Computer lab exercises #8

Comments on projects worth sharing:

1. Use BLINK whenever possible. It can save a lot of waiting and greatly accelerates explorations.

From a protein sequence entry in NCBI select "BLINK" under related information. (You might need to scroll down, in case the upper tables are expanded).

BLINK provides a GUI interface to pre-computed BLASTP searches.

Display Settings: GenPept

Send to:

tumor necrosis factor alpha [Homo sapiens]

GenBank: ACO37640.1

FASTA Graphics

Go to:

```
LOCUS       ACO37640                232 aa            linear   PRI 06-APR-2009
DEFINITION  tumor necrosis factor alpha, partial [Homo sapiens].
ACCESSION   ACO37640
VERSION     ACO37640.1  GI:226201421
DBSOURCE    accession FJ795028.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 232)
  AUTHORS   Guan,W.J., Ma,Y.H., Yu,L.L., Na,R.S. and Liu,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-FEB-2009) Academy of Agricultural Sciences, Institute
            of Animal Science, Quanmingyuan West, Beijing 100193, People's
            Republic of China
FEATURES             Location/Qualifiers
     source          1..232
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6p21.3"
                     /sex="male"
                     /tissue_type="placenta"
                     /country="China"
                     /collection_date="2008"
     Protein         <1..232
                     /product="tumor necrosis factor alpha"
                     /name="APC1 protein"
     Region          87..230
                     /region_name="TNF"
                     /note="Tumor Necrosis Factor; TNF superfamily members
                     include the cytokines: TNF (TNF-alpha), LT
                     (lymphotoxin-alpha, TNF-beta), CD40 ligand, Apo2L (TRAIL),
                     Fas ligand, and osteoprotegerin (OPG) ligand. These
```

Change region shown

Customize view

Analyze this sequence

Protein 3D Structure

Articles about the TNF gene

Pathways for the TNF gene

Reference sequence information

More about the TNF gene

Homologs of the TNF gene

LinkOut to external resources

Related information

BLink

Related Sequences

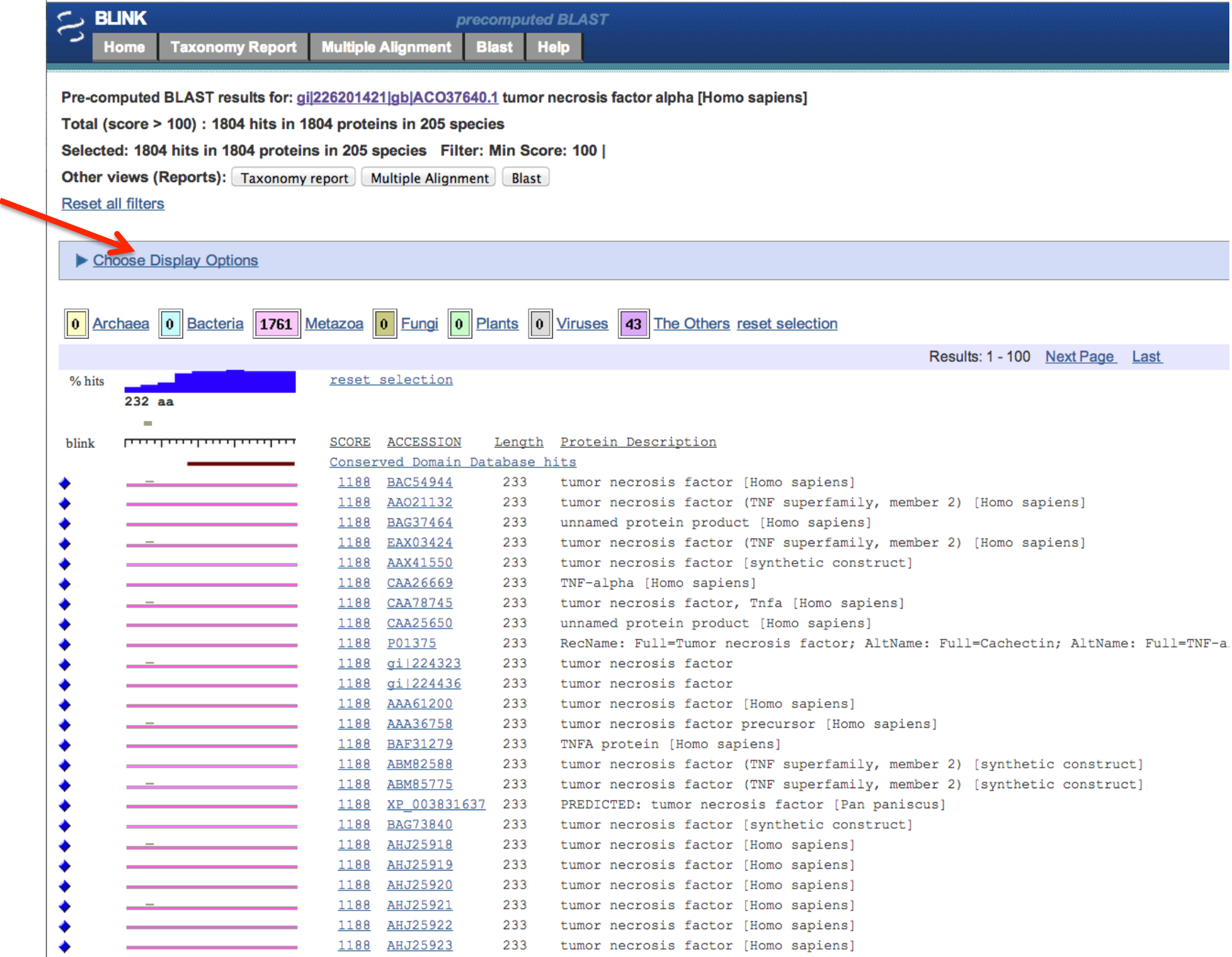
BioSystems

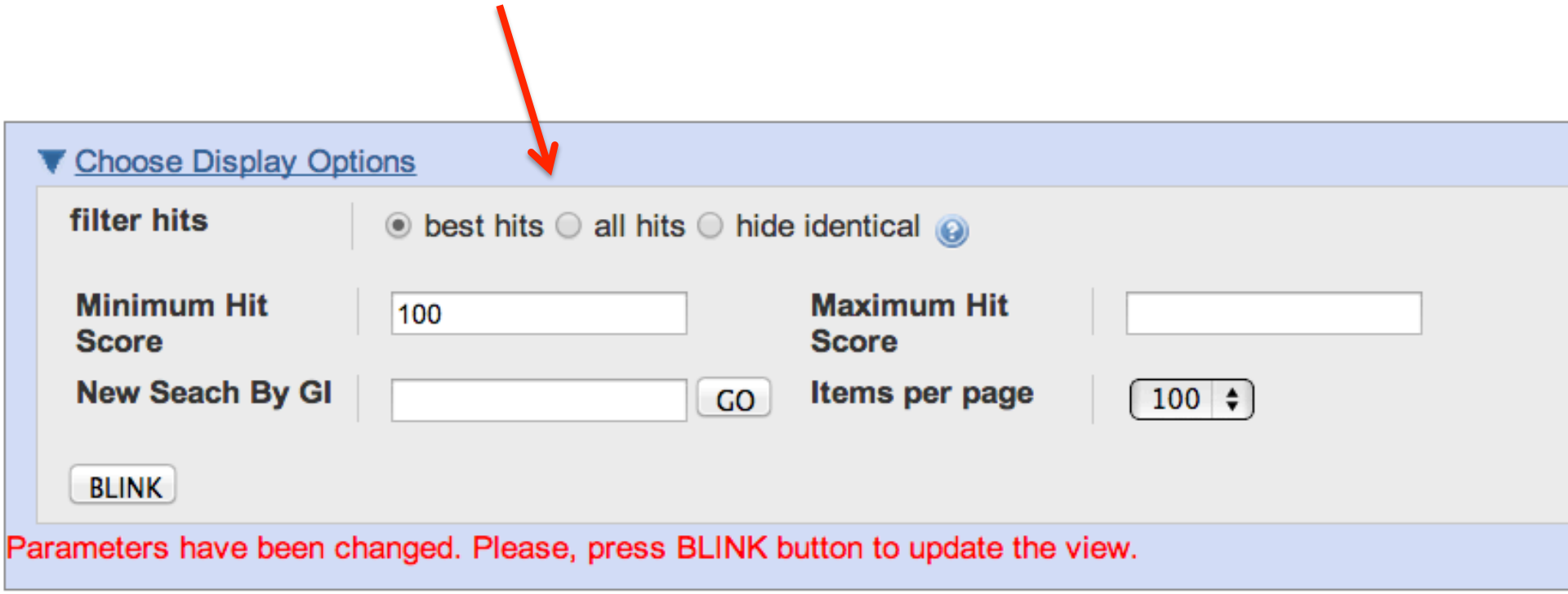
CDD Search Results

Conserved Domains (Concise)

Conserved Domains (Full)

Domain Relatives


BLINK
precomputed BLAST
Home
Taxonomy Report
Multiple Alignment
Blast
Help
Pre-computed BLAST results for: gi|226201421|gb|ACO37640.1 tumor necrosis factor alpha [Homo sapiens]
Total (score > 100) : 1804 hits in 1804 proteins in 205 species
Selected: 1804 hits in 1804 proteins in 205 species Filter: Min Score: 100 |
Other views (Reports): Taxonomy report Multiple Alignment Blast
Reset all filters
Choose Display Options
0 Archaea 0 Bacteria 1761 Metazoa 0 Fungi 0 Plants 0 Viruses 43 The Others reset selection
Results: 1 - 100 Next Page Last
% hits
232 aa
reset selection
blink
SCORE ACCESSION Length Protein Description
Conserved Domain Database hits
1188 BAC54944 233 tumor necrosis factor [Homo sapiens]
1188 AAO21132 233 tumor necrosis factor (TNF superfamily, member 2) [Homo sapiens]
1188 BAG37464 233 unnamed protein product [Homo sapiens]
1188 EAX03424 233 tumor necrosis factor (TNF superfamily, member 2) [Homo sapiens]
1188 AAX41550 233 tumor necrosis factor [synthetic construct]
1188 CAA26669 233 TNF-alpha [Homo sapiens]
1188 CAA78745 233 tumor necrosis factor, Tnfa [Homo sapiens]
1188 CAA25650 233 unnamed protein product [Homo sapiens]
1188 P01375 233 RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltName: Full=TNF-a
1188 gi|224323 233 tumor necrosis factor
1188 gi|224436 233 tumor necrosis factor
1188 AAA61200 233 tumor necrosis factor [Homo sapiens]
1188 AAA36758 233 tumor necrosis factor precursor [Homo sapiens]
1188 BAF31279 233 TNFA protein [Homo sapiens]
1188 ABM82588 233 tumor necrosis factor (TNF superfamily, member 2) [synthetic construct]
1188 ABM85775 233 tumor necrosis factor (TNF superfamily, member 2) [synthetic construct]
1188 XP_003831637 233 PREDICTED: tumor necrosis factor [Pan paniscus]
1188 BAG73840 233 tumor necrosis factor [synthetic construct]
1188 AHJ25918 233 tumor necrosis factor [Homo sapiens]
1188 AHJ25919 233 tumor necrosis factor [Homo sapiens]
1188 AHJ25920 233 tumor necrosis factor [Homo sapiens]
1188 AHJ25921 233 tumor necrosis factor [Homo sapiens]
1188 AHJ25922 233 tumor necrosis factor [Homo sapiens]
1188 AHJ25923 233 tumor necrosis factor [Homo sapiens]


Choose Display Options
filter hits
best hits
all hits
hide identical
Minimum Hit Score
100
Maximum Hit Score
New Seach By GI
GO
Items per page
100
BLINK
Parameters have been changed. Please, press BLINK button to update the view.

BLINK *precomputed BLAST*

Home | Taxonomy Report | Multiple Alignment | Blast | Help

Pre-computed BLAST results for: gi|226201421|gb|ACO37640.1 tumor necrosis factor alpha [Homo sapiens]

Total (score > 100) : 1804 hits in 1804 proteins in 205 species

Selected: 1804 hits in 1804 proteins in 205 species Filter: Min Score: 100 |

Other views (Reports): Taxonomy report | Multiple Alignment | Blast

Reset all filters

▶ Choose Display Options

0 Archaea | 0 Bacteria | 1761 Metazoa | 0 Fungi | 0 Plants | 0 Viruses | 43 The Others reset selection

Results: 1 - 100 Next Page Last

% hits

232 aa

reset selection

blink

SCORE	ACCESSION	Length	Protein Description
	Conserved Domain Database hits		
1188	BAC54944	233	tumor necrosis factor [Homo sapiens]
1188	AAO21132	233	tumor necrosis factor (TNF superfamily, member 2) [Homo sapiens]
1188	BAG37464	233	unnamed protein product [Homo sapiens]
1188	EAX03424	233	tumor necrosis factor (TNF superfamily, member 2) [Homo sapiens]
1188	AAX41550	233	tumor necrosis factor [synthetic construct]
1188	CAA26669	233	TNF-alpha [Homo sapiens]
1188	CAA78745	233	tumor necrosis factor, Tnfa [Homo sapiens]
1188	CAA25650	233	unnamed protein product [Homo sapiens]
1188	P01375	233	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltName: Full=TNF-a
1188	gi|224323	233	tumor necrosis factor
1188	gi|224436	233	tumor necrosis factor
1188	AAA61200	233	tumor necrosis factor [Homo sapiens]
1188	AAA36758	233	tumor necrosis factor precursor [Homo sapiens]
1188	BAF31279	233	TNFA protein [Homo sapiens]
1188	ABM82588	233	tumor necrosis factor (TNF superfamily, member 2) [synthetic construct]
1188	ABM85775	233	tumor necrosis factor (TNF superfamily, member 2) [synthetic construct]
1188	XP_003831637	233	PREDICTED: tumor necrosis factor [Pan paniscus]
1188	BAG73840	233	tumor necrosis factor [synthetic construct]
1188	AHJ25918	233	tumor necrosis factor [Homo sapiens]
1188	AHJ25919	233	tumor necrosis factor [Homo sapiens]
1188	AHJ25920	233	tumor necrosis factor [Homo sapiens]
1188	AHJ25921	233	tumor necrosis factor [Homo sapiens]
1188	AHJ25922	233	tumor necrosis factor [Homo sapiens]
1188	AHJ25923	233	tumor necrosis factor [Homo sapiens]

Lineage Report

root
. Amniota [vertebrates]
. . Mammalia [mammals]
. . . Theria [mammals]
. . . . Eutheria [placentals]
. . . . . Boreoeutheria [placentals]
. . . . . . Euarchontoglires [placentals]
. . . . Primates [primates]
. . . . Simiiformes [primates]
. . . . Catarrhini [primates]
. . . . Hominoidea [primates]
. . . . Hominidae [primates]
. . . . Homininae [primates]

. . . . Homo sapiens (man)	1188	16 hits	[primates]	tumor necrosis factor [Homo sapiens]
. . . . Gorilla gorilla gorilla (lowland gorilla)	1182	1 hit	[primates]	PREDICTED: tumor necrosis factor [Gorilla gorilla gorilla]
. . . . Pan troglodytes	777	1 hit	[primates]	tumor necrosis factor [Pan troglodytes]
. . . . Gorilla gorilla (gorilla)	775	1 hit	[primates]	tumor necrosis factor [Gorilla gorilla]
. . . . Pongo abelii (orang utan)	1153	1 hit	[primates]	PREDICTED: tumor necrosis factor [Pongo abelii]
. . . . Pongo pygmaeus (orang utan)	764	1 hit	[primates]	tumor necrosis factor [Pongo pygmaeus]
. . . . Nomascus leucogenys (White-cheeked Gibbon)	1150	1 hit	[primates]	PREDICTED: LOW QUALITY PROTEIN: tumor necrosis factor [Nomascus leuc
. . . . Papio sp.	1169	1 hit	[primates]	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltNam
. . . . Macaca mulatta (rhesus monkeys)	1143	14 hits	[primates]	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltNam
. . . . Papio ursinus (baboon)	1139	1 hit	[primates]	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltNam
. . . . Chlorocebus sabaeus	1139	1 hit	[primates]	tumor necrosis factor alpha [Chlorocebus sabaeus]
. . . . Macaca fascicularis (long-tailed macaque)	1131	2 hits	[primates]	tumor necrosis factor precursor [Macaca fascicularis]
. . . . Cercocebus atys	1130	1 hit	[primates]	tumor necrosis factor alpha [Cercocebus atys]
. . . . Papio anubis (baboon)	1119	1 hit	[primates]	tumor necrosis factor alpha precursor [Papio anubis]
. . . . Macaca sp.	761	1 hit	[primates]	tumor necrosis factor [Macaca sp.]
. . . . Callithrix jacchus (white ear-tufted marmoset)	1034	1 hit	[primates]	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltNam
. . . . Saimiri boliviensis boliviensis	993	1 hit	[primates]	PREDICTED: tumor necrosis factor [Saimiri boliviensis boliviensis]
. . . . Saimiri sciureus (South American squirre...)	988	1 hit	[primates]	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin; AltNam
. . . . Aotus nancymaae	775	1 hit	[primates]	tumor necrosis factor alpha [Aotus nancymaae]

...

. . . . Macropus eugenii	830	2 hits	[marsupials]	tumour necrosis factor alpha [Macropus eugenii]
. . . . Monodelphis domestica	758	1 hit	[marsupials]	PREDICTED: tumor necrosis factor-like isoform 1 [Monodelphis
. . . . Macropus rufus	751	1 hit	[marsupials]	tumor necrosis factor precursor [Macropus rufus]
. . . . Trichosurus vulpecula (silver-gray brushtail ...)	671	1 hit	[marsupials]	RecName: Full=Tumor necrosis factor; AltName: Full=Cachectin
. . . . Didelphis virginiana (Virginia opossum)	604	1 hit	[marsupials]	tumor necrosis factor precursor [Didelphis virginiana]
. . . Ornithorhynchus anatinus (duckbill platypus)	622	1 hit	[monotremes]	Tnf [Ornithorhynchus anatinus]
. . Chrysemys picta bellii	445	1 hit	[turtles]	PREDICTED: tumor necrosis factor [Chrysemys picta bellii]
. synthetic construct	1188	3 hits	[other sequences]	tumor necrosis factor [synthetic construct]

Pre-computed BLAST results for: gi|25952111|ref|NP_000585.2 tumor necrosis factor [Homo sapiens]

Matching gis: 27544420;27802685;510158303;363836827;189054614;118767860;119623829;375725640;383266295;380291495

Total (score > 100) : 1017 hits in 1017 proteins in 186 species

Selected: 1017 hits in 1017 proteins in 186 species Filter: Min Score: 100 | Best hits |

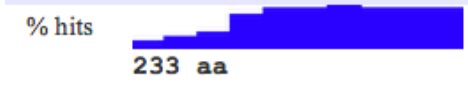
Other views (Reports): Taxonomy report | Multiple Alignment | Blast

Reset all filters

▶ Choose Display Options

0 Archaea | 0 Bacteria | 1002 Metazoa | 0 Fungi | 0 Plants | 0 Viruses | 15 The Others reset selection

Results: First Previous Page 101 - 186

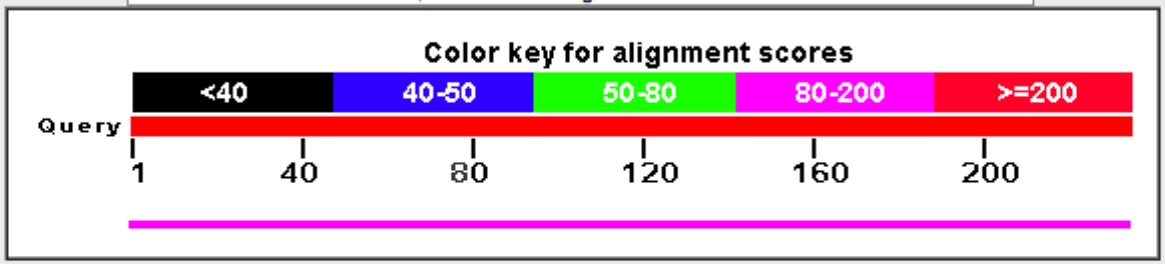


reset selection

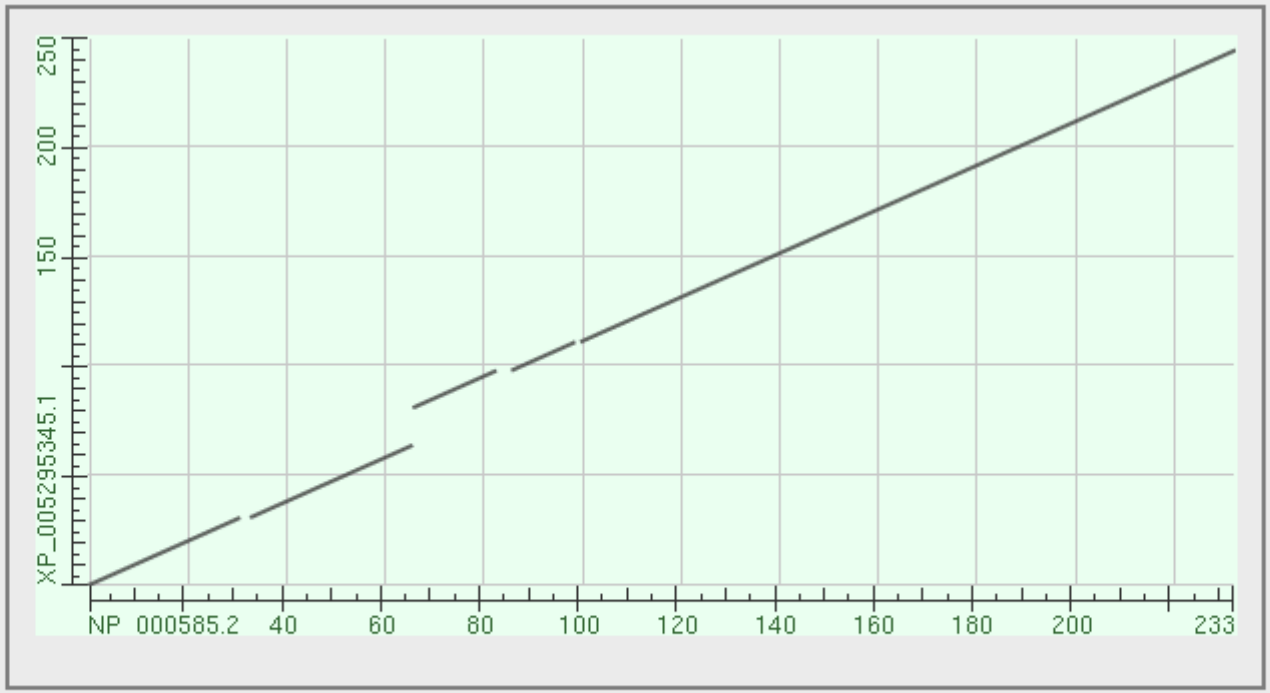
SCORE	ACCESSION	N	Tax
Conserved Domain Database hits			
484	CAD80055	1	Elephas maximus
478	CAD80061	1	Trichys fasciculata
466	CAD80060	1	Castor canadensis
462	CAD80059	1	Anomalurus sp. T-1787
460	CAD80057	1	Dipus sagitta
450	XP_005295345	10	Chrysemys picta bellii
423	CAD80058	1	Dipodomys merriami
392	XP_003218070	7	Anolis carolinensis
390	ABC88246	1	Phodopus sungorus
285	NP_001108250	3	Xenopus laevis
274	BAB68749	1	Mus musculus brevirostris
266	NP_001107143	11	Xenopus (Silurana) tropicalis
258	AAP94278	1	Acanthopagrus schlegelii
257	AGQ17907	1	Ginglymostoma cirratum

Distribution of 1 Blast Hits on the Query Sequence

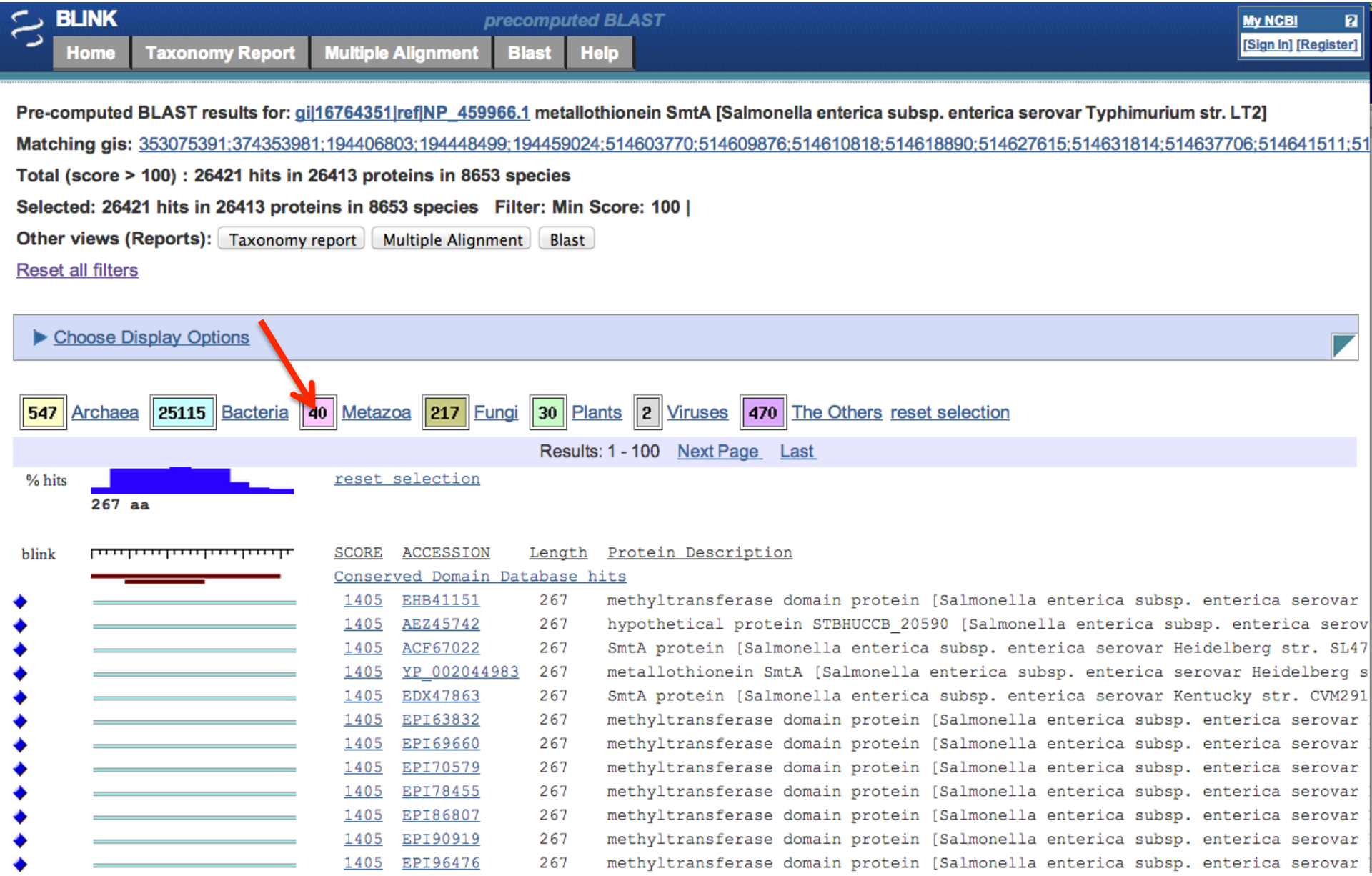
Mouse over to see the defline, click to show alignments



Plot of gi|25952111|ref|NP_000585.2| vs gi|530607321|ref|XP_005295345.1|



If one is looking for homologs in other phyla or domains,



Pre-computed BLAST results for: gi|16764351|ref|NP_459966.1 **metallothionein SmtA [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]**

Matching gis: 353075391;374353981;194406803;194448499;194459024;514603770;514609876;514610818;514618890;514627615;514631814;514637706;514641511;51

Total (score > 100) : 26421 hits in 26413 proteins in 8653 species

Selected: 40 hits in 40 proteins in 27 species Filter: Min Score: 100 | Included taxons: Metazoa; |

Other views (Reports): Taxonomy report | Multiple Alignment | Blast

Reset all filters

▶ Choose Display Options

Archaea | Bacteria | 40 Metazoa | Fungi | Plants | Viruses | The Others | reset selection

Results: 1 - 40

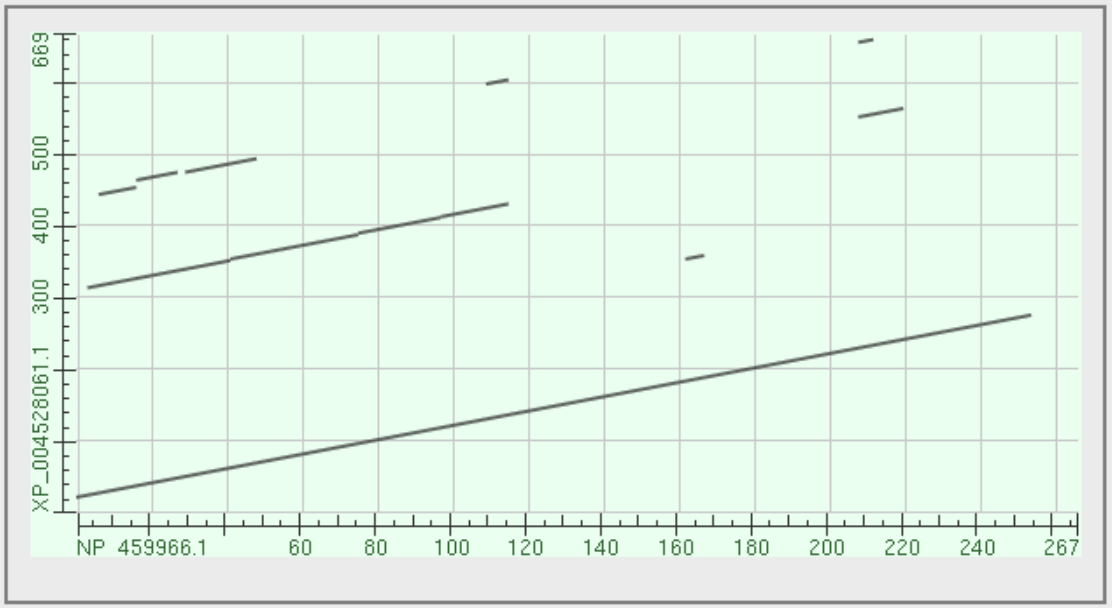
% hits

267 aa

reset selection

blink

SCORE	ACCESSION	Length	Protein Description
	Conserved Domain Database hits		
1109	XP_004528061	669	PREDICTED: chromosome partition protein MukF-like [Ceratitis capitata]
269			
935	GAA57985	361	3-demethylubiquinone-9 3-methyltransferase, partial [Clonorchis sinensis]
4430			
128	XP_004532210	249	PREDICTED: 3-demethylubiquinone-9 3-methyltransferase-like [Ceratitis capitata
467			
124	XP_001603088	289	PREDICTED: hexaprenyldihydroxybenzoate methyltransferase, mitochondrial-like [
207			
123	CAB16512	268	Protein COQ-3, isoform a [Caenorhabditis elegans]
123	NP_001041045	268	Protein COQ-3, isoform a [Caenorhabditis elegans]
116			



PREDICTED: chromosome partition protein MukF-like [Ceratitis capitata]

Sequence ID: ref|XP_004528061.1| **Length:** 669 **Number of Matches:** 7

Range 1: 23 to 276 GenPept Graphics ▼ Next Match ▲ Previous Match

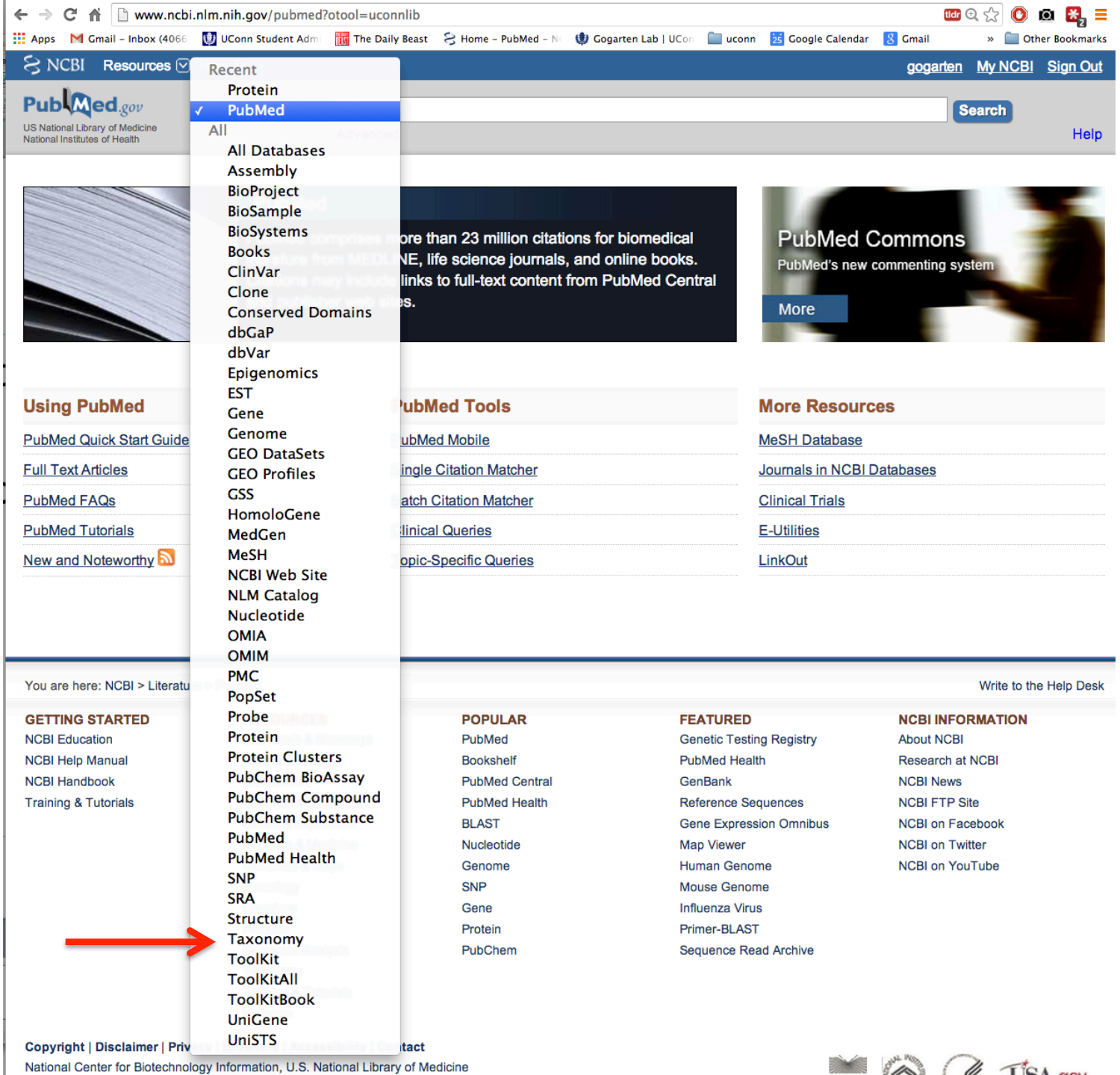
Score	Expect	Identities	Positives	Gaps
431 bits(1109)	1e-151	209/254(82%)	225/254(88%)	0/254(0%)

```
Query  1    MQDRNFDDIAEKFSRNIYGTTKGQLRQAILWQDLDRVLEEIGGRKLRVLDAGGGEGQTAI  60
            +QDRNFDDIAEKFSRNIYGTTKGQLRQAILWQDLDR+L   G   LR+LDAGGG GQTAI
Sbjct  23   VQDRNFDDIAEKFSRNIYGTTKGQLRQAILWQDLDRLLPACGAGPLRILDAGGGAGQTAI  82

Query  61   KMAERGHQVTLCDLSGEMIARARQAAEAKGVSKDMHFIQCPAQDVASHLESPVDLILFHA  120
            +MAERGH VTLCDLS EMIA A++AA+ KGVS  MHF+QC  QDVA HLESPVDLILFHA
Sbjct  83   QMAERGHHVTLCDLSSEMIALAKRAADEKGVSHRMHFVQCAIQDVAQHLESPVDLILFHA  142

Query  121  VLEWVADPVGVLETLWSVLRPGGALSLMFYNANGLLMHNMVAGNFDYVQAGMPKRKKRTL  180
            VLEWVA+P  VL+TLWS LRPGGALSLMFYNANGLL+HNMVA NFDYVQAGMPK+KKRTL
Sbjct  143  VLEWVAEPRTVLDTLWSTLRPGGALSLMFYNANGLLLHNMVATNFDYVQAGMPKKKKRTL  202

Query  181  SPDYPRDPAQVYQWLEAIGWQITGKTGVRVFHDYLREKHQQRDCYETLVELETRYCRQEP  240
            SPDYPRDP QVY WL+  GWQITGKTGVRVFHDYLREK +QRD Y  L+ELETRYCRQEP
Sbjct  203  SPDYPRDPQQVYGWLQEAGWQITGKTGVRVFHDYLREKEKQRDSYAQLLELETRYCRQEP  262

Query  241  YISLGRYIHVTAIK  254
            +ISLGRYIHVTA K
Sbjct  263  WISLGRYIHVTARK  276
```

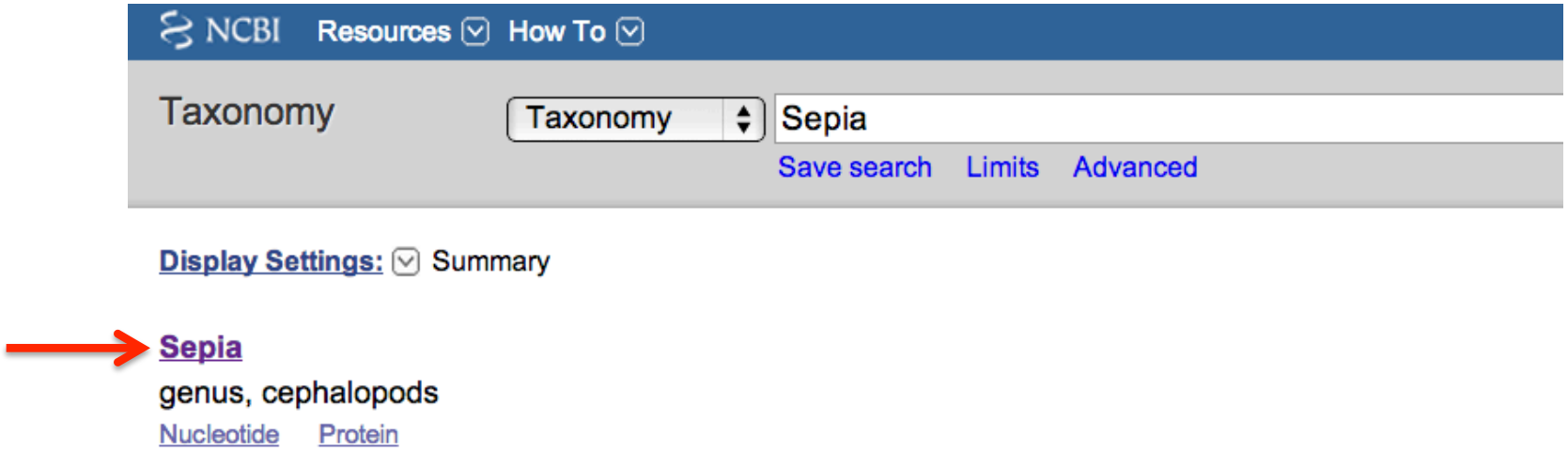
Comments on projects worth sharing:

2. The Taxonomy Browser at NCBI can facilitate finding genome and EST projects.

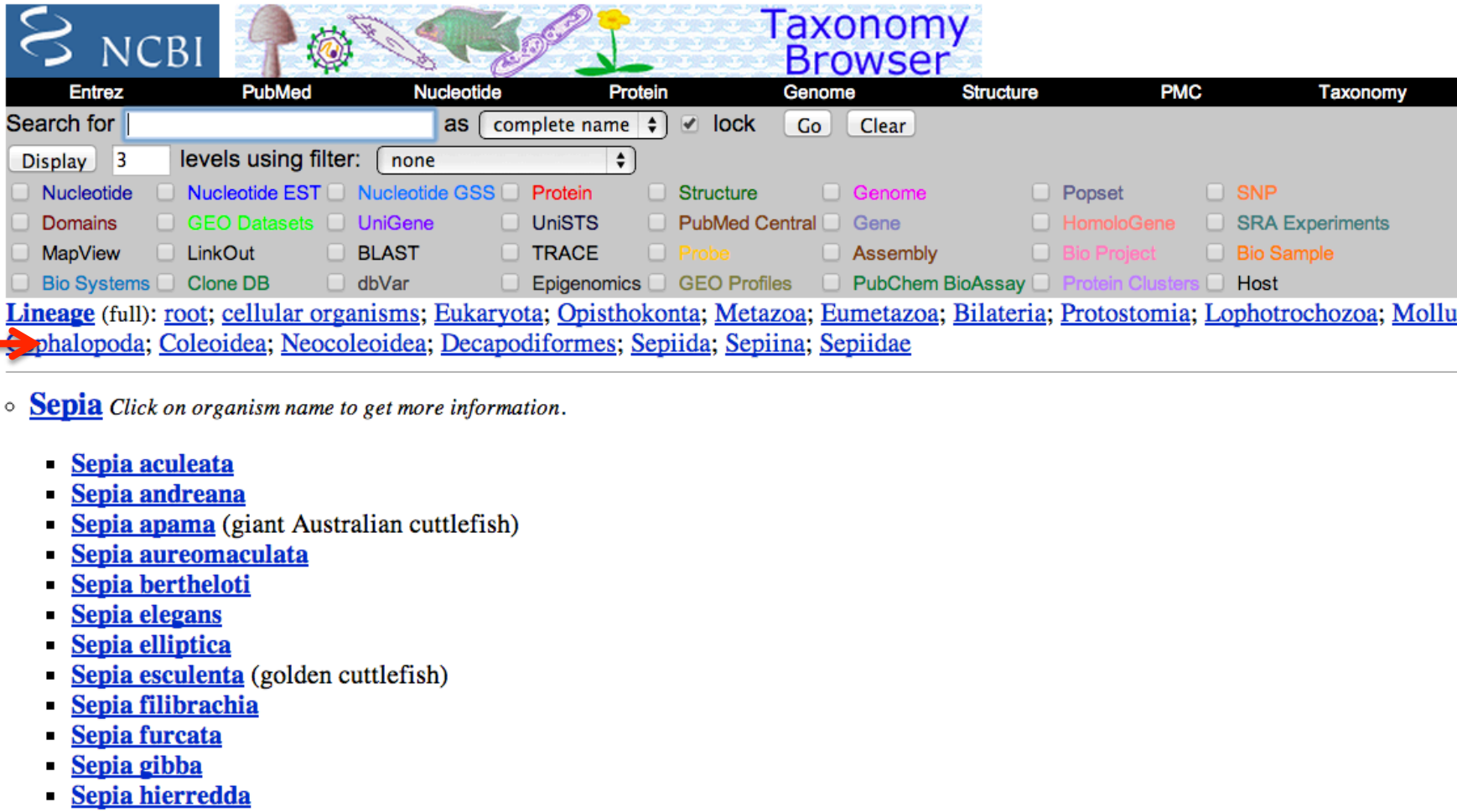
In entrez select taxonomy as databank


www.ncbi.nlm.nih.gov/pubmed?otool=uconnlib
Apps
Gmail - Inbox (4066
UConn Student Adm
The Daily Beast
Home - PubMed - N
Gogarten Lab | UCon
uconn
Google Calendar
Gmail
Other Bookmarks
NCBI
Resources
gogarten
My NCBI
Sign Out
PubMed.gov
US National Library of Medicine
National Institutes of Health
Search
Help
Recent
Protein
PubMed
All
All Databases
Assembly
BioProject
BioSample
BioSystems
Books
ClinVar
Clone
Conserved Domains
dbGaP
dbVar
Epigenomics
EST
Gene
Genome
GEO DataSets
GEO Profiles
GSS
HomoloGene
MedGen
MeSH
NCBI Web Site
NLM Catalog
Nucleotide
OMIA
OMIM
PMC
PopSet
Probe
Protein
Protein Clusters
PubChem BioAssay
PubChem Compound
PubChem Substance
PubMed
PubMed Health
SNP
SRA
Structure
Taxonomy
ToolKit
ToolKitAll
ToolKitBook
UniGene
UniSTS
ore than 23 million citations for biomedical
NE, life science journals, and online books.
links to full-text content from PubMed Central
s.
PubMed Commons
PubMed's new commenting system
More
Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy
ubMed Tools
ubMed Mobile
ingle Citation Matcher
atch Citation Matcher
linical Queries
opic-Specific Queries
More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities
LinkOut
You are here: NCBI > Literatu
Write to the Help Desk
GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Gene Expression Omnibus
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive
NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube
Copyright | Disclaimer | Priv
tact
National Center for Biotechnology Information, U.S. National Library of Medicine

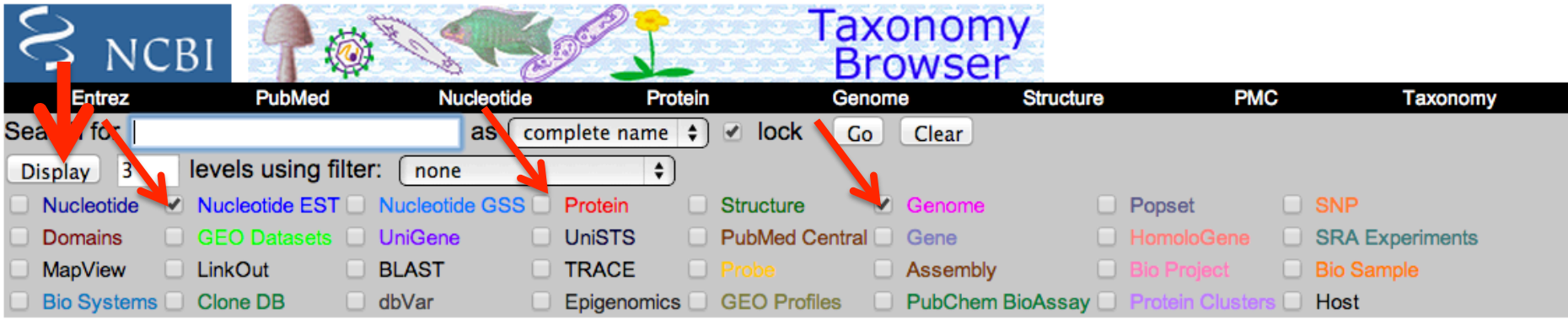
Search for an organisms you are interested in,



Select the group of organisms you are interested in



Check genomes and ESTs and whatever else you like, THEN click display



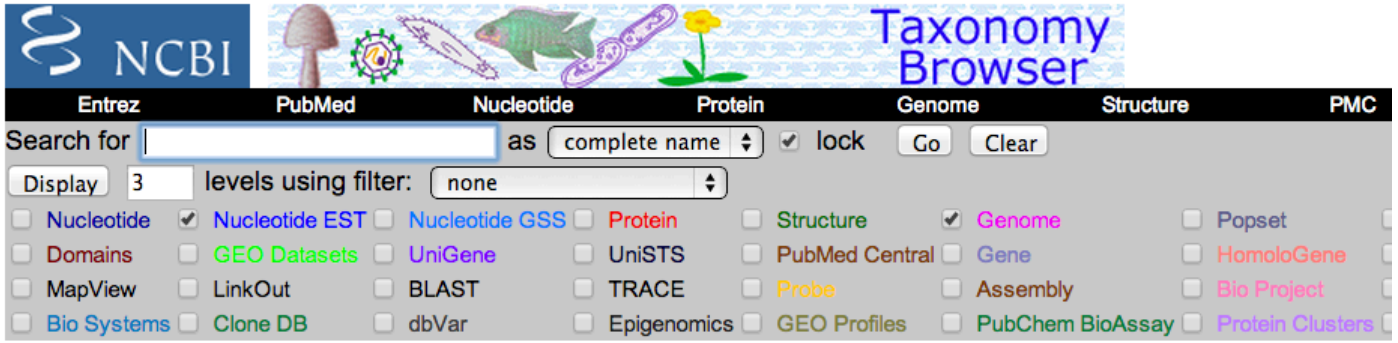
Lineage (full): root; cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Lophotrochozoa; Mollusc

---

- **Cephalopoda** (cephalopods) 114,034 36 *Click on organism name to get more information.*
  - **Coleoidea** 112,997 35
    - **Neocoleoidea** 112,997 35
      - **Decapodiformes** 112,994 24
      - **Octopodiformes** 3 11
  - **Nautiloidea** 1,037 1
    - **Nautilida** 1,037 1
      - **Nautilidae** 1,037 1
  - **environmental samples**
    - **Cephalopoda environmental sample**

Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant sc literature for the most reliable information.

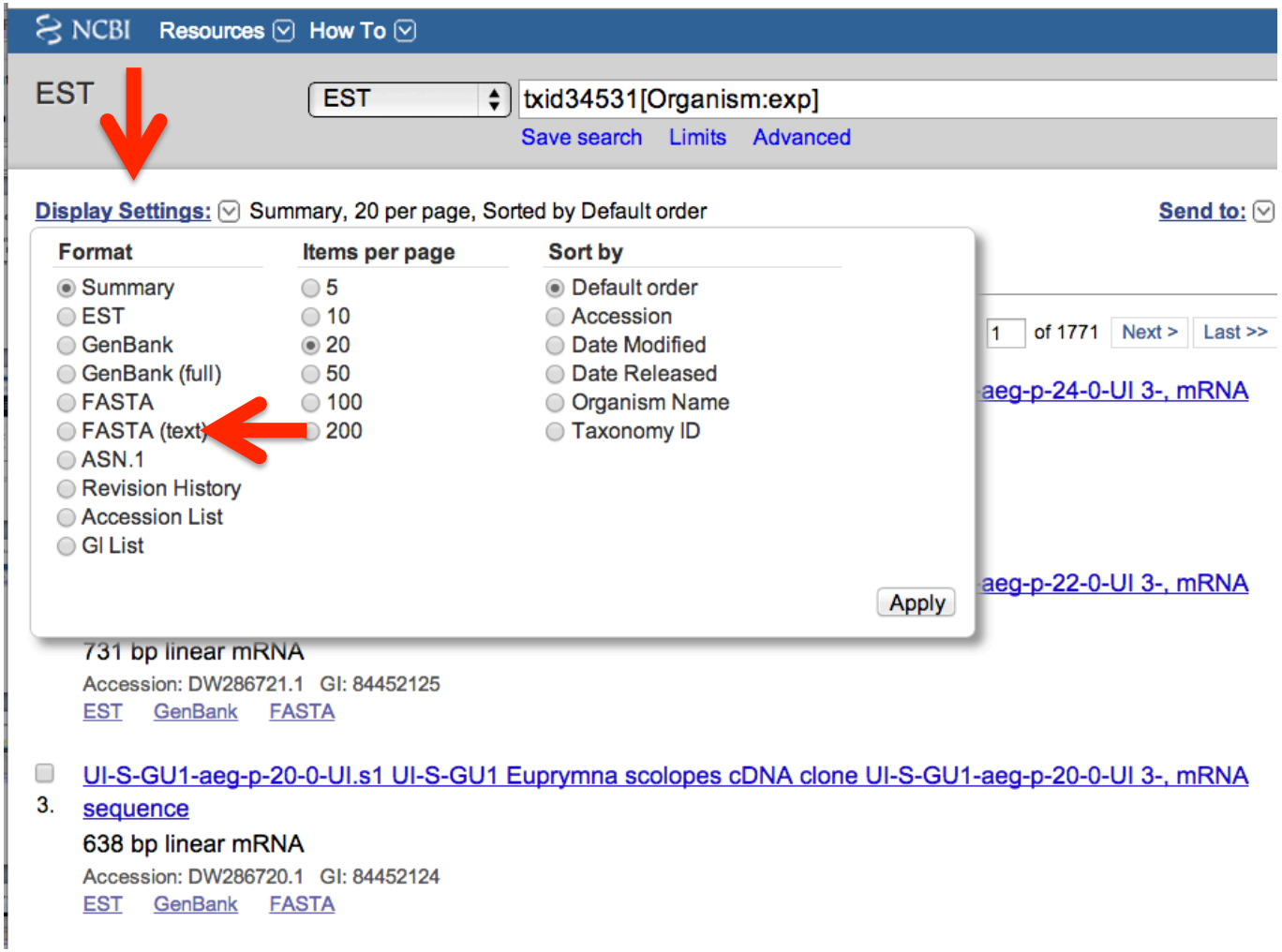
If you move up or down the taxonomic hierarchy, the checked items will be displayed for each Taxon



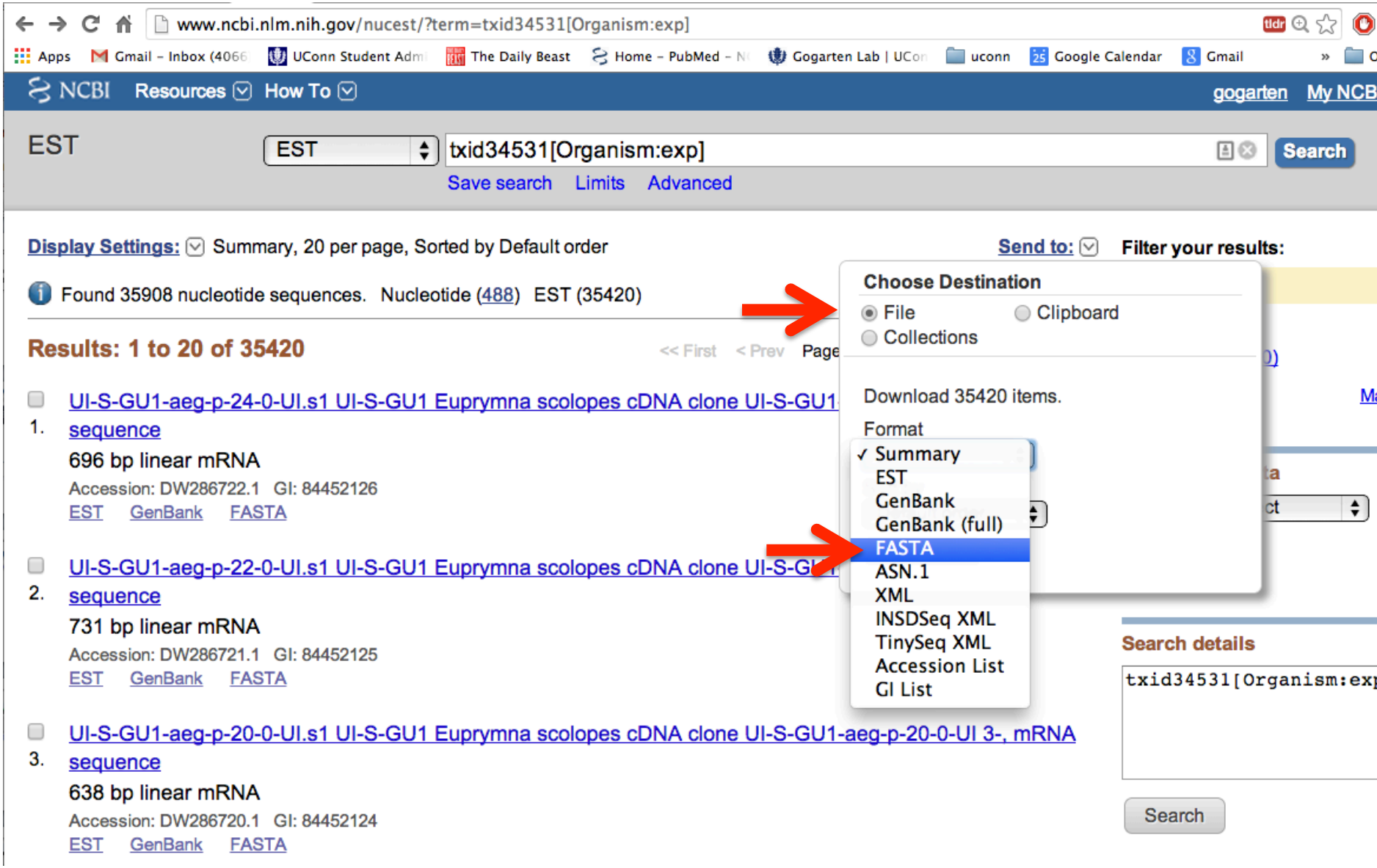
or

Retrieves all ESTs from Genbank

You can select format or send to file to get a multiple sequence fasta file with all sequences



You can select format or send to file to get a multiple sequence fasta file with all sequences



Short in class exercise

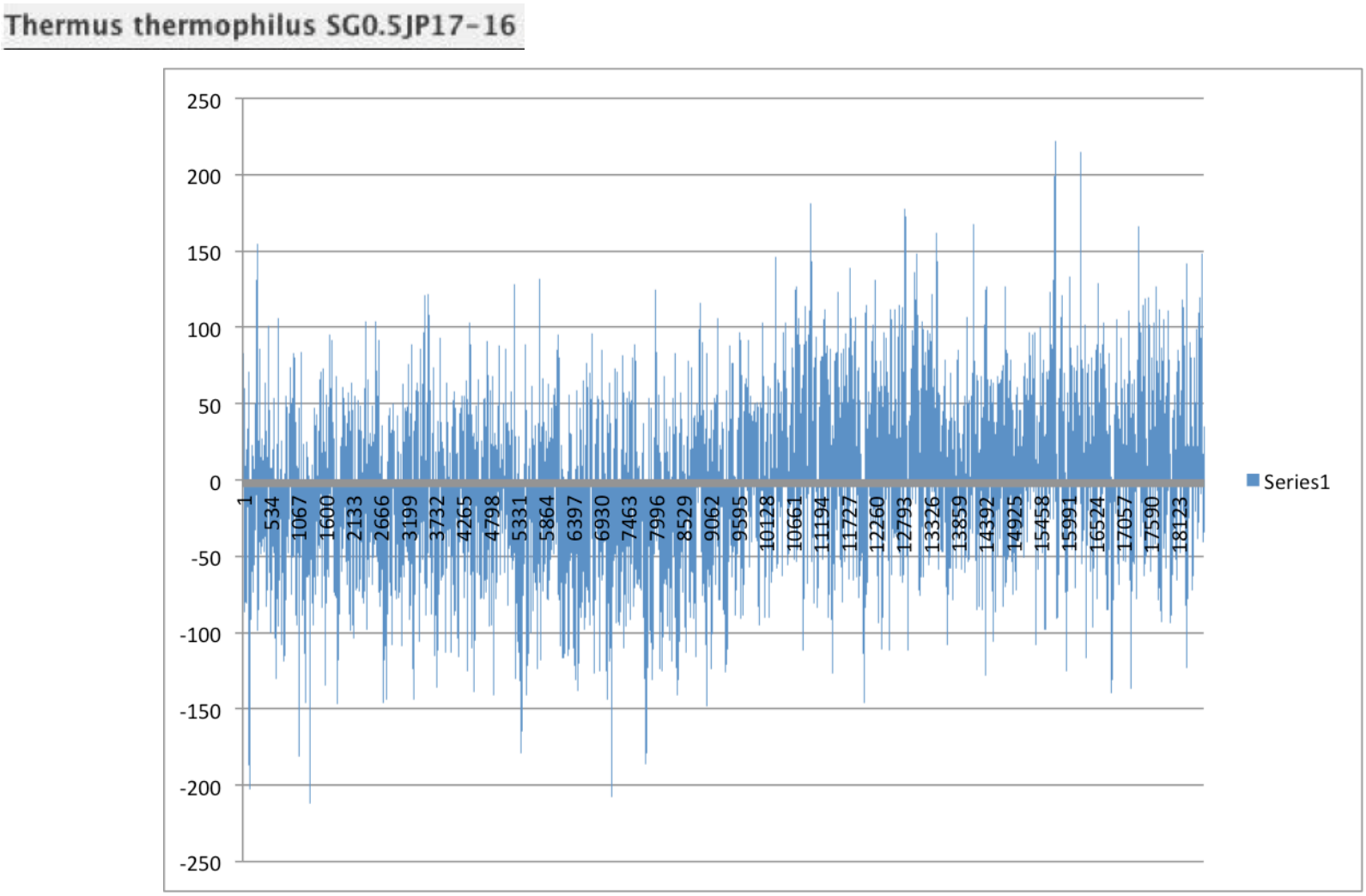
- Write and apply a script to calculate cumulative strand bias for a genome (fna file).

In DNA replication the two strands are not created equal.

See Drew Berry's Ted talk at https://www.youtube.com/watch?v=WFCvkkDSfIU for illustration (start at 3 min – 4.50 min, if not much time)

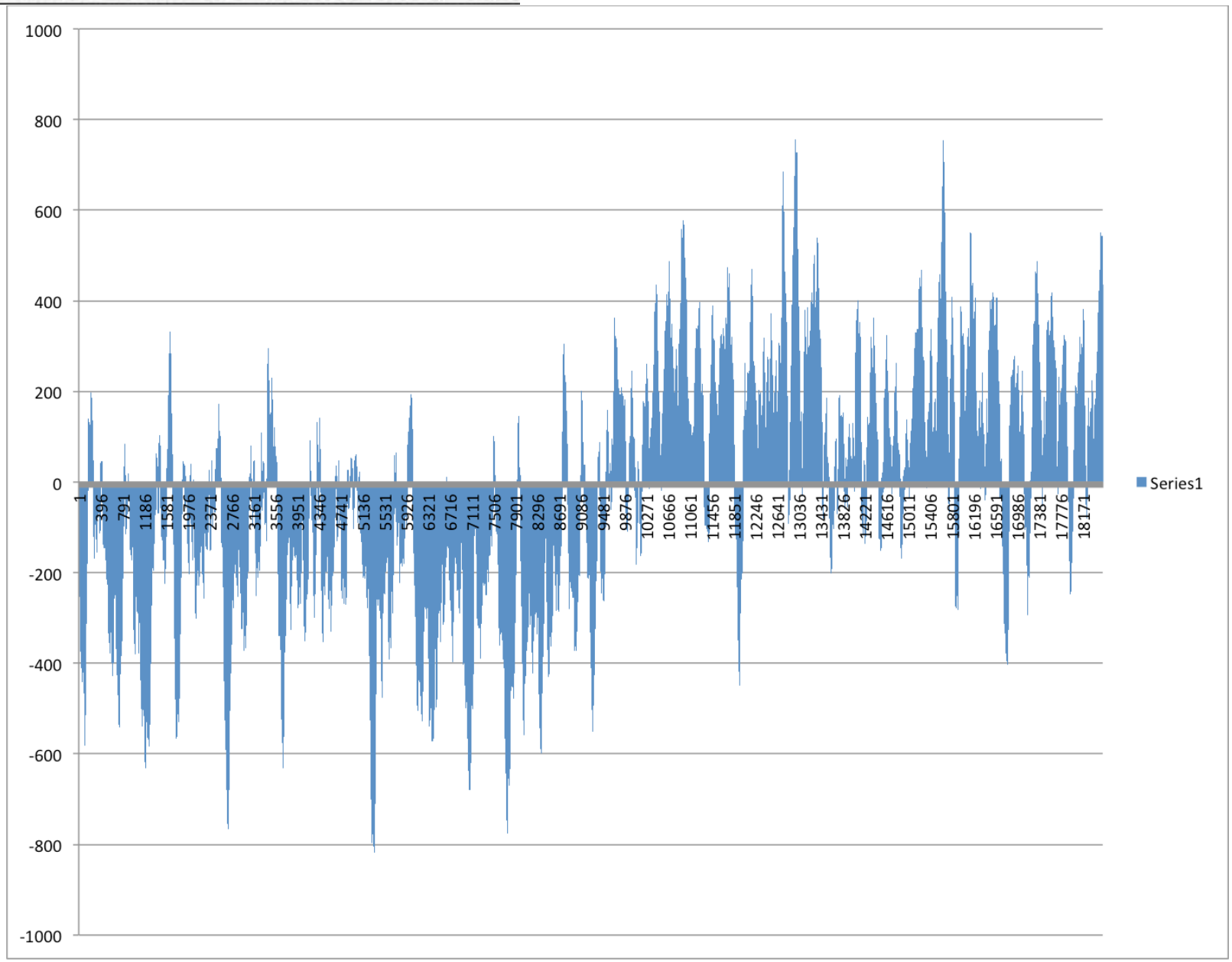
The differences between leading and lagging strand are reflected in the number or ORFs encoded on the strands and the presence of motifs that bind factors initiating and halting replication, and the composition with respect to nucleotides, and n-mers of nucleotides.

One way to look at strand bias, is to calculate the GC content in a rolling window.



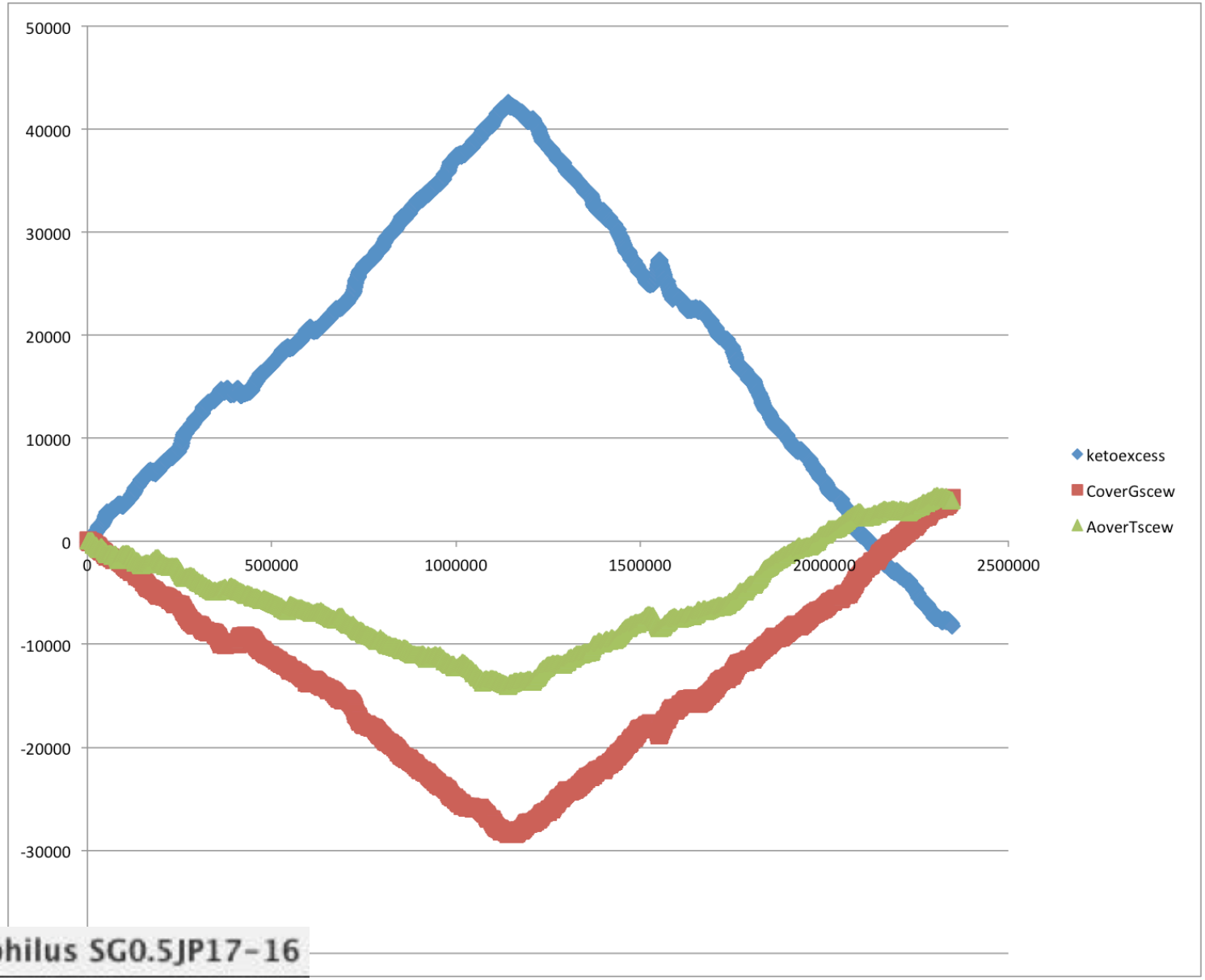
Window=1000 , printed every 100

Thermus thermophilus SG0.5JP17-16



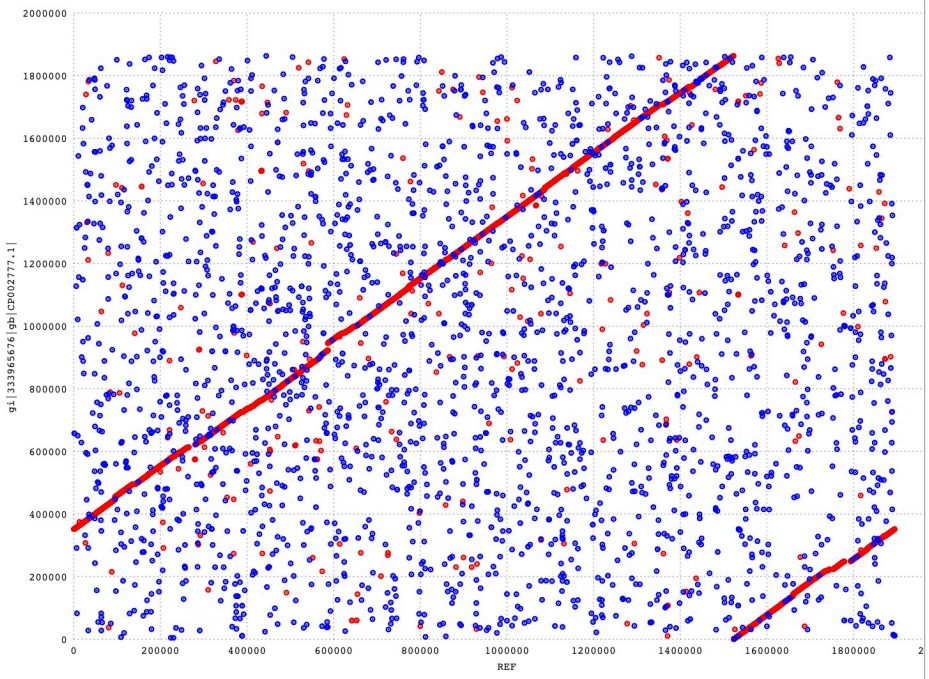
Window=10000 , printed every 100

Usually one plots the Cumulative Strand Bias to more clearly see the turning points

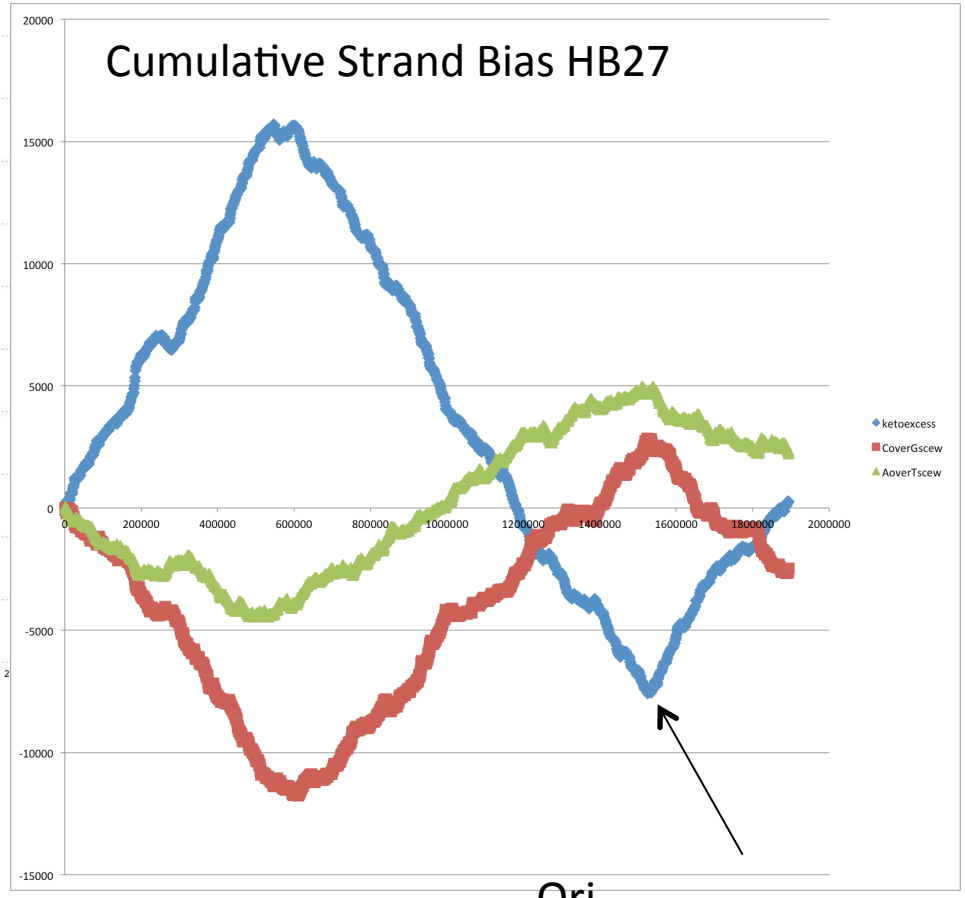


Thermus thermophilus SG0.5JP17-16

Usually, *.fna files of bacterial genomes start with the origin of replication, and the direction is chosen so that the first encoded protein is DnaA (chromosomal replication initiator protein). Sometimes things go wrong.



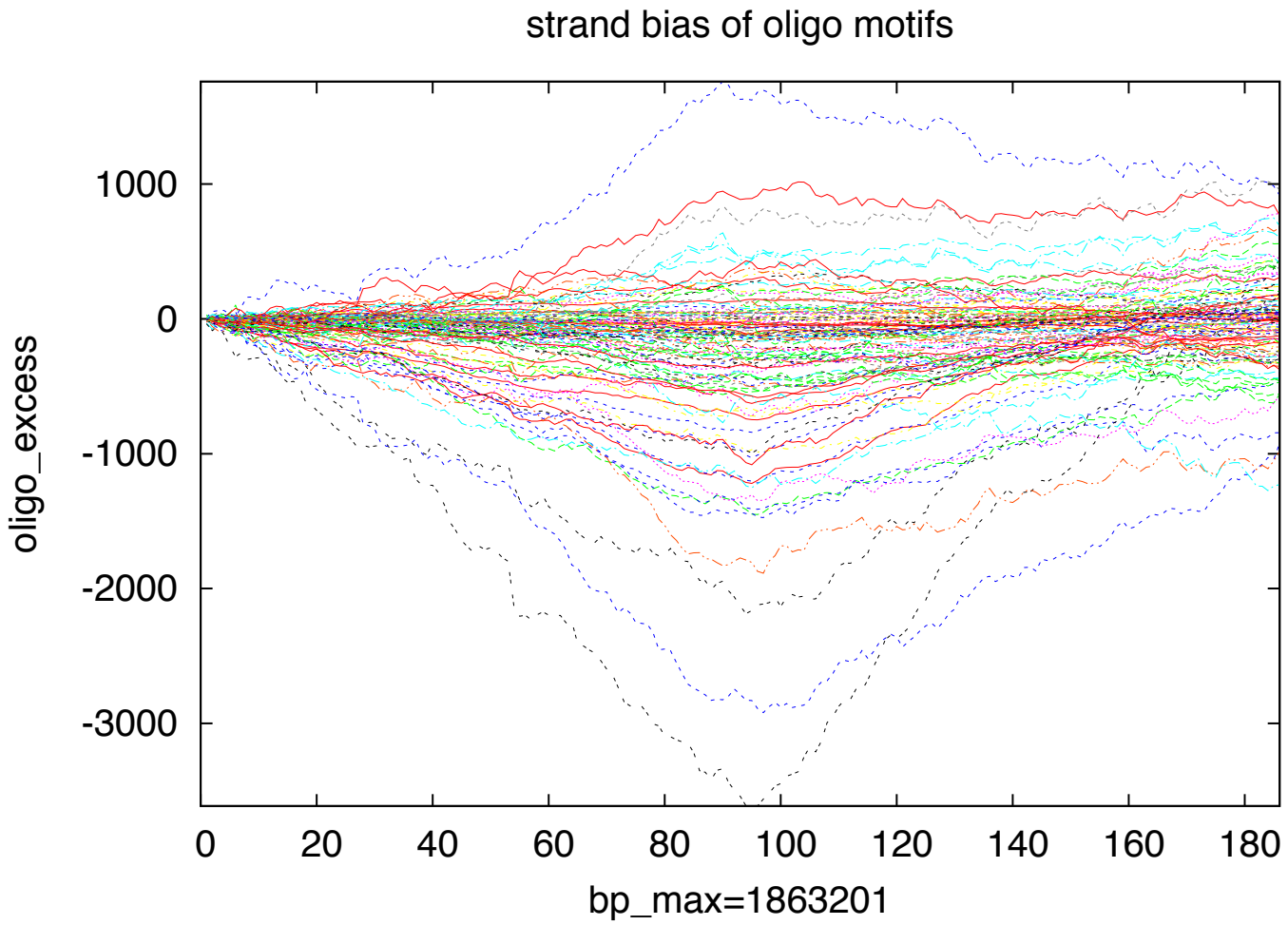
Mummer Plot: HB27 versus SG0



Ori
Should be here

The same can be done with oligonucleotide bias (how often does an oligonucleotide occur on one strand minus occurrence on the other strand)

Tetramer bias for *Thermus thermophilus* SG0



Download a bacterial or archaeal genome of your choice from
ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
We are only interested in the sequence of the main chromosome

Write a script that

1. Reads in a genome
2. Assigns the nucleotides of the genome to an array (e.g. @genome)
3. Goes through the @genome array and sequentially
   counts the numbers of Gs Cs As and Ts
4. Every 1000 (or 5000) nucleotides calculate the excess of Gs over Cs, As over Ts, and the excess of keto bases (G+T-A-C). Feel free to explore other biases.
5. Print the results into a table
6. Plot the columns of the table in Excel or gnuplot

Possibility for

1. Read in a genome

The easiest will be to plagiarize a script you already wrote. The following works, if the script and the *.fna file are in the same directory.

1a: open input and output file, reset stuff

```
#!/usr/bin/perl -w
#initialize genome name and base_hash
$my_genome = "";
%base_hash=();

#assign genome name to $my_genome
@dir=`ls`; # see P24 in the UNIC Perl primer
foreach (@dir) {
   if (m/\.fna$/) {if ($my_genome) {die "More then one genome in directory"} else {$my_genome=($_)}
   }
}
######
chomp ($my_genome);
print "\n\n$my_genome is the file name of the genome to be analyzed \n";

# open my genome for input
open (IN, "< $my_genome") or die "cannot open $my_genome:$!";

# open my_table for output
open (OUT, ">my_table" ) or die "cannot open my_table" ;
print OUT "number \tketoexcess\tCoverGscew\tAoverTscew\tbase_hash{A}\tbase_hash{T}\tbase_hash{G}\tbase_hash{C}\n";
# if we want to use exel, we can print a header in the first line";
# if we use gnuplot, we want to omit the header
```

Possibility for
1b: read genome into array
You have 2 possibilities either read and analyze the genome line by line, or read in everything and then start the analysis.

```
$header = <IN>;
#reads first line of file, next command test for fasta commentline

if ($header =~m/^>/) {print "\nthe analyzed genome has the following comment line:\n$header \n\n"};
if (!($header =~m/^>/)) {print "this is not in FASTA format \n\n";
            exit;}
###         exit - could have died instead;


### Read genome into array @genome
$number=0;

while (defined ($line=<IN>)){

#initialise @bases within loop

        @bases=();
    chomp($line);

    @bases=split(//,$line);

    foreach (@bases) {
       $number += 1;
       $genome[$number]=$_;
    }
```

The $_ is the variable the perl goes through in the foreach loop, see P13

Possibility for
1b': read genome into array
You have 2 possibilities either read and analyze the genome line by line, or read in everything and then start the analysis.

```
if ($header =~m/^>/) {print "\nthe analyzed genome has the following comment line:\n$header \n\n"};

if (!($header =~m/^>/)) {print "this is not in FASTA format \n\n";
            exit;}
###         exit - could have died instead;

$number=0;

while (defined ($line=<IN>)){

#initialise @bases within loop
# potential problem: this reads and analyses line by line.
#It might be better, especially if one wants to use nucleotide pairs or oligod, to read everthing in first
        @bases=();
    chomp($line);

    @bases=split(//,$line);

    foreach (@bases) {
```

....

Write a script that

1. Reads in a genome
2. Assigns the nucleotides of the genome to an array (e.g. @bases)
3. Goes through the @bases array and sequentially counts the numbers of Gs Cs As and Ts

```
    @bases=();
chomp($line);

@bases=split(//,$line);

foreach (@bases) {
   $number += 1; # counts position in genome
   $base_hash{$_} += 1;#counts As,Gs and Cs and Ts
```

```
print OUT "$number\t$ketoexcess\t$gcscew\t$atscew\t$base_hash{A}\t$base_hash{T}\t$base_hash{G}\t$base_hash{C}\n";
```

4. Every 1000 nucleotides calculates the excess of Gs over Cs, As over Ts, and the excess of keto bases (G+T-A-C). Feel free to explore other biases.

```
#every 1000 nucleotides, print stuff to file
      if ($number%1000==0){
      $gcscew=($base_hash{C}-$base_hash{G});
      $atscew=($base_hash{A}-$base_hash{T});
      $ketoexcess=($base_hash{G}+$base_hash{T})-($base_hash{A}+$base_hash{C});
```

5. Print the results into a table

```
print OUT "$number\t$ketoexcess\t$gcscew\t$atscew\t$base_hash{A}\t$base_hash{T}\t$base_hash{G}\t$base_hash{C}\n";
```

Close loop(s), close files

6. Plot the columns of the table in Excel or gnuplot

Links to info on gnuplot is here

If gnuplot is installed on you computer and able to communicate with your x11 terminal program, then

```
> gnuplot
```
Will invoke the gnuplot program

```
> plot "my_table" using 1:2 with lines
```
will plot the 2nd columns against the 1st

```
>set terminal x11
```
Will set the output to x11 (screen)

```
>set terminal png
```
Will set the output to a png file

```
>set terminal postscript
```
sets the output to a postscript file

```
> set out "myplot.png"
```
directs output to the file myplot.png

```
> set multiplot
```
Plots multiple curves into the same figure

```
> plot "my_table" using 1:4 with lines,\
> "my_table" using 1:2 with lines,\
> "my_table" using 1:3 with lines
```
plots multiple curves onto the same figure.

gnuplot is installed on the cluster, but you need to direct the output to a file, which is inconvenient (extra credit, if you can make the cluster output to an x11-window on your laptop)

Often one uses a perl script to do the plotting example here