BCB 5200 Introduction to Bioinformatics

Sequence Analysis Lab

Bioinformatics and Computational Biology
Saint Louis University

Batch BLAST jobs through web

- There are two methods to do batch BLAST jobs.
  - The first is through the web interface
  - the second is using the standalone BLAST binaries and downloaded NCBI databases.

- If you are going to submit a batch BLAST search on the web we recommend that you do not submit a file of more than 40 sequences.

BLAST Results

Edit and Resubmit | Save Search Strategies | ▷ Formatting options | ▷ Download | YouTube How to read this page | Blast report description

Job title: MH0068_GL0058265 (4176 letters)

Results for: 4:lcl|Query_133185 MH0068_GL0055100(2892bp) ▼

1:lcl|Query_133182 MH0068_GL0058265(4176bp)
2:lcl|Query_133183 MH0068_GL0019004(2217bp)
3:lcl|Query_133184 MH0068_GL0010430(1812bp)
4:lcl|Query_133185 MH0068_GL0055100(2892bp)
5:lcl|Query_133186 MH0068_GL0055096(2814bp)
6:lcl|Query_133187 MH0074_GL0034786(2679bp)
7:lcl|Query_133188 MH0068_GL0055841(3474bp)
8:lcl|Query_133189 MH0068_GL0027893(1392bp)
9:lcl|Query_133190 MH0068_GL0026512(4713bp)
10:lcl|Query_133191 MH0068_GL0012576(1923bp)
11:lcl|Query_133192 MH0006_GL0176780(1536bp)
12:lcl|Query_133193 MH0177_GL0078034(2880bp)
13:lcl|Query_133194 MH0095_GL0085358(2475bp)
14:lcl|Query_133195 MH0074_GL0048502(4395bp)
15:lcl|Query_133196 MH0068_GL0037198(1983bp)
16:lcl|Query_133197 MH0095_GL0040817(2124bp)
17:lcl|Query_133198 MH0095_GL0025062(2637bp)
18:lcl|Query_133199 MH0006_GL0031342(1584bp)
19:lcl|Query_133200 MH0074_GL0030661(3480bp)
20:lcl|Query_133201 MH0019_GL0006447(2280bp)

RID
Query ID
Description
Molecule type
Query Length

Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.7.0+ ▷ Citation

Other reports: ▷ S… tree of results]

Graphic Summ…

…0 Blast Hits on 1000 subject sequences

…e the title, click to show alignments

…y for alignment scores

<40 | 40-50 | 50-80 | 80-200 | >=200

Query

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - U6PWDXWJ014

Your search is limited to records matching entrez query: txid4932 [ORGN].

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

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

20 sequences (YAL015C Chr1 complement(126904..128103)...

Results for: 1:lcl|63955 YAL015C Chr1 complement(126904..128103) [1200 bp, 399 aa] DNA N-glycosylase and apurinic/apyrimidin...(400aa) What's this?

Query ID lcl|63955
Description YAL015C Chr1 complement(126904..128103) [1200 bp, 399 aa] DNA N-glycosylase and apurinic/apyrimidinic (AP) lyase involved in base excision repair, localizes to the nucleus and mitochondrion
Molecule type amino acid
Query Length 400

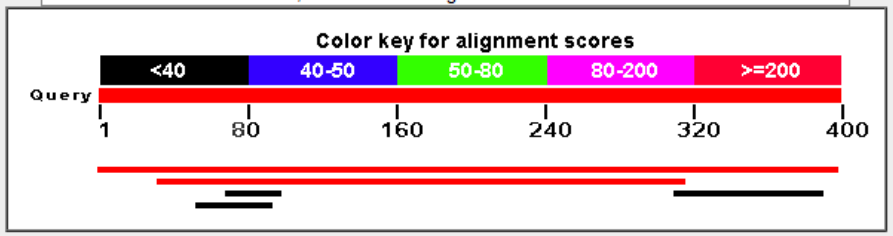
Database Name refseq_protein
Description NCBI Protein Reference Sequences
Program BLASTP 2.2.25+ Citation

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

Graphic Summary

Distribution of 5 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

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

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_009387.1	Ntg1p [Saccharomyces cerevisiae S288c]	818	818	99%	0.0	U G

Command-line BLAST+

Visit the BLAST site at NCBI ("help" tab) to find the URL for the BLAST+ download.

```
$ mkdir database # this creates a new directory
$ cd database/ # we navigate into that directory
# Enter the following, without arguments, to see a help document.
$ update_blastdb.pl
# Next get a list of all available databases
$ update_blastdb.pl --showall
$ update_blastdb.pl --showall | less
```

```
$ update_blastdb.pl refseq_protein
```

```
$ tar -zxvf refseq_protein.00.tar.gz
```

Command-line BLAST+

- Go to NCBI-Genome database
- Search “Escherichia coli [orgn]”
- Find NC_002695.1

Reference genomes: [see all organisms]

- ***Escherichia coli O157:H7 str. Sakai***

  Submitter: GIRC

  Human Pathogen

  Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes

  Environment: OxygenReq:Facultative, OptimumTemperature:37, Tempera

  Phenotype: Disease:Hemorrhagic colitis

Type	Name	RefSeq		Size (Mb)
Chr	-	NC_002695.1	A000007.2	5.5
Plsm	pOSAK1	NC_002127.1	AB011548.2	0
Plsm	pO157	NC_002128.1	AB011549.2	0.09

- ***Escherichia coli IAI39***

  Submitter: Genoscope

  Human Pathogen

  Morphology: Gram:Negative, Shape:Bacilli, Motility:No

  Environment: OxygenReq:Facultative, TemperatureRange:Mesophilic, Ha

  Phenotype: Disease:Urinary tract infection

Type	Name	RefSeq	INSDC	Size (Mb)
Chr	-	NC_011750.1	CU928164.2	5.13

- ***Escherichia coli str. K-12 substr. MG1655***

  Submitter: Univ. Wisconsin

00)

yama,K., Kuhara,S., Hattori,M.,
okoyama,K., Makino,K., Shinagawa,H. and

whole set of rRNA operons between an
ia coli O157:H7 Sakai strain and an
n MG1655
3), 315-324 (2000)

ota,Y., Yutsudo,C.H., Kimura,S.,
ori,M., Tatsuno,I., Abe,H., Iida,T.,
ashi,T., Yasunaga,T., Honda,T.,
.
e of the prophage VT2-Sakai carrying the
terohemorrhagic Escherichia coli O157:H7
reak
227-239 (1999)

ional Center for Biotechnology
MD 20894, USA

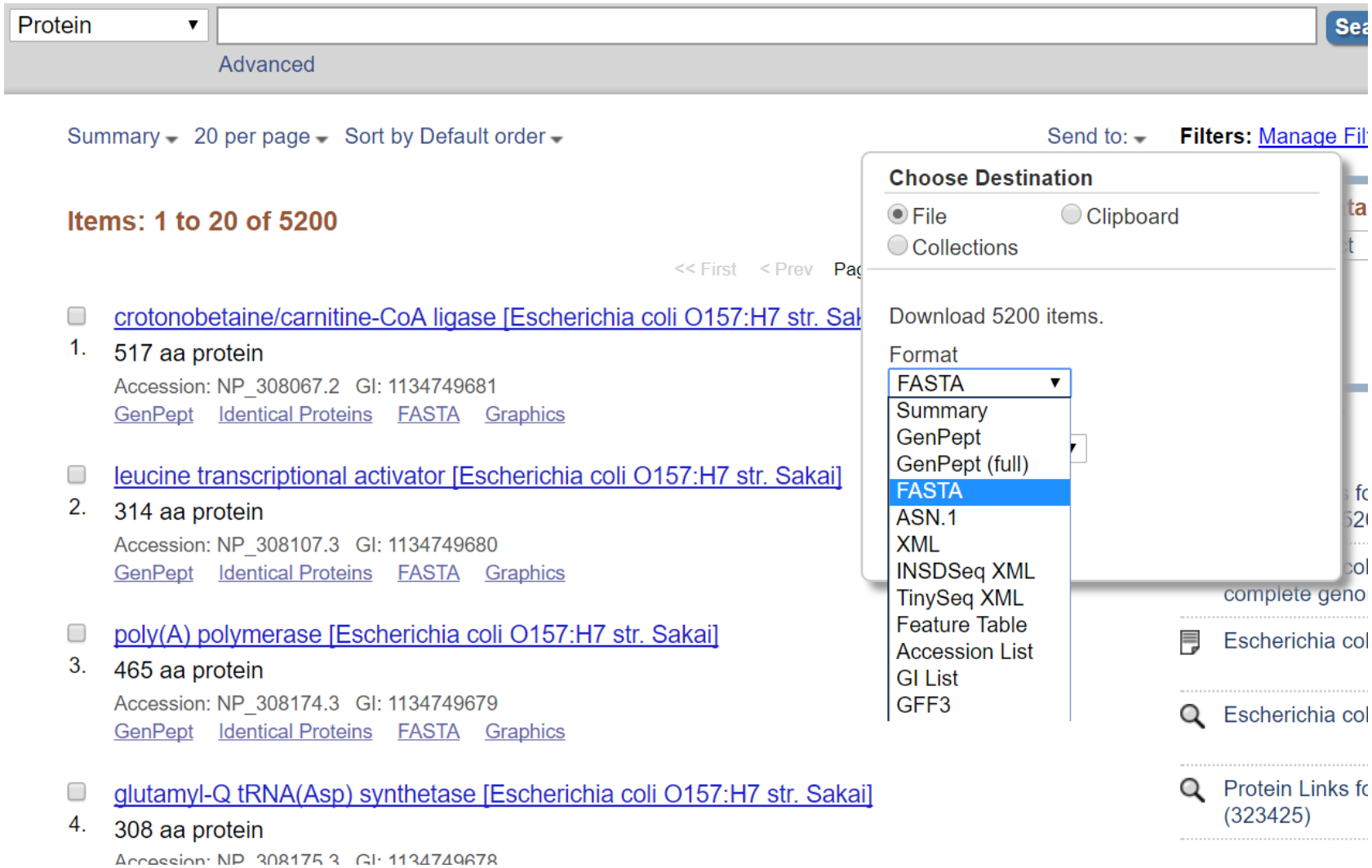
tation Project

Related information

- Assembly
- BioProject
- Components (Core)
- Full text in PMC
- Gene
- Genome
- Identical GenBank Sequence
- Probe
- Protein
- PubMed
- PubMed (Weighted)
- Reference Genome BioProject
- Representative Genome BioProject
- Taxonomy

Command-line BLAST+

- Download 5200 proteins in the fasta format



Command-line BLAST+

- Download 5200 proteins in the fasta format
- Go to terminal, log in with your account
- mkdir course
- cd course
- cat> Ecoli.O157.fasta
- Paste the sequences
- Ctrl+D for ending terminal line input

Command-line BLAST+

makeblastdb to make BLAST database (-help)

- makeblastdb -in Ecoli.O157.fa -parse_seqids -dbtype prot
- cat>RHS.fasta

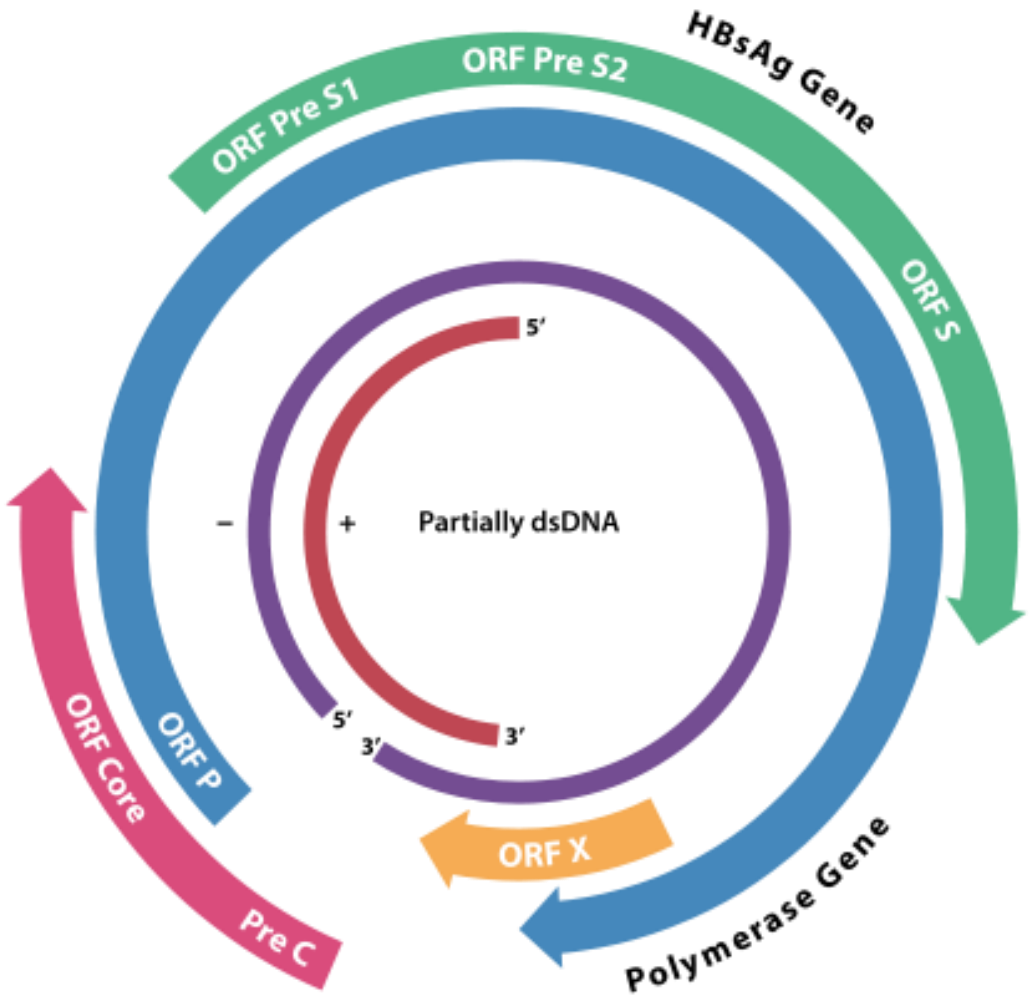
- blastp -query RHS.fasta -db Ecoli.O157.fasta -out RHS.Ecoli.O157.result

- less RHS.Ecoli.O157.result

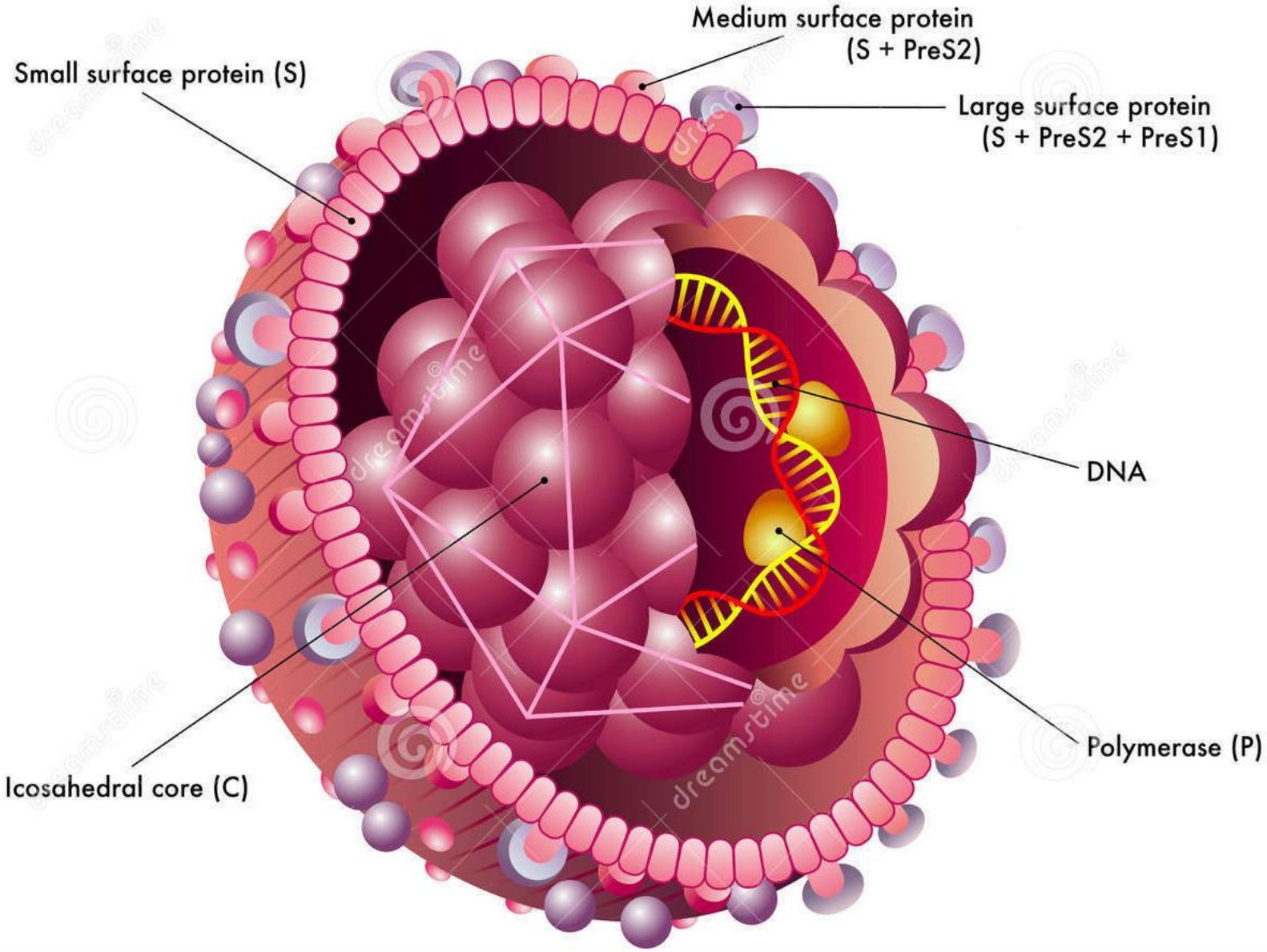
Understanding function of HBV surface antigen (CAA53344)

Hepatitis B virus (HBV)

- A double-stranded DNA virus belonging to the Hepadnaviridae family
- Causes human liver disease , hepatitis B
- Genome size: 3182 and 3248 bp depending on genotypes
- The genome encodes four overlapping open reading frames (ORFs):
  - Core protein (HBcAg, 2 isoforms)
  - Polymerase/reverse transcriptase (RT)
  - Surface proteins (HBsAg, 3 isoforms)
  - HBx (regulatory unit, transcriptional transactivator)



Hepatitis B virus



HBV genome

- https://www.ncbi.nlm.nih.gov/nuccore/NC_003977.2

Fasta format

```
>YP_009173869.1 large envelope protein [Hepatitis B virus]
MGQNLSTSNPLGFFPDHQLDPAFRANTANPDWDFNPNKDTWPDANKVGAGAFGLGFTPPHGGLLGWSPQA
QGILQTLPANPPPASTNRQSGRQPTPLSPPLRNTHPQAMQWNSTTFHQTLQDPRVRGLYFPAGGSSSGTV
NPVLTTASPLSSIFSRIGDPALNMENITSGFLGPLLVLQAGFFLLTRILTIPQSLDSWWTSLNFLGGTTV
CLGQNSQSPTSNHSPTSCPPTCPGYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVCPLIPGSSTTS
TGPCRTCMTTAQGTSMYPSCCCTKPSDGNCTCIPIPSSWAFGKFLWEWASARFSWLSLLVPFVQWFVGLS
PTVWLSVIWMMWYWGPSLYSILSPFLPLLPIFFCLWVYI
```

How to understand its function?

- PubMed search?

- BLAST search
  - Homologs?
  - Species distribution?
  - Any likely features?

Choice of BLAST programs?

- BLASTP
- PSI-BLAST
- DELTA-BLAST
- RPS-BLAST

- Tips on parameters
  - Database (NR vs RefSeq)
  - Max targets (20,000)
  - Threshold (0.05 vs 0.001)
  - Filter (low complexity)
  - Taxonomy report (restrict organism)

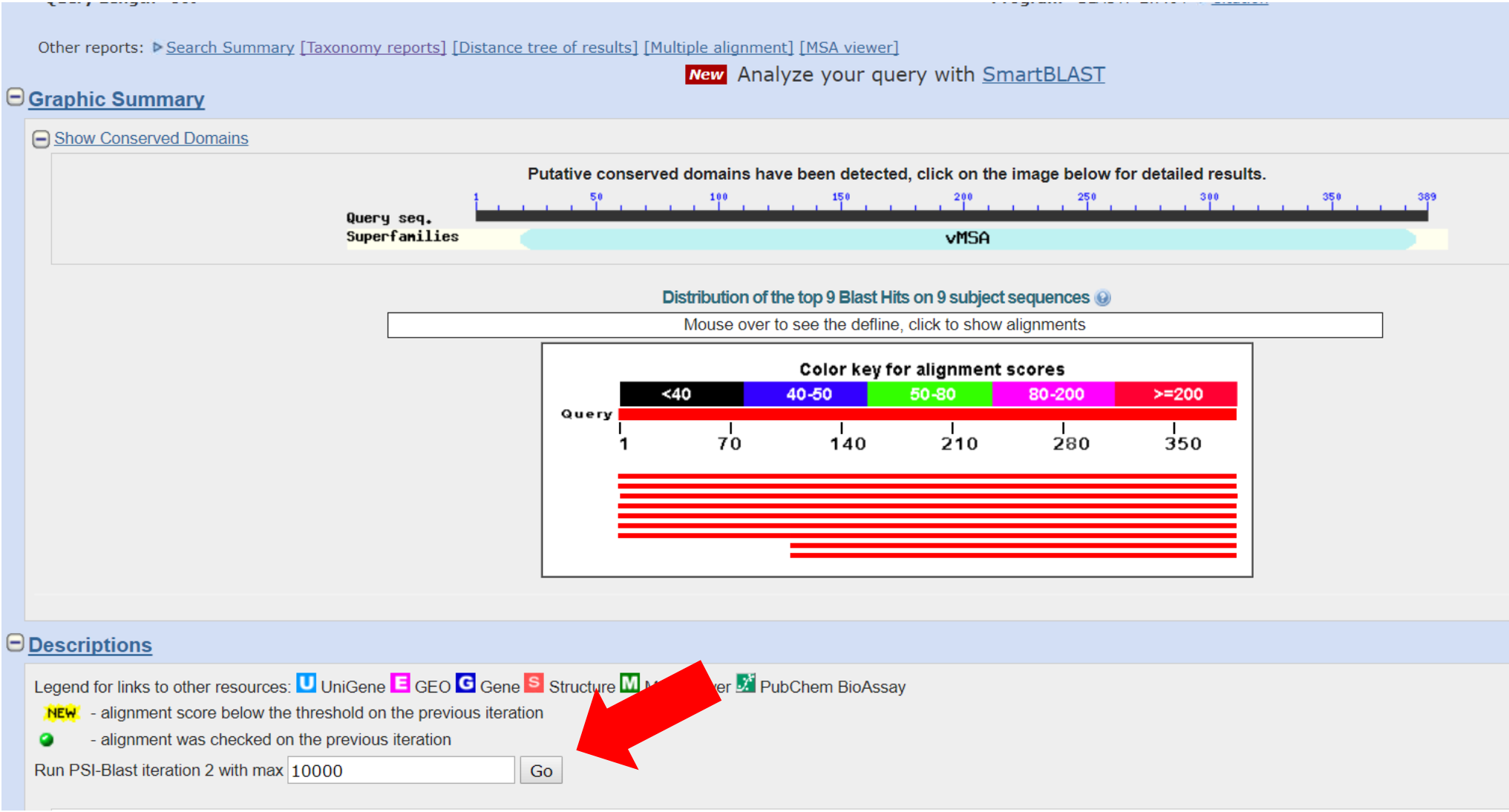
Iteration 1 of PSI-BLAST

1. Review the BLAST result from iteration 1 of PSI-BLAST
   → Go to Format options to change the parameters

2. Too many sequences
   → If you are interested in the human HBV mutations
   → If you are interested in a more broad functional features

3. Restrict hits from certain lineages, clades or species
   → What can you get?

Iteration 2 of PSI-BLAST



Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Ident	Links
AOG18115.1	PreS1/PreS2/S [Cloning vector pAAV2neo-1.3HBV(ayw)]	708	708	100%	0.0	99%	
ABB04016.1	PreS1 [synthetic construct]	708	708	100%	0.0	100%	
APA28776.1	large S protein [synthetic construct]	656	656	99%	0.0	91%	
AOG18108.1	PreS1/PreS2/S [Cloning vector pAAV2neo-1.3HBV(adr)]	646	646	100%	0.0	89%	
AAL13124.1	pre-S1/pre-S2 protein [synthetic construct]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
AAK58874.1	PreS1+PreS2+HBsAg [synthetic construct]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
ALV66577.1	large S protein [Heron hepatitis B virus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
AOG18116.1	PreS2/S [Cloning vector pAAV2neo-1.3HBV(ayw)]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
AOG18109.1	PreS2/S [Cloning vector pAAV2neo-1.3HBV(adr)]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AAT37474.1	Hepatitis B surface antigen/human papillomavirus EE7-FLAG hybrid protei	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AAT37475.1	Hepatitis B surface antigen/human papillomavirus EE7 (D1-35) hybrid pro	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW YP_009175035.1	preS1 surface protein [Woolly monkey hepatitis B Virus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	G
NEW AAC64339.1	surface antigen subtype ayw [Expression vector pUK-S]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW KRY25800.1	Large envelope protein [Trichinella spiralis]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AIT39699.1	DSV4 [synthetic construct]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW APA28777.1	S protein [synthetic construct]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AGH10171.1	surface antigen [Bat hepatitis virus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW WP_080445144.1	hypothetical protein [Bacillus cereus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AAM28153.1	surface antigen [synthetic construct]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AOG18110.1	S [Cloning vector pAAV2neo-1.3HBV(adr)]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AAL13120.1	S protein [synthetic construct]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW ALV66578.1	S protein [Heron hepatitis B virus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW ARM20234.1	surface antigen [Bat hepatitis virus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AAL62458.1	preS/S [Duck hepatitis B virus]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NEW AHC08491.1	S protein [synthetic construct]	357	357	58%	1e-120	90%	
NEW AGH10175.1	surface antigen [Bat hepatitis virus]	358	358	99%	2e-118	56%	
NEW YP_007678000.1	surface antigen [Bat hepatitis virus]	355	355	99%	5e-117	55%	G

Run PSI-Blast iteration 3 with max 10000 Go

Pay attention to:

1. New sequences are identified
   - Woolly monkey HBV
   - Bacillus cereus
   - Bat HBV
   - Duck HBV
2. E-value changed

Search reaches convergence in iteration 4

BLAST Results

Your results are being filtered to match entrez query: all [filter] NOT(txid10407 [ORGN]).

Edit and Resubmit Save Search Strategies ▸Formatting options ▸Download YouTube How to read this page Blast report c

PSI blast Iteration 4

Job title: (2) - Protein Sequence (389 letters)

RID XRWC3V1C015 (Expires on 10-11 06:00 am)

Query ID lcl|Query_175982
Description None
Molecule type amino acid
Query Length 389

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.7.0+ ▸Citation

No new sequences were found above the 0.005 threshold

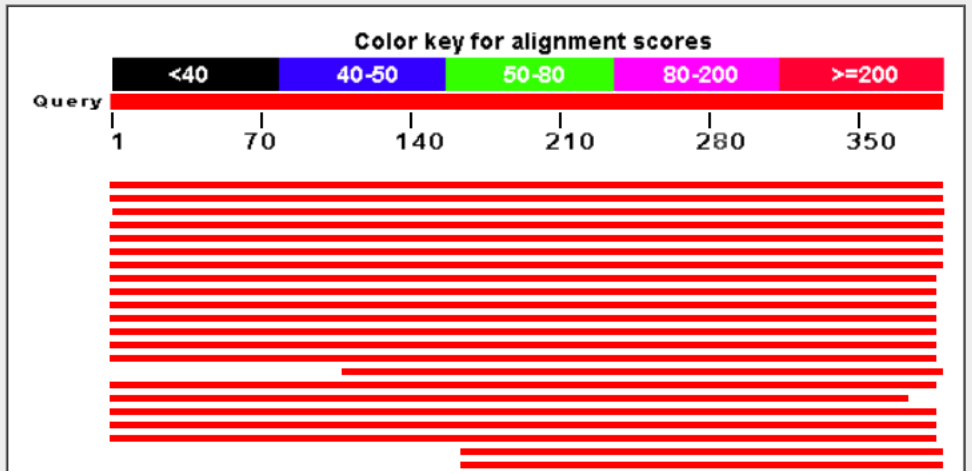
Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

Graphic Summary

Distribution of the top 34 Blast Hits on 34 subject sequences

Mouse-over to show defline and scores, click to show alignments



Now let’s use another query

```
>ALV66578.1 S protein [Heron hepatitis B virus]
MENTTSGFLGPLLVLQAGFFSLTRILTIPQSLDSWWTSLNFLGGAPTCPGQNSQSPTSNHSPTSCPPICP
GYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVCPLLPGTSTTSTGPCKTCTIPAQGTSMFPSCCCT
KPSDGNCTCIPIPSSWAFARFLWEWASVRFSWLSLLVPFVQWFVGLSPTVWLSVIWMMWYWGPSLYNILS
PFLPLLPIFFYLWVYI
```

Many more hits

NEW	WP_080445144.1	hypothetical protein [Bacillus cereus]	368	368	100%	1e-127	92%	
NEW	AAK58874.1	PreS1+PreS2+HBsAg [synthetic construct]	374	374	100%	4e-127	92%	
NEW	APA28776.1	large S protein [synthetic construct]	373	373	100%	5e-127	91%	
NEW	AAT37474.1	Hepatitis B surface antigen/human papillomavirus EE7-FLAG hybrid protei	368	368	100%	1e-125	92%	
NEW	AAT37475.1	Hepatitis B surface antigen/human papillomavirus EE7 (D1-35) hybrid pro	365	365	100%	2e-125	92%	
NEW	AAC64339.1	surface antigen subtype ayw [Expression vector pUK-S]	362	362	100%	5e-125	92%	
NEW	AOG18116.1	PreS2/S [Cloning vector pAAV2neo-1.3HBV(ayw)]	364	364	100%	5e-125	92%	
NEW	AAL62458.1	preS/S [Duck hepatitis B virus]	360	360	100%	2e-124	93%	
NEW	ABB04016.1	PreS1 [synthetic construct]	366	366	100%	3e-124	92%	
NEW	AOG18115.1	PreS1/PreS2/S [Cloning vector pAAV2neo-1.3HBV(ayw)]	365	365	100%	4e-124	92%	
NEW	YP_009045992.1	surface protein [Roundleaf bat hepatitis B virus]	269	269	100%	2e-88	68%	G
NEW	YP_009175036.1	surface protein [Woolly monkey hepatitis B Virus]	269	269	92%	2e-88	77%	G
NEW	AGH10171.1	surface antigen [Bat hepatitis virus]	275	275	98%	2e-88	65%	
NEW	YP_009175035.1	preS1 surface protein [Woolly monkey hepatitis B Virus]	271	271	92%	6e-87	77%	G
NEW	YP_009045996.1	surface protein [Horseshoe bat hepatitis B virus]	260	260	100%	8e-85	65%	G
NEW	AQT40957.1	surface protein [Horseshoe bat hepatitis B virus]	262	262	100%	4e-83	65%	
NEW	AGH10175.1	surface antigen [Bat hepatitis virus]	258	258	98%	7e-82	64%	
NEW	YP_007678000.1	surface antigen [Bat hepatitis virus]	257	257	98%	2e-81	65%	G
NEW	ARM20226.1	surface antigen [Bat hepatitis virus]	256	256	98%	6e-81	65%	
NEW	ARM20234.1	surface antigen [Bat hepatitis virus]	256	256	98%	7e-81	65%	
NEW	ARM20230.1	surface antigen [Bat hepatitis virus]	256	256	98%	9e-81	64%	
NEW	AIW47284.1	surface antigen [Bat hepatitis virus]	254	254	98%	7e-80	65%	
NEW	ARM20218.1	surface antigen [Bat hepatitis virus]	252	252	100%	3e-79	62%	
NEW	NP_955537.1	surface antigen [Ground squirrel hepatitis virus]	239	239	93%	2e-76	63%	G
NEW	NP_040995.1	hypothetical protein [Ground squirrel hepatitis virus]	239	239	93%	6e-76	63%	G
NEW	AAB08035.1	small envelope protein [Arctic ground squirrel hepatitis B virus]	235	235	93%	6e-75	64%	
NEW	AAB08034.1	middle envelope protein [Arctic ground squirrel hepatitis B virus]	235	235	93%	2e-74	64%	
NEW	ABI31620.1	major surface antigen [Woodchuck hepatitis virus]	233	233	93%	3e-74	62%	
NEW	P03144.1	RecName: Full=Large envelope protein; AltName: Full=L glycoprotein; Alt	239	239	93%	8e-74	63%	
NEW	AGW01291.1	surface protein [Tent-making bat hepatitis B virus]	232	232	99%	1e-73	59%	
NEW	YP_009046000.1	surface protein [Tent-making bat hepatitis B virus]	231	231	99%	1e-73	59%	G
NEW	NP_944491.1	surface protein [Woodchuck hepatitis virus]	231	231	93%	1e-73	62%	G
NEW	AAA46775.1	surface protein [Woodchuck hepatitis virus]	230	230	93%	4e-73	61%	
NEW	AAA46773.1	surface antigen [Woodchuck hepatitis virus]	229	229	93%	6e-73	61%	

...Duck hepatitis B virus isolate DHBVQCA34	viruses	47.4	1	Duck hepatitis B virus isolate DHBVQCA34 hits
...Duck hepatitis B virus strain China	viruses	47.4	1	Duck hepatitis B virus strain China hits
...Stork hepatitis B virus	viruses	47.0	4	Stork hepatitis B virus hits
...Hepatitis B virus duck/DHBV-16	viruses	45.8	1	Hepatitis B virus duck/DHBV-16 hits
...Duck hepatitis B virus brown Shanghai duck/S5	viruses	45.4	1	Duck hepatitis B virus brown Shanghai duck/S5 hits
..Roundleaf bat hepatitis B virus	viruses	269	5	Roundleaf bat hepatitis B virus hits
..Woolly monkey hepatitis B virus	viruses	269	6	Woolly monkey hepatitis B Virus hits
..Bat hepatitis virus	viruses	275	13	Bat hepatitis virus hits
..Hepatitis B virus Woolly monkey/Louisville	viruses	271	1	Hepatitis B virus Woolly monkey/Louisville hits
..Horseshoe bat hepatitis B virus	viruses	260	3	Horseshoe bat hepatitis B virus hits
..Ground squirrel hepatitis virus	viruses	239	4	[illegible]
..Arctic ground squirrel hepatitis B virus	viruses	235	3	[illegible]
..Woodchuck hepatitis virus	viruses	233	34	[illegible]
..Tent-making bat hepatitis B virus	viruses	232	5	Tent-making bat hepatitis B virus hits
..Woodchuck hepatitis virus w64 (ISOLATE PWS23)	viruses	229	1	Woodchuck hepatitis virus w64 (ISOLATE PWS23) hits
..Woodchuck hepatitis virus 7	viruses	231	1	Woodchuck hepatitis virus 7 hits
..Woodchuck hepatitis virus 59	viruses	231	1	Woodchuck hepatitis virus 59 hits
..Woodchuck hepatitis virus 8	viruses	231	1	Woodchuck hepatitis virus 8 hits
..Woodchuck hepatitis virus 1	viruses	229	1	Woodchuck hepatitis virus 1 hits
..Woodchuck hepatitis virus 2	viruses	225	1	Woodchuck hepatitis virus 2 hits
..Bluegill hepadnavirus	viruses	107	2	Bluegill hepadnavirus hits
..Ross's goose hepatitis B virus	viruses	66.2	6	Ross's goose hepatitis B virus hits
..White sucker hepatitis B virus	viruses	66.2	2	White sucker hepatitis B virus hits
..Snow goose hepatitis B virus	viruses	52.4	12	Snow goose hepatitis B virus hits
..Sheldgoose hepatitis B virus	viruses	51.2	6	Sheldgoose hepatitis B virus hits
..Tibetan frog hepadnavirus	viruses	51.2	4	Tibetan frog hepadnavirus hits

Check the taxonomy report

Search converged at Iteration 4

BLAST Results

Your search is limited to records that exclude: Human hepatitis B virus (taxid:10407) ▹Full Entrez Query

Edit and Resubmit Save Search Strategies ▹Formatting options ▹Download YouTube How to read this page Blast report description

PSI blast Iteration 4

Job title: (2) - ALV66578:S protein [Heron hepatitis B virus]

RID XS57YPUF015 (Expires on 10-11 08:31 am)

Query ID ALV66578.1
Description S protein [Heron hepatitis B virus]
Molecule type amino acid
Query Length 226

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.7.0+ ▹Citation

No new sequences were found above the 0.005 threshold

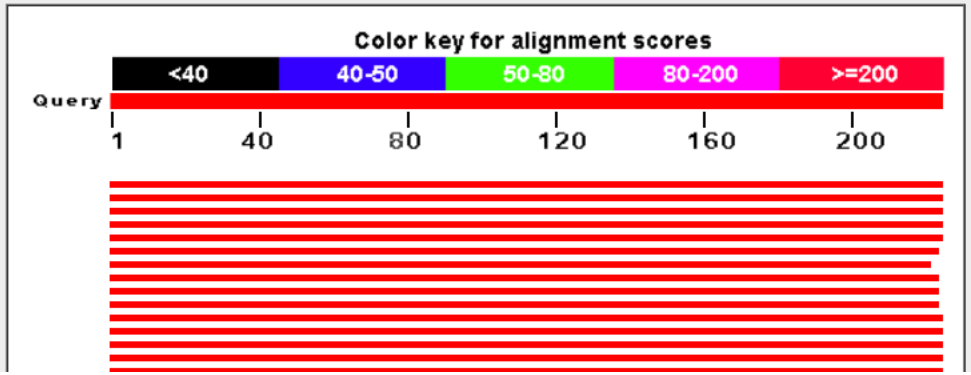
Other reports: ▹Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

Graphic Summary

Distribution of the top 100 Blast Hits on 199 subject sequences

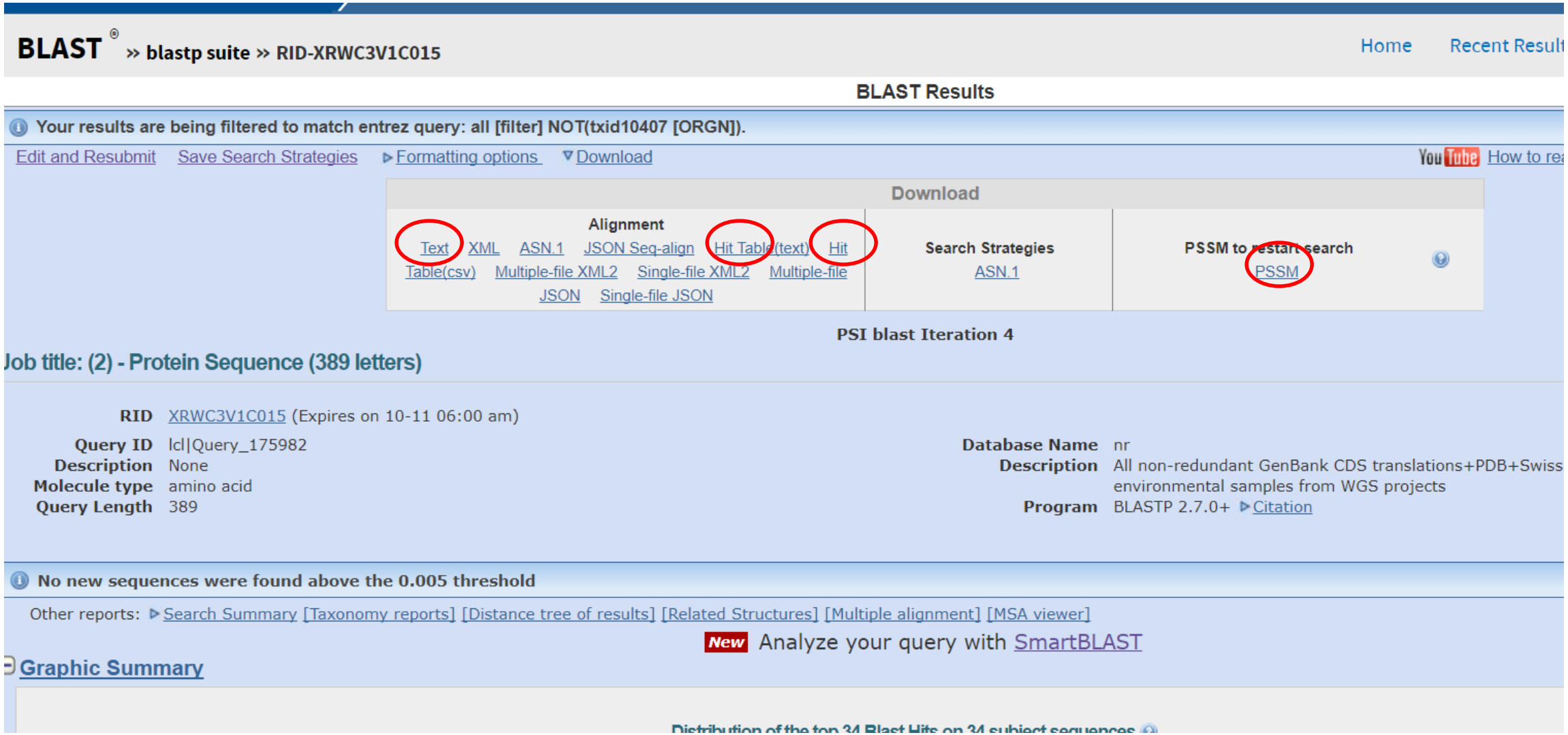
Mouse over to see the defline, click to show alignments



PSI-BLAST searches (HBV Surface protein)

- Heron hepatitis B virus
- Bat hepatitis B virus
- Woodchuck hepatitis virus
- Ground squirrel hepatitis B virus
- Bluegill hepadnavirus
- Tibetan frog hepadnavirus
- Duck hepatitis B virus
- Woolly monkey hepatitis B Virus
- Snow goose hepatitis B virus
- Trichinella spiralis

- Download the Text or Hit Table files



From sequence ids to fasta format

- Put the sequence id list to a text file in Unix
  cat>te
  paste your list
  Ctrl+D for ending terminal line input

- cat te | cut -f 1 -d ‘ ‘> HBVS.ac.list
  - (NOTE: Remove “prf||1803562A, prf||1305266A” in your file)

- cat HBVS.ac.list | epost -db protein -format acc | efetch -format fasta > HBVS.ac.fasta
  - Or use “grep -v '\|' HBVS.ac.list | epost -db protein -format acc | efetch -format fasta > HBVS.ac.fasta

- Or Batch Entrez
  - Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records
  - www.ncbi.nlm.nih.gov/sites/batchentrez

Multiple sequence alignment

- ClustalW or Omega: https://www.ebi.ac.uk/Tools/msa/clustalo/
- Kalign: https://www.ebi.ac.uk/Tools/msa/kalign/
- T-coffee: http://tcoffee.crg.cat/
- Muscle: https://www.ebi.ac.uk/Tools/msa/muscle/
- Promals: http://prodata.swmed.edu/promals/promals.php

BLASTClust to remove the highly-similar sequences

- BLASTClust is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences. The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster. In the case of proteins, the blastp algorithm is used to compute the pairwise matches; in the case of nucleotide sequences, the Megablast algorithm is used.

- blastclust –help

- cat HBVS.ac.fasta | blastclust –S 1.75 –L .9 > HBVS.bc

- echo "$(tail -n +2 te)" | cut -f 1 -d ' ' | less (remove the first line + get the first column ids)

- Generate multiple sequence alignment with these sequences
- Clustal Omega
- Promals3D:
  - Advantage: profile-profile comparison and secondary structural information
  - Note: PSI-BLAST iteration number set to 1
  - http://prodata.swmed.edu/promals3d/getResult2.php?name=QUERY_CBIuhw&email=empty_email&target_name=

Multiple sequence alignment editors

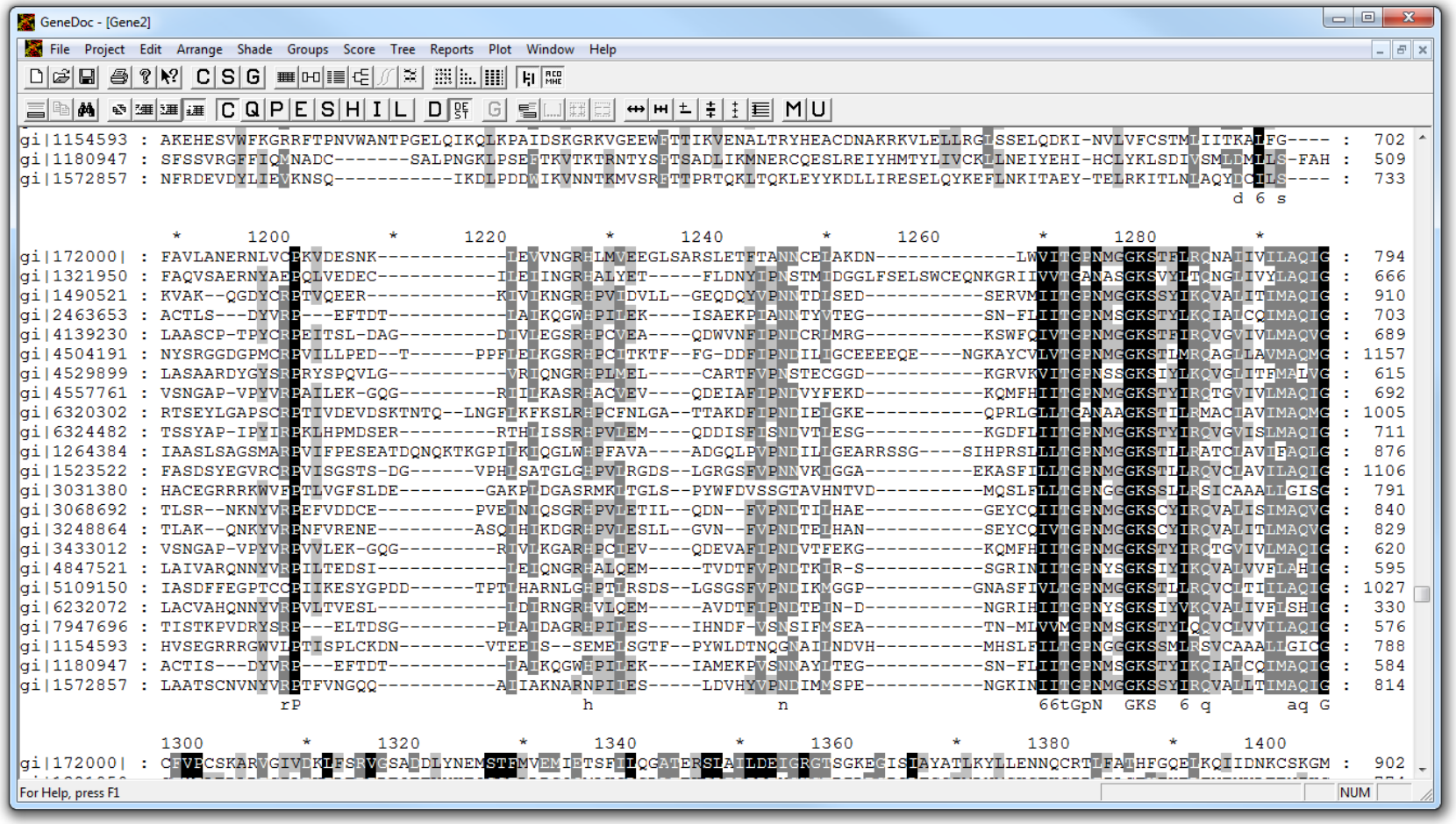
- BioEdit - MS-Windows
- Genedoc - MS-Windows
- EditSeq/MegAlign - Lasergene - Mac or MS-Windows
- DNA Strider - Macintosh
- Seq-Al - Macintosh
- ASAD - Excel - Macintosh or MS-Windows
- Chroma - Windows
- SeqPup - Mac. MS-Windows, X-Windows
- AliView - Mac. MS-Windows, Linux
- Boxshade: www.ch.embnet.org/software/BOX_form.html

MSA-Visualization and improvement

- GeneDoc (Windows)
- Download: http://genedoc.software.informer.com/download/
  - Arranging and Editing
    - GeneDoc's Grab and Drag arrangement mode allows you to move residues around like beads on a string
  - Shading Alignments
  - Reports: Stats, Score, Composition
  - Exporting and Copying Figures

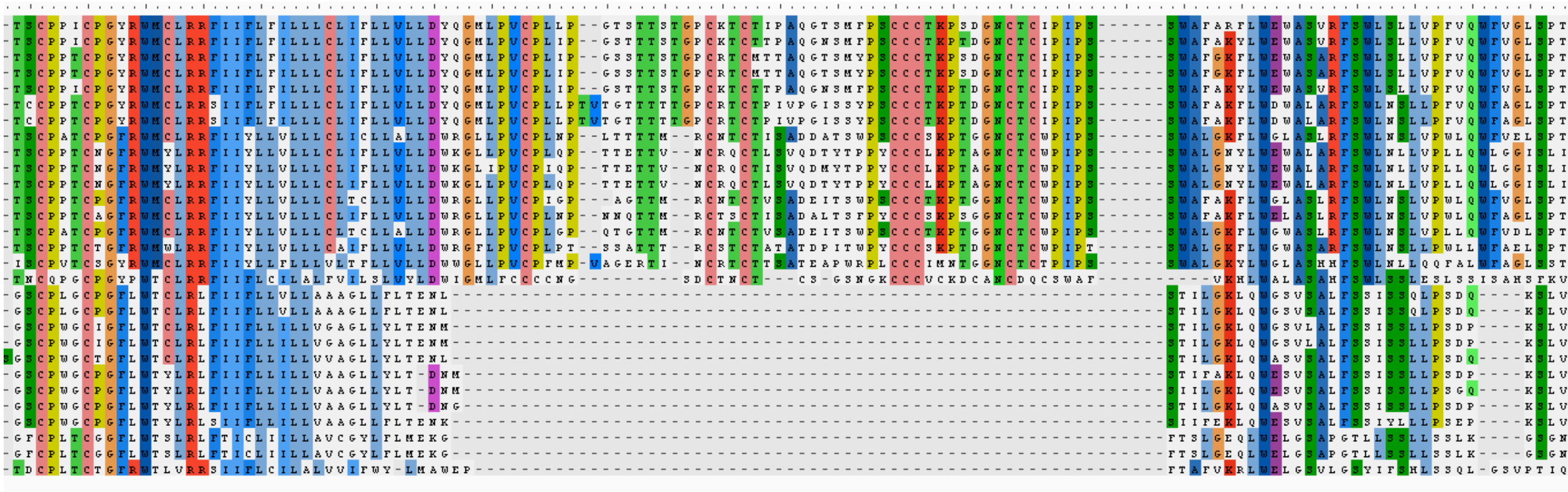
GeneDoc: Conservation Mode

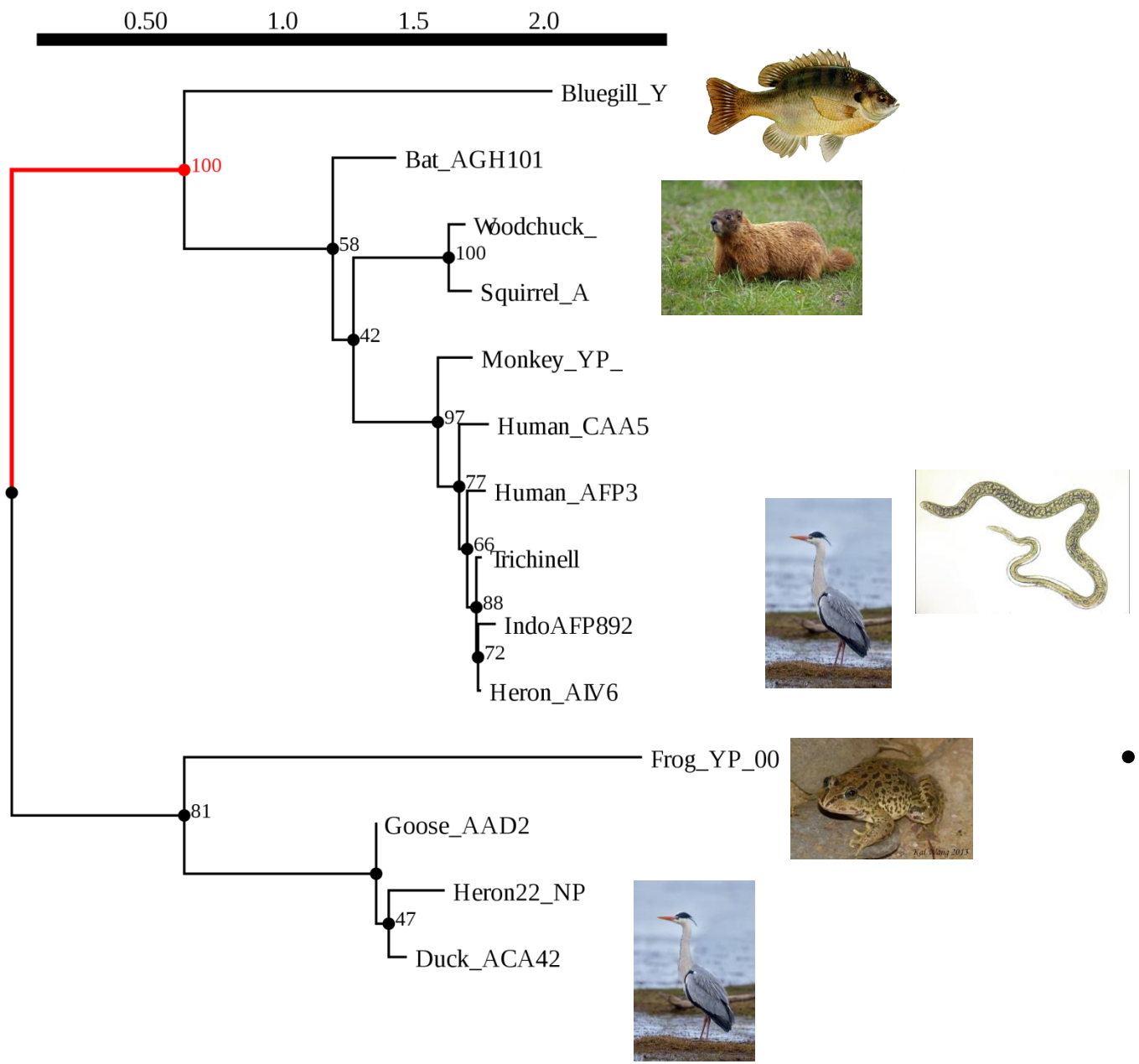
- GeneDoc (Windows): use import to load alignment file



GeneDoc: Property Mode

AliView alignment of HBV-S





- Problem:
  - Trichinella spiralis (pork worm)
  - Heron/Human?
  - Root

Our prediction on HBV surface protein

- A protein family can be found in many viruses found in mammals, bird, frog and even fish
- HBV-S has 4 TM region
- Conserved residues:
  - Conserved hydrophobic Trp (W) for TM
  - a Cys-rich domain between TM2-TM3
- The Cys-rich domain insert event during evolution
- Gene transfers

- Potential function:
  - The insert domain has multiple conserved Cys residues which potentially form disulfide bonds
  - Peptide hormone like function?
  - will contribute to binding host receptor during infection

ClustalW

```
CLUSTAL W (1.83) multiple sequence alignment

beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                            :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean       --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice          --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                .       . ..  *  .::        :           :              :

beta globin   NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin     YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin   SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean       QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice          HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
               :  :   ::  :        :      * .     .   :
```

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

Praline

(a) Praline multiple sequence alignment

```
beta globin      ..........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin        ...........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin      .............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          ..........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency      0000000000142654382579345734633643436244536864 33*35344*50063

beta globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..HCDKLH....VDP
myoglobin        HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL
soybean          A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD
rice             R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD
Consistency      3166354224776653*4368635424454451335634333542003335440000922

beta globin      ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH......
myoglobin        KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean          PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA........
rice             AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency      437448444982585423053365544 54*5546542644675432200 1000
```

Note also the changing pattern of gaps within the boxed region in these five different alignments.

MUSCLE

(b)

```
MUSCLE (3.6) multiple sequence alignment

beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin   --------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                            :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
myoglobin     HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                            . ..  *  .::        :                  :    :

beta globin   NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
myoglobin     YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
soybean       QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA--------
rice          HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
               :  :   ::  :        :      * .     .   :
```

Probcons

(c)

```
PROBCONS

                                                                                  ▼
beta globin   M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin     M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin   M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
              *          * :   :   :   :  .. .    .      ::   *     *.
                                 ▽                          ▼
beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
              .    :    . ..  *  .::       ::   .           *.  *      :

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean       PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice          AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

TCoffee

(d)

```
CLUSTAL FORMAT for T-COFFEE Version_5.13

beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                            :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
              .         . ..  *  .::        :               *.  *

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean       Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice          H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

HMMER

- **HMMER3** is a package to build and use HMMs developed by Sean Eddy (http://hmmer.wustl.edu/).
- Software available in HMMER3:
  - hmmbuild to build an HMM from a multiple alignment;
  - hmmalign to align sequences to an HMM model;
  - hmmsearch to search a sequence database with an HMM model;
  - jackhmmer to iteratively search sequence(s) against a protein database;
  - hmmscan to search protein sequence(s) against a protein profile database;
  - hmmemit to get sample sequences from a profile HMM;
  - hmmfetch to retrieve profile HMM(s) from a file

HMMER installation

- Visit http://linuxbrew.sh
- Follow the install instructions

- brew tap homebrew/science
- brew update
- brew install hmmer

Pfam download

- mkdir ~/data

- cd ~/data
- wget -c -t0 ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz
- gunzip Pfam-A.hmm.gz
- hmmpress Pfam-A.hmm

- cat RHS.fasta | muscle > RHS.muscle
- more RHS.muscle

- Copy the result to a text file
- Open it using any MSA viewer

Let make hmm profile using this alignment:

- hmmbuild RHS.hmm RHS.muscle

Search Ecoli genome using this HMM model:

- hmmsearch –E 1 RHS.hmm Ecoli.O157.fa > RHS.Ecoli.hmm.result
- Using more to see the result

Get the hit id and fasta sequences

- hmmsearch –E 1 RHS.hmm Ecoli.O157.fa | grep '>>' | cut -f 2 -d ' '|epost -db protein -format acc | efetch -format fasta > RHS.Ecoli.fasta

Identifying pfam domains in these sequences

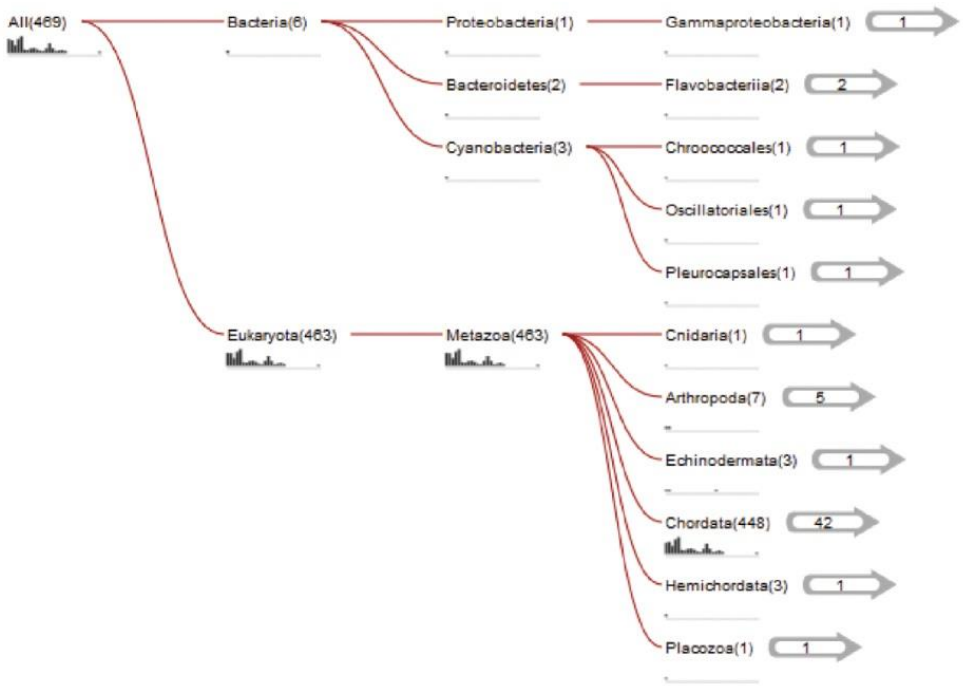
- hmmscan ~/data/Pfam-A.hmm RHS.Ecoli.fasta > RHS.Ecoli.pfam.result

HMMER is available online

(a) HMMER web output

	Target	Description	Species	E-value
>	301781474	PREDICTED: hypothetical protein LOC100482027	Ailuropoda melanoleuca	1.1e-154
>	4504349	hemoglobin subunit beta	Homo sapiens	3.7e-97
>	55635219	PREDICTED: hemoglobin subunit beta isoform 2	Pan troglodytes	3.7e-97
>	397496509	PREDICTED: hemoglobin subunit beta-like	Pan paniscus	3.7e-97
>	297689478	PREDICTED: hemoglobin subunit beta-like	Pongo abelii	3.5e-95

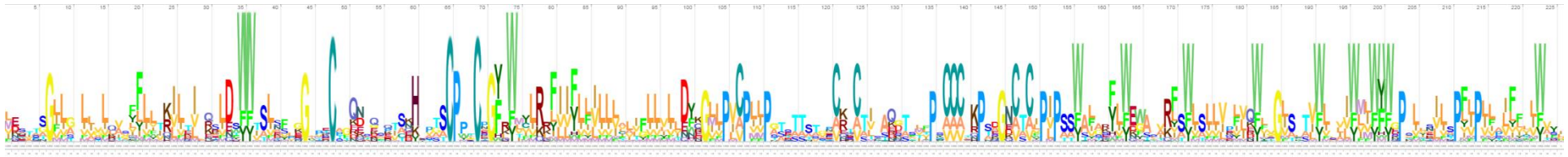
(b) HMMER phylogenetic output



- Domain architecture
  - Taxonomy
- Iterative manner

Surface protein (HBsAg)

- HMMER searches:
  - HBV S protein
  - Ground Squirrel hepatitis virus S protein
  - Duck hepatitis B virus S protein



HMMER software: build profiles, complement BLAST

Build a profile HMM (input is a multiple sequence alignment)

```
$ ./hmmbuild -h # provides brief help documentation
$ ./hmmbuild globins4.hmm ../tutorial/globins4.sto
```

Download a database to search (e.g. human RefSeq proteins)

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein
.faa.gz
$ gunzip human.protein.faa.gz
$ wc -l human.protein.faa
302761 human.protein.faa
```

Search an HMM against a database

```
$ ./hmmsearch globins4.hmm human.protein.faa > globins4.out
```

HMMER results

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  globins4.hmm
# target sequence database:        /mnt/reference/human.protein.faa
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---
    E-value  score  bias    Sequence                        Description
    ------- ------ -----    --------                        -----------
    3.3e-64  216.6   0.0    ref|NP_000509.1|     hemoglobin subunit beta [Homo sa
      7e-61  205.8   0.0    ref|NP_000510.1|     hemoglobin subunit delta [Homo s
    2.3e-60  204.2   1.3    ref|NP_000508.1|     hemoglobin subunit alpha [Homo s
    2.3e-60  204.2   1.3    ref|NP_000549.1|     hemoglobin subunit alpha [Homo s
    6.2e-60  202.8   0.3    ref|NP_976311.1|     myoglobin [Homo sapiens]
    6.2e-60  202.8   0.3    ref|NP_976312.1|     myoglobin [Homo sapiens]
    6.2e-60  202.8   0.3    ref|NP_005359.1|     myoglobin [Homo sapiens]
    4.8e-55  186.9   0.0    ref|NP_000175.1|     hemoglobin subunit gamma-2 [Homo
    1.4e-54  185.4   0.4    ref|NP_005321.1|     hemoglobin subunit epsilon [Homo
    2.1e-54  184.8   0.1    ref|NP_000550.2|     hemoglobin subunit gamma-1 [Homo
    4.9e-48  164.2   0.2    ref|NP_005323.1|     hemoglobin subunit zeta [Homo sa
    1.7e-40  139.7   0.1    ref|NP_005322.1|     hemoglobin subunit theta-1 [Homo
    1.8e-39  136.4   0.2    ref|NP_599030.1|     cytoglobin [Homo sapiens]
      5e-35  121.9   0.3    ref|NP_001003938.1|  hemoglobin subunit mu [Homo sapi
      3e-08   35.0   0.0    ref|NP_067080.1|     neuroglobin [Homo sapiens]
  ------ inclusion threshold ------
       0.14   13.4   0.0    ref|NP_001371.1|     dedicator of cytokinesis protein
       0.25   12.6   0.8    ref|NP_006737.2|     sex comb on midleg-like protein
       0.28   12.4   0.8    ref|NP_001032629.1|  sex comb on midleg-like protein
```

HMMER output includes scores, E values