BioFactHMM: MULTIDIMENSIONAL MODELING OF BIOLOGICAL DATA FROM HIDDEN MARKOV MODEL GENERATED DATASETS

Article *in* Indian Journal of Computer Science and Engineering · August 2020

DOI: 10.21817/indjcse/2020/v11i4/201104264

3 authors:

Manas Ranjan Pradhan
Skyline University College
20 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Beenu Mago
10 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Deepak Kalra
Skyline University College
18 PUBLICATIONS **6** CITATIONS
SEE PROFILE

BioFactHMM: MULTIDIMENSIONAL MODELING OF BIOLOGICAL DATA FROM HIDDEN MARKOV MODEL GENERATED DATASETS

Dr. Manas Ranjan Pradhan
Associate Professor, School of Information Technology, Skyline University College
Sharjah, United Arab Emirates
manas.pradhan@skylineuniversity.ac.ae

Dr. Beenu Mago
Assistant Professor, School of Information Technology, Skyline University College
Sharjah, United Arab Emirates
beenu.mago@skylineuniversity.ac.ae

Dr. Deepak Kalra
Associate Professor, School of Information Technology, Skyline University College
Sharjah, United Arab Emirates
deepak.kalra@skylineuniversity.ac.ae

Abstract-The ever growing biological research generates large volumes of biological data and knowledge bases ranging from clinical test results to genome analysis. The dynamic changes of genome sequences and complexity of these database and their relations have given lot of challenges to data analysis. There are many online databases are available for biological studies. It is essential that biological data can be analyzed in multidimensional way creating data warehouse and then online analytical processing. The method of multidimensional modeling, star schema is not sufficient for biological data as it cannot cater more relationships. The Snowflake schema though helpful in better relations among datasets than star schema but cannot model all data from all databases specially the hidden states of long new biological sequences or complex medical data. Looking at above scenario, the idea mentioned in this paper combined the efforts of generating datasets by HMM (Hidden Markov Model) from all types biological databases available online and use Fact Constellation schema of data warehouse modeling. Hidden Markov Model has adopted in this study to find newly datasets and help in analyzing relations between these datasets. Once the data sets generated the fact constellation schema of multidimensional modeling done for making data warehouse. Henceforth new proposed model in this work is called BioFactHMM schema specially proposed for biological data which is a mix of star and snowflake schema. This model desires to capture all semantics of bio sequence from various data sources using HMM. Then data warehouse modeling is done with design principles of Fact constellation schema. Subsequently, the analysis technique of OLAP cube is done to view the data and reports in a multidimensional way.

***Keywords**:* HMM, Multidimensional, Genome data, Biological data, Data Warehouse, Data Modeling, Fact Constellation, Biological databases

1. Introduction

The genome sequence analysis is done in various perspectives depend on the species, gene and protein. The analysis also done by many biology professionals with respect to problem areas of genome profiling and root cause analysis of diseases. There are three basic categories of biological databases are available depend on structure, sequence and other types of mix. The sequence based genome databases such as EMBL, GenBank and DDBJ are more popular because of DNA storage. [Pradhan,2019]. These databases predict information about biological functions. Due to huge storage of data, it is needed to use computational tools and techniques. Data can be extracted, processed and view in different format. In order to find meaningful information from data in many perspectives, the multidimensional modeling in fact necessary to analyze the data from various data sources. These data sources and information lies in them change dynamically. While we do sequence analysis, the HMM (Hidden Markov Model) which discuss in section 5 is used to generate data to facilitate multidimensional modeling. This paper focuses mostly on multidimensional modeling of genome sequences that uses HMM for various depository of molecular databases and subsequently analysis them using OLAP.The rest of the paper organized with section 2 literature review. Section 3 represents fundamental data warehouse modeling types (Star, Snow-flake, Fact-

Constellation). Section4 mentioned a structural diagram of various types biological databases. Section 5 describes the fundamental HMM process and various types of HMM used for generating sequence dataset from a genome space. Section 6 illustrate the proposed model and its stages. Section 7 highlighted the results and comparison with other models. Section 8 represents the challenges of dynamic genome repository. Section 9 mentioned conclusion and future scope of this research work.

2. Literature Review

The genome sequence analysis using HMM can generate many new sequences. As a fundamental principle of HMM the observables have been identified regularly. The new genome sequence test dynamically and produce markov chains. The hidden data sometime does not reflect while analyzing the huge database system of genome [Pradhan,2019]. So the use of HMM is essential to find all observables from hidden states considered as part of data sets in this study. As numerous datasets generated in genome or any biological data, the computational technology like Data warehousing technology is necessary to analyze and view data to meet the requirements. A data warehouse is defined as "a subject-oriented, integrated, non-volatile and time-variant collection of data in support of management's decisions" [Inmon ,1996]. The commonly used data warehouse models are star schema, snowflake schema and Fact-Constellation schema. The biological data modeling also has been taken into consideration all these types of schema. The conventional model of Star schema successful mostly in business context but seems to be not enough to build biological data. Looking at the need of biological data a specific model proposed called as BioStar schema [Wang and Ramanathan, (2005)]. In difference to star schema, snowflake schema contains more data object relation in a hierarchical manner. The adoption of snowflake schema mentioned in protein annotation model called COLUMBA. It is a unified protein annotation database and relevant data sources narrate proteins with appropriate dimension [Rother *et la*., (2004)]. COLUMBA takes 32 tables in a global schema. Irrespective of the above research work done by different authors, the data generally accessed through ETL (Extract, Transfer, Load) process into a multidimensional database. These data further convert to smaller data marts as per the need by filtering or aggregating. The analysis is done depending on the collection of DataMart. Online analytical processing (OLAP) is a technique to view data in the warehouse which allows many analytical queries on the data [Cabibbo and Torlone, (1998)]. The OLAP summarizes and aggregates data and find information in many dimensions [Marcel, (1999)]. There are two types OLAP operation called as roll-up and drill-down. Roll-up summarizes detailed data for the next higher level of a dimension whereas Drill-down navigate to detailed data of lower-level. OLAP also does pivoting, slicing and dicing on data cube [Vassiliadis, (1998)]. Though OLAP have been successfully applied to the business domain, but comes up with challenges when applied to biology [Dubitzky *et la*., (2001)]. The biological data warehousing is complex as deals with global living systems in contrast to business data. The great challenges are to capture, model and encode data for biological knowledgebase [Wang and Ramanathan, (2005)]. A conceptual model has developed by Markowitz and Topaloglou based on multi-dimensional analysis of microarray gene expression data [Markowitz and Topaloglou, (2001)]. This model is based on annotation and expression of gene. The use of star or snowflake schemas in this model were not detailed and not considered many clinical and genomic data. Pedersen *et al*. tried the clinical data to develop extended star-schema model [Pedersen *et al*., (2001)]. Though few problems of clinical data modelling were discussed and considered, this extended data model was not comprehensively covered microarray gene expression and other genomic data. Reviewing the above highlighted literature from various authors, this research work is aligning to propose its BioFactHMM model which is based on HMM and Fact constellation schema of multi-dimensional modeling.

3. Multidimensional Data Model

The data stores in a Multidimensional data model in the form of data cube. Generally three-dimensional cubes find in most of the cases. Users can view data multidimensional using data cube. The results are analyzed from various biological data sources in different ways and attributes are tracked and arranged accordingly. Data are extracted, transferred and load from many data sources to make a multidimensional database [Inmon, (1996)]. Then DataMarts are formed, which is basically smaller data stores as per user needs by aggregating or classifying data. The queries related to biological data can be answered easily from the multidimensional databases by OLAP.Users can view and analyze data from different dimensions which will be again discuss in later point of this paper.

There are typically three types of Multidimensional Models.

3.1. *Star Schema*

The star schema looks like a star structure. The table forms here are called facts and dimensions. The fact table located at center of star and other dimension tables join to it. Fact table is formed by taking key attributes from each dimension table. The advantage of this schema is that very easy to model but disadvantage in context to biological data is that it can incorporated only small number of tables. It is not so much flexible for analytical needs as it is a purpose oriented schema for particular view model and not allowing complex analytics. It does not support much relations among data tables as the cardinalities does not show many to many relations. The databases used are mostly not normalized in Star schema which means data integrity is not enforced properly. Fig1 describes below about star schema model.

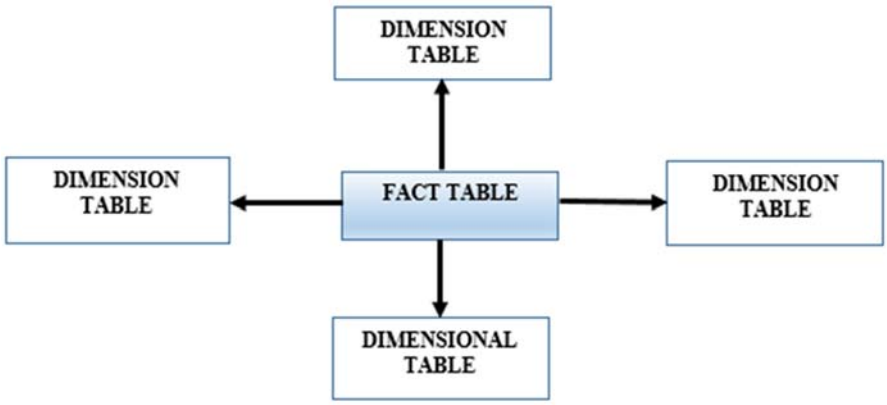

Fig 1. A Star Schema [Kimball and Ross , (2013)]

3.2. Snowflake Schema

The snowflake schema is differing from star schema with more number of normalized dimension tables. One-dimension table links with its subsidiary dimension table. The fact table is centrally located and connected to multiple dimension tables and subsequently to other sub tables so that redundancy can be reduced. In context to biological data this model also lost some of the hidden state sequence which may not be in normalized table. The fact table that generated may not occupied all the sequence states under study. The additional levels of attribute normalization add complexity to relations and joining. There is a significant poor performance in query processing when it deals with the joins to view in different dimensions. While doing operation like update and insert, data loads in this schema must be highly controlled and managed to avoid anomalies. Fig2 below shows the SnowFlake schema model.

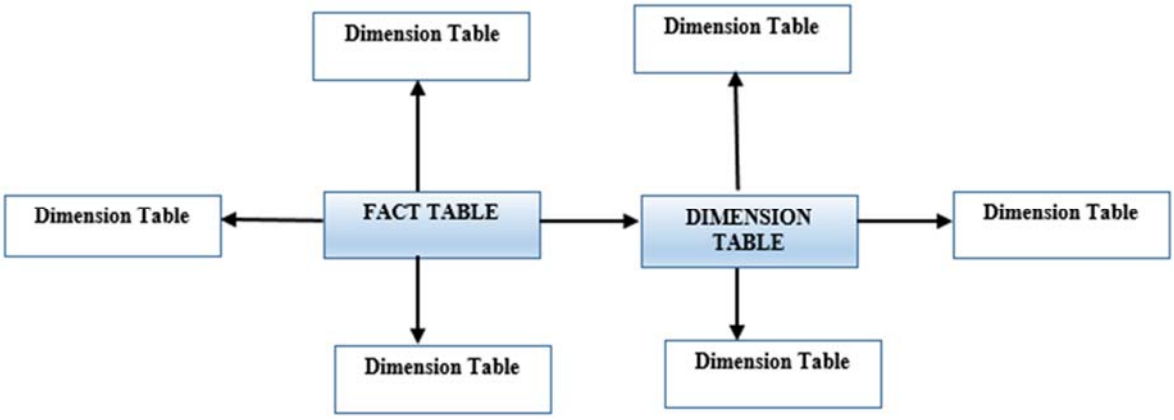

Fig 2. Snow-Flake Schema [Kimball and Ross, (2013)]

3.3. *Fact-Constellation Schema*

This type of schema has many fact tables and connected to many dimension tables. It is a combination of Star and Snowflake schema. The dimension tables are common to fact tables. Most of the disadvantage of start schema and snow-flake schema removed while running this schema. The relations among dimension tables occurs in one to many as well as many to many form. This schema takes the advantage of most normalized databases. Fig3 below shows a fact constellation schema model. The proposed model in this paper adopted Fact constellation schema. The Online Analytical Processing (OLAP) will be used based on this schema for data extract and view.

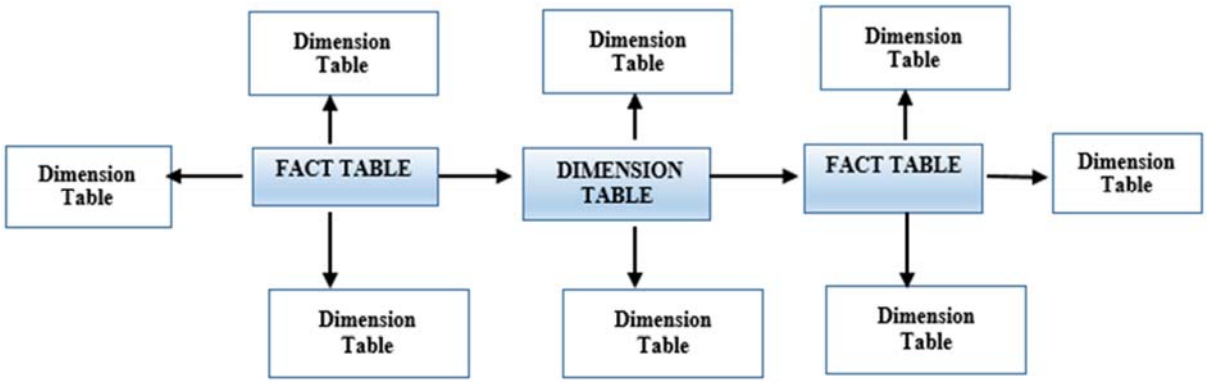

Fig 3. Fact Constellation [Kimball and Ross, (2013)]

4. Biological Databases

Biological databases have information about genome and medical data in the form of bio-libraries which have been collecting from various sources such as biological experiments, bio-literature and computational results [Toomula *et la*., (2011)]. A biological database is generally large and software used online to store, update, query, and retrieve data for specific biological question.

The new biological data collected and processes from researchers and professionals all over world. These databases get updated in real time [Toomula *et la*., (2011)]. The publicly available biological databases help the scientific community to search, study and analyze the data. [Pradhan, (2019)]

The classification of Biological database in a boarder sense as shown in below in Fig4.

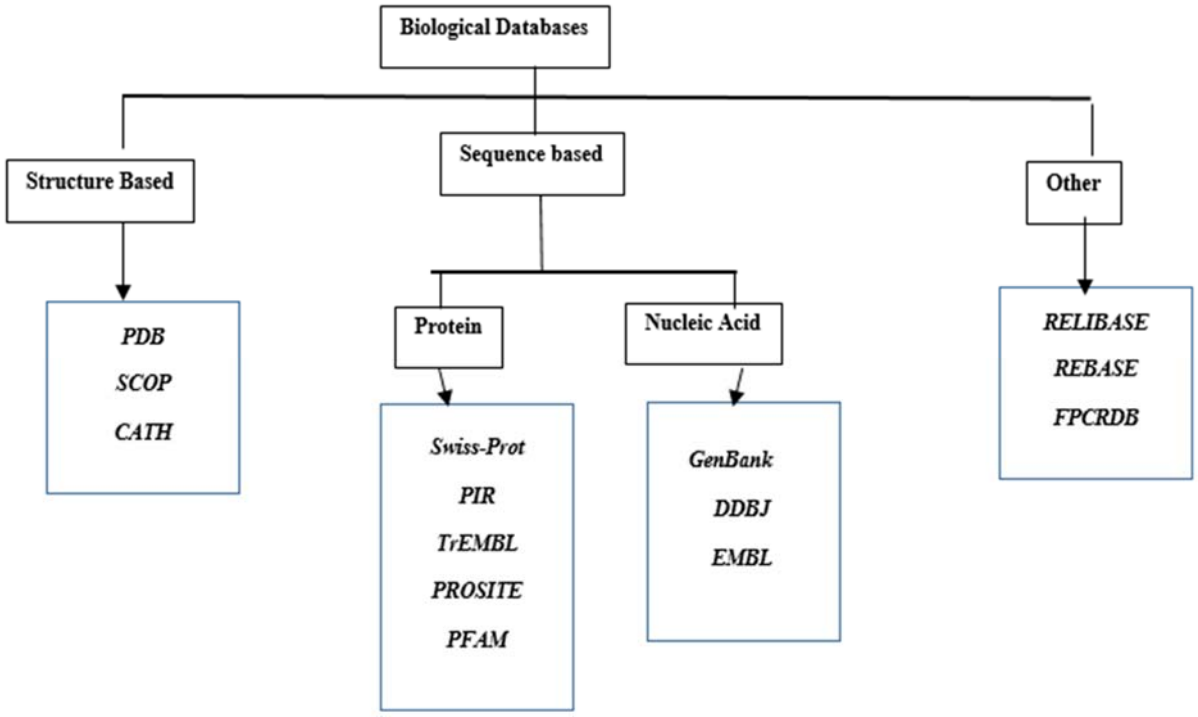

Fig 4. Biological Databases [Toomula *et la*., (2011)]

5. Hidden Markov Model (HMM)

The HMM is a statistical model mostly applicable to genome sequence data. It is based on markov chain formation to solve many alignment problems. This model facilitates operations of insertions and deletions [David, (2004)]. It works on the probabilistic functions of Markov chain states. All states of a sequence including hidden states can be tracked and used in this model [Rabiner, (1989)].

The HMM is represented as a triplet: $\lambda = (A, B, \Pi)$

A - the state transition probability matrix-Aij $= P(q_{t+1=j} \mid q_{t=i})$ ----------(1)

B- observation probability distribution- $b_j(k) = P(O_t = k \mid q_t = j)$, $i \leq k \leq M$ ----------(2)

Π - the initial state distribution

where,

N - Number of States-$Q = (q_1, q_2, q_3, \ldots . q_t)$

M - the number of Symbols (observables)-$O = \{O_1, O_2, O_3, \ldots \ldots O_t\}$

The three types of HMMs are described in the literature of Byung-Jun Yoon are as below [Pradhan, (2019), Yoon, (2009)].

5.1. *Profile-HMMs*

This type of model has been used extensively for profiling biological sequences as classifying protein or finding motifs [Pradhan,2019]. The libraries generally used here are: the PROSITE database and the Pfam database. The additional homologous belongs to same family can be found from given profile-HMM which is needed for relation to be maintained in a fact constellation schema of data warehouse preparation.

5.2. *Pair-HMMs*

This type of model manages pairwise sequence alignments of DNA and protein [Pradhan, (2019)]. The technique of progressive alignment helps in finding larger datasets by multiple alignment. This feature allows in constellation fact table to find the relation between dimensional tables very easily. Generalized pair hidden Markov model (GPHMM) may use to create fact tables. The pair-HMM can be used for complex sequence, structures though it is generally applicable for linear sequence [Yoon, (2009)]. These features help in finding numerous data tables that can link with fact table.

5.3. *Context-Free HMMs*

This type of model is based on contextual data and their useful correlations [Pradhan, (2019)]. This feature will help in creating data tables which have taken contextual attributes and later on take a vital part in fact table to analyze the results for OLAP.

6. Proposed Model: BioFactHMM

The proposed model BioFactHMM specially designed for biological data. This biological data structures available in many forms such as number, textual, image, tabular or XML structures [Toomula *et la*., (2011)].

A BioFactHMM schema is a triplet $T = H_s, D_i, F_i$ where H_s is the HMM generated sequence sets, D_i is a dimension data sets and F_i is a dataset for Fact table identity. The Fact dataset is generated from Dimension data set.D_i contains set of attributes (A_j). F_i contains all the primary keys derived from D_i.

6.1. *Stage-1 –HMM Produces Datasets*

The first stage of this model is to create as many as data sets using HMMs.The data that reflect in Data sets can be mix of numbers, text, image etc. In other way structured, unstructured as well as semi-structure data may reflect on the data tables. Fig 5 below shows how HMM integrate to find datasets from biological data.

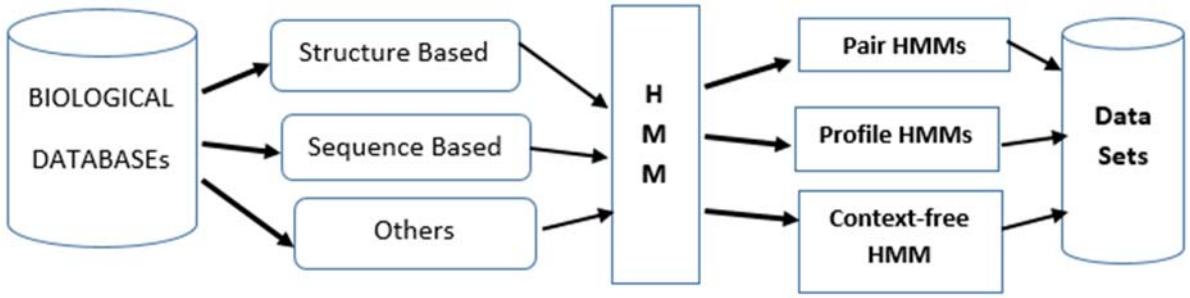

Fig. 5 Process of creating of datasets using HMM

The data tables can be formed from datasets. The data table is considered to create fact tables which will easy for practioners and researchers to see the relations. The genome data generally very complex and have many dimensions [Wang *et la*., (2005)]. The cardinality of relations find between fact-table to dimension-table or in between dimensional tables are generally man-to many type. The uncertainty in these relations will be taken care by HMM.The states that have created using HMMs take all patterns and relations into consider. The uncertainty factor will come less as even hidden states are taken into consideration. The complexity of incomplete data and wrong annotations can be tracked as the data structure has been created in data table is in understandable form.

Fig.6 shows how dataset(observable) for data tables can be generated without any state missing from genome sequence using a fundamental Markov chain.

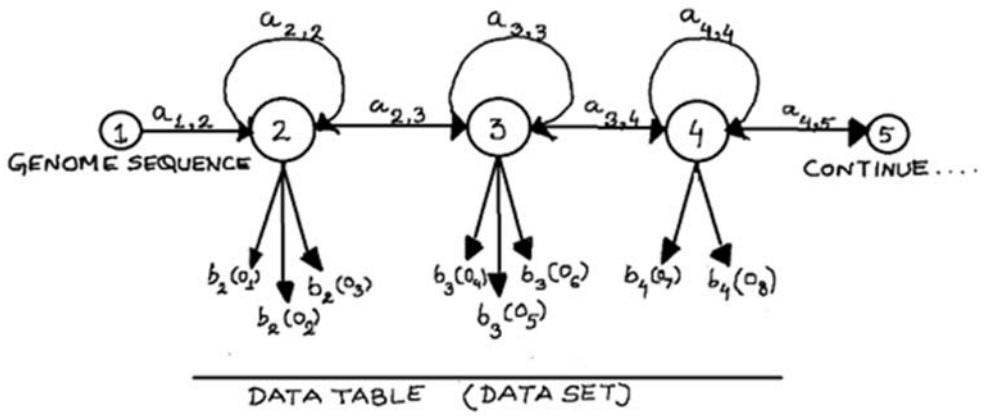

Fig 6. Datsets from Genome sequence using HMM

Many data sets which will use for dimensional table and can be analyzed as below .These represents H_s (HMM generated sequence states). Below tables depicts datasets from Observables.

Table 1: Dataset 1

$b_2(O_1)$	**$b_2(O_2)$**
$b_2(O_2)$	**$b_2(O_3)$**
$b_2(O1)$	**$b_2(O_3)$**

Table 2: Dataset 2

$b_3(O_4)$	**$b_3(O_5)$**
$b_3(O_5)$	**$b_3(O_6)$**
$b_3(O_4)$	**$b_2(O_6)$**

Table 3: Dataset 3

$b_4(O_7)$	**$b_4(O_8)$**

6.2. *Stage 2-Formation of Dimension tables*

The above data sets help in forming dimensional tables which basically available in the form of 1,2,3 or more dimensional forms. These represents D_i (Dimension tables). The below table 4 shows the formation of multiple dimension tables. Likewise, many data tables can be formed.

Table 4 : Multiple Dimension tables

Dimension Table1
$b_2(O_1)$
$b_2(O_2)$

Dimension Table2
$b_2(O_1)$
$b_2(O_2)$
$b_2(O_3)$

DimensionTable3
$b_3(O_4)$
$b_3(O_5)$

DimensionTable4
$b_3(O_4)$
$b_3(O_5)$
$b_3(O_6)$

DimensionTable5
$b_4(O_7)$
$b_4(O_8)$

6.3. *Stage-3-Creation of Fact table*

The fact table can be created by selecting common data items from data tables (dimension tables). As shown below the b2(O1), b3(O4) and b4(O7) are the key data points (in database concept called as primary key) which is connecting data tables (dimension table1, dimension table 2 and dimension table3). Fig7 below shows the creation of a Fact table.

For example

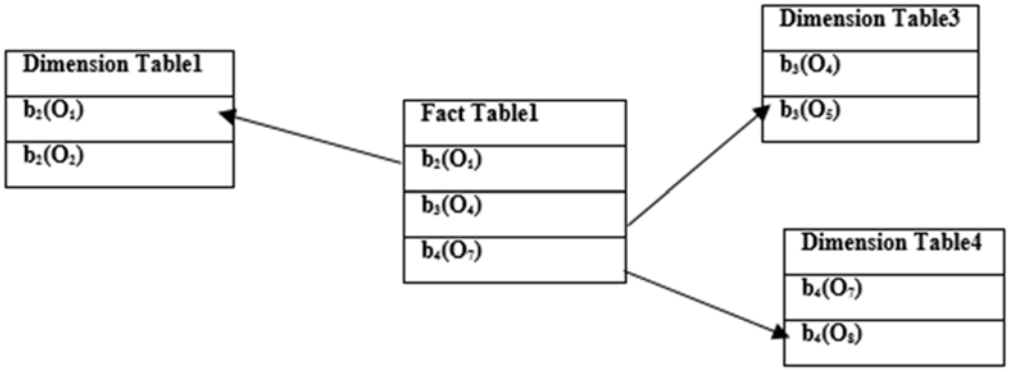

Fig 7. Creation of Fact Table

6.4. *Stage 4-Creation of Fact Constellation*

When we implement Fact Constellation schema in multidimensional model. It is a collection of many fact-tables means we need to create fact table1, fact table 2, fact table 3…etc. having some common dimension tables (e.g Dimension table1, Dimension table2). Many literatures mentioned it as schema of Galaxy. Due to its capacity of holding many datasets relations, it is used for biological data warehouse design and analysis. Below Fig.8 is an example for creating Fact Constellation.

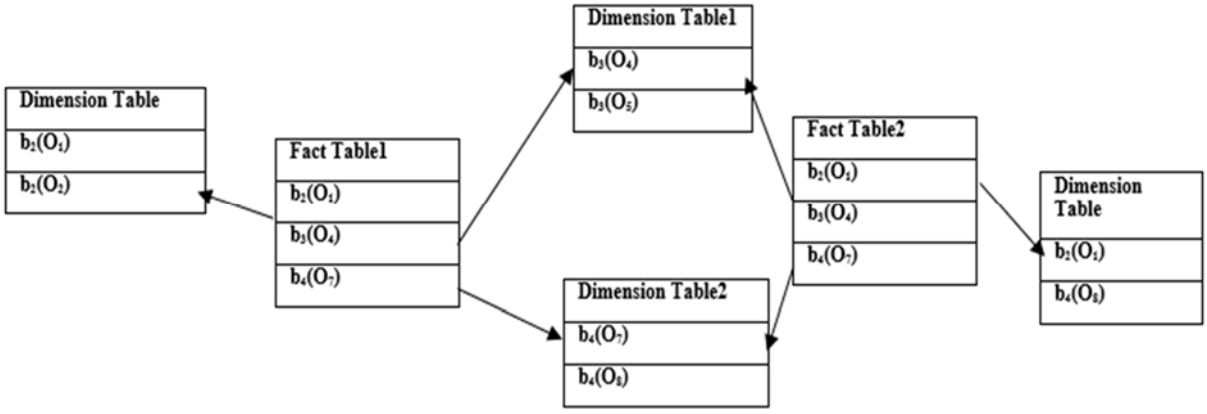

Fig 8. Fact Constellation Example

An example of fact constellation to genome data has shown in fig.9 below.

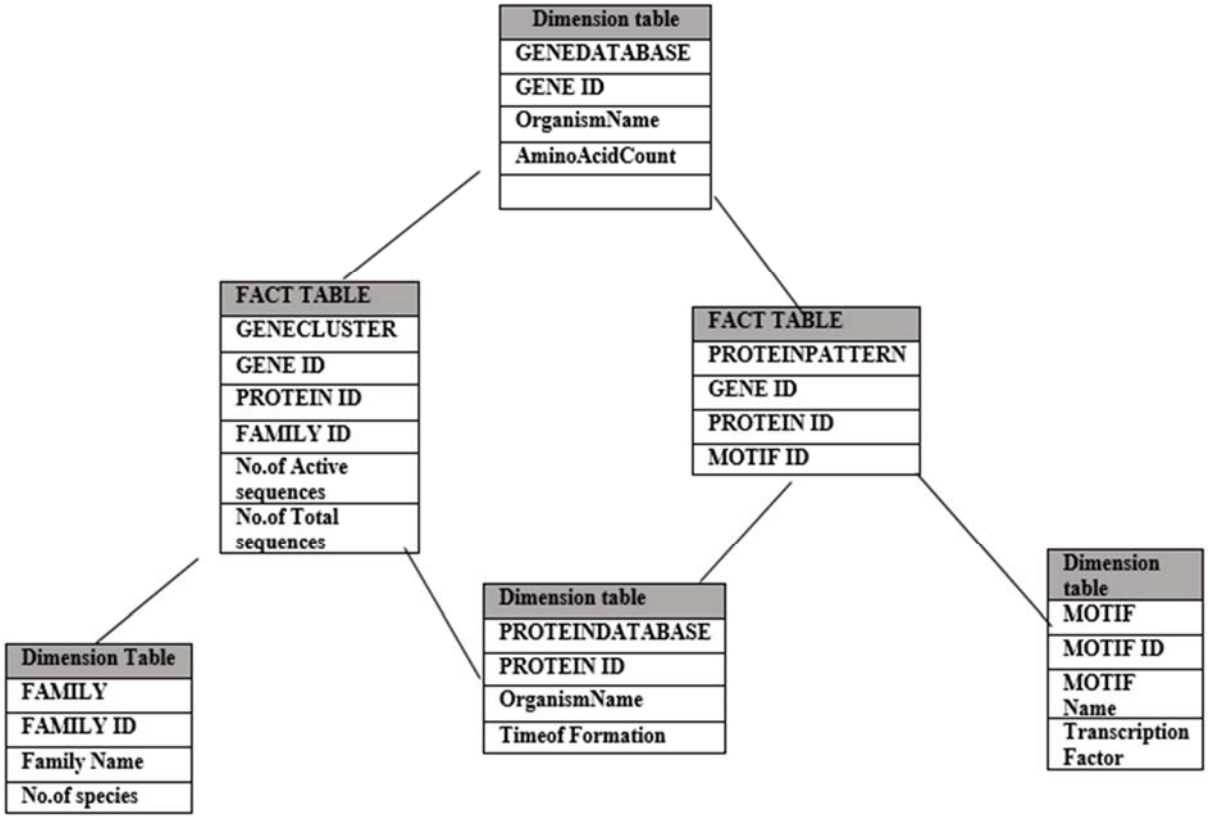

Fig 9. Example of Fact Constellation –Genome data

6.5. *Step 5- OLAP (Online Analytical Processing)*

OLAP is used for intelligence report generation. For business context, it helps to discover data, view unlimited report, make analytical solutions and predict "what if" conditions. It is basically a software that facilitates decision making and reporting from a data warehouse [Alkharouf *et la*., (2005)]. In biological context it used in multidimensional analysis of medical data and genome data. The online genome data can be analyzed very efficiently in multidimensional analysis. Some of the fundamental analysis queries are: Which of my targeted genome data most likely to go for identification of gene? What dataset of protein biggest impact on diseases? What is the gene pattern, motif patterns have been giving new results over a period of time? Next when we wish to run these kinds of many queries depending on different types of databases used, OLAP categorizes into two types of OLAP.One is M-OLAP (Multidimensional OLAP) and R-OLAP (Relational OLAP). OLAP tools does not store data in row-by-column format instead multidimensional database structures—known as Cubes. Analysts can slice and dice data on a Cube to produce a worksheet kind of view. Analyst need to group data in many ways for report generation.

Fig.10 shows that Data warehouse generally form by ETL operation on the data that has taken for study. Once it creates data warehouse with data sets, the data tables can be segregated to dimension table and fact table depend on what the user wishes to analysis. The OLAP cube set up with slicing and dicing operation to visualize the targeted result analysis.

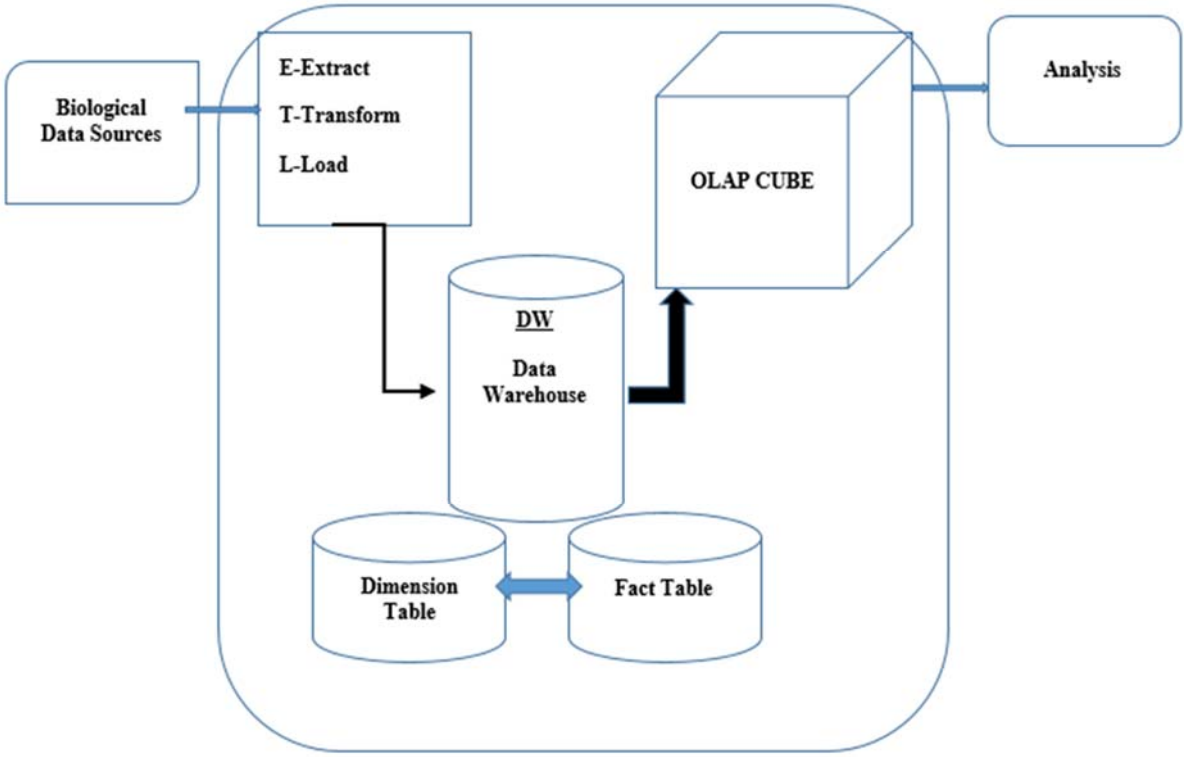

Fig 10. Data Warehouse Architecture [Alkharouf et la., (2005)]

The below Fig.11 shows a typical example of slicing and dicing inside an OLAP cube. The genome slice has made with genome channel and target objects. Inside genome channel the data sources are structural database, sequence database and other types of biological database. The database regions PDB, PROSITE, EMBL and RELIBASE have taken to analyze the gene, protein, motif and transcription factor.

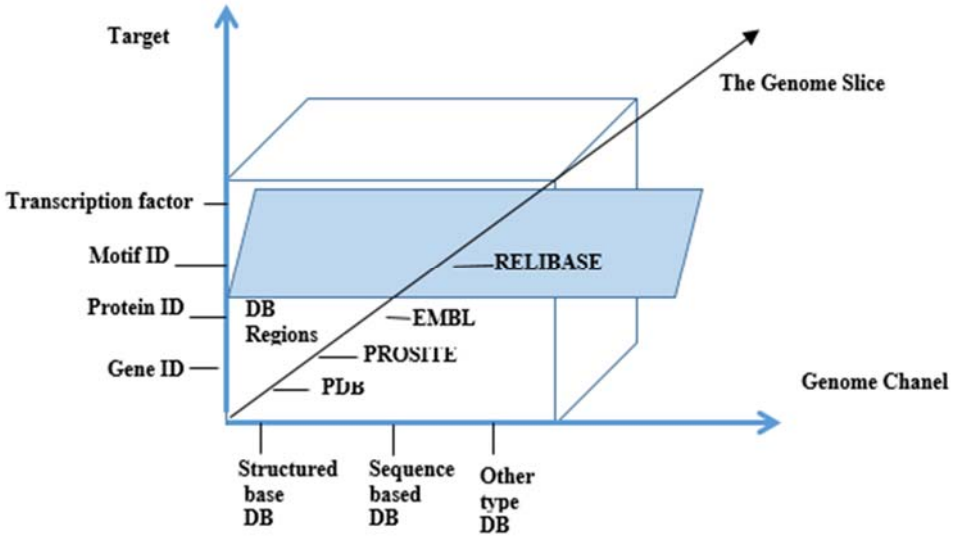

Fig 11. Slicing & Dicing on Genome data

If some prior knowledge about biological dataset known to user, OLAP enable to customized report about any scientific queries of interest due to its elastic nature [Alkharouf *et la*, (2005)]. Thus OLAP is an automatic analysis rather than manual which most biologists follow.

7. Results and Discussion

Due to overgrowing as well as dynamic changes in biological data, it is very much need to adopt multi-dimensional models. There is minimal research work available on this subject of interest. Markowitz and Topaloglou proposed a conceptual model for microarray gene expression data using star or snowflake schema [Markowitz and Topaloglou , (2001)]. This model is not so much detailed and not have taken clinical and genomic data [Markowitz and Topaloglou, (2001)]. Pedersen *et al*. tried to develop a model under extended star schema and has taken few cases clinical data [Pedersen *et al*., (2001)]. But it is not so efficient to cover microarray gene expression and other genomic data [Pedersen *et al.,* (2001)]. Wang *et al.* in their research work detailed a new model by taking star schema dedicatedly for clinical and genomic data and named it as BioStar [Wang *et al*., (2005)]. The data warehouse in BioStar integrate biomedical datasets related to human diseases. The researchers collected clinical data, image data, drug responses etc. from patients to be part of BioStar data warehouse [Wang *et al*., (2005)]. This model includes data from genome sequences & annotations, microarray gene expression and protein related information from public databases [Wang et al., (2005)]. Though their work is represented quite comprehensively and provided a framework of biomedical data warehouse, but the dynamic behavior of data generation and ever growing relations of databases have not noticed while proposing the star based model. Taking into consideration of above mentioned research works from different authors, the BioFactHMM model proposed in this paper have mentioned about HMM generated datasets where not a single data could be missed from genome space knowledge repository. The HMM will help to dynamically regenerate the probably datasets regularly. The Fact constellation schema of data warehouse has integrated so that the dimensional tables will link many fact tables at a time. As a matter fact the robust relation between dimension tables give the analytical aspects of OLAP better results. The OLAP generates all possibly results as more databases and knowledge bases can be taken into consideration.

8. Challenges of Genome Depository

This research work mentioned in this paper analyzes the multidimensional modeling of biological data.HMM helps in getting datasets from various depository of bio-molecular databases. The fact constellation schema tried to model data with maximum clarity. But still the challenge of genome repository and dynamic update of data gives a challenge to biological community. The accuracy and accessibility of data across all types of databases is difficult as data grow exponential. It is difficult for structural data (image based) to feed into dataset to form dimensional table and hence complex to analyze on multidimensional model. Though OLAP technology tries to bring the real time data processing, but the stored data in genome space may be incorrect or incomplete due to large scale submission of information [Lathe *et la*., (2008)]. Duplication and anomaly are other factors associated with redundancy complexity. The repository contains data which are neither annotated nor organized properly which gives a challenge to access and utilize right information for the researchers.

9. Conclusion and Future Scope

This paper tried to project HMM that can be useful for fact constellation schema of multidimensional modeling of biological data. The proposed model BioFactHMM explained various stages of creating the schema. The OLAP analysis of slicing and dicing highlighted with an example. The future work will be based on experimental analysis using computer programming language Python to establish the model in a Hadoop big data environment scenario. There are also dynamic changes in data in genome repository and uncertainty will be studied. The data modeling complexity of structure based data will be detailed in future scope. The efficiency of the BioFactHMM will be verified with other existing models.

References

[1] 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 & biotechnology, 2005(2), 181–188. https://doi.org/10.1155/JBB.2005.181

[2] Cabibbo L, Torlone R. (1998),” Querying multidimensional databases”, “Proceedings of the 6th International Workshop on Database Programming Languages, pp. 319–335.

[3] David Mount (2004), “Alignment of Pairs of Sequences,” Chapter 3, in Bioinformatics: Sequence and Genome Analysis, 2nd edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.

[4] Dubitzky W, Krebs O, Eils R. (2001),” Minding, OLAPing, and mining biological data: towards a data warehousing concept in biology”, Proceedings of Network Tools and Applications in Biology (NETTAB), CORBA and XML: Towards a Bioinformatics Integrated Network Environment, pp. 78–82.

[5] Inmon W H. (1996) “Building the Data Warehouse”,2nd Edition. John Wiley & Sons, New York.

[6] Kimball Ralph and Ross Margy,(2013) ,The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling ,3rd Edition,Wiely publication ISBN: 978-1118530801

[7] Lathe, W., Williams, J., Mangan, M. & Karolchik, D. (2008), “Genomic Data Resources: Challenges and Promises”. Nature Education Vol1, No.3, pp.2

[8] Marcel P. (1999),” Modeling and querying multidimensional databases: an overview”, Networking and Information Systems Journal, Vol 2, No5-6, pp.515–548

[9] Markowitz VM, Topaloglou T. (2001),” Applying data warehouse concepts to gene expression data management”, Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering (BIBE), pp. 65–72.

[10] Pedersen TB, Jensen CS, Dyreson CE. (2001),” A foundation for capturing and querying complex multidimensional data”, Information Systems,26(5):383–423.

[11] Pradhan, M. R (2019), "Genome Sequences analysis using HMM in Biological Databases",2019 International Conference on Digitization (ICD), Sharjah, United Arab Emirates, pp.272-275, doi: 10.1109/ICD47981.2019.9105756.

[12] Rabiner, L. R. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286.
[13] Rother Kristian, Müller Heiko and Trissl Silke (2004) “Columba: Multidimensional Data Integration of Protein Annotations”, In: Rahm E. (eds) Data Integration in the Life Sciences. DILS 2004. Lecture Notes in Computer Science, vol. 2994. Springer, Berlin, Heidelberg.
[14] Toomula Nishant, Kumar Arun, D Satish Kumar and B Vijaya Shanti (2011),” Biological Databases- Integration of Life Science Data”, Journal Computer Science & Systems Biology ,4(5): 087-092. doi:10.4172/jcsb.1000081
[15] Vassiliadis P. (1998), "Modeling multidimensional databases, cubes and cube operations," Proceedings. Tenth International Conference on Scientific and Statistical Database Management (Cat. No.98TB100243), Capri, Italy, pp. 53-62, doi: 10.1109/SSDM.1998.688111.
[16] Wang, L., Zhang, A., & Ramanathan, M. (2005),” BioStar models of clinical and genomic data for biomedical data warehouse design”, International journal of bioinformatics research and applications, 1(1), 63–80. https://doi.org/10.1504/IJBRA.2005.006903
[17] Yoon Byung-Jun (2009),” Hidden Markov Models and their Applications in Biological Sequence Analysis”, Current Genomics.; Vol.10 No. (6), pp. 402–415.