possible, only primer pairs that do not generate a valid PCR product on unintended sequences and are therefore specific to the intended template. Note that the specificity is checked not only for the forward-reverse primer pair, but also for forward-forward as well as reverse-reverse primer pairs.

Organism

Homo sapiens

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

Database

Refseq mRNA (refseq_rna)

Primer specificity stringency

At least 2 total mismatches to unintended targets, including

at least 2 mismatches within the last 5 bps at the 3' end

The larger the mismatches (especially those toward 3' end) are between primers and the unintended targets, the more specific the primer pair is to your template (i.e., it will be difficult to anneal to and amplify unintended targets). However, specifying a larger mismatch value may make it more difficult to find such specific primers. Try to lower the mismatch value in such case.

custom BLAST; focus on 3' end to avoid mispriming

Misprimed product size deviation

1000

Splice variant handling

☐ Allow primer to amplify mRNA splice varia

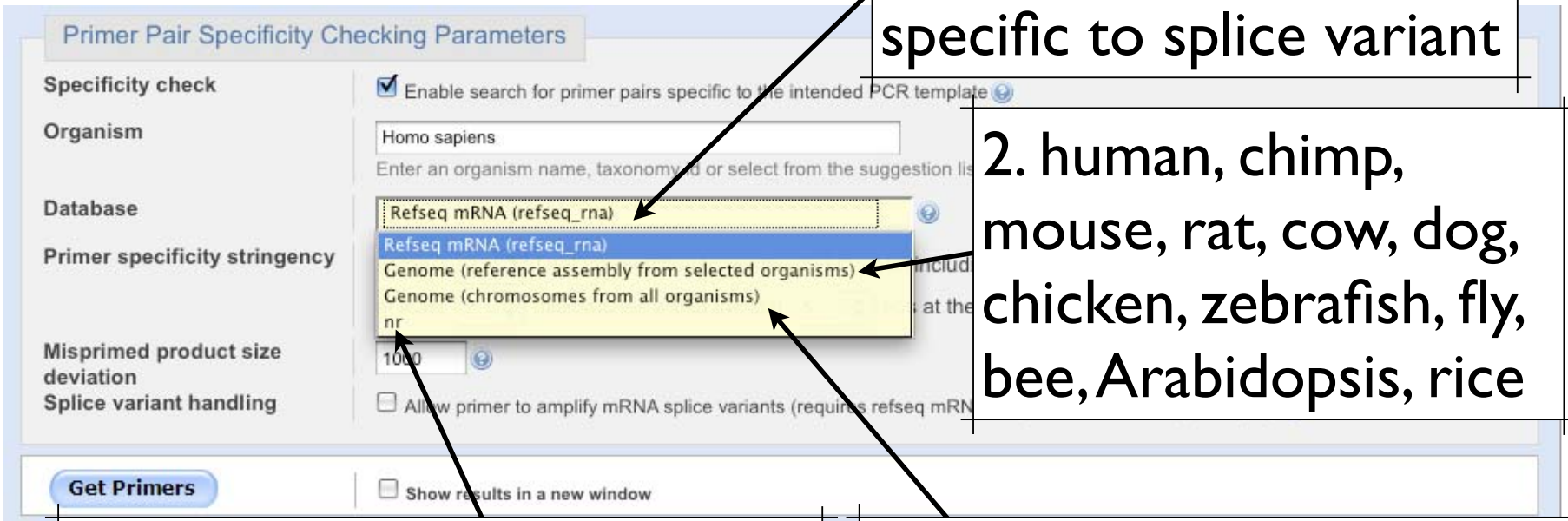
Get Primers

☐ Show results in a new window

▶ Advanced parameters

Primer-BLAST Specificity

Four BLAST nucleotide databases available for searching



1. with refseq template, specific to splice variant

2. human, chimp, mouse, rat, cow, dog, chicken, zebrafish, fly, bee, Arabidopsis, rice

3. all NC_ (includes above) + other organisms, microbes

4. nr = database with widest coverage of organisms

Primer-BLAST Advanced

Adjustable settings from Primer3 see Primer 3 Input Help:
http://fokker.wi.mit.edu/primer3/input-help-040.htm

Advanced parameters

Primer Pair Specificity Checking Parameters

Blast max number of hit sequences	250 (default)
Blast expect (E) value	1000 (default)
Max primer pairs to screen	3000 (default)

Primer Parameters

	Min	Opt	Max
PCR Product Tm			
Primer Size	15	20	27

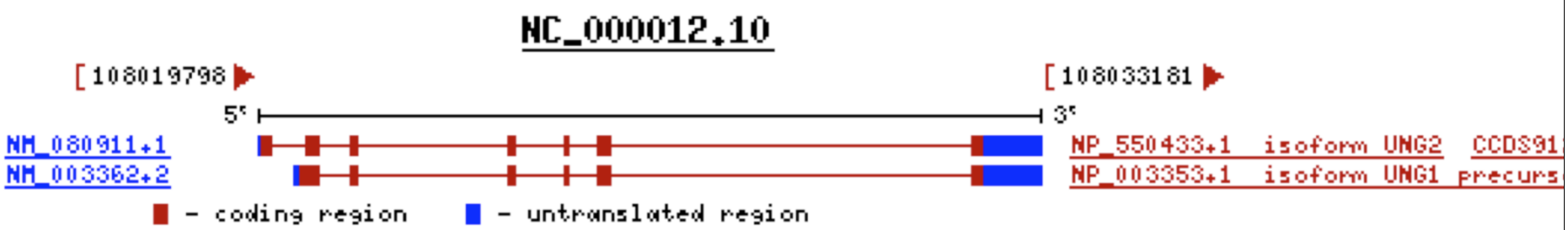
	Min	Max
Primer GC content (%)	20.0	80.0

GC clamp	0
Max self complementarity:	8.00
Max 3' end complementarity:	3.00
SNP handling	☐ Primer binding site may not contain known SNP
Repeat filter	Automatic Avoid repeat region for primer selection by filtering with repeat database
Low complexity filter	☑ Avoid low complexity region for primer selection
Concentration of monovalent cations	50.0
Concentration of divalent cations	0.0
Concentration of dNTPs	0.0
Salt correction formula:	Schildkraut and Lifson 1965
Annealing Oligo Concentration	50.0

Useful options specific to Primer-BLAST:

1. avoid regions that contain SNPs
2. avoid repetitive regions

Primer-BLAST example



Task #1: Use Primer BLAST to design primers specific to the UNG2 splice variant, NM_080911.

Task #2: Use Primer BLAST to design primers that will identify both splice variants.

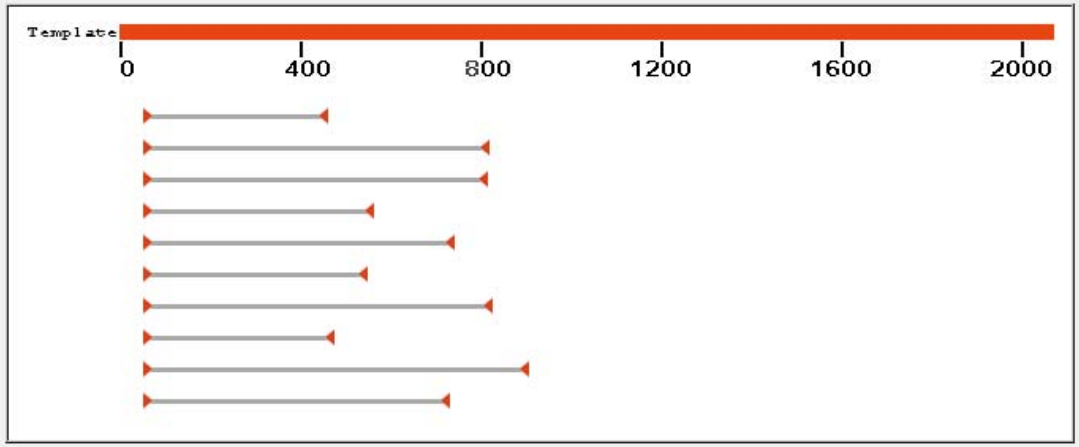
Task #3: Carry out a specificity check for one of your primer pairs. Will this primer pair (designed against the human UNG transcripts) also amplify transcripts from other primate species?

Task #1: Use Primer BLAST to design primers specific to the UNG2 splice variant, NM_080911.

enter NM_080911 as template

use all default settings

▼ Summary of primer pairs



▼ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	CTCCTCAGCTCCAGGATGAT	Plus	20	56	75	59.36	55.00%
Reverse primer	AGGTGAAGACTTGGTGTGGG	Minus	20	479	460	60.00	55.00%
Product length	424						

Products on intended target

>NM_080911.1 Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA

```
product length = 424
Forward primer  1    CTCCTCAGCTCCAGGATGAT  20
Template        56   ....................  75

Reverse primer  1    AGGTGAAGACTTGGTGTGGG  20
Template        479  ....................  460
```

Task #2: Use Primer BLAST to design primers that will identify both splice variants.

Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	CCCACACCAAGTCTTCACCT	Plus	20	460	479	60.00	55.00%
Reverse primer	CACCCCAACATCTGTCACTG	Minus	20	1407	1388	60.00	55.00%
Product length	948						

Products on intended target

>NM_080911.1 Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA

```
product length = 948
Forward primer  1     CCCACACCAAGTCTTCACCT  20
Template        460   ....................  479

Reverse primer  1      CACCCCAACATCTGTCACTG  20
Template        1407   ....................  1388
```

Products on allowed transcript variants

>NM_003362.2 Homo sapiens uracil-DNA glycosylase (UNG), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA

```
product length = 948
Forward primer  1     CCCACACCAAGTCTTCACCT  20
Template        488   ....................  507

Reverse primer  1      CACCCCAACATCTGTCACTG  20
Template        1435   ....................  1416
```

enter NM_080911 as template

☑ Allow primer to amplify mRNA splice variants

Task #3: Carry out a specificity check for one of your primer pairs. Will this primer pair (designed against the human UNG transcripts) also amplify transcripts from other primate species?

no template

use my own:

☑ forward primer

☑ reverse primer

☑ organism; specify primate

☑ database; specify nr

Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%
Forward primer	GCCTTGTTTTCTTGCTCTGG	20	59.99	50.00%
Reverse primer	CACCCCAACATCTGTCACTG	20	60.00	55.00%

Products on target templates

>AK291341.1 Homo sapiens cDNA FLJ76845 complete cds, highly similar to Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 1, mRNA

```
product length = 595
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        849  ....................  868

Reverse primer  1     CACCCCAACATCTGTCACTG  20
Template        1443  ....................  1424
```

>XM_001136198.1 PREDICTED: Pan troglodytes uracil-DNA glycosylase, transcript variant 1 (UNG), mRNA

```
product length = 595
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        925  ....................  944

Reverse primer  1     CACCCCAACATCTGTCACTG  20
Template        1519  ....................  1500
```

>XM_509349.2 PREDICTED: Pan troglodytes uracil-DNA glycosylase, transcript variant 2 (UNG), mRNA

```
product length = 595
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        848  ....................  867

Reverse primer  1     CACCCCAACATCTGTCACTG  20
Template        1442  ....................  1423
```

>XM_001104421.1 PREDICTED: Macaca mulatta similar to uracil-DNA glycosylase isoform UNG1 precursor, transcript variant 2 (LOC706816), mRNA

```
product length = 603
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        868  ....................  887

Reverse primer  1     CACCCCAACATCTGTCACTG  20
```

Things you can do to maximize the chance of finding primers specific for your template.

- Use refseq accession or GI (rather than the raw DNA sequence) as template whenever possible. Even if you are only interested in part of the sequence, you can still use the accession or GI but you do need to specify the range (use forward primer "From" field for your sequence start position and reverse primer "To" field for your sequence stop position). The reason is that an accession or GI carries accurate information about its identity which allows primer-blast to better distinguish between intended template and off-targets.

- Choose a non-redundant database (such as refseq_rna or genome database). The nr database contains redundant entries which can interfere with the process of finding specific primers.

- Specify an organism for database search if you are only amplifying DNA from a specific organism. Searching all organisms will be much slower and off-target priming from other organisms are irrelevant.

Credits

- Materials for this presentation have been adapted with permission from the following NCBI HelpDesk course materials:

  Field Guide Course Materials

  Advanced Workshop for Bioinformatics Information Specialists

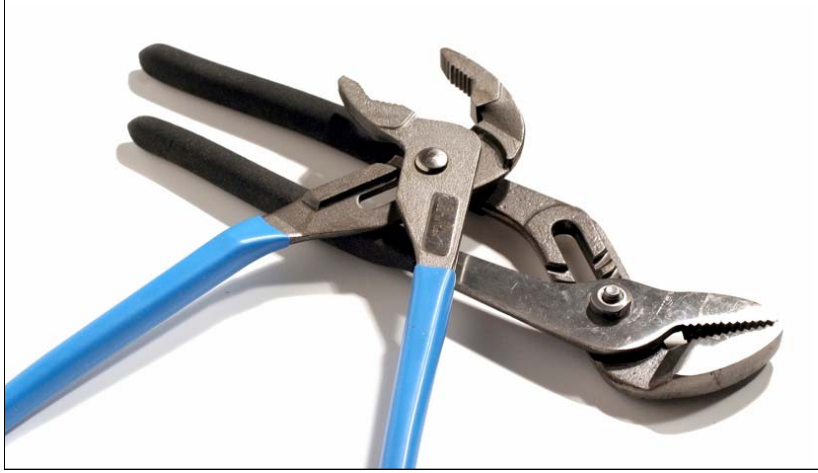
  NCBI News

- NCBI BLAST

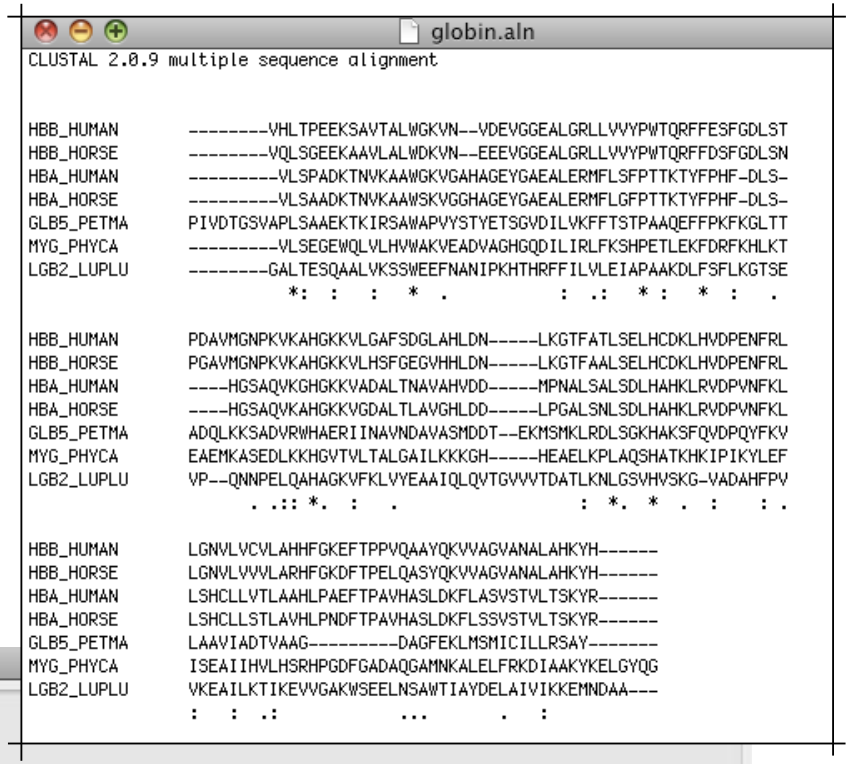
  http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

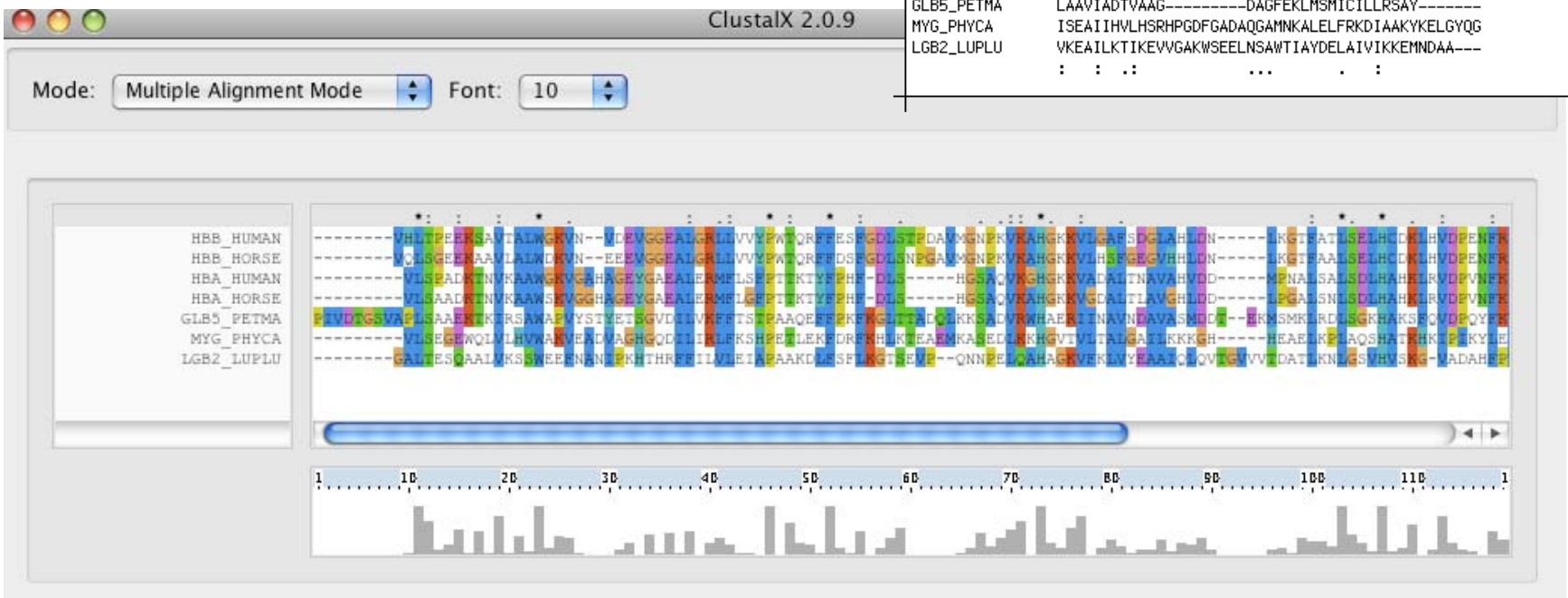
MSA

MSA = Multiple Sequence Alignments



Examples


globin.aln
CLUSTAL 2.0.9 multiple sequence alignment
HBB_HUMAN --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
HBB_HORSE --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
HBA_HUMAN ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
HBA_HORSE ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
MYG_PHYCA ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
LGB2_LUPLU --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
HBB_HUMAN PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
HBB_HORSE PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
HBA_HUMAN ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
HBA_HORSE ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
GLB5_PETMA ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
MYG_PHYCA EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
LGB2_LUPLU VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
HBB_HUMAN LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
HBB_HORSE LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
HBA_HUMAN LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBA_HORSE LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
GLB5_PETMA LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
MYG_PHYCA ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LGB2_LUPLU VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---


ClustalX 2.0.9
Mode:
Multiple Alignment Mode
Font:
10
HBB_HUMAN
HBB_HORSE
HBA_HUMAN
HBA_HORSE
GLB5_PETMA
MYG_PHYCA
LGB2_LUPLU

Multiple Sequence Alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

The sole purpose of multiple sequence alignments is to place *homologous positions* of *homologous sequences* into the *same column.*

Clustal

- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994)
  - CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.
    - Nucleic Acids Research, 22:4673-4680.

Differences between CLUSTAL and BLAST?

CLUSTAL	BLAST
• global alignment method	• local alignment method
- Align complete sequence	- Search for HSP
• Assumes homology	• Test for homology
• Complex gap penalties	• Simple gap penalties
• Slower	• Fast
• Align protein-protein or nucleotide-nucleotide only	• Translated searches

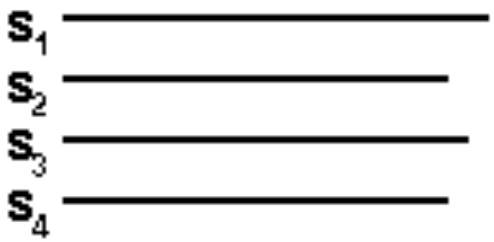
CLUSTAL Algorithm Steps

1. Pairwise alignment of each sequence pair
   - Number of comparisons depends on how many sequences
2. Compute distance matrix
   - Percent non-identity between each alignment pair
   - Lower distance means more similar
3. Construct a sequence similarity tree
   - Cluster sequences according to distance (similarity)
4. Progressive alignment of sequences according to a tree

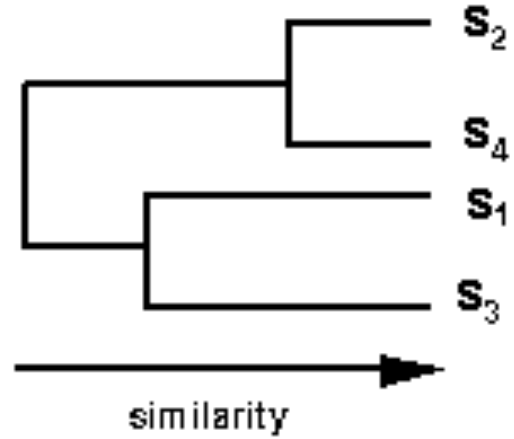
How does the Clustal algorithm actually work?

(A) Pairwise Alignment

Example – 4 sequences S_1 S_2 S_3 S_4



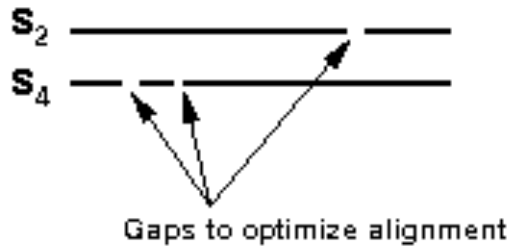
6 pairwise comparisons
then cluster analysis



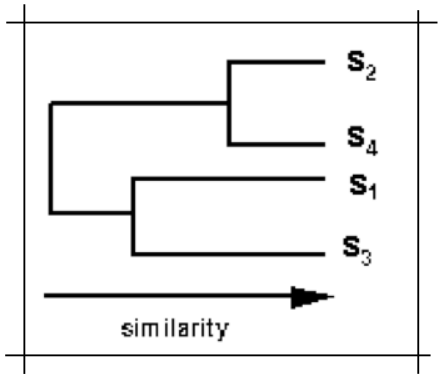
Which sequences would be aligned first?

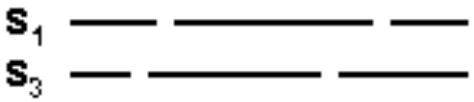
Steps in a Multiple Sequence Alignment continued …

(B) Multiple alignment following the tree from A

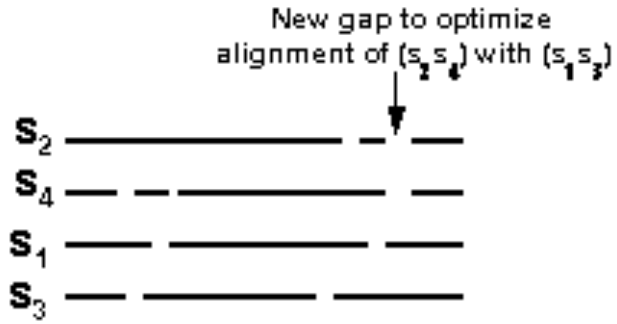


align most similar pair





align next most similar pair

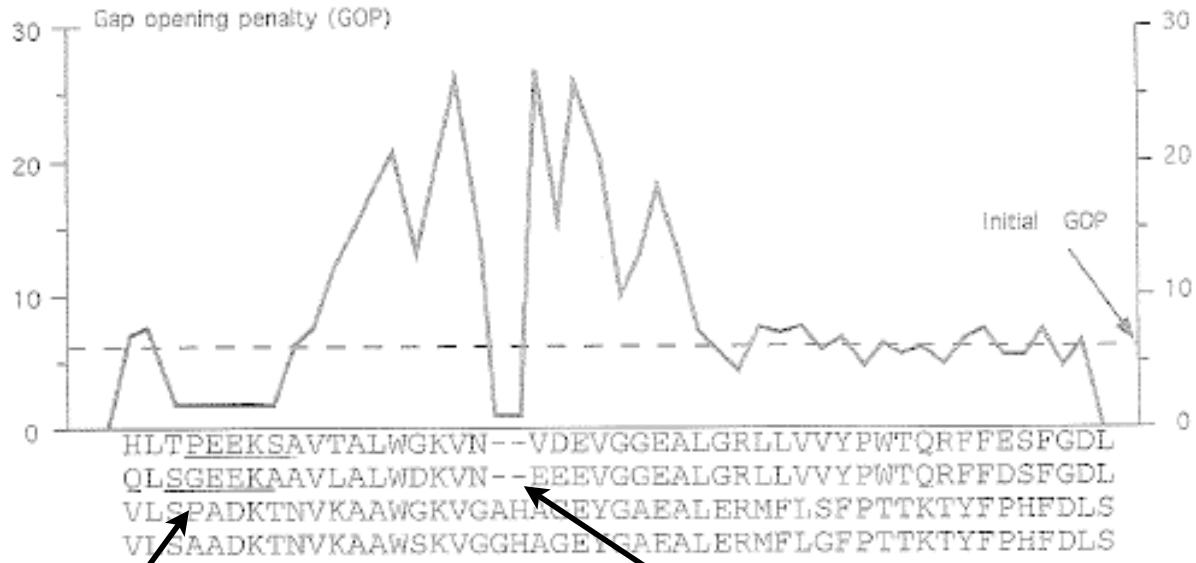


align alignments – preserve gaps

Position Specific Gap Penalities

- There are two type of gap opening penalities: gap opening and gap extension
  - Determined empirically by user
- Decrease penalties where gaps already occurs
- Increase penalties in adjacent positions to where gap already occurs
  - Encourage extension of gaps in loop regions vs. introduction of new gaps
- Increase or decrease gap penalties according to amino acid type
  - Increase penalties in stretches of hydrophobic residues

Gap Penalties Example



hydrophilic stretch

existing gap

Figure from Higgens et al, Methods in Enzymology 266: 383

Standard Multiple Sequence Alignment Approach

- Be as sure as possible that the sequences included are homologous
- Know as much as possible about the gene/ protein in question before trying to create an alignment (secondary structure, domains etc..)
- Start with an automated alignment: preferably one that utilizes some evolutionary theory such as CLUSTAL

http://www.ebi.ac.uk/Tools/clustalw2/index.html

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW2 Help
- ClustalW2 FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours

- Similar Applications
  - Align
  - Kalign
  - MAFFT
  - MUSCLE
  - T-Coffee

- ClustalW Programmatic Access

- www.clustal.org

Clustal Related Literature

Search for Clustal related literature in Medline...

EBI > Tools > Sequence Analysis > ClustalW2

ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

New users, please read the FAQ.

>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	yes	def	def
	ITERATION		NUMITER	
	none		1	

OUTPUT		PHYLOGENETIC TREE			
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
aln w/numbers	aligned	none	off	off	NJ

Enter or paste a set of sequences in any supported format: Help

http://www.ebi.ac.uk/Tools/muscle/index.html

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- Muscle Help
- Jalview Help

- Similar Applications
  - Align
  - ClustalW2
  - Kalign
  - MAFFT
  - T-Coffee

- Muscle Programmatic Access

EBI > Tools > Sequence Analysis

MUSCLE

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Download Software

MUSCLE multiple sequence alignment method with reduced time and space complexity

RESULTS	SEARCH TITLE	YOUR EMAIL
interactive	Sequence	
OUTPUT FORMAT	**OUTPUT TREE**	**OUTPUT ORDER**
FASTA	none	aligned

Enter or Paste a set of Sequences in any supported format: Help

Upload a file: Choose File no file selected Run Reset

If you plan to use these services during a course please contact us.

http://www.ebi.ac.uk/Tools/t-coffee/index.html

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- TCoffee Help
- Jalview Help
- Alignment
- Guide Tree
- Colours

- Similar Applications
  - Align
  - ClustalW2
  - Kalign
  - MAFFT
  - MUSCLE

- T-Coffee Programmatic Access

T-Coffee Related Literature

Search for T-Coffee related literature in Medline...
more

EBI > Tools > Sequence Analysis

T-Coffee

T-Coffee is a multiple sequence alignment program. Multiple sequence alignment programs are meant to align a set of sequences previously gathered using other programs such as blast, fast, sw ...
The main characteristic of T-Coffee is that it will allow you to combine results obtained with several alignment methods. For instance if you have an alignment coming from ClustalW2, an other alignment coming from Dialign, and a structural alignment of some of your sequences, T-Coffee will combine all that information and produce a new multiple sequence having the best agreement whith all these methods.
By default, T-Coffee will compare all you sequences two by two, producing a global alignment and a series of local alignments (using lalign). The program will then combine all these alignments into a multiple alignment.

TCoffee Multiple Sequence Alignment

Download Software

EMAIL	RESULTS	RUN NAME	MATRIX	ORDER
	interactive	Sequence	none	aligned

Enter or Paste a set of Sequences in any supported format: Help

Upload a file: Choose File no file selected Run Reset

Standard Multiple Sequence Alignment Approach

Examine alignment:

- Are you confident that aligned residues/bases evolved from a common ancestor?
- Are domains of the proteins/predicted secondary structures, etc. aligning correctly?
- Are most indels outside of known motifs or secondary structure?

  → No? May need to edit sequences and redo…

The Take Home Message

Why perform an MSA?

- Visualize trends between homologous sequences
  - Shared regions of homology
  - Regions unique to a sequence within a family
  - Consensus sequence
- As the first step in a phylogenetic analysis

The Take Home Message

How does one perform an MSA?

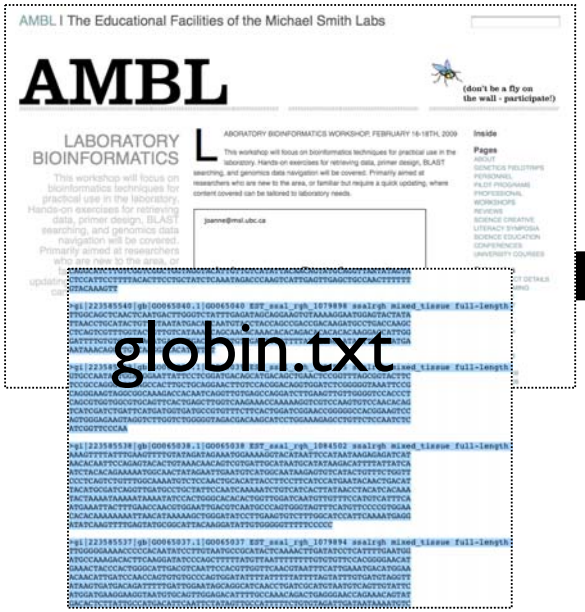
- By hand: too hard!
- Automated alignment: Fast, but doesn't necessarily produce the "correct" alignment

Best approach = Automated alignment with manual editing

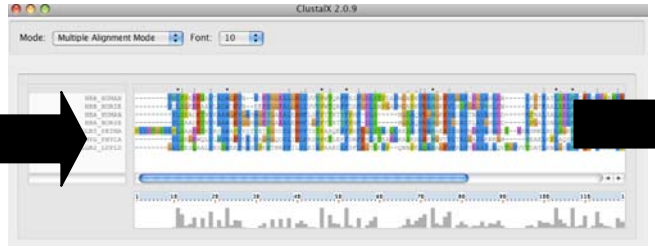
MSA

PRACTICAL EXERCISE: Comparing Sets of Protein Sequences

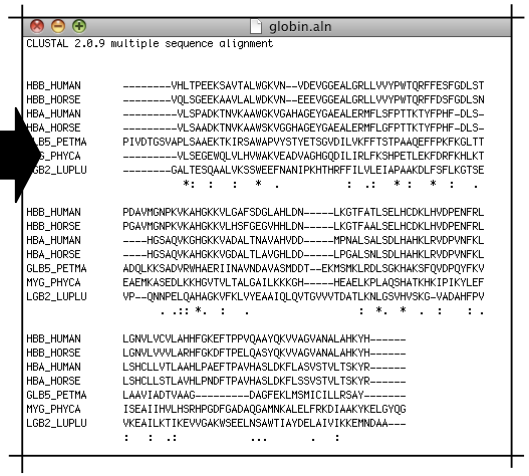
navigate to:
bioteach.ubc.ca/bioinfo2009



Clustal



We'll walk through install + do MSA #1 together



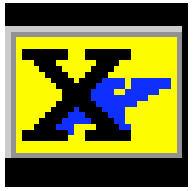
Install ClustalX on laptop

download program and install

Use ClustalX to generate MSA

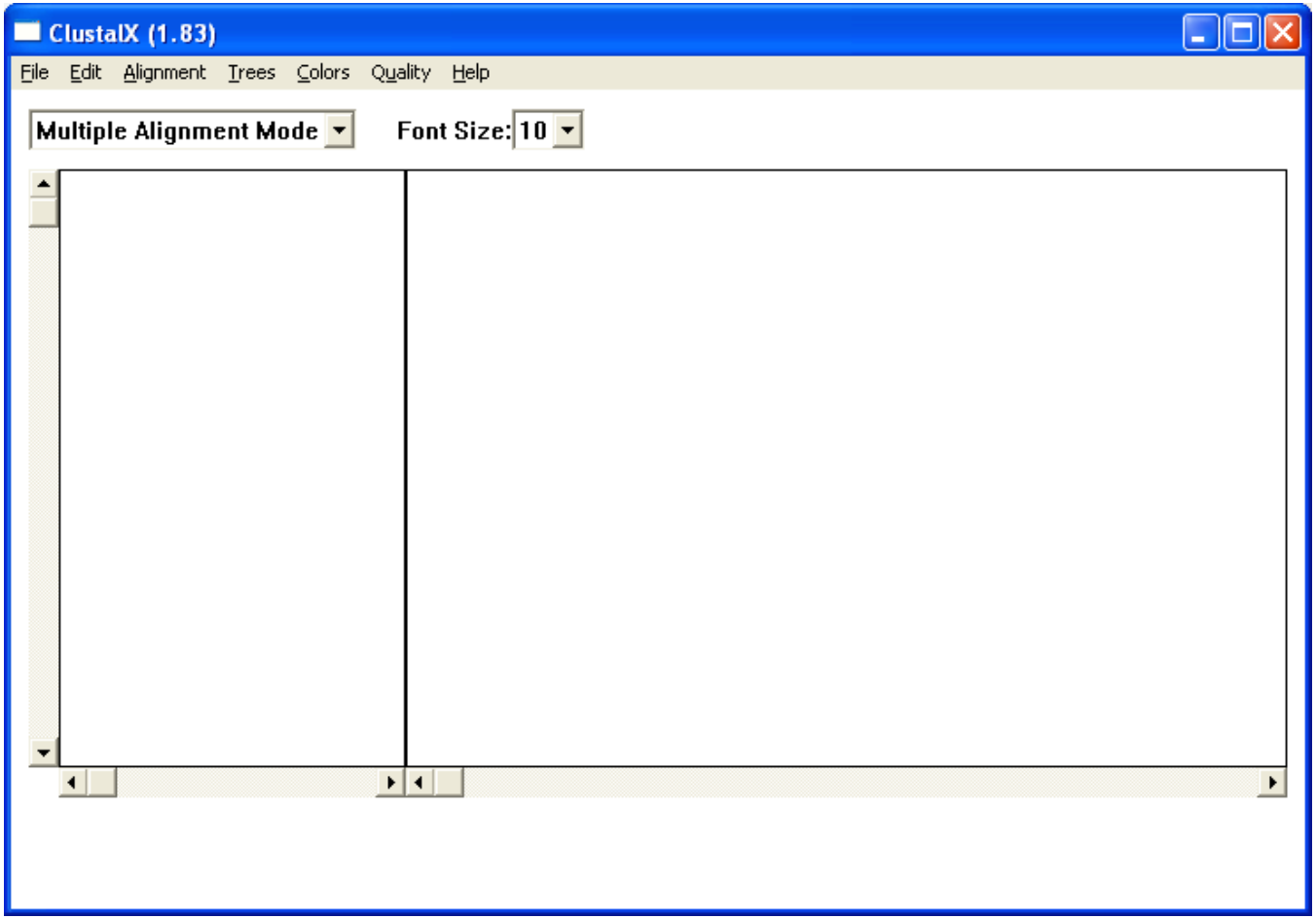
MSA #1: Use example sequences to generate alignment

MSA #2: Use your own sequences



Clustalx.exe

Open ClustalX



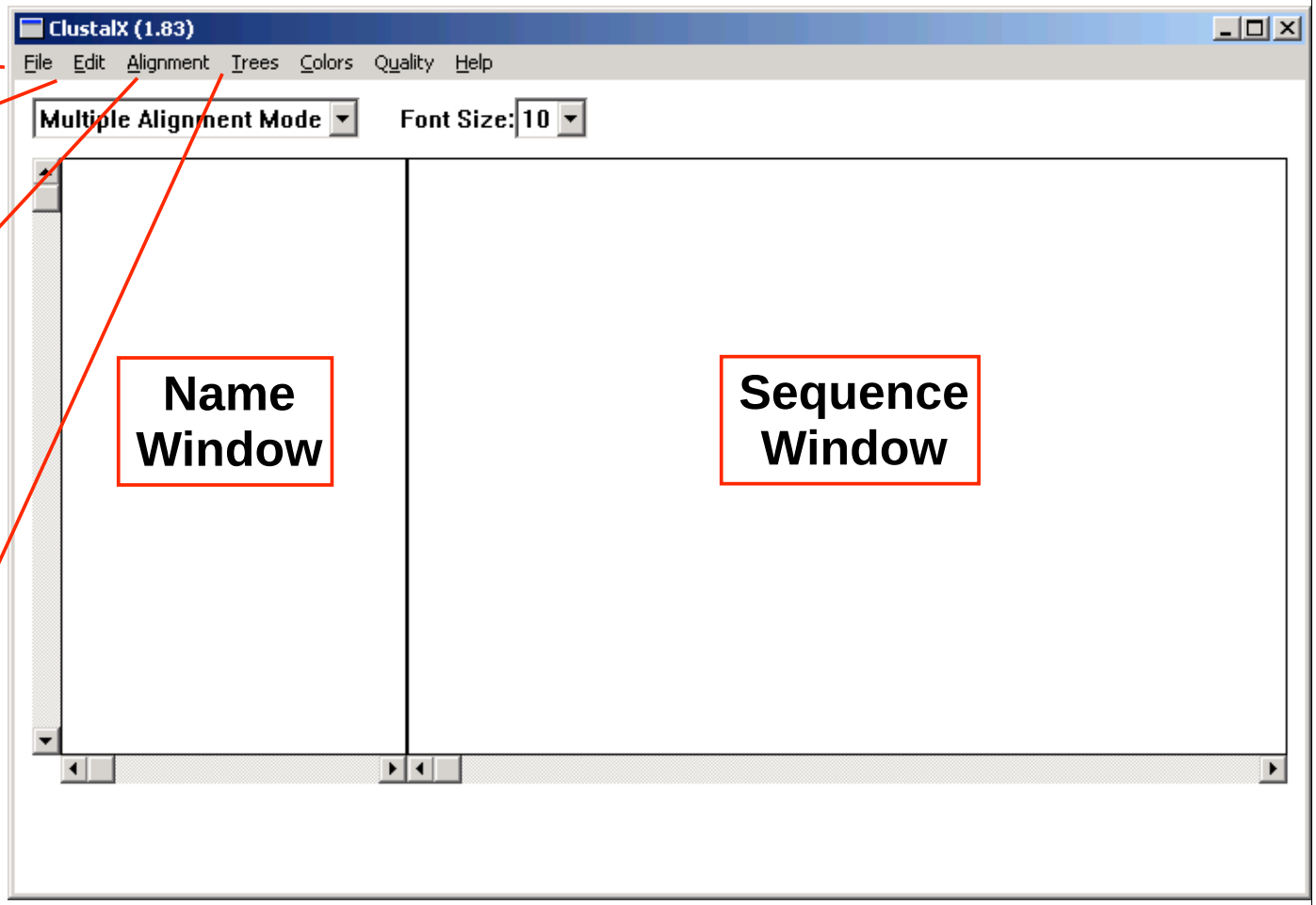
Starting up ClustalX

File:
-Load sequences

Edit:
-Remove all gaps

Alignment:
-Do complete alignment
-Alignment parameters

Trees:
-Bootstrapped NJ
-Output format options



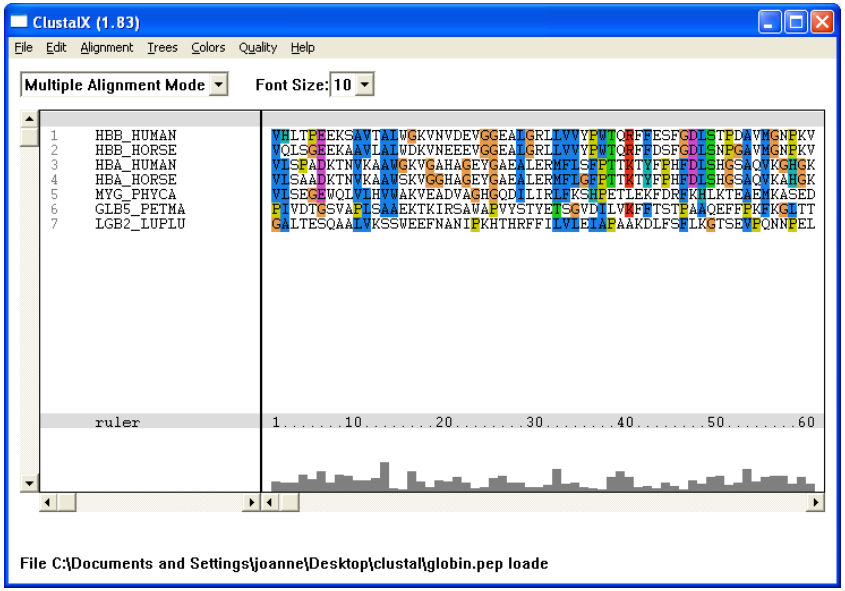
globin.pep - WordPad

File Edit View Insert Format Help eFax

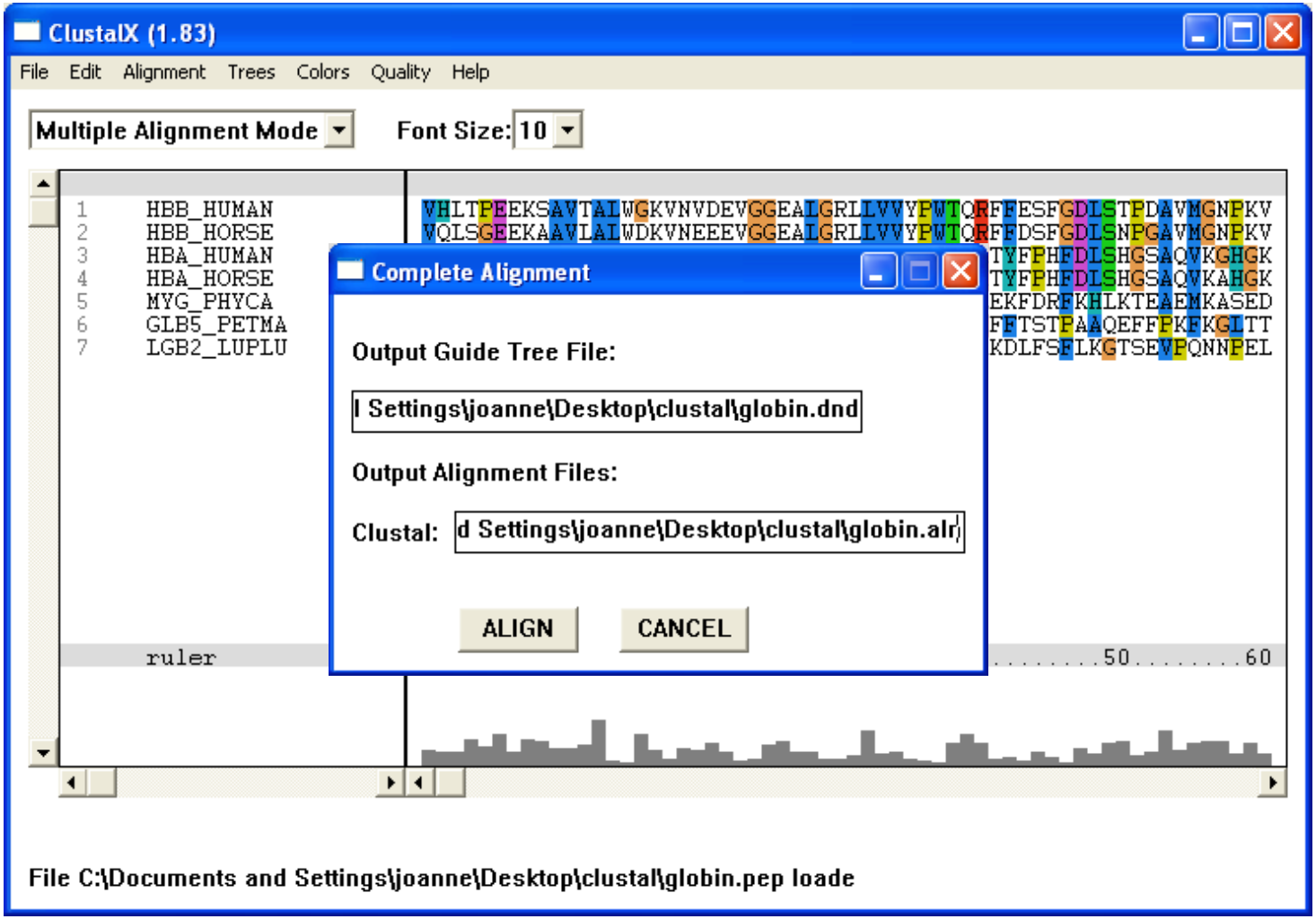
```
>P1;HBB_HUMAN
Sw:Hbb_Human => HBB_HUMAN
          VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
          PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP
          ENFRLLGNVL VCVLAHHFGK EFTPPVQAAY QKVVAGVANA LAHKYH*
C;ID   HBB_HUMAN      STANDARD;      PRT;   146 AA.
C;AC   P02023;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-APR-1993 (REL. 25, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN BETA CHAIN. . . .

>P1;HBB_HORSE
Sw:Hbb_Horse => HBB_HORSE
          VQLSGEEKAA VLALWDKVNE EEVGGEALGR LLVVYPWTQR FFDSFGDLSN
          PGAVMGNPKV KAHGKKVLHS FGEGVHHLDN LKGTFAALSE LHCDKLHVDP
          ENFRLLGNVL VVVLARHFGK DFTPELQASY QKVVAGVANA LAHKYH*
C;ID   HBB_HORSE      STANDARD;      PRT;   146 AA.
C;AC   P02062;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-MAR-1992 (REL. 21, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN BETA CHAIN. . . .

>P1;HBA_HUMAN
Sw:Hba_Human => HBA_HUMAN
          VLSPADKTNV KAAWGKVGAH AGEYGAEALE RMFLSFPTTK TYFPHFDLSH
          GSAQVKGHGK KVADALTNAV AHVDDMPNAL SALSDLHAHK LRVDPVNFKL
          LSHCLLVTLA AHLPAEFTPA VHASLDKFLA SVSTVLTSKY R*
C;ID   HBA_HUMAN      STANDARD;      PRT;   141 AA.
C;AC   P01922;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-FEB-1994 (REL. 28, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN ALPHA CHAIN. . . .

>P1;HBA_HORSE
Sw:Hba_Horse => HBA_HORSE
          VLSAADKTNV KAAWSKVGGH AGEYGAEALE RMFLGFPTTK TYFPHFDLSH
          GSAQVKAHGK KVGDALTLAV GHLDDLPGAL SNLSDLHAHK LRVDPVNFKL
          LSHCLLSTLA VHLPNDFTPA VHASLDKFLS SVSTVLTSKY R*
C;ID   HBA_HORSE      STANDARD;      PRT;   141 AA.
C;AC   P01958;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-MAR-1992 (REL. 21, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN ALPHA CHAINS (SLOW AND FAST). . . .
```

Load the sequences

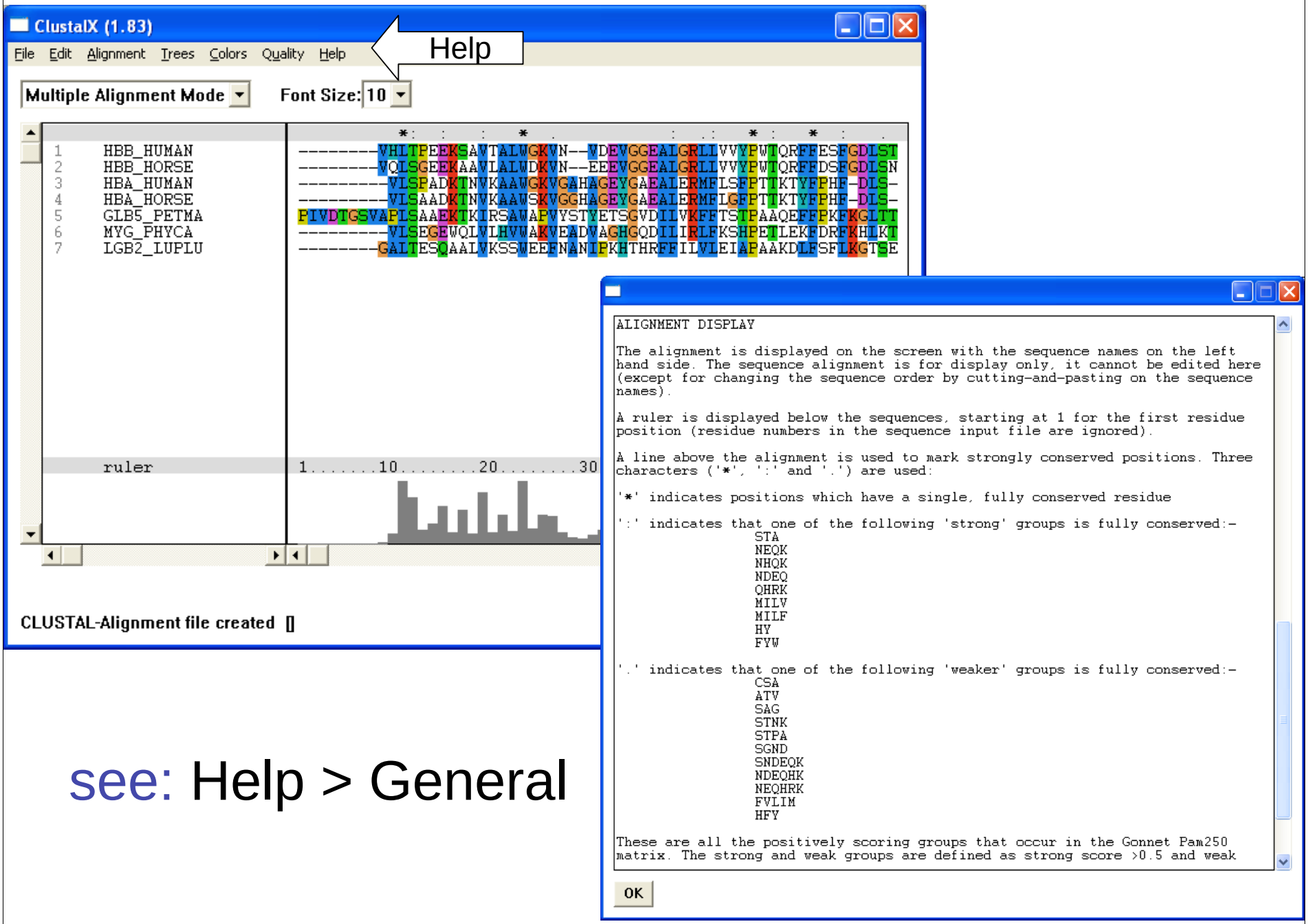
–globin.pep



Alignment > Do Complete Alignment



also see: Alignment > Alignment Parameters



see: Help > General

Expasy Proteomics Server

Site Map | Search ExPASy | Contact us

Search for Go Clear

ExPASy Proteomics Server

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References / Linking to ExPASy).

[Databases] [Tools & Software] [Education & Services] [Links]
[Announcements] [Mirror Sites] [Job openings]

Databases	Tools and software packages

• **UniProt Knowledgebase (Swiss-Prot and TrEMBL)** - Protein knowledgebase
• **ViralZone** - Portal to viral UniProtKB/Swiss-Prot entries new
• **PROSITE** - Protein families and domains
• **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis
• **World-2DPAGE Repository** - A public standards-compliant repository for gel-based proteomics data published in the literature
• **MIAPEGelDB** - A public repository for MIAPE Gel electrophoresis documents
• **ENZYME** - Enzyme nomenclature
• **UniPathway** - Metabolic pathways
• **SWISS-MODEL Repository** - Automatically generated protein models
 	
• Links to many other molecular biology databases
	• **Proteomics and sequence analysis tools**
	  ◦ Identification and characterization (Aldente, FindMod, Popitam, Phenyx, pI/Mw, ProtParam...)
	  ◦ DNA -> Protein
	  ◦ Similarity searches (BLAST...)
	  ◦ Pattern and profile searches (ScanProsite...)
	  ◦ Post-translational modification and topology prediction
	  ◦ Primary structure analysis
	  ◦ Secondary and tertiary structure tools (Swiss-PdbViewer...)
	  ◦ Alignment and Phylogenetic analysis
	• **Melanie / ImageMaster** - Software for 2-D PAGE analysis
	• **MSight** - Mass Spectrometry Imager
	• **Roche Applied Science's Biochemical Pathways**

Let's start at 9:00am

Genome Browsers
GEO - gene expression omnibus
Pathway Resources for Systems Biology

joanne@msl.ubc.ca