International Journal of Applied and Advanced Scientific Research (IJAASR)
Impact Factor: 5.255, ISSN (Online): 2456 - 3080
(www.dvpublication.com) Volume I, Issue I, 2016

COMPARISON OF POPULAR BIOINFORMATICS DATABASES

Abdulganiyu Abdu Yusuf*, Zahraddeen Sufyanu, Kabir Yusuf Mamman* & Abubakar Umar Suleiman***

* Bioresources Development Centre, Kano National Biotechnology Development Agency (NABDA), Abuja - Nigeria

** Faculty of Informatics and Computing, University Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia

Cite This Article: Abdulganiyu Abdu Yusuf, Zahraddeen Sufyanu, Kabir Yusuf Mamman & Abubakar Umar Suleiman, "Comparison of Popular Bioinformatics Databases", International Journal of Applied and Advanced Scientific Research, Page Number 19-28, Volume 1, Issue 1, 2016

Abstract:

Bioinformatics is the application of computational tools to capture and interpret biological data. It has wide applications in drug development, crop improvement, agricultural biotechnology and forensic DNA analysis. There are various databases available to researchers in bioinformatics. These databases are customized for a specific need and are ranged in size, scope, and purpose. The main drawbacks of bioinformatics databases include redundant information, constant change, data spread over multiple databases, incomplete information, several errors, and sometimes incorrect links. Also, standard database, naming conventions, and nomenclature are not clearly defined for many aspects of biological information. Hence, these make information extraction more difficult. In this paper, most widely used bioinformatics databases are presented. These databases are notable for their level of redundancy and annotation, structure coverage and accessibility. They are GenBank, Protein Information Resource **(**PIR), DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL), Protein Data Bank **(**PDB), Universal Protein Resource (UniProt), Swiss-Prot, Structural Classification of Protein (SCOP) and Class Architecture Topology Homology **(**CATH) databases. The key features of the databases are demonstrated and detailed comparisons of the databases were made based on primary and secondary form of databases, and their uniqueness were also highlighted. The databases are foundation stones of bioinformatics and are useful for performing a rigorous benchmarking.

Key Words: Bioinformatics, Databases & Information Technology

1. Introduction:

Bioinformatics is a sub-discipline of computational biology that deals with application of information technology to store, organize and analyze the vast amount of biological data which is available in the form of structure and sequence of proteins and nucleic acids [1]. Bioinformatics being an interdisciplinary field, it combines concepts and techniques from biological sciences, computer science, chemistry, physics, and mathematics. In recent years, bioinformatics is developing rapidly due to an emphasison the efficient use of databases to store, update, query, and retrieve biological data [2]. Bioinformatics database is a combined product of biotechnology and information technology, and plays a vital role in accelerating modern life research science. Due to perceived importance of bioinformatics databases, many countries including the United Kingdom, United State of America, China, Japan and India have been extending a lot of effort to study and construct bioinformatics databases, and private sectors do provide financial support to the governments. These efforts have already made a great contribution to national and regional growth in science and technology. Although several bioinformatics databases were reported from literature, many researchers are not aware of their availability, and these databases were not effectively utilized [3].Recently, most of the databases suffer from several problems unpredicted in early years when their sizes are much smaller. These problems range from lack of standard for annotation, to a large degree of redundancy in databases, and between databases. Some databases are regulated by users rather than by a central body, and sequences are not up to date [4].

This paper studied most frequently used bioinformatics databases. These databases have different efficacies and play important roles in different fields of biology and bioinformatics. They are published based on the following conditions; (i) publicly available, no restrictions, (ii) available, but with copyright, (iii) accessible, but not downloadable, (iv) academic, but not freely available and (v) commercial. The main reason of this study, is to combine the details of some prominent bioinformatics databases into one document. This is to make researchers in the field understand and compare the general concept of many bioinformatics databases that exist in the literature. The rest of the paper is organized as follows: section 2 briefly describes several bioinformatics databases. The key features of selected databases are presented in section 3, focusing mainly on regulating bodies, search tools, file formats, as well as their recent exponential growth. Section 4 compares the performance and uniqueness of the databases, and concluding remark is drawn in section 5.

2. Databases Used in Bioinformatics:

The recent progress in computer-based technologies has enabled bioinformaticists and biologists in general to cope with the information age. Information from different sources presented in the form of biological databases are generally maintained by various institutions. They vary widely in their mode of accessing, content

and format [5]. Biological data are being placed in a database, and the need for data analysis has made molecular biology databases vital tools for research. Thus, a particular database and the easiness to organize data are needed to keep research efficient and to get optimal output.

Characteristics and contents of bioinformatics databases are classified into primary or secondary databases. Primary databases hold primary sequence information, and contain three types of biological data: amino acid sequence, Ribonucleic Acid (RNA) and Deoxyribonucleic Acid (DNA) information [6]. The Primary databases contain extremely large amounts of data, which are updated very fast. They have a lot of users, data management systems with high efficiency, and require big hard disk space. Whereas secondary databases store the analysis of primary sequence information [6]. This classification contains less information than the primary databases and their updating rates are slower [7]. Unlike secondary databases, some important primary databases exchange data every day through the internet, in order to make the data comprehensive and authoritative. Many primary databases contents' make use of dominant or similar ways of data arrangement, while an arrangement of secondary database has a more original design [7].

Previous efforts have been made to maintain accurate lists of available bioinformatics resources, but most of them have not been sufficiently comprehensive due to the slow process of manual duration, or specialized requirements for resource inclusion. For instance, the Database of Databases (DoD) [6] and BioMed Central's Databases catalog (accessible at http://databases.biomedcentral.com/),were previously established but are no longer accessible. DBcat [8] was incorporated into a regular special issue from Nucleic Acids Research (NAR) journal [7], which listed databases that have been published at some point within that journal, and the Bioinformatics links directory [9], which is an associated special issue of the same journal, but focuses on web-services. In addition, some previous work has been relatively limited in scope, focusing either on a specific domain or journals such as phylogenetic software [10], Genome Biology and BMC Bioinformatics [11].

An automatic recognition and extraction of databases names' in full text from literature, enables repeatable discovery of trends in resource usage, allowing a comparison of the most used resources in each field and to discover if other resources are poised to replace their competitors over time. The nine (9) important bioinformatics databases from the literature are GenBank, PIR, EMBL, DDBJ, PDB, UniProt, Swiss-Prot, CATH and SCOP [6]. The main file formats used for the databases are ASN.1, EMBL, Swiss-Prot, FASTA, GenBank/GenPept, PHYLIP, and PIR. A short description of a few selected databases is described in the next section.

3. Description of the Databases:

GenBank: GenBank is a comprehensive public database of nucleotide sequences and supporting bibliography and biological annotation. Its content includes genomic DNA, high throughput raw sequence data, and sequence polymorphisms [12].The database is maintained by National Center for Biotechnology Information (NCBI) in collaboration with European Molecular Biology Laboratory (EMBL), DNA Data Bank of Japan (DDBJ), and European Bioinformatics Institute (EBI). Records in Genbank database can be updated only by the author or third party, if the author has given them permission and notified NCBI. The growth of GenBank is at an exponential rate, doubling every 18 months and its data are accessed and cited by millions of researcher's across the globe [13].As of 15th December 2015, the GenBank contains approximately 203,939,111,071 number of bases from 189,232,925 reported sequences [14] as obtained from Figure 1. Each sequence submitted to GenBank is assigned a unique GenBank identifier or GenBank accession number.

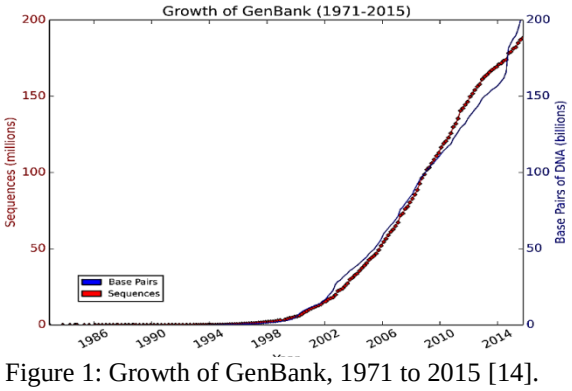

Figure 1: Growth of GenBank, 1971 to 2015 [14].

The Genbank database has four key functions. Firstly, the database frequently stores, updates sequences and annotations. Secondly, it enables querying and comparisons between species by using structured vocabularies. Thirdly, the database architecture allows integration of different biological datasets with the sequences. Finally, it provides a user interface, which can be used for visualization, access, downloading and searching of the data [15].

There are two ways to search for sequences in GenBank are:(i) using text-based keywords similar to a PubMed search, and (ii) using molecular sequences to search by sequence similarity using a Blast. To search

GenBank effectively using the text-based method requires an understanding of the GenBank sequence format. The search output for sequence files is produced as flat files for easy reading. The resulting flat files contain three sections; header, features, and sequence entry [13]. However, each record has a limit of 350k bytes. Longer sequences are broken into segments downloadable from the file transfer protocol (FTP) site. Downloadable versions of these sequence records are organized in several hundred files at the FTP site [16].Most of the sequences are free to use but some have copyright. Information exists in a record can be used to link with another record even from different database. The records can be retrieved by Accession number, Author, Taxonomy, Gene, Keywords and can also be redundant [12]. Details about GenBank can be found online at http://www.ncbi.nlm.nih.gov/genbank.

Universal Protein Resource (UniProt): The Universal Protein Resource is a comprehensive resource for protein sequence and annotation data. The mission of UniProt is to provide the research community with a high-quality, comprehensive, and freely accessible resource of protein sequence and functional information via World Wide Web (WWW) [17]. This database contains a huge amount of information about biological function of proteins curated from the research literature [17]. The PIR, Swiss-Prot and TrEMBL protein databases are merged to form the Universal Protein Resource (UniProt). The UniProt acts as a central resource of protein sequences and functional annotations collaborating with three databases, each of which addresses a key need in protein bioinformatics [18]. The components UniProt database is UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) and UniProt Archive (UniParc).

- ✓ **The UniProt Knowledgebase (UniProtKB):** provides the central database of protein sequences with accurate, consistent, rich sequence and functional annotation. UniProtKB incorporates a range of data from other resources as well [19].
- ✓ **The UniProt Reference Clusters (UniRef)** databases incorporate closely related sequences into a single record for speedy searches and recovery of the data. It provides comprehensive and non-redundant data collections based on the UniportKB in order to obtain complete coverage of sequence space at several resolutions [19].
- ✓ **The UniProt Archive (UniParc)** is also a comprehensive repository containing the history of all protein sequences. The protein sequences are retrieved from predominant, publicly accessible protein information resources. This database gathers all new and updated protein sequences. However, the UniParc contains only protein sequences and database cross-references. All other information must be retrieved from the source databases [19].

The UniProt databases can be accessed freely online via http://www.uniprot.org/ and downloaded in several formats through ptf://ptf.uniport.org/pub. A Blast search can also be done to create alignments. The UniProt database is used by thousands of scientists around the world every day and its website has been visited by over 400,000 visitors in 2014[20]. Figure 2 and Figure 3 depict the distribution of UniProt citations base on the year and according to research areas.

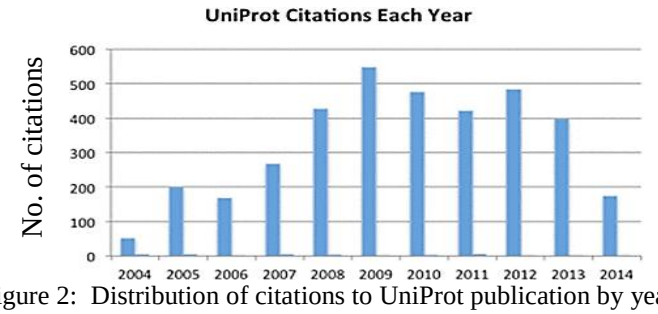

Figure 2: Distribution of citations to UniProt publication by year [20]

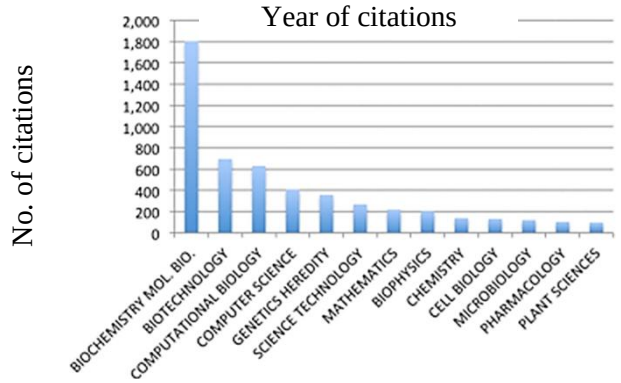

Figure 3: Distribution of number of UniProt publication, according to research areas [20]

Research areas

The main research areas covered by the publications include biochemistry and molecular biology (over 50%), biotechnology, computational biology, computer science and genetics, among others. Note publications may be classified in more than one research area.

Protein Data Bank (PDB): The Protein Data Bank is a primary database maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). It is an archive of three-dimensional structural data of large biological macromolecules such as protein and nucleic acid or their complexes[21]. It allows the user to view data both in plain text and through a molecular viewer using Jmol. Jmol is a free, open source molecule viewer for students, educators, and researchers in chemistry and biochemistry. The main goal of the PDB is to make the data as uniform as possible while improving data accessibility and providing advanced querying options [22].

PDB collects and integrates external data from KEGG Pathways [23], Gene Ontology (GO) [24], Enzyme NCBI resources [25] and Enzyme Commission. Also, it allows data extraction at query run time, which means implemented web services, can extract information as the query is executing. The PDB as a key resource for structural biotechnologists, it stores data that are obtained by X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron microscopy. Scientists from other areas search the PDB to have anidea about the structures of biological macromolecules. The PDB database is updated weekly [26]. As of27thDecember 2015, the breakdown of PDB components are presented in Table 1.

Table 1: Contents of PDB [26]

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid Complexes	Others	Total
X-ray crystallography	95636	1694	4817	4	102151
NMR	9840	1135	231	8	11214
Electron Microscopy	666	29	227	0	922
Hybrid	83	3	2	1	89

The table shows that, most of the data in PDB are determined by X-ray crystallography. More than 12% of structures are determined by protein NMR. Very few protein structures are determined by electron microscopy. The rate of protein structure determination by method and year is given in Figure 4.

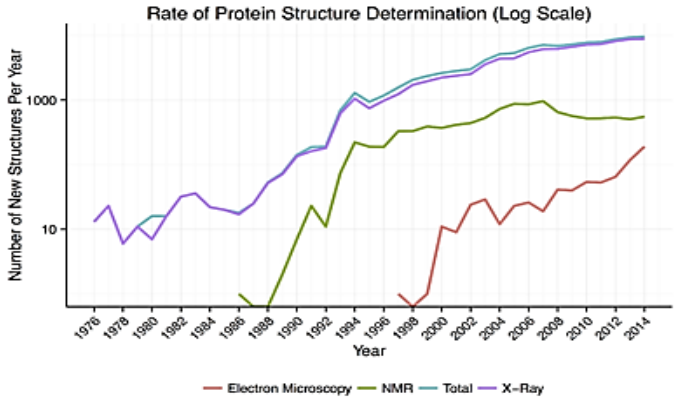

Figure 4: The protein structure determination rate by method and year [26]

The file format that is used by the PDB was called the pdb. It happened late innineties, the "macromolecular Crystallographic Information File" format (mmCIF), were introduced. Then, the details of the file format are presented in an XML version called Protein Databank Markup Language (PDBML). The structure files can be downloaded from the pdb web server in any of a fore mentioned formats [27].For accuracy, PDB files are checked in various aspects such as nomenclature, the chemistry of the polymer and ligands, or stereo-chemical validation. More details about PDB can be accessed online at www.rcsb.org/pdb.

The Protein Information Resource (PIR): The Protein Information Resource is an integrated public bioinformatics resource that supports proteomic and genomic research studies. It was established in 1984 by National Biomedical Research Foundation (NBRF) as a resource to assist researchers and consumers to identify and interpret protein sequence information. It has provided many analysis tools and protein databases to the scientific community, including the PIR-International Protein Sequence Database (PSD) of functionally annotated protein sequences. The PIR-PSD, was originally created as the atlas of Protein Sequence and Structure, and edited by Margaret Day off. It contains protein sequences that were highly annotated with functional, structural, bibliographic, and sequence data [28]. PIR also offers the PIRSF protein classification system [29] that classifies proteins, based on full-length sequence similarities and their domain architectures, to reflect their evolutionary relationships. Moreover, it supports a literature mining resource (iProLINK) in [30], which provides multiple annotated literature datasets to facilitate text mining research in the areas of literature-based database duration, named entity recognition, and protein ontology development. The PIR current version contains four distinct sections, which differ in quality of the data and the level of annotation:

- ✓ PIR1 - fully classified and annotated entries,
- ✓ PIR2 - preliminary entries, not thoroughly reviewed,
- ✓ PIR3 - unverified entries, not reviewed and
- ✓ PIR4 –entries in PIR4 are carefully reviewed, but are untranslated.

The new capabilities for searching PIR sequence databases include domain search, annotated-sorted search, combined global search and interactive text search. The databases and search tools can be accessed at http://pir.georgetown.edu/.

Swiss-Prot: Swiss-Prot was established in 1986and maintained collaboratively since 1987, by Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI). It is a protein sequence and knowledge database and serves as a hub for bimolecular information archived in 66 databases [31]. It is well known for its minimal level redundancy, high quality of annotation, post-translational modifications, use of standardized nomenclature, domains structure, and certain level of integration with other databases. Its format is very similar to that of the EMBL nucleotide sequence database [31]. Since Swiss-Prot is a protein sequence database, its repository contains amino acid sequence, protein name and description, taxonomic data, and citation information. If additional information is provided with the data, such as protein structures, and diseases associated with the protein, then Swiss-Prot provides a table where these data can be stored. Swiss-Protalso combines all information retrieved from publications reporting new sequence data, review articles, and comments from enlisted external experts. In Swiss-Prot, data are stored in two classes; core data and annotations. Core data are the protein sequences, where as annotation consists of the description of the following items: (i) function(s) of the protein, (ii) post-translational modification(s), (iii) domains and sites, (iv) disease(s) associated with deficiencies and (v) secondary structure[32].Swiss-Prot is weekly updated and it is freely distributed as aflat file[33].As of6thJuly 2016, UniProt/Swiss-Prot contains 551705 sequence entries comprising of 197114987 amino acid abstracted from 245756 references as depicted in Figure 5.

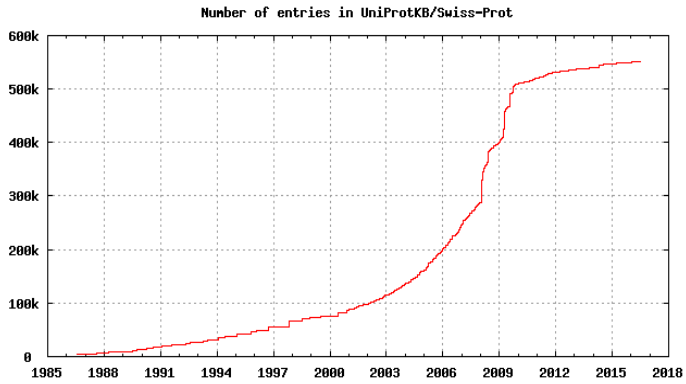

Figure5: Exponential growth of UniProtKB/Swiss-Prot [33]

European Molecular Biology Laboratory (EMBL): The EMBL is a DNA sequence database created in 1980 at the European Molecular Biology Laboratory (EMBL) in Heidelberg. It was maintained since 1994 by European Bioinformatics Institute (EBI)- Cambridge [34]. The DNA and RNA sequences are directly submitted to the EMBL nucleotide sequence database by individual researchers as well as by genome sequencing projects and patent applications, and the database is produced and maintained collaborating with both GenBank and DDBJ. The international collection of sequence data is exchanged between EMBL, GenBank, and DDBJ on a daily basis and a knowledge of global sequence information can be retrieved from any of the three entries. EMBL-EBI data resources are freely available and cover the entire range of biological sciences from raw DNA sequences to curated proteins, chemicals, structures, systems, pathways, and ontologies [34]. Its growth is exponential, on version 3.12.01contains 15,386,184,380 bases in 14,370,773 records (see Figure 6) [35]. EMBL supports several retrieval tools such as;srsfor text-based retrieval, Blast and fasta for sequence based retrieval. Details about EMBL can be found athttp://www.ebi.ac.uk/embl.html.

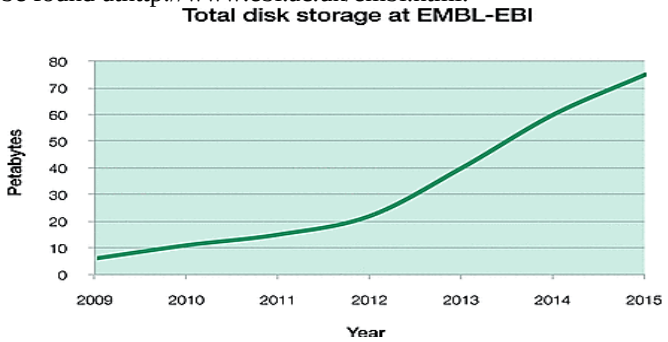

Figure 6: Installed storage at EMBL-EBI [35]

DNA Data Bank of Japan (DDBJ): DNA Data Bank of Japan is a nucleotide database hosted in Japan and is accepting DNA submission from mainly Japanese researchers. It was started in 1984 at the National Institute of Genetics (NIG) in Mishima, and it is an annotated collection of all publicly available nucleotide and protein sequences. The database is maintained inthe instituteby a team of researchers lead by Takashi Gojobori[36].Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country. The objective of DDBJ is to support and promote the sharing and use of biological data as a public resource. They work in close collaboration with GenBank and EMBL,and the three databases store almost identical data. DDBJ also provides various search and analysis tools through the website;http://www.ddbj.nig.ac.jp/. Between July 2010 and June 2011, 18296211 entries and 13576228536 base pair were released from INSD as core traditional nucleotide flat file. DDBJ contributed 12.7% of the entries and 10% of the base pairs added to the core nucleotide data of INSD during this period. Most of the nucleotide data were submitted by Japanese researchers; the rest came from China, Korea, Taiwan and other countries [36]. The major enhanced functions of DDBJ are:(i) DRA data import, (ii) New annotation workflows in the analytical process, (iii) Deletion policy query and result data, (v) MD5 checksum for download files and (iv) Usage statistics information.

Class Architecture Topology Homology (CATH): The CATH Protein Structure Classification method is a semi-automatic, hierarchical classification of protein domains maintained by Institute of Structural and Molecular Biology (ISMB). The database was first published in 1997 by Orengo et. al, [37]. The accompanying website (http://www.cathdb.info/) provides an easy-to-use entry to the classification, allowing for both browsing and downloading of data. The latest version of CATH (version 4.0) adds annotations for over 62000 new structural CATH domains and over 100 new superfamilies. This increase was enabled by a number of improvements to the automated assignments made in the CATH update workflow. CATH carries many broad features with its principal rival, Structural Classification of Protein (SCOP). However, there are also many areas in which the detailed classifications in the two databases differ greatly [38]. The name CATH is an acronym of four main levels classified hierarchically; Class (C), Architecture (A), Topology (T), Homologous Super family (H). The description of the four main levels of the CATH is given in Table 2.

Table 2: Levels and Description of CATH

Level	Description
Class (C)	Class is derived from secondary structure contents of proteins. It is assigned for more than 90% of protein structures automatically.
Architecture (A)	Architecture describes the gross orientation of secondary structures, independent of connectivity in proteins.
Topology (T)	A large-scale grouping of topologies which share particular structural features.
Homologous Super family (H)	Homologous super families of CATH cluster the proteins with highly similar structures & functions.

In CATH the Class level classification is done on the basis of the following 4 criteria: secondary structure contents of proteins, secondary structure alternation score, secondary structure contacts of proteins, and percentage of parallel strands in proteins [38].

Structural Classification of Protein (SCOP): The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of known protein structures. The whole concept is based on similarities of the amino acid sequences and three-dimensional structures of the proteins. The database was originally published in 1995 and it is usually updated at least once yearly by Alexei G. Murzin [39]. The classification of proteins in SCOP is on hierarchical levels as shown in table 3;

Table 3: Classification of proteins in SCOP is on hierarchical levels

Level	Description
Class	It is the general structural architecture of the protein domains
Fold	It represents the similar arrangement of regular secondary structures but without evidence of evolutionary relatedness.
Super family	It represents whether the protein structures have sufficient structural and functional similarities to each other to infer a divergent evolutionary relationship but not necessarily a detectable sequence homology. Example, the variable and constant domains of immunoglobulin's
Family	Proteins belonging to the same family share some sequence similarity.

This classification of proteins in SCOP is more significantly based on the human expertise. It is generally acknowledged that SCOP gives a better justified classification. The human expertise decides whether some proteins are evolutionarily related and therefore should be placed in the same super family, or their similarity is only a result of structural constraints present in the proteins to classify them to the same fold [39]. The SCOP version 1.75C classifies 167,547 domains, 59,514 PDB entries, 7 class, 1961 Super family, 4493families, as of October 2013. The SCOP database can be accessed at http://scop.mrc-lmb.cam.ac.uk/scop/.

4. Comparison of the Databases:

The outstanding bioinformatics databases described in section 3 are categorized into primary and secondary databases. In this section, a general comparison is made between the primary and secondary databases. Then, the performance of selected databases iscompared according to above categorization. Moreover, the uniqueness of the databases was also highlighted. Finally, the main features of the aforementioned data sets are summarized in one representation. The chief differences between the databases highlighted in this study are;(i) Most primary databases are financially supported by the governments and society funding bodies, while some secondary databases are financially supported by enterprises and individuals, besides the government and society funding bodies. (ii) In order to make the data comprehensive and authoritative, some important primary databases will exchange data every day through Internet, while secondary databases cannot exchange data in most cases. (iii) The amount of data in primary databases is huge and updated quickly, whereas the data in secondary databases are based on a primary database and often require software for exploitation. The updated rate of the secondary database's data is far slower than that of the primary database. (iv) Many primary databases contents make use of prevalent or similar ways to the arrangement of data, while the arrangement of secondary database content has a more original design. (v) Experimental results are submitted directly to the database by researchers, and the data are essentially archival in nature. In contrast, secondary databases consist of data derived from the analysis of primary data such as sequences and secondary structures.

Primary Databases: Experimental results are submitted directly to the database by researchers, and the data are essentially archival in nature. From the study, GenBank, DDBJ, PDB, SWISS-PROT, PIR, EMBL, and UniProt are exampes of primary databases. Other primary databases that were not explained or mentioned in this study include Medline, IMEx, Ensembl, Genome Server, Genome-MOT, Mutations, and IMGT. They are not as challenging as the corresponding primary databases considered in this study.

Secondary Databases: In contrast, secondary databases consist of data derived from the analysis of primary data such as sequences, secondary structures. Examples include SCOP, CATH, and UniProt. Other examples are PROSITE, BLOCKS, PRINTS, OMIM, RefSeq, Pfam, Taxon, MIPS, InterPro, TrEMBL, NRL-3D, GOA and GenPept. After studying details of the databases, it can be concluded thatmost of the databases have a web-interface to search for data. The Common mode to search is by Keywords, or Cross-references that help to navigate from one database to another easily. Retrieval systems help to extract rich information from multiple databases. The most widely used biological data bank resource on the World Wide Web is the genomic information stored in the National Institutes of Health's GenBank, U.S.A, the European Bioinformatics Institutes' EMBL, and Japan's National Institute of Genetics DNA Databank of Japan [13, 31, 36]. The three databases, under the direction of the International Nucleotide Sequence Database Collaboration (INSDC), gather, maintain, and share mainly nucleotide data, each catering to the needs of the region in which it is located.

Protein Data Bank has some better representation such as very large macromolecules and X-ray structures refined with multiple models. The most difficult problem for the PDB is that, the files are not uniform hence, contain numerous inconsistencies and errors. There are inconsistent residue numbering and missing values for experimental parameters [27]. The PIR is the only sequence database that provides context cross-references between its own database entries. These cross-references assist search in exploring relationships such as subunit associations in molecular complexes, enzyme-substrate interactions, as well as in browsing entries with shared features and annotations. A recent effort to combine SWISS-PROT, TrEMBL and PIR led to the creation of the UniProt database, which has larger coverage than any one of the three databases, while at the same time maintaining the original SWISS-PROT feature of low redundancy, cross-references, and a high quality of annotation. Swiss-Prot has poor sequence coverage, highly structured, excellent annotation. The SWISS-PROT database has some legal restrictions: the entries are copyrighted, but freely accessible to academic researchers. Commercial companies must buy a license fee from SIB. The uniqueness about SCOP is that it embeds a theory of evolution as defined by experts, rather than the necessarily more limited set of rules implemented by a series of algorithms and automatic tools. The structural classifications of proteins are generally obtained from SCOP and CATH. The main features of the aforementioned databases are summarized in Table 4.

Table 4: An overview of the selected databases for bioinformatics

Names	Search Tools	File Formats	Regulating Body	Year	Country	Category
GENBANK	Blast	ASN.1, XML	NCBI	1982	USA	Primary
EMBL	SRS, Blast, Fasta	Flat-File	EMBL, EBI, SRS	1980	UK	Primary
UNIPROT	Fasta, Blast	Flat-File	EBI, SIB	2003	UK, Switzerland	Primary/Secondary
PDB	Fasta	PDB, XML	RCSB	1971	USA	Primary
PIR	Fasta, Blast	PIR	NBRF	1984	USA	Primary
SWISS-PORT	Blast	SWISS	SIB	1986	Switzerland	Primary
DDBJ	Blast, Fasta	Flat-File	NIG	1984	Japan	Primary
CATH	Fasta	Fasta	ISMB	1997	UK	Secondary
SCOP	Fasta, Blast		LMB,CPE	1994	UK	Secondary

The table gives a clear and thorough description of the databases in one representation. The parameters include; the database name, search tools, file formats, regulating bodies, year of creation with countries and the categorization into primary or secondary database, in that order. It can be observed that the most common tools used for search are Blast and Fasta. Also, most of the databases originated from UK and USA.

5. Conclusion:

We presented key features of the selected bioinformatics databases such as the regulating bodies, year of creation with countries, availability, file format, recent advances and their uniqueness. Then, the performance of the databases was compared based on the primary and secondary database. The scope of this paper is limited to some notable bioinformatics databases, other databases were taken into consideration for the sake of comparison. One thing that worth mentioning here is, there are certainly other very important and useful biological databases which store information about proteins. But the databases that have been discussed here are the most important ones and used by the majority of biologists and biotechnologists. This review may serve as a good starting material for researchers interested to use and process the biological information. It will also give them room to debate on some questions like; (i) which types of databases are good for annotation? (ii) How are the databases accessible? and (iii) does primary databases better than secondary databases in term of redundancy and errors. There are a lot of bioinformatics databases containing a lot of valuable information, and the best databases do not contain everything. Finally, there is no doubt that bioinformatics databases for efficient research have a significant impact in biological sciences and betterment of human lives.

6. Acknowledgement:

This study is supported by National Biotechnology Development Agency (NABDA), Abuja, Nigeria and University Sultan Zainal Abidin (UniSZA), Kuala Terengganu, Malaysia.

7. References:

1. Ramsden Jeremy J, Bioinformatics: An Introduction 2nd edition (Springer-Verlag Limited, London), 2009.
2. Duck G, Nenadic G, Brass A, Robertson DL, Stevens R. Extracting Patterns of Database and Software usage from the Bioinformatics Literature. BMC Bioinformatics. 2014 Aug; 30(17):i601–i608. doi: 10.1093/bioinformatics/btu471 PMID: 25161253.
3. Babu, P. A., Boddepalli, R., Lakshmi, V. V., & Rao, G. N. (2005). Dod: Database of databases–updated molecular biology databases. In silico biology, 5(5, 6), 605-610.
4. Duck, G., Nenadic, G., Brass, A., Robertson, D. L., & Stevens, R. (2013). bioNerDS: exploring bioinformatics' database and software use through literature mining. BMC bioinformatics, 14(1), 1.doi: 10.1186/1471- 2105-14-194 PMID: 23768135.
5. Duck G, Nenadic G, Brass A, Robertson DL, Stevens R. bioNerDS: Exploring Bioinformatics' Database and Software use through Literature Mining. BMC Bioinformatics. 2013; 14(1):194. doi: 10.1186/1471- 2105-14-194 PMID: 23768135.
6. Köhler, Jacob. "Integration of life science databases." Drug Discovery Today: BIOSILICO 2.2 (2004): 61-69.
7. Babu, P. A., Udyama, J., Kumar, R. K., Boddepalli, R., Mangala, D. S., & Rao, G. N. (2007). DoD2007: 1082 molecular biology databases. Bioinformation, 2(2), 64-67.Available from: http://www.ncbi.nlm.nih.gov/pmc/.doi: 10.6026/97320630002064.
8. Galperin MY, Cochrane GR. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Research. 2011 dec; 39(Database issue):D1–D6. doi: 10.1093/nar/gkq1243 PMID: 21177655.
9. Discala C, Benigni X, Barillot E, Vaysseix G. DBcat: A Catalog of 500 Biological Databases. Nucleic Acids Research. 2000 Jan; 28(1):8–9. doi: 10.1093/nar/28.1.8 PMID: 10592168.
10. Fox, J. A., Butland, S. L., McMillan, S., Campbell, G., & Ouellette, B. F. (2005). The Bioinformatics Links Directory: a compilation of molecular biology web servers. Nucleic acids research, 33(suppl 2), W3-W24.doi: 10.1093/nar/gki594 PMID: 15980476
11. Eales, J. M., Pinney, J. W., Stevens, R. D., & Robertson, D. L. (2008). Methodology capture: discriminating between the" best" and the rest of community practice. BMC bioinformatics, 9(1), 1.doi: 10.1186/1471- 2105-9-359 PMID: 18761740.
12. Duck G, Nenadic G, Brass A, Robertson DL, Stevens R. bioNerDS: Exploring Bioinformatics' Database and Software use through Literature Mining. BMC Bioinformatics. 2013; 14(1):1. doi: 10.1186/1471- 2105-14-194 PMID: 23768135.
13. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic acids research, 41(D1), D36-D42.
14. National Center for Biotechnology Information (NCBI). GenBank Release Notes 213.0.http://www.ncbi.nlm.nih.gov/genbank/release/213/. Accessed on 20th July 19, 2016.

15. Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, Mooney P, Tivey A, Berriman M, Hall N, Rutherford K, Parkhill J. (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. Nucleic acids research, 32(suppl 1), D339-D343.
16. File Transfer Protocol (FTP) site for GenBank Nucleotide Sequence. ftp://ftp.ncbi.nih.gov/genbank/. Accessed on 20th July, 2016.
17. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ. (2005). The universal protein resource (UniProt). Nucleic acids research, 33(suppl 1), D154-D159.
18. O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, A., & Apweiler, R. (2002). High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Briefings in bioinformatics, 3(3), 275-284.
19. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 23(10), 1282-1288. doi:10.1093/bioinformatics/btm098.
20. Oxford University Press. UniProt: A Hub for Protein Information. ," Nucleic Acids Research, 2014. doi:10.1093/nar/gkq989.
21. Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, Westbrook J. (2001). The PDB data uniformity project. Nucleic Acids Research, 29(1), 214-218.
22. Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, Zhang Q, Knezevich C, Xie L, Chen L, Feng Z, Green RK. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic acids research, 33(suppl 1), D233-D237.
23. M. Kanehisa and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nuc. Acids Res., 28(1): 27–30, 2000.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. (2000). Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.
25. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO. (2004). Database resources of the National Center for Biotechnology Information: update. Nucleic acids research, 32(suppl 1), D35-D40.
26. RCSB Protein Data Bank. http://www.rcsb.org/pdb/home/home.do. Accessed 19th July 19, 2016.
27. Berman, H. M. (2008). The protein data bank: a historical perspective. Acta Crystallographica Section A: Foundations of Crystallography, 64(1), 88-95. doi:10.1107/S0108767307035623.
28. C. H. Wu, L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Z. Hu, R. S. Ledley, P. Kourtesis, B. E. Suzek, C. R. Vinayaka, J. Zhang, W. C. Barker, The Protein Information Resource, Nuc. Acids Res., 31: 345–347, 2003.
29. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, Ledley RS.. (2004). PIRSF: family classification system at the Protein Information Resource. Nucleic acids research, 32(suppl 1), D112-D114.
30. Hu, Z. Z., Mani, I., Hermoso, V., Liu, H., & Wu, C. H. (2004). iProLINK: an integrated protein resource for literature mining. Computational biology and chemistry, 28(5), 409-416
31. Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A, Castro M. M. (2006). EMBL nucleotide sequence database: developments in 2005. Nucleic acids research, 34(suppl 1), D10-D15.
32. J. T. L. Wang, C. H. Wu, and P. P. Wang, Computational Biology and Genome Informatics, Singapore: World Scientific Publishing, 2003.
33. UniProtKB/Swiss-Prot Release Statistics. http://web.expasy.org/docs/relnotes/relstat.html. Accessed on 20th July 2016.
34. Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., & Thornton, J. (2014). The european bioinformatics institute's data resources 2014. Nucleic acids research, 42(D1), D18-D25.
35. Gibson, R., Alako, B., Amid, C., Cerdeño-Tárraga, A., Cleland, I., Goodgame, N., ten Hoopen, P., Jayathilaka, S., Kay, S., Leinonen, R. and Liu, X., 2016. Biocuration of functional annotation at the European nucleotide archive. Nucleic acids research, 44(D1), pp.D58-D66. Doi:10.1093/nar/gkv1311.
36. K. Okubo, H. Sugawara, T. Gojobori, and Y. Tateno, DDBJ in Preparation for Overview of Research Aactivities behind Data Submissions Nuc. Acids Res., 34(1): D6–D9, 2006.
37. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. Structure, 5(8), 1093-1109. doi:10.1016/S0969-2126(97)00260-8.
38. Cuff, A.L., Sillitoe, I., Lewis, T., Clegg, A.B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. and Orengo, C.A., 2011. Extending CATH: increasing coverage of the protein

structure universe and linking structure with function. Nucleic acids research, 39(suppl 1), pp.D420-D426. doi:10.1093/nar/gkq1001.

39. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin, "SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data," Nucleic Acids Research, 32(suppl 1), 2004, pp. D226-D229. doi:10.1093/nar/gkh039.