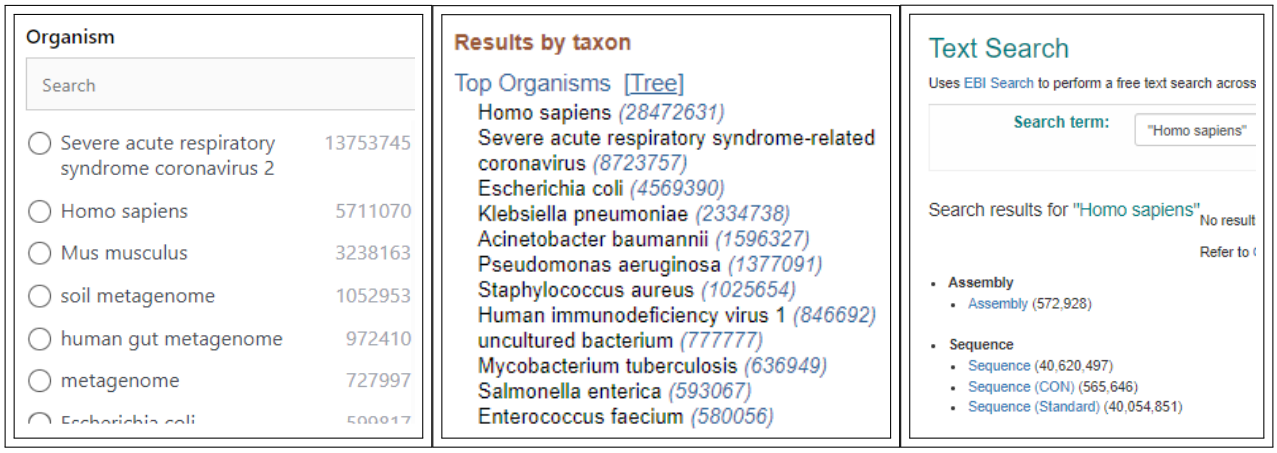
PRACTICAL 2

Question 1. How many "Homo sapiens" sequences deposited in DDBJ? Compare with Genbank and EMBL
http://www.ddbj.nig.ac.jp/
http://www.ncbi.nlm.nih.gov/genbank/
http://www.ebi.ac.uk/embl/

Solution.

DDBJ	Genbank	EMBL
5,711,070	28,472,631	40,620,497



(A) DDBJ (B) Genbank (C) EMBL

FIGURE 1. Number of "Homo sapiens" sequences deposited in each databases

Question 2. What is the GC-content of the **AY330867**?

Solution. Firstly I searched for the sequence, whose accession number is AY33087, on the **ENA (European Nucleotide Archive)**. On the right portion of the search, there is an option to download the **FASTA file**, which I downloaded. Later, I uploaded it to the **Seq2Feature** tool to compute the nucleotide content. The GC-content of **AY330867** is **46.527778%**

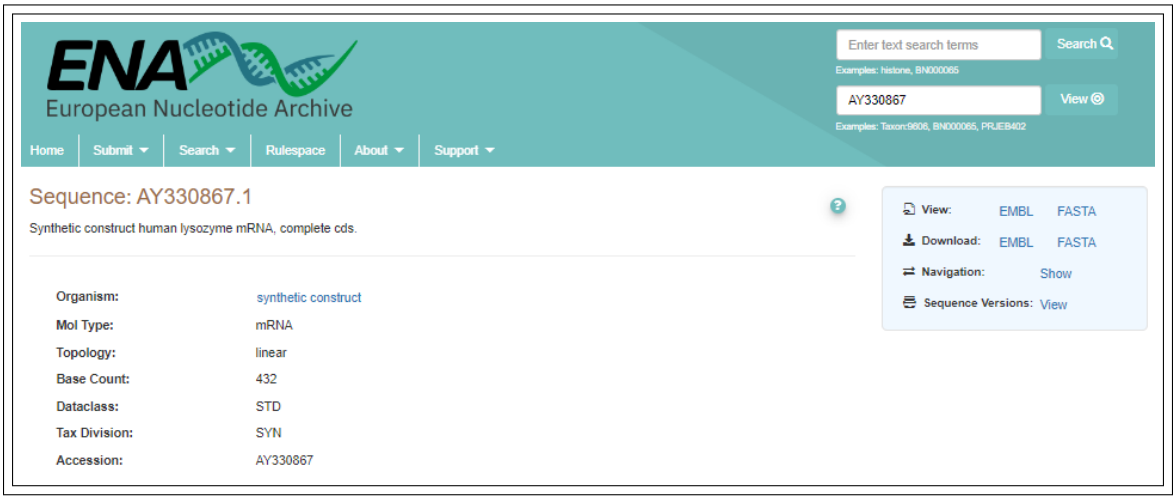


FIGURE 2. Text search on the ENA browser

Your input seq is:

ATGAGGAAGAAGCGGAGACAGCGACGAAGAATGAAGAAGGTCTTTGAAAGGTGTGAGTTGGCCAGAACTCTGAAAAGATTGGGAATGGATGGCTACAGGGGAATCAGCCTAGCAAACTGGATGTGTTTGGCCA
AATGGGAGAGTGGTTACAACACACGAGCTACAAACTACAATGCTGGAGACAGAAGCACTGATTATGGGATATTTCAGATCAATAGCCGCTACTGGTGTAATGATGGCAAAACCCCAGGAGCAGTTAATGCCTGT
CATTTATCCTGCAGTGCTTTGCTGCAAGATAACATCGCTGATGCTGTAGCTTGTGCAAAGAGGGTTGTCCGTGATCCACAAGGCATTAGAGCATGGGTGGCATGGAGAAATCGTTGTCAAAACAGAGGTGTCCGT
CAGTATGTTCAAGGTTGTGGAGTGTAATGA

Nucleotide Content:

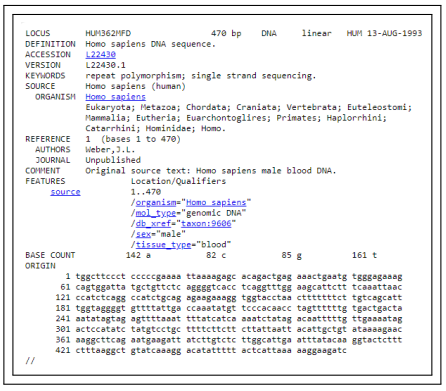
	Nucleotide content in %
AT_content	53.472222
Adenine_content	30.324074
Cytosine_content	16.666667
GC_content	46.527778
Guanine_content	29.861111
Keto_GT_content	53.009259
Purine_AG_content	60.185185
Thymine_content	23.148148

FIGURE 3. Nucleotide content on the Seq2Feature tool

Question 3. Compare the contents in **DDBJ**, **Genbank** and **EMBL**

Solution. The contents in DDBJ database are as follows:

- Locus
- Name
- Source
- Accession number
- Version number
- Keywords
- Authors
- Reference
- Journal
- PUBMED index
- Features that include exon, variation, gaps, mRNA
- Nucleotide sequence
- Frequency of bases (A,T,C,G)
- Protein sequence (translated)

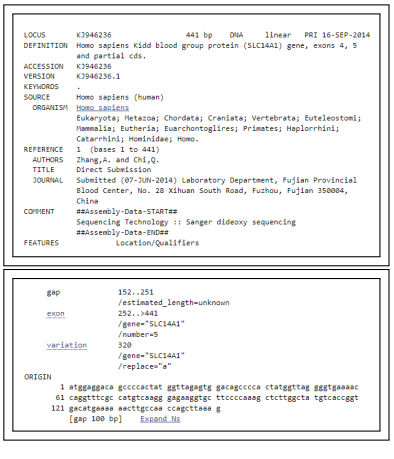


```
LOCUS       HUM362MFD               470 bp    DNA     linear   HUM 13-AUG-1993
DEFINITION  Homo sapiens DNA sequence.
ACCESSION   L22430
VERSION     L22430.1
KEYWORDS    repeat polymorphism; single strand sequencing.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 470)
  AUTHORS   Weber,J.L.
  JOURNAL   Unpublished
COMMENT     Original source text: Homo sapiens male blood DNA.
FEATURES             Location/Qualifiers
     source          1..470
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /sex="male"
                     /tissue_type="blood"
BASE COUNT          142 a           82 c           85 g          161 t
ORIGIN
        1 tggcttccct cccccgaaaa ttaaaagagc acagactgag aaactgaatg tgggagaaag
       61 cagtggatta tgctgttctc aggggtcacc tcaggtttgg aagcattctt tcaaattaac
      121 ccatctcagg ccatctgcag agaagaaagg tggtacctaa cttttttct tgtcagcatt
      181 tggtaggggt gttttattga ccaaatatgt tcccacaacc tagttttttg tgactgacta
      241 aatatagtag agtttttaaat tttatcatca aaatctatag acaatttttg ttgaaaatag
      301 actccatatc tatgtcctgc ttttcttctt cttattaatt acattgctgt ataaaagaac
      361 aaggcttcag aatgaagatt atcttgtctc ttggcattga atttatacaa ggtactcttt
      421 ctttagggct gtatcaaagg acatattttt actcattaaa aaggaagatc
//
```

FIGURE 4. DDBJ

The contents in Genbank database are as follows:

- Locus
- Definition
- Source
- Accession number
- Version number
- Keywords
- Title
- Authors
- Reference
- Journal
- PMID
- Features that include exon, variation, gaps, mRNA
- CDs (Coding sequence)
- Frequence of bases (A,T,C,G)
- Protein sequence (translated)

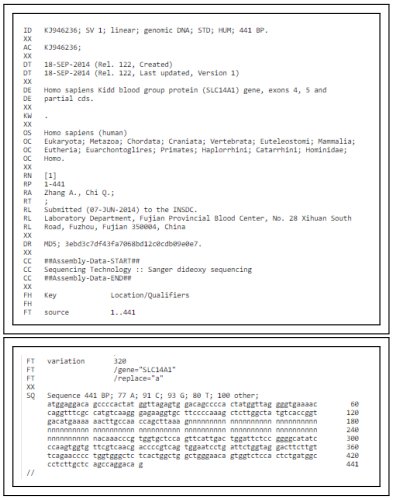


```
LOCUS       KJ946236                441 bp    DNA     linear   PRI 16-SEP-2014
DEFINITION  Homo sapiens Kidd blood group protein (SLC14A1) gene, exons 4, 5
            and partial cds.
ACCESSION   KJ946236
VERSION     KJ946236.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 441)
  AUTHORS   Zhang,A. and Chi,Q.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-JUN-2014) Laboratory Department, Fujian Provincial
            Blood Center, No. 28 Xihuan South Road, Fuzhou, Fujian 350004,
            China
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     gap             152..251
                     /estimated_length=unknown
     exon            252..441
                     /gene="SLC14A1"
                     /number=5
     variation       320
                     /gene="SLC14A1"
                     /replace="a"
ORIGIN
        1 atggaggaca gccccactat ggttagagtg gacagcccca ctatggttag gggtgaaaac
       61 caggtttcgc catgtcaagg gagaaggtgc ttccccaaag ctcttggcta tgtcaccggt
      121 gacatgaaaa aacttgccaa ccagcttaaa g
          [gap 100 bp]    Expand Ns
```

FIGURE 5. Genbank

The contents in EMBL database are as follows:

- Identification line
- Source
- Accession number
- Sequence version
- Date of first entry
- Description
- Keywords
- Organism species
- Organism classification
- References
- Author
- CDs (Coding sequence)
- Repeat regions
- Frequence of bases (A,T,C,G)
- Protein sequence (translated)



```
ID   KJ946236; SV 1; linear; genomic DNA; STD; HUM; 441 BP.
XX
AC   KJ946236;
XX
DT   18-SEP-2014 (Rel. 122, Created)
DT   18-SEP-2014 (Rel. 122, Last updated, Version 1)
XX
DE   Homo sapiens Kidd blood group protein (SLC14A1) gene, exons 4, 5 and
DE   partial cds.
XX
KW   .
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-441
RA   Zhang A., Chi Q.;
RT   ;
RL   Submitted (07-JUN-2014) to the INSDC.
RL   Laboratory Department, Fujian Provincial Blood Center, No. 28 Xihuan South
RL   Road, Fuzhou, Fujian 350004, China
XX
DR   MD5; 3ebd3c7df43fa7068bd12c0cdb09e0e7.
XX
CC   ##Assembly-Data-START##
CC   Sequencing Technology :: Sanger dideoxy sequencing
CC   ##Assembly-Data-END##
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..441
```



```
FT   variation       320
FT                   /gene="SLC14A1"
FT                   /replace="a"
XX
SQ   Sequence 441 BP; 77 A; 91 C; 93 G; 80 T; 100 other;
     atggaggaca gccccactat ggttagagtg gacagcccca ctatggttag gggtgaaaac       60
     caggtttcgc catgtcaagg gagaaggtgc ttccccaaag ctcttggcta tgtcaccggt      120
     gacatgaaaa aacttgccaa ccagcttaaa gnnnnnnnnn nnnnnnnnnn nnnnnnnnnn      180
     nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn      240
     nnnnnnnnnn nacaaacccg tggtgctcca gttcattgac tggatttctcc gggccatatc      300
     ccagtggtgg ttcgtcaacg acccggtcag tggaatcctg attctggtag gacttcttgt      360
     tcagaacccc tggtgggctc tcactggctg gctgggaaca gtggtctcca ctctgatggc      420
     cctcttgctc agccaggaca g                                                441
//
```

FIGURE 6. EMBL

Question 4. Get the papers about **discrimination of beta-barrel membrane proteins.** https://pubmed.ncbi.nlm.nih.gov/

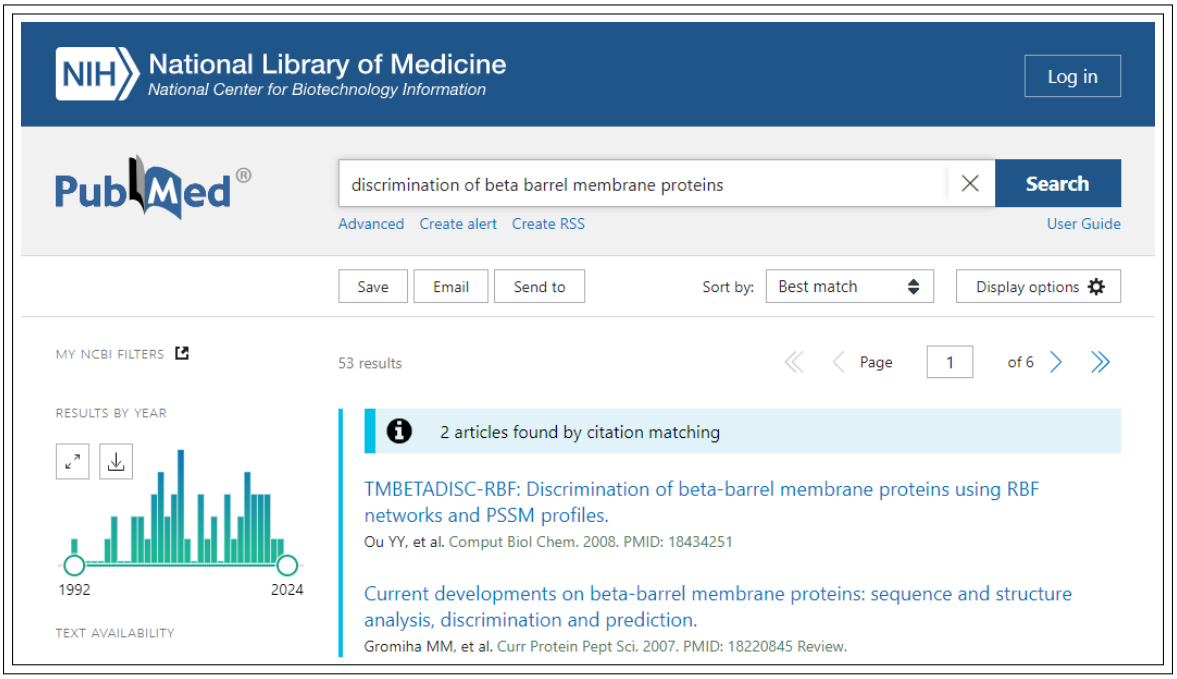


FIGURE 7. PUBMED papers about "discrimination of beta barrel membrane proteins

Solution. The list of papers is:

- TMBETADISC-RBF: Discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles
- Current developments on beta-barrel membrane proteins: sequence and structure analysis, discrimination and prediction
- BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane beta-barrel Proteins
- Current developments on beta-barrel membrane proteins: sequence and structure analysis, discrimination and prediction.

  and many more

Question 5. Find the papers published by any author (E.g.) Kihara D

Solution. The author chosen is **Daisuke Kihara**

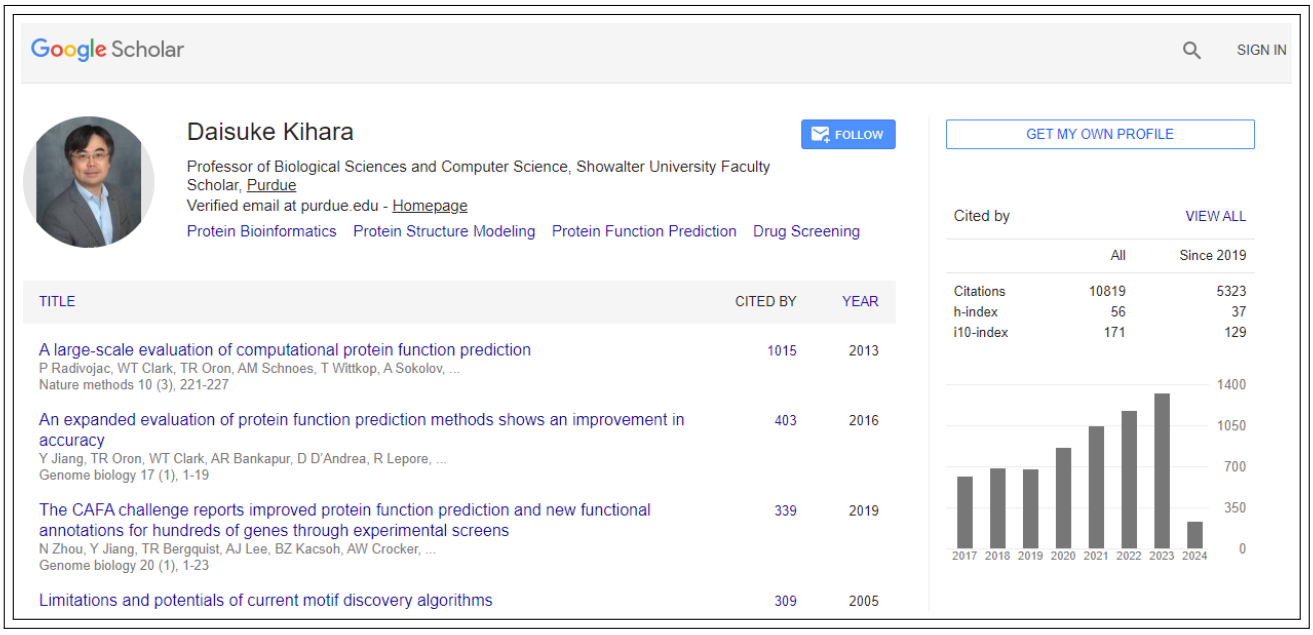


FIGURE 8. Google scholar search for Daisuke Kihara

The list of his papers is:

- A large-scale evaluation of computational protein function prediction
- An expanded evaluation of protein function prediction methods shows an improvement in accuracy
- Limitations and potentials of current motif discovery algorithms
- Protein-protein docking using region-based 3D Zernike descriptors

and many more

Question 6. How many related articles are listed in PUBMED for the paper, **Cell 2008 Dec 26;135(7):1158-9?**

FIGURE 9. Similar articles to Ceel 2008 Dec26;135(7):1158-9?

Solution. The PMID number is 19109882. There are 27 similar articles. Some of them have been listed below:

- Evolution. Tinkering inside the organelle.
- Transport proteins (carriers) of mitochondria.
- Systematic analysis of the twin cx(9)c protein family.
- Mitochondrial permeability transition pore opening as a promising therapeutic target in cardiac diseases.

  and many more

Question 7. List the papers published in the journal "**Nature**" for the year 2024. Check the list in **SCOPUS** and **PUBMED**

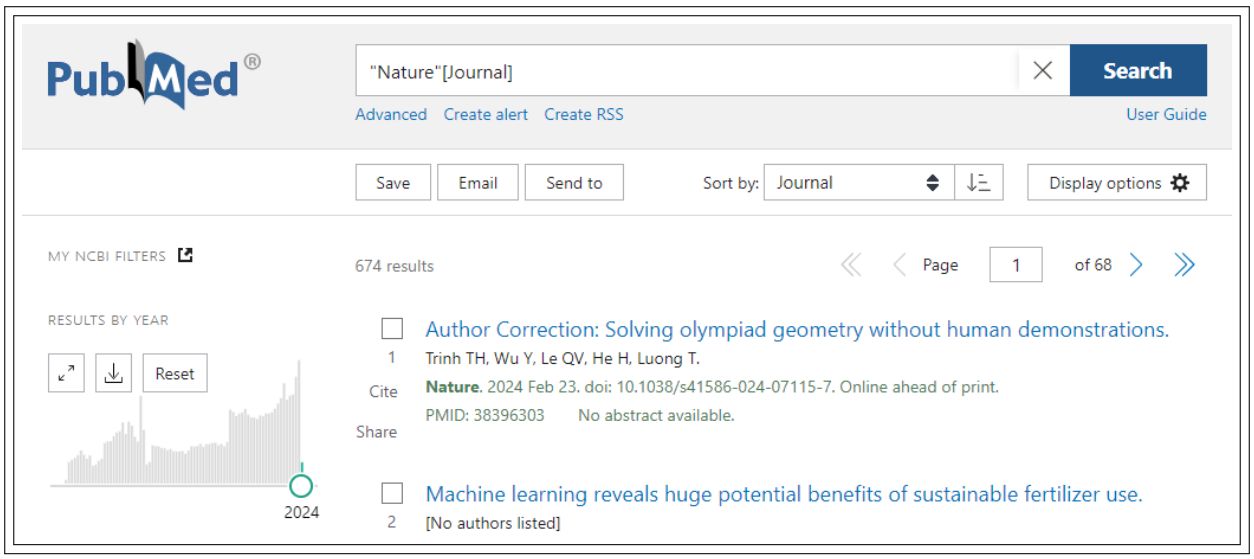


FIGURE 10. Nature publications on PUBMED

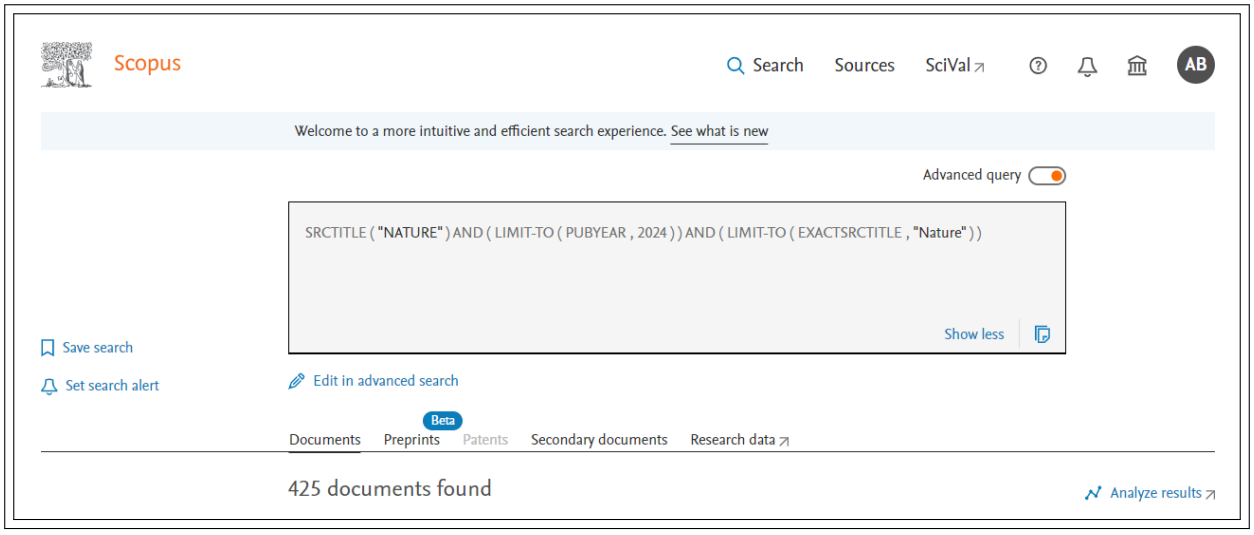


FIGURE 11. Nature publications on SCOPUS

Advanced query search

PUBMED has **674** publications
SCOPUS has **425** publications

Question 8. Find the h-index and number of citations for "**Burkhard Rost**"

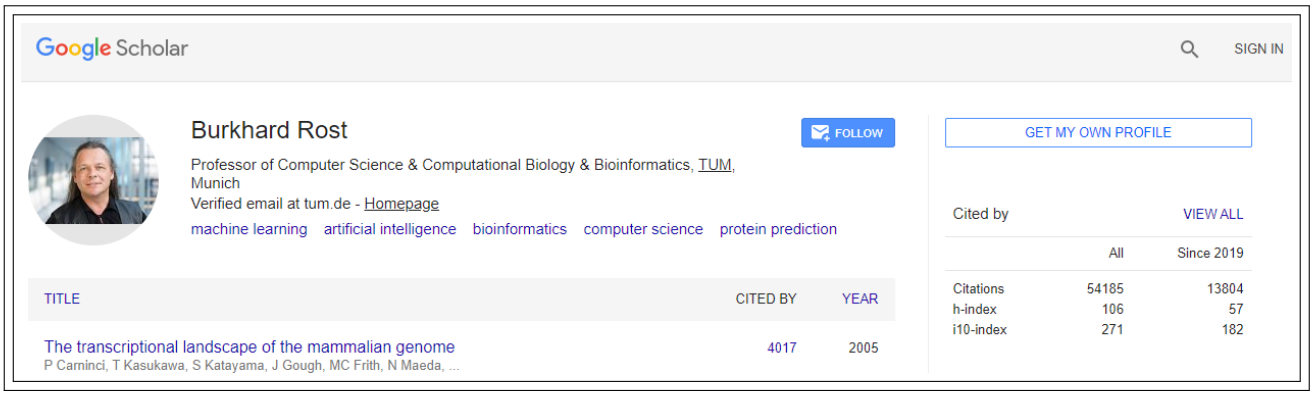


FIGURE 12. Google Scholar search for Burkhard Rost

Solution. The h-index of Burkhard Rost is **106**
The total number of citations is **54185**

Question 9. Find the class of the enzyme EC 1.7.2.3 and its function
http://www.brenda-enzymes.org/

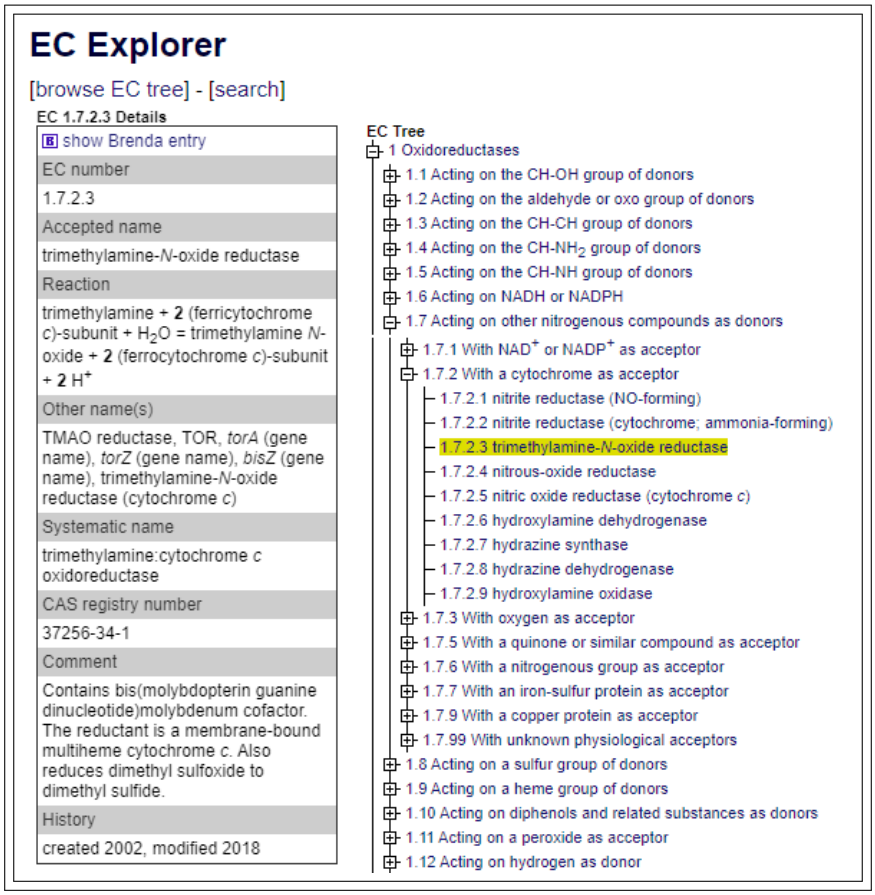
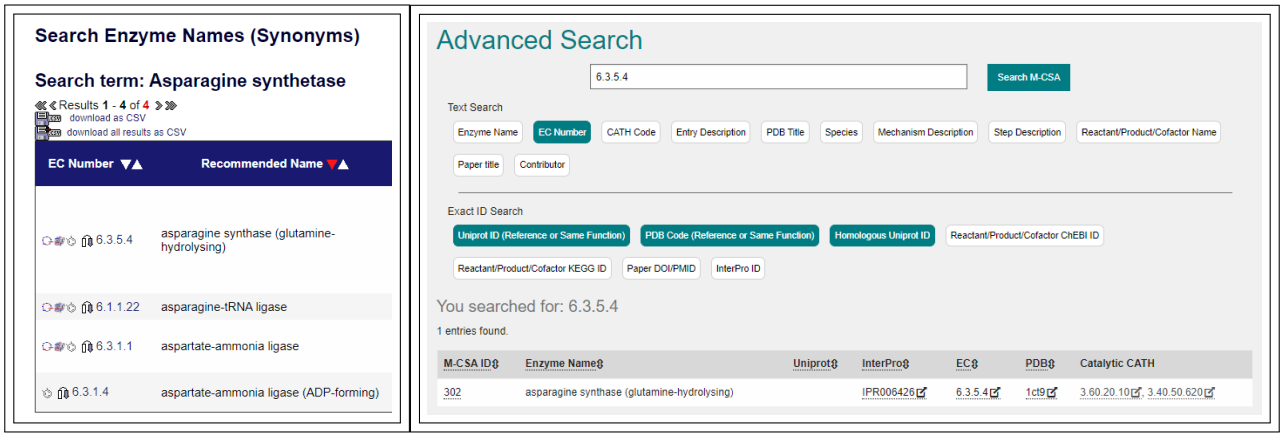


FIGURE 13. BRENDA search for enzyme EC 1.7.2.3

Solution. EC 1.7.2.3 is the EC number for the enzyme **trimethylamine-N-oxide reductase**. It is an **Oxidoreductase**. It contains a bis (molybdopterin guanine dinucleotide) molybdenum cofactor. The reductant is a membrane-bound multiheme cytochrome c. It reduces dimethyl sulfoxide to dimethyl sulfide.

Question 10. Find the catalytic site residues in **Asparagine synthetase**. *Hint*: Find the EC number and search in Catalytic site atlas https://www.ebi.ac.uk/thornton-srv/m-csa/



(A) BRENDA enzyme search (B) Catalytic site atlas search

FIGURE 14. Seq2Feature tool for nucleotide content

Solution. First I searched for the enzyme on the BRENDA to get its EC number. The EC number of **Asparagine synthetase** is **6.3.5.4**. On searching the EC number on the catalytic site atlas, in the mechanism sector, I found the catalytic site residues. They are

- **Cys2 (N-term)**
- **Leu51 (main-C)**
- **Thr322, Arg325**
- **Cys2**
- **Gly76 (main-N), Asn75**

Catalytic Residues Roles

UniProt	PDB* (1ct9)		
Cys2 (N-term)	Ala1A (N-term)	Acts as a general acid/base to activate the cysteine nucleophile.	proton acceptor, proton donor
Leu51 (main-C)	Leu50A (main-C)	Helps stabilise the reactive intermediates formed.	hydrogen bond acceptor, electrostatic stabiliser
Thr322, Arg325	Thr321A, Arg324A	Bind and stabilise the phosphate groups of the ATP and reactive intermediates formed.	hydrogen bond donor, electrostatic stabiliser
Cys2	Ala1A	Acts as a catalytic nucleophile in the glutaminase domain reaction.	covalently attached, hydrogen bond acceptor, nucleofuge, nucleophile, proton acceptor, proton donor
Gly76 (main-N), Asn75	Gly75A (main-N), Asn74A	Forms the oxyanion hole that stabilises the reactive intermediates and transition states formed.	hydrogen bond donor, electrostatic stabiliser

FIGURE 15. Catalytic site atlas residue for EC 6.3.5.4

Question 11. Find the scientific name, the taxonomy ID, and the number of chromosomes for the following organisms:
http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root
Organisms: **Human**, **Cat**, **Dog**, **domestic guinea pig**, and **Thale cress**

Solution. Below, I have tabulated each of the properties for each of the given organisms.

Human

Scientific Name: *Homo sapiens*
Taxonomy ID: 9606
Number of chromosomes: 46

Cat

Scientific Name: *Felis catus*
Taxonomy ID: 9685
Number of chromosomes: 38

Dog

Scientific Name: *Canis lupus familiaris*
Taxonomy ID: 9615
Number of chromosomes: 78

Domestic guinea pig

Scientific Name: *Cavia porcellus*
Taxonomy ID: 10141
Number of chromosomes: 64

Thale cress

Scientific Name: *Arabidopsis thaliana*
Taxonomy ID: 3702
Number of chromosomes: 10

The properties have been computed by the **Taxonomy browser**. The number of chromosomes could be extracted from any other database too.

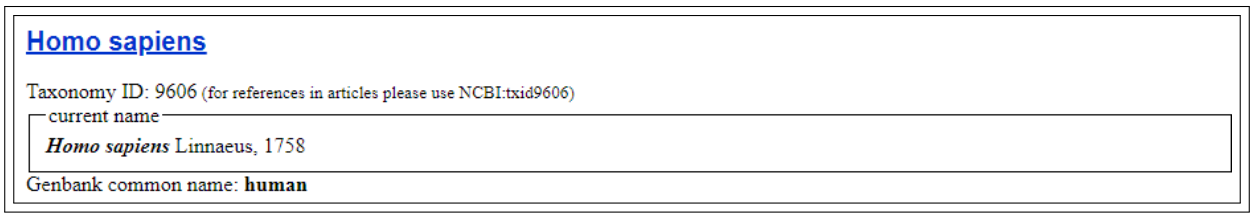


FIGURE 16. Taxonomy Browser: Human

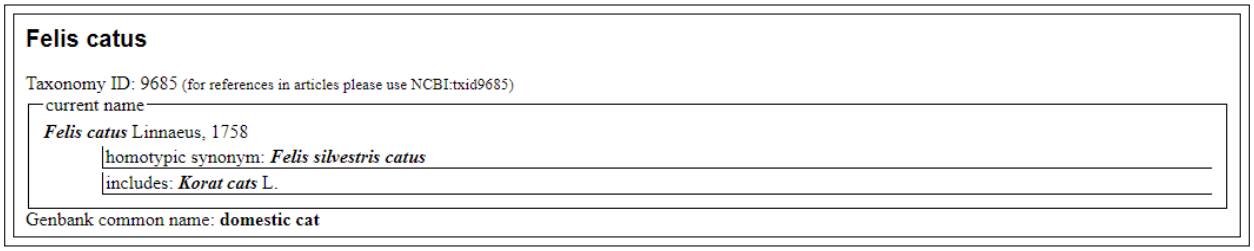


FIGURE 17. Taxonomy Browser: Cat

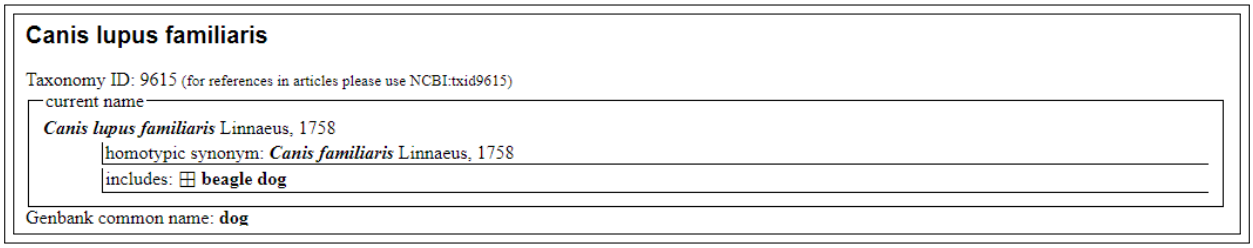


FIGURE 18. Taxonomy Browser: Dog

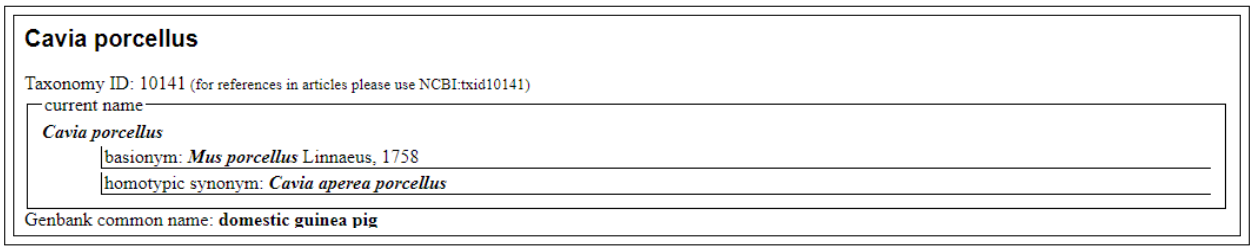


FIGURE 19. Taxonomy Browser: domestic guinea pig

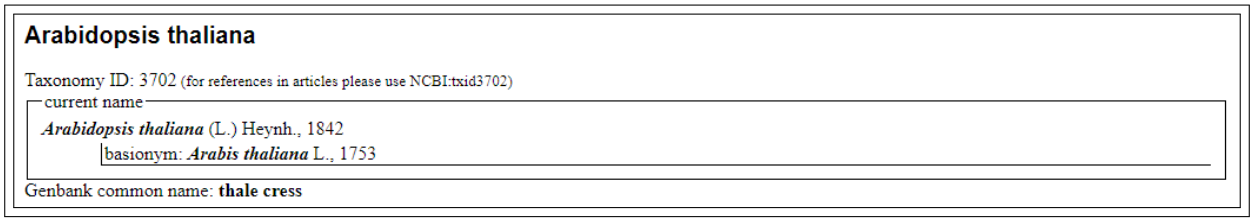


FIGURE 20. Taxonomy Browser: Thale cress

Question 12. What are NCBI E-utilities? Give the syntax for fetching a record in FASTA format using E-utilities.
http://www.ncbi.nlm.nih.gov/books/NBK25500/

Solution. NCBI E-utilities are a set of tools that allow you to access and manipulate data from the **NCBI databases**, such as **PubMed**, **PMC**, **Gene**, **Nuccore** and **Protein**. You can use E-utilities to perform tasks such as searching, linking, downloading, and converting data formats. E-utilities use a fixed URL syntax for their operations, and can be called from a web browser, a command line, or a custom program. Syntax for fetching a record in FASTA format:

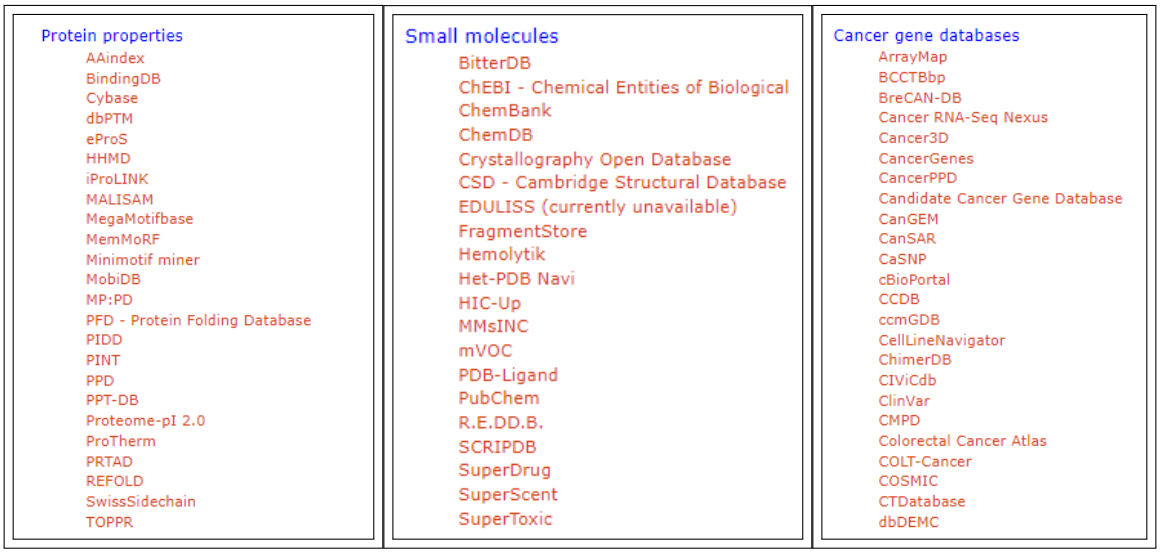
Fetch a record in FASTA format

```
efetch.fcgi?db=<database>&id=<uid_list>&rettype=fasta&retmode=<
    retrieval_mode>
# Entrez database (&db)
# List of UIDs (&id)
# Retrieval type (&rettype), [here fasta]
# Retrieval mode (&retmode)

# Example
efetch.fcgi?db=nuccore&id=34577062,24475906&rettype=fasta&retmode=text
# Database is nuccore
# UIDs are 34577062 and 24475906
# rettype is fasta
# retmode is text
```

Question 13. List two databases under each of the following category.
a. Protein properties
b. Small molecules (Structure related)
c. Cancer gene databases
Hint: Use Nucleic Acids Research (NAR) – 'database category list'
https://www.oxfordjournals.org/*our_journals*/nar/database/c/

Solution. Below are the databases available in each of the categories specified in the question:



(A) Protein properties (B) Small molecules (C) Cancer gene

FIGURE 21. Database Category Lists

Protein Properties

AAindex
BindingDB
Cybase

Small molecules (structure related)

BitterDB
ChemBank
Crystallography Open Database

Cancer gene databases

BCCTBbp
BreCAN-DB
Cancer RNA-Seq Nexus