Molecular biological databases: evolutionary history, data modeling, implementation and ethical background

Dr. M. P. N. S. Cooray MBBS, MSc (BMI)
Medical Officer (Health Informatics), National Institute of Health Sciences, Kalutara, Sri Lanka
E-Mail address: mpnscooray@yahoo.com

Sri Lanka Journal of Bio-Medical informatics 2012;**3**(1):2-11
DOI: http://dx.doi.org/10.4038/sljbmi.v3i1.2489

Abstract

Influence from the field of computer and information sciences in early 1970s was one of the major factors leading to the development of computer based repositories for biological data. Since the emergence of the first ever computer based molecular biological database 'Protein Data Bank' in 1971, biological database domain has grown rapidly in terms of information content and volume, database modeling, implementation and integration. The key driving forces of this growth are the amount and diversity of data generated from advancing biological research technologies and the challenges imposed on data modeling by the inherent properties of biological data and the concept of 'Electronic Data Publishing' introduced in early 1990s.

Due to the data modeling challenges driven by inherent properties of biological data, multiple modeling tools such as 'Enhanced Entity Relationship' diagrams and 'Unified Modeling Language' are used to model biological mini worlds. In order to be compatible with diverse data models, multiple implementation approaches are also used by the biological database developers, viz: 'flat files', 'XML', 'relational databases', 'object oriented databases', 'ASN.1'. Adherence to simplicity and conservative technology provides a better approach to biological database modeling and implementation.

Even though the biological databases differ in their internal data model, implementation approach and biological scope, almost all of them share similar three tier web architecture. This similarity is the basis of the current three major strategies used for database integration viz: 'Link/hypertext integration', 'View integration' and 'Data warehousing'.

Keywords - Protein Data Bank; Electronic Data Publishing; GenBank; Biological Databases

Emergence and evolution of computer based molecular biological databases

The "Atlas of protein sequence and structure" compiled by the late Margaret Dayhoff in 1965 is the first published molecular biological information content which resembles the features of a modern day biological database[(1)].

With the rapid growth of computer science and information science in the early 1970s, biological scientists were influenced to create freely accessible computer based repositories for biological data. This trend was initiated in 1971 by the development of a repository for protein structure data at the Brookhaven National Laboratory and were stored in laboratory notebooks and punch cards[(2)]. This repository was named as the Protein Data Bank (PDB) which is the leading database of protein 3D structure now. In 1982, a decade since the start of the first computer based molecular biological database, PDB, two major nucleotide sequence databases were developed simultaneously. While United States National Institute of Health (NIH) initiated the GenBank project, the European Molecular Biology Laboratory (EMBL) started establishing its own sequence data bank. Within a couple of years GenBank and EMBL started collaborating. By

the mid 1980s, the DNA Data Bank of Japan (DDBJ) also joined in to form the International Nucleotide Sequence Database Collaboration (INSDC)[3].

With the emergence of the concept of Electronic Data Publishing[4] in early 1990s the scope of biological databases started to grow into fields of data visualisation and data publishing. They adopted the role of a new kind of primary literature where findings are submitted directly to the database[4].

Traditional biological publications were based on information and knowledge derived from experimental data, but with the advancement of experimental research many findings which are close to the data end of the spectrum was generated. Sharing an enormous amount of such data generated through current technologies like ultra high throughput sequencing is impossible without the support of a publishing database which facilitates electronic data publishing[4,5].

The best example to describe above evolutionary changes in molecular biological databases is the changes adopted by GenBank, to facilitate the storage and publishing of sequence data generated by the Human Genome Project[4]. This milestone marks the development in databases of post genomic era. Currently the number of base pairs in GenBank doubles about every 18 months and includes approximately 110 million sequences and 200 billion base pairs[3].

Therefore, the demand from the biological research community for electronic data publication and the load of data generated by molecular biological research were the key driving forces of current growth and evolutionary changes in design and implementation architecture of biological databases.

Database issue of Nucleic Acids Research (NAR) in 2010 includes descriptions of 58 new data resources and updates 73 previously published data resources. The online Database Collection that accompanies the issue holds a total of 1230 data resources[6] which represents a 5% growth in the number of biological databases during the period from 2009 - 2010[6].

The above growth and evolution of biological databases which relates to the data and information content, data organisation (model), accessibility and interoperability are critical factors for future research advances in molecular biology, genomic medicine, pharmacogenomics and bioinformatics[7].

Properties of biological data

The data driven growth and evolution of biological databases are not merely due to the rapid rate of data generated by advanced biological research, but greatly due to the challenges presented to the database developers by the inherent properties of biological data and nature of data users etc[7].

It is important to understand these properties of biological data ranging from research articles to complex metabolic pathways[7] before developing solutions for any biological database problem. High level complexity[7,8] of biological data compared to most of the other database domains presents a challenge in modeling data substructures and their relationships. Inaccurate modeling could lead to loss of information leading to total failure of the design.

Variability of amount and range[7,8] of data presented in the biological world forces biological data models to adopt flexibility to data types and value ranges to include 'outlier values' thus preventing information loss. Sometimes it is necessary to use multiple data types to represent a single piece of data according to the context of use.

Schema changes with time[7,8] and needs recursive designing process and re-release of databases with updates.

The above discussed challenges are imposed by the inherent properties of biological data itself and the following properties represent the challenge created by the biological data users.

Requirement of representing the same data in different ways[7,8] leads to building of multiple conceptual schemas for a single physical schema.

Limited requirement of access to the database[7,8] enforces the development of strong security models.

Preference to stay away from technical aspects[7,8] of the database by its users impose challenges in developing Graphical User Interfaces (GUI), pre-defining standard queries and development of tools for user query building.

Other challenges arise from the analytical context of data where designers need to model meta data[8] for data analysis and building of archival capabilities[7,8] for data validation and analytical purposes. Challenges imposed by the properties of biological data lead to development in novel techniques in data modeling and database research[8].

Tools of modeling

Even though the inherent complexity of biological data is proven to be challenging, precise understanding of the mini world (Domain under consideration) and accurate modeling are key to successful and sustainable database building. Being simple and conservative[9] provides a practical approach to biological data modeling.

Entity Relationship (ER) based modeling

Since its introduction[10] ER modeling is very popular in database community for its ability in modeling high level conceptual schemas[8] (implementation independent model). ER models are better suited to model well defined entities with simple relationships[11]. Due to the rarity of such entities in the molecular biological world, ER modeling is hardly used alone to model such domains.

Entity Category Relationship (ECR) model[12] introduced in 1985, opens the path for the development of Enhanced/Extended Entity Relationship (EER) model[8]. EER modeling is capable of capturing specialisation and generalisation relationships in the biological world. These EER models work as useful tools to map high level conceptual schemas to relational database

schemas[8]. This capability is built in to most of the common Computer Aided Software Engineering (CASE) tools which supports efficient database building.

Even though the EER models are capable of capturing simple molecular biological relationships, to model constructs such as ordered relationships, functional processes and 3D structures which are common to molecular world, an extension to EER[11] model was introduced in 2007. This model is capable of modeling higher level abstractions such as cells, tissues, organs and biological systems without compromising its ability to act as a relatively easy mapping tool to build relational database schemas.

Unified Modeling Language[14] (UML)

UML was developed in 1967 by simplification and collaboration of major object oriented development methods developed since the development of first object oriented language Simula-67. This product of Rational Software Corporation was unanimously adopted in 1997 by the membership of Object Management Group (OMG).

UML is a general purpose visual modeling language that captures information about the static structure and dynamic behaviour of a system which is ideal to model molecular biological scenarios. It provides a set of tools to model implementation, independent (conceptual) schemas for multiple object oriented languages. This enables CASE tool mapping of UML models to software components and reverse engineering of current module to UML models. This added software support and ability of modeling molecular biological data, makes UML a highly recommended tool among biological database developers[9,14].

UML provides conceptual schemas which can be accurately matched to object database, relational database and object-relational database schemas.

Structured Query Language – Data definition Language (SQL – DDL)

This is a SQL description of a relational database table structure which can be used as principle data model description as a master of what a database stores[9].

eXtensible Markup Language – Data Tag Description (XML – DTD)

This can also be used as principle data model descriptor, apart from SQL – DDL or high level UML views[9].

The important practical aspect is to firmly adhere to a single master data model to prevent branching towards modeling ‘all of biology’ in a single database design which can lead to instability in the database structure[9].

Implementation approaches

After having a stable primary data model of the mini world, next most important decision is about the implementation approach used to build the database. Following are the common

approaches to implement biological databases[15].

Flat files

A flat file database is built around a single table. This table contains all the details with fields for each and every parameter. Each record is specified in a single line and parameters/fields are separated by specified delimiters.

Even though the implementation appears direct and simple, maintenance of such databases is painstaking as they are prone to data corruption, errors and data redundancy. Due to data redundancy and different flat file internal formats, it is difficult to integrate two databases. File specific parses or data converters are required to convert flat file data into a common flexible form such as XML.

XML

XML is an advanced kind of flat file format with greater support to represent complex nested data structures. It also possesses the ability to contain data definitions and supports introduction of new definitions and tags as required. Therefore, XML is scalable. In practice XML proves to be fast to access and reliable, thus making it popular among web based database applications.

Availability of generic XML parsers to convert data into XML based databases can act as input and output formats of other implementation approaches such as relational databases.

Relational databases

Relational database systems were developed after the introduction of relational model[16] in 1970. The relational database is a collection of relations which resembles a table of values or a flat file[9] to some extent. In this table of values each row is named a 'Tuple' or record, a column header is called an 'Attribute' and the table is called the 'Relation'.

Relations differ from ordinary tables and flat files as they are not sensitive to the ordering of tuples and ordering of attributes and values within a tuple. In relational databases various restrictions can be applied to the data. They are called 'constraints' and classified into inherent model-based constraints, explicit schema based constraints and application based constraints.

Relational database implementations are currently one of the more successful ways of implementing a biological database[9].

Object oriented databases

Object oriented databases were developed due to two main reasons - modeling challenges imposed by more complex data domains such as biological research, geographical information systems, advanced multimedia systems etc.[8] and the requirement for seamless integration with Object Oriented Programming Languages[8] (OOPL).

An object database is a collection of objects which represent an instance of an abstract or concrete entity in the real world. Each object comprises of attributes and methods. These objects in the database follow the rules of Inheritance, Polymorphism and Data Encapsulation which forms the basis of object oriented programming concepts[15].

Even though the representation of complex molecular biological data is more accurate with the object databases, due to the immaturity of commercial products and their enterprise architecture, they are still not common in the biological database domain[15].

Abstract Syntax Notation (ASN.1)

This format was originally used to describe the messages of communication protocols of top layers in Open System Interconnection (OSI) model. It contains syntax and a description of how a data type is physically represented in a sequential file or a data stream[17]. With the progression of Human genome project National Centre for Biotechnology Information (NCBI) adopted this format to represent its sequence data and ASN.1 which is one of the major file formats in GenBank.

Approach	Advantages	Disadvantages	Current Databases
Flat File	Easy to implement, Commonly used	Difficult access, Difficult validation, Difficult integration	MITOMAP EMBL DDBJ
ASN.1	Relatively easy implementation, Standardised types and descriptions	Difficult integration	GenBank, OMIM CDD
XML	Flexibility, Improved access, Faster Response, Well established among web communities	As a DBMS less mature	SwissProt GO
Relational	Scalability, Reliability, Easy implementation	Multiple joins can reduce performance Enhance integration	GDB SwissProt SMART MITOMAP GO
Object Oriented	Easy implementation, Abstract data type support, Integration with OOPL	Document factorisation Enhance integration	MITOMAP

Table 1. Comparison of database implementation approaches[7,15]

Biological Database Integration

Doubling of biological data every 18 months[3] and a 5% growth in the number of databases[6] each year, resulted in scattering of biological knowledge in several hundreds of distinct databases[18]. Due to the differences in technical and political contextual aspects of biological databases, it is becoming unrealistic to solve a complex biological query by adhering to a single database[18].

This growth in number and content of biological databases lead to the development of research in biological database integration.

Strategies of Biological Database Integration

Even though the biological databases differ in their internal data structure, implementation approach and biological scope, almost all of them share similar three tier software architecture[19]. This similarity is the basis of the current strategies used for integration.

Strategies used for biological database integration[18,19,20] can be classified into three major categories.

- Link integration[19] or Hypertext navigation[20]

  Due to the compatibility with the nature of Internet and World Wide Web, it is by far the most successful strategy in current use. Integration depends on hypertext links from one database to another. Due to the unrestricted nature of external linking in web pages and minimum requirement of intercommunication with external sources, this approach is easy to implement[19]. Even though ease of implementation is its strength and key to popularity, there are disadvantages related to currency, accuracy, relatedness and validity of external source.

  Web services addressing certain issues by maintaining ease of implementation and improved validity are a variant of the link integration method[19].

- View integration[19] or Unmediated queries[20] with federated databases[20]

  This approach leaves information in source databases similar to above strategy, but builds a software environment which handles queries directed to multiple databases or a higher level integrated view of physically distinct databases by software means. Bottleneck of query response lies with the slowest responding data source[19,20].

  Usability of this approach is limited due to technical and political issues governing participating source databases[19].

- Data Warehousing[18,19]

  Building a single database with a unified data model to accommodate all the data in external sources is the basis of this approach. Compared to above two methods this is the most demanding strategy in terms of physical and technical resource requirement and intercommunication of participating sources.

The major problem of data warehousing is the dynamic nature of its data model and maintaining the currency of its data. Data model should be changed frequently to adopt new information from external sources and the need to build mechanisms to update the warehouse regularly which is difficult with dynamic data models.

Despite the provision of a solid theoretical basis for interoperability through similar three tier software architecture, it is still difficult to implement these solutions due to the ambiguity of ontologies and unique identifiers used to denote data[19]. Development and sharing of unified ontology for biological database domain and introduction of a global identification standard for molecular biological data will resolve most of these problems of biological database integration.

Social and ethical issues of biological databases

Recent growth in advance genomic research technologies have resulted in multiple genome wide association studies and publication of individual genome data sets. With the availability of such meticulous data dimensions of research ethics related to issues of privacy, confidentiality and consent needed to be broadened[22].

Discrepancies in the rate of advancement in research technologies and the rate of upgrading of available ethical framework was seen as the major cause for the ethical, legal and social implications raised by novel technologies[21].

Therefore, ethicists were forced to re-engineer the ethical framework with the advancement of science. Genetic privacy: right to protection from non-voluntary disclosure of genetic information[22] is a new concept introduced during recent decades due to the advancement in genetics and information technology. Health information privacy concept was re-evaluated with the advancement of medical informatics leading to novel mechanisms of information storage[22].

Due to the genetic resemblance in individual communities, conventional individual based privacy and consent protection was not sufficient to cover the whole community. Concepts of community based consent and privacy were needed to overcome these issues which could lead to categorical discrimination based on genetic information[22].

Efforts to re-engineer current ethical framework resulted in important recommendations[23] which is related to human genomic studies. Major recommendations were issued relating to returning research results, obligations to third party relatives and future use of data[23].

Therefore, biological database designers and research communities should adhere to a strong security framework which involves hardware, software and live-ware in order to protect the privacy and confidentiality of research subjects.

References

1. Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: Juggling between evolution and stability. Briefings in Bioinformatics 2004; **5**(1): 39-55.
Doi: http://dx.doi.org/10.1093/bib/5.1.39

2. Bourne PE, Westbrook J, Berman HM. The Protein Data Bank and lessons in data management. Breifings in bioinformatics 2004; **5**(1): 23-30.
Doi: http://dx.doi.org/10.1093/bib/5.1.23

3. National Institute of Health, April 2008. Retrieved in October 2010:
Available from: http://www.nih.gov/news/health/apr2008/nlm-03.htm

4. Robbins RJ. Biological databases: A new scientific literature. Publishing Research Quaterly 1994; **10**: 3-27.
Doi: http://dx.doi.org/10.1007/BF02680434

5. Petterson E, Lundeberg J, & Ahmadian A. Generations of sequencing technologies. Genomics 2009; (93): 105-111.
Doi: http://dx.doi.org/10.1016/j.ygeno.2008.10.003

6. Cochrane GR, Galperin MY. The 2010 Nucleic Acids Research Database Issue and online Database collection: a community of data resources. Nucleic Acids Research 2010; **38(**D1-D4): 1-4.
Doi: 10.1093/nar/gkp1077

7. Navathe SB, Patil U. Genomic and Proteomic Databases and Applications: A Challenge for Database Technology. In Lecture Notes in Computer Science. 2004; **2973:** 81-98.

8. Elmasri R, Navathe. Genome Data Management. In Fundamentals of Database Systems 2007; **5:**1042-54.

9. Birney E, Clamp M. Biological database design and implementation. 2004; ***5***(1): 31-38.

10. Peter P, Chen S. The Entity Relationship Model - Toward a Unified View of Data. ACM Transactions on Database Systems 1976; **1**(1): 9-36.
Doi: http://dx.doi.org/10.1145/320434.320440

11. Elmasri R, Ji F, Fu J. Modeling Biomedical Data. In C. Jake, & S. Amandeep S. (Eds.), Biological Database Modeling 2007; 25-50.
Doi: http://dx.doi.org/10.1504/IJBRA.2007.015008

12. Elmasri R, Weeldreyer J, Hevner A. The category concept: An extension to the entity-relationship model. Data Knowledge Engineering 1985; **1**(1): 75-116.
Doi: http://dx.doi.org/10.1016/0169-023X(85)90027-8

13. Rumbaugh J, Jacobson I, Booch G. 1999. The Unified Modeling Language Reference Manual. Pearson Education.

14. Paton NW, Khan SA, Hayes A, Moussouni F, Brass A, *et al*. Conceptual modeling of genomic information. Bioinformatics 2000; **16**(6): 548-557.

Doi: http://dx.doi.org/10.1093/bioinformatics/16.6.548

15. Hasegawa H. 2008. Genome Databases Current Implementation Practices. Retrieved in October 2010.
Available from: https://docs.google.com/viewer?a=v&q=cache:aY4x8Bh-T9MJ:www.cs.helsinki.fi/u/jplindst/biodb2008/HASEGAWA_BioDB_seminar.pdf+&hl=en&gl=lk&pid=bl&srcid=ADGEESgjRWtBC1cvcDYhuN8ZyqgP7HPulZ9kTfbFH-qcGSwRD9mqQRl_1YPvIKBp3EbWiPk5B4Zvd2DB1NPftuqPH-IhQiOoslqlle7ZxmfZS_lxBQxzf9q7FupFZk3Xw7fZhPyjhVou&sig=AHIEtbRxfopyRcqrWk_oZ-kAFBEeQLVE9w

16. Codd E. A Relational model for large shared data banks. CACM 1970; **13**(6): 377-87
Doi: http://dx.doi.org/10.1145/362384.362685

17. Buneman P, Davidson SB, Hart K, Overton C, & Wong L 1995. A Data Transformation System for Biological Data Sources. 21st VLDB Conference. Zurich, Switzerland.
Doi: http://dx.doi.org/10.1089/cmb.1995.2.557

18. Philippi S. Light-weight integration of molecular biological databases. Bioinformatics 2004; **20**(1): 51-57.
Doi: http://dx.doi.org/10.1093/bioinformatics/btg372

19. Stein LD. Integrating biological databases. Nature Reviews Genetics 2003; **4:** 337-45.
Doi: http://dx.doi.org/10.1038/nrg1065

20. Karp PD. A Strategy for Database Interoperation. Journal of Computational Biology 1995; **B**(4): 573-586.
Doi: http://dx.doi.org/10.1089/cmb.1995.2.573

21. Lunshof JE, Chadwick R, Vorhau, DB, & Church GM. From genetic privacy to open consent. Nature Reviews Genetics 2008; **9**: 406-11.
Doi: http://dx.doi.org/10.1038/nrg2360

22. The Genetic Information Nondiscrimination Act of 2008: Information for Researchers and Health Care Professionals. Parlimentary Act, Department of Health and Human Services – USA, 2009

23. McGuire AL. Caulfield T, Cho MK. Research ethics and the challenge of whole-genome sequencing. Nature Reviews Genetics 2008; **9**: 152-156.
Doi: http://dx.doi.org/10.1038/nrg2302