International Journal of Pharma and Bio Sciences V1(2)2010

MICROBIAL GENOME DATABASES: A USER'S PERSPECTIVE

JERRINE JOSEPH, SAMEER HASSAN, VASANTHI RAJENDRAN AND VANAJA KUMAR*

Tuberculosis Research Centre, Chennai, INDIA.

***Corresponding author** vanajakumar@trcchennai.in

ABSTRACT

The challenge of assimilation of the accumulating wealth of information necessitates the urge for having a comprehensive database. Such database can be used as an important tool in assisting scientists to decipher a host of biological phenomena ranging from the structure of biomolecules and their interaction to the whole metabolism of organisms and the evolution of species. This knowledge helps facilitate the fight against diseases and assists in the development of medications. The efforts of data integration through data warehouse or data federation are needed to answer the complex analytical queries in the bioscience. The aim of this paper is to deal with the data integration of various databases.

KEY WORDS

Data integration; Bioscience; Comprehensive database; Microbial genome and Annotation.

INTRODUCTION

Science has historically been divided into disciplines for the convenience of learning and of managing the complexity. As a result, the scientific data have been traditionally collected and managed by categories. With the advent of high throughput technologies in biology, the data accumulation has been growing exponentially. With the current state of web technologies and data management technologies in computer science it becomes possible to understand organisms systematically by comprehensive data analysis and information extraction[1] .The requirements for comprehensive information extraction and data analysis across diverse disciplines are becoming emergent for the need of our understanding in system biology[3]. The current data marts or databases, most of which are originally designed for the purpose of data storage or repository, become incompetent in answering

MICROBIAL GENOME DATABASES: A USER'S PERSPECTIVE

comprehensive questions related to discovery or decision-making in large scale[4]. Although well structured solutions for data integration in industry already exist, the technologies for biological data integration are still in its infant stage and special care needs to be taken with respect to the unique properties of biological data.

Microbial genome annotation generally refers to a process of assigning biological meaning to the raw sequence data by identifying gene regions or functional features and determining their biological functions. Gene annotation is a combination of automated methods that generate a "preliminary" annotation in terms of predicted genes (also called Open Reading Frames or ORFs, which represent the sequence of DNA or RNA located between the start codon and stop codon sequence) and associated functions along with pathways based on sequence similarity or profile searches. The result of a preliminary (baseline) annotation is often sparse, with numerous genes not having associated functions or pathways. Consequently, several techniques are employed for annotating genes as well as validating baseline annotations. The most effective annotation techniques involve comparative multi-genome analysis based on observed biological evolutionary phenomena: pairs of genes with related (coupled) functions are often both present or both absent within genomes; tend to be co-located (on chromosomes) in multiple genomes; might be fused into a single gene. Doing justice for the comparison of the genomes from different species or within the same species is another mammoth task *in vivo*. The same task has been made very easy due to the aid of bioinformatic tools like genome browsers. They have become a boon to help in comparative genomics visualization with greater ease, efficiency and accuracy.

Bioinformatic tools themselves have undergone evolution over a period of time and come up with different versions supporting different databases. The scenario is that apart from the scientific purpose lot of technical aspect of the programming, scripting, server interface and operating systems at the end user PCs play a pivotal role in the access to the information. These drawbacks actually curtail the scientific pursuit of the enthusiastic researcher.

CHARACTERISTICS OF BIOLOGICAL DATABASES

The characteristics of biological databases are derived from their unique origin and history. Originally, most biological databases were devised by a group of scientists who have limited database background. The major purpose of these databases was data storage rather than information extraction. Furthermore, biological data is hierarchical by nature and its data types are tightly correlated with the specific technologies of data acquisition[6]. As a result, the databases are sporadic, data types are heterogeneous and span diverse domains. For example, biological data for human species traverse multiple levels, such as organism, organ, tissue, cell, organelle and pathways or networks, and span diverse domains, such as genomics, transcriptomics, proteomics, pheneomics, localizeomics, ORFeomics, pharmacogenomics, pharmacogenetics clinical trials, etc[2]. The nature of biological data, plus its unique evolution over history leads to some special features of molecular databases, which is summarized in detail below.

1. Molecular data are highly heterogeneous due to the inherent complexity in biological system and a wide array of technologies used to study them. The new terminology "omics" is a real-life reflection of this reality[2]. The classification of databases based on

the contents of data can easily be divided into the following main categories: genome database, gene databases, gene expression databases, protein databases, protein-protein interaction databases, pathway databases, etc. Each of these database types themselves can be easily divided into several subtypes of databases. For example, the protein database includes protein sequence databases, protein structure databases, protein signature, domains and profile databases.

2. The data volume is large with unique data types, and data accumulation is on-going and far from complete. For instance, the estimated human gene number in total is 20,000 to 25,000 [8]. Without considering individual differences or ethnic differences, theoretically, a completed gene expression profiling database should contain expression profiles of all these genes in all human organs/tissues, cell types in particular cases, covering various development stages or time lines without considering any stress effects. Viewed in the context of other molecular types such as DNA and proteins and the various types of technologies used to study them, the volume of biological data becomes extremely high. The data types are dictated by the biotechnologies used in experiments[6]. Data sources in bioscience are highly dynamic. The data dimensions in biosciences are expanding rapidly as a result of the development and the innovation of new technologies. To pace with these changes, new data types or databases are emerging all the time and the existing databases continuously restructure their formats to incorporate the new data, which leads to the multiple generations/releases of legacy databases yearly. For example, GenBank, one of the major genome, gene and protein data repository systems, make their release bimonthly [1]. This type of deep hierarchical structure is very common in biology and it could be difficult to model and inefficient to query using traditional relational models [9].

3. Lack of standardization in data formats and in controlled vocabularies in scientific domains. Molecular databases are highly heterogeneous due to their original formation and history. As a result, the database schema of the similar biological data types by different databases will be quite different due to the technologies used. One will find that the across-platform comparison is almost impossible at the data-analysis level. Additionally, data formats vary over different domains and over different projects. The vocabularies in describing biological objects are ambiguous due to the fact of widely used synonyms and homonyms. We end up with a vast mosaic of databases in one biological domain with different formats typically using non-standard query software specific for that particular database [4]. These databases and systems often do not have an explicit database schema, which is conventionally considered as a formalized catalogue of all interrelated tables in a database with well-defined attributes and well-structured indices of these tables, which is prevalent in industry databases [5]; http://img.jgi.doe.gov/).

4. The database management applications and data-access tools for biological databases are at their infant stages .Lack of standardization in data formats and the dynamics in data types hamper the development of application tools in biological database management systems [6]. Hence the retrieval efficiency is low and complicated, and heterogeneous applications need to be developed to handle the information extraction and analysis.

REQUIREMENTS FOR BIOLOGICAL DATABASES AND APPLICATIONS

MICROBIAL GENOME DATABASES: A USER'S PERSPECTIVE

The unique features of data or databases in biosciences hold some interesting requirements from biologists to the databases in biosciences. For example, to answer comprehensive biological queries one often needs to traverse a wide range of object domains from many heterogeneous databases and a user must click through many interfaces and must make efforts to manage intermediate results[6]; http://genome.jgipsf.org/microbial/index.html).

Furthermore, a data integration technology that recognizes which parts of two data sources have the same meanings or overlapped domains is desirable. The detailed requirements of databases in biosciences can be specified below. 1) The heterogeneous features of biological databases require that data models and database management systems in biosciences are capable of handling data types and are flexible in dealing with data types[10]. Otherwise the possible constraints of data types and values placed on databases and database management systems could result in the exclusion of unexpected types. 1) Diminishing the reliability of query results or data analysis results. 2) The highly dynamic feature of biological databases challenges the database and application development community in biology to support database schema evolution and data object migration for improving information flow between generations/releases of databases [6]; http://img.jgi.doe.gov/. Currently, the ability to extend the database schema to meet the requirements of frequent changes in the biological setting is unsupported in most relational and object database management systems [9]. However, this sort of tracking in history is important for biological researchers to be able to access and verify previous results. Therefore, mechanisms for aligning different biological databases with similar contents or different versions of formats should be supported. Data alignment tools and data integration tools based on the various biological workflow for legacy databases should be available[6]. 3) The deep hierarchical nature of biological data causes some concerns whether relational schemas will meet the challenge for efficient data representation and retrieval in highly integrated data warehouse[9].

CHALLENGES OF DATA INTEGRATION IN BIOSCIENCE

The problems of modeling, storing and querying data in bioscience is not solved satisfactorily yet[3]. One of the challenges we face is to represent the relationships in bioscience in a precise and unambiguous manner. Obviously, developing a single global data schema for data integration seems impossible and difficult[4]. One proposal for solving this problem is to develop mediated schemas which focus and represent one domain of knowledge each to further integrate into a mediated schema to represent the global and complex knowledge.

The biological raw data is distributed amongst many different general and specialized databases. Each database provides information on particular organisms, but do not and are not able to deal in depth covering all the features of the genome annotation, protein prediction and guiding to biotherapeutic discovery as the ultimate goal. The flow of the path from gene annotation for any pathogen, should logical progress from genomics to proteomics to metabolomics to finally drug discovery. Unfortunately few of the existing biological databases are not in position to project the entire sequence of the flowchart on a single platform for any single organism.

Hence, it is left to the discretion of the user to choose specific database for specific organism for only partial information. The onus lies on the user to

gather the required data from the various databases available and then critically evaluate and assimilate the information into knowledge to enable mankind benefit through it.

Microbial Genome Data Sources

Microbial genomes sequenced by organizations worldwide, follow a similar annotation process, and end up in one of several microbial genome data sources, such as EBI Genome Reviews [11], CMR [13] and RefSeq [14]. Furthermore, additional genome annotation details such as protein families and pathways reside in multiple specialized data sources such as UniProt (protein sequences and functions), InterPro (protein families and domains), COG (clusters of orthologous genes), and KEGG (pathway maps). The result of diverse annotation methods, curation techniques, and functional characterization employed across microbial genome data sources. An additional problem in dealing with these sources is the difficulty of determining the coherence and completeness of their data. Data *coherence* regards the quality of annotations: although inherently imprecise, these annotations can be qualified in terms of "biological coherence" rules. For example, predicted genes with overlapping sequences often indicate errors in gene prediction and need to be manually reviewed and corrected. Problems related to data coherence are caused by the high cost in terms of time and expertise needed to validate and correct annotations manually. Data *completeness* regards the extent and coverage of functional characterization and depends on the diversity of the genomes included in a data source and the depth of integration of genome annotations collected from diverse sources [12]. The list of the genome databases available online is tabulated in Table 1.

Table 1.

The features of the various microbial databases on line is classified and given as a tabular format

DATABASES	ORGANISMS	TOOLS	FEATURES	WEBSITE
TIGR-Comprehensive Microbial Resource (CMR)	Bacteria (526), Archaea (43)& Viruses (3)	• Genome homology Graph • Protein Scatter plot • GC comparison graph	Genome and proteome data analysis made user friendly by graphical visualization of the output	http://cmr.jcvi.org/cgi-bin/CMR

MICROBIAL GENOME DATABASES: A USER'S PERSPECTIVE

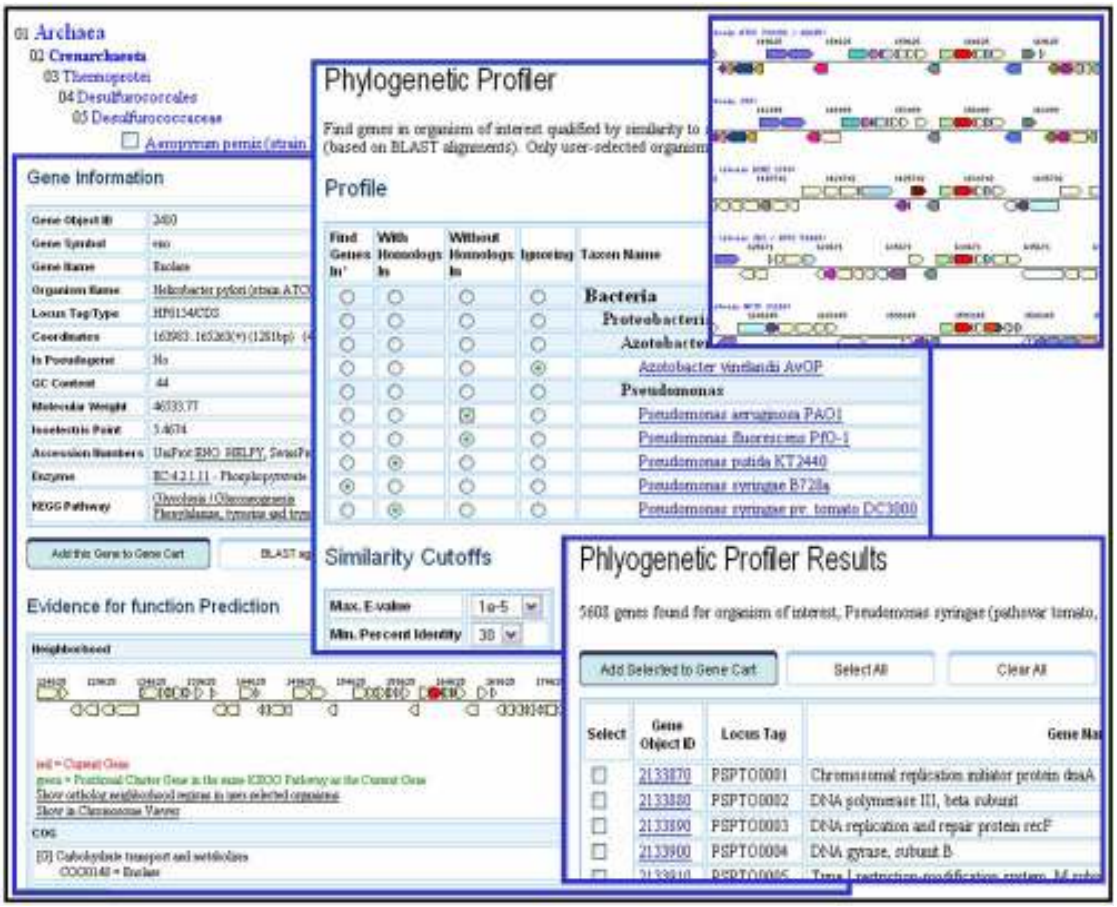
NCBI	Phages (255), Plasmids (2020)&Viruses (2886)	• GenePlot • BLAST • gMap • Map Viewer • TaxPlot	Genome and proteome sequences of most organisms and clustered based on homology	http://www.ncbi.nlm.nih.gov/sites/entrez
GOLD	Archaea (109) Bacteria (l,2931), Eukaryotes (1119) , metagenomes ongoing genomes (128), 1069 complete published genomes.		GOLD is a comprehensive resource for accessing information on genome and metagenomics	http://igweb.integratedgenomics.com/GOLD/.
NMPDR	Archaea (47), Bacteria (725) and Eukaryotes (29)	Rapid Annotation using Subsystem Technology	The **NMPDR** provides curated annotations for comparative analysis of genomes and biological subsystems.	http://www.**nmpdr**.org
IMG	Eukaryotic sites (111**)**, Prokaryotic microbial sites **(**498) with 335 complete.	MyIMG gene annotation	Explores the microbial genomes along its 3 main dimensions (genomes, genes and functions).	http://genome.jgi-psf.org/
MBGD	Bacteria (744), Archaea (54), Eukaryotes (16)	• DomClust • CGAT • CoreAligner	**MBGD** is a database for comparative analysis of completely sequenced microbial	http://mbgd.genome.ad.jp

genomes.

For example using IMG database the output of homology search of a gene against the database is shown both in tabular and graphical forms thereby providing a better insight into gene arrangement, neighbouring genes, phylogenetic profiling etc (Figure 1).

Figure 1.

IMG Web Data Explorer: Data Analysis Example.



MICROBIAL GENOME DATABASES: A USER'S PERSPECTIVE

CONCLUSION

Effective data analysis across biological data management systems involves providing support for seamless composition of analysis operations, which in turn requires a systematic process for analyzing the data structure and operations of the application domain. Lack of such collaborations often leads to poor use of data management technologies or misunderstood requirements which can result in "sterile pursuits of insignificant or misunderstood problems". A systematic development process, starting with requirements analysis, provides the framework needed for specifying analysis workflows including documentation for data structure and operations. Following such a process is time consuming and requires resources that may not be available to academic groups. The challenges of data integration in biosciences have to be dealt with in order to come up with effective data management system. The need is for an immediate addressal to this issue by developing databases which cater to multiple use of the research community by integrating various data sources on a single window.

REFERENCES

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool, J. Mol. Biol 215 (3): 403-10 (1990).
2. A. Bernal, U. Ear, and N. Kyrpides. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide, Nucleic Acids Res 29 (1): 126-7 (2001).
3. Bloom, T. and Sharpe, T. (2004) 'Managing Data from High-Throughput Genomic Processing: A Case Study' *Proc. of the 30th VLDB Conference*
4. P. M. Bowers, M. Pellegrini, M. J. Thompson, J. Fierro, T. O. Yeates, and D. Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution, Genome Biol 5 (5): R35 (2004).
5. Branka, M.A., Venkatesh, T.V. and Goodman, N. (2001) 'Bioinformatics: Getting Results in the Era of High-Throughput Genomics' *Cambridge Healthtech Institute Report*.
6. Davidson, S.B., Crabtree, J., Bunk, B., Schug, J., Tannen, V. and Stoeckert, C. (2001) 'K2/Kleisli and GUS:Experiments in Integrated Access to Genomic Data Sources', *IBM Systems Journal*, Vol. 40, pp.512-531.
7. M. Y. Galperin. The Molecular Biology Database Collection: 2005 update, Nucleic Acids Res 33 (Database issue): D5-24 (2005).
8. M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, .D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource, Nucleic

Acids Res 32 (Database issue): D258-61 (2004).
9. L. Hauser, F. Larimer, M. Land, M. Shah, and E. Uberbacher. Analysis and annotation of microbial genome sequences, Genet. Eng. (N Y) 26 225-38 (2004).
10. H. V. Jagadish, and F. Olken. Database management for life science research: summary report of the workshop on data management for molecular and cell biology at the National Library of Medicine, Bethesda, Maryland, February 2-3, 2003, Omics 7 (1): 131-7 (2003).
11. P. Kersey, L. Bower, L. Morris, A. Horne, R. Petryszak, C. Kanz, A. Kanapin, U. Das, K. Michoud, I. Phan, A. Gattiker, T. Kulikova, N. Faruque, K. Duggan, P. McLaren, B. Reimholz, L. Duret, S. Penel, I. Reuter, and R. Apweiler. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes, Nucleic Acids Res 33 (Database issue): D297-302 (2005).
12. R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides. The ERGO genome analysis and discovery system, Nucleic Acids Res 31 (1): 164-71 (2003).
13. J. D. Peterson, L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White. The Comprehensive Microbial Resource, Nucleic Acids Res 29 (1): 123-5 (2001).
14. K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res 33 (Database issue): D501-4 (2005).