PROQUEST

https://proquest.ngrok.app

Large Language Model-Based Protein Sequence and Annotation Analysis System Developed for UniProt

Melike Akkaya, Rauf Yanmaz, Sezin Yavuz

Advisor: Prof. Dr. Tunca Doğan

Biological Data Science Laboratory, Department of Computer Engineering,
Hacettepe University, 06800, Ankara, Turkey

HACETTEPE UNIVERSITY

HACETTEPE U. BIOLOGICAL DATA SCIENCE LAB

EMBL-EBI

UniProt

Introduction

The need to query large biological databases effectively and intuitively in biomedical research is steadily increasing. Although traditional query tools offer accuracy, they limit user accessibility due to the high level of technical expertise required.

> *"We have always observed that users, especially those new to UniProt, tend to explore the database using natural language queries... By understanding the meaning of these queries and 'translating' them into advanced queries, this capability will help users reach the expected results."*
>
> — Aurélien Luciani (EMBL-EBI)

In this project, a user-friendly system based on a large language model (LLM) has been developed, enabling natural language querying of the UniProt[1] protein sequence and annotation database.

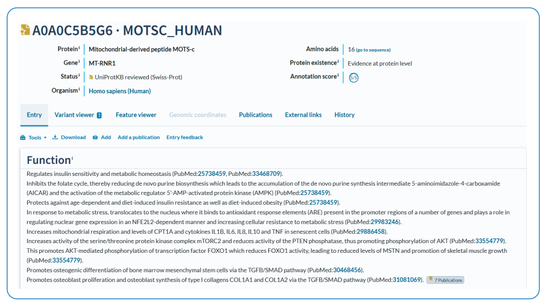


Figure 1: Example of Protein Information Page

1) LLM-Based Solr Query Generation:

User queries in natural language are automatically converted into Solr queries for the UniProt database.

- **Prompt Engineering:**
UniProt's query, search, and result fields were analyzed to enable Solr query generation by LLMs. Search and result data were moved to an **SQLite-based helper database** to optimize prompts within token limits.

- **Model Selection & Strategy:**
Advanced LLMs were tested. Query variability was reduced using multiple trials and adjustable temperature settings, offering flexibility to users.

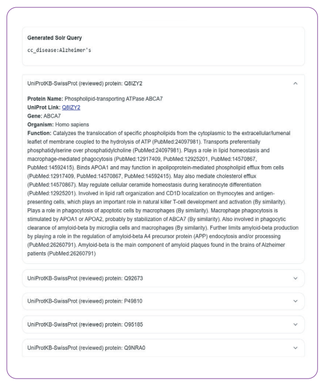


Figure 6

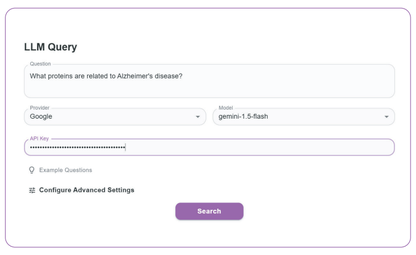


Figure 5

Users can interact with the system simply by writing questions in natural language, without needing to know the structure of the database or any query language (Figure 5).

The system translates these questions into technical queries suitable for UniProt, then analyzes the results and presents the most accurate answers in a clear and understandable way (Figure 6).

Figure 2 shows how difficult it is to understand the content of the data example used when setting up the vector database at this stage.

```
ID   MOTSC_HUMAN             Reviewed;          16 AA.
AC   A0A0C5B5G6;
DT   13-APR-2016, integrated into UniProtKB/Swiss-Prot.
DT   29-APR-2015, sequence version 1.
DT   02-OCT-2024, entry version 28.
DE   RecName: Full=Mitochondrial-derived peptide MOTS-c {ECO:0000303|PubMed:25738459};
DE   AltName: Full=Mitochondrial open reading frame of the 12S rRNA-c {ECO:0000303|PubMed:25738459};
GN   Name=MT-RNR1 {ECO:0000312|HGNC:HGNC:7470};
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
OX   NCBI_TaxID=9606 {ECO:0000312|EMBL:AJM13597.1};
RN   [1] {ECO:0000312|EMBL:AJM13597.1}
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA], FUNCTION, SUBCELLULAR LOCATION, TISSUE
RP   SPECIFICITY, AND MUTAGENESIS OF GLU-5 AND GLY-7.
RX   PubMed=25738459; DOI=10.1016/j.cmet.2015.02.009;
RA   Lee C., Zeng J., Drew B.G., Sallam T., Martin-Montalvo A., Wan J.,
RA   Kim S.-Y., Mehta H., Hevener A.L., de Cabo R., Cohen P.;
RT   "The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis
RT   and reduces obesity and insulin resistance.";
RL   Cell Metab. 21:443-454(2015).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RX   PubMed=7219534; DOI=10.1038/290457a0;
RA   Anderson S., Bankier A.T., Barrell B.G., de Bruijn M.H.L., Coulson A.R.,
RA   Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H.,
RA   Smith A.J.H., Staden R., Young I.G.;
RT   "Sequence and organization of the human mitochondrial genome.";
RL   Nature 290:457-465(1981).
RN   [3]
RP   ROLE IN MRSA INFECTION.
RX   PubMed=29096170; DOI=10.1016/j.molimm.2017.10.017;
RA   Zhai D., Du J., Zhang Y., Xu C., Ruan B., Yang Y., Lei X., Xiang A., Lu H.,
RA   Zhu Y., Yan Z., Hou Z., Li Q., Wang L., Lv J.;
RT   "MOTS-c peptide increases survival and decreases bacterial load in mice
RT   infected with MRSA.";
RL   Mol. Immunol. 92:151-160(2017).
RN   [4]
RP   FUNCTION, AND INDUCTION.
RX   PubMed=29884458; DOI=10.18632/aging.101463;
RA   Kim S.J., Mehta H.H., Wan J., Kuehnemann C., Chen J., Hu J.F.,
RA   Hoffman A.R., Cohen P.;
RT   "Mitochondrial peptides modulate mitochondrial function during cellular
RT   senescence.";
```

Figure 2: Sample flat file content

Figure 10 shows that users can optionally provide a protein sequence along with their question, without needing to know the query structure.

Figure 11 presents the resulting list of the most relevant and logical answers generated in response to such a query.

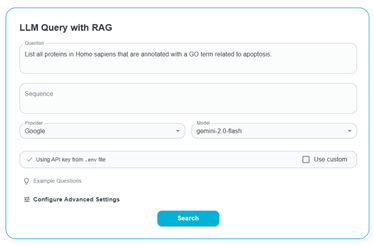
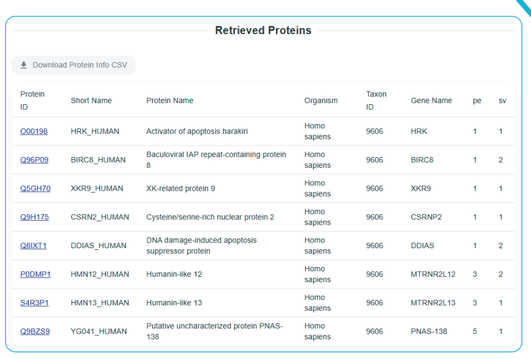


Figure 10



Protein ID	Short Name	Protein Name	Organism	Taxon ID	Gene Name	pe	sv
O00198	HRK_HUMAN	Activator of apoptosis harakiri	Homo sapiens	9606	HRK	1	1
Q96P09	BIRC8_HUMAN	Baculoviral IAP repeat-containing protein 8	Homo sapiens	9606	BIRC8	1	2
Q5GH70	XKR9_HUMAN	XK-related protein 9	Homo sapiens	9606	XKR9	1	1
Q9H175	CSRN2_HUMAN	Cysteine/serine-rich nuclear protein 2	Homo sapiens	9606	CSRNP2	1	1
Q8IXT1	DDIAS_HUMAN	DNA damage-induced apoptosis suppressor protein	Homo sapiens	9606	DDIAS	1	2
P0DMP1	HMN12_HUMAN	Humanin-like 12	Homo sapiens	9606	MTRNR2L12	3	2
S4R3P1	HMN13_HUMAN	Humanin-like 13	Homo sapiens	9606	MTRNR2L13	3	1
Q9BZS9	YG041_HUMAN	Putative uncharacterized protein PNAS-138	Homo sapiens	9606	PNAS-138	5	1

Figure 11

1) LLM-Based Query Transformation:

To evaluate the results, queries were written in three different categories: simple, complex, and curator writing.

Query Type	Gemini-1.5-flash	Meta-llama-3.1	codestral-latest	Deepseek-r1
Simple Queries	42/51 (82.35%)	39/51 (76.47%)	45/51 (88.24%)	47/51 (92.16%)
Complex Queries	40/50 (80.00%)	35/50 (70.00%)	43/50 (86.00%)	44/50 (88.00%)
Curator- Written Queries	14/19 (73.68%)	12/19 (63.16%)	15/19 (78.95%)	18/19 (94.74%)

Table 1: Table showing the performance of selected models in different query types

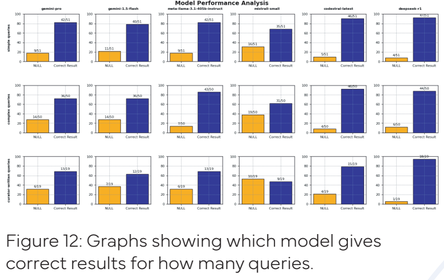


Figure 12: Graphs showing which model gives correct results for how many queries.

3) RAG (Retrieval-Augmented Generation)-Based Functional Explanation Generation

To generate more meaningful, context-aware protein explanations, the RAG (Retrieval-Augmented Generation) method was applied.

The process has two main steps: **retrieval** and **generation**.

- In the retrieval step, user queries were embedded using the nomic-embed-text-v1[7] model and matched with similar documents via ChromaDB.[8]

- Additionally, keyword-based search used two methods: trigram-based indexing with SQLite FTS5,[9] and sparse vector matching using Pinecone's BM25Encoder.[10]

In the generation step, retrieved data and user prompts were combined to produce context-relevant responses via a language model.

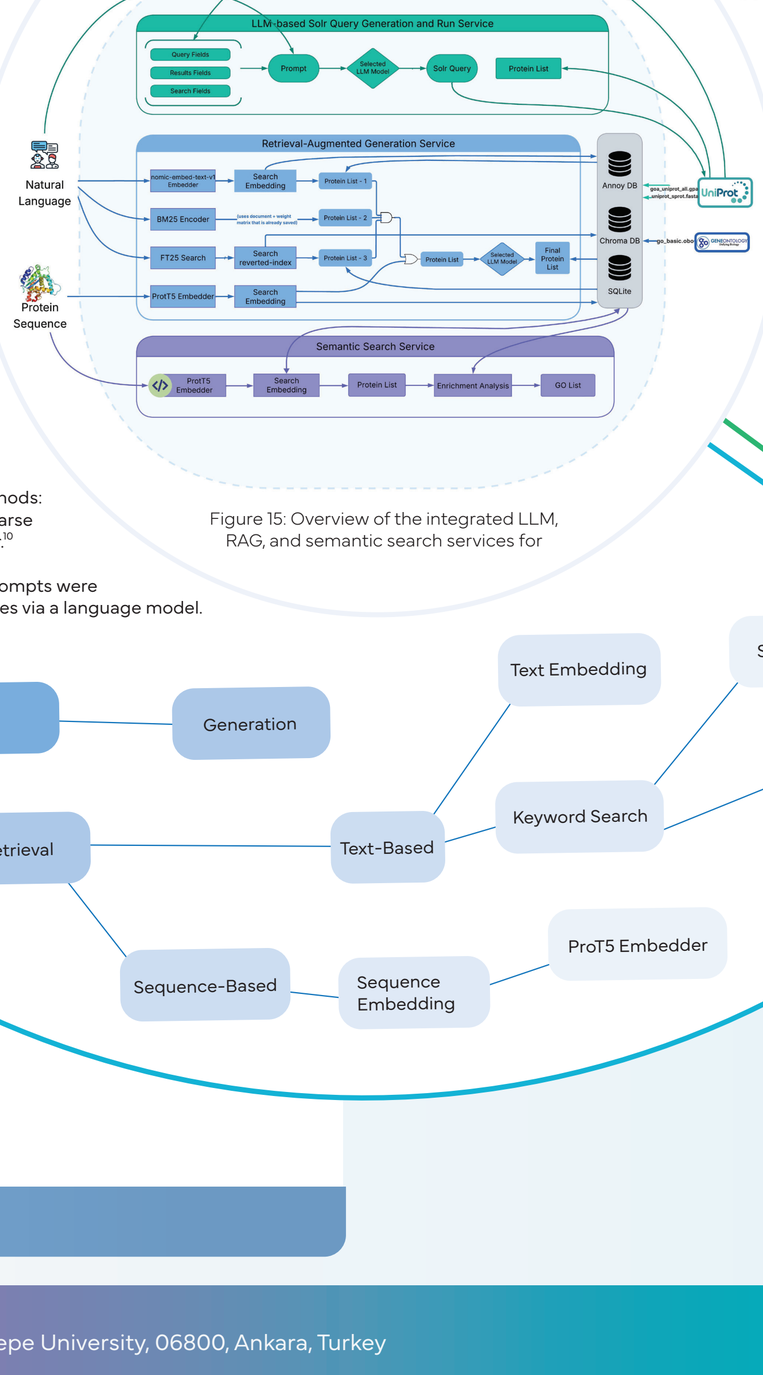


Figure 15: Overview of the integrated LLM, RAG, and semantic search services for

3) RAG (Retrieval-Augmented Generation) Component:

Data

The system is composed of three main components, each supported by specific data types and sources:

1)LLM-Based Query Transformation: User queries in natural language are adapted to UniProt's structure using documentation on query, search, and result fields as guidance.

2)Semantic Search: Precomputed ProtT5 embeddings[2] from the UniProt Reviewed (Swiss-Prot) dataset (per-protein.h5) are used. Protein descriptions are taken from the uniprot_sprot.fasta file. For GO term analysis, only curated terms (excluding those labeled with IEA) from the go_basic.obo[3] file are included.

3)Retrieval-Augmented Generation (RAG): To generate context-aware explanations, the plain text uniprot_sprot.dat file is used. Since sequence embeddings are directly utilized, the raw sequence data is not separately processed.

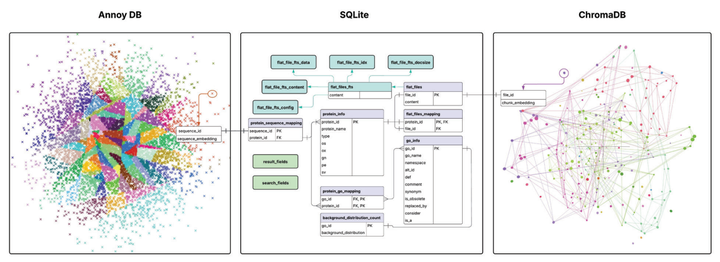


Figure 3: ER diagram of the databases used in the system

2) Semantic Search and Embedding-Based Matching

To compare a given protein sequence with others in the UniProt database in a biologically meaningful and fast way, a vector-based search system was developed. This system uses 1024-dimensional embeddings generated by UniProt's ProtT5 protein language model.

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\cdot\sqrt{\sum_{i=1}^{n} B_i^2}},$$

These vectors are indexed using the **Annoy library**[5] for fast approximate nearest neighbor searches. Angular distance is used for similarity, and cosine similarity for ranking results. Since Annoy cannot store metadata, an **SQLite database**[6] is used to link embedding IDs with protein information. This setup offers a portable, efficient, and file-based search infrastructure.

- **Similarity Queries:**
The user-provided protein sequence is embedded using ProtT5 and queried via Annoy to retrieve the most similar proteins in UniProt. The results are presented along with biological data such as **functional descriptions, GO terms, organism,** and **subcellular location.**

At this stage, two different metrics were used to evaluate the relationships among the identified GO terms:

Enrichment Score: Evaluates whether a GO term appears more frequently than expected in the user's gene list.

$$\text{Enrichment Score} = \frac{\frac{m}{n}}{\frac{M}{N}}$$

N: Total number of genes
M: Number of genes associated with the relevant GO term
n: User-provided gene list
m: Number of genes in the user list associated with the GO term

P-value: Measures the likelihood that the observed frequency of a GO term happened by chance. It's calculated using a hypergeometric distribution based on how many genes in a random list would be linked to the GO term.

$$P = \sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

2) Semantic Search:

As shown in Figure 12, tests conducted on 1,000 randomly selected sequences were compared with results obtained using the BLAST[4] (Basic Local Alignment Search Tool) on the GenBank database. In **96.8% of the cases**, the correct match was found in the first position, and in nearly all cases, it was among the top 7 results.

Figure 13: A chart showing the rank position of the correct protein match in the results for a sample query.

- The searches returned results within seconds and, as shown in Figure 14, performed significantly faster than traditional alignment-based methods.

- The system output was enriched with **functional descriptions** of similar proteins, **GO terms, organism information**, and **location data.**

Two key metrics were used to evaluate system performance:

1) **Spearman Correlation Coefficient:** Measures the similarity in ranking between vector-based search results and BLAST results. A value closer to 1 indicates stronger consistency between the two methods.

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

2) **Common Protein Metric:** Measures the proportion of shared proteins identified by both methods. Ranges between 0 and 1; higher values indicate greater similarity.

Common Protein Metric =

$$\frac{2 \times \text{Number of Common Proteins}}{\text{Proteins from Vector Search} + \text{Proteins from BLAST Search}}$$

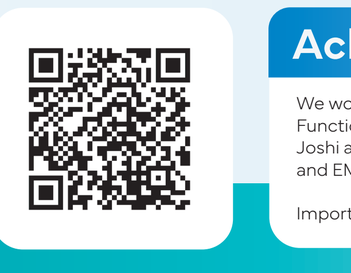


Figure 14: A graph showing the average speed performance of the system based on tests with 10,000 sequences, compared to currently used alternatives.

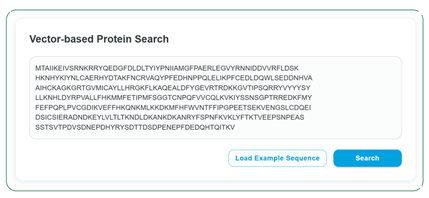
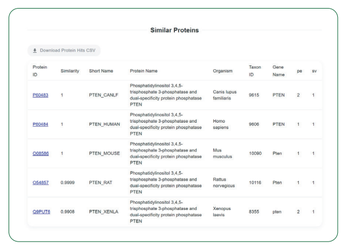
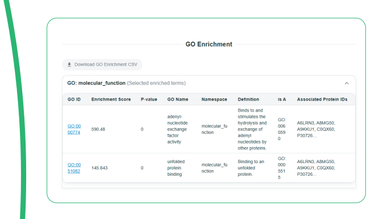


Figure 7



Protein ID	Similarity	Short Name	Protein Name	Organism	Taxon ID	Gene Name	pe	sv
P60483	1	PTEN_CANLF	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	Canis lupus familiaris	9615	PTEN	2	1
P60484	1	PTEN_HUMAN	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	Homo sapiens	9606	PTEN	1	1
Q08860	1	PTEN_MOUSE	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	Mus musculus	10090	Pten	1	1
Q64857	0.9999	PTEN_RAT	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	Rattus norvegicus	10116	Pten	1	1
Q9PUT6	0.9908	PTEN_XENLA	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	Xenopus laevis	8355	pten	2	1

Figure 8



GO ID	Enrichment Score	P-value	GO Name	Namespace	Definition	Is A	Associated Protein IDs
GO:0019104	590.48	0	adenyl-nucleotide exchange factor activity	molecular_function	Binds to and stimulates the hydrolysis and exchange of adenyl nucleotides by other proteins.	GO:0060589	A8LEN2, A8MQ50, A9MGU1, C0QX60, P35720, ...
GO:0005515	148.843	0	unfolded protein binding	molecular_function	Binding to an unfolded protein.	GO:0044183	A8LEN2, A8MQ50, A9MGU1, C0QX60, P35720, ...

Figure 9

Figure 7 shows a system where the user can perform a search using only the available sequence information.

Figure 8 presents a list of proteins that are most similar to the given sequence and are therefore likely to be related.

Figure 9 displays the Gene Ontology (GO) terms associated with the relevant protein, grouped by their respective namespaces for easier interpretation by the user.

General Evaluation

This study presents a comprehensive system that can convert natural language queries in the UniProt database into Solr-based structural searches and perform semantic similarity analysis on protein sequences. This structure, which is faster and more user-oriented compared to traditional tools like BLAST, facilitates access to protein function information for both technical and non-technical users.

The system was developed in collaboration with the **EMBL-EBI UniProt** team, in line with real user needs and query habits.

Acknowledgements

We would like to thank Maria-Jesus Martin, the Protein Function Development team leader, team members Vishal Joshi and Aurelien Luciani, and the entire UniProt DB team and EMBL-EBI.

Important Links: https://proquest.ngrok.app