Proceedings of the Pakistan Academy of Sciences 52 (2): 91–102 (2015)

ISSN: 0377 - 2969 (print), 2306 - 1448 (online)

Pakistan Academy of Sciences

Research Article

A Computational Framework for Unified Biological Databank to Overcome Heterogeneity of Biological Data Format

Muhammad Usman Ghani Khan, Sana Amanat, Abdul Nasir, Razi Iqbal, and Muhammad Idrees

Bioinformatics Lab, Al-Khwarizmi Institute of Computer Science,
Department of Computer Science and Engineering,
University of Engineering and Technology,
Lahore, Pakistan

Abstract: Building block of life is DNA which produces RNA and protein sequences; all these molecular blocks lie in the category of biological data. Biological data which comprises various organisms like animals, plants, viruses, bacteria and humans,throughout the globe has heterogeneous nature. This heterogeneity is caused by different parameters such as storage mechanism, information representation and content format. The abundance of heterogeneity creates a havoc towards optimized exploitation, integration, storage and retrieval of existing biological data. This work introduces a computational framework for achieving a unified format for biological data, which can accommodate different formats of storage; and data presentation. An information retrieval system for biological data has been developed which sheds light on different recompenses gained by this unified format of the biological data.

Keywords: Biological data, heterogeneous format, unified format, integration, parser

1. INTRODUCTION

Biological information is electronically available in the form of databanks under certain categories of organisms such as Homo sapiens, plants, virus, bacteria and animals [1]. These databanks have towering versatility of biological data formats as compared to other information [2]. These biological data are available in scrappy form under certain molecular categorization like DNA, RNA and protein sequences.

There are multiple sources of molecular databases, e.g., DNA Databank of Japan (DDBJ) [3], National Centre for Biotechnology Information (NCBI) [4], European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) and some others [4, 5]. These databanks provide heterogeneous [6] biological data under certain formats. It is observed that this heterogeneity leads to numerous hurdles towards the storage, retrieval, data mining [2], semantic search, integration [7], discovery and optimized utilization of existing biological data [8].

Upon critical inspection of different databanks, i.e., GenBank, DDBJ and EMBL-EBI, we formalize three levels of heterogeneity in relation to biological data which are storage mechanism, content format and information representation format. Storage mechanism: leads to the storage techniques used to store molecular information physically such as txt, dat, seq and web pages. Content format: expresses the methodology to store a sequence such as FASTA [9], GenBank, and EMBL. Information Representation: depicts the format in which biological information is presented to the user, e.g., as text and images. Table 1 shows three levels of heterogeneity for existing biological databanks, i.e. storage, presentation and content.

Received, April 2014; Accepted, April 2015
*Corresponding author: Muhammad Usman Ghani Khan; Email: usman.ghani@uet.edu.pk

Table 1. Three levels of biological data format heterogeneity.

Molecule type	Data bank	Storage format	Presentation	Content format
DNA	DDBJ	Dat, Seq, Fasta	Text	Fasta, GenBank
	ENA	Seq	Text	EBML
	NCBI	A, Fasta	Text	Fasta, GenBank
Human Genes	GeneCards	Web Page	Text, Image	-

So far, it seems as if there is no such versatile databank with unified format that can accommodate all levels of heterogeneity, with respect to biological data. It is important to note that it is not a complete list of available formats across the globe, but it provides the most common list of biological databanks with their format information. To cope with heterogeneity, we propose a unified format which is capable to accommodate all levels of heterogeneity.

1.1 Contributions and Goals

Development of unified biological databank with unified format has following goals and contributions:

1. To provide unified biological repository.
2. To provide unified format for heterogeneous biological data formats.
3. Integration of all molecular data at a single point.
4. Generation of a biological databank service in Pakistan.
5. 24/7 Retrieval of unified biological databank.
6. Availability of updated biological information.
7. Download services for databank.
8. Addition of molecular data discovered at wet labs for sharing with the public.
9. Provide an efficient, authorized and scalable hub for biological research purposes.

2. MATERIALS AND METHODS

This section covers the physical arrangement of databank, which is composed of various modules and processes such as data acquisition process, computational modules (download server, biological data processing engine, data server and reporting server), security mechanism of databank and interface for internal and external communication. Furthermore, the conversion from heterogeneous to unified format is also discussed in this section.

2.1 Enactment of Unified Biological Databank

Physically unified biological databank [10] is based on four core components. Fig.1 shows the abstract overview of unified biological databank.

2.1.1 Data Acquisition Process

There are various sources of biological data such as wet lab and dry lab [11]. By considering dry lab, the data are available in diverse forms i.e. web pages, images, videos and text files. In the context of text file format there are various existing databanks which provide molecular information for different organisms like *Homosapiens*, *Musmusculus*, *Rattusnorvegicus*, *Oryzasativa* and others [12]. These databanks provide information to the public for further utilization per requirements. In this work we have downloaded various data sets for numerous molecules and organisms. Some of the considered data sources for our work are given below:

1. DDBJ : ftp://ftp.ddbj.nig.ac.jp/ddbj_database/
2. ENA: ftp://ftp.ebi.ac.uk/pub/databases/embl/cds/
3. NCBI: ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/
4. Ensemble: ftp://ftp.ensembl.org/pub/release72/fasta/homo_sapiens/dna/
5. Vector Base: https://www.vectorbase.org/downloads
6. RNAfam: ftp://ftp.sanger.ac.uk/pub/databases/Rfam

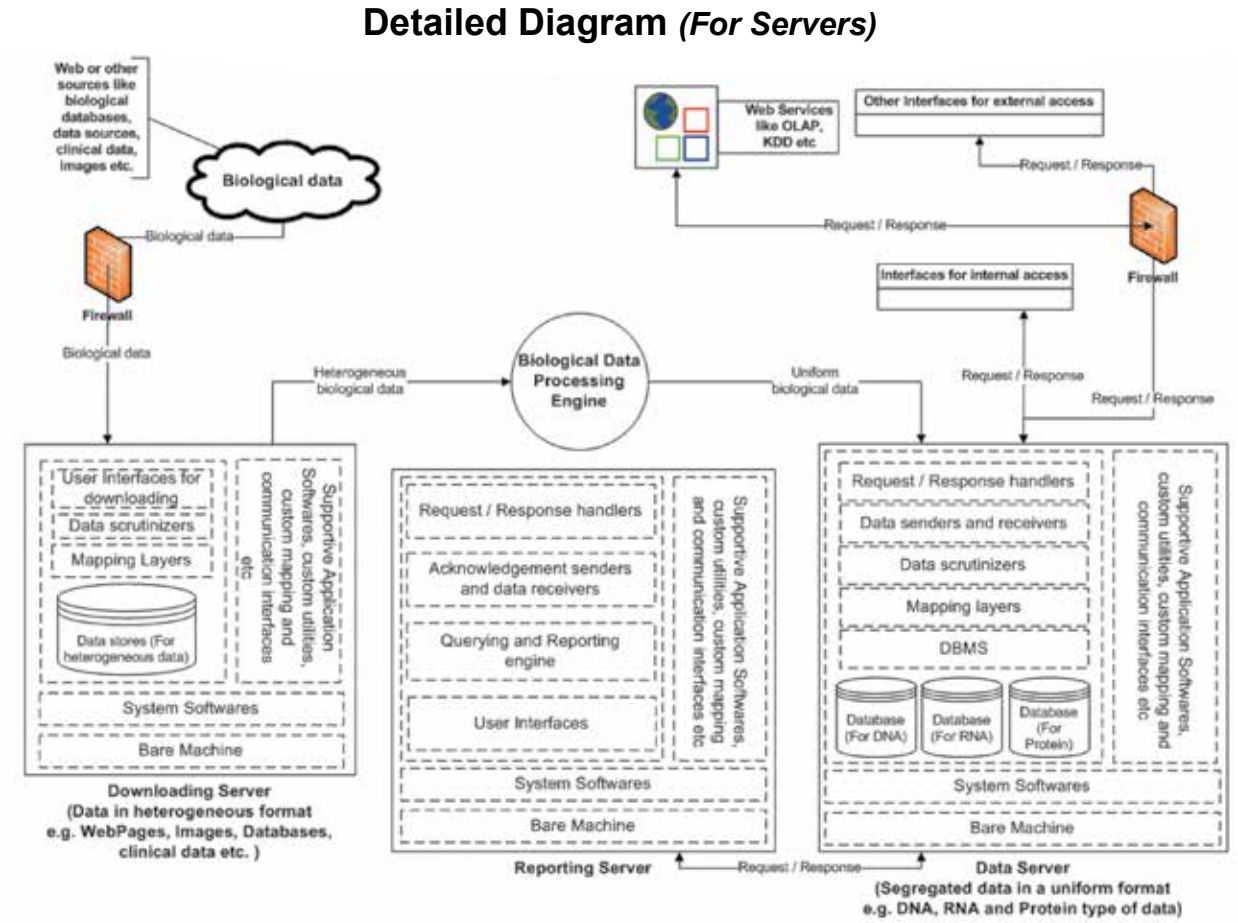

Fig. 1. Proposed framework for unified biological databank.

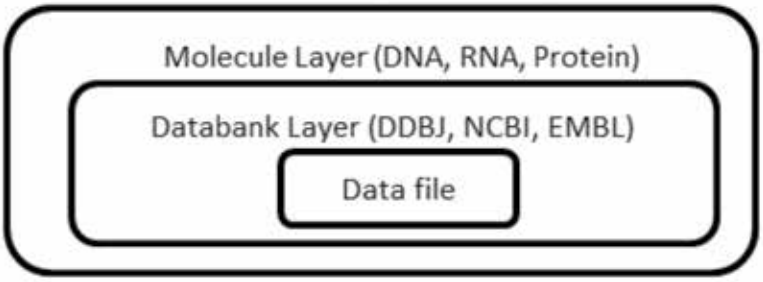

Fig. 2. Abstract view of data mapping layers.

7. miRbase: ftp://mirbase.org/pub/mirbase/CURRENT/

The above list consists of the most common but not all databanks which are publically available.

2.1.2 Download Server

The outcome of data acquisition process is the data source links through which we can download our required data on our download server. These data source links are stored into local index file and a request is send to these links one by one under file transfer protocol. On the behalf of request, a response is generated by the server [13] for the links which communicate with the download server and firewall to make the internet traffic more secure. After negotiating the protocol destination file server provides a list of all those files which exist on that link.

Firstly, download server sends a request to the data source link under file transfer protocol to start the communication. In response to the sent request, data source file acknowledges and negotiates the underlying protocol to start the transfer of data. As soon as download server receives the acknowledgement from download server, it starts receiving the list of all those files which exist on that link. This transfer continues till the end of last file. Next step of this process is to read the content of those files which exist in the list of those files received from particular data source link.

Local index file which contains the list of all those files which exists on a particular data source link is read out. The first index of local index file is the name of the file which exists at index 0 on that data source link. After getting the first index from local index file it is searched on download server that either this file already exists or not, if that file does not exist then a new file is created in the name of 0^{th} index file. Next step is to read the file content from source server and write that content into newly created file at download server. This process goes on till the end of the file.

Other than text files, biological information which is available on web pages is also gathered using the hypertext transfer protocol (HTTP) [14]. To gather the interested set of biological information, a request is sent to the particular web server under http and the response is generated by the web server which permits the download server to receive the requested data. Download server has several parsers which read the content of that particular web page.

Since there exist a lot of other information in addition to biological information, so utilities to scrutinize the data are applied for filtering, sorting [15], trimming [16] and examining the required biological information from the web pages. At the end of this process the content and data source files are stored on download server on underlying mapping layers which determines the hierarchy in which data has to be stored physically as shown in Fig. 2.

Some of the download utilities interact with the underlying system software throughout the process from request to the storage of data on particular memory location of data server. System software interacts with the bare machine to accomplish the task of storage.

2.1.3 Biological Data Processing Engine

The main purpose of this databank is to provide biotic information in unified form using some standards and rules. This engine takes heterogeneous data which is in FASTA, GenBank and EBI format from data server and on the behalf of incoming format it chooses an appropriate parser that converts it into unified format.

Rule 1: If incoming stream has FASTA format then extracted identifier, organism and sequence are to be converted into unified format.

Rule 2: If incoming stream has DDBJ format the command is to extract all tags which are in upper case for conversion into unified format.

Rule 3: If incoming stream has EBI format, it

commands to extract all tags such as ID, XX, and SQ to convert those into unified format.

2.1.3.1 Data server: The outcome of biological data processing engine is the unified data which has to be stored on data server. To start the storage on data server biological data processing sends a request to the data server using request handler. After getting the response from data server it receives the data from biological data processing engine and starts receiving the data. On this received unified data, scrutinizer algorithms are applied using utilities which examine the type of molecule, databank and organism. After examining the incoming data, mapping layer communicates with the underlying database management system which defines the rules to store the unified biological data under certain molecule categories such as DNA, RNA and protein. System software interacts with bare hardware to make possible the storage of unified biological data on data server. Rules which are applied to store the unified data are as under:

Rule 1: If incoming data has DNA molecule and organism is animal then store it into DNA repository under "wildlife" category.

Rule 2: If incoming data has DNA molecule and organism is Homo sapiens then store it into DNA repository under "human" category.

Rule 3: If incoming data has Protein molecule and organism is animal then store it into protein repository under wildlife category.

Rule 4: If incoming data has RNA molecule and organism is animal then store it into RNA repository under wildlife category

2.1.3.2 Report server: To make sure the availability of unified biological databank a user interface, i.e., web page is provided via report server. A user can give input and sends request to generate the desired report using a secured channel such as firewall to control the incoming and outgoing network traffic. The process from user input to the generation of report is as follows:

Whenever user provides the input a request handler of report server sends a request to the data server to enable the communication for data exchange. Data server sends an acknowledgement using response handler indicating a ready state for data transfer. After negotiating the communication protocols the query analyzer performs the examination on requested query.

The user's query and input parameters are provided to the query engine which executes the incoming query and communicates with the data server to fetch the required data from unified database. At this stage report server and data server simultaneously communicate with each other using message passing to execute the query and transfer of data. Data server transfers the requested data which is received on report server to generate the reports.

2.1.3.3 Options: To generate report there are many options available such as:

1. *Generate report on web page:* If the users avails this option then generated report has to be showed on the web page and user can easily approach it.
2. *Generate report which is downloadable:* In this option if the user does not want to receive the report of web page and is desirous to download this report for later use, then pdf based reports be provided to facilitate the user.
3. *Sends the generated report via email:* If user wants to receive its generated report into mail box then the user has to provide his valid email address.

2.1.3.4 Levels: There are many users of this databank and every user wants the information on report according to his level and requirements. There are some of the levels of report content such as:

(a) Biologists, pathologists: These users require more detailed report because they deal with the

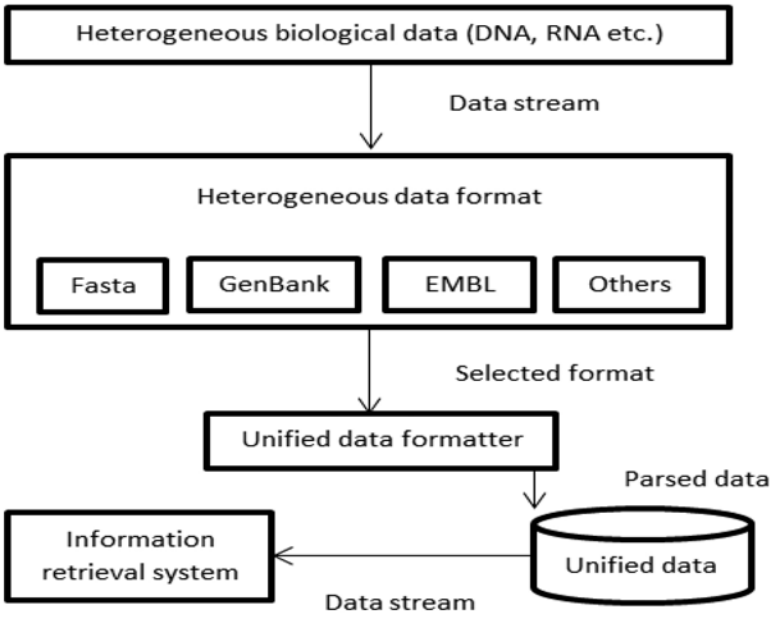

Fig. 3. Conversion steps from heterogeneous to unified format.

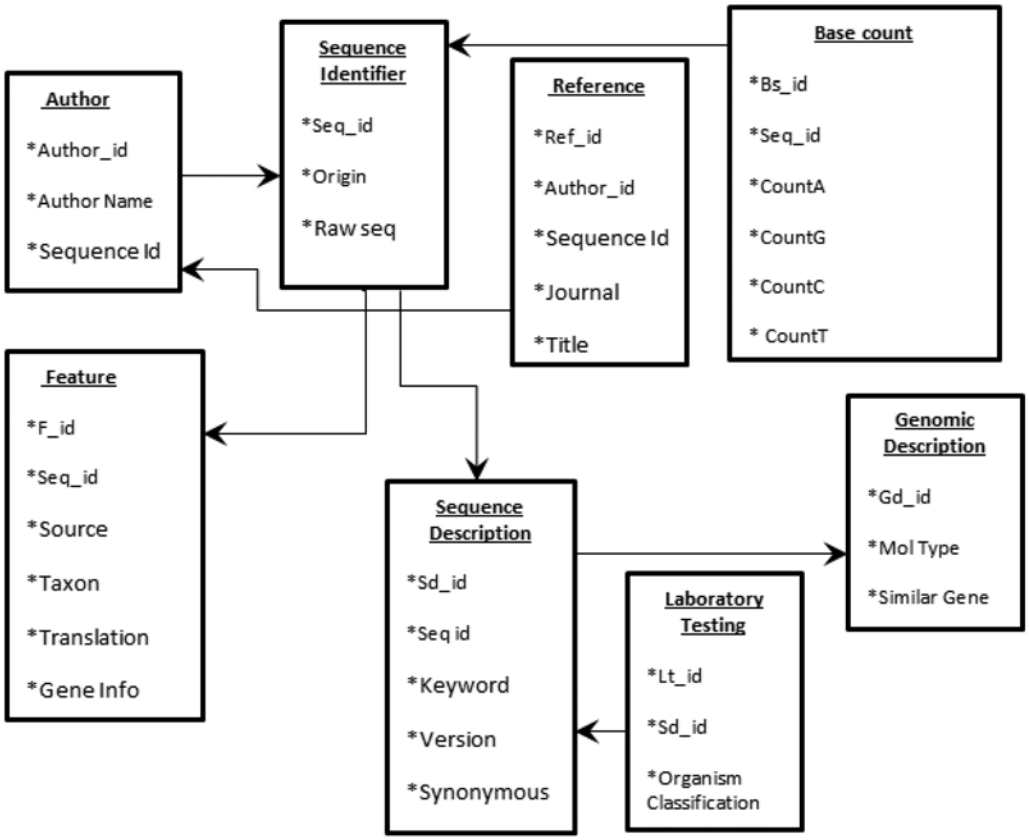

Fig. 4. Physical structure of unified format.

invention, diagnose, repairing and treatment of molecular level substance.

(b) *Researchers:* They deal with the invention and research of biological data at molecular level for all organisms. So they required report which contains the more information about taxonomy.

(c) *Layman, newbies:* They use this databank for studies, for DNA matching and for learning purpose.

2.2 Exertion of Unified Format

To convert the heterogeneous format into unified format for biological data a framework is proposed which receives heterogeneous biological data, analyze it, selects a particular parser, converts it into unified format and stores it into unified databank. Framework of unified format is shown in Fig. 3. For internal processing of unified formatter, we designed numerous parsers for each type of incoming format based on heterogeneous data stream. Upon careful inspection of input data stream of heterogonous nature, a particular parser is selected by the unified formatter that automatically converts and accommodates all the fields into unified format to overcome three levels heterogeneity. Outcome of this unified data formatter is the unified biological data without any kind of heterogeneity treated as unified data set. This data set is accessible via information and retrieval system using different searching parameters. Detail of this framework is as follows:

2.2.1 Heterogeneous Data Format

Biological data is available in heterogeneous formats such as FASTA, GenBank and EMBL. Some formats are less descriptive like FASTA and other formats provides sequence related information in details, most common features of formats are identifier, organism and raw sequence.

2.2.2 Unified Data Formatter

Working of data formatter is as follows:

Downloading of Heterogeneous biological data: According to Table 2 all downloading links of various databanks are maintained in a file and files are downloaded on our down server using the following algorithm.

2.2.2.1 Download files: Get file list form the link by sending request to the ftp server.

Step1- > IF: File does not exist or time; start download.
Else: Goto Step2

Step2-> IF: Time or date of file that exist in our system is less than time on the ftp server. Delete the file and start download again.
Else: Goto Step3

Step3- > IF: Size of the file that exists on our system, is less then size on the ftp server. Delete the file and start download again.
Else: Goto next iteration of loop (Goto next file.)
It is important to store the files in some layered fashion physically. We maintained this approach using the following algorithm.

2.2.2.2 Hierarchy maintenance: Read all the databanks from the file databank.txt

Table 2. Various attributes of existing heterogeneous format.

Format	Fields
FASTA	Definition line, sequence
GenBank	Accession, Version, Reference, Author, Origin, Reference, Keywords, Source, Title, Journal,Remark,CommentFeatures,Basecount,Consortium
EMBL	Accession, Project, Date, Keyword, Organism classification, Organelle, Reference Number, Reference Position, Reference Comment, reference cross-reference, Reference Group, Reference Author, Reference Title, Reference Location, Third Party Annotation, Feature Header, Feature Table

Step1-> IF: Folder of the databank in the databank. txt, present. Go to Step2.
Else: Create the Folder and its txt file for databanks and Go to Step2.Checking for all databanks in the .txt file of the particular databank using threads. For each Data bank of databank in threads= number of banks of a particular databank.

Step2-> IF: Data bank folder exist then
Go to Step3
Else: Create Folder and Go to step 3.

Step3- > Read links from the .txt file of the databanks and then start to download files.

Upon careful inspection of these downloaded data files it can be argued that these files contains most common formats such as FATSA, EBML and GenBank format and has these following fields which are shown in Table 2.

2.2.3 Unified Format

Unified format holds all the fields of heterogeneous formats as mentioned in Table 2 and some of the additional fields. Unified format contains all the fields of FASTA, GenBank and EMBL format. Detail of all fields of unified format with physical structure is as shown in Fig. 4.

All these tables are linked via a sequence identifier which plays a role as primary field and all other fields are considered to be the secondary fields because they provide data related to the primary field. This schema contains various tables such as author table which contains information of author and sequence identifier. Sequence table stores the information about sequence only such as raw sequence which is a combination of nitrogen bases, its original unique identifier and the identifier which we used to store this sequence in our database.

Other tables contains the information of base pairs of raw sequence, definition, organism name, molecule type and the information of publication with respect to author, journal title, sequence. Furthermore the additional information in form of annotation is also stored in feature table and header table format. Detail of some of the extra fields in unified format is as below:

2.2.4 Secondary Identifier

This is the unique identifier which is assigned to every record which has to be entered into our databank. Format of this field is as: name of original data bank, name of organism, actual unique identifier. Actual identifier could be strain, accession number, genome identifier or any unique property that can distinguish every record, i.e.:

"Name of reference databank + name of organ/ organism+ uniquely identified feature e.g. DDBJ_ HomosapienDNA_AB1001"

CountA: Contains total count of adenine base in a sequence.

CountG: Contains total count of guanine base in a sequence.

CountC: Contains total count of cytosine base in a sequence.

CountT: Contains total count of thymine base in a sequence.

2.2.5 Parser Rule

There are certain rules which are followed while converting heterogeneous format into unified format.

Rule 1: Tag matching for GenBank format: If data stream has any tag locus, accession, version, keyword then split the stream and store the chunk of stream from current Index+1 to the length of stream. If data stream does not have any tag then concatenates it with the previous stream.

Rule 2: Tag matching for FASTA format; If data stream starts with > symbol then split the stream by special letters and store each splitting region from current Index to the length of stream by the increment of 1.

Rule 3: Tag matching for EMBL format. If data stream has any tag ID, OS, FT, PA, SQ then split the stream and store the chunk of stream from current Index+1 to the

length of stream: If data stream does not have any tag then concatenates it with the previous stream.

2.2.6 Unified Dataset

Outcome of the parsing is the unified biological data which is stored in a repository under molecule category such as DNA, RNA and protein sequence. This unified dataset is actually known as unified biological databank.

2.2.7 Information Retrieval Tools

Unified biological databank is made accessible via information retrieval engine. Where users can input any query such as accession number, raw sequence, author name, organism, specie and other parameters. Against these parameters information retrieval engine searched the repository and show the results to the user.

2.2.8. Demo

We have developed a prototype of unified biological databank, which will be deployed at the link http://kics.edu.pk/edanbioinformatics/home.html. Currently, the web page indicates the objectives and main milestones of the project, while the link and accessibility to the utilization of software/databases will be provided to public at the completion of the project**.** This application initially downloads existing heterogeneous biological data at download server and converts it into unified format using biological processing. Unified biological data is stored on our data server, from which user can download unified data and generate reports. Initially searching criteria from unified biological repository is accession number, which can be enhanced in the upcoming versions.

In order to use this application, the user has to enter the accession number of a sequence as input to our system. In addition to input, the user has to select the target database from which the search will be carried out. After providing input the result will be shown on the screen consisting of accession, organism and sequence as shown in Fig. 5.

3. RESULTS AND DISCUSSION

There are multiple formats of biological data and it is analyzed that FASTA, GenBank and EMBL formats are fiasco to accommodate all types of molecule information by compensating all levels of heterogeneity. It is clear that unified format is more versatile which has enough capability to accommodate variety of formats without losing any useful information. It accommodates all fields of heterogeneous formats, all formats of files such as seq, fa, fasta, xls, txt and others using parsing rules. So we can argue that unified format tackles all issues of heterogeneity.

3.1 Evaluation

To validate this unified format, we store various datasets of FASTA, GenBank and EMBL format in unified format by using numerous parsers. This unified repository is accessible via information retrieval system. Information retrieval system takes input which is passed to the information retrieval engine to perform analysis. Searching criteria is extracted from the input query for processing and the required data is fetched based on provided searching criteria.

The output of query processor displays on graphical user interface in a unified manner.

3.2 Demonstration

Initially, we used a single parameter; accession number as searching criteria keeping in mind the fact that every record has an accession number. On the behalf of provided input retrieval system provides the name of author, sequence ID and author ID. This result is purely based on user end, which fields or information they want in return. We will extend the searching criteria in next versions. Fig. 5 shows the sample of output based on unified format.

To evaluate the performance of this tool, a task based evaluation strategy was planned, where an end user is asked to search and retrieve some specific record according to his requirement.

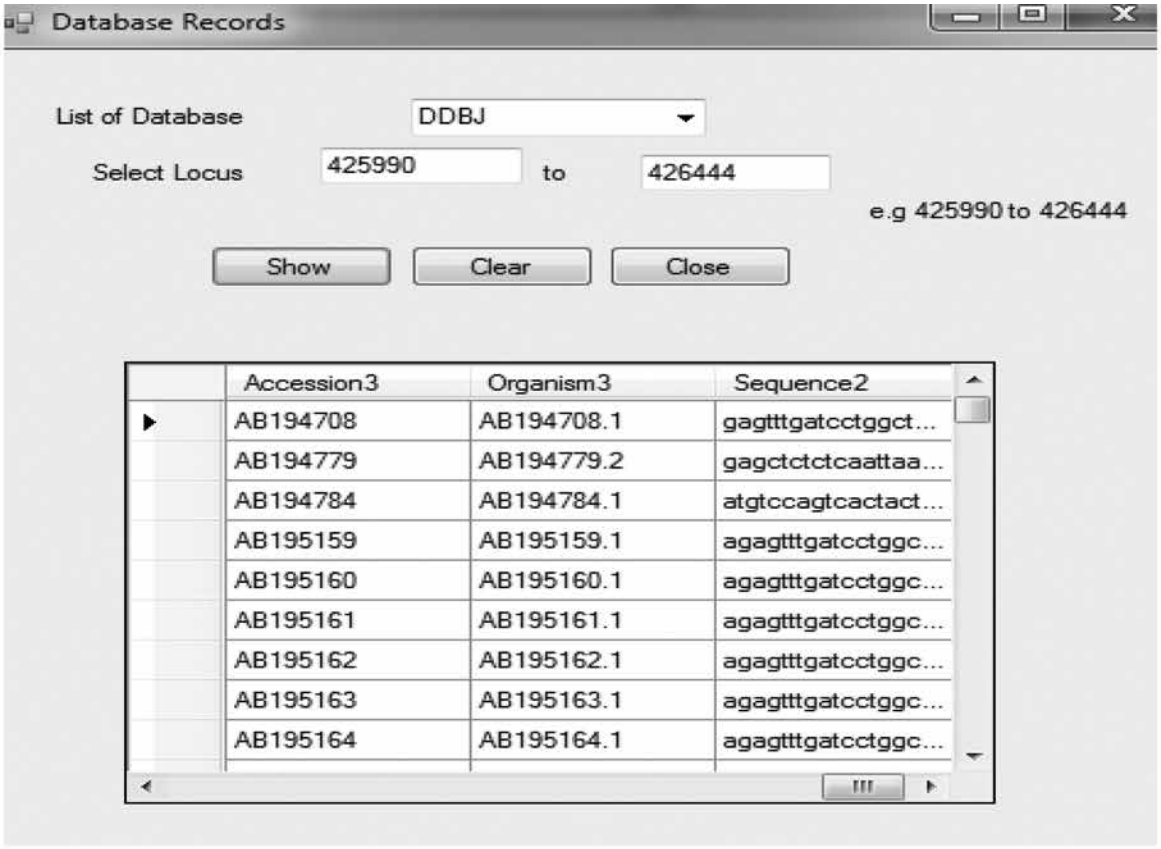

Fig. 5. Sample screen of the developed software.

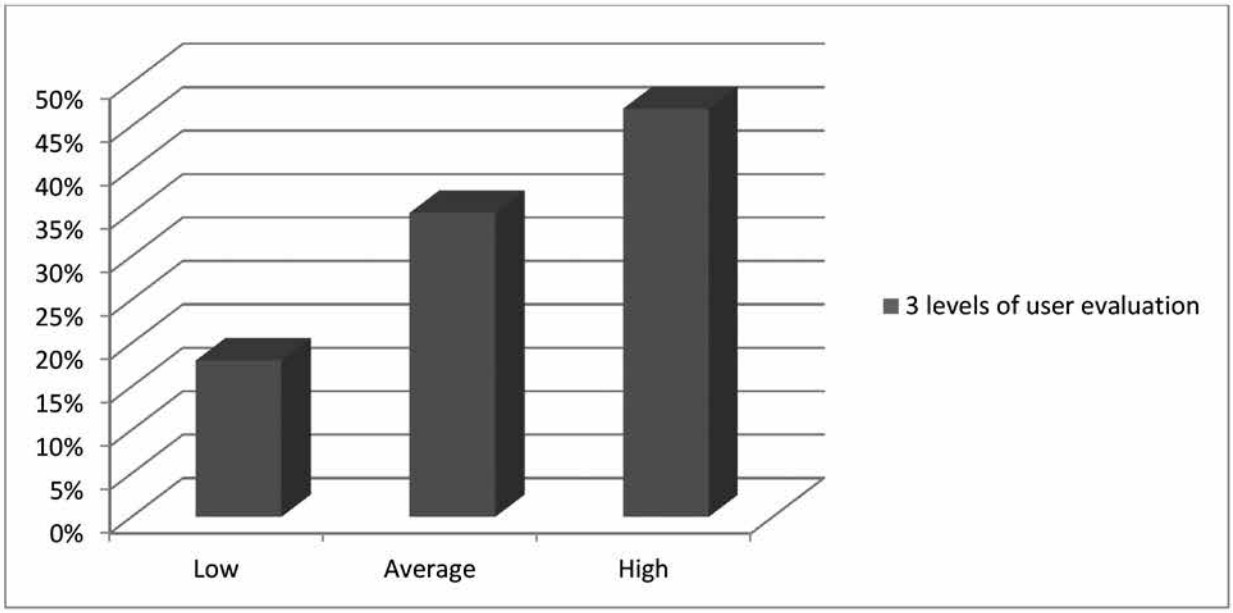

Fig. 6. Evaluation of information retrieval system based on various queries provided by the user.

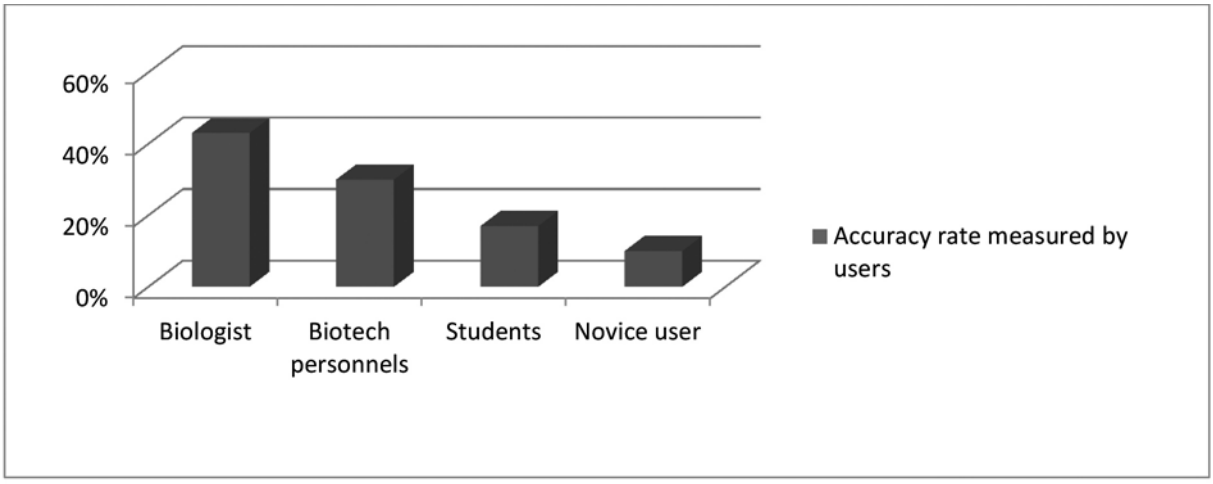

Fig. 7. Role-based evaluation.

3.2.1 Task Based Evaluation

A group of testers used this tool according to their role some and evaluate the performance of this tool as showed in Fig 6.

3.2.2 Relevance of Searching Criteria

This evaluation was performed to find the relevance of retrieved information. Users were given a questionnaire with the following three questions.

i. Retrieved information is highly relevant.
ii. Retrieved information is relevant.
iii. Retrieved information is not relevant.

In this type of evaluation the searching result was tested. There are certain levels of relevance such as highly similar, average and low showed the degree of searching result as per desire and searching criteria. Fig.7 shows the overall degree of relevance of search result.

To tackle certain issues of heterogeneity, we proposed a unified format, which holds all the fields of heterogeneous format, which is enough adept to accommodate data of all file formats.

There are many heterogeneous biological databanks and sequence alignment tools such as NCBI, DDBJ and BLAST, respectively. The major drawback in these databanks is that these are providing heterogonous biological data which is a major issue in integration and semantic search. These databanks are confined to specific organism sequences only [6, 13]. Furthermore, BLAST analysis has become a ubiquitous method of interrogating new sequence data, but these are the major limitation to using BLAST alone as a discriminating tool and its output is often skewed [16].

This project will basically deal with the establishment of a biological databank for the purpose of serving the community because in Pakistan such type of databank does not exist as yet. The need of this project urged due to some of the factors, i.e., economic stability, employment opportunity for undergrads and graduates, distance learning, agricultural and forensic science development. Setting up a new laboratory with equipment and trained staff for research purpose is highly costly job, by using this databank scientist can take access biological data efficiently and effectively. Newbies in bioinformatics and biotechnology can use this databank for study and research work. Even some pharmaceutical companies can sue it for drug discovery and their experimental work. This project can also help in detection of dengue, polio and other viruses which directly attacks human DNA structure [17– 20].

4. CONCLUSIONS

Globally, biological information is available in

segregated form with heterogeneous flavor and in Pakistan a unified biological databank is not available. Heterogeneity nature of biological data is a main hindrance towards the semantic searching and integration of various biological data sets. To address these issues, a unified format might be proposed which holds all the fields of heterogeneous format, which is enough adept to accommodate data of all file formats. As an implementation we provide unified biological databank which is composed of four processing machine e.g. download server, data server, reports server and biological data processing engine. Firstly, heterogeneous data is acquired and stored on data server, which is further provided to the engine to convert it into unified form. This unified data is stored on data server which communicates with the reports server to generate reports on the behalf of user's input. To make available this unified biological databank an interface is provided which receives the input from user and displays the result.

5. REFERENCES

1. Coral, D.V., H.G. Karl & S. Sandor. cDNA2Genome: A tool for mapping and annotation cDNAs. *BMC Bioinformatics* 4(1): 1471-2105 (2003).
2. Min-Huang, H., Y.S. Chang, M.C. Cheng, L. Kuang-Li & S.M. Yuan. *A Unified, Adjustable, and Extractable Biological Data Mining-Broker* (Springer) 2690(1): 773-777 (2003).
3. Kosuge, T., M. Jun, K. Yuichi, F. Takatomo, K. Eli, O. Osama, K. Okubo, T. Takagi & N. Yasukaza. DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Research* 42(D1): 44-49 (2014).
4. Wollbrett, J., P. Larmande, F.D. Lamotte & M. Ruiz. Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases. *BMC Bioinformatics* 14(1): 14-126 (2013).
5. Amir, E.D., K.L. Davis, M.D. Tadmore, F.S. Erin, H.L. Jacob, C.B. Sean, K.S. Daniel, K. Smita, P.N. Garry & P. Dana. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology* 31(6): 545-554 (2013).
6. Cao, S.L., L.H. Qin, W.Z. Zhong, Y.Y. Zhu & Y.X. Li. Semantic Search among Heterogeneous Biological Databases Based on Gene Ontology. *Acta Biochimica et Biophysica Sinica* 36(5): 365–370 (2004).
7. Alison, C., T.J. Cruz & M. Dumontier. Ontology-based querying with Bio2RDF's linked open data. *Journal of Biomedical Semantics* 4 (Suppl 1): 115-127 (2013).
8. David, S.W. J. Timothy, A.C. Guuo, M. Wilson, C. Knox, Y. Liu, D. Yanninck, R. Mandal, F. Aziat, E. Dong, S. Bouatara, S. Igor, A. David, J. Xia, P. Liu, F. Yallou, B. Trent, P.P. Rolando, R. Eisner, F. Allen, V. Neveu, G. Russ & S. Augustin. HMDB 3.0 — the human metabolome database in 2013. *Nucleic Acids Research* 1065: 801-807 (2013).
9. Rose, P.W., B. Bojan, C. Bi, W.F. Bluhm, D. Dimitris, S.G. David, A. Prlic, Q. Martha, G.B. Quinn, D.W. John, Y. Jasmene, Y. Benjamin, Z. Christine, M.B. Helen & E.B. Philip. The RCSB protein data bank: Redesigned web site and web services. *Nucleic Acids Research* 39 (Suppl 1): 392-401 (2011).
10. David, B.K. DNA Microarrays in the undergraduate microbiology lab: Experimentation and handling large datasets in as few as six weeks. *Journal of Microbiology & Biology Education* 8.1:3-12 (2007).
11. Dennis, A., I. Benson, M. Karsch, J. David, J.O. Lipman & L.W. David. GenBank. *Nucleic Acids Research* 41.D1: 36–48 (2013).
12. Walker, G.K. & K. Gordon. Responding to hypertext transfer protocol (http) requests. *U.S. Patent Application* 13/838,744.
13. Block, M., M. Strenge, C. Mohr, G. Boris & F. Franz. Integrated Application Server and Data Server Processes with Matching Data Formats. *U.S. Patent Application:* 13/919,921.
14. Chen, B.P. A quick sorting algorithm adaptive to massive data with high repetition rate. *Advanced Materials Research* 834: 1002-1005 (2014).
15. Shrestha, R.K., B. Lubinsky, V.B. Bansode, M.B. Moinz, G.P. McCormack & S.A. Travers, QTrim: A novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics* 15(1): 1-6 (2014).
16. David, A., S.A. Garry & R. Jacques. Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* 6(2): 1-7 (2005).
17. Shah, A. & S. Ahsan. Problems and issues in managing biological data. In: *International Symposium on Bio-Inspired Computing 2005 (BIC 05)*, Johor, Malaysia. Sept 6-10, 2005 (2005).
18. Idrees, M. & M. Khan. SMGCD: Metrics for biological sequence data. *Nucleus* 51(1): 125-131(2014).
19. Idrees, M., M. Khan & A. Shah. Unified Data Model for Biological Data. *Mehran University Research Journal of Engineering & Technology* 3(3): 261-277 (2014).
20. Ghani, M.U., P.I. Khan, K.H. Asif, A. Nasir, M.J. Arshad & S. Amanat. An Agent-based CBIR system for medical images. *Journal of Faculty of Engineering & Technology* 21(2): 39-45 (2014).