UDC 004.651: 004.652 *Reviews*

UNIPROT DATABASE — UNIVERSAL INFORMATION RESOURCE OF PROTEIN SEQUENCES

Kulyyassov A. T.*

National Center for Biotechnology,13/5, Korgalzhyn road, Nur-Sultan, 010000, Kazakhstan
**kulyyasov@biocenter.kz*

ABSTRACT

Protein sequences are stored in public databases such as the UniProt Knowledgebase (UniProtKB), where curators add bioinformatics data, including prediction of structure and function of biomolecules and experimental results. Protein function prediction can be done using sequence similarity searches, but an alternative approach is to use protein signatures that classify proteins into families and domains. The main protein signature databases are accessible through the integrated InterPro database, which provides the UniProtKB sequence classification. In addition to characterizing proteins through protein families, many researchers are interested in analyzing the complete set of proteins from the genome (i.e., the proteome), and there are databases and resources providing unreduced sets of proteomes and analyzes of proteins from organisms with fully sequenced genomes. This article reviews the tools and resources available on the Internet for characterizing both individual proteins and analysis of the entire proteome.

Keywords: Association-Rule-Based Annotator (ARBA), European Bioinformatics Institute (EBI), The European Molecular Biology Laboratory (EMBL), The DNA Data Bank of Japan (DDBJ), Gene Ontology Annotation (GOA), Global Proteome Machine (GPM), Mass spectrometry (MS), proteomics, Liquid Chromatography tandem Mass Spectrometry (LC–MS/MS), Multiple reaction monitoring (MRM), National Institutes of Health (NIH), Protein Data Bank (PDB), PRoteomics IDEntifications (PRIDE), Protein Information Resource (PIR), Post-translational modification (PTM), Swiss Institute of Bioinformatics (SIB), the Universal Protein Resource (UniProt), the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProt), the UniProt Reference (UniRef).

1. Introduction

The rapid development of proteomics, driven by the widespread use of mass spectrometers in research institutes, medical clinics, commercial companies and other organizations around the world, has led to a significant increase in the amount of data generated. Despite this, the need to reliably identify analyzed proteins and quickly obtain information about their properties and functions remains unchanged. This requires a database of protein sequences that does not contain redundant data (or with a minimum level of redundancy), with maximum coverage, including splice isoforms, disease variants and post-translational modifications. Sequence archiving is an important feature in order to be able to interpret and maintain the results of a proteomic set. Stable identifiers, consistent nomenclature, and convenient dictionaries are very useful for identifying proteins. An important requirement is also the provision of detailed information on protein function, biological processes, molecular interactions and pathways, with cross-references to relevant external sources. This article shows how the UniProt database meets these criteria.

2. Databases of protein sequences

A number of new technologies in protein science make it possible to quickly identify a large number of proteins in a complex, map their interactions in a cellular context, determine their location in a cell, and analyze their biological activity. Protein sequence databases play a vital role as a central resource for storing the data resulting from these efforts and making them freely available to the scientific community. Data from large-scale experiments are often no longer published in the usual sense, but are deposited in a database. This means that protein sequence databases are the most comprehensive resource of protein information available to scientists.

In order to take full advantage of the various resources, it is necessary to distinguish between them and determine the types of data they contain. Universal protein databases cover proteins of all kinds, while specialized data collections contain information about a particular family or group of proteins, or about proteins related to a particular organism. Universal protein sequence databases can be divided into three categories:

- simple sequence data archives, or sequence repositories, where data is stored with little or no manual intervention in record creation;
- specialized databases that contain information in a specific area of research or description of certain parameters of biomolecules, such as 3D structures;
- databases created by experts, in which the original data is supplemented by additional information obtained from sources such as published scientific literature.

One of the protein sequence database families, the Universal Protein Resource (UniProt), will be discussed in detail.

2.1. Sequence data archives

A number of protein sequence databases act as

repositories of these sequences. These databases add little or no additional information to the sequence records they contain, and generally make no effort to provide users with an unreduced collection of sequences. An example is the GenBank Genetic Products Databank, or GenPept, created by the National Center for Biotechnology Information [1]. Database entries are derived from translations of sequences contained in the Nucleotide Database jointly maintained by DDBJ [2], EMBL Nucleotide Sequence Database [3], and GenBank [4] and contain minimal annotation that was extracted primarily from the corresponding nucleotide entry. The records lack any additional annotation and the database does not contain proteins derived from amino acid sequencing. It represents a redundant view of the world of proteins, which means that each protein can be represented by multiple records, and no attempt is made to group these records into a single database record. The NCBI Entrez Protein database [1] is another example of a sequence repository. The database contains sequence data translated from DDBJ/EMBL/GenBank nucleotide sequences, as well as sequences from UniProt, RefSeq, and Protein Data Bank (PDB). The database differs from GenPept in that many of the records contain additional information, but most of the annotated data has been extracted from curated databases, so little new information has been added to the records that cannot be found in other data collections. As with GenPept, the collection of sequences is redundant. A more ambitious approach is taken in the collection of reference sequences (RefSeq) created by NCBI [5]. The aim of the project is to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcripts (RNA) and protein products, for the main organisms under study. NCBI staff provide ongoing follow-up, with review status indicated on each entry. However, most records are generated automatically with minimal manual intervention, so the database is closer to a sequence repository than any of the curated databases discussed below.

The search for protein sequences and peptides can also be performed using proteomic databases and repositories of spectral libraries [6–9]. For example, to select candidate peptides in experiments, large databases designed for storing and exchanging experimental proteomic data, such as the Global Proteome Machine (GPM) and the Proteomics IDEntifications (PRIDE) databases, can be used [10–13].

2.2. Specialized databases

In addition to protein sequence archives, there are many specialized databases available to the life science community. Some of them focus on one specific aspect of proteins or protein groups or families, or on a particular organism, while others seek to combine and use existing resources to the fullest. They vary in size and the amount of data they contain. One example of the first type of database is the Protein Data Bank (PDB), which archives three-dimensional structural data [14]. Another example is the Gene Ontology (GO) knowledge base, the world's largest source of information about gene functions [15]. This knowledge is also machine-readable and can be used for computational analysis of large-scale molecular biological and genetic experiments in biomedical research. GOA, a Gene Ontology Annotation project [16], annotates proteins to GO terms. Data on protein interactions can be found in the information resources IntAct [17], BioGRID [18], and STRING [19].

Other examples are information resources containing a large amount of data obtained in the course of experiments on targeted proteomics [20]. These databases can help select suitable candidate peptides for the development of peptide quantitation methods using the Multiple reaction monitoring (MRM) method. The largest databases are PeptideAtlas [21] and SRMAtlas [22], which contain information on peptides obtained from proteomic experiments. Other useful resources are the PeptideTracker [23], which contains quantitative information on proteins from various biological matrices, and the Panorama Public database [24], which provides complete information for the development of MRM assays.

2.3. Universal curated databases

While repositories are an important means of getting sequences to the user as quickly as possible, it is clear that when additional information is added to the sequence, this greatly increases the value of the resource for users. Curated databases take basic sequence information and enrich it by adding additional information from various sources such as the scientific literature. This information is extracted and verified by curators and biological experts before being added to databases, which means that the data in these collections can be considered highly reliable. In addition, significant efforts are being made to maintain unreduced datasets by consolidating all reports for a given protein sequence into a single record.

3. Goal and structure of UniProt

The main goal of the UniProt database is to give access to comprehensive, high quality, and freely available information of protein sequence and functional that is essential for modern biological research. The UniProt database was created by the UniProt Consortium, which includes groups from the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR), and the Swiss Institute for Bioinformatics (SIB). The National Institutes of Health (NIH) is the main supporter of UniProt. The European Commission and the Swiss Federal Government are also providing an additional funding for its activities.

UniProt databases consist of several components (Figure 1, A):

• The UniProt archive (UniParc) provides a stable, non-reduced collection of sequences that contains the entirety of publicly available protein sequence data;

• The UniProt Knowledge Base (UniProtKB) is a central database of protein sequences with numerous

and accurate sequencing results and functional annotation;

• UniProt NREF (UniRef) databases provide non-reduced collections of data based on the UniProt knowledge base to fully cover the sequence space at multiple resolutions. UniProt Reference Clusters (UniRef) provides clustered sets of sequences from the UniProt knowledge base (including isoforms) and selected UniParc records. This allows you to hide redundant sequences and get full coverage of the sequence space with three resolutions:

✓ UniRef100 combines identical sequences and subfragments with 11 or more residues from any organism into one UniRef record;

✓ UniRef90 is built by clustering UniRef100 sequences in such a way that each cluster consists of sequences that have at least 90% sequence identity and 80% overlap with the longest sequence (aka start sequence);

✓ UniRef50 is created by clustering original UniRef90 sequences that have at least 50% sequence identity and 80% match with the longest sequence in the cluster;

• The UniProt Proteomes database provides proteomes for species with fully sequenced genomes. The proteome is the complete set of proteins that are expressed by an organism;

• Annotation systems are systems used to automatically annotate proteins with high accuracy:

- UniRule (Expert curated rules)
- ARBA (system generated rules)

The UniProt database is gaining more and more popularity, as there is a growth trend in the titles and abstracts of publications, such as in the PubMed information resource (Figure 1, B). As of January 2022, the UniProt database contains over 225 million sequence records. The data in UniProt are distributed according to the taxonomic classification with the maximum number of sequences for bacteria (60%) and eukaryotes (34%). The remaining small fractions of 3% each corresponds to archaea and viruses (Figure 1, C). The distribution of the number of sequences within eukaryotes according to the UniProt database is shown in Figure 1, D. The largest share is made up of mammals (34%), among the occurring human sequences they make up 10%. Further, 21% of data on plants, 18% on fungi, 10% on other vertebrates, 5% on insects, 3% on nematodes, and 9% on other species.

The UniProtKB Knowledge Base consists of two sections: section Swiss-Prot that are fully annotated manually, where data are obtained from literature information extraction, computational analysis, and curator evaluation, and a section TrEMBL with computationally analyzed records awaiting full manual annotation (Figure 1, A).

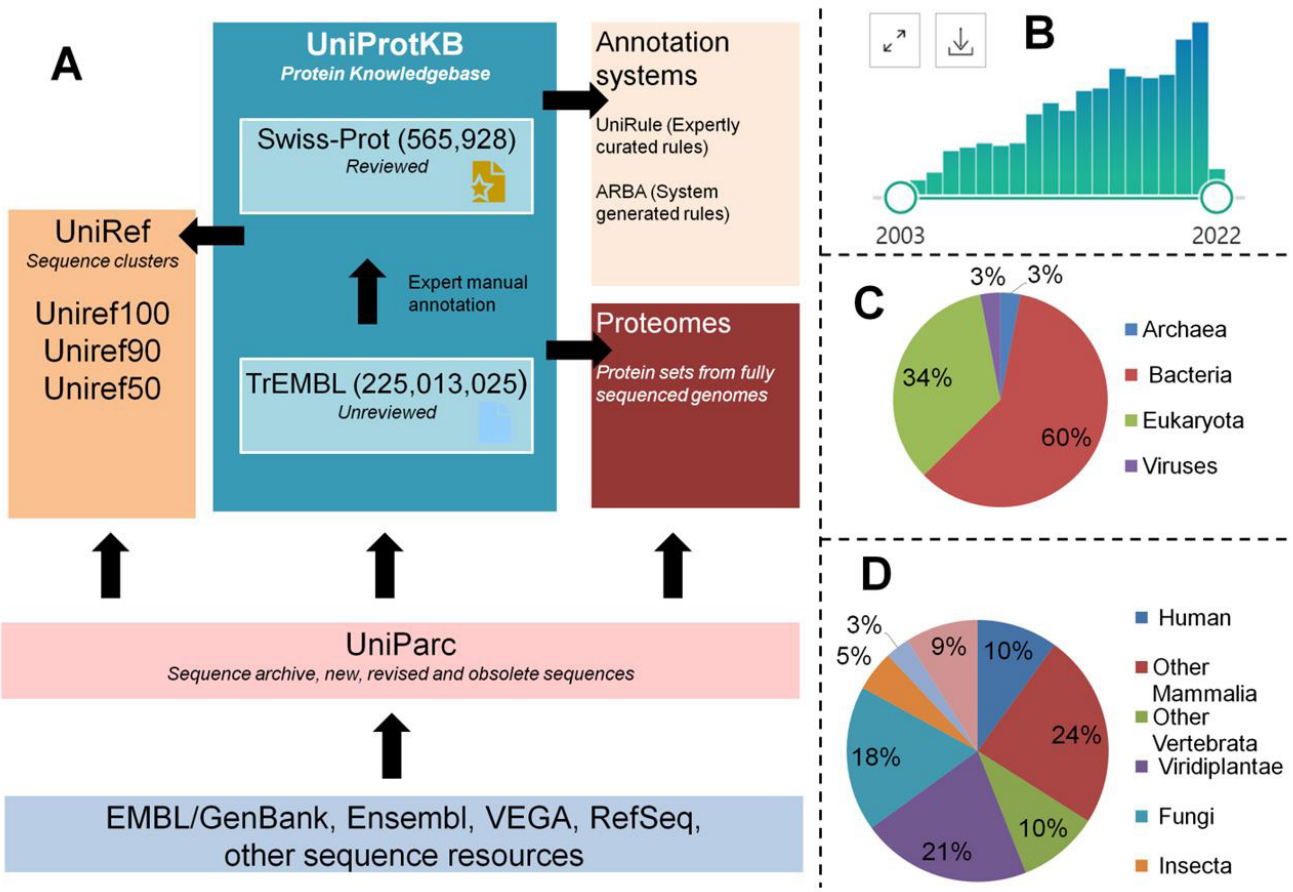

Figure 1. — **Structure and content of the UniProt protein sequence database.** A: Components of the UniProt structure. B: Dynamics of the number of publications in PubMed for 2003–2022 by the keyword UniProt in the title and abstract of the article (197 publications in 2021). C: taxonomic distribution of sequences in Swiss-Prot. D: Taxonomic distribution of sequences among eukaryotes in Swiss-Prot. The numerical values given in the block diagram correspond to the data for January 2022.

4. High quality annotation

In addition to collecting the basic data required for each UniProt entry (consisting primarily of amino acid sequence, protein name or description, taxonomic data, and citation information), as much annotation information as possible is attached to the protein. This is performed both automatically and manually.

4.1. Manual annotation by curators based on literature data and sequence analysis

Sequences for which new functional, structural and/or biochemical data have been published receive a high manual annotation priority. In UniProt, an annotation consists of a description of the following elements:

- ✓ the function(s) of the protein;
- ✓ information on enzyme-related processes (cofactors, metabolic pathway, catalytic activity, regulatory mechanisms);
- ✓ biologically significant domains and sites;
- ✓ post-translational modifications (PTM);
- ✓ molecular weight determined by mass spectrometry;
- ✓ subcellular(s) location(s) of the protein;
- ✓ tissue-specific protein expression;
- ✓ developmentally specific protein expression;
- ✓ secondary structure;
- ✓ Quaternary structure;
- ✓ interactions;
- ✓ splice isoform(s);
- ✓ mature protein products;
- ✓ polymorphism(s);
- ✓ similarity with other proteins;
- ✓ the use of protein in the biotechnological process;
- ✓ diseases associated with protein deficiencies or abnormalities;
- ✓ the use of protein as a pharmaceutical product;
- ✓ sequence conflicts, etc.

This annotation is contained in comment lines (CC), feature tables (FT) and keyword tables (KW). Comments are classified by topic, making it easy to extract specific categories of data from the database. In order to obtain the most up-to-date and extensive knowledge about the protein, information is taken not only from publications reporting new sequence data, but also from review articles. In addition, outside experts are also brought in to provide comments and updates on specific groups of proteins.

4.2. Automatic classification and annotation

With the rapid growth of sequence databases, there is a growing need for reliable functional characterization and annotation of newly predicted proteins. To cope with such large amounts of data, faster and more efficient means of protein sequence characterization and annotation are needed. One promising approach is automatic large-scale functional characterization and annotation, which is created with limited human involvement.

For data annotation, the InterPro tool [25] is used to recognize domains and classify all protein sequences in UniProt into families and superfamilies. InterPro is an integrated resource of protein families, domains, and sites that combines the efforts of its databases: Pfam [26], PROSITE [27], SMART [28], and others [29]. In UniProtKB/TrEMBL records, the domains predicted by the previously listed databases are used for automatic domain annotation.

The rule-based semi-automated UniRule [30] computational annotation system annotates experimentally uncharacterized proteins based on similarity to known experimentally characterized proteins, adding properties such as protein name, functional annotation, catalytic activity, pathway, GO terms, and subcellular location. The number of UniRules used for annotation has grown to 6768 rules in total (release 2020 04).

To complement the UniRules creation process led by UniRules experts, the Association-Rule-Based Annotator (ARBA), a self-learning annotation system for automatic classification and annotation of proteins UniProtKB [31], was recently introduced (release 2020 04). It replaced the previous rule-based SAAS system. ARBA is trained on UniProtKB/Swiss-Prot, then uses rule search methods to generate concise annotation models with maximum representativeness and coverage based on InterPro group membership and taxonomy properties. ARBA uses a data exclusion set that screens out data not suitable for computational annotation (such as specific biophysical or chemical properties) and generates human-readable rules for each release, which are available at https://www.uniprot.org/arba/. 22,894 ARBA rules were used to annotate 87,325,890 proteins in release 2020 04, increasing the total coverage of rule-based annotation systems from 35% to 49% in UniProtKB/TrEMBL. In addition, in release 2020 04, over 15 million uncharacterized protein names have been enhanced with InterPro member database signatures, updating their name to "domain X protein" in accordance with the International Guidelines for Protein Nomenclature (https://www.uniprot.org/docs/International_Protein_Nomenclature_Guidelines.pdf).

5. Data integration

UniProtKB combines datasets by matching these data to corresponding protein sequence records, displaying the mappings using the ProtVista visualization tool [32], and uploading them via FTP and API [33]. Clinically relevant sources of variation (e.g., 100K genomes, gnomAD and ClinVar SNPs) are mapped to protein features and variants using a pre-calculated mapping of genomic coordinates for amino acids at the beginning and end of each exon and converting UniProt position annotations to their genomic coordinates [34]. Functional positional annotations from the UniProt human reference proteome are currently mapped to

the corresponding genomic coordinates in the GRCh38 version of the human genome for each release of UniProt.

UniProt further aggregates and visualizes unique and non-unique peptides identified from proteomic mass spectrometry data deposited through the ProteomeXchange [10] consortium (e.g. PeptideAtlas [35], MassIVE [36] and jPOST [37] and other large-scale initiatives (CPTAC [38], ProteomicsDB [39], MaxQB [40], ETD and CTDP [41].

Aligning variants with protein characteristics such as functional domains and active sites, ligand binding sites and PTMs in the UniProt entry can provide mechanistic insight into how particular variants can lead to disease or drug or pathogen resistance.

6. Examples of using UniProt

6.1. Annotated transcription factor Nanog data

Depending on the goals and objectives of the study, specialists from various fields (biologists, doctors or biochemists) can select the information they need, as well as sort the search by various criteria, such as keywords, taxonomy, diseases or subcellular localization. As an example, one can demonstrate the search for information on pluripotency transcription factors. Figure 2 shows the results of the Nanog protein (Q9H9S0) data analysis in the UniProt system. Each cell of the search result contains links to articles and publications, such as "Protein function and role" (Function, Figure 2 A), or links to other databases, such as in the "Interaction" cell (Interaction, Figure 2 B). Amino acid sequences can also be loaded as FASTA files, or multiple sequences can be added to the basket at once (Add to basket, Figure 2C) and used for comparison and alignment with other sequences. The spatial structure of the protein in 3D format (Figure 2D) also contains links to the PDB databases [42] containing the results of X-ray diffraction and NMR experiments, as well as the Alphafold database [43], where the spatial structure of the protein was established using artificial intelligence algorithms. For cell biologists, subcellular localization of the studied protein is also significant information (Figure 2, E) with indication and references to the relevant literature. The Uniprot functional table also contains other data, for example: information on diseases and phenotypes (Pathology/Biotech), post-translational modifications of proteins (PTM/processing), information on gene expression at the mRNA or protein level in cells or tissues of multicellular organisms (Expression), information about the similarity of sequences with other proteins, and about the domain(s) present in the protein (Family/Domains).

6.2. Alignment of Sox2 and Oct4 DNA-binding domains in various species

The pluripotency transcription factors contain structure domains, assisting them in binding to

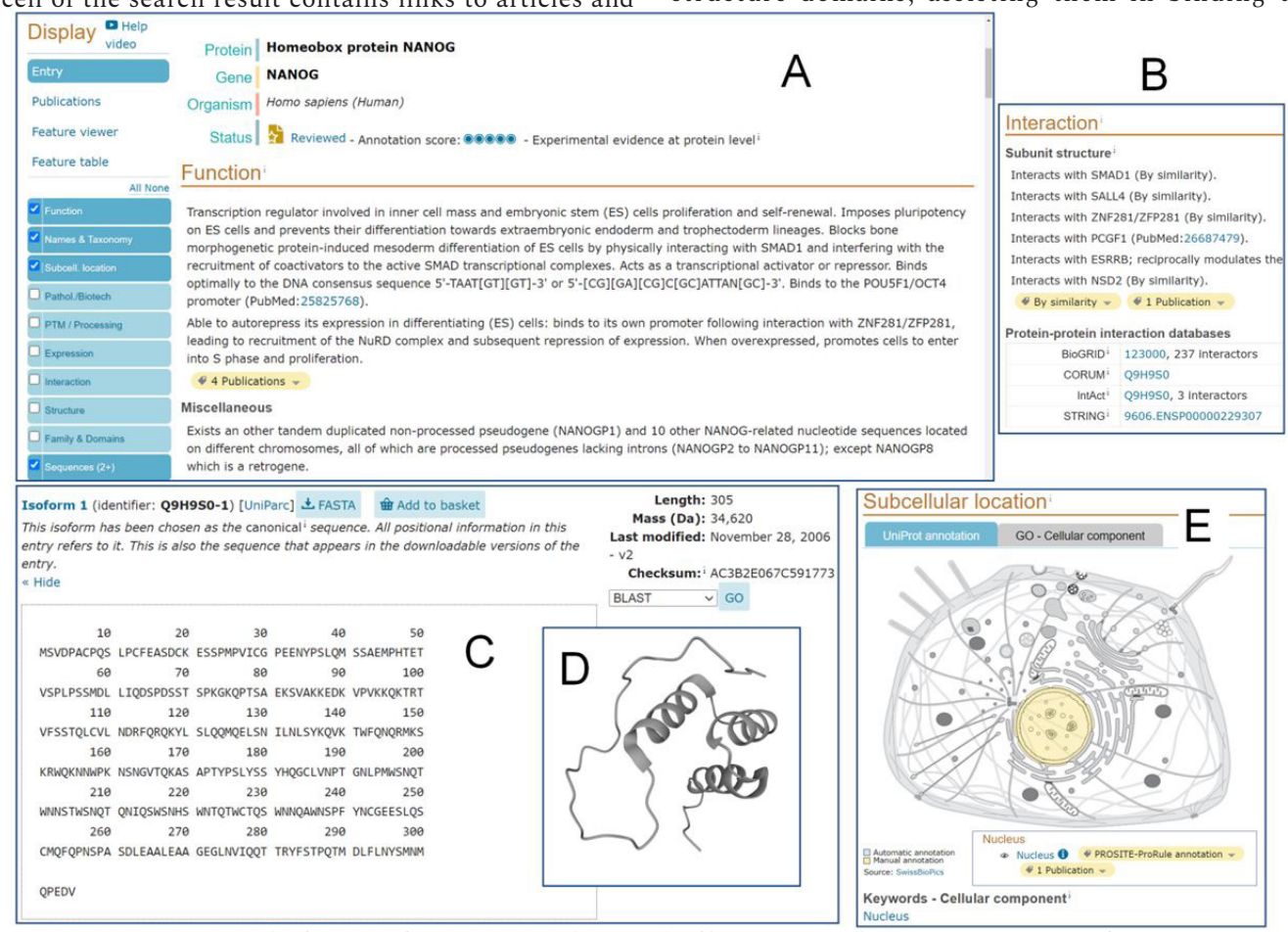

Figure 2. — **Structured information from UniProt on the example of human Nanog pluripotency transcription factor (UniProtKB Q9H9S0).** A: Functional annotation. B: Protein-protein interactions. C: Amino acid sequence, length and molecular weight. D: Spatial structure of the molecule. E: Subcellular localization.

specific regions of DNA. Proteins Sox2 (SRY-box 2), Oct4 (Octamer-binding transcription factor 4), and NANOG form the core of the transcriptional network that controls cell pluripotency [44] and are key in the induction of pluripotency in somatic cells [45–50]. For example, Sox2 binds to the C(T/A)TTGTC DNA sequence, while Oct4 recognizes the ATGC(A/T)AAT consensus sequence. Basically, they bind together on a composite motif formed by the superimposition of individual binding sites [51], known as the canonical motif. Thus, the direct interaction between these key transcription factors is DNA-dependent [52, 53], which involves DNA-binding domains such as POUS (POU-specific domain) and POUHD (homeodomain) of the OCT4 protein and HMG (high-mobility group domain) in Sox2 [54].

These DNA-binding domains are evolutionarily conserved, as seen when comparing their amino acid sequences in different animal species (Figure 3). To start, the sequences of DNA-binding domains Sox2 and Oct4 from various species of organisms found in the Uniprot database were first placed in the basket (Add to basket), and then aligned (Align). Sequence logo — a method of graphical representation of the conservation of nucleotides (in a strand of RNA or DNA) or amino acids (in proteins) was obtained using the Sequence logo program available at https://weblogo.berkeley.edu/logo.cgi.

Conclusion

Modern research is moving away from single-gene to whole-genome research, and approaches based on bottom-up hypotheses are being replaced by top-down exploratory studies. Proteins are studied in their context in the cell to determine which molecules they interact with and in which biological systems they play a role. This "systems biology" approach is much more relevant to real life and is likely to provide a deeper understanding of the molecular biology of organisms than a focus on a single gene. To achieve this goal, researchers need access to large and complete datasets with as much functional information as possible. It is hoped that a combination of tools such as those described here will help biologists shed light on the biological functions of newly discovered proteins and systems. Only then can the data be used in full for medical or commercial purposes.

Acknowledgments

This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan

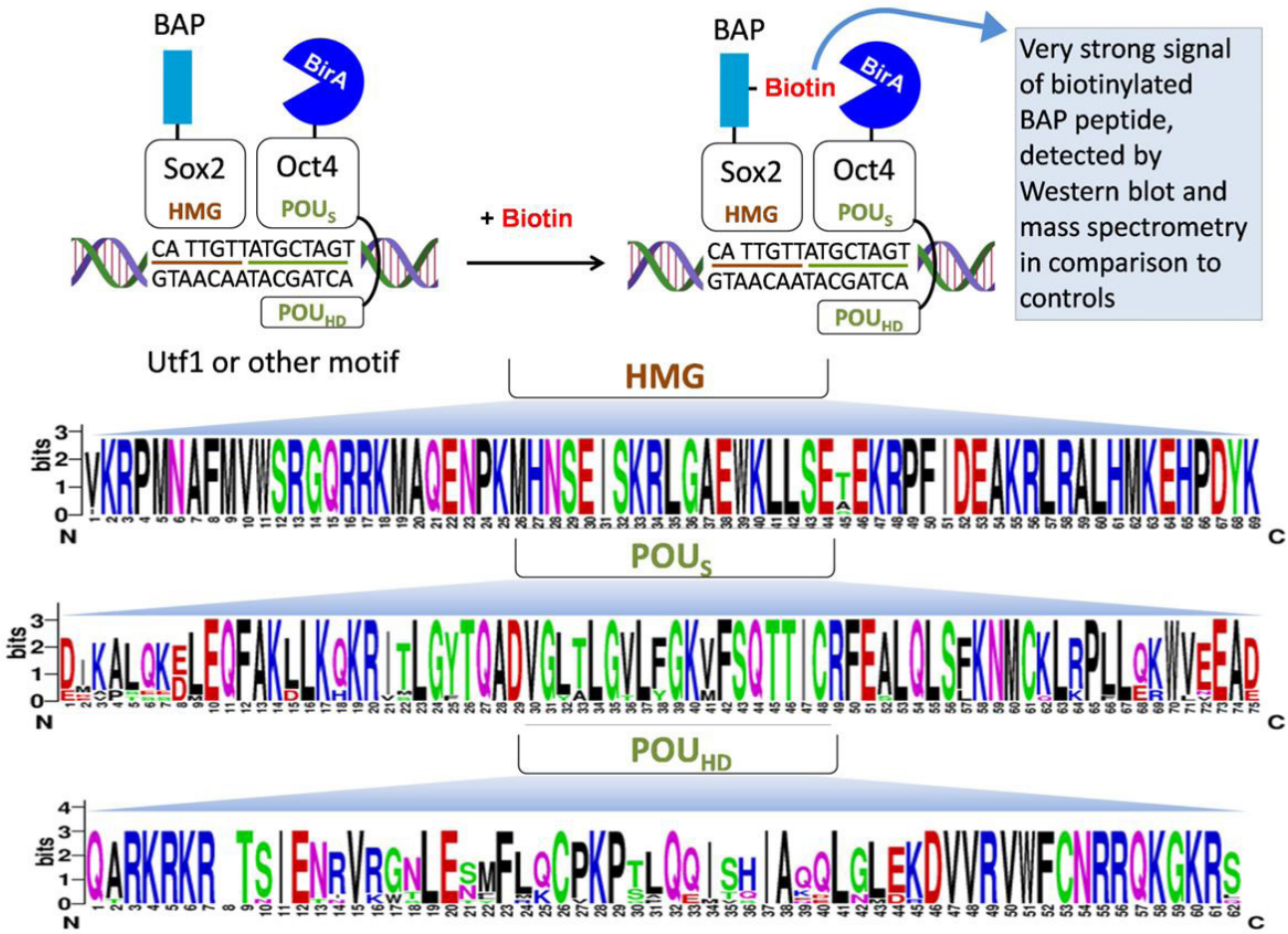

Figure 3 — Detection of protein-protein interaction using biotin ligase and biotin acceptor peptide. The HMG box domain of Sox2 and POUS with POUHD domains of Oct4 allow tight binding on composite DNA motifs such as Utf1 or others. The direct contact of the biotin acceptor peptide (BAP) and wild-type BirA results in site-specific biotinylation of the target molecule [55]. The level of BAP biotinylation can be detected using Western blotting or quantified using MRM mode of LC–MS/MS [56] followed by raw data processing with Skyline software [57]. Sequence Logo of HMG DNA-binding domain sequences of some mammalian species (human, cow, sheep, goat, pig, mouse), chicken, frog and fish. Sequence Logo of POUS and POUHD domain sequences of some mammalian species (human, cow, sheep, goat, cat, dog, rat, mouse), chicken and frog.

(AP09259838 "Application of new proteomics methods in studying the mechanism of action of pluripotency transcription factors expressed in mammalian cell lines" for 2021–2023).

REFERENCES

1 Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., … Yaschenko, E. Database resources of the National Center for Biotechnology Information // Nucleic Acids Res. – 2007. – Vol. 35, № Database issue. – P. D5-12.

2 Okubo, K., Sugawara, H., Gojobori, T., & Tateno, Y. DDBJ in preparation for overview of research activities behind data submissions // Nucleic Acids Res. – 2006. – Vol. 34, № Database issue. – P. D6-9.

3 Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., … Apweiler, R. EMBL Nucleotide Sequence Database in 2006 // Nucleic Acids Res. – 2007. – Vol. 35, № Database issue. – P. D16-20.

4 Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L. GenBank // Nucleic Acids Res. – 2007. – Vol. 35, № Database issue. – P. D21-5.

5 Pruitt, K. D., Tatusova, T., Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins // Nucleic Acids Res. – 2007. – Vol. 35, № Database issue. – P. D61-5.

6 Bhowmick, P., Roome, S., Borchers, C. H., Goodlett, D. R., Mohammed, Y. An Update on MRMAssayDB: A Comprehensive Resource for Targeted Proteomics Assays in the Community // J Proteome Res. – 2021. – Vol. 20, № 4. – P. 2105-2115.

7 Martens, L. Public proteomics data: How the field has evolved from sceptical inquiry to the promise of in silico proteomics // EuPA Open Proteom. – 2016. – Vol. 11. – P. 42-44.

8 Perez-Riverol, Y., Alpi E., Wang, R., Hermjakob H., Vizcaino, J. A. Making proteomics data accessible and reusable: current state of proteomics databases and repositories // Proteomics. – 2015. – Vol. 15, № 5-6. – P. 930-49.

9 Uszkoreit J., Winkelhardt D., Barkovits K., Wulf M., Roocke S., Marcus K., Eisenacher M. MaCPepDB: A Database to Quickly Access All Tryptic Peptides of the UniProtKB // J Proteome Res. – 2021. – Vol. 20, № 4. – P. 2145-2150.

10 Deutsch, E. W., Bandeira, N., Sharma V., Perez-Riverol, Y., Carver, J. J., Kundu, D. J., Garcia-Seisdedos, D., Jarnuczak, A. F., Hewapathirana, S., Pullman, B. S., Wertz, J., Sun, Z., Kawano, S., Okuda, S., Watanabe, Y., Hermjakob, H., MacLean, B., MacCoss, M. J., Zhu, Y., Ishihama, Y., Vizcaino, J. A. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics // Nucleic Acids Res. – 2020. – Vol. 48, № D1. – P. D1145-D1152.

11 Perez-Riverol, Y., Csordas, A., Bai J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Perez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A. F., Ternent, T., Brazma, A., Vizcaino, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data // Nucleic Acids Res. – 2019. – Vol. 47, № D1. – P. D442-D450.

12 Fenyo, D., Beavis, R. C. The GPMDB REST interface // Bioinformatics. – 2015. – Vol. 31, № 12. – P. 2056-8.

13 Jones, P., Cote, R. G., Martens, L., Quinn, A. F., Taylor, C. F., Derache, W., Hermjakob, H., Apweiler, R. PRIDE: a public repository of protein and peptide identifications for the proteomics community // Nucleic Acids Res. – 2006. – Vol. 34, № Database issue. – P. D659-63.

14 Burley, S. K., Bhikadiya, C., Bi C., Bittrich, S., Chen, L., Crichlow, G. V., Duarte, J. M., Dutta, S., Fayazi, M., Feng, Z., Flatt, J. W., Ganesan, S. J., Goodsell, D. S., Ghosh, S., Kramer Green, R., Guranovic, V., Henry, J., Hudson, B. P., Lawson, C. L., Liang, Y., Lowe, R., Peisach, E., Persikova, I., Piehl, D. W., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Vallat, B., Voigt, M., Westbrook, J. D., Whetstone, S., Young, J. Y., Zardecki, C. RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D // Protein Sci. – 2022. – Vol. 31, № 1. – P. 187-208.

15 Gene Ontology C. The Gene Ontology (GO) project in 2006 // Nucleic Acids Res. – 2006. – Vol. 34, № Database issue. – P. D322-6.

16 Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology // Nucleic Acids Res. – 2004. – Vol. 32, № Database issue. – P. D262-6.

17 Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., Hermjakob, H. IntAct--open source resource for molecular interaction data // Nucleic Acids Res. – 2007. – Vol. 35, № Database issue. – P. D561-5.

18 Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., Tyers, M. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions // Protein Sci. – 2021. – Vol. 30, № 1. – P. 187-200.

19 Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets // Nucleic Acids Res. – 2021. – Vol. 49, № D1. – P. D605-D612.

20 Kulyyassov, A., Fresnais, M., Longuespee, R. Targeted liquid chromatography-tandem mass spectrometry analysis of proteins: Basic principles, applications, and perspectives // Proteomics. – 2021. – Vol. 21, № 23-24. – P. e2100153.

21 Deutsch, E. W., Sun, Z., Campbell, D., Kusebauch, U., Chu, C. S., Mendoza, L., Shteynberg, D., Omenn, G. S., Moritz, R. L. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet // Journal of Proteome Research. – 2015. – Vol. 14, № 9. – P. 3461-3473.

22 Kusebauch, U., Campbell, D. S., Deutsch, E. W., Chu, C. S., Spicer, D. A., Brusniak, M. Y., Slagel, J., Sun, Z., Stevens, J., Grimes, B., Shteynberg, D., Hoopmann, M. R., Blattmann, P., Ratushny, A. V., Rinner, O., Picotti, P., Carapito, C., Huang, C. Y., Kapousouz, M., Lam, H., Tran, T., Demir, E., Aitchison, J. D., Sander, C., Hood, L., Aebersold, R., Moritz, R. L. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome // Cell. – 2016. – Vol. 166, № 3. – P. 766-778.

23 Mohammed, Y., Bhowmick, P., Smith, D. S., Domanski, D., Jackson, A. M., Michaud, S. A., Malchow, S., Percy, A. J., Chambers, A. G., Palmer, A., Zhang, S., Sickmann, A., Borchers, C. H. PeptideTracker: A knowledge base for collecting and storing information on protein concentrations in biological tissues // Proteomics. – 2017. – Vol. 17, № 7. – P. 1-6.

24 Sharma, V., Eckels, J., Schilling, B., Ludwig, C., Jaffe, J. D., MacCoss, M. J., MacLean, B. Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline // Mol Cell Proteomics. – 2018. – Vol. 17, № 6. – P. 1239-1244.

25 Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H. Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S. Y., Finn, R. D. InterPro in 2019: improving coverage, classification and access to protein sequence annotations // Nucleic Acids Res. – 2019. – Vol. 47, № D1. – P. D351-D360.

26 Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., Bateman, A. Pfam: The protein families database in 2021 // Nucleic Acids Res. – 2021. – Vol. 49, № D1. – P. D412-D419.

27 Sigrist, C. J., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I. New and continuing developments at PROSITE // Nucleic Acids Res. – 2013. – Vol. 41, № Database issue. – P. D344-7.

28 Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J., Bork, P. SMART 5: domains in the context of genomes and networks // Nucleic Acids Res. – 2006. – Vol. 34, № Database issue. – P. D257-60.

29 Lautenbacher, L., Samaras, P., Muller, J., Grafberger, A., Shraideh, M., Rank, J., Fuchs, S. T., Schmidt, T. K., The M., Dallago, C., Wittges, H., Rost, B., Krcmar, H., Kuster, B., Wilhelm M. ProteomicsDB: toward a FAIR open-source resource for life-science research // Nucleic Acids Res. – 2022. – Vol. 50, № D1. – P. D1541-D1552.

30 MacDougall, A., Volynkin, V., Saidi, R., Poggioli, D., Zellner, H., Hatton-Ellis, E., Joshi, V., O'Donovan, C., Orchard, S., Auchincloss, A. H., Baratin, D., Bolleman, J., Coudert, E., de Castro, E., Hulo, C., Masson, P., Pedruzzi, I., Rivoire, C., Arighi, C., Wang, Q., Chen, C., Huang, H., Garavelli, J., Vinayaka, C. R., Yeh, L. S., Natale, D. A., Laiho, K., Martin, M. J., Renaux, A., Pichler, K., UniProt, C. UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase // Bioinformatics. – 2020. – Vol. 36, № 17. – P. 4643-4648.

31 UniProt C. UniProt: a worldwide hub of protein knowledge // Nucleic Acids Res. – 2019. – Vol. 47, № D1. – P. D506-D515.

32 Watkins, X., Garcia, L. J., Pundir, S., Martin, M. J., UniProt C. ProtVista: visualization of protein sequence annotations // Bioinformatics. – 2017. – Vol. 33, № 13. – P. 2040-2041.

33 Nightingale, A., Antunes, R., Alpi, E., Bursteinas, B., Gonzales, L., Liu, W., Luo, J., Qi, G., Turner, E., Martin, M. The Proteins API: accessing key integrated protein and genome information // Nucleic Acids Res. – 2017. – Vol. 45, № W1. – P. W539-W544.

34 McGarvey, P. B., Nightingale, A., Luo, J., Huang, H., Martin, M. J., Wu, C., UniProt C. UniProt genomic mapping for deciphering functional effects of missense variants // Hum Mutat. – 2019. – Vol. 40, № 6. – P. 694-705.

35 Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., Aebersold, R. The PeptideAtlas project // Nucleic Acids Res. – 2006. – Vol. 34, № Database issue. – P. D655-8.

36 Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome // Cell Syst. – 2018. – Vol. 7, № 4. – P. 412-421 e5.

37 Moriya, Y., Kawano, S., Okuda, S., Watanabe, Y., Matsumoto, M., Takami, T., Kobayashi, D., Yamanouchi, Y., Araki, N., Yoshizawa, A. C., Tabata, T., Iwasaki, M., Sugiyama, N., Tanaka, S., Goto, S., Ishihama, Y. The

jPOST environment: an integrated proteomics data repository and database // Nucleic Acids Res. – 2019. – Vol. 47, № D1. – P. D1218-D1224.

38 Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., Madhavan, S., Ketchum, K. A. The CPTAC Data Portal: A Resource for Cancer Proteomics Research // J Proteome Res. – 2015. – Vol. 14, № 6. – P. 2707-13.

39 Samaras, P., Schmidt, T., Frejno, M., Gessulat, S., Reinecke, M., Jarzab, A., Zecha, J., Mergner, J., Giansanti, P., Ehrlich, H. C., Aiche, S., Rank, J., Kienegger, H., Krcmar, H., Kuster, B., Wilhelm, M. ProteomicsDB: a multi-omics and multi-organism resource for life science research // Nucleic Acids Res. – 2020. – Vol. 48, № D1. – P. D1153-D1163.

40 Schaab, C., Geiger, T., Stoehr, G., Cox, J., Mann, M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database // Mol Cell Proteomics. – 2012. – Vol. 11, № 3. – P. M111 014068.

41 Fornelli, L., Toby, T. K., Schachner, L. F., Doubleday, P. F., Srzentic, K., DeHart, C. J., Kelleher, N. L. Top-down proteomics: Where we are, where we are going? // J Proteomics. – 2018. – Vol. 175. – P. 3-4.

42 Zardecki, C., Dutta, S., Goodsell, D. S., Lowe, R., Voigt, M., Burley, S. K. PDB-101: Educational resources supporting molecular explorations through biology and medicine // Protein Sci. – 2022. – Vol. 31, № 1. – P. 129-140.

43 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., Hassabis D. Highly accurate protein structure prediction with AlphaFold // Nature. – 2021. – Vol. 596, № 7873. – P. 583-589.

44 Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., Young, R. A. Core transcriptional regulatory circuitry in human embryonic stem cells // Cell. – 2005. – Vol. 122, № 6. – P. 947-56.

45 Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors // Cell. – 2007. – Vol. 131, № 5. – P. 861-72.

46 Takahashi, K., Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors // Cell. – 2006. – Vol. 126, № 4. – P. 663-76.

47 Takahashi, K., Yamanaka, S. A decade of transcription factor-mediated reprogramming to pluripotency // Nat Rev Mol Cell Biol. – 2016. – Vol. 17, № 3. – P. 183-93.

48 Yamanaka, S. Induced pluripotent stem cells: past, present, and future // Cell Stem Cell. – 2012. – Vol. 10, № 6. – P. 678-684.

49 Yamanaka, S., Blau H. M. Nuclear reprogramming to a pluripotent state by three approaches // Nature. – 2010. – Vol. 465, № 7299. – P. 704-12.

50 Chambers, I., Tomlinson S. R. The transcriptional foundation of pluripotency // Development. – 2009. – Vol. 136, № 14. – P. 2311-22.

51 Esch, D., Vahokoski, J., Groves, M. R., Pogenberg, V., Cojocaru, V., Vom Bruch, H., Han, D., Drexler, H. C., Arauzo-Bravo, M. J., Ng, C. K., Jauch, R., Wilmanns, M., Scholer, H. R. A unique Oct4 interface is crucial for reprogramming to pluripotency // Nat Cell Biol. – 2013. – Vol. 15, № 3. – P. 295-301.

52 Merino, F., Ng, C. K. L., Veerapandian, V., Scholer, H. R., Jauch, R., Cojocaru V. Structural basis for the SOX-dependent genomic redistribution of OCT4 in stem cell differentiation // Structure. – 2014. – Vol. 22, № 9. – P. 1274-1286.

53 Tapia, N., MacCarthy, C., Esch, D., Gabriele, Marthaler A., Tiemann, U., Arauzo-Bravo, M. J., Jauch, R., Cojocaru V., Scholer H. R. Dissecting the role of distinct OCT4-SOX2 heterodimer configurations in pluripotency // Sci Rep. – 2015. – Vol. 5. – P. 13533.

54 Kulyyassov, A., Kalendar, R. In Silico Estimation of the Abundance and Phylogenetic Significance of the Composite Oct4-Sox2 Binding Motifs within a Wide Range of Species // Data. – 2020. – Vol. 5, № 4.

55 Kulyyassov, A., Ogryzko, V. In Vivo Quantitative Estimation of DNA-Dependent Interaction of Sox2 and Oct4 Using BirA-Catalyzed Site-Specific Biotinylation // Biomolecules. – 2020. – Vol. 10, № 1.

56 Kulyyassov, A., Shoaib, M., Pichugin, A., Kannouche, P., Ramanculov, E., Lipinski, M., Ogryzko, V. PUB-MS: A Mass Spectrometry-based Method to Monitor Protein-Protein Proximity in vivo // Journal of Proteome Research. – 2011. – Vol. 10, № 10. – P. 4416-4427.

57 Kulyyassov, A. Application of Skyline for Analysis of Protein-Protein Interactions In Vivo // Molecules. – 2021. – Vol. 26, № 23.

УДК 004.651: 004.652

БАЗА ДАННЫХ UNIPROT — УНИВЕРСАЛЬНЫЙ ИНФОРМАЦИОННЫЙ РЕСУРС БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Кулыясов А. Т.*

Ұлттық биотехнология орталығы, Қорғалжынтасжолы, 13/5, Нұр-Сұлтан, 010000, Қазақстан
**kulyyasov@biocenter.kz*

АБСТРАКТ

Последовательности белков хранятся в открытых базах данных, таких как UniProt Knowledgebase (UniProtKB), куда кураторы добавляют результаты экспериментов, данные биоинформатического анализа, включающие предсказание структур и функции биомолекул. Предсказание функции белка может быть сделано с помощью поиска сходства последовательностей, но альтернативным подходом является использование белковых сигнатур, которые классифицируют белки по семействам и доменам. Основные базы данных белковых сигнатур доступны через интегрированную базу данных InterPro, которая обеспечивает классификацию последовательностей UniProtKB. Помимо характеристики белков через белковые семейства, многие исследователи заинтересованы в анализе полного набора белков из генома (т.е. протеома), и существуют базы данных и ресурсы, предоставляющие нередуцированные наборы протеомов и анализы белков из организмов с полностью секвенированными геномами. В этой статье рассматриваются инструменты и ресурсы, доступные в Интернете для определения характеристик, как отдельных белков, так и анализа всего протеома.

Ключевые слова: Association-Rule-Based Annotator (ARBA), European Bioinformatics Institute (EBI), The European Molecular Biology Laboratory (EMBL), The DNA Data Bank of Japan (DDBJ), Gene Ontology Annotation (GOA), Global Proteome Machine (GPM), Mass spectrometry (MS), proteomics, Liquid Chromatography tandem Mass Spectrometry (LC–MS/MS), Multiple reaction monitoring (MRM), National Institutes of Health (NIH), Protein Data Bank (PDB), PRoteomics IDEntifications (PRIDE), Protein Information Resource (PIR), Post-translational modification (PTM), Swiss Institute of Bioinformatics (SIB), the Universal Protein Resource (UniProt), the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProt), the UniProt Reference (UniRef).

ӘОЖ 004.651: 004.652

UNIPROT МӘЛІМЕТТЕР БАЗАСЫ-АҚУЫЗ ТІЗБЕГІНІҢ ӘМБЕБАП АҚПАРАТТЫҚ РЕСУРСЫ

Құлыясов А. Т.*

Национальный центр биотехнологии, Кургальжинское шоссе, 13/5, Нур-Султан, 010000, Казахстан
**kulyyasov@biocenter.kz*

ТҮЙІН

Ақуыздар тізбегі UniProt Knowledgebase (UniProtKB) сияқты қоғамдық дерекқорларда сақталады, онда кураторлар болжамды ақпарат пен эксперименттік деректерді қосады. Ақуыз функциясын болжауды тізбектің ұқсастығын табу арқылы жасауға болады, бірақ балама тәсіл-ақуыздарды отбасылар мен домендер бойынша жіктейтін ақуыз қолтаңбаларын қолдану. Ақуыз қолтаңбаларының негізгі дерекқорлары UniProtKB тізбегін жіктеуді қамтамасыз ететін InterPro интеграцияланған дерекқоры арқылы қол жетімді. Ақуыз тұқымдастары арқылы ақуыздарды сипаттаумен қатар, көптеген зерттеушілер геномнан алынған ақуыздардың толық жиынтығын (яғни протеоманы) талдауға қызығушылық танытады және протеомалардың жиі емес жиынтығын және толық реттелген геномдары бар организмдерден ақуыздарды талдауды қамтамасыз ететін мәліметтер базасы мен ресурстар бар. Бұл мақалада жеке ақуыздардың да, бүкіл протеоманың да сипаттамаларын анықтау үшін Интернетте қол жетімді құралдар мен ресурстар қарастырылған.

Кілті сөздер: Association-Rule-Based Annotator (ARBA), European Bioinformatics Institute (EBI), The European Molecular Biology Laboratory (EMBL), The DNA Data Bank of Japan (DDBJ), Gene Ontology Annotation (GOA), Global Proteome Machine (GPM), Mass spectrometry (MS), proteomics, Liquid Chromatography tandem Mass Spectrometry (LC–MS/MS), Multiple reaction monitoring (MRM), National Institutes of Health (NIH), Protein Data Bank (PDB), PRoteomics IDEntifications (PRIDE), Protein Information Resource (PIR), Post-translational modification (PTM), Swiss Institute of Bioinformatics (SIB), the Universal Protein Resource (UniProt), the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProt), the UniProt Reference (UniRef).