CHAPTER 5

Using Computational Methods on Biological Databases

Yunus Emre GOKTEPE
Necmettin Erbakan University, Türkiye

Yusuf UZUN
Necmettin Erbakan University, Türkiye

Tayfun TASDEMIR
Necmettin Erbakan University, Türkiye

To Cite This Chapter:
Goktepe, Y.E., Uzun, Y. & Tasdemir, T. (2023). Using Computational Methods on Biological Databases. In H. Arikan, Y. Uzun & I. Cetin (Eds.), *Current Studies in Basic Sciences, Engineering and Technology 2023* (pp. 44-51). ISRES Publishing.

Introduction

The first basic analyses of proteins were made in 1828s. The molecules that Gerardus Johannes Mulder, a Dutch chemist, examined in these analyses were first given the name "protein", derived from the ancient Greek word πρωτεῖος (proteios), by the Swedish chemist Jöns Jakob Berzelius in 1838 (Braun and Gingras, 2012).

The first study on the interactions between proteins was reported by Hedin (Hedin, 1906), who examined the inhibition and kinetic properties of trypsin, an enzyme found in the digestive tract. Studies accepting that protein interactions have vital importance in the formation of enzyme activities required for organisms began to be carried out in the 1960s (Crick and Orgel, 1964). The importance of protein interactions was understood later, and studies have made significant progress with the development of the yeast two-hybrid protein interaction method (Fields and Song, 1989). This method was developed as the first of the protein interaction prediction methods called high-throughput because it works with a large number of inputs and outputs.

It is known that proteins play vital roles in all biological processes and are at the center of the processes (Zeng et al., 2020). They perform important functions by making interactions with other proteins and breaking those interactions. Finding these interactions is an important process and it takes too much time and cost to make predictions about

interacting proteins by in-vitro studies (Chen et al., 2019).

Computational methods can help and support biological experiments. There are plenty of studies to assist this research (Wang et al., 2020). For example, there are several types of research on protein structure prediction (Kulhman and Bradley, 2019), prediction of protein cellular attributes (Chou, 2001), and prediction of protein interactions (Wei et al., 2017; Goktepe & Kodaz, 2018).

In this study, a web-based model is offered to investigate sequences, get properties of proteins, and make predictions of possible interactions (Taşdemir, 2023). One capability of this model is to expedite protein search. It is also able to find proteins that resemble a given sequence partition. Researchers can gather selections of amino acid properties easily with this model and use them in their computational studies. Moreover, the interaction prediction facility of the model can be used by biologists to decide which protein pair should they choose to make an interaction test in the laboratory.

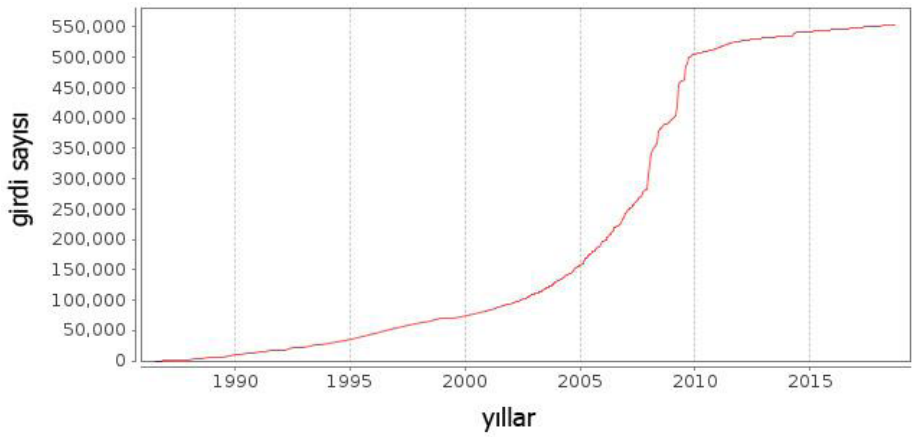
Databases

In recent years, a significant amount of protein interaction information has been obtained through studies. Protein interaction data revealed by researchers with different experimental or computational methods are added to databases managed by universities and institutes. Although there are over 100 databases created for this purpose, almost all of the data is in the few most widely used databases.

The data on protein sequence, function, and interactions obtained by new research are constantly increasing. UniProtKB/Swiss-Prot 2019_02 version published on 13 February 2019 by UniProt, a comprehensive database of protein sequences and functions, contains 559,228 sequence information. Figure 1 shows the increase in the number of sequence entries contained in the UniProtKB/Swiss-Prot database between 1985 and 2019 (UniProt, 2019).

Figure 1.

Increase of protein data in UniProtKB/Swiss-Prot database over years (UniProt, 2019)



With the motivation of technological advances, the studies presented by researchers interested in bioinformatics and biomedical issues are increasing day by day. One of the most important databases in which studies on this subject are kept is Medline (Medline, 2023). In 2022, 1,369,611 new citations were added to the Medline database established by the U.S. National Library of Medicine (NLM), and the increase in the number of citations included in the database over the years can be seen in Table 1.

Table 1.
Number of PUBMED production statistics (Medline, 2023)

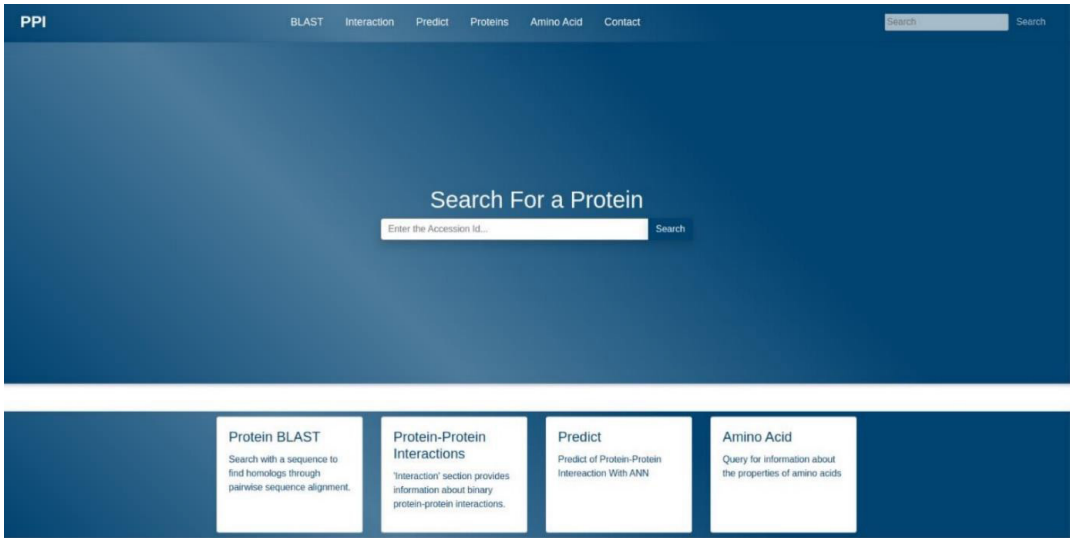
	FY2022	FY2021	FY2020	FY2019	FY2018
MEDLINE Citations Indexed (Annual)	1,369,611	1,291,807	952,919	956,39	904,636
MEDLINE Citations Cumulative Total	29,807,639	28,444,654	27,149,277	26,196,358	25,239,968
MEDLINE Journal Titles	5,282	5,282	5,274	5,243	5,251
PubMed Citations (Annual)	1,714,780	1,733,089	1,514,199	1,366,447	1,329,148
PubMed Citations Cumulative Total	34,693,538	33,136,289	31,563,992	30,178,674	28,934,389
PubMed Searches	2.58 Billion	2.57 Billion	3.3 Billion	3.1 Billion	3.3 Billion
Web/Interactive	1.283 Billion	1.186 Billion	1.076 Billion	896 Million	831 Million
Script/E-Utilities	1.303 Billion	1.391 Billion	2.2 Billion	2.2 Billion	2.5 Billion

It is often indicated that the known protein networks are far from being complete yet (Kotlyar et al., 2015). Since the number of data used in studies on this subject is constantly increasing, we can say that the need for information accession models such as the one suggested in this study has also increased.

Web-Based Protein Information Accession Model

On the main screen, there is a navigation bar as seen in Figure 2. The main screen of the project provides access to the offered features (Taşdemir, 2023). Using this bar, a blast search can be done, or some useful information can be accessed such as protein information, amino acid sequences, and interactions of proteins. Interaction predictions of a query protein can be obtained using the “Predict” menu of the bar.

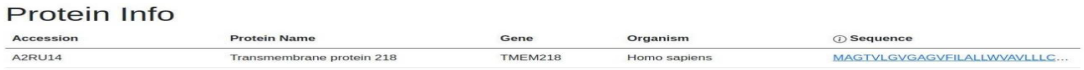
Figure 2.
The main screen of the project



Accessing Protein Data by ID

Information about a query protein can be obtained by entering its protein_id (accession) data at the center of the main screen. Accession code, full name, gene and organism information, and sequence data of the protein are shown in the result page (Figure 3). All of this information can be obtained in json form by clicking on the sequence data.

Figure 3.
Search result of a sample query protein

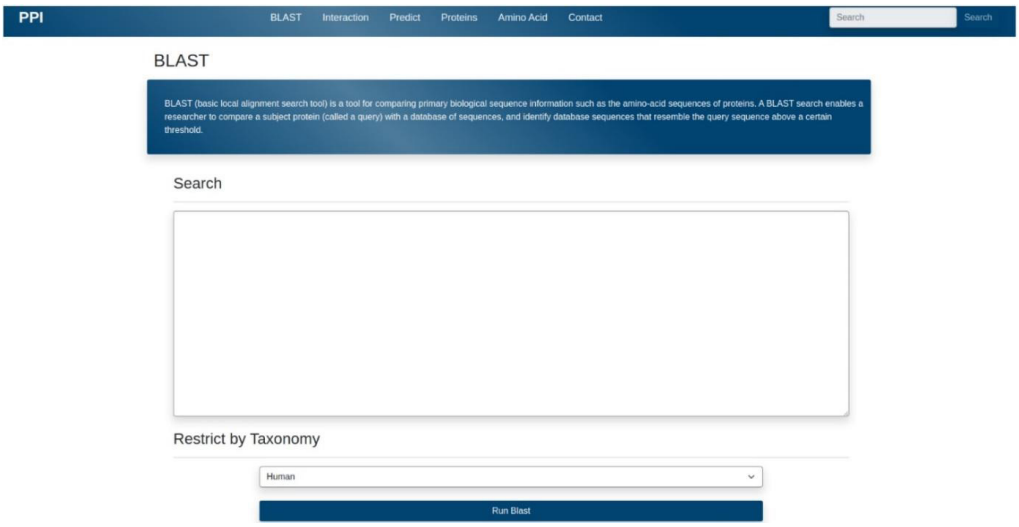


BLAST Search

On the BLAST (Basic Local Alignment Search Tool) screen (Figure 4), an algorithm was written that gives the similarity of protein sequences (Blast, 2023). This algorithm works as follows;

- Input a sequence (or part of a sequence),
- Choose an organism type,
- Compare all proteins with the one entered by the user.
- Proteins that are similar with a max-margin of difference of 5% to the query sequence are listed as output.

Figure 4.
BLAST search interface of the project



Searching for Known Interactions

Known interactions of a query protein can be listed via its ID (accession number) or full name. A list of interactions is listed on a result page. Moreover, interaction can be searched by giving a sequence or part of a sequence. In such a situation, proteins that involve a given sequence are taken into account and known interactions of these proteins are listed in the result page by using pagination (Figure 5).

Figure 5.
Interactions are listed on a result page with pagination

Interaction Info

Protein1		Protein2	
Accession1	Interactor1	Accession2	Interactor2
A1A4S6	EBI-1390944	Q6P5Z2	EBI-1384335
Q6P5Z2	EBI-1384335	A1A4S6	EBI-1390944
Q6P5Z2	EBI-1384335	Q9UNA1	EBI-1390913
Q6P5Z2	EBI-1384335	Q08379	EBI-618309
Q6P5Z2	EBI-1384335	Q15323	EBI-948001
Q6P5Z2	EBI-1384335	Q5JR59-3	EBI-11522433
Q6P5Z2	EBI-1384335	Q8ND90	EBI-302345
Q6P5Z2	EBI-1384335	Q9UBB9	EBI-1105213
Q6P5Z2	EBI-1384335	Q8N1B4	EBI-2799833
Q6Y5D8	EBI-4396535	Q29502	EBI-4406512
Q6Y5D8-1	EBI-4396677	Q29502	EBI-4406512
Q29502	EBI-4406512	Q6Y5D8	EBI-4396535
Q29502	EBI-4406512	Q6Y5D8-1	EBI-4396677
Q5T5U3	EBI-1642518	P29692	EBI-358607
Q5T5U3	EBI-1642518	P63104	EBI-347088

Showing 1 to 15 of 160 results

‹ 1 2 3 4 5 6 7 8 9 10 11 ›

AAIndex Properties of Amino Acids

AAindex is a widely used database that contains several indices about the physicochemical

and biochemical properties of amino acids. AAindex is a widely used database in research that contains several indices about the physicochemical and biochemical properties of amino acids. It consists of 3 parts, the first of which is AAIndex1 which contains an amino acid index of 20 numeric values. The latter is AAIndex2 which contains an amino acid mutation matrix and the last part is AAIndex3 gives statistical protein contact potentials (Kawashima et al., 2008).

In this section of the project, a sequence or part of a sequence is given, and requested amino acid properties or properties are selected. The system finds numerical values of requested properties and lists them as a result using pagination as seen in Figure 6.

Figure 6.
Numerical values of amino acid index properties

Seq / AA	ANDN920101	ARGP820101	ARGP820102	CHOC750101	CHOC760102	CHOC760103	CHOP780214	CHOP780215	CHOP780216
M	4.52	1.18	2.67	170.80	31.00	0.40	0.01	0.06	0.51
A	4.35	0.61	1.18	91.50	25.00	0.38	0.04	0.06	0.64
G	3.97	0.07	0.49	66.40	23.00	0.36	0.19	0.15	1.63
I	3.95	2.22	1.45	168.80	18.00	0.60	0.01	0.06	0.29
I	3.95	2.22	1.45	168.80	18.00	0.60	0.01	0.06	0.29
K	4.36	1.15	0.06	171.30	97.00	0.03	0.07	0.10	1.13
K	4.36	1.15	0.06	171.30	97.00	0.03	0.07	0.10	1.13
Q	4.37	0.00	0.72	161.10	71.00	0.07	0.04	0.10	0.84
I	3.95	2.22	1.45	168.80	18.00	0.60	0.01	0.06	0.29
L	4.17	1.53	3.23	167.90	23.00	0.45	0.04	0.07	0.36
K	4.36	1.15	0.06	171.30	97.00	0.03	0.07	0.10	1.13
H	4.63	0.61	0.31	167.30	43.00	0.17	0.09	0.05	0.77
L	4.17	1.53	3.23	167.90	23.00	0.45	0.04	0.07	0.36
S	4.50	0.05	0.97	99.10	44.00	0.22	0.13	0.11	1.52
R	4.38	0.60	0.20	202.00	90.00	0.01	0.10	0.09	1.05

Showing 1 to 15 of 60 results

‹ 1 2 3 4 ›

Making Predictions

In the prediction section, there is an algorithm that predicts about 2 query proteins to provide information to researchers on whether there is a high probability to be an interaction between them or not.

Interaction prediction is made by IDs of query proteins and the system gives a result that includes whether there is a possible interaction or not and the accuracy of this prediction as seen in Figure 7.

Figure 7.
Interaction prediction result of 2 query protein

PPI Predict Info

Accession1	Accession2	Accuracy	Predict Ratio	Result
A0PK11	Q8N6S5	94.39	100.00	Interaction

This prediction algorithm is developed with Python using the Keras module of

the Tensorflow library. Keras is a widely used high-level neural network library.

Conclusions

Proteins play an important role in almost every process in cells of organisms. Thus, there is a lot of research using them and researchers are in search of ways to access the properties of proteins. This study exhibits a new system that makes an effective search for proteins. It can make BLAST searches, enabling researchers to find similar proteins that resemble a query sequence with a %5 difference tolerance. It also gathers various physicochemical and biochemical properties of amino acid sequences which are widely used in biological and bioinformatics research.

Moreover, given 2 query proteins the proposed system makes predictions whether there could be a possible interaction between them or not.

Gathering information about proteins is labor-intensive work. It takes too much time to select a potential interacting protein pair in laboratory studies. This study would help researchers by searching proteins, gathering useful information about them, and making predictions about protein pairs.

References

Blast, “BLAST Developer Information”, (2023). https://blast.ncbi.nlm.nih.gov/.

Braun, P. and Gingras, A. C., 2012, History of protein-protein interactions: From egg-white to complex networks, Proteomics, 12 (10), 1478-1498.

Chen, C., Zhang, Q., Ma, Q., & Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. Chemometrics and Intelligent Laboratory Systems, 191, 54-64.

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics, 43(3), 246-255.

Crick, F. H. and Orgel, L. E., 1964, The Theory of Inter-Allelic Complementation, J Mol Biol, 8, 161-165.

Fields, S. and Song, O. K., 1989, A Novel Genetic System to Detect Protein Protein Interactions, Nature, 340 (6230), 245-246.

Göktepe, Y. E., & Kodaz, H. (2018). Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing, 303, 68-74.

Hedin, S. G., 1906, Trypsin and Antitrypsin, Biochem J, 1 (10), 474-483.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36, D202-D205. [PMID:17998252].

Kotlyar, M., Pastrello, C., Pivetta, F., Lo Sardo, A., Cumbaa, C., Li, H., Naranian, T., Ding, Z. Y., Vafaee, F., Broackes-Carter, F., Petschnigg, J., Mills, G. B., Jurisicova, A., Stagljar, I., Maestro, R. ve Jurisica, I., (2015). In silico prediction of physical protein interactions and characterization of interactome orphans, Nature Methods, 12 (1), 79-84.

Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and

design. Nature Reviews Molecular Cell Biology, 20(11), 681-697.

PubMed, 2023. National Center for Biotechnology Information, National Library of Medicine, https://www.ncbi.nlm.nih.gov/pubmed, 22.06.2023.

MEDLINE, 2023. MEDLINE PubMed Production Statistics https://www.nlm.nih.gov/bsd/ medline_pubmed_production_stats.html, 22.06.2023.

Taşdemir, T., 2023, Biyoinformatik Veriler Üzerinde Derin Öğrenme Uygulaması, Department of Computer Engineering, Seydisehir Ahmet Cengiz Engineering Faculty of Necmettin Erbakan University.

UniProt, 2019, https://www.uniprot.org/statistics/Swiss-Prot: [01.01.2019].

Wang, Y., You, Z., Li, L., & Chen, Z. (2020). A survey of current trends in computational predictions of protein-protein interactions. Frontiers of Computer Science, 14, 1-12.

Wei, L. Y., Xing, P. W., Zeng, J. C., Chen, J. X., Su, R. ve Guo, F., 2017, Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier, Artificial Intelligence in Medicine, 83, 67-74.

Zeng, M., Zhang, F., Wu, F. X., Li, Y., Wang, J., & Li, M. (2020). Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics, 36(4), 1114-1120.

Authors

Yunus Emre GOKTEPE, PhD, is an Assistant Professor at Necmettin Erbakan University. The author's areas of expertise are bioinformatics, artificial intelligence and machine learning. He is still serving as a faculty member of the Department of Computer Engineering, Seydisehir Ahmet Cengiz Engineering Faculty at Necmettin Erbakan University in Konya, Turkey.
E-mail: ygoktepe@erbakan.edu.tr,**ORCID:** 0000-0002-8252-2616

Yusuf UZUN, PhD, is an Assistant Professor of Computer Engineering at Necmettin Erbakan University in Konya, Turkey. He holds a PhD in Mechanical Engineering from Necmettin Erbakan University. His main areas of interest are artificial intelligence, autonomous systems and augmented reality applications. He also works as the Rector's Advisor at Selcuk University.
E-mail: mailto:yuzun@erbakan.edu.tr, **ORCID:**0000-0002-7061-8784

Tayfun TASDEMIR, is a Computer Engineer and working as a Backend Developer at MGS Software, Istanbul, Turkey. He has the bachelor's degree in the Department of Computer Engineering, Seydisehir Ahmet Cengiz Engineering Faculty at Necmettin Erbakan University in Konya, Turkey. His main areas of interest are artificial intelligence, web programming and machine learning.
E-mail: mailto:tayfuntasdemir3@gmail.com, **ORCID:** 0009-0009-4668-5280

Similarity Index
The similarity index obtained from the plagiarism software for this book chapter is 16%.