Assignment 2

1. To find the number of “Homo sapiens” sequences deposited in DDBJ, Visit the given site http://www.ddbj.nig.ac.jp/ which takes you to DDBJ home site. Click “Services”, select “database” in the tags section and select “DDBJ Search” and leaving the search field as blank, select the organism as “Homo sapiens” and click enter. We get a total of 5079578 sequences of homo sapiens in DDBJ database. Similar process is followed in GenBank and in EMBL and the number of sequences found in GenBank is 3,106,484 (for nucleotides) and 13,436,072 (for proteins) and the number of sequences found in EMBL is 34,297,208.

2. Using Seq2Feature, the GC_content for the given sequence would be as follows:
   ```
   AY330867.1
   46.527778
   ```

3. By comparing the contents of DDBJ, GenBank, and EMBL, one thing that came clear is that GenBank has the most organised and vast amount of data and many search options to choose throughout the ontology of the database.
   The ontology of DDBJ is:
   ```
   LOCUS
   DEFINITION
   ACCESSION
   VERSION
   KEYWORDS
   SOURCE
         ORGANISM
   REFERENCE
         AUTHORS
         TITLE
         JOURNAL
   COMMENT
   FEATURES
   BASE COUNT
   ORIGIN
   ```
   The advanced search of EMBL works as follows:
   ```
   Data Type → Query → Inclusion/Exclusion → Fields → Data Filters
   ```
   The search categories in GenBank are given in the following link:
   Search: all[filter] - NLM (nih.gov)

4. There are two papers found by citation matching of the topic “discrimination of beta barrel membrane proteins”, the citations of those papers are:

   i. Ou, Y. Y., Gromiha, M. M., Chen, S. A., & Suwa, M. (2008). TMBETADISC-RBF: Discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles. Computational biology and chemistry, 32(3), 227–231. https://doi.org/10.1016/j.compbiolchem.2008.03.002

ii. Gromiha, M. M., & Suwa, M. (2007). Current developments on beta-barrel membrane proteins: sequence and structure analysis, discrimination and prediction. Current protein & peptide science, 8(6), 580–599. https://doi.org/10.2174/138920307783018712

5. List of Papers published by Elofsson A. can be found in the link given below:
Elofsson A - Search Results - PubMed (nih.gov)
There are 157 results found with the given author

6. When searched for the paper "Cell 2008 Dec 26;135(7):1158-9", the result is a paper which citation is given below

i. Imai, K., Gromiha, M. M., & Horton, P. (2008). Mitochondrial beta-barrel proteins, an exclusive club?. Cell, 135(7), 1158–1160. https://doi.org/10.1016/j.cell.2008.12.017

When scrolled down for the related articles, there is the section "Similar Articles" which contained the related articles which the citations are given below:

ii. Alcock, F., Clements, A., Webb, C., & Lithgow, T. (2010). Evolution. Tinkering inside the organelle. Science (New York, N.Y.), 327(5966), 649–650. https://doi.org/10.1126/science.1182129

iii. Wohlrab H. (2009). Transport proteins (carriers) of mitochondria. IUBMB life, 61(1), 40–46. https://doi.org/10.1002/iub.139

iv. Walther, D. M., Papic, D., Bos, M. P., Tommassen, J., & Rapaport, D. (2009). Signals in bacterial beta-barrel proteins are functional in eukaryotic cells for targeting to and assembly in mitochondria. Proceedings of the National Academy of Sciences of the United States of America, 106(8), 2531–2536. https://doi.org/10.1073/pnas.0807830106

v. Longen, S., Bien, M., Bihlmaier, K., Kloeppel, C., Kauff, F., Hammermeister, M., Westermann, B., Herrmann, J. M., & Riemer, J. (2009). Systematic analysis of the twin cx(9)c protein family. Journal of molecular biology, 393(2), 356–368. https://doi.org/10.1016/j.jmb.2009.08.041

vi. Mullauer, F. B., Kessler, J. H., & Medema, J. P. (2009). Betulinic acid induces cytochrome c release and apoptosis in a Bax/Bak-independent, permeability transition pore dependent fashion. Apoptosis : an international journal on programmed cell death, 14(2), 191–202. https://doi.org/10.1007/s10495-008-0290-x

Note that these are the most similar articles related to this paper, there are a total of 28 papers related to this article.

7. When checked for the list in SCOPUS, the list of journals published in “Nature” in the year 2022 are 2,199. From PubMed, there are a total of 401 Journals Published in “Nature” in the year 2022. The list of the papers are given below in links
   PubMed: "Nature"[Journal] - Search Results - PubMed (nih.gov)
   SCOPUS: SRCTITLE(Nature) AND (LIMIT-TO (PUBYEAR,2022)) – Search - SCOPUS

8. For the author Rost Buckhard, the h-index is 82 and the number of citations is 31,092.

9. Using Brenda, the enzyme EC 3.4.11.4 is Tripeptide Aminopeptidase which EC tree is:
   ```
   3 = Hydrolases
   3.4 = Acting on peptide Bonds (peptidases)
   3.4.11 = Aminopeptidases
   3.4.11.4 = Tripeptide aminopeptidase
   ```
   The function of this enzyme is to form a tripeptide by releasing the N-terminal residue.

10. catalytic site residues in Asparagine synthetase are:

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

11. The Scientific name, taxonomy ID and number of chromosomes of the given organisms are as follows:

Organism name	**Scientific Name**	**Taxonomy ID**	**Number of Chromosomes**
Human	Homo sapiens	9606	46
Cat	Felis catus	9685	38
Dog	Canis lupis familiaris	9615	78
House Mouse	Mus musculus	10090	40
Onion	Allium cepa	4679	16
Thale cress	Arabidopsis thaliana	3702	10

12. NCBI E-Utilities (Entrez Programming Utilities) have a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search and retrieve the requested data.
    All NCBI Utilities begin with the following string:
    https://eutils.ncbi.nlm.nih.gov/entrez/eutils/
    These URLs are direct requests to servers that are used only by the E-Utilities and that are optimized to give users the best performance
    For basic downloading of full records, the following syntax is used:
    `efetch.fcgi?db=<database>&id=<uid_list>&rettype=<retrieval_type>&retmode=<retrieval_mode>`
    for downloading some data in FASTA format, the following syntax is used:
    `efetch.fcgi?db=<database>&id=<uid_list>&rettype=`**`fasta`**`&retmode=<retrieval_mode>`
    (in the above syntax, the change of the basic syntax is bolded with "FASTA" signifying that when the word "fasta" is inserted in that place we would get the information in FASTA format)

13. Two databases for the following categories:
    a. Protein Properties:
        i. Cybase
        ii. eProS

    b. Small Molecules (structure related):
        i. ChemBank
        ii. Hemolytik

    c. Cancer gene databases:
        i. ArrayMap
        ii. CancerGenes