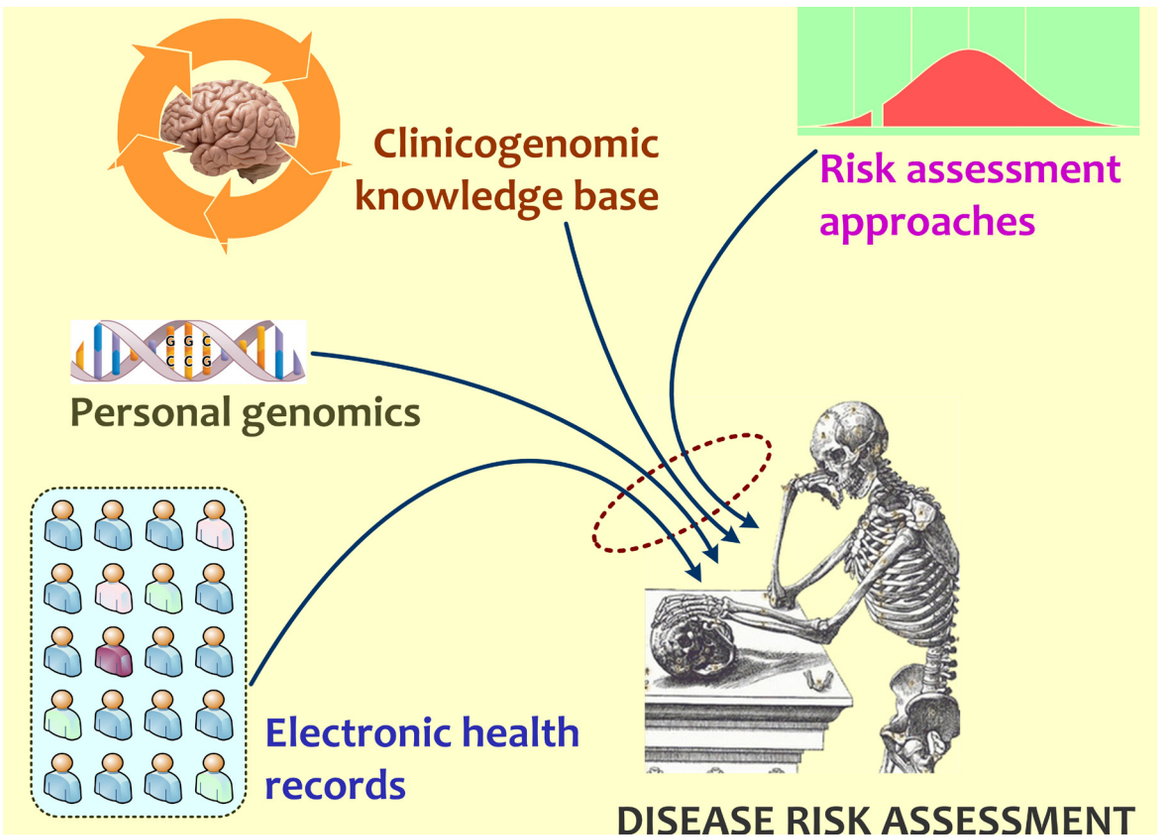
ISSN 2249-894X
ORIGINAL ARTICLE
REVIEW OF RESEARCH

# NECESSITY OF OLAP FOR UNEARTHING ALL KNOWLEDGE OF A MULTIDIMENSIONAL GENOMIC DATA




**AAKRITI SHRIVASTAV**
DataSol Consultancy, Hill Road, Nagpur

**ABSTRACT:-**

The enormous amount of genomic data maintained in the databases is often not utilized to its full potential. This is largely because of the technological as well as non technological factors. Also, the underdevelopments of methods for analysing huge amount of data are also a big problem. In view of this difficulty efforts are needed to develop new techniques and technologies that can be widely used so that all the data available can be utilized for new knowledge generation. To achieve this objective though there are few options, Online Analytical Processing or OLAP is one of the promising technique that can be used for data management as well as processing. OLAP helps user to easily and selectively extract and view data from different perspectives. In view the above; literature published in the reputed journals has been reviewed in this study. The literature review has been carried out by following

deductive reasoning methodology and is presented in chronological order. The results of this study indicate that different researchers have taken different approach to manage the information and improve quality of data and provide specific method to represent the data analysis results. Overall, it is evident from the literature that OLAP offers a meaningful option, especially for managing genomic data.

**KEYWORDS:** Genomic data, techniques and technologies, Online Analytical Processing or OLAP, data management

## 1.0 INTRODUCTION

The surge in the quantity of genomic data has opened numerous avenues by which one can extract information to for generating varied type of knowledge. Besides, the latest development in the field of genetics has resulted in enormous increase in the speed of data generation. Coupled with these technologies the information and communication technology has also provides a platform for easier sharing of this information. Like other data sets, the genetics data also has inherent variation. Moreover, the data generation often takes place vis-à-vis specific objectives, which may alter during the course of a particular study. Although, many online and offline tools are available to analyze genomic data sets, there exists a void, especially, when the multidimensionality of the data sets as well as the objectivity is concerned. Hence, in view of this, it is necessary that newer techniques be studies for their possible use in the field of genetics so as to acquire new knowledge as early as possible. One such technique is online analytical processing or OLAP, which enables a user to easily and selectively extract and view data from different points-of-view.

Such a multidimensional analysis (using OLAP) solution can provide researchers with the insight needed to make more informed decisions and build better models for predictions. By consolidating summarized genomic information from volumes of heterogeneous data and presenting this data to users in a meaningful concept, besides, multidimensional analysis offers great potential for improving and coordinating decision-making across the varied fields of research. OLAP is valuable because of its flexibility. Once the facts and dimensions are defined within the OLAP server, OLAP tools provide an easy way to analyze data by simply dragging and dropping dimensions and facts into the appropriate locations. In the backdrop of above information, the research efforts pertaining to the relational and OLAP data sources and its integration with different applications have been reviewed in this study.

## 2.0 Methodology

In the backdrop of the aims and objectives of this study, the literature published in standard journals and form reputed sources was collected. The process of review was based on the standard method. For the purpose of the review aspects like the research question being posed, theoretical background, methodology used, findings, conclusions, etc. were considered. The discussion is presented in a chronological order, so that it also indicates the underlying pattern of evolution of thoughts and ideas in the selected domain i.e. use of OLAP in different fields. The discussion on the basis of the review is presented in the following section.

## 3.0 Discussion

It has been reported (more specifically in the corporate sector related literature)

that OLAP allows users to analyse database information from multiple database systems at one time. While relational databases are considered to be two-dimensional, OLAP data is multidimensional, which means the information can be compared in many different ways. For example, a researcher might compare their results/finding of genomic data of one organism with that of other characteristics, which might be stored in a different database. Hence, in order to process database information using OLAP, an OLAP server is required to organize and compare the information. Researchers can analyze different sets of data using functions built into the OLAP server. In view of its powerful data analysis capabilities, OLAP processing is frequently used for data mining, which aims to discover new relationships between different sets of data. Thus, in view of its utility, the literature has been specifically reviewed and the discussion is presented hereunder

Dayhoff et al., (1994) have reported about the work related to developing, testing and evaluating the benefits of physicians' workstations as an aid to medical data capture in an outpatient clinic setting. In this study, the authors described the uses of the physician's workstation vis-à-vis data handling and use with an emphasis on the data collection cost recovery process with respect to the above mentioned workstation. In addition to this Hettler et al., (1997) suggested initiation of work with OLAP in the healthcare set ups, which will open a world of opportunity for data-mining across the enterprise, which can aid in lowering the healthcare costs. Though content-based visual information retrieval has been one on the most vivid research areas in the field of computer vision, Müller et al., (2004) have stated that no general breakthrough has been achieved with respect to large varied databases with documents of differing sorts and with varying characteristics.

Gene expression databases contain a wealth of information, but current data mining tools are limited in their speed and effectiveness in extracting meaningful biological knowledge from them. Coleman et al., (2004) on the basis of their studies reported that multidimensional analysis tools are a practical step toward actually using the varied information that is already being captured in various systems. Furthermore, the authors stated that it is relatively easy to extend the existing system to incorporate other pertinent data, thereby enabling the ongoing analysis of other aspects of the workflow. In addition to above, Alkharouf et al., (2005) reported that OLAP can be used as a supplement to cluster analysis for fast and effective data mining of gene expression databases. Authors found that compared to traditional cluster analysis of gene expression data, OLAP was more effective and faster in finding biologically meaningful information.

Furthermore, Wang et al., (2005) have reported that data warehousing and online analytical processing technologies can be applied to integrate as well as mine the biomedical data. However, they reported that the task of capturing and modelling diverse biological objects and their complex relationships is the main difficulty to achieve this. In view of the problems encountered in the healthcare sector, Parmanto et al., (2005) noted that the typical multidimensional data warehouse designs that are frequently seen in other industries are often not a good match for data obtained from healthcare processes. Hence, they suggested that multidimensional data analysis should be explored for such problems. In addition to above, Trissl et al., (2005) reported that integration of databases can help in unearthing the real and full potential of the protein data. The multidimensional data processing is particularly important as the information pertaining to the structural and functional aspects of proteins based on certain properties of the proteins, such as sequence features, fold classification, or functional annotation is stored in different databases.

Kaya et al., (2005) proposed a novel multiagent learning approach to handle the

huge data related problems. Their approach is based on utilizing the mining process for modular cooperative learning systems, which incorporates fuzziness and OLAP based mining to effectively process the information reported by agents. Daumer et al., (2007) has stated that variations in the disease course can be studied in details if the available data can be processed using an OLAP tool. Given that the OLAP has a potential to address multiple queries at a time, the development of it (OLAP) is very essential to explore the health related data in general and that of genetics in particular. Similarly, Mangalampalli et al., (2007) have also proposed the use of electronic health records for quick, reliable, secure, real-time and user-friendly information. Parmanto et al., (2008) have reported that the potential of multidimensional data analysis should be explored to discover new knowledge, which was not previously planned. Kehl et al., (2008) explained multidimensional indices and hierarchical relationships between related types of data and that could be integrated within our data warehouse. The authors have stated that OLAP is a mature technology in the financial sector, but it has not been used extensively for scientific analysis. Furthermore, Satpute et al., (2008) reported that the format or semantics of the grammar can also help in faster retrieval of the data, especially when a new algorithm is used.

Prevedello et al., (2008) proposed a novel concept of Business Intelligence for managing the information available in data repository, which could benefit from the ability to integrate and visualize data (e.g. information reflecting complex workflow states) from all of their imaging and information management systems in one composite presentation view. To achieve this OLAP appears to the ideal system. Ebidia et al., (2009) described the process of developing OLAP capacity from data generated in an on-line transaction processing system and stated that if used and managed appropriately, OLAP has tremendous potential for meeting data visualization demands.

Linking genotypic and phenotypic information is one of the greatest challenges of current genetics research and in view of it Nuzzo et al., (2009) presented the design and architecture of the Phenotype Miner, a software system able to flexibly manage phenotypic information, and its extended functionalities to retrieve genotype information from external repositories and to relate it to phenotypic data. Authors reported that a comprehensive, integrated and automated workbench for genotype and phenotype integration can facilitate and improve the hypothesis generation process underlying modern genetic studies.

Bernier et al., (2009) stated that OLAP, which is central to the field known as Business Intelligence, a key field for such decision-support systems, is now being used in different domains. Besides, the variations of basic OLAP along with spatial analysis have resulted in SOLAP (Spatial OLAP) and an altogether new research area. Moreover, Youssif et al., (2010) have reported that at the moment number of various formats are used for medical images, which are created, transmitted, and analyzed, for different types of purpose. Besides, all the data is not stored in one place, which demands that a multidimensional analysis be carried out.

Raistrick et al., (2010) designed a framework upon the Web Model-View-Controller-based architecture in which the reusable and extractable models can be conveniently adapted to other hospital information systems and which allows for efficient database integration. Lin et al.,(2011) established a real time online health and decision support system with the novel information technology integrating modelized architecture and Web services for clinical infometrics and concluded that its implementation can help the researchers to get more and more data for different objectives.

## 4.0 Conclusions

Biomedical research is now generating enormous amounts of data, ranging from clinical test results to microarray gene expression profiles. Besides, the scale and complexity of these datasets gives rise to considerable challenges in data management and analysis. Hence, the new systems are needed to address the data management and analysis challenges. In the backdrop of literature reviewed in this chapter, it is evident that different approaches are tried by different researchers to manage the huge quantity of data available in the molecular biology domain, however absence of a specific, standardized method for data utilization, (especially data that is present in digital form) warrants that newer frameworks, approaches should be developed to recap benefits from the accumulated data. For this, a concept such as OLAP is available, which is provided by a number of vendors and can work with any relational database management system. The utility of OLAP is well documented by different authors, however, given its importance, there appears a lot of scope for research in multidimensional data analysis. Furthermore, the benefits of OLAP allow the database to support multiple levels of analysis, which is imperative for appropriate decision making. Compiling such sets using current web resources is tedious because the necessary data are spread over many different databases and hence, an online OLAP can help the researchers in a big way. Thus, it is concluded that OLAP shows great promise for the dynamic data analysis for e.g. that of protein analysis for bioengineering and biomedical applications. In addition, OLAP may have similar potential for other scientific and engineering applications involving large and complex datasets, which should be explored in future studies.

## 5.0 References


• Dayhoff, R., Kirin, G., Pollock, S., Miller, C., & Todd, S., (1994). Medical data capture and display: the importance of clinicians' workstation design. Proc Annu Symp Comput Appl Med Care., pp 541–545.

• Hettler M., (1997). Data mining goes multidimensional, Healthc Inform., 14(3): 43-46, 48, 51-56.

• Ebidia A., Mulder C., Tripp B. & Morgan MW.,( 1999). Getting Data Out of the Electronic Patient Record: Critical Steps in Building a Data Warehouse for Decision Support. Proc AMIA Symp., pp 745-749.

• Muller H., Michoux N., Bandon D., & Geissbuhler A., (2004 Feb). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inform., 73(1) :pp 1–23,

• Coleman R M., Ralston M D., & Szafran A.,(2004 September). Multidimensional Analysis: A Management Tool for Monitoring HIPAA Compliance and Departmental Performance, BS journal of digital imaging,, 17(3), pp 196–204.

• Alkharouf N W., Jamison D C., & Matthews B F., (2005). Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases, Journal of Biomedicine and Biotechnology, 2005(2): pp 181–188.

• Wang L., Zhang A., & Ramanathan M., (2005). BioStar models of clinical and genomic data for biomedical data warehouse design, Int J Bioinform Res Appl., 1(1): pp 63–80.

• Parmanto B., Scotch M., & Ahmad S., (2005). A Framework for Designing a Healthcare Outcome Data Warehouse, Perspectives in Health Information Management., 2: pp 3.

• Trissl S., Rother K., Müller H., Steinke T., Koch I., Preissner R., Frömmel C., Leser U.,(2005), Columba: an integrated database of proteins, structures, and annotations , BMC

Bioinformatics., 6: pp 81.
• Kaya M., & Alhajj R.,(2005). Fuzzy OLAP association rules mining-based modular reinforcement learning approach for multiagent systems, IEEE Trans Syst Man Cybern B., 35(2): pp 326-38.
• Mocanu C., & Mocanu M., (2007). Electronic medical record--interface specifications with medical informatics systems, Oftalmologia., 51(4): pp 14-19.
• Daumer M., Neuhaus A., Lederer C., Scholz M., Wolinsky J S., & Heiderhoff M., (2007). Prognosis of the individual course of disease - steps in developing a decision support tool, Multiple Sclerosis for the Sylvia Lawry Centre for Multiple Sclerosis Research BMC Medical Informatics and Decision Making,, Volume 7: pp 11.
• Mangalampalli A., Rama C., Muthiyalian R., Jain AK.,& Parinam AM.,( 2007). High-end clinical domain information systems for effective healthcare delivery, Int J Electron Healthc., 3(2): pp 208-219.
• Parmanto B., Paramita MV., Sugiantara W., Pramana G., Scotch M.,& Burke DS., (2008). Spatial and multidimensional visualization of Indonesia's village health statistics, Int J Health Geogr., 7: pp 30.
• Kehl C., Simms AM., Toofanny RD.,& Daggett V.,( 2008). Dynameomics: a multi-dimensional analysis-optimized database for dynamic protein data, Protein Eng Des Sel., 21(6): pp 379-386.
• Satpute S., Katkar V., & Sahare N, (2008). Data Extraction of XML Files using Searching and Indexing Techniques, World Academy of Science, Engineering and Technology 39. [Available at www.waset.org/journals/waset/v15.php]
• Nuzzo A., Riva A., & Bellazzi R.,( 2009). Phenotypic and genotypic data integration and exploration through a web-service architecture, BMC Bioinformatics,, 10(Suppl12): pp S5.
• Bernier E., Gosselin P., Badard T., & Bedard Y., (2009).Easier surveillance of climate-related health vulnerabilities through a Web-based spatial OLAP application,International Journal of Health Geographics., 8: pp18.
• Prevedello L M., Andriole K P., Hanson R., Pauline Kelly, & Ramin Khorasani, (Apr 2010). Business Intelligence Tools for Radiology: Creating a Prototype Model Using Open-Source Tools , journal of digital imaging., 23(2): pp 133–141.
• Youssif AAA., Darwish A.A. & Mohamed R.A.,( March 2010). Content based medical image retrieval based on pyramid structure wavelets. IJCSNS., 10 :pp 3.
• Raistrick C A., Day I N M., & Gaunt T R., (October 2010). Genome-Wide Data-Mining of Candidate Human Splice Translational Efficiency Polymorphisms (STEPs) and an Online Database, PLoS ONE, 5 (10) e13340.
• Lin H C., Wu H C., Chang C H., Li T C., Liang W M., & Wang J Y W.,( 2011). Development of a real-time clinical decision support system upon the web mvc-based architecture for prostate cancer treatment, BMC Medical Informatics and Decision Making., 11:16. doi:10.1186/1472-6947-11-16