Mining the Information for Structure Based Drug Designing by Relational Database Management Notion

R. BALAJEE and M.S. DHANARAJAN*

Department of Bioinformatics, Sathyabama University, Chennai, India.

*Jaya College of Arts and Science, Thiruninravur- 602 024, India.

balajeer_2002@yahoo.com

Received 6 January 2009; Accepted 5 March 2009

Abstract: Structure based drug design is a technique that is used in the initial stages of a drug discovery program. The role of various computational methods in the characterization of the chemical properties and behavior of molecular systems is discussed. The field of bioinformatics has become a major part of the drug discovery pipeline playing a key role for validating drug targets. By integrating data from many inter-related yet heterogeneous resources, informatics can help in our understanding of complex biological processes and help improve drug discovery. The determination of the three dimensional properties of small molecules and macromolecular receptor structures is a core activity in the efforts towards a better understanding of structure-activity relationships.

Keywords: Cancer, Cancer genetics, Structure based drug discovery, Data model, Drug designing

Introduction

As the death toll from infectious disease has declined in the western world, cancer has become the second ranking cause of death rate, led only by heart disease. Current estimates project that one person in three in the United States will develop cancer, and that one in five will die of it. From an immunological perspective, cancer cells can be viewed as altered self-cells that have escaped normal growth regulating mechanisms.

The interaction between genes and environment plays an important role in cancer predisposition. Cancer results from both hereditary and non-hereditary factors. Under normal conditions, cells divide for various reasons such as replacement of aging tissue increased metabolic demands *etc.* However, when cell division is unchecked, the result is disastrous leading to the development of cancer. Cancer may thus be defined as an abnormal

proliferation of cells, which results from uncontrolled cell division. The three general mechanisms which cause cancer are

- To damage in DNA repair mechanism.
- To transformation of a normal gene into an oncogene.
- To loss of function of tumor suppressor gene.

Objectives

1. Develop information management data model for cancer
 - Proteome, structural, clinical and genomic data
2. Develop interface with the knowledge base
 - GUI based front end.
3. Develop efficient information retrieval from the knowledge base
 - Key word based retrieval
 - Query based retrieval

Data model

Data model will capture the data for proteome, structural, clinical and genomic data along with images, which can be used for subsequent analysis. It is object-oriented and provides navigational links between the different modules illustrated in the Figure 1. The significance of the data model is that it captures all the information related to the cancer. The data include not only textual information and reports but also structural and clinical types, which can be compressed, annotated and stored in the database. The annotation is hierarchical for representing information.

System architecture

Figure 1. System architecture.

User interface

The user interface includes both the presentation and query manager. The presentation manager along with query manager efficiently will present data to the user in a graphic environment. The presentation manager will provide support for browsing the data while the query manager will support advance query features to present the data along with its statistics. The query manager will provide a set of predefined operators and template query to assist in information retrieval and visualize the results.

Cancer – data (or) knowledge based mining

It consists of two types:

- (i) Automatic document clustering.
- (ii) Compiling classified information (or) Extract information.

(i) Automatic document clustering

It consists of document involving the task of identifying the document dealing with cancer information. Gathering the information and to form a group of tasks. Our document classifier system augments the performance of many tools available for performing this task.

(ii) Information extraction

Compiling the document to extract information, Text mining involves parsing documents of prespecified tasks that are written in programming language.

Description of the cancer data model

In the following section the conceptualization of the cancer data model is described. On the top of the class hierarchy we have the concept disease. All the diseases are categorized under this class. In future one can develop such data model for other diseases. This disease can be the main class and the entities, which share common features, can be designed as derived classes. Under the main concept disease, cancer becomes the first derived object. Data features common to all types of cancers can be organized under this class. Various types of cancers and their associated distinct features can be derived from this class.

Cancer disease

The genetic constitution of mankind hardly changes within a century. The changes take place among human populations in different parts of the world. The variation affects their geographical variation in the larger extent of environmental effects. The environmental effects may be exogenous or endogenous.

The background of human cancer has describes 15% caused by virus have to be taken with more than one gain of salt. This should be analyzed and provided a rough estimate of how much attention to be paid for the achievement to prevent a disease.

Breast cancer is a major lethal cancer for females in the Western world. A majority of cases have occurred in Postmenopausal woman. Only the least number of young are affected in this disease. The risk factors include exposure to estrogens, ionizing radiation, cigarette smoking and high fat diet. A study has revealed that 10-20% is ascribed to hereditary factors. The genes to be studied are BRCA1 and BRCA2 affecting certain receptors *etc.* Continuous monitoring of these genes with mutation will lead to potentially problematic and developed without controversies.

Cancer stages

Once the nature/type of the cancer is identified the stage of its maturation has to be determined which is generally divided into four stages according to size.

Stage 1: Tumor small in size ; Less than 2 cm in diameter ; No involvement of lymph nodes.

Stage 2: Size of the tumor increases up to 5 cm with or without the involvement of lymph nodes.

Stage 3: Tumor cells spread to auxiliary lymph nodes but not to other parts of the body.

Stage 4: The cancer cells spread to other parts of the body.

Risk factors

The clinical history, risk factors are important from clinical point of view. Cancer risk factor can be organized as shown below.

Figure 2a. Other diagnostic methods. **Figure 2b.** Other diagnostic methods.

Cancer therapy

Once the diagnosis is done the treatment regimen has to be decided based on the decision in diagnosis. Data management of cancer therapy is very important for clinicians. It includes a wide range of treatment procedures. Each method is defined as an object and the features of each method (defined as entities) corresponding are also defined. The data organization of cancer chemotherapy is as follows

Figure 3. Cancer therapy (1).

Figure 4. Cancer therapy (2).

Cancer genetics

The Cancer Genetics Network (CGN) is a research resource for investigators conducting research on the:

- Genetic basis of human cancer susceptibility,
- Integration of this information into medical practice, and
- Behavioral, ethical, and public health issues associated with human genetics.

Genetic polymorphism

Genetic polymorphism is the occurrence together in the same locality of two or more discontinuous forms of a species in such proportions that the rarest of them cannot be maintained just by recurrent mutation. It is sometimes called balancing selection, and is intimately connected with the idea of heterozygote advantage.

In genetic polymorphism the predisposition factors for cancers are defined and the data organization is shown below

Figure 5. Genetic polymorphism

Mutation detection methods

Identifying mutations that lead to cancer requires novel detection methods. Intensive research over the years has led to new detection techniques.

- Nucleotide sequencing
- DNA micro array technology

Advantages of such data models

- Efficient organization of data – Hence information retrieval is lot easier.
- The above framework when integrated with the patient information management will lead to efficient clinical data management.
- Clear statistics of treatment success and failure can be computed automatically.

The data type involved in drug designing cycle. **Figure 6.** Drug designing cycle.**Table 1.** Data types

Genome	1. Gene Function
	2. Single Nucleotide Polymorphism (SNP) Data
	3. Cloning
	4. Calculated Orthologs
	5. Gene Interaction
	6. Phenotype.
Proteome	1. Isoform
	2. Disease Association
	3. Physical Properties
	4. Protein-Protein interaction
	5. Post-Translational Modification
	6. Molecular and Cellular Function
Clinical	1. Drug, Xenobiotics effect
	2. Transgenic and Knockout animal
	3. Transgenic and Knockout Result (with references)
	4. Expression and Location of Organs
Structural	1. Fold, class
	2. PDB structure
	3. Nearest structural homolog & Structural Domains
	4. Sequence & Structural Neighbors
	5. Protein Ligand Interaction.

Collection methodology

The list of Public Domain Databases (PDD) is used to collect the information for above Data Types. List of Resources used to collect information for the different data types:

- Gene Function – National Center for Biotechnology Information (NCBI)
- SNP Data - National Center for Biotechnology Information (NCBI)
- Cloning - National Center for Biotechnology Information (NCBI)
- Calculated Orthologs - National Center for Biotechnology Information (NCBI)
- Gene Interaction – PUBMED, OMIM.
- Phenotype - National Center for Biotechnology Information (NCBI) Locus Link.
- Isoform - National Center for Biotechnology Information (NCBI)
- Disease Association – PUBMED, OMIM, GENCARDS.
- Physical Properties – Genatlas
- Protein – Protein Interaction - Pubmed
- Post Translational Modification – Through Databases like NCBI, Swissprot.
- Drug - National Center for Biotechnology Information (NCBI)
- Xenobiotics – National Center for Biotechnology Information (NCBI)
- Drugs & Xenobiotics Effect – PUBMED, OMIM, Gencards.
- Animal model - PUBMED, OMIM, Gencards.
- Expression in Organs – PUBMED
- Location of Expression – Gencards / Source.

Conclusion

The recent move towards the study has proved that structure based drug designing can be made through a bridge between bioinformatics and Cheminformatics approaches. The field of bioinformatics has become a major part of the drug discovery pipeline playing a key role in validating drug targets. By integrating data from many inter-related yet heterogeneous resources, informatics can help in our understanding of complex biological processes and help improve drug discovery. This follows in the form of genomics explosion and the huge increase in the number of potential drug targets with recent advanced techniques and approaches. However, to expand the business in pharmaceutical industry and to implement the drug discovery approach in Biotechnology industry the study of genomics and Proteomics method is essential. Infact, the incorporation of Information technologists played subsequently. In the next 5-10 years, there will be a definitely boom in the market based on bioinformatics and Cheminformatics approach to drug designing. The above steps are used in drug designing which make a bridge between bioinformatics and Cheminformatics approaches.

References

1. Agrawal R, Imielinski T and Swami A. Database mining: A performance perspective IEEE Transactions on Knowledge and Data Engineering, 1993, **5**(6).
2. Gehrke J, Ramakrishnan R and Ganti V Rainforest, A framework for fast decision tree construction of large datasets. Proc Int Conf on Very Large Data Bases (VLDB), 1998.
3. Moore A W and Lee M, *J Artificial Intelligence Res.*, 1998, **8**, 67-91.
4. Du Mouchel W, Volinsky C, Johson T, Cortes C and Pregibon D, Squashing flat files flatter, ACM Int Conf on Knowledge Discovery and Data Mining, 1999.

5. Ross B J, Gualtieri A G, Fueten F and Budkewitsch P, Hyperspectral image analysis using genetic programming, Proceedings of the Genetic and Evolutionary Computation Conference, (GECCO'02), 2002, 1196.
6. Sarawagi S, Thomas S and Agrawal R, Integrating association rule mining with relational database systems, alternatives and implications, ACM SIGMOD Int Conf on Management of Data, 1998.
7. Shafer J, Agrawal R and Mehta M, SPRINT, A scalable parallel classifier for data mining, Proc 22nd Int. Conf on Very Large Data Bases, 1996.
8. Carlos Ordonez, Programming the K-means Clustering Alogrithm in SQL, Proc ACM Int Conf on Knowledge Discovery and Data Mining, 2004, 823-828.
9. Augen J, *Drug Discovery Today*, 2002, **7**, 315-323.
10. Jun Xu and Arnold Hagler, *Molecules*, 2002, **7**, 566-600.
11. Graham Richards W, *Pure Appl Chem.*, 1994, **66**(8), 1589-1596.