Goal-oriented Schema in Biological Database Design

Ping Chen
Department of Computer Science
University of Helsinki
Helsinki, Finland 00014
EMAIL: pchen@cs.helsinki.fi

***Abstract*—In this paper, I reviewed current research status in database design and presented a new idea, which is called goal-oriented schema, in database design proposed by Lei et al. using a case study from biological data management. Goal-oriented strategy shows its advantages in database design over traditional requirement-based design schema. This schema is promising in the development of database design.**

I. INTRODUCTION

Over the last decades, a huge amount of biological data has been accumulated as the rapid development of biotechnology. In order to understand and explain biological phenomena from the data, people are now focusing more on data analysis originating from their former work and using those results to direct their experiments. Thus, we need a tool to organize all the data, biological databases having been considered as such a tool to assist scientists in data management.

As of 2006, there are over 1000 public and commercial biological databases, containing genomic, proteomic and metabolomic data. There are different kinds of biological databases based on their different functions, such as sequence databases (DDBJ, EMBL, GenBank), genome databases (Ensembl), protein sequence databases (UniProt, Swiss-Prot, Pfam), protein structure databases, protein-protein interaction databases and microarray databases. A good database design schema is now playing an important role in organizing biological data to satisfy more requirements from users [1].

Standard database development contains [2, 4] requirements analysis, logical design and physical design. Requirements analysis results in a conceptual schema about how data to be stored. In recent years, goal-oriented approaches [3, 4] in requirement analysis have been widely used and proved to be an effective way in database design. This approaches focuses on modeling stakeholders' goals, exploring a space of alternatives and selecting one on the basis of criteria [4]. Goal analysis in database design would display not only the meaning of the data, but also user groups and the purposes of the database.

Here, I will give a review mainly on goal analysis in biological database design, using a case study in microarray data management. In section 2, I give an introduction to database design process which adds a goal analysis phase before the conceptual schema design. In section 3, I focus on the current status of biological database design based on the case study and exhibit its evolution. In section 4 and 5, I mainly concentrate on the goal analysis of biological database design. In section 6, I conclude and give my opinions on database design driven by stakeholder goals.

II. GOAL-ORIENTED DATABASE DESIGN PROCESS

In the past few years, database researchers have developed many design strategies and produced different kinds of database design processes. In 1999, P.Atzeni et al. presented a design strategy based on the types of modeling constructs, such as entities and relationships [6]. In another case, a step by step strategy was proposed to build a database, including top-down, bottom-up, inside-out and mixed strategies [7]. In general, database design consists of steps of requirement analysis, conceptual design, logical design, physical design, database implementation and maintenance.

In 2007, Lei Jiang et al. [5] presented a goal-oriented design strategy. It has several steps which start from a group of stakeholders and their high-level goals [Fig1].

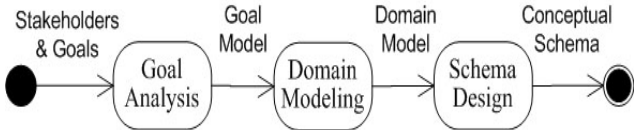

Fig 1: Goal-oriented design process

February 26, Helsinki
58308110 Seminar: Management of Biological Databases (3 cu), spring 2008
Department of Computer Science
University of Helsinki
Email: pchen@cs.helsinki.fi

Goals are collected and then proceeded a goal analysis phase to produce a goal model. More detail about the goal analysis is showed in section 3. Based on such a goal model, requirement analysis integrates a set of alternative data requirements, a particular one chosen to generate the conceptual model. Conceptual design is an essential step to transform the data requirements into a data description model, which displays the "real world" [Fig2]. Entity-relationship model (ER) is a good example of the conceptual model [2], which explicitly displays the relationship among entities. The conceptual schema [9] reflects all the changes during evolution, while logical schema describes structure of the database and is relatively stable. The logical model has a feature of tables, holding primary key and foreign key in it [Fig3]. Logical design follows a design of physical structure, which including designs for data storage structure and storage methods. Finally, based on the logical model and physical model, a test is needed before the database open to the public. Meanwhile, it should be maintained during its operation. All the steps in this goal-oriented strategy are driven by the goal model which is created in the first step.

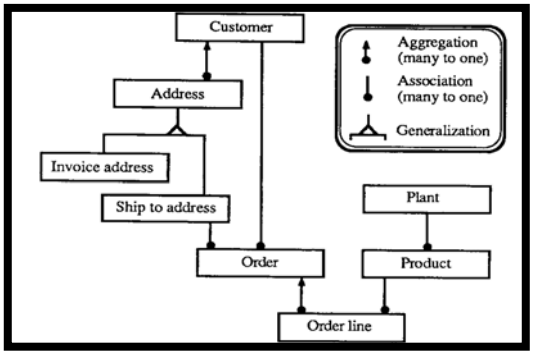

Fig 2: Example of conceptual schema

The conceptual model shows entities and the relationship between them, transforming the data requirements into a data description model.

```
Customer (CustId, CustName, CreditLimit)
Address (CustId, AdrId, AdrKind, Address)
InvoiceAddress (CustId, AdrId, AttnOf)
ShipToAddress (CustId, AdrId, ShippingMode)
Plant (PlantId, PlantLocation)
Product (ProdCode, Prodname, PlantId)
Order (OrderId, CustId, OrderDate, OrderTotal)
OrderLine (OrderId, LineId, Qty, ProdCode, LineTotal)

Underlined:Primary Key
In Italics: Foreign Key
```

Fig 3: Example of logical schema

The logical model has a feature of tables, holding primary key and foreign key in it. Primary key can specify each record in a table, while the foreign key can link to other tables.

III. DESIGN OF BIOLOGICAL DATABASE : A CASE STUDY OF MICROARRAY DATABASE

Up to now, huge amounts of biological data have been collected from different biological sources. Biological data is produced in a digital form which needs to be stored in a database, which is supposed to satisfy different user groups in their own researches. A well-designed biological database is a powerful tool which can contribute a lot to biological researches. So, the scheme of biological database design is quite important.

Here, I use a case study of microarray data management to illustrate the schema in the biological database design.

Microarray technique is a high-throughput method which is commonly used for gene expression profiling on the level of transcriptome, monitoring expression levels of thousands of genes simultaneously. A DNA microarray (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots called probes, commonly representing single genes, arrayed on a solid surface by covalent attachment to a chemical matrix. Samples are total RNA extracted from cells and labeled with dye. Qualitative or quantitative measurements with DNA microarrays utilize the selective nature of DNA-DNA or DNA-RNA hybridization under high-stringency conditions and fluorophore-based detection. After hybridization, a scanner can detect the fluorescent intensity and produce a digital image of the gene expression profiling [Fig 4, 5, 6].

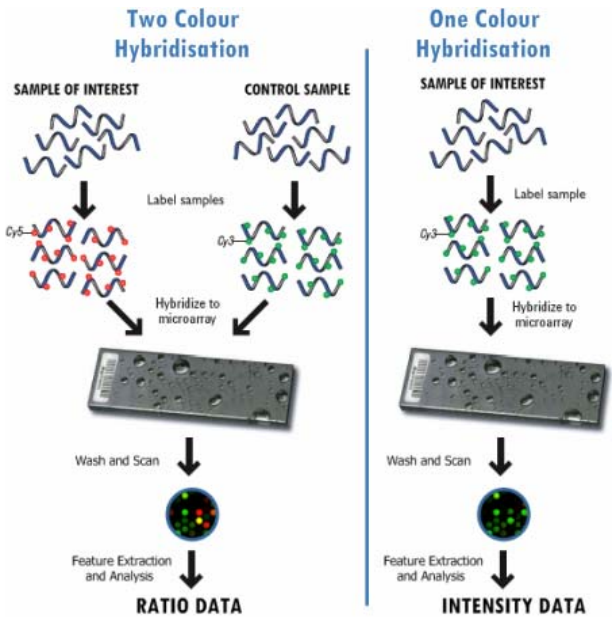

Fig 4: Two-color microarray (left) VS one-color microarray (right) (In a two-color microarray, two samples are labeled differently with Cy5 and Cy3 fluorescent dyes. In a one-color microarray, only one sample can be measured and be labeled with Cy3.)

According to different gene expression profiling techniques, microarrays can be categorized into two groups, one-color and two-color microarrays.

One-color microarrays, also called single-channel microarrays, are designed to give estimations of the absolute levels of gene expression. Therefore, the comparison of the two sets of conditions requires two separate single-dye hybridizations. Since only a single dye

is used, the data collected represent absolute values of gene expression. These may be compared to other genes within a sample or to reference normalizing probes used to calibrate data across the entire array and across multiple arrays.

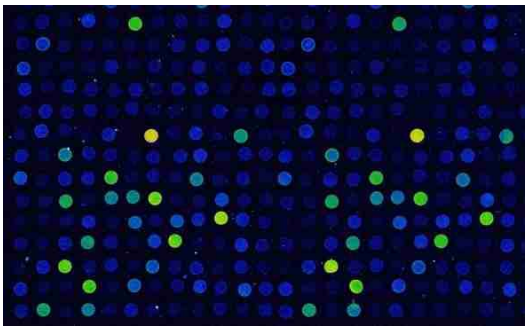

Fig 5: Digital image of one-color microarray

Two-color microarrays are typically hybridized with cDNA prepared from two samples that the researchers wish to compare (such as diseased tissue versus healthy tissue). These samples are labeled with two different fluorescent dyes (Cy3/Cy5) with different emission wavelength. The two Cy-labelled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

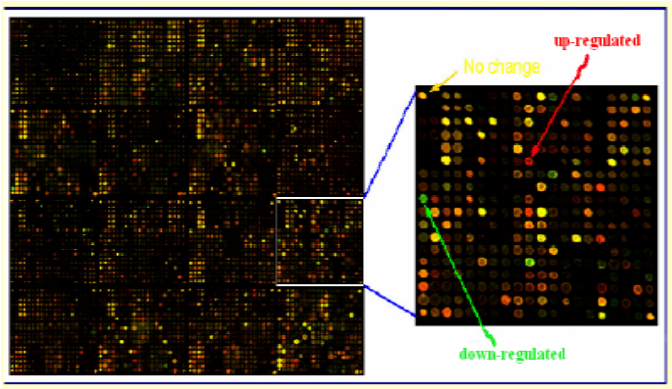

Fig 6: Digital image of two color microarray

In two-channel microarrays, sample I is labeled with Cy5 (Red) and sample II is labeled with Cy3 (Green). In the image, red spots mean these genes are up-regulated in sample I; green spots mean that they are down-regulated and yellow ones represent that there is no change between the two samples.

In 2000, D.J.Lockhart et al. presented that such a technology has made it possible to produce large amounts of gene expression data at a time [8]. The management for gene expression data is urgent for gene expression analysis. In 2001, Markowitz, V.M et al. proposed applying data warehouse concepts to gene expression data management [1]. He indicated that data management for gene expression data should satisfy the requirements of data acquisition and analysis and modeled the gene expression data into three spaces, which are sample space, gene annotation space and gene expression data space. It shows the importance for requirement analysis. Later in 2006, Lei Jiang et al. [4] proposed a new idea of goal analysis using a case study of biological data design.

Design of biological database has been a focus more and more people are concentrating on. Traditional database design starts with requirement analysis which reflects user requirements for the data structure. Lei Jiang et al. exhibited an evolving design of biological database using a case of 3Sdb (Small subset of sample database) [4]. 3Sdb is a repository of data on biological samples in gene expression experiments, which stores information on samples and their donors. In requirement analysis, one of major requirements is data acquisition. So, a good organization of the data can contribute a lot to satisfy different user groups. Schema on how to organize data from samples and their donors has evolved over a period of 18 months, four versions of conceptual design coming out before goal-oriented database design strategy.

Version I organizes three main concepts and several sub-concepts, including Sample, Study Group and Donor. Concept Sample includes all the biological samples, holding a long set of attributes. Concept Study Group represents a group of samples with a set of experiment parameters. And concept Donor and its sub-concepts are designed to organize donor information, such as diagnoses, medications and family history.

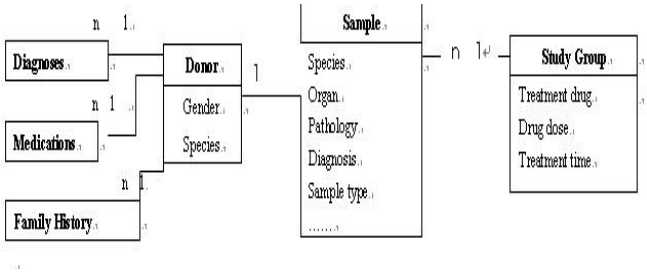

Fig7. Design for biological sample (v1)

Version I organizes three main concepts and several sub-concepts, including Sample, Study Group and Donor.

Version II specialize the concept Sample into different sample types, such as Tissue Sample, Cell Culture. In each sub-concept, there is a list of attributions to specialize different queries.

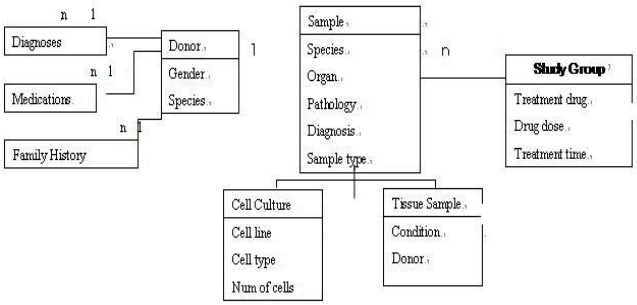

Fig8. Design for biological sample (v2)

The concept Sample into different sample types, such as Tissue Sample, Cell Culture.

In version III, a new concept *Matched Sample* is introduced, which represents a set of samples coming from same donor or from the same biopsy. Two new concepts *Donor Visit* and *Visit Update* are introduced in the Donor

profile. In this new profile, each sample is associated with a donor visit and each visit is updated in the concept of *Visit Update*. A donor can give his sample by different donor visits with different diagnosis information by each visit update. Fig 9 shows the relationship between them.

In version IV, the concept *Treatment* is separated from the concept Study Group, which allows multiple treatments used in the same study group.

Four versions of the 3Sdb conceptual schema show the evolution over the time period before the appearance of the goal-oriented design schema. Along with the new design strategy proposed, biological database design has trended to start from goal analysis [4, 10, and 11].

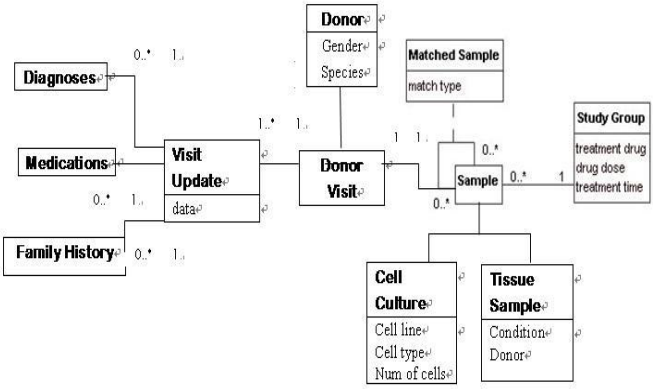

Fig9. Design for biological sample (v3)

Several concepts (Matched Sample, Donor Visit and Visit Update) are introduced in this version.

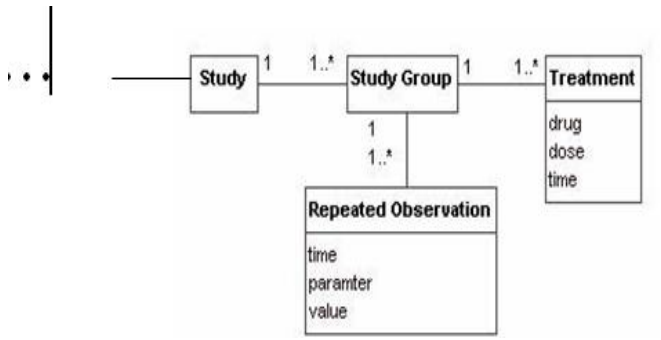

Fig10. Design for biological sample (v4)

In this version, the concept *Treatment* is separated from the concept Study Group.

As shown above, conceptual schema of 3Sdb has been modified during the evolution of database design. In 2006, Lei Jiang et al. revisited the design progress and put a goal analysis into the step of requirement analysis. They continued the case study of 3Sdb by introducing a goal analysis step in a new version of 3Sdb design.

The goal analysis aims to build a goal model, starting with a set of high-level goals of stakeholders. In the case of 3Sdb, the top goal is to collect and organize data of biological samples, which is an entry point of goal analysis using certain goal reasoning technique. Lei shows two techniques used in goal analysis, AND/OR decomposition and means-end analysis.

AND/OR decomposition constructs a goal model by refining the goals into a set of sub-goals with alternative ways to achieve the top goal [Fig 11]. As is shown in this model, the top-level goal is to correlate sample and donor conditions with gene expression data. In order to achieve this goal, the top goal is decomposed into three sub-goals, which are to correlate gene expression with normal organs, to correlate gene expression with diseases and drugs and to correlate gene expression with other factors, all having a relationship of AND decomposition with the top goal. In the second step, a sub-goal 1.2 is refined into 4 sub-goals 2.1, 2.2, 2.3, 2.4 of itself, still holding AND decomposition type. In the last step, the model defines that in order to achieve the sub-goal 2.2, one of the sub-goals 3.1, 3.2, 3.3 of itself should first be achieved.

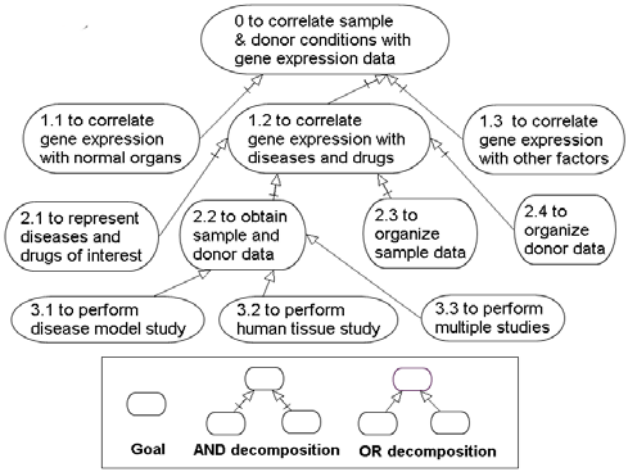

Fig11. A goal model from AND/OR decomposition goal analysis

AND/OR decomposition constructs a goal model by refining the goals into a set of sub-goals with alternative ways to achieve the top goal.

Means-end analysis is another type of goal analysis which describes a relationship between goals and methods achieving them. This technique is well explained in Fig 12, showing different means to achieve each goal. Lei gave an example of goals 3.1 and 3.2. In this model, disease model study can be performed by using animal models, cell cultures or both of them. And human tissue study can be performed by using samples from patients.

The goal model produced from the goal analysis shows alternative data requirements and provides multiply ways for setting relationships between different data. Compared with other design strategy in the case study of 3Sdb, the goal-oriented design process has exhibited its advantages, not only on behalf of more comprehensive information it provided from alternative data requirements, but also on the generation of schemas with rich and explicit data semantics.

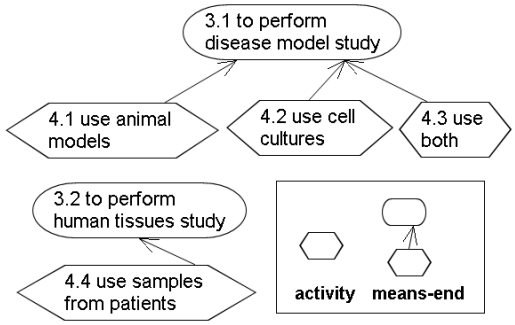

Fig12. A goal model from Means-end goal analysis

Means-end analysis is another type of goal analysis which describes a relationship between goals and methods achieving them.

IV. STEPS IN GOAL ANALYSIS

Later in 2007, Lei Jiang et al. mentioned a design process of the goal model in more detail [5].

In the first step, the main purpose for this step is to identify high-level goals of each stakeholder with a list of stakeholders as input, goal identification as its task and a list of top goals of each stakeholder as an output.

In the second step, a list of top goals generated in the first step is input in order to produce a goal model by goal analysis. The techniques used in goal analysis has already been explained in the case study of 3Sdb using the technique of AND/OR decomposition. A more complicated example of a portion of goal model is showed in Fig 13, which explicitly demonstrates a set of highly alternative data requirements in the goal model.

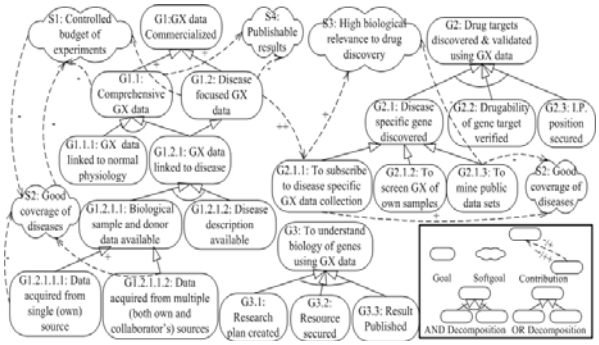

Fig13. A more complicated example of a portion of goal model

In the third step, the objective is to select a design alternative by goal evaluation with the input of goal model created in the second step. The output in this step is a set of leaf-level goals in the goal model, whose collective fulfillment achieves the aggregate top goals.

In the fourth step, it aims to identify initial set of domain notions from goals we select. To achieve each goal, specific datasets are needed. Domain notions represent potential application data requirements [Table1].

Goals	Domain Notions
G1	gene, gene expression
G1.2	disease
G1.2.1	linked(gene expression, disease)
G1.2.1.1	biological sample, donor
G1.2.1.2	
G1.2.1.1.2	sample source, collaborator

Table1. Domain notions

In the fifth step, the purpose is to identify and select plans to achieve a goal by goal operationalization and plan evaluation. A method called "Means-end analysis" is used in this step, proposed in 2006 by Lei et al [Fig 14].

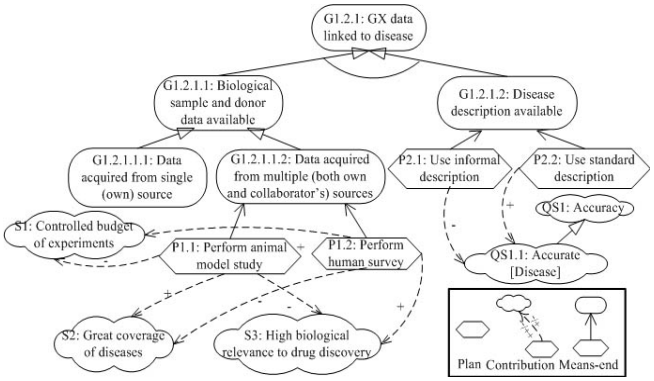

Fig14. A goal model with enriched plans

The last step of goal analysis is to expand the set of domain notions using plans and to construct the domain model for the target database. The domain model finally gives a framework of relationships among all the domains originated from former steps, which is essential in the construction of conceptual schema. Example of a domain model shows in Fig15.

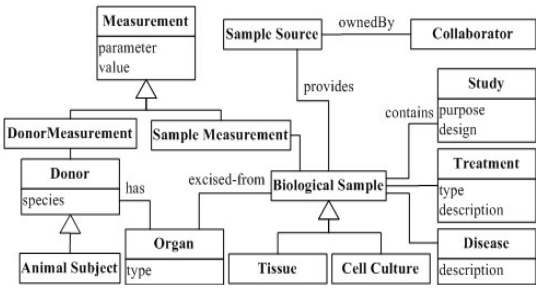

Fig15. Example of a domain model

V. CONCLUSION

In recent years, the notion of database has been proposed and applied in different fields. As a large number of data keeps coming out at a rapid speed in the real world, people are now concentrating on finding a good design schema to manage all the data. Although different database design strategies have been proposed in the past few years, database design schema is still keeping developing as requirement changes all the time.

Combined with biological data management, a new strategy of goal-oriented database design was proposed by Lei Jiang et al. in 2006 [4]. In this paper, I have mainly focused on this goal-oriented approach in database design. Compared with conventional database design strategy, goal-oriented schema shows its advantages on data management. Firstly, goal model, a product of goal analysis, provides a set of alternative sub-goals to achieve the top goal, which makes it feasible to integrate data in an

alternative way. From this model, the relationship between all the data is more explicit and meaningful. Secondly, a domain schema designed based on the goal model gives a refinement for the follow step of conceptual model design, which shows a better transmission in the design process compared with the former requirement-based conceptual model. Thirdly, on the behalf of biological data management, this approach can greatly satisfy biologists not only on the explicit function of a certain database, but also on structure of the data organization, where each data set can be traced back to its goal.

From a case study of biological database design used in this paper, goal-oriented database design strategy has showed its advantages in data management. In the future, maybe more and more database designers will adopt this schema in their own database design. As the world varies from time to time, database design will keep improving in this process. Driven by goal, integrating more factors in database design, it is promising towards the development of database design and a more perfect schema will come out in the near future.

REFERENCES

[1] V. M. Markowitz and T. Topaloglou, “Applying Data Warehousing Concepts to Gene Expression Data Management,” presented at the 2nd IEEE International Symposium on Bioinformatics & Bioengineering, Bethesda, USA, Nov. 4-6, 2001.

[2] C. Batini, “Conceptual database design: an entity-relationship approach,” Benjamin/Cummings Pub. Co., Redwood City, **USA**, 1991.

[3] J. Mylopoulos, “From Object-Oriented to Goal-Oriented Requirements Analysis,” presented at Communications of the ACM, New York, USA, Jan, 1999.

[4] Lei Jiang, “Incorporating Goal Analysis in Database Design: A Case Study from Biological Data Management,” presented at 14th IEEE International Requirements Engineering Conference, Minneapolis/St.Paul, USA, Sep.11-15, 2006.

[5] Lei Jiang, “Goal-Oriented Conceptual Database Design,” presented at 15th IEEE International Requirements Engineering Conference, Delhi, India, Oct 15-19 2007.

[6] P.A. Ng, “Further Analysis of the Entity-Relationship Approach to Database Design,” Software Engineering, vol. 7, pp. 85-99, Jan/Feb, 1981.

[7] T. M. Connolly and C. E. Begg, “Database Solutions: A step by step guide to building databases”. Addison Wesley, 2003.

[8] D.J. Lockhart and A.E. Winzeler, “Genomics, Gene Expression, and DNA Arrays”, Nature, 405, pp. 827-836, 2000.

[9] Qing Li and Dennis McLeod, “Conceptual Database Evolution Through Learning in Object Databases,” Knowledge and Data Engineering, Vol.6, pp.205-224, Apr 1994.

[10] R. Gustas, (1996).Goal Driven Enterprise Modelling: Bridging Pragmatic and Semantic Descriptions of Information Systems. Information modelling and knowledge bases VII,[Online] pp. 73 – 91. Available: http://portal.acm.org/

[11] A. Dardenne, “Goal-Directed Requirements Acquisition,” Elsevier Science Publishers B. V. ,Amsterdam, The Netherlands,, 1993, pp.3-50