Alfarama Journal of Basic & Applied Sciences
https://ajbas.journals.ekb.eg
ajbas@sci.psu.edu.eg
Faculty of Science Port Said University
http://sci.psu.edu.eg/en/
ISSN 2682-275X
July 2022, Volume 3 Issue II
DOI:10.21608/AJBAS.2022.105632.1078
Submitted: 17-01-2022
Accepted: 27-03-2022
Pages: 346 - 362

Perspectives review and Challenges in Biological Databases Integration

Monica Fawzy[1,*], Noha E. EL-Attar[2], [3]Shady Y. El-mashad and [4]Wael A. Awad

[1]Department of Mathematics and Computer Sciencce, Faculty of Science, Port Said University, Egypt.
[2]Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt.
[3]Faculty of Engineering (at Shoubra), Benha University, Benha, Egypt.
[4]Faculty of computer and information, Damietta University, Damietta El-Gadeeda, Egypt.
*** Corresponding author:** monica.fawzy2012@gmail.com

ABSTRACT

With the enormous and rapid increase in biological data, there has become a significant need in developing biological databases, which help in various researches related to human diseases, drugs Manufacturing, and others based on the stored data. The set prepay to convey back imply distance from fundamental databases has been yoke of the most crucial issues in this field. In this paper, we present a review of biological databases with its types: - primary databases, secondary databases and specialized databases with examples of all of them and a brief description of them and we highlight some perspective, similarities and differences of widely utilized biological databases. Also, we have displayed what researchers have found in developing the biological databases. In addition, the paper presents some of the main challenges that face the biological databases developers such as; big data issues, data errors, and integration of data and state of art on how they can be overcome.

Keywords

B-number, Biological Database, Bioinformatics, DNA.

1. INTRODUCTION

Biological databases can be defined as libraries of life-science information that may be collected from different types of resources, computational analysis, high-throughput experiment technologies, published literature and experimental experiments are only a few examples. [1]. This biological information can be in the form of raw data or measurements stored or exchanged in a digital form in files or databases [2].

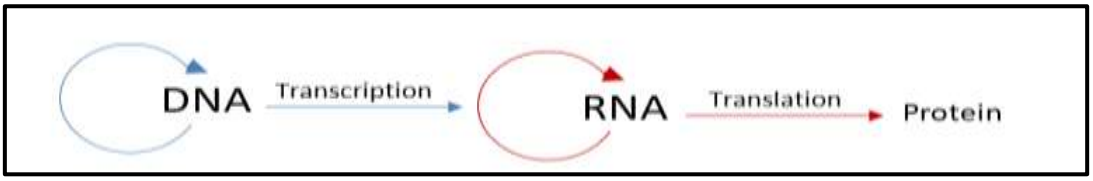
Constructing the biological databases is considered one of the significant issues that is handled by the bioinformatics science. Generally, bioinformatics can be defined as “Bioinformatics is a branch of computer science that includes storing, retrieving, processing, and disseminating data related to biomolecules such as RNA, DNA, and proteins” [1].

The "core dogma" of biology governs the transmission of genetic information in general.; it presents the process of transcribing DNA sequences are translated into RNA, which is ultimately translated into proteins. as shown in **(Fig. 1)**. In order to study the biological databases well, the features of the biological raw data of DNA, RNA, and Proteins need to be defined.

Initially, DNA (Deoxyribonucleic Acid) is the genetic material in humans and almost all other organisms. Most DNA is located in the cell nucleus. Ribonucleic acid (RNA) is a polymer molecule that plays a necessary component role in different biological processes, gene coding, decoding, control, and

expression, to name a few. Nucleic acids, such as DNA and RNA, compose the four primary macromolecules required for all known living forms, along with lipids, proteins, and carbohydrates.

Protein is large biomolecules or macromolecules. It contains a long chain or more of the essential amino acid residue. Proteins have a wide range of roles in animals, including metabolic process stimulation, DNA replication, response to stimuli, cell and organism production, and the movement of molecules from one area to another. Proteins differ primarily in their amino acid sequence, which is determined by the nucleotide sequence in their genes. Proteins normally fold into a precise 3D shape that influences their activity. [4].



(Fig. 1): DNA to Protein Transaction Process

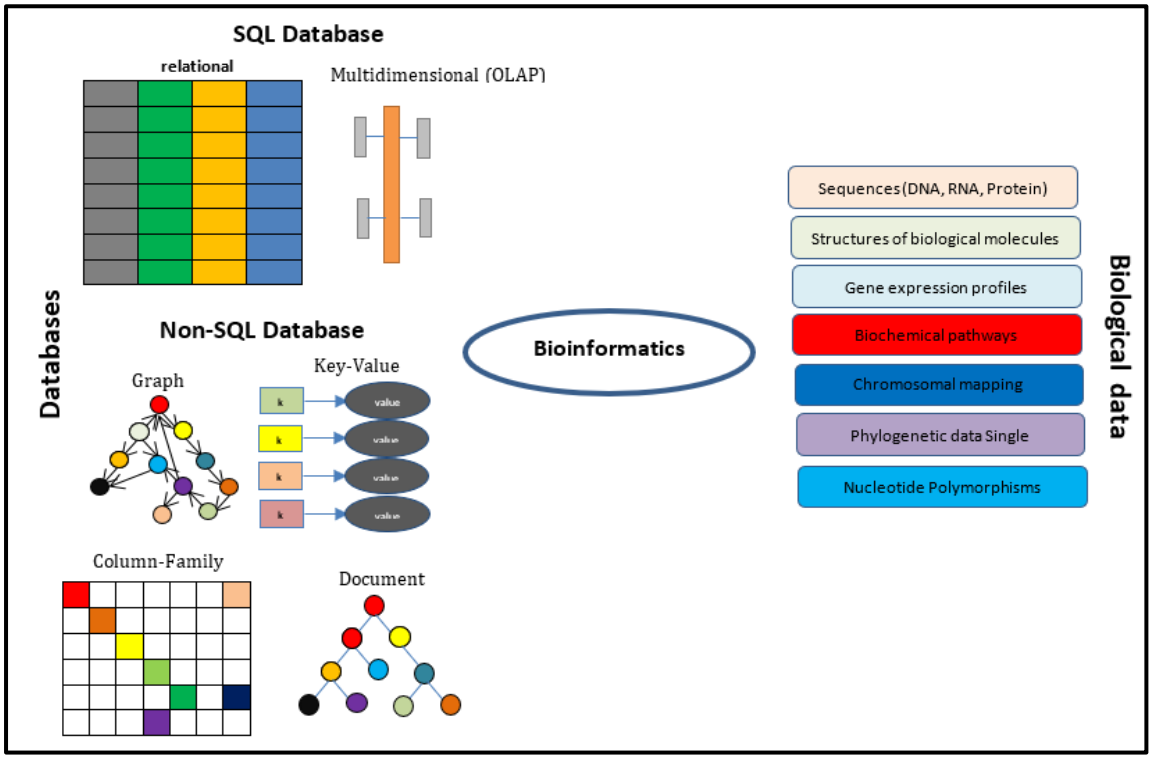
In general, bioinformatics science is interested in two significant fields; the first field is concerned with developing the computational tools and biological databases, while the second one is about how to apply and use these tools and databases to create biological knowledge to enhance the recognition methods inside the living systems [1].

The biological computational tools include several types of developed software for interpreting DNA sequence, structural, or functional analysis. In addition, it is interested in constructing and curating of biological databases which support significant analysis types, such as: molecular structural analysis molecular functional analysis and molecular sequence analysis. Sequence alignment, sequence database searching and style finding, genetics and meadows construction, reconstructing evolutionary links, and genome aggregation and comparison are all part of molecular sequence analysis. The examination, comparison, categorization, and prediction of protein and nucleic acid structure are all part of molecular structural analysis. Finally, the molecular functional analysis handles the methods of determining the gene expression patterns, predicting the protein-to-protein interaction, predicting the subcellular protein localization, rebuilding the metabolism pathways, and the simulation processes. Indeed, these three analysis issues are not individual processes; it might be integrated together to accomplish perfect results [1].

2. BIOLOGICAL DATABASES

Biological data is usually collected from various sources across different structural and functional boundaries. Biological data do not restrict to limited types of information, there is a wide range of data as declared in (Fig. 2). The biological data may appear in the form of genetic sequences, protein functions, medical findings, ecological questions, and others. So, existing of biological databases helps the different types of biological data stakeholders in dealing with the collected biological data which is well organized, easily accessed, centrally managed, and flexibly updated [1].

Databases serve as a repository for information. It's used to store and arrange data so that it's easy to find information using various search criteria. Generally, there is a wide range of databases as declared in (Fig. 2) [5].



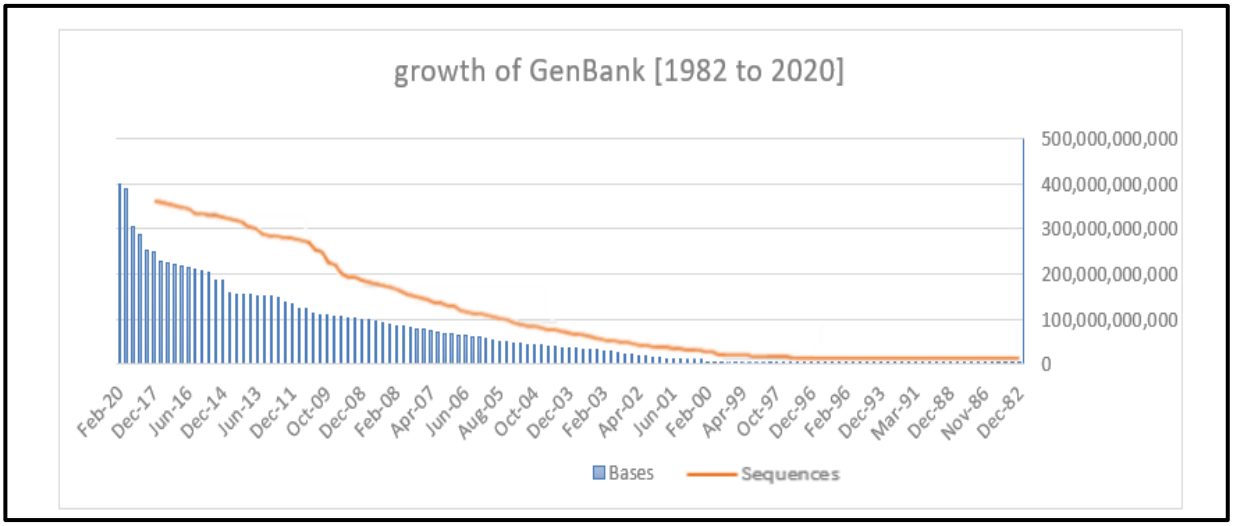
(Fig. 2): Relation between types of databases and biological data (Databases :SQL database or Non-SQL database…..Biological database :sequences, structures of biological molecules, gene expression profiles, biochemical pathways, chromosomal mapping , phylogenetic data single and nucleotide polymorphisms)

The current point of view for developing the biological databases is based on dividing the functions of databases: Primary, secondary, and specialized database into three groups. Before that, they divided the databases into SQL database and Non-SQL database.

The major databases mostly contain experimentally produced data such as nucleotide sequences, protein sequences, and macromolecular structures. The source of data for primary databases are directly submitted from experiments that produce original biological data.

The primary databases are considered the raw materials for the other databases types. Most of the Governments and community funding groups support primary databases financially. GenBank, European Molecular Biology Laboratory (EMBL), and DNA Data Bank of Japan (DDBJ) are the three primary public sequence databases that preserve and disseminate raw nucleic acid sequences [7,1, 8]. GenBank [7] is an annotated repository of all publicly valid nucleotide sequences and their protein translations that is free to the public. The National Center for Biotechnology Information (NCBI) produces and maintains it as part of the International Nucleotide Sequence Database Collaboration (INSDC).

GenBank is growing at an exponential rate, doubling every 18 months. Furthermore, as stated in, its data is viewed and cited by millions of scholars worldwide (Fig. 3).

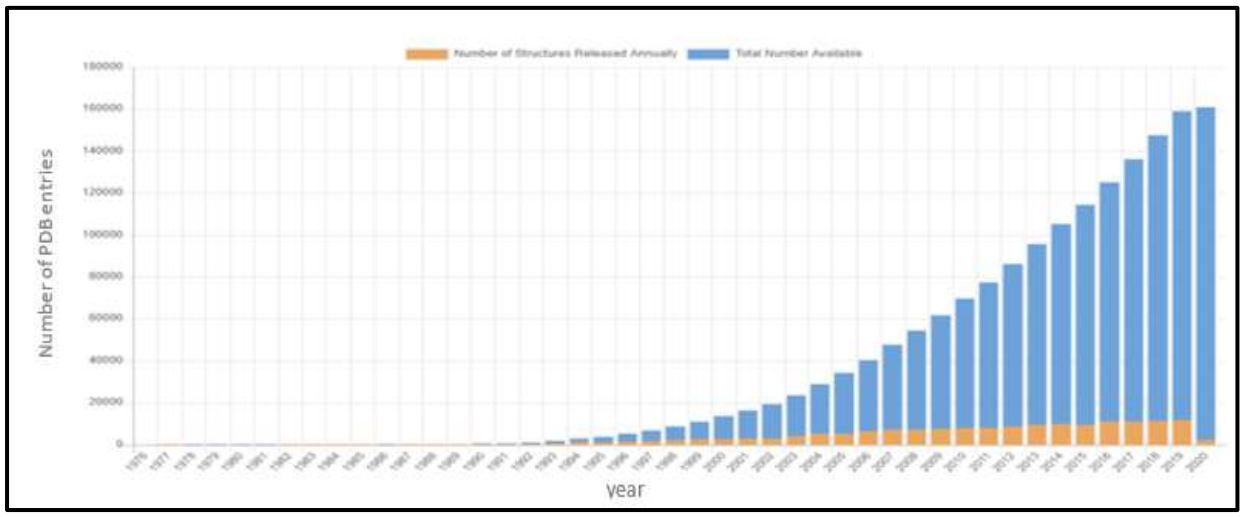


(Fig .3):The Growth of GenBank (1982 - 2020)

The European Molecular Biology Laboratory (EMBL) [1] is a molecular biology research institute financed by 27 member countries. One prospective member state and two associate member states [3]. EMBL was founded in 1974 as an international organisation supported by public research funds from its member countries. Approximately 85 separate groups at EMBL conduct research throughout the gamut of molecular biology. EBML is the name of nucleotide sequence database in Europe, but it is a bit confusing with the name of the institute itself. Therefore, currently European Nucleotide Archive (ENA) is more commonly used.

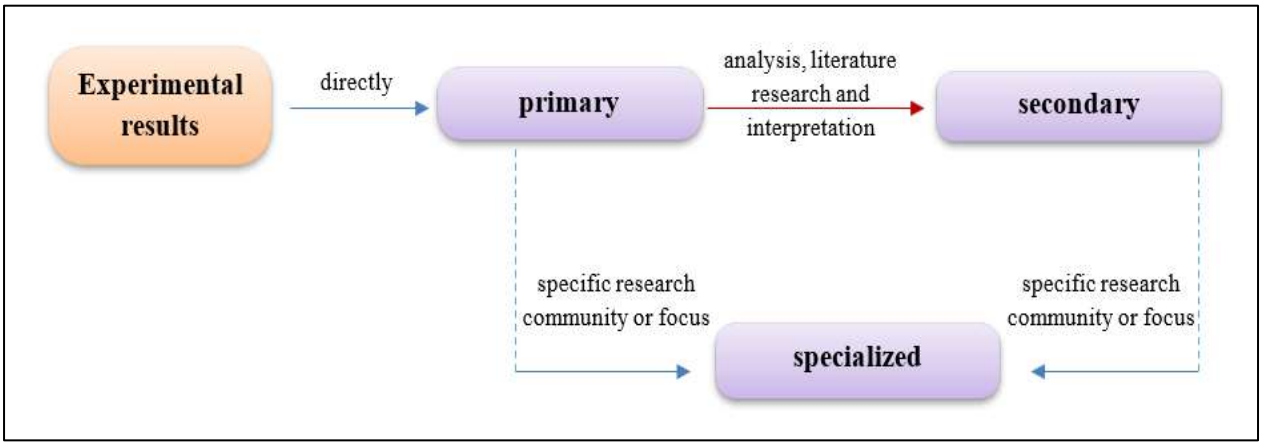
The DNA Data Bank of Japan (DDBJ) is a biological database that stores DNA sequences. It is housed in the National Institute of Genetics (NIG) in Japan's Shizuoka prefecture. DDBJ is also a participant in the International Nucleotide Sequence Database Collaboration (INSDC). It exchanges data on a daily basis with the European Molecular Biology Laboratory at the European Bioinformatics Institute and GenBank at the National Center for Biotechnology Information. As a result, these three databanks have the identical information. They all accept nucleotide sequence submissions and then share fresh and updated data on a regular basis in order to achieve optimal synchronization. Because they include original sequencing data, these three databases are designated as primary databases [1].

Over the last decade, secondary databases have been regarded as the reference library for molecular biologists, giving a plethora of (sometimes inaccurate) information regarding every genetic product or genetic product under investigation by the scientific community. Data in secondary databases is the product of data analysis, literature search and interpretation. Some individuals and companies financially support secondary databases, besides some government bodies that support primary databases. Protein Data Bank (PDB) is an example of a secondary database. PDB's growth is expanding year after year, and its data is viewed and cited by millions of scholars worldwide, as stated in (Fig. 4).



(Fig .4): The Growth of PDB (1976 - 2020)

Specialized databases usually cater to a certain scientific community or are focused on a single organism. These databases may contain DNA or RNA sequences, as well as other types of information for a given organ [1]. (Fig. 5) displays the relationship among the three categories of the biological databases.



(Fig .5):Relationships among the biological databases types

3. COMMON ISSUES IN PUBLIC BIOLOGICAL DATABASE

The ongoing enhancement of biological databases is a major objective of bioinformatics science. The content of these repositories may contain some inaccuracies due to the quick nature of this improvement and the massive volume of data output. We acknowledge that mistakes and omissions can occur at both the sequence and metadata levels in open databases, but for the purposes of highlighting some of the various data integrity difficulties that might arise, we will primarily focus on sequence and taxonomy data concerns. The reassembly of a misassembled Francisella Tularensis genome and the detection of single nucleotide faults in a reference Tobacco mosaic virus (TMV) genome are two high-profile instances of sequence errors [6]. On the other hand, because databases of protein sequences are obtained from genome sequencing, via genome annotation, and in silico translation, the integrity issues for the proteomics database are often identical to those for genomics. A sequence database mistake is unlikely to result in the incorrect detection of a protein present in the sample (false positive), but it might easily result in the failure to identify a protein present in the sample (false negative) (false negative). This is especially important in terms of identifying precise peptide signatures for use in focused experiments [6].

In addition to the mentioned data errors and integrity problems, there are other types of obstacles that may face the biological databases developers such as, redundant data, a steady flux, spreading data from several databases, insufficient information, incorrect links, and unclear naming terms and annotations. All of these challenges make the retrieval and extraction of information more complex. As a result, in order to

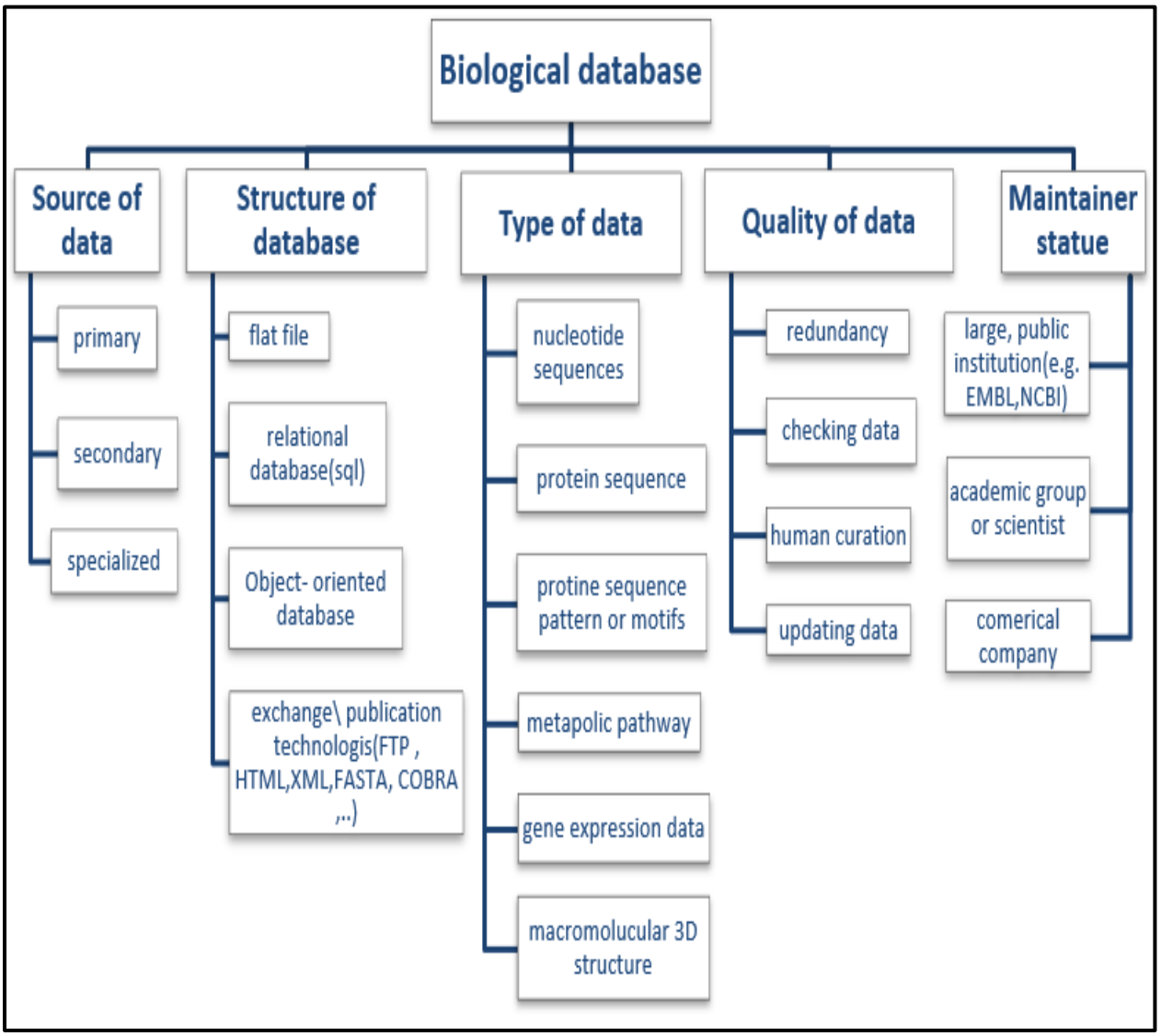
make the data more complete and dependable, the major core databases must interchange data on a regular basis over the internet. However, in many circumstances, most secondary databases are unable to communicate data. Although, the data in the secondary databases is derived from the primary databases, the rate of data update within secondary databases is relatively slow for major databases. This is due to the fact that the amount of data in major databases is often enormous and continually updated [1].

4. POPULAR BIOLOGICAL DATABASES FOR HUMAN RESEARCH

The departed development in computer-based technologies has enabled bioinformaticists and biologists in typical to cope with the information age. The variety in the biological Information sources and structures, made various biological institutions tend to build, develop, and maintain their own biological databases. Hereby, the molecular biology databases have become crucial tools for research.

However, the importance of the biological databases is not limited to keep the biological data, it extends to how the data will be organized is such a way that keeps the search process efficient to find the optimal output. Generally, the database types may vary in their content structure, format, and mode of data accessing [7].

(Fig. 6) displays the classification of the popular biological databases according to their structure, source, type of data, quality of data, and maintainer statue. Also, some of the existing biological databases are presented and classified in table (1) according to their types (i.e. primary, secondary, and specialized) with a brief description for each one.



(Fig .6):Schematic Representation of Biological Databases (the classification of the popular biological databases according to their structure, source, type of data, quality of data, and maintainer statue.)

Table 1: Description for Different Types of Biological Databases

Category	Name	Data modeling	Brief description	URL
Primary Database	GenBank [8]	Flat file	A large public database of nucleotide sequences that may be used for bibliographical and biological annotation. It contains genomic DNA, raw sequencing data from,sequence polymorphisms and high throughput.	https://www.ncbi.nlm.nih.gov/genbank/
	The European Molecular Biology Laboratory (EMBL) [1]	Nosql	An international organisation with over 80 separate research groups encompassing the gamut of molecular biology and Europe's flagship laboratory for the biological sciences — a DNA sequence database.	https://www.embl.de/
	Database and the DNA Data Bank of Japan (DDBJ) [9]	Flat file	All publicly accessible nucleotide and protein sequences are annotated in this database.	https://www.ddbj.nig.ac.jp/index-e.html
Secondary database	Protein Data Bank (PDB) [10]	Flat file (XML)	It keeps track of the three-dimensional structures of big biological molecules like DNA and protein, as well as their complexes.	https://www.rcsb.org/
	The Protein Information Resource (PIR) [11]	FASTA format	A comprehensive general bioinformatics library that aids genetic and protein research.	https://proteininformationresource.org/
	UniProt knowledgebase [12]	Custom flat file, FASTA, GFF, RDF, XML.	It manually gathers annotated data from literature and computer analysis based on values.	https://www.uniprot.org/help/uniprotkb
	neXtProt[13]	FTP	A knowledge platform for human proteins that may be accessed online. It aspires to provide a complete resource for knowledge on human proteins, including their function, subcellular localization, expression, relationships, and role in illnesses.	https://www.nextprot.org/proteins/search
	InterPro [14]	FTP	Predictive model resources, or "signatures," are collected from large protein signature databases and reflect protein domains, families, regions, frequency, and locations.	https://www.ebi.ac.uk/interpro/
	Pfam [15]	Relational	Database to represent the serial alignment of multiple protein families and hidden Markov models	https://pfam.xfam.org

	ConsensusPathDB [16]	BioPAX ,PSI-MI, SBML	Database of functional molecular interactions, fusion of information about protein interactions, signs of genetic interactions, metabolism, gene regulation, and drug-drug interactions in humans	http://consensuspathdb.org/
	23andMe [17]	Relational model	The largest research database that can be reconnected with information on genotypes and patterns in the world. By inviting clients to take part in the research, a new research model has been created that helps speed genetic discovery and provides the ability to obtain new insights more quickly about disease treatments	https://research.23andme.com/
	International HapMap Project [18]	Flat file (XML)	A database built on collaboration between university centres, nonprofit biomedical research organisations, and private firms in Canada, China, Japan, Nigeria, the United Kingdom, and the United States.	https://www.genome.gov/10001688/international-hapmap-project
	Online Mendelian Inheritance in Man [19]	Flat file	A regularly updated database of human genetics, illnesses, and genetic features, with a focus on the genotype-phenotype link.	https://omim.org/
	miRBase [20]	Relational model	A biological database that archives sequences and comments of microRNA	http://www.mirbase.org/
	Human Protein Atlas [21]	Relational model	With various integration approaches, such as antibody-based imaging, mass spectrometer-based proteins, transcriptomics, and system biology, it assigns all human proteins in cells, tissues, and organs.	https://www.proteinatlas.org/
	Structural Classification of Protein (SCOP) [22]	Flat file	Provides a clear and comprehensive account of the links between known protein structures.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102479/
	Class Architecture Topology Homology (CATCH) [23]	Relational model	Performs a semi-hierarchical classification of protein domains.	http://www.cathdb.info/
	Neuroscience Information Framework [24]	Relational model	A repository of global neuroscience web resources, containing experimental, clinical, and infectious neuroscience databases, knowledge databases, atlases and genetic / genomic resources, and providing many reliable links via the Neuroscience Portal at Wikipedia	https://neuinfo.org/

	International Nucleotide Sequence Database Collaboration (INSDC) [25]	Relational model	A long-term founding initiative operating between DDBJ, EMBL-EBI and NCBI. INSDC covers a range of raw data readings, through alignment and aggregations to job annotations, enriched with contextual information related to samples and experimental configurations.	http://www.insdc.org/
	Ensemble [26]	FTP	A database of genomic annotations for chordates and model organisms that is kept up to date.	https://www.ensembl.org/index.html
	RefSeq [27	Relational model	Is a complete, integrated, non-redundant, well-annotated set of sequences derived from sequence data accessible in the redundant archival database GenBank. It includes genomic DNA, transcripts, and proteins.	https://www.ncbi.nlm.nih.gov/refseq/
	ChEMBL [28]	Relational model	A database with a wide-ranging bioactivity of biologically active biologically active compounds that contain medications for functional correlation, Absorption, excretion, metabolism ,distribution, and toxicity in vivo are all factors to consider.	https://www.ebi.ac.uk/chembl/
	KEGG [29]	Flat file (XML)	From a high-level and genetic viewpoint, this resource can help you comprehend facility operations, cells, and organisms.	https://www.genome.jp/kegg/
	Model organism database [30]	FASTA	It is a knowledge resource that contains in-depth biological data on model organisms that have been thoroughly examined. This model allows researchers to easily find basic information about large groups of genes, efficiently plan experiments, integrate their data with existing knowledge, and build new hypotheses	https://www.genome.gov/10001837/model-organism-databases
	RegulonDB [31]	XML	One of the most comprehensive databases for understanding the transcriptional regulatory network in any organism. Initially, it was created as the primary database source for information on the Escherichia coli k-12 clone regulatory network.	http://regulondb.ccg.unam.mx/
	Personal Genome Project [32]	XML	A big, long-term group study aimed at sequencing and releasing the full genomes and medical information of 100,000 participants to allow personal genomics and personal medicine research.	https://my.pgp-hms.org/

	Pathogen-Host Interaction database [33]	XML, FASTA	A biological database containing structured information about genes that have been shown in experiments to alter the outcome of host-host interactions.	http://www.phi-base.org/
	Encyclopedia of Life[34]	Flat File	Is a free, collaborative online encyclopaedia that aims to chronicle all of the world's 1.9 million living species. It is based on existing databases as well as contributions from professionals and non-experts all across the world.	https://eol.org/
	GEO[35]	Flat file and relational model	Is a public repository for functional genomics data that accepts MIAME-compliant data uploads. Data that is based on an array or a sequence is acceptable. Users can utilise the tools to search for and download experiments and curated gene expression profiles.	https://www.ncbi.nlm.nih.gov/geo/
Specialized database	Saccharomyces Genome Database [36]	FASTA	A molecular biology and genetics resource in leaven. Yeast or emergent leaven, Saccharomyces cerevisiae	https://www.yeastgenome.org/
	SNPedia [37]	XML	Human genetics is being researched by a Wiki. This article discusses the impact of genetic variations, referencing peer-reviewed scientific articles.	https://www.snpedia.com/index.php/SNPedia
	UCSC Malaria Genome Browser [38]	FTP, Relational model	A bioinformatics research tool to study the malaria genome.	https://genome.ucsc.edu/
	FlyBase [39]	Relational model	On the reference genome assembly, Flybase employs a generic genome browser to show genome annotations and genome-aligned evidence.	https://flybase.org/
	Protein-protein Interaction Databases [40]	Relational model	a database that is open-source and a collection of tools for storing data, displaying and analyzing rich arranged molecular reaction data in standard formats accepted by members of the community	https://www.ncbi.nlm.nih.gov/pubmed/25859942
	Polymorphism and Mutation Databases [41]	Flat file, FASTA	A collection of small genetic variants from various species. The search tool on the dbSNP webpage allows you to query variants by simple terms or complicated searches. Reference The SNP Cluster Report includes an allele summary, mapping information in HGVS nomenclature, a gene-centric view, a map table with chromosomal coordinates, a variation view, and a link to the 1000 Genomes Browser.	http://www1.biologie.uni hamburg.de/bonline/library/genomeweb /GenomeWeb/human-gen-db-mutation.html

	PeptideAtlas [42]	Object- oriented	Pepteatlas is a statistically validated technique and structure for archiving proteomic data that allows data interchange and integration with genomic data.	http://www.peptideatlas.org/
	PRIDE [43]	Relational model	A collection of mass-spectrometry-based proteomics data, including the identification of proteins, peptides, and post-translational changes documented in peer-reviewed journals.	https://www.ebi.ac.uk/pride/archive/
	MEROPS [44]	Object- oriented	A database that contains information about peptides and the proteins that bind to them.	https://www.ebi.ac.uk/merops/
	Gene Wiki [45]	Flat file (XML)	A collection of community-written Wikipedia articles regarding human genes found in the NCBI Gene database.	https://en.wikipedia.org/wiki/Gene_Wiki
	EcoCyc [46]	HTML	In E. coli, a database is used to record information on proteins and molecular interaction pathways.	https://ecocyc.org/
	Sex-Associated Gene Database (SAGD) [47]	Relational model	A constructed library of generically classified RNA-seq datasets containing female and male biological copies of the same condition, with datasets routinely re-analyzed using established methodologies.	http://bioinfo.life.hust.edu.cn/SAGD#!/
	Super-Enhancer database (SEdb) [48]	sam format	A wide range of active transcription enhancers with enhanced chromatin properties associated with the enhancer	http://www.licpathway.net/sedb/
	Virtual Metabolic Human (VMH) [49]	COBRA format	VMH enables complicated searches of its material, allowing for quick analyses and interpretations of complex data emerging from biological investigations.	https://vmh.life/
	Adaptive Laboratory Evolution (AleDB) [50]	Object- oriented	A method frequently used in biology to study molecular evolution and adaptive changes in microbial groups over a long period of time and under specific growth conditions. With recent breakthroughs in next-generation sequencing technologies and lower sequencing-associated costs, ALE has become more popular as a biotechnology tool	https://aledb.org/
	Catalog of all Germline microsatellites (CAGm) [51]	Relational model	Microsatellites are made up of base pairs 1-6 that repeat side by side to create an array; there are about a million of them in the human genome, most of which are found in gene introns, exons, and regulatory regions.	http://bidd2.nus.edu.sg/CMAUP/

	WormBase [52]	Flat file (XML)	An online biological resource that provides information on the biology and genome of the worm Caenorhabditis elegans, as well as other similar nematodes.	https://wormbase.org//
	Xenbase [53]	FTP	MOD is a model organism database that includes informatics tools as well as genetic and biological information on Xenopus frogs.	http://www.xenbase.org /entry/
	AceDB [54]	Hierarchical object	Acedb is a database system designed primarily for processing genome and bioinformatic data. It comes with a number of useful tools for manipulating, displaying, and annotating genomic data.	http://www.sanger.ac.u k/ science/tools/acedb
	STRING [55]	Relational model	A database of protein-protein interactions that are known and expected.	https://string-db.org/
	EdgeExpressDB [56]	Relational model	A database for analysing biological networks and comparing massive high-throughput expression datasets.	http://fantom.gsc. riken.jp/4/ edgeexpress/about/
	BioGRID[57]	Relational model	A data repository for interactions that has been gathered through extensive curation efforts. All data is publicly available through a search index and may be downloaded in specified forms.	https://thebiogrid.org/

5. APPROACHES FOR IMPROVING THE BIOLOGICAL DATABASES

In order to protect and secure biological databases against any intentional errors, the researchers have resorted to enhance biological databases by several methods such as, constructing validation-specific tools, database enablement specialised ontologies for self-validation which promoting internal control by allowing only skilled curators and annotators to enter data [58].

The problem of biological database integration also have encouraged several researchers to solve it, for instance:

The Biology Workbench [59]: It is the first online application to provide users with a combined environment for tools and data. The BW offers web-based search interfaces through thirty-three (33) databases, save the search results, and send the saved sequence to (66) sequence analysis and presentation programmes [59]. For diverse biological datasets, this method is still the most used. [60].

NC Bioportal Project [61]: It is a bioinformatics portal that may be customised. Bioportal is a website that connects static PISE interfaces with grid computing resources [62].

Thick Client Solutions [63]: Another design approach is smart client software (i.e., packages that must be downloaded and installed locally). These tools allow you to create workflows by connecting tools in a certain order and pipelining data through the process.

Anabench (Biology Workbench) [64]: It is a Web/CORBA-based workbench that allows for the flexible integration of bioinformatics sequence analysis applications. This online application gives you access to tools like the EMBOSS suite and a workflow pipelining system [65], but it doesn't provide you access to data from public sources.

On the other hand, the importance of the biological databases motivated researchers to create and develop new biological databases. Also, it has encouraged them to solve the issues existing in the current databases such as data error and lack of integrity. Generally, there are no standards that can be used for

biological databases development; thus, it is become essential to define a list of IDs genes for each database. Currently, this list of gene identifiers is embedded in a separate file for each database. A class in Java has been developed to read these files, connect to the online services, and then return all accessible information concerning these genes. Following the data retrieval procedure, it is critical to examine each file's structure and create an analyzer to transform the information into XML files. This enables users to perform searches and navigation via the databases easily. This approach has facilitated the data integration process. Also, the procedure of saving data in XML files is a viable replacement for relational databases at the moment. It has become a popular way for representing data changes, publishing it through web, and also building versatile syntax that permits the identical information to be represented in numerous ways [66].

As the knowledge integration is considered one of the most significant obstacles in the science of bioinformatics, recent research directions have concerned with how to apply the integration between the current existing databases. Some databases have achieved the integration via hyperlinks connected to other databases. For instance, more than 160 different databases are linked from the UniProt database. However, the user is left with the decision of which of the offered links to pursue in order to uncover meaningful relationships. Other databases have taken an orthogonal approach to overcome this problem, instead of referring connections to other sources, they just transfer the appropriate material from other sources into the database. There are certain disadvantages to this strategy as well. It creates unnecessary information (which might result in significant storage space consumption). Most importantly, this strategy may result in the utilisation of stale and out-of-date results. [65].

There are also additional issues to consider when considering data integration, such as discrepancies in the amount of genes detected in various databases. Pereira, R. et al. 2014 [66] concentrated on creating a framework for integrating four existing databases: NCBI, EcoCyc, KEGG, and RegulonDB. These databases were chosen because they include information that might be useful in TRN-related research. When comparing KEGG and NCBI, because they both contain the same amount of genes, there are frequently a lot of coincidences. When comparing the EcoCyc database to RegulonDB, the same thing happens. As a result, massive amounts of data had to be gathered in order to accomplish data integration in TRNs. His suggested integration process is centred on producing a B-number, which is often used as a gene identification, as a common feature for all databases. However, the B-number does not apply to all genes. This entails gathering all information about TRNs that is available across different databases and putting it in a single repository known as the KREN.

Other researchers have also noted how to integrate and access diverse biological datasets in an easy-to-use manner that does not need knowledge of the location or source of the underlying data, as well as the technical language used to get the data (e.g., SQL or SPARQL). This technology is inspired by the work submitted by Malick et al. in 2013 [67], which explicitly focused on searching for keywords in relational databases. This technique begins with a description of the layers that comprise an integrated data access system from the bottom up. Base data, data model, integration, and presentation layer are the four basic levels of a data integration access system. However, one disadvantage of this technique is that, in the absence of a presentation layer, such as an easy-to-use query interface, the data contained in global ontology can only be accessible using specialized query languages such as SPARQL. As a result, in order to access or retrieve the data, users must first understand the applicable query language. Table 2 summarizes some of the existing biological database that handled the integration problems with displaying their shortage points.

Table 2: Literature Reviews on Biological Databases and its drawbacks

Database Name	Description	Drawbacks
Gene Wiki Database [68]	I compiled a list of Wikipedia pages regarding human genes that were created by members of the public	Have a collection of community-written Wikipedia entries about human genes that aren't connected to other databases like Gene Wiki
RefSeq Database [69]	The presented fixed database was created using sequence data from the redundant archive database GenBank	Although they are not part of the INSDC, they are produced from INSDC sequences to provide non-redundant curated data that represents our current understanding of known genes. Sequence data from several INSDC records is included in certain instances. As a result, it's a complicated database
Class Architecture Topology Homology (CATH)[70]	Protein domain categorization that is semi-automated and hierarchical.	Without integrating it with another database, I offered a database for small genetic variants from other species
NCBI, EcoCyc, KEGG and RegulonDB [29]	Use B-numbers a gene identifier	Not all genes have a B-number
OMA and Bgee[71]	Use four layers: Base data layer, Data model layer, Integration layer and presentation layer	The data contained in the global ontology is largely available through a technical query language, such as SPARQL, due to the lack of a presentation layer, such as a user-friendly query interface

6. CONCLUSION

In this paper, we have presented a comprehensive review on the major biological databases. Also, we have displayed what researchers have found in developing the biological databases and the challenges they have faced in building new generation of biological databases. The traditional approach to retrieve information from biological databases has been one of the most important topics in this area. Currently, it is getting more difficult as there are a large number of new heterogeneous databases, often sporting different structures, formats, and ways of storing information as well as different ways of providing this information to users. These factors along with the dispersion of information entail the fact that biological data integration process is a very hard task to perform.

REFERENCES

[1] [1] Xiong, J., *Essential Bioinformatics, first edition, copyright by Cambridge University Press, New York*, 2006.doi: 10.1017/CBO9780511806087.

[2] [2] Phillips R., *Communications of the ACM*. USA,2014.doi: 10.1145/2398356.2398365.

[3] [3] Cabestany,Joan, *the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence*, LNCS 5517, pp. 820–828, 2009. doi/10.5555/646370.759075.

[4] [4] Berg, J., Tymoczko, J. and Stryer, L., *Biochemistry, chap3, copyright by W. H. Freeman, New York*, 2002.doi: 10.1042/EBC20160094.

[5] [5] Sarinder K Dhillo, *Biological Databases, Kuala Lumpur, Malaysia* ,2018. doi: 10.1002/0471250953.bi0101s27.

[6] [6] Cooper, B., *Genome Biol*. 15: R67,2014. doi:10.1186/gb-2014-15-5-r67

[7] [7] Yusuf A., Sufyanu Z., Mamman K., Suleiman A., *International Journal of Applied and Advanced Scientific Research*, Page Number 19-28, Volume 1, Issue 1, 2016. doi:10.5281/zenodo.154757

[8] [8] Agarwala R, et al. *NCBI Resource Coordinators*,2015. doi: 10.1093/nar/gkv1290.

[9] [9] *Y. Tateno, T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara, T. Gojobori. Nucleic Acids R*es. 30 (1): 27–30,2002. doi: 10.1093/nar/30.1.27

[10] Berman H., Battistuz T., Bhat T., Bluhm W., Bourne P., Burkhardt K., Feng Z., Gilliland G., Iype L., Jain S., Fagan P., Marvin J., Padilla D., Ravichandran V., Schneider B., Thanki N., Weissig

H., Westbrook J., Zardecki C., *ACTA Crystallographica Section D-Biological Crystallography 58*,2002.doi: 10.1007/978-1-4020-6754-9_13596.

[11] W. C. Barker, J. Garavelli, Hongzhan Huang, P. McGarvey, B. C. Orcutt, G. Srinivasarao, C. Xiao, L. Yeh, R. Ledley, Joseph F. Janda, F. Pfeiffer, H. Mewes, A. Tsugita, Cathy H. Wu, *Nucleic*

[12] Apweiler R., Bairoch A., Wu C., Barker W., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M., Natale D., O'Donovan C., Redaschi N., Yeh Y., *Nucleic acids research* 32. suppl_1, D115-D119,2004. doi: 10.1093/nar/gkaa1100.

[13] Lane L., Argoud-Puy G., Britan A., Cusin I., Duek P., Evalet O., Gateau A., Gaudet p., Gleizes A., Masselot A., Zwahlen C., Bairoch A., *Nucleic acids research* 40.D1 D76-D83,2012. doi: 10.1093/nar/gkr1179.

[14] Mulder N., Apweiler R., Attwood T., Bairoch A., Barrell D., BatemanA., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Sam Griffiths-Jones, Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Sandra E. Orchard, Pagni M., Peyruc D., Chris P. Ponting, Jeremy D. Selengut, Servant F., Christian J. A. Sigrist,4 Robert Vaughan,1 and Evgueni M. Zdobnov, *Nucleic acids research* 31.1 ,2003. doi: 10.1093/nar/gkg046

[15] Finn R., Coggill P., Eberhardt R., Eddy S., Mistry J., Mitchell A., Potter S., Punta M., Qureshi M., V egas AS., Salazar G., Tate J., Bateman A., *Nucleic acids research* 32. suppl_1 ,D138-D141,2004. doi: 10.1093/nar/gkm960.

[16] Kamburov, Atanas, *Nucleic acids research* 41. D1: D793-D800, 2013. doi: 10.1093/nar/gks1055.

[17] Servick, Kelly. *Can 23andMe have it all?*, Vol 349, Issue 6255 • pp. 1472-1477 • DOI: 10.1126/science.349.6255.1472

[18] International HapMap Consortium. *The international HapMap project*, Nature 426.6968 ,2003. doi: 10.1038/nature02168.

[19] Ada Hamosh, Alan F. Scott, Joanna Amberger, David Valle, and Victor A. McKusick. *Nucleic acids research* 15.1: 57-61, 2000.doi: 10.1002/ajmg.a.62407.

[20] Kozomara, Ana, and Sam Griffiths-Jones ,*Nucleic acids research* 42.D1: D68-D73,2014. doi: 10.1093/nar/gkt1181.

[21] Uhlén M. , Björling E., Agaton C., Szigyarto C., Amini B., Andersen E., Andersson A., Angelidou P., Asplund A., Asplund C., Berglund L., Bergström K., Brumer H., Cerjan D., Ekström M., Elobeid A., Eriksson C., Fagerberg L., Falk R., Fall J., Forsberg M., Björklund M., Gumbel K., Halimi A., Hallin I., Hamsten C, Hansson M., Hedhammar M., Hercules G., Kampf K., Larsson K., Lindskog M., Lodewyckx W., Jan Lund, Joakim Lundeberg, Magnusson K., Malm E., Nilsson P., Odling J., Oksvold P., OlssonI., Oster E., Jenny Ottosson, Linda Paavilainen, Anja Persson, Rimini R., Rockberg J., Runeson M., Sivertsson A., Sköllermo A., Johanna Steen, Maria Stenvall, Fredrik Sterky, Strömberg S., Sundberg M., Tegel H., Tourle S., Wahlund E., Waldén A., Wan J., Wernérus H., Westberg J., Wester K., Wrethagen U., Lan Xu L., Hober S., Pontén F.,*Molecular & cellular proteomics* 4.12: 1920-1932, 2005. doi: 10.1074/mcp.M500279-MCP200.

[22] A G Murzin , S E Brenner, T Hubbard, C Chothia., *Journal of molecular biology* 247.4: 536-540, 1995. doi: 10.1093/nar/25.1.236.

[23] Røgen, Peter, Fain B., Proceedings of the National Academy of Sciences 100.1: 119-124, 2003. doi: 10.1073/pnas.2636460100.

[24] D. Gardner , D. H. Goldberg , B. Grafstein , A. Robert, *Neuroinform* (2008) 6:149–160, 2008. doi: 10.1007/s12021-008-9024-z.

[25] Arita M., Mizrachi I., Cochrane G., *Nucleic acids research* 44.D1: D48-D50, 2016. doi: 10.1093/nar/gkaa967.

[26] T Hubbard , D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, T Down, R Durbin, E Eyras, J Gilbert, M Hammond, L Huminiecki, A Kasprzyk, H Lehvaslaiho, P Lijnzaad, C Melsopp, E Mongin, R Pettett, M Pocock, S Potter, A Rust, E Schmidt, S Searle, G Slater, J Smith, W Spooner, A Stabenau, J Stalker, E Stupka, A Ureta-Vidal, I Vastrik, M Clamp., *Nucleic acids research* 30.1: 38-41, 2002. doi: 10.1093/nar/30.1.38.

[27] Pruitt K., Tatusova T., Ostell J., *The NCBI Handbook 2* ,2002. doi: 10.1093/nar/gkv1189.

[28] Gaulton A., Bellis L., Bento A., Chambers J., Davies M., Hersey A., Light Y., McGlinchey S., Michalovich D., Al-Lazikani B., John P Overington., *Nucleic acids research* 40.D1: D1100-D1107, 2012. doi: 10.1093/nar/gkv1189.

[29] Kanehisa, Minoru. The KEGG database, Novartis Foundation Symposium. Chichester; New York; John Wiley;1999, 2002. doi:10.1002/0470857897.

[30] Rhee, Yon S.,*Nucleic acids research* 26.1: 55-59., 2003. doi: 10.1093/nar/gkg076.

[31] Angrist, Misha. *US National Library of Medicine National Institutes of Health* 6.6: 691-699, 2009. doi: 10.2217/pme.09.48.

[32] Winnenburg R., K. Baldwin T., Urban M., Rawlings C., Köhler J., Kim E. Hammond-Kosack. *Nucleic acids research* 34. suppl_1: D459-D464, 2006. doi: 10.3389/fpls.2015.00605.

[33] J. Michael Cherry, *Nucleic acids research* 26.1: 73-79, 1998. doi: 10.1101/pdb.top083840.

[34] Parr, C.S., Wilson, N., Leary, P., Schulz, K.S., Lans, K., Walley, L.,Corrigan JR.,*Biodiversity Data Journal, (2), e1079. Advance online publication*, 2014. doi:10.3897/BDJ.2.e1079.

[35] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Edgar, *Biology, Nucleic Acids Res*, 37 (Database issue), D885-D890, 2009. doi: 10.1093/nar/gks1193.

[36] Cariaso, Mike, and G. Lennon, *Biology, Nucleic Acids Res*, 2010.doi: 10.1007/978-1-4419-9863-7_1039.

[37] M. Mangan, Jennifer M. Williams, Lathe S., D. Karolchik, W. Lathe ,*Biotechnology annual review* 14: 63-108 , 2008.doi: 10.1016/S1387-2656(08)00003-3.

[38] FlyBase Consortium., *Nucleic Acids Research* 27.1: 85-88, 1999. doi: 10.1093/nar/27.1.85.

[39] Xenarios, Ioannis, and Eisenberg D., *Current Opinion in Biotechnology* 12.4: 334-339, 2001.doi: 10.1007/978-1-4419-9863-7_1046.

[40] Elizabeth M. Smigielski, K. Sirotkin, M. Ward, S. Sherry, Nucleic acids research 28.1: 352-355, 2000. doi: 10.1093/nar/28.1.352.

[41] Deutsch EW, Eng JK., Zhang H., King NL., Nesvizhskii AI., Lin B., Lee H., Yi EC., Ossola R., Aebersold R., *Proteomics* 5.13: 3497-3500, 2005. doi: 10.1002/pmic.200500160.

[42] Martens L., Hermjakob H., Jones P., Adamski M., Taylor C., States D., Gevaert K., Vandekerckhove J., Apweiler R., *Proteomics* 5.13: 3537-3545,2005. doi: 10.1002/pmic.200401303.

[43] Rawlings, Neil D., Alan J., Barrett, Bateman A., *Nucleic acids research* 38. suppl_1: D227-D233, 2010. doi: 10.1093/nar/gkh071.

[44] Huss III, Jon W., *PLoS biology* 6.7 ,2008. doi: 10.1371/journal.pbio.0060175.

[45] Keseler I., Mackie M., Zavaleta SA., Billington R., Martínez C., Caspi R., Fulcher C., GamaCastro S., Kothari A., Krummenacker M., Latendresse M., Muñiz-Rascado L., Ong Q., Paley S., Peralta-Gil M., Subhraveti P., David A Velázquez-Ramírez , Weaver D., Collado-Vides J., Paulsen I., Peter D Karp., Nucleic acids research 30.1: 56-58, 2002. doi: 10.1128/ecosalplus.ESP-0009-2013.

[46] Shi M., Zhang N., ShiC., Liu C., Luo Z., Wang D., Guo A., Chen Z., *Nucleic acids research* 47.D1: D835-D840, 2019. doi: 10.1093/nar/gky1040.

[47] Oldridge DA.,Wood AC., Weichert-Leahey N., Crimmins I., Sussman R.,Winter C., McDaniel LD.,
   a. Diamond M., Hart L., Zhu S., Durbin A., Abraham B., Anders L., Tian L., Zhang S., Wei J.,
   b. Khan J., Bramlett K., Rahman N., Capasso M., Iolascon A., Gerhard D., Auvil JG., Young R., Hakonarson H., Diskin S., Look aT., Maris JM., *Nature*, 528.7582: 418-421, 2015. doi:10.1038/nature15540.

[48] Noronha A., Daníelsdóttir A., Gawron P., Jóhannsson F., Jónsdóttir S., Jarlsson S., Brynjólfsson S., Sch neider R., Thiele I., Fleming R., *Computer science, Bioinformatics* 33.4 :605-607, 2017. doi: 10.1093/bioinformatics/btw667.

[49] Phaneuf PV., Gosting D., Palsson BO., Feist AM., *Nucleic acids research* 47.D1: D1164-D1171, 2019. doi: 10.1093/nar/gky983.

[50] Kinney K., Titus-Glover K., Wren JD., Varghese RT., Michalak P., Liao H., Anandakrishnan R., Pulenthiran A., Kang L., *Nucleic acids research* 47.D1 : D39-D45,2019. doi: 10.1093/nar/gky969.

[51] Todd W. Harris, Lee R., Schwarz E., Bradnam K., LawsonD., Chen W., Blasier D., Kenny E., Cunningham F., Kishore R., Chan J., Muller H., Petcherski A., Thorisson G., Day A., Bieri T., Rogers A., Chen C., Spieth J., Sternberg P., Durbin R., Lincoln D. Stein, *Nucleic acids research* 31.1: 133-137, 2003. doi: 10.1093/nar/gkg053

[52] Bowes JB., Snyder KA., Segerdell E., Gibb R., Jarabek C., Noumen E., Pollet N., Peter D. Vize, *Nucleic acids research* 36. suppl_1: D761-D767, 2007. doi: 10.1093/nar/gkm826.

[53] Kimball R, Ross M, Mundy J,. *The kimball group reader*. John Wiley & Sons, New York, NY,2015

[54] Severin, J., Waterhouse AM., Kawaji H., Lassmann T., van Nimwegen E., Balwierz PJ., Hoon MJ., Hume DA., Carninci P., Hayashizaki Y., Suzuki H., Daub CO., Forrest AR., *Genome Biology*, 10(4), R39,2009. doi: 10.1186/gb-2009-10-4-r39

[55] Stark C., Breitkreutz B.J., Reguly T., Boucher L., Breitkreutz A., Tyers M., *Nucleic acids research*, 34(suppl_1), pp. D535–D539,2006. doi: 10.1093/nar/gkj109.

[56] Parr C.S., Wilson N., Leary P., Schulz K.S., Lans K., Walley L.,Corrigan r., *Biodiversity Data Journal*, (2), e1079. Advance online publication,2014. doi:10.3897/BDJ.2.e1079.

[57] Barrett T., Troup D.B., Wilhite S.E., Ledoux P., Rudnev D., Evangelista C., Edgar R., *Nucleic Acids Res*earch, 37 (Database issue), D885-D890, 2009. doi: 10.1093/nar/gkn764.

[58] Caswell J., Gans JD., Generous N., Merkley E., Johnson C., Oehmen C., Omberg K., Purvine E., Tayl or K., Ting CL., Wolinsky M., Xie G., *Front. Bioeng. Biotechnol*. 7:58, 2019. doi:10.3389/fbioe.2019.00058

[59] Subramaniam S., *PROTEINS: Structure, Function, and Genetics*, 1998.doi: 10.1145/360262.360265.

[60] Kohler J., *Drug Discovery Today: BIOSILICO* 2, 61–69 ,2004.doi: 10.1016/S1741-8364(04)02392-3.

[61] RENCI, *international conference on Digital government research*, 2007.doi: 10.1145/1146598.1146708.

[62] Letondal C., *Bioinformatics* 17(1), 73–82 ,2001. doi:10.1093/bioinformatics/17.1.73.

[63] Rifaieh R., *Third International Workshop*, DILS 2006, Hinxton, UK, July 20-22, 2006. Proceedings , 2007.doi: 10.1007/978-3-319-69751-2.

[64] Badidi, E., De Sousa, C., Lang, B.F., Burger, G.: *United Arab Emirates University* 4, 63–72 ,2003.doi: 10.1186/1471-2105-4-63

[65] Badidi, E., De Sousa, C., Lang, B.F., Burger, G., *International Conference on Innovations in Information Technology* ,50(7), 785–793 ,2004.

[66] Pereira,R. and Mendes,R, *International Journal of Bioscience, Biochemistry and Bioinformatics*, Vol. 4, No. 5, September 2014.doi: 10.7763/IJBBB.2014.V4.368

[67] Malick R., Ussama M., and Azim MK., *Journal of Information Technology & Software Engineering*, Issue 1, 3:115, 2013.doi: 10.4172/2165-7866.1000115.

[68] C Orengo , A Michie, S Jones, D T Jones, M B Swindells, J M Thornton., *biology*, S*tructure* 5.8: 1093-1109, 1997. doi: 10.1016/s0969-2126(97)00260-8.

[69] Nuala A.**,** O'Leary, Mathew W.**,** Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K.**,** Kodali, Wenjun Li**.**, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy and Kim D. Pruit, *Nucleic Acids Res*. 2016 Jan 4; 44(Database issue): D733–D745.Published online 2015 Nov 8. doi: 10.1093/nar/gkv1189

[70] Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, *Nucleic Acids Research*, Volume 39, Issue suppl_2, 1 July 2011, Pages W546–W550, doi: 10.1093/nar/gkn877.

[71] Sima AC., Zbinden E., Anisimova M., Gil M., Stockinger H., Stockinger K., Rechavi MR., s*erveur académique Lausannois*, Volume 2019, 2019, baz106, 07 November 2019 .doi: 10.1093 /baz106