ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 15, Issue, 01, pp.23206-23210, January, 2023
DOI: https://doi.org/10.24941/ijcr.44531.01.2023

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

PRESENT SCENARIO OF DATA INTEGRATION IN BIO-INFORMATIC: DATA WAREHOUSING

***Rabindra Kumar Mishra[1], Ankit Kumar Jena[2] and Sanjay Kumar Dey[3]**

[1]Department of Basic Science & Humanity, GIET University, Gunupur, Rayagada, Odisha, India 765022
[2,3]Department of Biotechnology, GIET University Gunupur, Rayagada, Odisha, India 765022

ARTICLE INFO

Article History:
Received 16th October, 2022
Received in revised form
19th November, 2022
Accepted 15th December, 2022
Published online 20th January, 2023

Key words:
Atlas, ontology, API, Biowarehouse, BIOZON, COLUMBA , VINEdb.

****Corresponding Author:***
Rabindra Kumar Mishra

ABSTRACT

The biological data warehouse is shown here, and it is stored locally. It provides the most comprehensive forum for (a) biological sequence integration, (b) interactions among molecules, (c) Understanding of homology, (d) annotations for gene sequence, and (e) biological ontologies. For bioinformatics research and development, this framework provides both data and application infrastructure. This study defines an internet frame of reference for building database warehouses that incorporate multiple gatherings of bioinformatics datasets into a particular database managing model. A description of Atlas, Biowarehouse, BIOZON, COLUMBA, and VINE dBs to the data warehouse design has been given to validate t(DBMS), allowing queries to span multiple database servers. This paper is based on data extraction and Integration from varied sources and alternative proposals for processing the consolidated data, data warehousing, and integrating data into information. Data source as well as the architecture of a biological data warehouse proposition.

Citation: ***Rabindra Kumar Mishra, Ankit Kumar Jena and Sanjay Kumar Dey.*** **2023.** "Present scenario of data integration in bio-informatic: data warehousing*". International Journal of Current Research*, 15, (01), 23206-23210

INTRODUCTION

Data integration has been proposed using a variety of approaches, which are five groups in total: data warehousing, federated online databases, Integration based on services, conceptual Integration, and wiki-based Integration (Bilal Ben Mahria, 2021). One of the most significant bioinformatics resources is the data warehouse. This is subject-specific, structured, and non-volatile, making high-level analysis easier (Bilal Ben Mahria, 2021) Information is gathered, and data is mined for new information. Since the data warehouse is updated regularly, any data written to it will be lost during the upgrade. Data warehouses have some benefits, but they also have several drawbacks: they are costly to build, integrating changes and revisions from data sources into the warehouse can be challenging and time-consuming, and they are challenging to maintain (Prashila Dullabh, 2020). Data warehouses concentrate on data transformation, obtaining available data from a multiplicity of different sources, transforming it, and bring it in into the data warehouse. Atlas, Biowarehouse, BIOZON, COLUMBA, and VINEdb are examples of data warehouses (Sarinder K. Dhillon, 2019).

Atlas: Atlas is a robust, adaptable data warehouse that offers both data and application infrastructure platforms for bioinformatics application and analysis. The Atlas framework is needed on an interpersonal information model merged into a single entity using a SQL language in a series of program Interfaces (Sarinder K.Dhillon, 2019). Atlas program is easily accessible through http://bioinformatics.ubc.ca/.ubc.ca/ atlas/.There are five major parts of the Atlas system 1) the source data, 2) the ontology system, 3) the relational data models, 4) the APIs 5) the applications.(5)The data sources are divided into four categories: 1)' sequence,' 2)'molecular interactions,' 3)' gene-related resources,' and 4)'ontology.' Table 1 lists the sources and URL being used Atlas (Thomas Triplet, 2014)

Relational data models: The configuration of the statistical representations of the source data involved in Atlas is described by relational statistics models. MySQL, an internet relational folder managing model, is used to implement the data models described here (Galperin, 2012).

Ontology: Ontologies are divided into two categories: Atlas-defined ontologies and external ontologies(8). Ontologies that remain implemented to describe the notion andcorrelation create directly contained by Atlas, as well as those indirectly identified by the Gen Bank Sequence(7-8). Atlas specified ontologies are feature data models. External ontologies include the Proteomics Standards Initiative Molecular Interaction Standard measured terminology, NCBI Taxonomy for species categorization, GO for gene clarification, and the PSI-MI for gene annotation, and others described in the below table(7-8). There are three tables in this part of the Atlas ontology: One that describes ontology source and category, as well as terms and meanings(8). An ontology that keeps track of long-term collaborations. Ensure data integrity; international key restrictions are used. In comparison to these closely integrated ontologies, two other external frames of reference, GO and NCBI Taxonomy, are generated as separate MySQL databases. Foreign keys are not implemented in these ontologies. As a consequence, when ontology relations are modified, citations to removed relations that are considered ineffective remain in the system until the entire data set is reloaded (Thomas, 2006). A broad list of the ontologies are obtainable through http://bioinformatics.ubc.ca/atlas/ontology/.

Table 1. Atlas Data resource

Resource	URL
Uni Prot	ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/
H P R D	http://www.hprd.org/download/
MINT	http://mint.bio.uniroma2.it/mint/
DIP	http://dip.doe-mbi.ucla.edu/dip/Download.cgi
Gen Bank Seq.	ftp://ftp.ncbi.nih.gov/ncbi-asn1/
Gen Bank Ref Seq	ftp://ftp.ncbi.nih.gov/refseq/
Gen Bank Seq.	ftp://ftp.ncbi.nih.gov/ncbi-asn1/
Gen Bank Ref Seq	ftp://ftp.ncbi.nih.gov/refseq/
NCBI Taxonomy	ftp://ftp.ncbi.nih.gov/pub/taxonomy/
BIND	ftp://ftp.blueprint.org/pub/BIND/current/bindflatfiles/bindindex/
G O	http://www.godatabase.org/dev/database/archive/latest/
Homolo Gene	ftp://ftp.ncbi.nih.gov/pub/HomoloGene/
O M I M	ftp://ftp.ncbi.nih.gov/repository/OMIM/
Gene	ftp://ftp.ncbi.nih.gov/gene/
Locus Link	ftp://ftp.ncbi.nih.gov/refseq/LocusLink/

Application programming interfaces: Loader and retrieval are two types of APIs. Shown in Fig1.The Molecular Connections element hasa collection of loader APIs for which we have established our relational models (Aaron Birkland and Golan Yona, 2006) The loader APIs inhabit requests of connection representations in the Atlas database and are used to construct loading applications (Birkland, 2006) The retrieval APIs are used to retrieve data from Atlas. They're needed for creating practice cover requests like the Atlas toolbox apps. The SeqLoad and Seqloader modules are closely linked to the NCBI C++ Toolkit; they are only written in C++. Other groups can be found in Java. A popular class is responsible for lower-level data transfiguration inmutually the data loader and the tools for recovery (Birkland, 2006; Sohrab, 2005) This class contains methods for converting internal Atlas identifiers to externally referenced public identifiers, such as GI numbers and bio id to ontology id. Both the APIs profit from inheriting a standard identifier conversion class since it gives them the resources they need to mix data (Silke, 2005). The Biological Sequences part of Atlas achieves the Seq class's universal identifiers and hash maps. Both the Seqloader and SeqGet classes inherit this class that specifies the loader and retrieval approaches jointly (Benson, 2004). The potential to monitor a flow production depending on the form of the molecule is another function of the Biological Sequences API. API users simply call higher-level retrieval methods to determine which molecule type to screening and SeqGet can handle the stream management logistics as shown in Fig. 1. Since all of the basic programs are published under the General Public License and any designer may use it to model future API development (Benson, 2004; Rother, 2004).

Applications: The Atlas toolbox is a gathering of uses that usually carry out sequence and feature retrieval errands using the C++ API. For parameter entry, the applications use a command-line interface. They're all essential UNIX command-line utilities. The application and its function are explained in table 2. The source code for the toolbox applications also includes strong examples of applications created using the APIs. These toolbox apps can also be used by software developers as a point of departure for a personal application by using the APIs.

Bio Warehouse: Bio Warehouse is a free and open-source application for incorporating the asset of life science database into a particular objective database executive scheme for information processing, exploration, and scooping. Shown in Table 3.Bio warehouse, in some cases, can be built with either the Oracle or MySQL rules. There are two ways to use Bio Warehouse: (i) Through an Internet SQL query, the user can makean inquiry about two community Bio Warehouse servers controlled by SRI International, Public House, and Ecoli House. (ii) Handlers can also proceed with the Bio Warehouse software supply and establish the Bio Warehouse illustration with the section of sustained Bio Warehouse DBs that they want (Apweiler, 2004) This technique grants entry into databases that SRI is unable to reconstruct. When new DB versions are loaded, it gives each user power. Users may also make a broad hardware layout to the Bio Warehouse request and attach confidential data to the Bio Warehouse request (Galperin, 2004). The Bio Warehouse is filled with loader programs that convert a source database's flat-file representation into the warehouse schema. Each source database supported by Bio Warehouse has its loader (Galperin, 2004) Run the application once it has been loaded into a Bio Warehouse case as shown in Table3.A number of loaders are designed to work with a Particular System Rather Than An Individual Database.

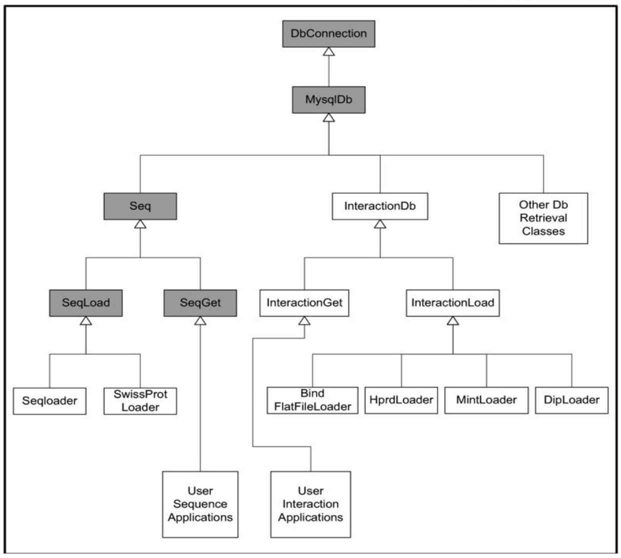

Fig. 1. API Architecture

Biozone: Biozone offers comprehensive information on different biological data. About 100 million records and 6.5 billion connections records. The database can be accessed through a progressive network through http://biozon.org (Hristidis, 2002). Table 4 lists the various type of documents available in Biozone and its Association type listed in Table 5. We establish a filecategorization thatcorrelates toother fields ofinformation to describe the biological meaning of documents (18). Every document has been arranged ata number oflevels based on its context or source, and each document has been grouped at a fewlevels based on its significance, source, and its ideas (Hristidis, 2002)

Columba: The COLUMBA database makes it possible to create protein structure data sets for a variety of structural studies. Combining responses on a variety of structural databases that are not currently protected by other reforms can be done here (Wilkinson, 2002). Allowing information from both a large and small number of protein structures to be used effectively PDB, KEGG, Swiss-Prot, CATH, SCOP, Gene Ontology, and ENZYME are among the twelve databases that COLUMBA is currently incorporating. Keyword searches or data source-specific online types may be used to search the database. For several structure-based studies, the COLUMBA database makes it easier to construct protein structure data sets.(19) It enables the combination of querying on a variety of structures that are not currently protected by means of additional biological research. As a result, in sequence together with a huge amount of protein structures and a bounded number of protein structures can be used effectively. http://www.columba-db.de. COLUMBA is built on the Postgre SQL DB system. As shown in Table 6, it presently integrates data from twelve different databases. The source data is available in many formats. Toinhabit COLUMBA through an inapplicable presentation, we exercise parsers written in Python and Perl, and respectively (Wheeler, 2000). We use our parser for PDB, which was derived from the Bio Python project.

Table 2. Atlas toolbox applications

Application	Purpose	Enter	Output
gi2seqentry	Retrieve sequences given a GenInfo identifier	GI Numbers	GBFF, EMBL, GFF, TABLE, ASN.1, GBSEQ
ac2seq	Retrieve sequences given an accession	DNA&Protein Accession Nos	FASTA format
feat2seq	Retrieve sub-sequences that span characteristics	Feature type and qualifier	FASTAformat
gi2seq	Retrieve sequences given a GenInfo identifier	GI Numbers	FASTA format
tax2seq	Retrieve sequences by taxonomy	The scientific name of a taxon	FASTA format
tech2seq	Retrieve sequences by sequencing method	Sequencing method	FASTA format
techtax2seq	Recover sequences based on taxonomic classification and sequencing method.	Sequencing method and NCBI taxoid/Taxo	FASTA format
Taxonomy			
ac2tax	Retrieve taxonomy given an accession number	GenBank Accession number (string)	NCBI taxon identifier
gi2tax	Retrieve taxonomy given a GI identifier	GI identifier	NCBI taxon identifier
tax2gi	Retrieve GenInfo identifiers held by taxon identifier	NCBI taxon identifier	GI identifier
Loader			
Fastaloader	FASTA sequence data loader	Sequences in FASTA format	
Seqloader	ASN.1 sequence data loader	GenBank/RefSeq	ASN.1
Feature			
ac2feat	Retrieve type	GenBank Accession numbers	An attribute in GFF or Table format
gi2feat	Retrieve type	GI No.	An attribute in GFF or Table format
ID Converters			
ac2gi	Convert an accession No. to a GI identifier	GenBank Accession number (string)	GI identifier
gi2ac	Convert a GI identifier to an accession No.	GI identifier	Accession No. (string)

Table 3. Bio Warehouse loaders designed

resource DB	Type of information
BioCyc	Genomes, genes, proteins, metabolic pathways, reactions, compounds
Comprehensive Microbial Resource	Genomes, genes, proteins, reactions
Kegg	metabolic pathways, reactions, compounds, Genomes, genes, proteins
MetaCyc. Ontology	The MetaCyc ontology of metabolic pathways.
NCBI Taxonomy	classification of a taxonomical organism.
BioPAX format	The BioPAX format is used to store information about biological pathways and protein-protein interaction. This loader can currently only process BioPAX Level 2 data (protein interactions).
EMP DB	Reactions, proteins
Eco2dbase	E. coli 2D protein gel database
GenBank – bacteria only	proteins and bacterial genes
Gene Ontology	The standardized language for relating to genes and gene annotation characteristics
MAGE-ML format	Gene expression datasets are represented in theMAGE-ML file format.
Swiss-Prot and TrEMBL	Protein knowledge ebase

Table 4. Types of sequence and documents

Document Type	sequence Type	*Atom Type*
protein sequence	String	amino acids
nucleic acid sequence	String	nucleic acids
protein family	Set	Proteins
Pathway	Set	protein families
unigene cluster	Set	nucleic acids (ESTs)
domain family	Set	Domains
Interaction	Set	proteins, nucleic acids
Descriptor	Text	Characters
Structure	List	3D coordinates
Domain	ordered pair	sequence coordinates

Table 5. Particular category of relations in the Biozone database

Association type	Indicating document	Specified document
Similarity	Protein	Protein
Manifests	Protein	Structure
encodes. Nucleic	nucleic acid	Protein
encodes. Unigene	unigene cluster	Protein
contains. Interaction	Interaction	protein, DNA
contains. enzyme-family	enzyme family	Protein
comprises. Domains	Domain	Protein
describes. Go	go term	Protein
hierarchy. Go	go term	go term
contains. Unigene	unigene cluster	nucleic acid
contains. Pathway	Pathway	enzyme family
Describes	Descriptor	any object
contains. domain-family	domain family	Domain
expresses. Unigene	unigene cluster	Tissue

Table 6. COLUMBA Design

Resource	URL	Analyzed by
Boehringer	http://us.expasy.org/tools/pathways	personal
K E G G	http://www.genome.jp/kegg	Personal
P D B	http://www.rcsb.org/pdb	Bio Python
S C O P	http://scop.berkeley.edu	Bio Python
C A T H	http://www.biochem.ucl.ac.uk/bsm/cath	Personal
D S S P	Computed	Personal
ENZYME	http://us.expasy.org/enzyme	Bio Python
Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy	bioSQL
Swiss-Prot	http://www.expasy.org/sprot	bioSQL
G O	http://www.geneontology.org	bioSQL
G O A	http://www.ebi.ac.uk/GOA	Different model
PISCES	http://dunbrack.fccc.edu/PISCES.php	Personal

We use the BioSQL project's parsers and schema to transmit Swiss-Prot, Gene Ontology Similarly, other analyzed in table 6.

Vinedb: This data warehouse was created to work with and analyze combined life science data. The web application and basic infrastructure are platform-independent according to agrowing open-source data warehouse architecture. A simulation component is also included in the system, enabling communal graphical searching of the included data (Berman, 2000; Ashburner et al., 20002) VINEdb can be found at http://tunicata.techfak.unit. The simulation approach is significant because it is user understandably and established a good connection between the data and the patron.

Conclusion

We created a biological data warehouse to provide high-throughput, scalable data access via SQL, API-level queries, and last user appliance queries. Bio Warehouse is made up of a worldwide communication schema by a collection of loader functions that resolves bioinformatics databases and uploads their information into that schema. Users can download and install the toolkit. SQL can be used to retrieve previously disparate data that has now been centralized in a relational model. The Atlas architecture's ability to integrate data at two levels is one of its main advantages. The first level employs a standard data model to combine data from various sources that are identical. The APIs, ontologies, and methods used at the second level are used to link different data types. The Bio Warehouse is filled with loader programs that convert a source database's flat-file representation into the warehouse schema. Each source database supported by Bio Warehouse has its loader. Run the application once it has been loaded into a Bio Warehouse case. Few loaders are designed to work with a particular format rather than a sole database. The COLUMBA database makes it possible to create protein structure data sets for a variety of structural studies. Combining responses on a variety of structural databases that are not currently protected by other reforms can be done here. A simulation component is also included in the system, enabling interactive graphical exploration of the integrated data through VINEdb.

Acknowledgements

We would like to show our gratitude to the Dr. N.V.J. Rao, Registrar GIET University, Gunupur, Rayagada, Odisha 765022 for sharing their pearls of wisdom with us during the course of this research.

REFERENCES

Bilal Ben Mahria*, Ilham Chaker and Azeddine Zahi Ben Mahria et al. J. (2021). A novel approach for learning ontology from a relational database: from the construction to the evaluation Big Data, 8:25,https://doi.org/10.1186/s40537-021-00412-2

Prashila Dullabh, Lauren Hovey, Krysta Heaney-Huls, Nithya Rajendran1 & Adam Wright. Published online: 22.01.2020. 59, Applied Clinical Informatics. Vol. 11 No. 1/2020, DOI https://doi.org/ 10.1055/s-0039-1701001. ISSN 1869-0327.

Sarinder K. Dhillon. (2019) in Encyclopedia of Bioinformatics and Computational Biology,

Sarinder K. Dhillon. (2019). Encyclopedia of Bioinformatics and Computational Biology Volume 2, Pages 96-117,Biological Databases Author links open overlay panel.

Dermeval D, Vilela J, Bittencourt II, Castro J, Isotani S, Brito P, Silva A. (2016). Applications of ontologies in requirements engineering: a systematic review of the literature. Requirements Eng. 21:405–37. 2.

Thomas Triplet, Gregory Butler. (July 2014. A review of genomic data warehousing systems, Briefings in Bioinformatics, Volume 15, Issue 4, Pages 471–483, https://doi.org/10.1093/bib/bbt031.

Galperin MY, Ferna´ndez-Sua´rez XM.(2012) nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res 40:D1–8.

Triplet T, Butler G. (2011). Systems biology warehousing: challenges and strategies toward effective data integration.In: 3rd International Conference on Advances in Databases, Knowledge, and Data Applications. IARIA, 34–40.

Thomas J Lee1, Yannick Pouliot, Valerie Wagner, Priyanka Gupta, David WJ Stringer-Calvert, Jessica D Tenenbaum and Peter D Karp. (2006). BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics, 7:170 doi:10.1186/1471-2105-7-170,PP.1-14.

Aaron Birkland and Golan Yona*. (BMC Bioinformatics 2006).BIOZON: a system for unification, management, and analysis of heterogeneous biological data, 7:70 doi:10.1186/1471-2105-7-70, PP.1-24.

Birkland A, Yona G. (BMC Bioinformatics 2006). BIOZON: a system for unification, management, and analysis of heterogeneous biological data. 7:70.

Sohrab P Shah, Yong Huang, Tao Xu, Macaire MS Yuen, John Ling &BF Francis Ouellette, *BMC Bioinformatics* volume 6, Article number: 34 (2005), Atlas – a data warehouse for integrative bioinformatics, DOI: 10.1186/1471-2105-6-34,PP.1-16.

Silke Trißl, Kristian Rother, Heiko Müller, Thomas Steinke, Ina Koch, Robert Preissner, Cornelius Frömme and Ulf Leser. (BMC Bioinformatics 2005). Columba: an integrated database of proteins, structures, and annotations, 6:81 doi:10.1186/1471-2105-6-81,PP.1-11.

Benson D, Karsch-Mizrachi I, Lipman D, Ostell J & Wheeler D. (2004). GenBank: update. *Nucleic Acids*, (32 Database):D23–26. 10.1093/nar/gkh045

Rother K, Müller H, Trissl S, Koch I, Steinke T, Preissner R, Frömmel C & Leser U. (2004). Columba: Multidimensional Data Integration of Protein Annotations. In DILS, Volume 2994 of Lecture Notes in Computer Science Edited by: Rahm E. Springer; 156-171.

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 L. (2004). UniProt: the

Universal Protein knowledgebase. *Nucleic Acids*, (32 Database):115–119. 10.1093/nar/gkh131

Galperin MY. (2004). The molecular biology database collection: update. Nuc Acids Res 2004, 32:D3-22

Hristidis V, Papakonstantinou Y & Discover. (2002). Keyword search in relational databases. VLDB.

Wilkinson MD, Links M & BioMOBY. (2002). An Open Source Biological Web Services Proposal. Briefings in Bioinformatics , 3(4):331-341.

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA & Rapp BA. (2000). Database resources of the National Center for Biotechnology Information. Nuc Acids, 28(1):10-14.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE. (2000). The Protein Data Bank. Nucleic Acids, 28:235-242.

Ashburner M, Ball CA, Blake JA & et al. (2000). Gene ontology: a tool for the unification of biology. Nat Genet 25: 25–29.

****