Abstract

Bioinformatics a subject which has emerged in phylogenetic tree of subject involves solving molecular biology problem with the use of computational methods. To aid researchers in obtaining, organizing and managing bioinformatics information, we have proposed and developed a model for information integration that utilizes URL based advanced data integration methods heterogeneous information. The central theme of the paper is information clustering methodology which enables users to integrate information which are pertinent to their research and developmental activities.

Keywords: Bioinformatics, Information Integration, Information Clustering, Data Mining, Web Usage.

Introduction

Bioinformatics, a term coined for the applications of computer science in biology is now emerging as a major element in contemporary biology. There is a paradigm shift in biological research as an extensive use the computers, software tools and computational models. Walter Gilbert, a renowned scientist, described this shift in biology as follows: “The new paradigm, now emerging, is that all of the "genes" will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture only then turning to experiment to follow or test that hypothesis.” Bioinformatics deal with the exponential growth in biological data has led to the development of primary and secondary databases of nucleic acid sequences, protein sequences and structures. These databases are available and hosted as web servers such as GenBank–Genetic Information, SWISS-PROT, Protein Data Bank (PDB), Protein Information Resources (PIR), along with many others related to Human Genome Project, Mouse, Bacteria and agricultural fields. These databases are available as public domain information and hosted on various Internet servers across the world. Basic research and modeling is done using these databases with the help sequence analysis tools like BLAST, FASTA, CLUSTALW, etc., and the modeled structures are visualized using visualization tools such as WebLab, MOLMOL, RASMOL, etc. As the network speed increases, the need has become greater to build integrated system that provides online access to databases. In addition to the heterogeneity and wide distribution of the molecular biology and bioinformatics databases a large number of software tools for sequence analysis, molecular modeling, structure prediction, database searching, similarity display, data display and reporting along with many others items of importance are coming up. Bioinformatics play an important role for the integration of broad disciplines of Biology to understand the complex mechanisms of the cell. Bioinformatics also aids the way in which biomedical investigators use the information in their testing. As the number of and size of molecular databases, bibliographic literature pertaining to these databases grows and their uses increases, the associated lacks expertise in the area of bioinformatics to choose right kinds of information. All people working in the area of bioinformatics relies on some how some or more popular search engine and Meta search engines giving the result when chooses for some popular keywords of bioinformatics. The results obtained in few lakhs and browsing through the navigational pages is another tedious job for obtaining the desired result. The story behind these is the scientists are overwhelmed by the range and volume of the bioinformatics resources available to them, and frequently lack the expertise to use them (Yarfitz and Ketchell, 2000). This range of research resources and continue to be expand and develop into complexity and numbers. So it is essential to understand the users about what information they need from the resources, how they search for information and how the information is used.

As the field of Bioinformatics is increasing there is
also an increase in the challenges in front of library and information center for providing quality services as and when desired by the bioinformatics scientists. Such challenges as elucidated by McMullen & Denn are related to structure, function and communication of molecules, different task involved in sequence analysis and protein structure and function prediction, heterogeneity of bioinformatics information, access platform, networking of information, awareness about the bioinformatics resources are to mention few. These issues are universal for each individual working in the field of bioinformatics. Increasing demand of information services and resource access of bioinformatics realizes the single platform access of information of the subject. In this paper we have discussed the link based information resources integration in the subject of bioinformatics known as iBIRA – \textit{Integrated Bioinformatics Resource Access}.

\textbf{The Technologies of iBIRA}

Each educational resource management centers relies on different categories of resources, ranging from print media to non-print materials, fee based as well as free and open access. The information is provided by in person as well as distant media via email, forums, discussion groups, blogging etc. As the popularity of internet is increasing there is also an increase in utilization and used as media of information transfer. Bioinformatics is a subject which relies on huge amount of biological data in the forms of nucleic acid structures, coding sequences; maps etc needed continuous changes in the data form. The information centers involved in bioinformatics research continuously working in this area for updating and revival of information every moment. People associated with bioinformatics may not be able to grasp all the information which has been recently changed or going to change as newer and newer projects has been launched and coming up in near future. The proposed iBIRA model which relies on link update for the information associated with bioinformatics. The iBIRA task model for web based bioinformatics information integration is represented here as a means of working with different forms of information. The objective of the model is to design, implement, evaluate and deploy a suite of tools and network based services that reliably improve the productivity of research in bioinformatics society by providing proactive information for solving problems related to bioinformatics activities. This is based on managing bioinformatics information using URLs crawling in which the information can be stored in the form of URLs for assembling and integrating information. The bioinformatics information represented in variety of component will link to the resources functioning in different web repositories. Fig.1 shows the different approaches of data fetching of bioinformatics and biological information.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{ibira.png}
\caption{Prototype of the iBIRA}
\end{figure}

The Bioinformatics as an application of information science finding its root in information communication of genetic codes involved prediction of structures of molecules in disease along with protein structures resulting in development of Biological databases. These analysis tools and structure prediction of molecules are studies with different software and analysis tools, and Institutions involved in such activities. Associations, forums, community centers, are playing important role in spread of bioinformatics knowledge. But at the same time library and information centers are also involved in sharing of information through different ways, subscription of journals, literature databases, citation reports, bibliographic services, to mention few. But over to it, there are still more and more issues yet to consider for immediate access of information on daily basis, through RSS feed, blogging, content access, email alerts, involving integrated access of information at single platform.
**Data and Metadata**
Most of bioinformatics information are stored at remote repositories and are implemented as databases. Additionally, bioinformatics information is stored in multiple and heterogeneous databases. Distributed information system for querying results over such databases is an essential component to facilitate information integration. In this work, bioinformatics information is accessed through an information interface, where the information is clustered into different categories of information components. This will enable users to retrieve multiple information at the same time and information can be visualized as an individual component, along with a service that is used to provide individual users with access to (i) correlated information at multiple subject categories; (ii) the results of the clustering will help to choose only that information which is very near to query term; (iii) distributed querying can interact further with the other related people over blogs, discussion forums, shares with bookmarking tools etc. The theme discussed above, is based on metadata that is structured according to ontologies. In biological sciences, it is also customary to create annotations in free text form. Such metadata contains invaluable information assembled by database curators. It has been provided with data sheet that is optimized with maximum standard information for the respective field.

**Conclusion**
Integration of molecular biology and bioinformatics databases and information resources requires a user friendly interface for coherent access of information. URL based information integration is one of the aspects which has been undertaken in this project to cluster heterogeneous information under one umbrella of iBIRA. The principle applied in this project can be further explored for the incorporation of more aspects of the information in the field of bioinformatics, as it is changing and reshaping rapidly and posing a challenge in front of library and information centers to provide task-based information services.

**References**