A Basic Bioinformatics Session

Lecture – 5

Prof. Sanjeeva Srivastava & TAs Team
BSBE, IIT Bombay

How to analyse DNA, Protein sequence?

Outline

- How to study a gene using bioinformatics?
- How to find out sequence similarity between proteins
- How to build an evolutionary relationship between species
- Power of data analysis and visualization using Orange and Python
- Let's explore and navigate expression of a gene/protein in human brain diseases

Bioinformatics Session Activity-I

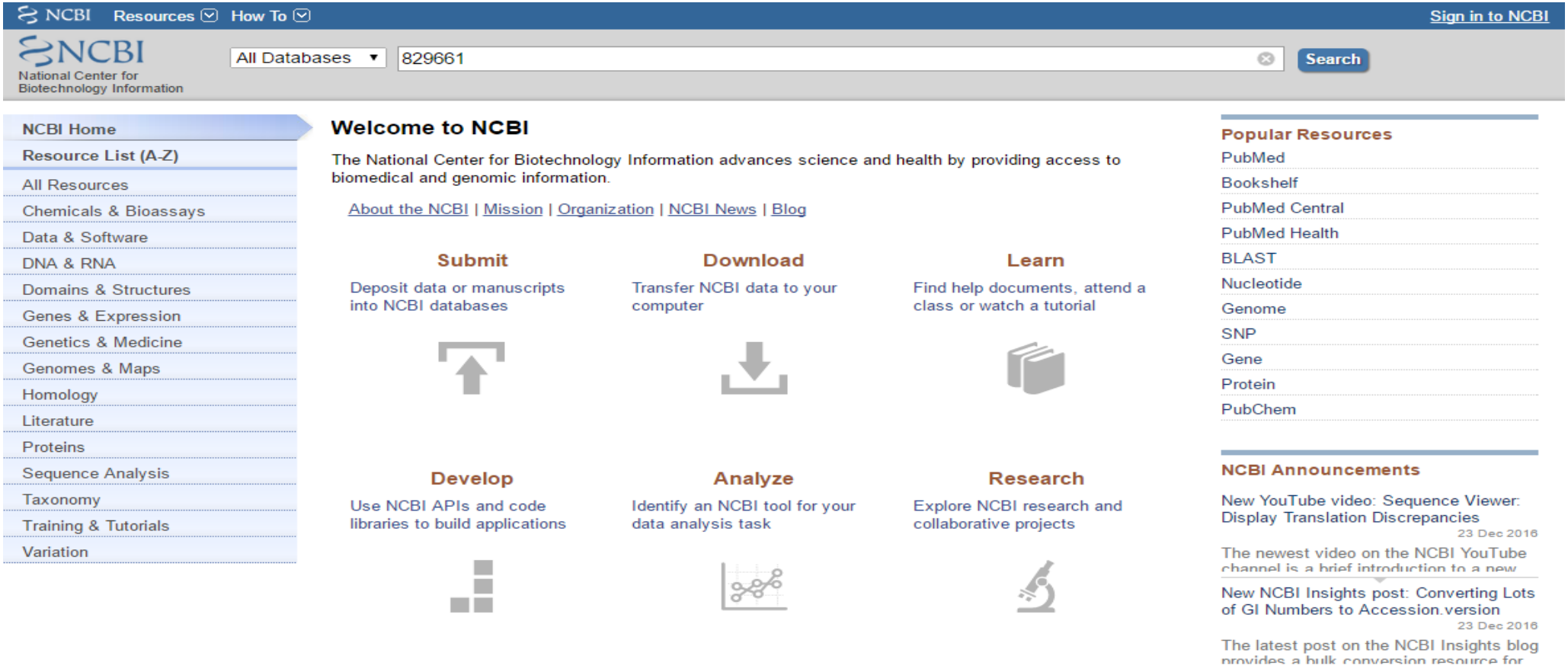
Objective – How to study a gene using bioinformatics?

Activity - I

Gene ID	829661
Date of last update	
Gene symbol	
Gene description	
Locus tag	
Location	
Size of gene (bp)	
Organism	
Superkingdom	
Size of Chromosome (bps)	
# Genes on chromosome	
Flanking gene to the left on genome	
Flanking gene to the right on genome	
EC#	

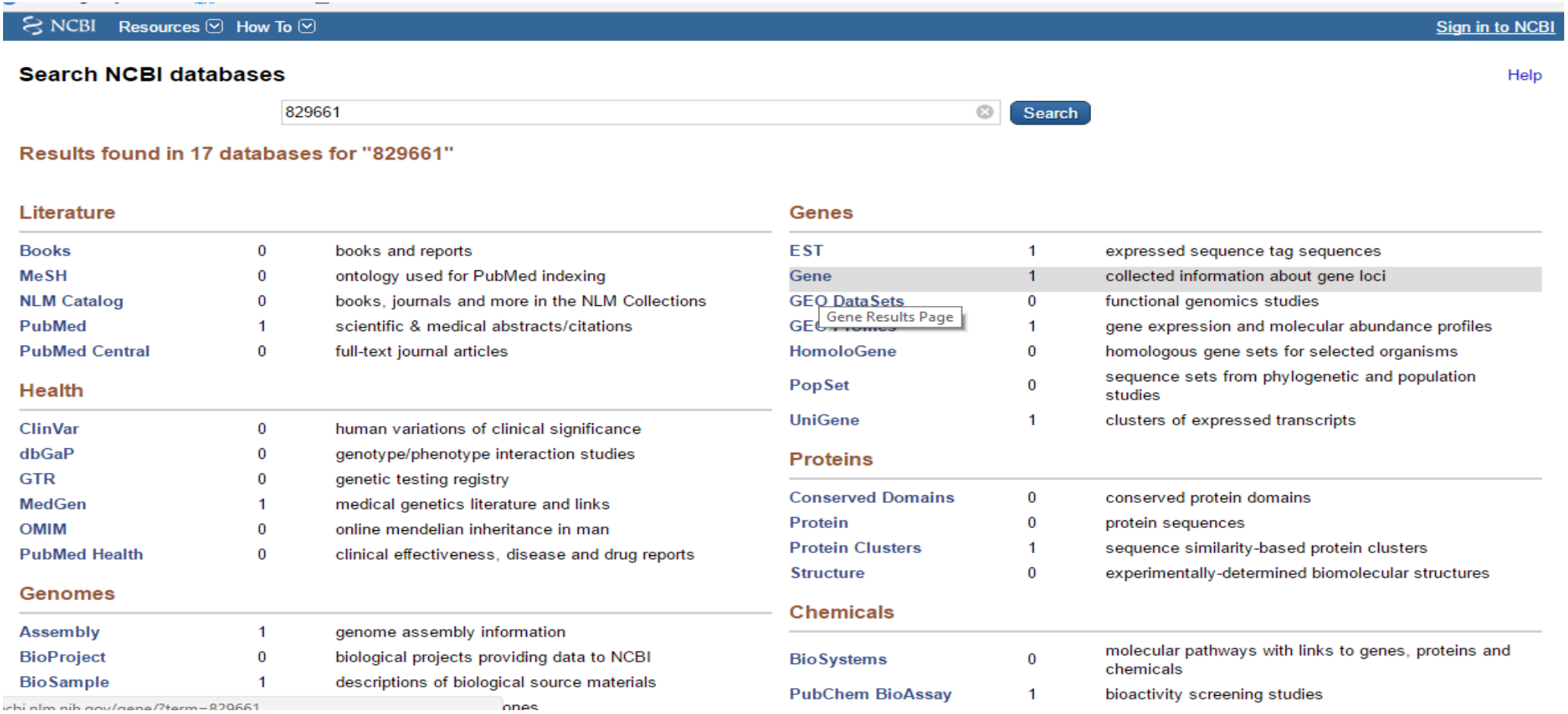
Activity - I

Step 1 https://www.ncbi.nlm.nih.gov



Activity - I

Step 2



Activity - I

Steps 3 & 4

Gene ID: 829661, updated on 14-Dec-2016

Summary

Gene symbol	CAT2
Gene description	catalase 2
Primary source	Araport:AT4G35090
Locus tag	AT4G35090
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	CATALASE; catalase 2; T12J5.2
Summary	Encodes a peroxisomal catalase, highly expressed in bolts and leaves. mRNA expression patterns show circadian regulation with mRNA levels being high in the subjective early morning. Loss of function mutations have increased H2O2 levels and increased H2O2 sensitivity. Mutants accumulate more toxic ions yet show decreased sensitivity to Li+. This decreased sensitivity is most likely due to an insensitivity to ethylene. Note that in Queval et al. (2007) Plant Journal, 52(4):640, SALK_057998 is named as cat2-1, SALK_076998 is named as cat2-2; in Bueso et al. (2007) Plant Journal, 52(6):1052, SALK_076998 is named as cat2-1. TAIR has adopted the nomenclature consistent with that in Bueso et al. (2007) after consultation with the authors: SALK_076998 (cat2-1), SALK_057998 (cat2-2).

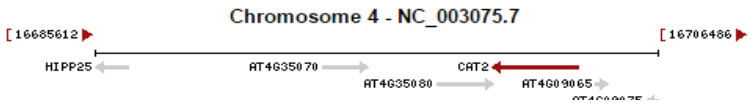
Use this for the first few questions

Genomic context

Location: chromosome: 4 — See CAT2 in Map Viewer

Exon count: 7

Sequence: Chromosome: 4; NC_003075.7 (16700312..16703504, complement)



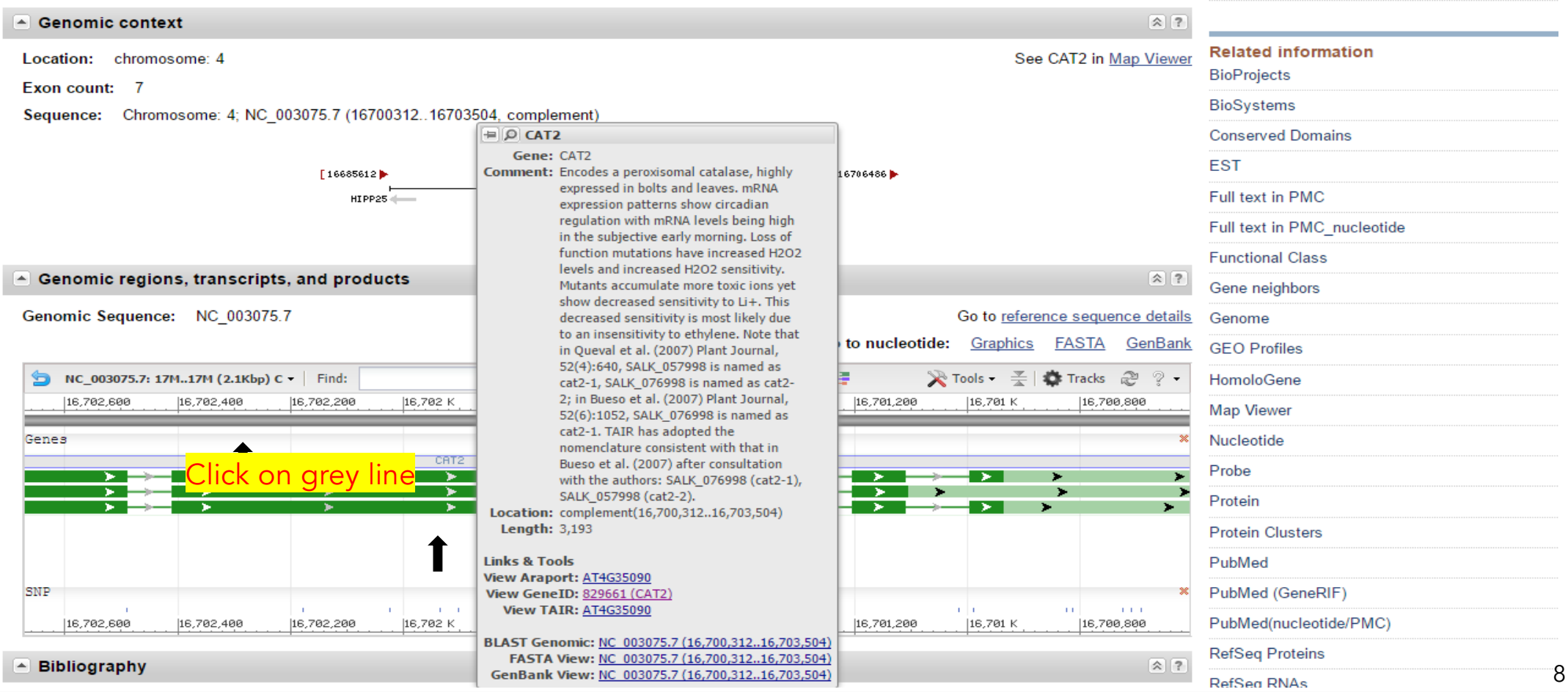
Genomic context
Genomic regions, transcripts, and products
Bibliography
Variation
Pathways from BioSystems
Interactions
General gene information
Markers, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links

Genome Browsers
Map Viewer

Related information
BioProjects
BioSystems
Conserved Domains
EST
Full text in PMC
Full text in PMC_nucleotide

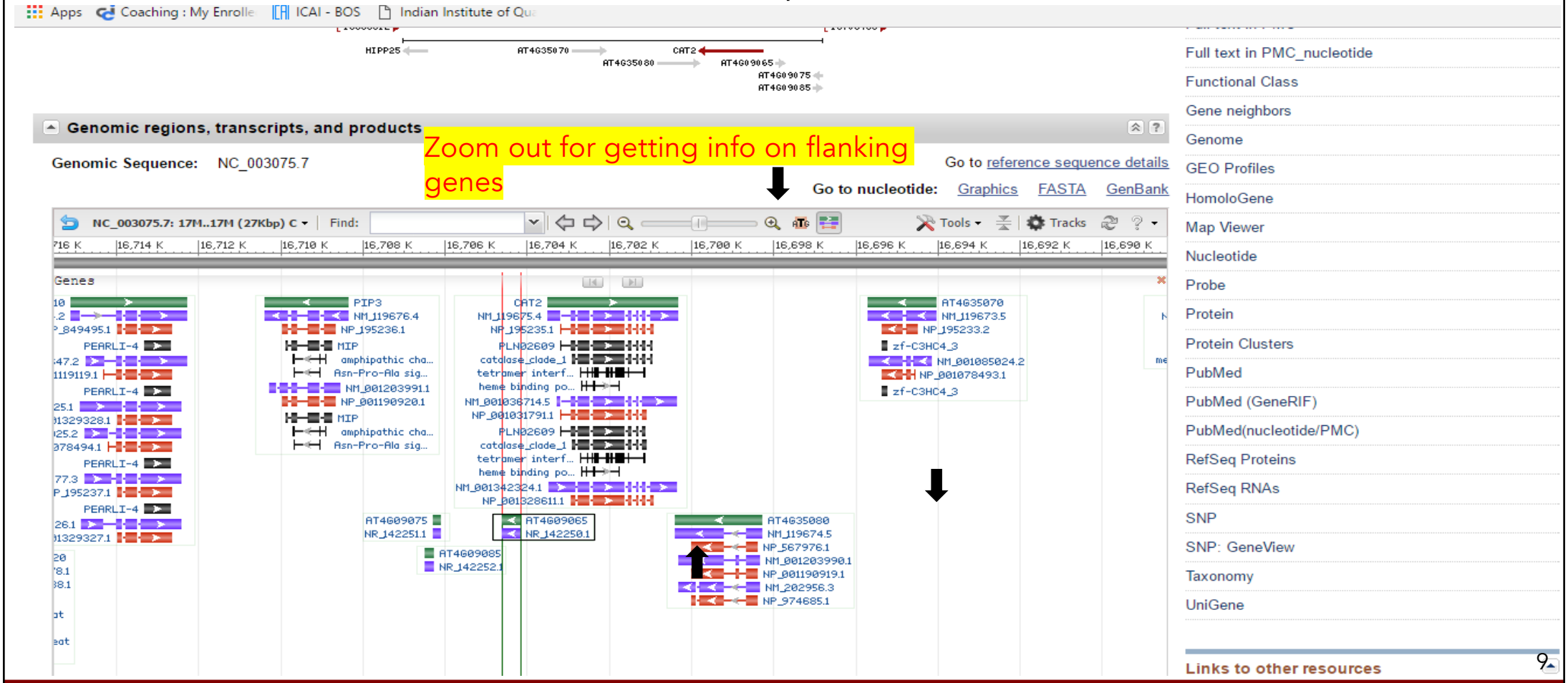
Activity - I

Step 5



Activity - I

Step 6



Activity - I

Step 7

Full Report ▾ Send to: ▾ Hide sidebar >

Showing Current items.

CAT2 catalase 2 [*Arabidopsis thaliana* (thale cress)]

Gene ID: 829661, updated on 14-Dec-2016

Summary

Gene symbol	CAT2
Gene description	catalase 2
Primary source	Araport:AT4G35090
Locus tag	AT4G35090
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	CATALASE; catalase 2; T12J5.2
Summary	Encodes a peroxisomal catalase, highly expressed in bolts and leaves. mRNA expression patterns show circadian regulation with mRNA levels being high in the subjective early morning. Loss of function mutations have increased H2O2 levels and increased H2O2 sensitivity. Mutants accumulate more toxic ions yet show decreased sensitivity to Li+. This decreased sensitivity is most likely due to an insensitivity to ethylene. Note that in Queval et al. (2007) Plant Journal, 52(4):640, SALK_057998 is named as cat2-1, SALK_076998 is named as cat2-2; in Bueso et al. (2007) Plant Journal, 52(6):1052, SALK_076998 is named as cat2-1. TAIR has adopted the nomenclature consistent with that in Bueso et al. (2007) after consultation with the authors: SALK_076998 (cat2-1), SALK_057998 (cat2-2).

Click

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Variation
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Genome Browsers

- Map Viewer

Activity - I

Step 8

Apps Coaching : My Enrolle ICAI - BOS Indian Institute of Qu

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for as complete name lock Go

Display 3 levels using filter: none

Arabidopsis thaliana

Organism

Taxonomy ID: 3702
Genbank common name: **thale cress**
Inherited blast name: **eudicots**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 1 (Standard)
Other names:
common name: **thale-cress**
common name: **mouse-ear cress**
authority: **Arabidopsis thaliana (L.) Heynh.**

Lineage (full)
cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis

Super Kingdom

Click

Entrez records	
Database name	Direct links
Nucleotide	361,009
Nucleotide EST	1,529,700
Nucleotide GSS	752,876
Protein	287,613
Structure	985
Genome	1
Popset	1,174
SNP	1,069,597
Domains	51
GEO Datasets	42,792
UniGene	30,633
PubMed Central	45,035
Gene	43,798
HomoloGene	10,445
SRA Experiments	21,929
Probe	206,829
Assembly	12
Bio Project	3,158
Bio Sample	38,588
Bio Systems	1,021

Activity - I

Step 9

Publications

1. A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. Pucker B, et al. PLoS One 2016 Oct 6
2. Phased diploid genome assembly with single-molecule real-time sequencing. Chin CS, et al. Nat Methods 2016 Dec
3. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Berlin K, et al. Nat Biotechnol 2015 Jun

More...

Genome Links for Gene (Select 829661) (1) Genome

txid3702[Organism:noexp] (1) Genome

CAT2 [Arabidopsis thaliana] Gene

829661[uid] AND (alive[prop]) (1) Gene

See more..

Representative (genome information for reference and representative genomes)

Reference genome: [see all organisms]

- *Arabidopsis thaliana TAIR10*

Submitter: The Arabidopsis Information Resource (TAIR)

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Nuc	Chr	1	NC_003070.9	CP002684.1	30.43	35.9	12,646	-	238	1,969	9,693	930
Nuc	Chr	2	NC_003071.7	CP002685.1	19.7	35.9	7,590	2	95	1,341	6,303	1,047
Nuc	Chr	3	NC_003074.8	CP002686.1	23.46	36.3	9,468	2	92	1,411	7,618	1,080
Nuc	Chr	4	NC_003075.7	CP002687.1	18.59	36.2	7,422	-	77	1,072	5,838	835
Nuc	Chr	5	NC_003076.8	CP002688.1	26.[illegible]	35.9	10,987	-	123	1,410	[illegible]13	951
MT	Chr	MT	NC_001284.2	Y08501.2	0.366924	44.8	117	3	21	-	131	-
Chl	Chr	Pltd	NC_000932.1	AP000423.1	0.154478	36.3	85	7	37	-	129	-

Chromosomes

Click on chromosome name to open Genome Data Viewer

Obtain EC number info from the net

Activity - I

Model Answer

	Answers
Gene ID	829661
Date of last update	14-Dec-16
Gene symbol	CAT2
Gene description	catalase 2
Locus tag	AT4G35090
Location	Chromosome 4
Size of gene (bp)	3193bp
Organism	Arabidopsis thaliana
Superkingdom	Eukaryota
Size of Chromosome (bps)	18.59 Mb
# Genes on chromosome	5838bps
Flanking gene to the left on the genome	AT4G09065
Flanking gene to the right on the genome	AT4G35080
EC#	1.11.1.6

Bioinformatics Activity - II

1. To find out sequence similarity between proteins
2. To build an evolutionary relationship between species

Activity - II

Basic Local Alignment Search Tool (BLAST)	
Accession number	NP_000508.1
Protein Name	
Organism Name	
Sequence Length	
Uniprot id	
No. Of BLAST hits	
PDB id	
No. Of BLAST hits	

Multiple Sequnce Alignment (MSA)	
Accession No.s	a) NP_000508.1, b) NP_001070890.2, c) NP_032244.2, d) AFM89055.1, e) ABW88850.1
Evolutionary related	

Activity - II

NCBI Resources How To Sign in to NCBI

Protein Protein Search

Advanced Help

GenPept Send to:

hemoglobin subunit alpha [Homo sapiens]

NCBI Reference Sequence: NP_000508.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_000508                142 aa            linear   PRI 04-JAN-2017
DEFINITION  hemoglobin subunit alpha [Homo sapiens].
ACCESSION   NP_000508
VERSION     NP_000508.1
DBSOURCE    REFSEQ: accession NM_000517.4
KEYWORDS    RefSeq.
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 142)
  AUTHORS   Pang W, Weng X, Ye X, Long J, Wu S, Sun L, Wei C, Chen M, Tang W,
            Qiu S and Zhang C.
  TITLE     Identification of a variation in the IVSII of alpha2 gene and its
            frequency in the population of Guangxi
  JOURNAL   Gene 583 (1), 24-28 (2016)
   PUBMED   26930363
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure

Structure Of The Human Hemoglobin Mutant Hb Providence (a-gly-c:v1m;

PDB: 5SW7

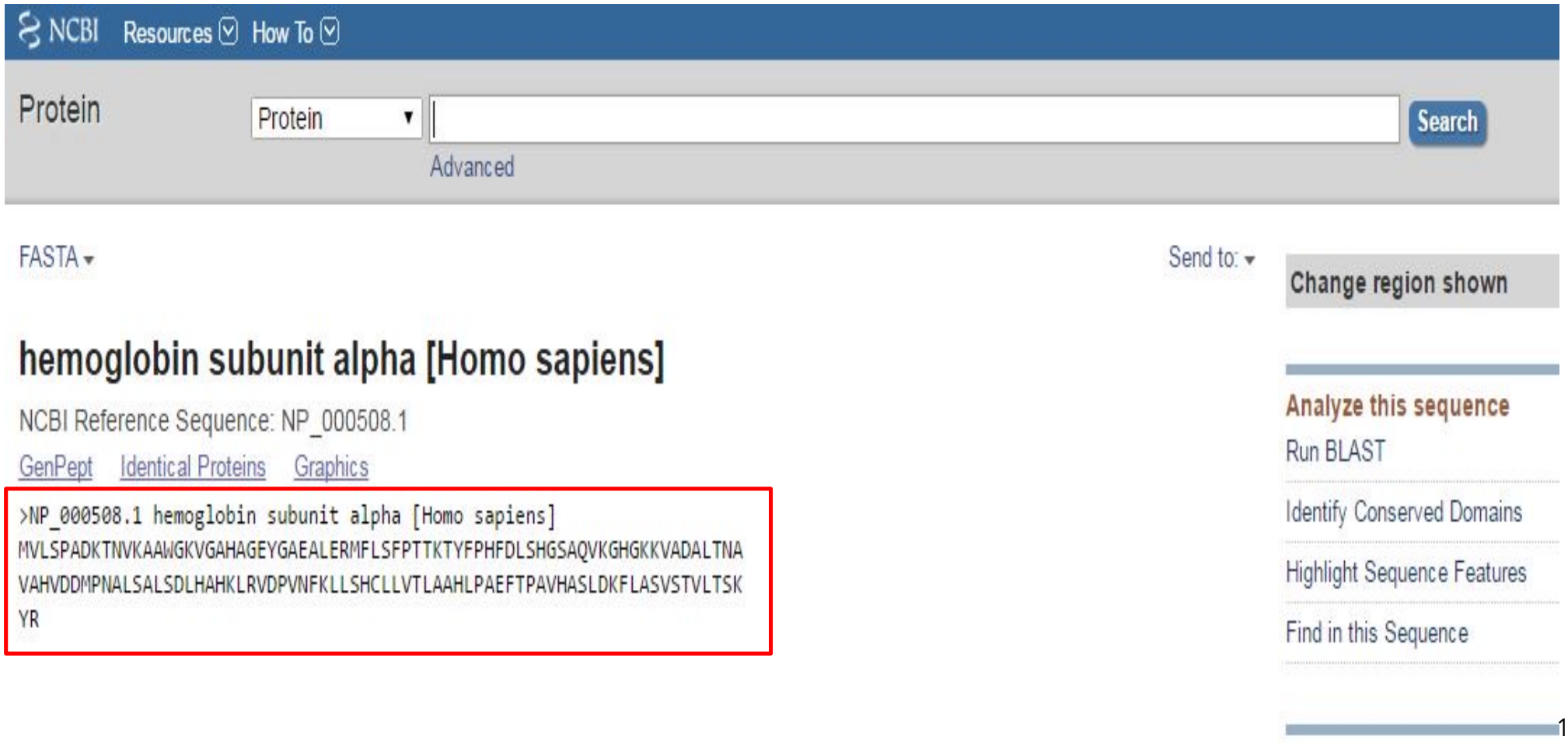
Source: Homo sapiens

Method: X-Ray Diffraction

Resolution: 1.85 Å

See all 222 structures

Activity - II


NCBI Resources How To
Protein
Protein
Search
Advanced
FASTA
Send to:
Change region shown
hemoglobin subunit alpha [Homo sapiens]
NCBI Reference Sequence: NP_000508.1
GenPept Identical Proteins Graphics
>NP_000508.1 hemoglobin subunit alpha [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNA
VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSK
YR
Analyze this sequence
Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence

Activity - II

NCBI Resources How To Sign in to NCBI

Protein Protein Search

Advanced Help

GenPept Send to:

hemoglobin subunit alpha [Homo sapiens]

NCBI Reference Sequence: NP_000508.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_000508               142 aa            linear   PRI 04-JAN-2017
DEFINITION  hemoglobin subunit alpha [Homo sapiens].
ACCESSION   NP_000508
VERSION     NP_000508.1
DBSOURCE    REFSEQ: accession NM_000517.4
KEYWORDS    RefSeq.
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 142)
  AUTHORS   Pang W, Weng X, Ye X, Long J, Wu S, Sun L, Wei C, Chen M, Tang W,
            Qiu S and Zhang C.
  TITLE     Identification of a variation in the IVSII of alpha2 gene and its
            frequency in the population of Guangxi
  JOURNAL   Gene 583 (1), 24-28 (2016)
   PUBMED   26930363
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure

Structure Of The Human Hemoglobin Mutant Hb Providence (a-gly-c:v1m;

PDB: 5SW7

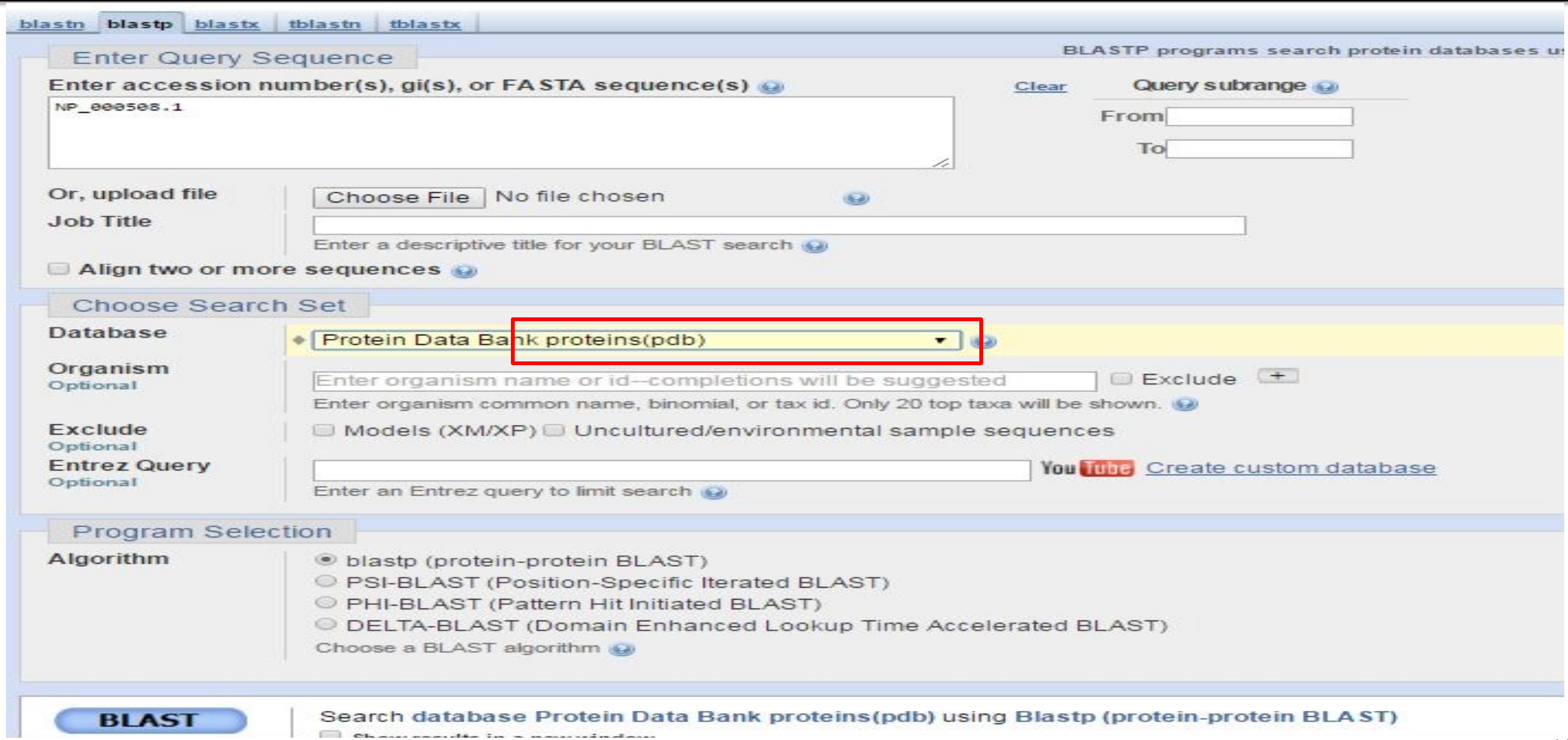
Source: Homo sapiens

Method: X-Ray Diffraction

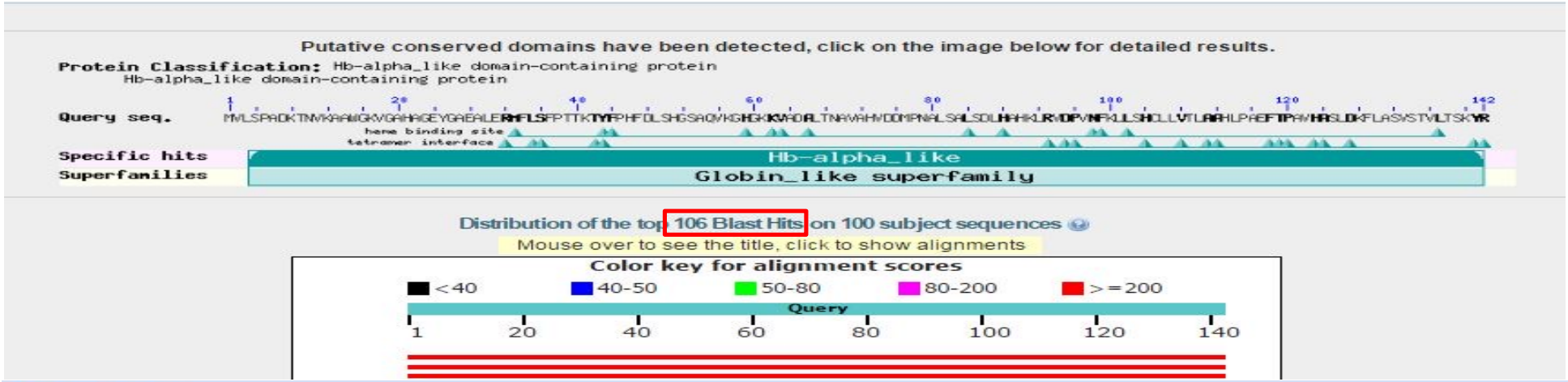
Resolution: 1.85 Å

See all 222 structures...

Activity - II


blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein databases u
Enter accession number(s), gi(s), or FASTA sequence(s)
NP_000508.1
Clear
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Protein Data Bank proteins(pdb)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
YouTube
Create custom database
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Protein Data Bank proteins(pdb) using Blastp (protein-protein BLAST)

Activity - II



Descriptions

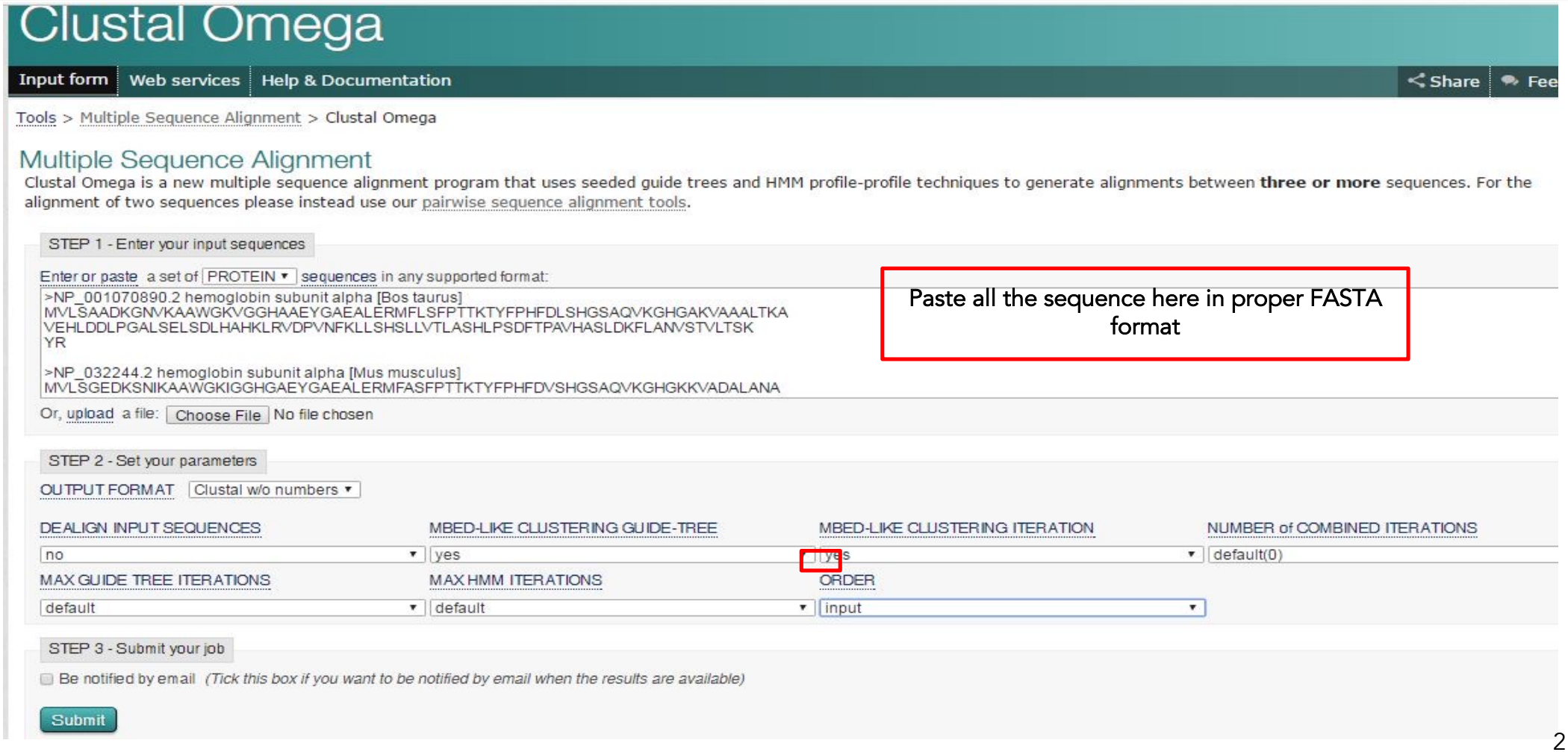
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Chain A, Hemoglobin (Alpha + Met) Variant	286	286	100%	2e-101	100%	1BZ1_A
☐	Chain B, A Cis-Proline In Alpha-Hemoglobin Stabilizing Protein Directs The Structural Reorganizatic	286	286	100%	3e-101	100%	3IA3_B
☐	Chain A, Hemoglobin Thionville: An Alpha-Chain Variant With A Substitution Of A Glutamate For Val	284	284	100%	2e-100	99%	1BAB_A

Activity - II


Clustal Omega
Input form
Web services
Help & Documentation
Share
Tools > Multiple Sequence Alignment > Clustal Omega
Multiple Sequence Alignment
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.
STEP 1 - Enter your input sequences
Enter or paste a set of PROTEIN sequences in any supported format:
>NP_001070890.2 hemoglobin subunit alpha [Bos taurus]
MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGAKVAAALTKA
VEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTPAVHASLDKFLANVSTVLTSK
YR
>NP_032244.2 hemoglobin subunit alpha [Mus musculus]
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHGKKVADALANA
Paste all the sequence here in proper FASTA format
Or, upload a file: Choose File No file chosen
STEP 2 - Set your parameters
OUTPUT FORMAT Clustal w/o numbers
DEALIGN INPUT SEQUENCES
no
MBED-LIKE CLUSTERING GUIDE-TREE
yes
MBED-LIKE CLUSTERING ITERATION
yes
NUMBER of COMBINED ITERATIONS
default(0)
MAX GUIDE TREE ITERATIONS
default
MAX HMM ITERATIONS
default
ORDER
input
STEP 3 - Submit your job
Be notified by email (Tick this box if you want to be notified by email when the results are available)
Submit

Activity - II

Clustal Omega

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-I20170109-113913-0104-48805500-oy

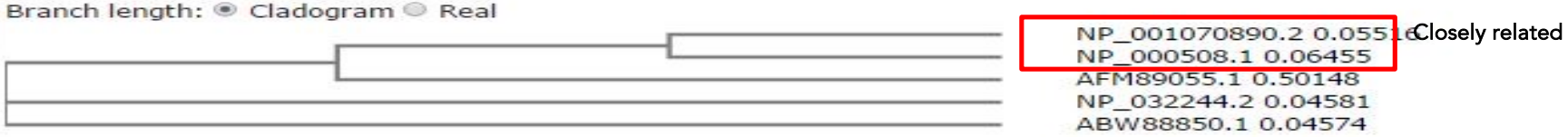
Alignments | Result Summary | Phylogenetic Tree | Submission Details

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Download Phylogenetic Tree Data

Branch length: Cladogram Real



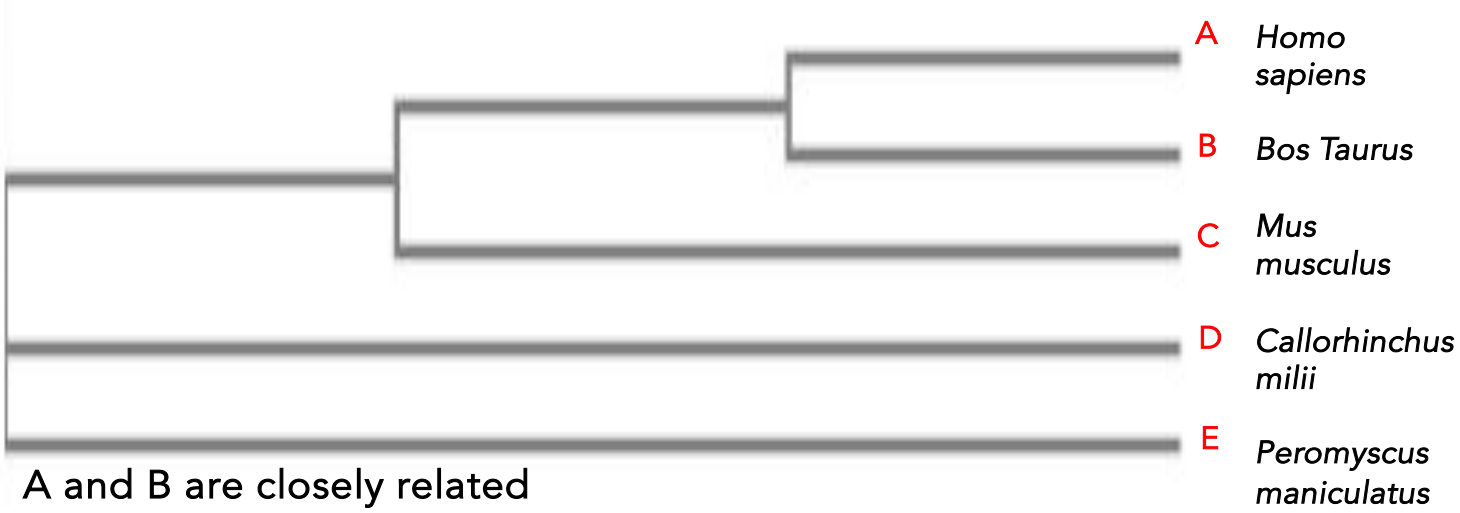
Closely related

Tree Data

```
(
(
(
NP_001070890.2:0.05516,
NP_000508.1:0.06455)
:0.02139,
AFM89055.1:0.50148)
:0.01030,
NP_032244.2:0.04581,
ABW88850.1:0.04574);
```

Activity - II

Phylogeny: how to interpret it?



A and B are closely related

While C is somewhat related to A and B

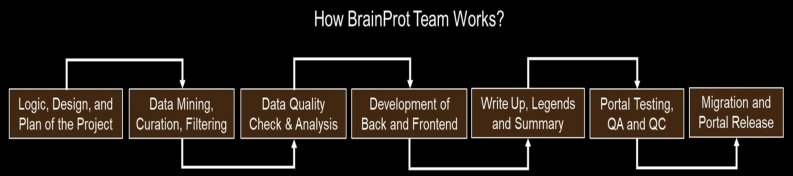
D and E are distantly related from all

Activity - II

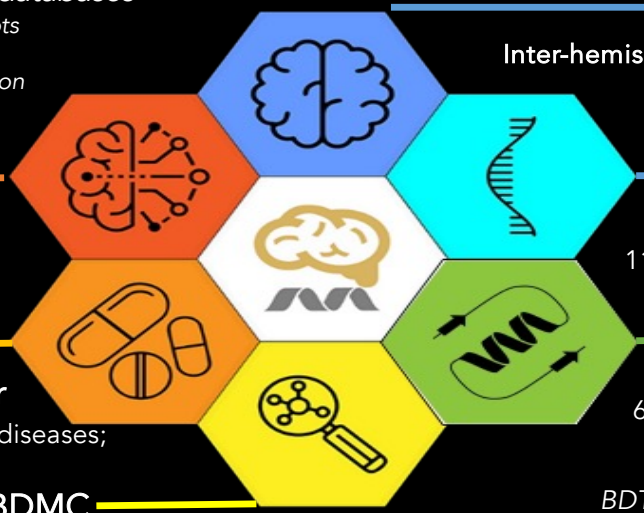
Basic Local Alignment Search Tool (BLAST)	
Accession number	NP_000508.1
Protein Name	hemoglobin subunit alpha
Organism Name	Homo sapiens
Sequence Length	142 amino acid
Uniprot id	P69905.2
No. Of BLAST hits	100
PDB id	1BZ1_A
No. Of BLAST hits	106

Multiple Sequnce Alignment (MSA)	
Accession No.s	a) NP_000508.1, b) NP_001070890.2, c) NP_032244.2, d) AFM89055.1, e) ABW88850.1
Evolutionary related	a) & b)

BrainProt™ a multi-functional knowledgebase of Human Brain Diseases



- *56 brain diseases, 14 popular databases*
- *Manually curated from 1300 manuscripts*
- *Meta-data information ~1900 samples*
- *Inter-disease & inter-data set comparison*



IBPM
Inter-hemispheric Brain Proteome Map

HBDA
Human Brain Disease Atlas

BDTM
Brain Disease Transcriptome Map
11 brain diseases are integrated

BDDF
Brain Disease Drug Finder
Clinical trial information of 56 brain diseases; drug identification of 11 diseases

BDPM
Brain Disease Proteome Map
6 brain diseases are integrated

BDMC
Brain Disease Marker Curator

BDTM, BDPM - 51 omics datasets; total of ~1900 samples, 6000 DEGs, 3500 DEPs

Case Study-1: Gliomas

HBDA

Human Brain Disease Atlas - Index Table

Click on any disease's name to access its data in our portal!

Disease Name	MeSH ID	MedGen UID
Essential Tremor	D020329	78725
Fragile X Syndrome	D005600	8912
Friedreich Ataxia	D005621	5276
Frontotemporal Dementia	D057180	83266
Frontotemporal Lobar Degeneration	D057174	148228
Gaucher Disease	D005776	42164
Gilles De La Tourette Syndrome	D005879	21219
Glioblastoma	D005909	42228
Glioma	D005910	9030
Huntington's Disease	D006816	5654

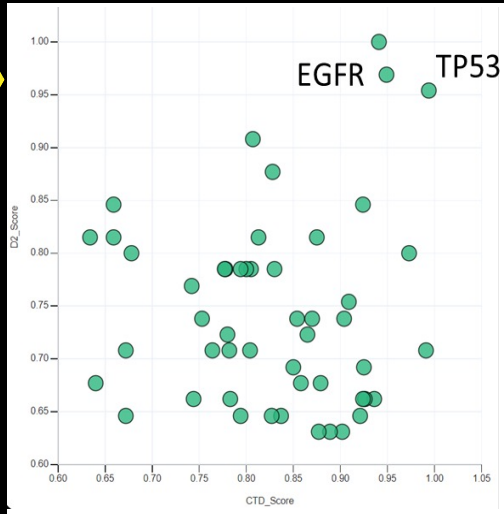
Search

Rows per page: 10 21 - 30 of 56

Disease Resources:

Disease Name:	Glioma
MeSH ID:	D005910
MedGen UID:	9030
UMLS:	C0017638
Disease Ontology (DO):	0060108
NCI:	C3059
GARD:	6513
KEGG:	hsa05214
MSeqDR:	HP:0009733
Monarch Initiative:	0021042
OMIM:	137800

BDMC

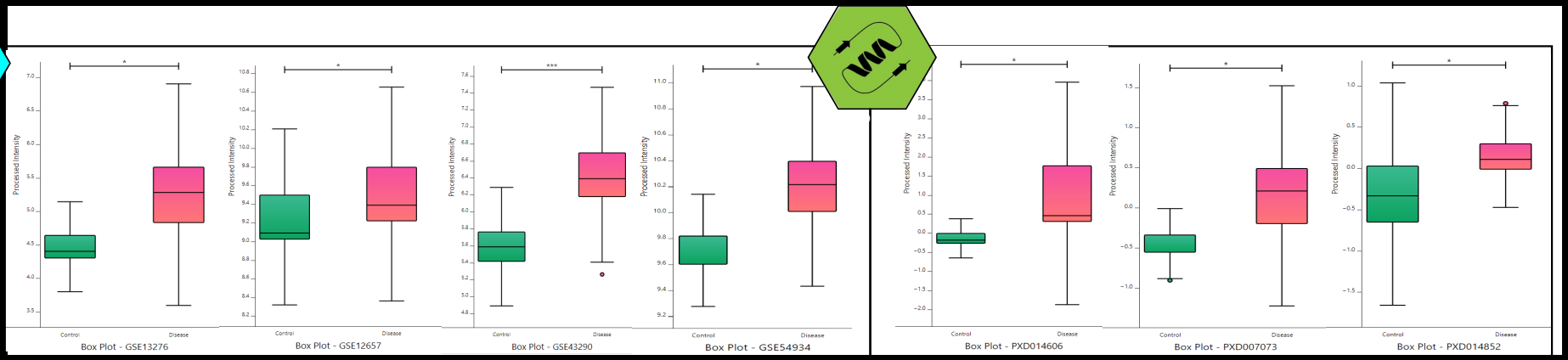


Search

Gene Name	SwissProt Accession	DISGENET Score	Harmonizome Score	Pubpular Score	CTD Score	Disease 2.0 Score	BDMC Score	Frequency	CTD Marker	eDGAR Marker	Disease 2.0 Marker
TP53	P04637	1	0.958	0.729	0.994	0.954	9.125	5	1	1	1
PTEN	P60484	0.918	0.958	0.804	0.807	0.908	9.193	5	1	1	1
IDH1	O75874	0.88	1	1	0.131	1	9.685	5	1	1	1
EGFR	P00533	0.834	0.956	0.809	0.949	0.969	9.114	5	1	0	1
MGMT	P16455	0.793	0.945	0.982	0.941	1	9.32	5	1	0	1
AKT1	P31749	0.707	0.804	0.719	0.52	0.923	7.408	5	0	0	1
CDKN2A	Q8N726	0.854	0.686	0.712	0.805	0.785	8.287	5	1	0	1
VEGFA	P15692	0.686	0.689	0.68	0.386	0.892	8.134	5	1	0	1
PROM1	O43490	0.661	0.906	0.838	0.828	0.877	8.574	5	1	0	1
GFAP	P14136	0.662	0.868	0.89	0.531	0.785	7.457	5	0	0	1

Case Study-1: Gliomas (contd..)

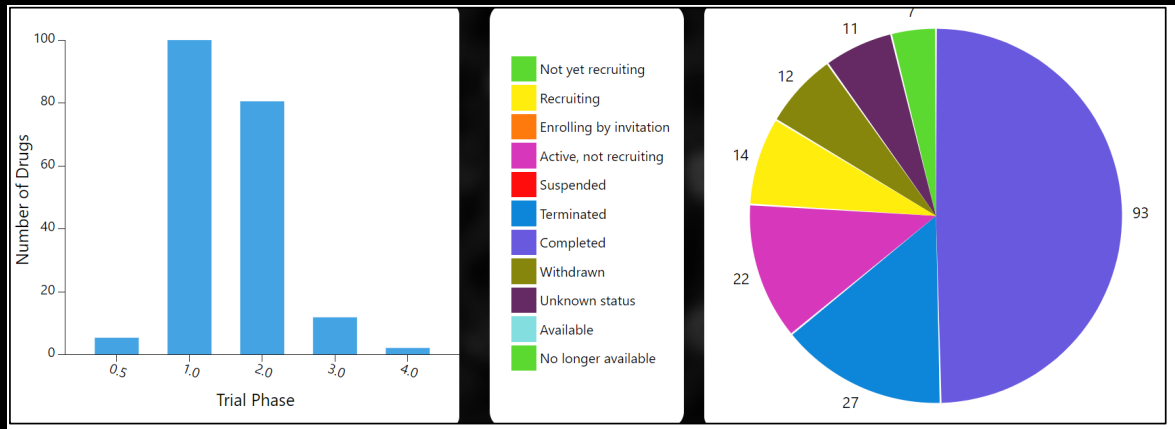
BDTM



Glioblastoma | Glioma | Meningioma | Glioma | Meningioma

BDDF

Clinical Trial Information of Glioma



Top Drugs available for TP53

TP53

Uniprot ID	Gene Name	ChEMBL ID	Common Name	Score
P04637	TP53	CHEMBL1201733	PAZOPANIB HYDROCHLORIDE	0.016
P04637	TP53	CHEMBL558752	RAF-265	0.007
P04637	TP53	CHEMBL521686	OLAPARIB	0.148

How close are we?

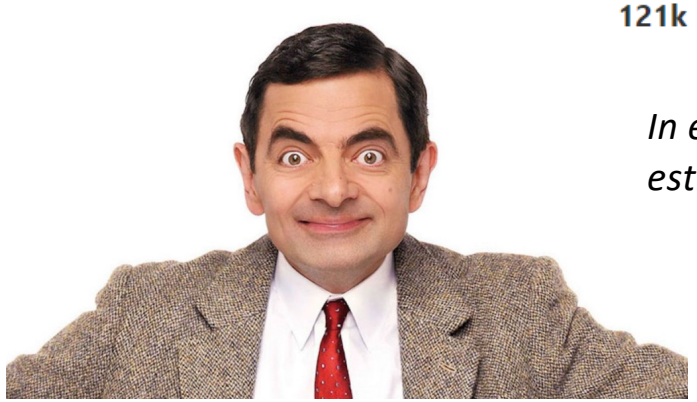
Review | Open access | Published: 10 September 2020

Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species

Maria V. Suntsova & Anton A. Buzdin

BMC Genomics **21**, Article number: 535 (2020) | Cite this article

121k Accesses | **27** Citations | **153** Altmetric | Metrics



In early works, the divergence of human and chimpanzee genomes was estimated as roughly 1%

How close are we?

Eighty percent of proteins are different between humans and chimpanzees

Galina Glazko, Vamsi Veeramachaneni, Masatoshi Nei, Wojciech Makałowski

Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA

Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

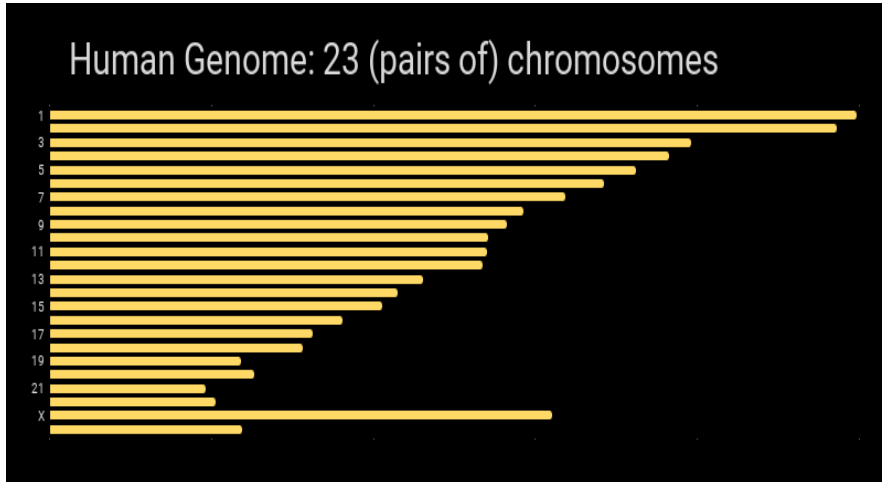
Genes and Genomes

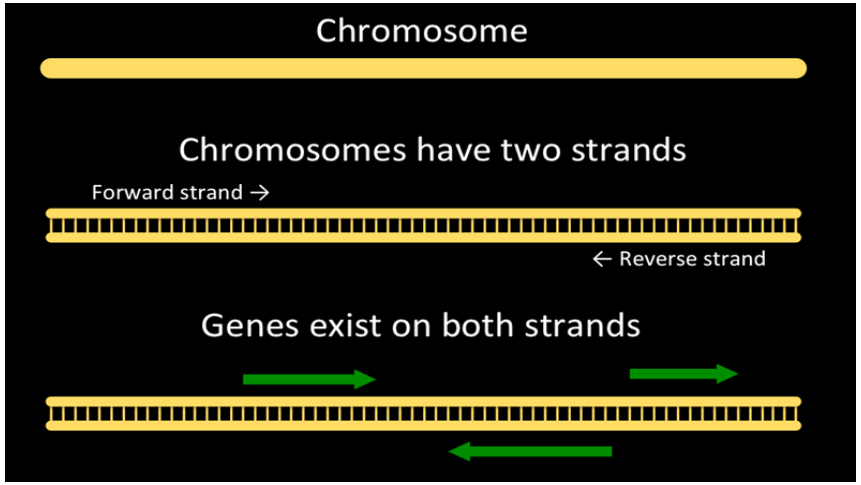
Genome: Collection of all DNA present in an organism; For humans, the genome consists of a person's chromosome

Chromosome: Largest unit of DNA organization in an organism; Linear chromosomes (humans); circular Chromosomes (bacteria)

Strand: Chromosomes are double-stranded; One is known as forward strand, typically drawn on top moving from left to right and and the other is reverse strand; Genes can occur on either strand

Gene: A section of DNA on chromosome strand that creates a functional molecule





Basics Of Sequence Alignment

- Retrieval of sequences from biological databases based on similarity
- Submission of a query sequence- Pairwise alignment with all sequences in the database
- *Sensitivity*- Retrieve as many correct hits as possible
  Specificity/Selectivity- Exclude incorrect hits

- **BLAST- Basic Local Alignment Search tool-Stephen Altschul in 1990 in NCBI**
- Find high-scoring *ungapped segments* among related sequences
- Threshold indicates pairwise similarity beyond random chance

BLAST Scores

E-value - $E = m \times n \times P$

Bit Score $S' = (\lambda \times S - \ln K) / \ln 2$

m= total number of residues in the database

n = number of residues in the query sequence

P= probability that an HSP is done by random chance

E-value- likelihood that alignment has occurred by chance

Lower the E-value, more significant the match is

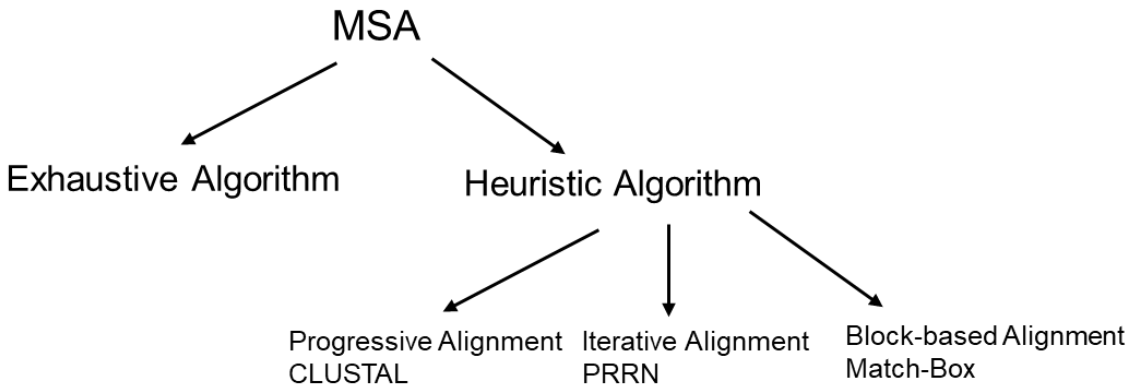
E-value >10 distant relation
E-value 0.01-10 >Not significant but a hint of homologous relationship

Multiple Sequence Alignment

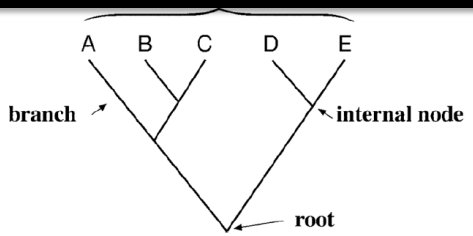
MAFFT (**M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform) is a high speed multiple sequence alignment program.

We have recently changed the default parameter settings for MAFFT. Alignments should run much more quickly and larger DNA alignments can be carried out by default. Please click the 'More options' button to review the defaults and change them if required.

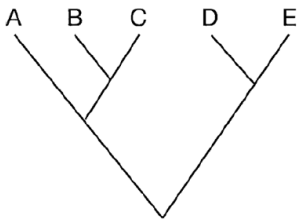
- To align multiple related sequences to achieve optimal matching of the sequences
- Numerous pairwise alignment to a single alignment to identify the conserved regions-sequences and motifs

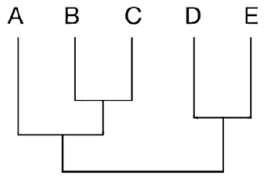


Representing evolutionary relationships between organisms

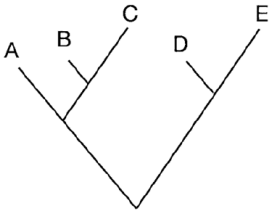


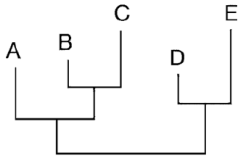
Tree Representation



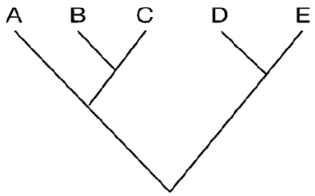


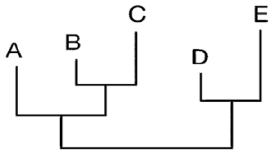
Cladogram





Phylogram





- *Newick format*- To provide information of tree topology to computer programs

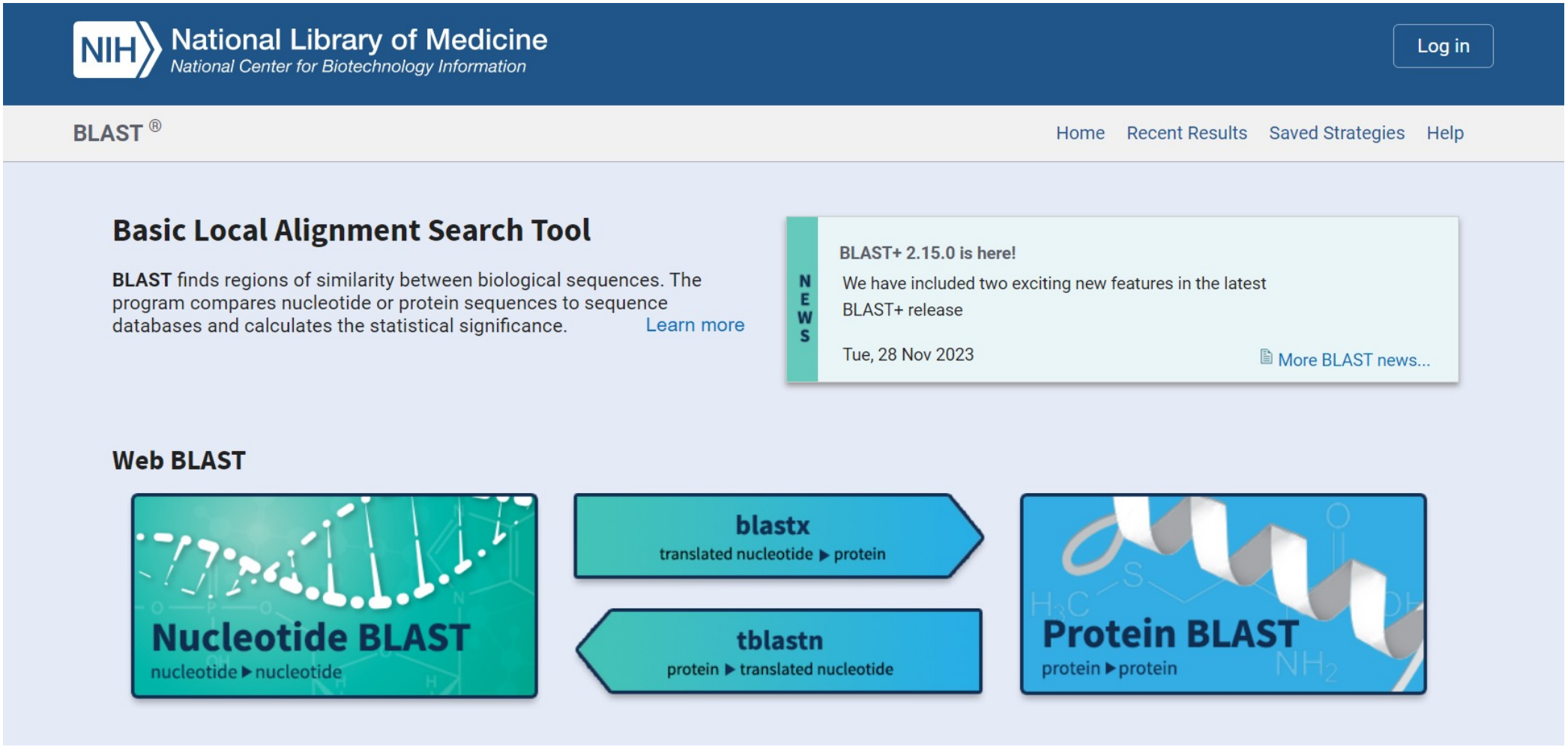
- trees are represented by taxa included in nested parentheses

(((B,C),A),(D,E))

(((B:1,C:2),A:2),(D:1.2,E:2.5))

Newick format

How to find similarities between two sequence?



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Clustal Omega

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback

Tools > Multiple Sequence Alignment > Clustal Omega

Service Announcement

We will be retiring these pages on the 31st of January, 2024. Please visit our new website at **www.ebi.ac.uk/jdispatcher** and share your **feedback** with us!

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

https://www.ebi.ac.uk/Tools/msa/clustalo/

Let’s try to explore and navigate the expression of a gene/protein in Human Brain Diseases

https://www.brainprot.org/

The Power of Data Analysis and Visualization using Orange and Python

orange
DATA MINING

Screenshots Examples Download Blog Docs Workshops Search

Data Mining Fruitful and Fun

Open source machine learning and data visualization.

Download Orange 3.36.2

Next Lecture…

Molecular basis of inheritance & Flow of information