Bioinformatics Data Solutions

Shobharani¹, Daharani N.V.², P. Malathi³, Anitha J.⁴, Shailaja L. K.⁵ and A. J. Prasad⁶

¹,²,³,⁴,⁵Department of M.C.A., Dr. Ambedkar Institute of Technology, Bangalore 560056
shobharanir@rediffmail.com¹; Professor⁶, pjaajp@gmail.com⁶
dharaninv@yahoo.com²,bfgcpm@rediff.com³, alochan@rediffmail.com³ and shailaja_lk@yahoo.co.in⁵

ABSTRACT

The combining of research results from different bioinformatics repositories requires extracting of information from the source databases, transforming it and then loading it into the data warehouse. The need to merge existing repositories involves data warehousing. Data integration solution aims to provide an integrated access for users to query across heterogeneous databases. To ease the task of mapping the source schemas with the integration schema, over which the users formulate their queries, many researchers have used schema matching algorithms. The requirement to reconstruct large-scale integrative networks from relevant databases and provide an improved method for parameter estimation needs to be investigated. This may be seen as a method of associating a model with several data sets. Mathematical models are important for understanding how various bioinformatics systems function. The construction of such models involves the integration of qualitative and quantitative information describing the various entities of the data repositories. However this information is distributed amongst a number of databases that suffer from the inconsistent naming of the entities. The approach for integrating distributed data which can address issues associated with the mathematical modeling of bioinformatics systems requires resolving of conflicts between heterogeneous data sources. Data integration systems are based on i) the global schema, ii) the heterogeneous set of source schemas, and iii) the mapping that maps queries between the source and the global schemas. When users use queries over the data integration system, they make queries over the global schema and the mapping then asserts connections between the elements in the global schema and the source schemas. The bioinformatics data solution management problem requires experimental data from multiple sources in diverse formats to be drawn together to generate mathematical models of the system from which the data was generated. In order to achieve this, the experimental manipulations must be coherent and provide a thread by which the gathered data may be linked. This requires a data model and database implementation to capture, store and allow retrieval of both numerical and image data from multiple assay types that are performed in the course of a systems bioinformatics study. Instead of capturing all data gathered for each assay performed, one aims only to capture those key results that are necessary for the construction of mathematical models.

KEYWORDS

bioinformatics data repositories, bioinformatics data solution management, bioinformatics Web services, data warehousing, mathematical modeling of bioinformatics systems, heterogeneous data sources, ranked search system approach to bioinformatics, CLOUD data management.

1. INTRODUCTION

The term bioinformatics first came into use in the 1990s and was originally synonymous with the management and analysis of DNA, RNA and protein sequence data. Computational tools for sequence analysis had been available since the 1960s, but this was a minority interest until advances in sequencing technology led to a rapid expansion in the number of stored sequences in databases such as GenBank. Now, the term has expanded to incorporate many other types of biological data, for example protein structures, gene expression profiles and protein interactions. Each of these areas requires its own set of databases, algorithms and statistical methods. Bioinformatics is largely, although not exclusively, a computer-based discipline. Computers are important in bioinformatics for two reasons:

i) Many bioinformatics problems require the same task to be repeated millions of times. For example, comparing a new sequence to every other sequence stored in a database or comparing a group of sequences systematically to determine evolutionary relationships. In such cases, the ability of computers to process information and test alternative solutions rapidly is indispensable.

ii) Computers are required for their problem-solving power. Typical problems that might be addressed using bioinformatics could include solving the folding pathways of protein given its amino acid sequence, or deducing a biochemical pathway given a collection of RNA expression profiles. Computers can help with such problems, but it is important to note that expert input and robust original data are also required. Gathering and managing data from diverse and heterogeneous datasets is central to many systems approaches to bioinformatics. We need to easily identify and gather the information that we require and then integrate and analyze the experimental and reference data sets which can come from myriad databases with a wide variety of formats and access methods. Data collection and pre-processing using bioinformatics Web services, including text mining enables a method for analysing data and generating visualizations that support bioinformatics data management solutions. The development of techniques for a suitable bioinformatics data solution model requires consideration of
the growing number of models existing and with their increasing complexity. The annotation of computational models enhances the semantic description of the modeled system. One requirement for the reuse of existing models is the ability to efficiently retrieve them from a model repository. The user obtains an unranked list of models for his query, e.g. sorted by model ID in Bioinformatics-Models Database or sorted by the author's last name or date in the Model Repository. A ranked search result will help the user to find the models relevant to his work. Hence it is necessary to use sophisticated retrieval techniques to sort the search results regarding different aspects. One needs to build on existing model annotations within computational bioinformatics models. Those annotations are extracted and then are used to define similarities between a query and a model. Each annotation entity (e.g. model publication, type of model encoding and so on) represents a feature in the method. A query then sets up a “query model” which will be mapped against the model base. It is necessary to develop an integrative approach, combining public data and software resources for the rapid development of large-scale text mining tools targeting complex bioinformatics information. The user supplies input in the form of identifiers used in relevant data resources to refer to the concepts of interest. By doing so, the user is freed from providing any other knowledge or terminology concerned with these concepts and their relations, since they are retrieved from these and cross-referenced resources automatically. The terminology acquired is used to index the literature by mapping concepts to their synonyms, and then to textual documents mentioning them. The indexing results and the previously acquired knowledge about relations between concepts are used to formulate complex search queries aiming at documents relevant to the user’s information needs. A bioinformatics data warehouse stores and integrates bioinformatics research data to provide a software infrastructure for bioinformatics R&D. Ontologies serve to define the concepts and relationships both within a system and between systems.

2. Data Warehousing for Integrative Bioinformatics:

The central issue of this Bioinformatics data management field is the transformation of often distributed and unstructured biological data into meaningful information. Bioinformatics Data Warehousing techniques arrange data in a structured form. This facilitates the retrieval of a subset of the data according to the user query. To organize data in a structured form definitively helps the data management solution. Bioinformatics Data warehousing provides the data integration and formatting functionality required by Bioinformatics data mining applications. A Bioinformatics data warehouse is a decision support database that is essential for R&D data search. The Bioinformatics data warehouse is an architectural construct that addresses the growing need of information for research purposes. Bioinformatics data warehouse serves as data infrastructure for bioinformatics research and development. The goal of the system is to provide data, as well as a software infrastructure for bioinformatics research and development. The system is based on relational data models.

The goal in bioinformatics is to integrate data from disparate sources of heterogeneous biological information. The integration of disparate, heterogeneous biological sources of data enables new scientific inferences which are useful for bioinformatics analyses, and for discovering scientific relationships between data. Most public repositories of biological data focus on deriving and providing one particular type of data.

Working with publicly available biological data can be challenging due to the volume and complexity of the data types. With the proliferation of massive, publicly available data sets, researchers need a way to readily access this data. Querying distributed data has inherent limitations such as the server resource restrictions of the remote resource, concerns of secure data transmission over the internet, and of course the actual logistics of querying distributed resources. In such an environment, the distributed search space is difficult to process in a high-throughput way, and requires complex queries to tie together the heterogeneous data. Consequently, there is a need for a data integration solution that facilitates search and retrieval in an efficient, flexible, high-throughput manner.

3. CLOUD DATA MANAGEMENT

Cloud computing describes both a platform and a type of application. A cloud computing platform dynamically configures, and reconfigures, servers as needed. Cloud applications are applications that are extended to be accessible through the Internet. These cloud applications use large data centers and powerful servers that host Web applications and Web services. CLOUD data management methodology can be applied to bioinformatics data solutions. Some of the advantages of CLOUD data management are: i) On-Demand Provisioning, ii) Ask for what you need, exactly when you need it, iii) Pay only for what you use, iv) Scalability is possible - Scale up or down depending on usage needs, v) Utilize the skills, knowledge and resources of experts. The procedure that is applicable can be obtained from the methods available in Amazon technology. Hundreds of thousands of registered customers distributed over many countries leverage Amazon’s technology based websites for their selling and buying requirements. Tens of millions of active customer accounts belonging to various countries across the world such as India, China, Japan, Germany, France, US, UK, Canada, etc. use Amazon websites and Amazon technology for their buying, selling and retail activities. Amazon’s network Developers & IT Professionals are involved for hosting web-scale solutions. CLOUD Attributes are abstract resources not tied to physical hardware and can be flexible as one’s needs demand. Some of the parameters that are important are: i) Web site hosting, ii) Application hosting, iii) Internal IT application hosting iv) Quick and effective marketing campaigns, v) Content delivery and media distribution vi) High performance computing, vii)
batch data processing, and large scale analytics, viii) Storage, backup, and disaster recovery. The importance of Data Integration for “CLOUD” are: i) Improved Adaptability and Agility, ii) Respond to business needs in near real-time, iii) Functional Reusability, iv) Eliminate the need for large scale rip and replace, v) Independent Change Management, vi) Focus on configuration rather than programming, vii) Interoperability instead of point-to-point integration, viii) Loosely-coupled framework, services in network. The requirements for integrating with “CLOUDS” are: i) The ability to handle larger data sets, ii) The ability to handle and resolve data inaccuracies and inconsistencies, iii) The ability to do data manipulation efficiently and inexpensively, iv) The ability to provide visibility into the lineage of data, v) The ability to decouple data access from the implementation.

4. BIOINFORMATICS DATA MANAGEMENT-1 (PRESCRIPTION ASSISTANCE SOLUTIONS)
The simplest example that can be considered for bioinformatics data management can be based on prescription assistance solutions. In order to prescribe a medicine, either a Doctor already knows all the details about it or he/she needs to spend some time searching for that information. Several books are at the Doctor’s disposal as well as Web sites, such as the electronic Medicine Compendium, which contains information about licensed medicines. Also independent software packages with extra search capabilities can be used. Despite all the available information, the provided search functionalities are usually based in keywords or class-oriented. With the goal of speeding up the prescription process, a prescription assistance data management system that allows searching for information about medicines can be thought of. The software requirements will have to be based on the following priorities: i) to understand the doctors’ prescribing needs; ii) to design a relational database that will store information about medicines; iii) to extract information about medicines from the Web and load it into the database. Organizations must be able to audit, track and report information in a timely and accurate way. The metadata repository is an excellent vehicle for adhering to these rules. With the repository, one can easily monitor the location and progress of bioinformatics data.

5. BIOINFORMATICS DATA MANAGEMENT-2 (WEB SERVICES)
Bioinformatics scientists can locate web services that implement the analysis relevant for their scientific experiments. Users will be able to perform basic operations such as checking the accessibility of the web service, invoking it using sample inputs, and browsing the results returned allows service providers to register and advertise their web services. Novice and expert bioinformatics web services users can also register or share web services or interact with service providers. This facility allows users to discover bioinformatics web services by keyword-based search or by browsing services in the catalogue by their functions, popularity, provenance, reliability. The Advanced search capabilities allows users to provide more information about the functionality of the target web service, the kinds of data it takes as input and/or the format of the results it delivers, its reputation, reliability and so forth. It is possible to couple the traditional “form-filling” search mechanisms with a modern facet-based “shopping” style web interface similar to Amazon technology. In the biopharmaceutical industry drug discovery protocols, lab results, results from clinical trials and a host of other information are archived, maintained and reproduced at any time. Distributed systems implement software to access heterogeneous databases that are dispersed over the internet.

6. BIOINFORMATICS DATA MANAGEMENT-3 (ALTERNATIVE MEDICINE)
Traditional Indian Medicine (like Ayurvedic medicine), which is a type of alternative medicine, is receiving growing attention from patients and bioinformatics researchers in other countries. In spite of this growing attention, this has not been included as part of standard care in many countries mainly due to a lack of scientific evidence for its efficacy and safety. In addition, many of the documentations about this are not available easily, creating a barrier to patients, scientists, and physicians. A database that captures knowledge about traditional herbs, including scientific evidence reported in scientific publications about the association of these herbs with diseases will be useful for bioinformatics data solutions. The goal is to integrate data from disparate sources of heterogeneous herbal information. Data integration assembles targeted data for bioinformatics analyses, and enables to discover scientific relationships between data. Most public repositories of herbal data focus on deriving and providing one particular type of data.

7. BIOINFORMATICS DATA MANAGEMENT-4 (SHARED SCIENTIFIC DATA)
One can explore a range of topics somewhat off the beaten track of classical data integration architectures, including graph-based modeling and integration, annotation, structure inference, data and work flows, and data integration for bioinformatics systems. The techniques for managing and understanding of shared scientific data, and experience in modeling, annotating and sharing of bioinformatics systems models will be important. Web services have gained importance as a means for packaging existing data and computational resources in a form that is amenable for use and composition by bioinformatics data management applications. The main issues that hinders the wide adoption and use of web services is the difficulty in locating the “appropriate” web service, i.e., the web service that performs the analysis the bioinformatics scientist is interested in.

8. BIOINFORMATICS DATA MANAGEMENT-5 (VIRTUAL DATABASE)
Due to explosive growth in the bioinformatics literature, researchers are confronted by the dual problems of information overload and information overlook. Text mining techniques
provide an avenue to manage this growth and overcome these challenges through enhanced semantic annotation and focused search and information extraction. A key issue for text mining is the ambiguity, variability, and specificity of terminology used in the bioinformatics domain. Researchers desire focused results relevant to their specific domain of study. To resolve these issues and achieve rapid adaptation to new biological domains, one has to develop general purpose text mining techniques and integrate information from diverse domain-specific structured data sources, such as databases and ontology. The integration of diverse large-scale data resources through workflows and text mining within a graph-based framework enables user-driven integration of Web Services. Due to the huge amount of information coming from diverse and heterogeneous sources, and to the different interests and purposes; the design and development of a data integration framework is very crucial to better cope with the current advance and challenges in bioinformatics. Such framework will allow the scientists to focus on the scientific analysis while the information management expert concentrates on the data management part. One needs consider the following steps to define a virtual (integrated) database: i) Define a virtual database, using any relational database, ii) Select the needed information from the different data sources, iii) Filter the data, iv) Apply reformatting, v) Decide whether to perform an update or a data replacement.

9. BIOINFORMATICS DATA MANAGEMENT-6 (INDEXING OF THE BIOINFORMATICS DATA COLLECTIONS)

Users start with information needs. Bioinformatics data solution is finding material (such as research articles on Bioinformatics data collections based on Indexing, and hypertext, etc.) of an unstructured nature that satisfies an information need from within large collections stored on server computers. The term “unstructured data” refers to data which does not have easy-for-a-computer structure. It is the opposite of structured data, the example of which is a relational database. In web search, the search is over a large number of publications stored on several server computers. The Web is unprecedented in many ways: unprecedented in scale, unprecedented in the almost-complete lack of coordination in its creation, and unprecedented in the diversity of backgrounds and motives of its participants. Each of these contributes to making web search complicated. Some aspects of parallel and distributed search in web-scale search systems will also be relevant. When computer is used to do linear scan through Bioinformatics data collections, it is referred to as grepping after the Unix GREP command. To avoid linearly scanning the collections for each query, it is required to index the Bioinformatics data collections.

10. CONCLUSION

Organizations managing bioinformatics data need a formal structure or framework. Without a framework, bioinformatics data management can easily get out of hand, leading to serious problems with data integrity and information dissemination. Source transaction systems generate, capture and store the data. Source transaction systems send data to a data integration hub. The integration hub is used to aggregate data from multiple sources, standardize and format the data if possible and then send it to a data warehouse. The data warehouse is a central repository of data. Data is stored permanently in the warehouse and accumulates over time. Data marts are used to extract specific subsets of information from the warehouse. The data can be tailored and designed for presentation to individual researchers. Business intelligence systems include data mining tools, statistical analysis tools, visualization applications, molecular modeling tools and a host of applications used by the bioinformatics researchers. The metadata repository maintains a record of data flow through the system. The combination of a centralized data warehouse and distributed, specialized data marts gives the ability to archive vast quantities of data and then disseminate it to specific groups. Data marts can be set up with predefined queries that extract only the most appropriate data for a research group. Database abstraction is the integration of multiple databases and database models into a single, unified view of the databases and enables full data integration for discovering novel relationships and patterns in bioinformatics networks. Put another way, database abstractions are virtual enterprise databases that are comprised of many real physical databases. Database abstraction software places a layer of software (middleware) between the physical distributed databases and the applications that will be viewing the data. This layer connects to the back-end databases using available interfaces and maps the physical databases to a virtual database model that exists only in the software. The application uses this virtual database to access the required information, and the database abstraction software handles the collection and distribution of the data as needed to the physical databases. The advantage of using database abstraction software is the ability to bind many different data types into one unified enterprise-wide data model, including schema and content. Thus, we're able to access information contained in many different data stores as a single database with a new schema, a database that we define in middleware to meet the needs of the problem domain. Thus, as the physical databases change, and the virtual exposed schemas need to change, and one only needs to make changes that are more about configuration and mapping than programming. The explosion of information made available to enterprise applications by the broad-based adoption of Internet standards and technologies has introduced a clear need for an information integration platform to help harness that information and make it available to enterprise applications.

11. FUTURE SCOPE

A text-centric search engine will not be able to do a good job with proteomic data in bioinformatics. Data-centric approaches are suggested for data collections with complex structures that mainly contain non-text data. The future of bioinformatics is
integration. For example, integration of a wide variety of data sources such as clinical and genomic data will allow us to use disease symptoms to predict genetic mutations and vice versa. The integration of GIS data, such as maps, weather systems, with crop health and genotype data, will allow us to predict successful outcomes of agriculture experiments. Another future area of research in bioinformatics is large-scale comparative genomics. For example, the development of tools that can do 10-way comparisons of genomes will push forward the discovery rate in this field of bioinformatics. Along these lines, the modeling and visualization of full networks of complex systems could be used in the future to predict how the system (or cell) reacts to a drug for example. A technical set of challenges faces bioinformatics and is being addressed by faster computers, technological advances in disk storage space, and increased bandwidth. Finally, a key research question for the future of bioinformatics will be how to computationally compare complex biological observations, such as gene expression patterns and protein networks. Bioinformatics is about converting biological observations to a model that a computer will understand. This is a very challenging task since biology can be very complex. This problem of how to digitize phenotypic data such as behavior, electrocardiograms, and crop health into a computer readable form offers exciting challenges for future bioinformaticians. Large volume of data and complexity in problem solving inspire research in data mining and modern heuristics. Data mining (i.e., knowledge discovery) is the process of automating information discovery. It is the process of analyzing data from different perspectives, summarizing it into useful information, and finding different patterns (e.g., classification, regression, and clustering). Many problems are difficult to be solved analytically in a feasible time. Therefore, researchers are trying to find search techniques or heuristics to get a good enough or satisfactory solution in a reasonable time. Microarrays are a new technology to investigate millions of genes simultaneously, which present new statistical problems because the data is very high dimensional with very little replication. Therefore, methods for multiple testing become very important. Microarrays offer an exciting entry point for statistical analysts into areas of bioinformatics. Evolutionary computation techniques are commonly used to address problems with large search spaces, where an exhaustive search is not applicable in practice. Due to this fact, many machine learning and optimization tasks have to be solved by using genetic algorithms (GAs) or evolutionary algorithms (EAs). Estimation and modeling problems often turn out to be intractable by standard numerical methods. An approach to solving the above problems consists in applying optimization heuristics such as evolutionary algorithms (simulated annealing), neural networks, genetic algorithms, tabu search, hybrid methods, etc. Combining neural network technology with genetic algorithms, statistics, and fuzzy logic have to be used to find optimal or near-optimal solutions for a wide range of problems.

REFERENCES
[2]. Jason TL Wang, Mohammed J Zaki, ... Data Mining in Bioinformatics - Book- 2005
[3]. Jiawei Han, Micheline Kamber…. Data Mining: Concepts and Techniques – Book – 2006
[4]. http://www.biomedcentral.com/1471-2105/6/34
[5]. Yi-Ping Phoebe Chen; Bioinformatics technologies; - 2005
[6]. Christoph W. Sensen; Essentials of genomics and bioinformatics - 2002
[7]. Ajit Kumar Roy, Niranjan Sarangi; Applied bioinformatics, statistics & economics in fisheries research- 2008
[8]. Bryan P. Bergeron; Bioinformatics computing- 2003
[9]. Mehmet M. Dalkilic; Data mining and bioinformatics: first international workshop ; ... – 2006
[10]. Sushmita Mitra, Tinku Acharya; Data mining: multimedia, soft computing, and bioinformatics – 2003
[12]. N. A. Kolchanov, Ralf Hofestädt, Luciano Milanesi; Bioinformatics of genome regulation and structure II: Volume 2 – 2006
[13]. Hui-Huang Hsu; Advanced data mining technologies in bioinformatics – 2006
[14]. Tomasz G. Smolinski, Marianna G. Milanova, Aboul-Ella Hassanien; Computational Intelligence in Biomedicine and Bioinformatics: ... - 2008