The Journal of Animal & Plant Sciences, 25(2): 2015, Page: 337-345
ISSN: 1018-7081

A REVIEW: CONCEPTUAL DATA MODELS FOR BIOLOGICAL DOMAIN

M. Idrees[1] and M. U. G. Khan[2]

Department of Computer Science and Engineering[1,2]
Bio-informatics Lab, Al-Khwarizmi Institute of computer Science[1,2]
University of Engineering and Technology, Lahore, Pakistan.
Corresponding Author: midrees10@gmail.com

ABSTRACT

This paper demonstrates the survey and review of conceptual data models and the novel data modeling techniques of biological data. The term conceptual data modeling is used in broad categories in this sense. The biological data, its concepts and frameworks have diversity of expressiveness under the umbrella of bioinformatics. If we consider the biological data a single field of research, it is not possible to handle all these things efficiently and completely. For provision of highly maintainable and efficient solutions, which will have less cost and complexity, we must reduce its scope by making its sub domains in bioinformatics. Keep in mind the aforementioned reasons, we considered only the concept of central dogma of molecular biology; produces sequence biological data (DNA, RNA and protein structures); to describe this reviewed study of conceptual modeling. Our objectives are to provide a current state of art study of conceptual data models for a sequence biological data. Based on this research, we will propose a uniform data model for biological data for unification purposes. In this review paper, we provide the analysis and post-mortems of existing conceptual biological data models, and present their comparison, provided on the basis of conceptually proposed methodologies, Meta data, modeling methods and other critical aspects, necessary for sequence data. This study provides us the cutting edge for the integration of biological data.

Key words: Bioinformatics, Conceptual data model, Biological data, Central dogma.

INTRODUCTION

Bioinformatics is a broad domain that enables to encompass the use of biological data and its management with the help of information technology. It is a multi-disciplinary field used to involve the fields of molecular biology, computational biology, neuroscience, statistics, and information technology; and is interlinked to many other disciplines as well as described by Zhou and Song (2005). Mainly, it entails the creation and advancements of biological databases, statistical analysis techniques, computational algorithms, theoretical and formal problem solving approaches and the management of biological data. Human Genome Project (HGP) described by Jagadish and Olken (2006); Mouse Genome Project (MGP) reported by Topaloglou (2004) that sequencing automated tools have accelerated a large amount of diversities of biological data and created a lot of interest in it for research highlighted by Topaloglou (2004); Jagadish and Olken (2006).

Three biological objects DNA, RNA, and protein structures which play an active role during our lives, are considered to be the fundamental pillars of prokaryotic and eukaryotic organisms as discussed by Idrees and Khan (2014a). Prokaryotic organisms that may consist of single cell and the eukaryotic organisms contain multiple cells in their bodies. The cell is considered the most important and basic unit of organisms, either they are prokaryotic or eukaryotic explained by Shapiro (2009). Prokaryotic cells are much smaller and simpler than eukaryotic cells. These cells are being constructed and destructed in entire life of the organism through metabolism reactions. The cells are enclosed themselves through cell membranes in the bodies of organisms. There exist millions of biological constituents and anti-bodies in organisms. In our bodies, approximately 60 trillion cells have been discovered which contain 320 diversities of cell types explained by Cohen (2004). Nucleus of a cell contains all types of genetic information, in the forms of chromosomes among them. Due to these chromosomes, the formation of human-beings, its inherited characteristics, working, monitoring, controlling and functioning of each and every part is made possible reported by Kari *et al*. (2001); Cohen (2004).

As shown in the figure 1, central dogma is used to present the information flow among biological entities. The three entities named as DNA, RNA and Protein structures, play an important part for creation of sequence biological data. DNA structure provides basis for building blocks and encodes information for living organisms. DNA molecule consists of four nitrogen bases named as adenine (A), cytosine (C), guanine (G) and thymine (T). These bases make base pairs among each other during construction of DNA molecules as discussed by Idrees *et al*. (2014b). The difference between DNA and RNA is that it contains uracil (U) instead of thymine (T) in RNA structure. A similar type of base pairing

exists in RNA molecules. Mostly RNA molecule exists in single structure or strand. It is unstable molecular structure, because as soon as DNA is converted into RNA, immediately it is translated into protein structures, due to the formation of amino acids. Protein structures are most occurring organic compounds available in the nucleus of a cell in rigid, strong and hard forms described by Idrees and Khan (2014a). They comprise 50% overall dry weight of the cell and are available in different shapes and sizes, most likely in forms of bones, tendons, nails, tissues, muscles, and fibrous. Different types of protein structures get involved in primary, secondary, tertiary and quarterly structures. They form the structural parts of bodies and changing their shapes in a regular way inside the cells. Protein structures are made up of 20 different types of essential amino acids, those folds into unique three dimensional structures. These amino acids are used as a building block of different types of protein structures discussed by Kari *et al*. (2001); Idrees *et al*. (2014b).

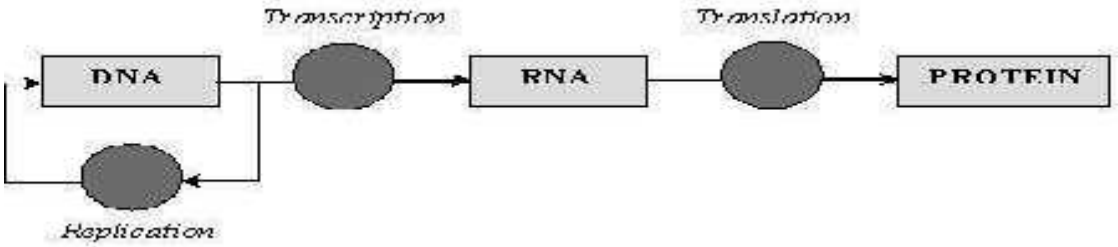

Figure 1. Central Dogma of Molecular Biology

Sequencing projects, omics technologies and biological entities are contributing and producing uncontrollable data in the current era continuously, which have been creating much more hurdles for research community due to its management explained by Cohen (2004); Topaloglou (2004); Jagadish and Olken (2006); Nair (2007); Huacarpuma (2011); Tenazinha (2011). This production is categorized as a structured, semi-structured or unstructured data. Mostly, it consists of semi-structured nature, because it does not have static schema due to its diversity of formats discussed by Keet (2003a).It is mostly available in forms of sequences, structures, vectors, scalars, graphs, images, equality, inequality, conventional and unconventional sort of queries (similarity query, sequence similarity, pattern matching and finding). Ubiquitous inconsistencies and uncertainties that lies in biological data, have a dire need for its curation, provenance, and for its integration reported by Cohen (2004); Topaloglou (2004); Ghalayini *et al*. (2006); Jagadish and Olken (2006); Graves (2012). As we analyze the nature of biological data, we generally come up with three main characteristics that differentiate it from other types of data, namely; complexity, heterogeneity, and highly dynamic nature highlighted by Keet (2003b); Shah *et al*. (2009).

Data modeling for biological data needs flexibility, dynamism, and power of expression due to its complex structures, substructures, huge amount of data types, and irregular and hierarchical relationships between different classes of biological species described by Ghalayini *et al*., (2006); Macedo and Porto (2007); Mesiti (2009); Tenazinha (2011); Graves (2012). Traditional data modeling approaches mostly used only for storage purposes and are not so good for defining the complex structure of biological systems from anatomical point of view and from system's functionality point of view, which is the most basic requirement for biological data. In addition, if a change occurs then that change could invalidate some information already maintained by the system, or in worst case, invalidate the whole system. Therefore, as the biological data nature evolves with time, there is always a change in the value of the data as well as in the structure that describes the data. So, traditionnel data models are not good in handling the schema evolution situations. Whenever there is a change, old value is replaced with the new one, making a problem for users to refer to the previous one as reported by Ghalayini *et al*. (2006); Shah (2009); Graves (2012).

Some systems are also modeled using the class based approach. This approach overcomes the problem of defining the complex structure of biological systems; but it also lacks in providing diversity of data types although it does provide a static knowledge mechanism. These problems are also discussed in Shah (2009). The author describes that traditional data models have become unfruitful for biological data. The unsuitability of these models is due to its high volume, dynamic nature and evolving behavior. Furthermore, the traditional data models having their static schemas, so they are unsuitable to model the dynamic biological structures classified by Jung (1995); Shin (1995); Keet (2003a); Zhou and Song (2005); Ghalayini *et al*. (2006); Macedo and Porto (2007); Shah (2009); Graves (2012);. It also suggests that sincere efforts are required for refinement and modeling using a common platform for standardization highlighted by Bauer and Paton (2002); Idrees *et al*. (2014b).

In this paper, we present a survey of conceptual data models for biological data like DNA, RNA and protein structures. The objective of this survey is to understand the molecular concepts, their functions and

biological process in a major. It will prove a smooth effort towards design and development of data model for biological data and for its integration. In the next section, we will describe and present the materials and methods of existing conceptual data modeling approaches for biological entities, their associations and for their semantics relationships. After that, we will present the results and discussions of conceptual data models for biological data. The comparative analysis encompasses different attributes, which will describe the novelty and understanding of already proposed data models. At the end, future directions and references will be included.

MATERIALS AND METHODS

A data model is an abstraction for data representation, its management and for manipulation. It is a transforming mechanism, from virtual reality into real world discussed in Graves (2012). Most of the existing conceptual data models are comprised on the basis of traditional data modeling techniques, like relational data model (RDM), object oriented data model (OODM), object relational model (ORM), entity relationship diagrams (ERD), enhanced entity relationship diagrams (EERD) and data flow diagrams (DFD). Concepts of abstractions, compositions and inheritance are used to elaborate their metadata and schema design. Some models have used unified modeling languages (UML) as a tool for implementation of various aspects. In last but not least, the existing models for biological data are fragmented, unauthentic and non-standardized as discussed in Peckham and Maryanski (1988); Keet (2003a); Zhou and Song (2005); Graves (2012). A detailed survey and their analysis are presented by defining a sort of parameters, extracted from these data models. Commonly, conceptual data models are qualitative in nature. It includes the concepts, functions, processes and their interactions within the molecular biology of under observation data cited in Dori and Choder (2007). Ontology describes invariant features of scientific biological structures, whereas conceptual data model design prominent features. It is used to determine the relationships and paved the way for storage, retrieval and analysis within the systems. It basically integrates the available information of a specific domain and establishes relations mostly with structural constituents described by Ghalayini *et al*. (2006). Gene ontology (GO) reported in Peckham and Maryanski (1988); exploits three constituents of biological data which represents its biological process, cellular components and molecular functions. It provides control vocabularies for the establishment of semantic relationships and conceptual data modeling for biological domains. Its objectives are to maintain the controlled vocabularies which are commonly used in all types of organisms to establish standards for biological terms described by Peckham and Maryanski (1988); Rochfort (2005). Data modeling for biological domain is facing difficulties due to its two inherent characteristics; domain complexity and its evolving nature. Domain complexity can incorporate complex concepts, their interactions, and different data types. The latter property needs a formal language for its expression. The design and establishment of biological schema is a complex task discussed in Ghalayini *et al*. (2006). Its domain specific functions could be characterized as local, as well as global/integrated functions. Local functions cover the domain of an individual's contributions like enzyme is considered as a catalyst for initiation of reactions and domain specific applications. Integrated functions are classified on the basis of biological domains and can form a part of whole system i.e. enzyme is considered as a biological entity, that might perform several functions and may have several sub-functions highlighted in Ghalayini *et al*. (2006). The formal mechanisms based on mathematical modeling explore the foundations of secondary structures of protein and gene mutations described in Gao and Yu (2009). This intelligent technique provides the modeling capabilities of gene mutations, protein secondary structures, and amino acids based on mathematical interpretations of group theory. The proposed framework basically, describes different biological functions of gene mutations in the terminology of computational and molecular biology explained in Gao and Yu (2009). Hierarchical structures in modeling biological systems depict the classifications of hierarchical structures into logical layers. A layer contains parallel biological concepts, their interactions and functions subsequently. A five layers model is represented as sub-cellular, cellular, organs, tissues and organisms in this research for its modeling and understanding of biological systems reported in Bassaingthwaighte (2002). The significant contribution of GIMS based on class diagrams implemented in UML modeling object technology. The Norman (2000); described beautifully that common objectives of ontologies are exploration of complex terminologies, not the description of conceptual studies for storage data. The objective of this research is conceptual description of concept based genome data, transcriptome data and proteome data in detail. The work is implemented in genome information management systems (GIMS) experiments.

We focus on scanning techniques and methodologies presented in recent era by research community towards biological conceptual data models. To understand the construction of molecular systems and their reactive behavior to the environment is still an open challenge for cellular systems reported in Tenazinha and Vinga (2011). XML data models are not suitable for semi structured data natures as highlighted in Gupta (2011); Graves (2012). Research community in bioinformatics domain is in effort to map the biological information on

conceptual data model. The design and development of universal and dynamic schema is non-trivial task because of evolving nature of data. Due to domain complexity and inherent semantics, there is a need of special language to facilitate in design a fully featured data model encompassing the complete requirements of biological domains. It provides us enriched semantic relationships and constraints concluded by Macedo *et al*. (2007).

In Erin Bornberg *et al*. (2002); is supposed to exploit the idea of conceptual data model for bioinformatics in general. They present the macromolecules as an entity types such as enzymes, reactions, biopolymers, organs, DNA, RNA and protein structures. These entity types might consist of several instances of micro molecules like protein contains primary, secondary, tertiary, motifs, etc. Several types of simple and complex relationships are defined to understand the domain concepts of biology by using unified modeling language (UML) modeling tool. This model is used to communicate the concepts of biological molecules but it does not provide a standard way for storage, retrieval and manipulation discussed by Bauer and Paton (2002). One of the most significant contributions has established a variety of hierarchies from cell (root) to top (biosphere) level for the understanding of biological concepts. The biological concepts are modeled and included genomes, proteomes and transcriptoms. A diversity of relationships for intra-links and inter-links has been defined. On the basis of these hierarchical levels, a variety of biosystems are designed. Diverse features with tiny importance are omitted from the details. The designed hierarchical biosystems did not follow any standard methodology, and based on poorly designed data models due to lack of structural hierarchies of these layered approaches. Much short comings exist among such kinds of biosystems. Commonly, they have a lack of flexibility, formalisms and inconvenience for managing and incorporating new data types in Miginsky (2006). It is supposed to be necessary to integrate the biological sources and data at shared platform. The construction of virtual laboratory is its optimum solution for the research community in order to solving this problem. The virtual library is accessible through a single global access point. The research says that object oriented modeling techniques are useful for solving problems of hierarchical data. A metadata based data model is proposed in this research for data and source integration. The meta-model is based on two concepts terms and properties. Terms are used to denote the similar objects for metadata model and properties are used to represent their internal characteristics. The metadata model integrates only light weight or index based biological data reported in Miginsky (2006). In biozone systems framework reported by Birkland (2006); biological objects like DNA, RNA and protein structures are modeled by using combinations of graph and class based model. It exploits that network and hierarchical data models are easy for modeling. These data models increase the complexity especially in the case of multiple parents. It describes that biological data is filled by redundancies, inconsistencies, diversities and heterogeneities. So it is complex to manage such type of biological data in network and hierarchical structures highlighted in Birkland (2006).

Sungwon Jung *et al*. (1995); state that many models provides a smooth path of tree structures for biological data only at data definition language (DDL) level but not provides its implementation details. A graph based schema is used to model the hierarchical biological structures only for their classification and management used by Jung (1995). Graph based data model is also represented in a formal language by using graph in theory Graves (2012). In this research, genomic data is modeled through edges, vertices and vertex names to tailor a graph based data model. Generic genome based data model has proposed through graph structures. This data model is extendable i.e. not static. Edges and nodes may be added or deleted at any time. It incorporates many granularities among them used in Graves (2012). The Dong (1995); explained the strengths and weaknesses of OO and relational models for genome and E.Coli data. The author advocates that complex objects are formed by the combinations of simple and small objects. They found that most of existing data banks like DDBJ, PDB, EMBL, GDB, and GDSB are using relational model for data management. Some data banks like MBASE are using object oriented approaches for their handling. It is supposed to use the object oriented concepts like inheritance, polymorphism and encapsulation to model the E.Coli data have described by Shin (1995). Marta Gozelaz *et al*. (2009); addresses the integration of genomic, proteomic, transcriptomics and amino acid sequence data banks that have been fragmented at diverse locations by the use of semantic ontologies. It is stated that conceptual representation of semantic ontologies are useful for the sources integration. He states due to the use of semantics that new relationships between these biological concepts are arising and are proving helpful for their integration described in Smedley (2009). Jason Swedlow describes that mostly problems occurred in modeling of biological data are due to its massive scaling, multi-dimensional attributes, such as time, space, formats, imaging , vector, audios, videos, sequences, structures, cells, terminologies, tissues, organs and motifs. He beautifully describes that only way of standard integration and shared platforms is only due to OBO, GO, SO used in Huacarpuma (2011). Jamescliffod suggests that by using temporal dimensions in relational model, we may come in a position to solve the problem of evolutionary and dynamic natures of biological data. Temporal dimensions in network, hierarchical model are not much suitable. It mostly adjusted in relational model.

We may use relative time, absolute time and periodic time values as constant parameters in objects identity from its birth to death used by Ghalayini (2006); Clifford (1982). Jake *et al*. (2003); advocated the issues of genomic schema, genomic schema fragments and the current dimensions of genomic data modeling and these concepts have been described in entity relationship diagrams (ERD) by using Erwin. All the modeling concepts discussed in this paper are based on biological sequence data, either it may be derived or annotated in Keet (2003a, b). Keet *et al*. (2003a, b) also used to describe and compare the different modeling approaches like ERD, OO and relational with ORM methodology. He states that ORM technique is much better than other modeling approaches, because it is used only when data is unstructured or semi-structured in nature advocated by Keet (2003a, b). Dove Dori *et al* has addressed the transcriptomic fragment of life cycle of mRNA and its internal representations through the usage of object process modeling (OPM) techniques in detail. The reviewed the conceptual modeling approaches based on qualitative modeling and quantitative modeling. Through OPM it describes the biological object with entity, process and by its states. Entity represents the uniqueness of objects of molecular biology. Processes are used to denote the changes in states during its life cycle and states are stable ends like create, destroy, modify, generate, etc. discussed in Dori and Choder (2007). Avital (2008); focus on the inter object and intra-object based modeling for biological sequences. He made scenario based diagrams to model the complicated concepts. The author is used to discuss that the integration of simple systems resulted produce a large complex system. Mostly they have modularity approaches to support their model advocated in Sadot (2008). Daniel (2005); tries to model the biological concepts by the use of object oriented approaches. He supposed to be considered that complex biological systems are most reactive to the environment due to their dynamic states. He states their behaviors and reactions by modeling their dynamic nature through class diagrams, sequence diagrams, collaboration diagrams and activity diagrams used in Shegogue and Zheng (2005). Robin Cruz in conceptual model for high throughput transcriptome sequence pipeline processes is presented. The conceptual model is based on four phases of sequence pipelining which are filtering phase, mapping phase, assembly phase and annotation phase. The output yielded from one phase is used as input in next phase. It takes short sequences of base pairs having maximum length 100 bp. The research is mainly focused on modeling and pipelining the sequences of biological process rather than biological data reported by Huacarpuma (2011). The generalized model is proposed for unification and querying of biological entities based on structural data types like tree, fragments and subparts of fragmented and annotated data. The proposed data model is CoCoDAG, and it is restricted on directed acyclic graph for biological objects, sub-objects, links and various semantics used by Gupta (2011). Graphs are further classified as directed, undirected, cyclic, acyclic, hyper, nested, DAG, trees, structure graph, hybrid natures, contact graph. Metabolism, gene regulatory and signaling pathways etc. Many kinds of graph and sub graph queries like isomorphism, homomorphism, and homeomorphism which are difficult to handle in traditional data models reported in Jagadish and Olken (2006). This model has minimized the learning cost and misperceptions of the concepts in the medical data and merging the small subsets of sample data into it. This conceptual model has solved the issues of unit inconsistency, term inconsistency and dimension limitation in genotypes and phenotypes of organisms discussed in Zhou and Song (2005). Elmasri *et al* addresses that bimolecular objects are modeled from micro-level to macro-level. Different abstraction levels have been defined to describe the conceptual hierarchies of biological data and concepts by using enhanced entity relationship diagrams (EER). The schematic bio-ontologies are developed to lay the foundations of conceptual data models. It provides a smooth way for development of conceptual data model through the design of semantic ontologies for proteomic data and 3-d protein structures. The author has presented the issues in modeling the biological concepts and the integration of biological data lying at diverse locations. Temporal object is characterized by two parameters, based on structure and state. Both the parameters may change after a specific period of time. But at the same time it keeps history of the relevant parameters as well. This model is elaborated by defining the RoF examples for the prokaryotic and Eukaryotic organisms reported in Shah *et al*. (2009).This proposed conceptual model addresses the modeling methods of molecular biology and its domain concepts by defining Meta data, classes and packaged. It has provided the smooth stability and maintainability at conceptual level of molecular biology. The model has been classified into two categories, operational model and knowledge model. Operational model is an abstract class based model that provides the scope of the overall conceptual model for molecular biology. The knowledge model is a Meta data based category that is used to describe the structures and concrete concepts of the molecular biology domain. The objectives of the proposed model are on stability, flexibility and maintainability of biological frameworks at concept level advocated in Busch and Wedemann (2009).

Based on the above aforementioned exhaustive survey, we have characterized these data modeling techniques and define a number of parameters, to summarize them in a table below. The summary of proposed features in the following table is helpful for

data modeling community to understand the current state of art in this domain. Though it provides bird's eye view in bioinformatics domain, rather it provides the smooth path for data integration as well. In short, our suggested parameters are useful in understanding the modeling techniques adopted by different researchers, how they provided their conceptual data models, their complexity levels, and other key points as well.

RESULTS AND DISCUSSION

In Table 1, we have provided the detailed analysis and comparison between existing available conceptually proposed data models for biological domain (Central dogma) especially. The models are analyzed by defining the different parameters for above said domain. In our tabular structure major features involved are types of data model, modeling technique, schema level, Meta data, application or domain area, difficulty level and query languages proposed in these models. Most of the data models have been proposed based on simple concepts. Modeling techniques specifies, the methods adopted to describe the required data. Schema level denotes the semantics and constraints imposed on these models. Meta data is used to represent the defined attributes, their functions. Application area or domain addresses the chosen application for certain data model. Difficulty level represents either its understanding the concepts and query language denotes the storage and retrieval methods applied in the given data models. The attributes which are necessary for the elaborations of these data models are included as shown in the above table. We included those data models which are closely relevant to our research area. In future, we intend to extend these parameters for further elaborations of these conceptual data models.

The following graph has been drawn based on the heuristics available for conceptual data models. We have divided most relevant conceptual data models to our domain, as isolated models for DNA, RNA and Protein structures, miscellaneous, genomics, collaboration of DNA and protein structures and in the last combination of DNA, RNA and protein structures in general.

Table 1. Analysis and comparison of existing conceptual data models for biological data

Data Models/ Characteristics	Data Model Used	Modeling Method	Schema Level	Metadata Representation	Application Area/Domain	Difficulty Level	Query language
A Conceptual Model for Transcriptome High-Throughput Sequencing Pipeline Huacarpuma (2011)	Conceptual model	Sequencing Pipeline model	Constraints Based for SRS	Specify attributes only	Transcriptome high throughput (DNA) data	More or Less	No
Conceptual Modeling for Applied Bioscience: The Bacteriocin Database Keet (2003a)	Conceptual model	ERM, OOM and ORM	Object and state based	Class based concepts	Bacteriocin database	Less	No
Deriving Conceptual Data Models from Domain Ontologies for Bioinformatics Ghalayini (2006)	Conceptual model	Mapping rules based	Semantics and Ontology based	Reverse engineering approach based	Description of general biological concepts	Less	No
Graph Data Models for Genomics Graves (2012)	Conceptual model	Graph based	Constraints based	Using tree like structures and graph schema language	Genomic data	less	Graph query algorithms
Temporal Object Oriented System (TOS) for Modeling Biological Data Macedo and Porto (2007)	Conceptual model	Temporal object based (TOS)	History based	Based on structure and state of object	Prokaryotic and eukaryotic organisms	Less	Temporal query language
Biological data and conceptual modeling methods Keet (2003b)	Conceptual Model	ERM, OOM and ORM	Object and state based	ER Diagram, ORM, OO concepts	Bacteriocin, Micro-organism	Less	No
Conceptual Modeling of Genetic Studies and Pharmacogenetics Zhou and Song (2005)	Conceptual model	Unified Modeling Language (UML)	semantics	Packaged diagrams, class diagrams	Clinical and Biomedical Data	Less	No

Conceptual data modeling for bioinformatics Bauer and Paton (2002)	Conceptual model	ERM-OOM	Semantics based	Defined attributes, classes and entities	DNA and Protein structures	Less	No
Conceptual Modeling in System Biology Fosters Empirical Findings: The mRNA Lifecycle Dori and Choder (2007)	Conceptual Model	Object Process Methodology (OPM)	Object and State based	System Diagrams, Object process Diagram	mRNA	Less or More	No
Proteins, Cells, Organs: Hierarchical Structuring in Modeling Biological Systems Bassaingthwaighte (2002)	Conceptual model	Layered approach		Hierarchy based	Cellular structures	Very Low	No
Conceptual Modeling of Genomic Information Paton *et al*. (2000)	Hybrid model (more conceptual less physical model)	Unified modeling language (UML) and GIMS	Ontology based	Class diagrams	Genome, transcriptome and proteome data and concepts	Less	No
A Unified Data Model and Declarative Query Language for Heterogeneous Life Sciences Data Gupta (2011)	Conceptual model	Directed Acyclic Graph	Semantics based	Highly based on abstraction, sub-fragmented parts considered as an objects, model objects and semantics	Generic Life sciences domains	More	XD Path based Query Lang-uage
BIOZONE: a system for unification, management and analysis of heterogeneous biological data Smedley (2009)	Conceptual model	Graph based	Objects and classes	Class based concepts	DNA, RNA and Protein Structures	More	No
Genomic data modeling Chen and Carlis (2003)	Conceptual model	ERM	High level Schema elements Sequence based	Entities, attributes, and keys, tabular structures	Genomics	More	No
Toward Verified Biological models Sadot (2007)	Physically implementation	State based	Module based	State and behavior based formalism	Cell, Fate	Too High	No
Object-Oriented Behavior Modeling for Biological Systems Shegogue and Zheng (2005)	Conceptual model	OOM	Behavior capturing using sequence, state and activity diagrams	Objects and state defined, static and dynamic behavior	Genome of Influenza A.	Less	No
Modeling Genomic Data with Type attributes, balancing stability and maintainability Busch and Wedemann (2009)	Conceptual model	Package and class based	Semantic based	Operational and knowledge based	Biological structures	Less or More	No

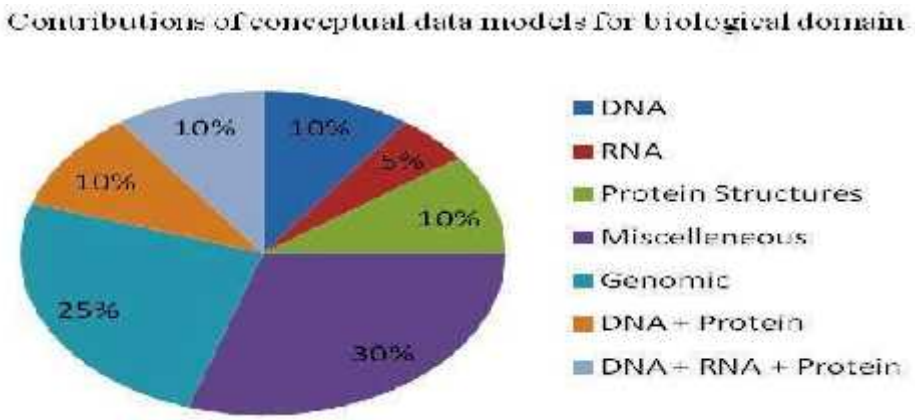

Figure 2. Analysis of existing conceptual data models for biological domain

Conclusions and Recommendations: Traditional data models are unfruitful to manage the dynamic behaviors, evolving structures and its diversity of data types and formats of biological data in a uniform way advocated by Rochfort (2005); Dori and Choder (2007); Idrees and Khan (2014a); Idrees *et al*. (2014b).

Mostly conceptual data models are independent of implementation details. They have used conceptual modeling methods and approaches for elaborations of biological concepts. The conceptual data models proposed for biological data in the existing literature, nevertheless a single model; capture the complete information in a unified conceptual manner. Due to lack of nomenclatures, their standards, disagreements on different biological definitions, terms and diversity in describing biological concepts, there exists large amount of uncertainties and abundance inconsistencies among them. In short, current modeling techniques, methods and tools are proved to be unfruitful and in-efficient, to model the behavior, structure, process, entities and their interactions at micro-level as well as at macro-level of current biological systems. The different behavior and understand-ability of each computer scientist and bioinformaticians in the biological system has created a lot of difficulty for other research community as well. There is a dire need of uniform data model for biological data that will facilitate and incorporate its data types, formats and its evolutionary nature for storage, retrieval and manipulations in a unified fashion.

REFERENCES

Busch, N. and G. Wedemann (2009). Modeling genomic data with type attributes, balancing stability and maintainability. BMC Bioinformatics, 10(97): 1-16.

Birkland, A. and G. Yona (2006). BIOZONE: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinformatics, 7(70): 1-24.

Bauer E. B. and N. W. Paton (2002). Conceptual data modeling for bioinformatics. Briefings in Bioinformatics, 3(2): 166-180.

Bassaingthwaighte, J. B. (2002). Proteins, cells, organs: hierarchical structuring in modeling biological systems. Proceedings of the Second Joint EMBS/BMES Conference, Houston, TX, USA (3): 2176-2177.

Cohen, J. (2004). Bioinformatics – An Introduction for Computer Scientists. ACM Computing Surveys, 36(2): 122 – 158.

Clifford, J. (1982). A model for historical database. Information systems, 1(1): 1-54.

Chen, J. Y. and J. V. Carlis (2003). Genomic data modeling. Information Systems, 28(1): 287-310.

Dori, D. and M. Choder, (2007). Conceptual Modeling in System Biology Fosters Empirical Findings: The mRNA Lifecycle. PLOS | ONE, 2(9): e872.

Elmasri, R., F. Li and J. Fu (2008). Modeling Biomedical Data. In: Chen, J.; Sidhu A S, (eds.) Biological Database Modeling, 25-49.

Ghalayini, E. H., M. Odeh and R. McClatchey (2006). Deriving Conceptual Data Models from Domain Ontologies for Bioinformatics. ICTTA'06, arXiv:cs/0603037 1(2): 3562 – 3567.

Graves, M. (2012). Graph Data Models for Genomics. Submitted to ACM Transactions on Database Systems, 1-42.

Gao, R. and J. Yu (2009). Mathematical Models of Protein Secondary Structures and Gene Mutations. Proceeding of the 2009 IEEE, International Conference on Mechatronics and Automation, 1(1): 4327-4332.

Gupta, S. (2011). A Unified Data Model and Declarative Query Language for Heterogeneous Life Science Data. SDSC TR-2011(3):1-11.

Huacarpuma, R. C., M. Holanda and M. E. Walter (2011). A Conceptual Model for Transcriptome High-Throughput Sequencing Pipeline. Advances in

Bioinformatics and Computational Biology, LNCS, 1(6832): 71-74.

Idrees, M. and M. U. G. Khan (2014a). SMGCD: Metrics for biological sequence data. The Nucleus, 51(1): 125-131.

Idrees, M., M. U. G. Khan and A. Shah (2014b). Unified Data Model for Biological Data. The Mehran University Research J. of Engineering and Technology, 33(3): 261-277.

Jagadish, H. V. and F. Olken (2006). Database Management for Life Sciences Research. SIGMOD Record, 33(2): 15-20.

Jung, S., S. Perkins, Y. Zhong, S. Prmanik and J. Beaman (1995). A New Data Model for Biological Classification. Bioinformatics, 11(1): 237-246.

Kari, L., R. Kitto and G. Gloor (2001). Focus: Toward Soft Hardware, A computer scientist's guide to molecular biology. Soft Computing, 5(1): 95 – 101.

Keet, C. M. (2003a). Biological Data and Conceptual Modeling Methods. J. of Conceptual Modeling, 29(1): 1-14.

Keet, C. M. (2003b). Conceptual Modeling for Applied Bioscience: The Bacteriocin Database. J. of Conceptual Modeling, 1(1): 1-25

Macedo, J. A. F. D. and F. Porto (2007). Dealing with Some Conceptual Data Models Requirements for Biological Domains. 21ST International Conference on Advanced Information Networking and Applications Workshops (AINAW' 07), (1): 142 651-656.

Mesiti, M., E.J. Ruiz, I. Sanz, R.B. Llavori, P. Parlasca, G. Valentini and D. Manset (2009). XML-Based approaches for the integration of heterogeneous bio-molecular data. BMC Bioinformatics, 10(Suppl 12): S7 1-18.

Macedo, J. A. F. D., F. Porto, S. Lifschitz and P. Picouet (2007). A Conceptual Data Model Language for Molecular Biology Domain. Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07), 231-236.

Miginsky, D. S., V. V. Suskolov, D. A. Rasskazov, N. L. Podkolodny and N. A. Kolochanov (2006). Architecture of software toolkit for storing and operating with biosystems models. Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure, BGRS'2006, 3(1): 292-295.

Miginsky, D S., V. V. Suskolov, V. V. Labuzhsky, A. G. Nikittin and I. G. Tarancev (2006). Object-oriented approach to bioinformaics software resource integration. Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure. BGRS'2006, 3(1): 288-291.

Nair, A. S. (2007). Computational Biology & Bioinformatics: A gentle Overview. Communications of the Computer Society of India, 31(1): 1-13.

Paton, N. W., S. A. Khan, A. Hays, F. Moussouni, A. Brass, K. Eilbeck, C. A. Goble, S. J. Hubbard and S. G. Oliver (2000). Conceptual modeling of genomic information. BIOINFORMATICS, 16(6): 548-557.

Peckham, J. and F. Maryanski (1988). Semantic data Models. ACM Computing Surveys, 20(3): 153-190.

Rochfort, S. (2005). Metabolomics Reviewed: A New "Omics" Platform Technology for Systems Biology and Implications for Natural Products Research. J. Nat. Prod, 68(12): 1813-1820.

Shah, A. A., S. Ahsan and A. Jaffer (2009). Temporal Object-Oriented System (TOS) for Modeling Biological Data. J. of American Science, 5(3): 63-73.

Shapiro, J. A. (2009). Revisiting the Central Dogma in the 21st Century. Ann. N.Y. Acad. Sc., 6(28): 1178.

Shin, D. G. (1995). Comparative Study of Relational and Object-Oriented Modeling of Genomic Data. Proceedings of the 28th Annual Hawaii International Conference on System Sciences, IEEE, (5): doi:1060-3425, 81-90.

Smedley, D., S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson and A. Kasprzyk (2009). BioMart-biological queries made easy. BMC Genomics, 10(1): 1-12.

Swedlow, J. R., S. E. Lewis and I. G. Goldberg (2006). Modeling data across labs, genomes, space and time. NATURE, CELL BIOLOGY, 8(2006): 1190-1194.

Sadot, A., J. Fisher, D. Barak, Y. Admanit, M. J. Stern, E. J. Hubbard and D. Harel (2008). Toward Verified Biological models. IEEE/ACM Trans Comput Biol Bioinform, 9(1): 67-103.

Shegogue, D. and W. J. Zheng (2005). Object-oriented biological system integration: a SARS coronavirus example. Bioinformatics, 21(10): 2502-2509.

Tenazinha, N. and S. Vinga (2011). A Survey on Methods for Modeling and Analyzing Integrated Biological Networks. IEEE/ACM transactions on computational biology and bioinformatics (TCBB), 8(4): 943-958.

Topaloglou, T. (2004). Biological Data Management: Research, Practice and Opportunities. Proceedings of the 30th VLDB Conference, 1233-1236.

Zhou, X. and II. Song (2005). Conceptual Modeling of Genetic Studies and Pharmacognetics. In: O. Gervasi *et al*. (eds.) ICCSA 2005, LNCS 3482, Springer, Heidelberg, 3482(1): 402-415.