Int. J. Drug Dev. & Res.
| October - December 2013 | Vol. 5 | Issue 4 | ISSN 0975-9344 |
www.ijddr.in

Biological Databases

Kaviena Baskaran

1st Year BDS, Saveetha Dental College & Hospital

Corresponding Authors:
E-mail: kavitju@yahoo.com

Abstract:
Biology has entered a new era in distributing information based on database and this collection of database become primary in publishing information. This data publishing is done through Internet Gopher where information resources easy and affordable offered by powerful research tools. The more important thing now is the development of high quality and professionally operated electronic data publishing sites. To enhance the service and appropriate editorial and policies for electronic data publishing has been established and editors of article shoulder the responsibility. [1]

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

Review Paper

Introduction

What is electronic data publishing (EDP)? Can electronic publishing be trusted and what is the role of editor in EDP? Let us discuss EDP and its role in biology, showing how in some areas of biology EDP has evolved from an electronic version of a traditional review into a new kind of primary literature. Traditional biological publishing provides information and knowledge, not data. Now biological research is generating more information that is close to the data end of the spectrum. To accommodate this, large databases has been introduced to support EDP for molecular biology. For example, Genbank® and GSDB collect nucleotide sequences and PIR, PDB stores data information related to protein [1]

The Human Genome Project

The international Human Genome Project was the first science project in biology provides a compelling argument for EDP. The main objective of this project is to construct high resolution genetic map of human genome. Also to determine complete sequence such as gather, store, distribute and analyses data produce on DNA. A creation of proper technology is needed to achieve this objective. [1]

Databases as Publishing

Early Database Development

In the early stage, the prominent biological databases, such as GenBank or PIR, were almost same as the review article. Important details were collected from the literature by single researcher who then compiled and published them in a form that supported further use and analysis. [2]

The Database Scaling Problem

Primary-literature status editorial involvement is also facilitated by direct data submission. For the first time, editorial quality control could be applied to the sequence information itself. [3]

Example of Electronic Data Publishing

The Genome Database

Delici.C stated that "Publication-on-demand changes the role of the database from publication to publisher. The user interacts with the interface to determine what information is available, then decides what to "buy" and places an on-line order, either for a one-time publication, or for a subscription to a specified review, effectively designed by the user but carried out by the database staff and the research community as they populate the full database from which the review is extracted."[4]

Gopher

At John Hopkins, a year ago, the Gopher server were established initially provides electronic version of subjects from the book Mathematics and Biology.[5] To retrieve information from one server is easy via Gopher. [6] The common possible problem with the Gopher is the size and dynamism of the resource. It would be difficult to locate information if you do not know where to look. This problem was recognized immediately after the development of a system that lets you search all of the Gopher menus in the world with a single query and then new service was added by Veronica. [7]

Emergence and Evolution of Computer Based Molecular Biological Databases

The compilation by late Margaret Dayhoff in 1965, The "Atlas of protein sequence and structure" is the initial published molecular biological information content which resembles the features of a modern day biological database [8]

In the 1970's, the fast growth of computer science and information science, influenced the biological scientists to create freely accessible computer based repositories for biological data. Started 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 [9]

While Genbank Project initiated by United States National Institute of Health (NIH), the European Molecular Biology Laboratory (EMBL) started establishing its own sequence data bank. Few years later, GenBank and EMBL started collaborating. Later in mid 1980s, with the collaboration of the DNA Data Bank of Japan (DDBJ), the International Nucleotide Sequence Database Collaboration (INSDC) was formed.[10]

IN early 1990's with the appearance of Electronic Data Publishing concept the scope of biological databases started to develop into fields of data visualization and data publishing.[11]

Traditional biological printings were based on findings and knowledge derived from experimental data, but with the advancement of experimental research many information which are close to the data end of the spectrum was created. It's almost impossible to share huge amount of such data developed through current technologies like ultra-high throughput sequencing without the support of a publishing database which facilitates electronic data publishing. [11][12]

The best example to express 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.[11] This success marks the development in databases of post genomic era. [10]

In 2010, Database issue of Nucleic Acids Research (NAR) includes descriptions of 58 new data resources and updates 73 previously published data resources. The online Database Collection which accompanies this issue holds a total of 1230 data resources which represents 5% growth in the number of biological databases during the 2 years period from 2009. [13]

Properties of Biological Databases

Beside the rapid rate of data generated by advanced biological research, the challenges presented to the database developers by the inherent properties of biological data and nature of data users also contributes in the data driven growth and evolution of biological databases. It is important to digest these properties of biological data ranging from research articles to complex metabolic pathways before developing solutions for any biological database problem. [14]

From the analytical context of data, other challenges may arise where designers need to model meta data for data analysis and constructing of archival capabilities for data validation and analytical purposes. [15]

Entity Relationship (ER) Based Modeling

Since its introduction, the ER modeling is very popular in database community for its ability in modeling high level conceptual schemas (implementation independent model). [15][16] ER models are more compatible to model well defined entities with simple relationships [17].

Entity Category Relationship (ECR) model created in 1985, opens the way for the development of Enhanced/Extended Entity Relationship (EER) model. [15][18]

Even though the EER models could capture 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 model was introduced in 2007. [17]

Unified Modeling Language (UML)

UML, the common purpose visual modeling language that captures information about the static structure and dynamic behavior of a system which is ideal to model molecular biological scenarios. The ability of modeling molecular biological data and also with the added software support, it makes UML a highly recommended tool among biological database developers. [19][20]

Structured Query Language – Data Definition Language (SQL-DDL)

This is a SQL description of a relational database table structure. This can be used as principle data model description as a master of what a database stores. [19]

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

Apart from SQL – DDL or high level UML views, this can also be used as principle data model descriptor. The important practical aspect is to firmly support to a single master data model to avoid branching towards modeling of all biological substances in a single database design which can lead to instability in the database structure.[19]

Implementation approaches

The next most important decision after having a single primary data model of the mini world is about the implementation approach used to build the database. The below are the common approaches to implement biological databases. [21]

Relational databases

After introduction of relational model in 1970, the Relational database systems were developed. [22] It is a collection of relations which resembles a table of values or a flat file[19] to some extent. Currently these Relational database implementations become one of the more successful ways of implementing a biological database. [19]

Object oriented databases

The are two main reasons for the development of this Object oriented databases, modeling challenges imposed by more complex data domains such as biological research, geographical information systems, advanced multimedia systems etc. and the requirement for seamless integration with Object Oriented Programming Languages(OOPL). [15]

Abstract Syntax Notation (ASN.1)

Initially this format was used to elaborate the messages of communication protocols of top layers in Open System Interconnection (OSI) model. It consist syntax and an elaboration of how a data type is physically represented in a sequential file or a data stream. [23]

Biological Database Integration

A 5% growth in the number of databases each year and Doubling of biological data every 18 months, resulted in scattering of biological knowledge in several hundreds of distinct databases. Because of the differences in technical and political contextual aspects of biological databases, it becomes unrealistic to solve a complex biological query by adhering to a single database. [24]

Strategies of Biological Database Integration

Link integration or Hypertext navigation

The minimum requirement of intercommunication with external sources and unrestricted nature of external linking in web pages causes this approach to be easy to implement. Web services addressing certain issues by maintaining ease of implementation and improved validity are a variant of the link integration method. [25]

View integration or Unmediated queries with federated databases

Bottleneck of query is the slowest responding data source. [25][26] Technical and political issues governing participating source databases cause the limitation of the usability of this approach. [25]

Conclusion

There are various unanswered question raised on the Electronic data publishing especially on the safety, liability and also the consensus, as does the traditional literature. With continue editing in EDP, will it remain the authorship same as the print literature? With editorial policies and procedure how EDP will become an edited communication system? Scientist realised that, it is possible with having reliable scientific literature through establishing professional editing standard. The same applies for EDP. Projects such as GDB/GSDB or PIR, the scientist are clear with the editorial policies. However, Sites which are not properly monitored such as Gopher, the information will come and go making tracking difficult.

References

1) Robbins, Robert J.. 1994. Biological databases: A new scientific literature. *Publishing Research Quarterly, 10:3-27.*

2) Culliton, B. J. 1990. Mapping terra incognita (humani corporis). Science 250:210-212.

3) Cuticchia, A. J., K. H. Fasman, D. T. Kingsbury, R. J. Robbins, and P. L. Pearson. 1993. The GDB Human Genome Data Base Anno 1993. Nucleic Acids Research. 21:3003-3006.

4) DeLisi, C. 1988. The human genome project. American Scientist 76:488-493.

5) Fickett, J. W. 1989. The database as a communication medium, in R. R. Colwell [Ed.],

Biomolecular Data: A Resource in Transition. New York: Oxford University Press, 295-302.

6) Gilbert, W. 1991. Towards a paradigm shift in biology. Nature 349:99.

7) Haeckel, E. 1897. The Evolution of Man, Volume 1, Third Edition. New York: D. Appleton and Company.

8) Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: Juggling between evolution and stability. Briefings in Bioinformatics 2004; 5(1): 39-55.

9) Bourne PE, Westbrook J, Berman HM. The Protein Data Bank and lessons in data management. Breifings in bioinformatics 2004; 5(1): 23-30.

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

11) Robbins RJ. Biological databases: A new scientific literature. Publishing Research Quaterly 1994; 10: 3-27.

12) Petterson E, Lundeberg J, & Ahmadian A. Generations of sequencing technologies. Genomics 2009; (93): 105-111.

13) 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.

14) 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.

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

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

17) Peter P, Chen S. The Entity Relationship Model - Toward a Unified View of Data. ACM Transactions on Database Systems 1976; 1(1): 9-36.

18) Elmasri R, Ji F, Fu J. Modeling Biomedical Data. In C. Jake, & S. Amandeep S. (Eds.), Biological Database Modeling 2007; 25-50.

19) Elmasri R, Weeldreyer J, Hevner A. The category concept: An extension to the entity-relationship model. Data Knowledge Engineering 1985; 1(1): 75-116.

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

21) Hasegawa H. 2008. Genome Databases Current Implementation Practices. Retrieved in October 2010.

22) Codd E. A Relational model for large shared data banks. CACM 1970; 13(6): 377-87

23) Buneman P, Davidson SB, Hart K, Overton C, & Wong L 1995. A Data Transformation System for Biological Data Sources. 21st VLDB Conference. Zurich, Switzerland.

24) Philippi S. Light-weight integration of molecular biological databases. Bioinformatics 2004; 20(1): 51-57.

25) Stein LD. Integrating biological databases. Nature Reviews Genetics 2003; 4: 337-45.

Article History: ------------------------

Date of Submission: 13-07-2013

Date of Acceptance: 29-07-2013

Conflict of Interest: NIL

Source of Support: NONE