ISSN No. 0976-5697

Volume 1, No. 3, Sept-Oct 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Development of Information Model Containing Specialized Information

Dr.Lyazat Naizabayeva*
Associate professor of Computer Engineering Department,
Kazakh-British Technical University,
Almaty, Kazakhstan
naizabayeva@gmail.com

JelinaChadiarova
Bachelor of Computer Engineering Department,
Kazakh-British Technical University,
Almaty, Kazakhstan
j.chadiarova@gmail.com

Abstract: The paper describes the use of CASE-technology in the development of the cardiovascular diseases (CVD) database, which contains information about genes and the relevant human proteins. The physical design of the database is done using a professional engineering package MS SQL Server. In the created distributed database data is organized and structured, the security of a client-server database is organized, user privileges are assigned. The client application is created in Borland C++ Builder programming environment. A qualitative data processing using engine is suggested, it uses a universal search, which allows you to find relevant information about the genes responsible for disease development, gene-candidate synonyms, its markers, the chromosome area map.

Keywords: Computer-Aided Software / System Engineering (CASE), ER-model, MS SQL Server, Structured Query Language, Client application, Borland C++ Builder.

I. Introduction

Experts around the world are constantly working on identifying the genes, relevant proteins, as well as molecular markers of cardiovascular (CVD) and oncological (OD) diseases, as the most common diseases with a high mortality today [10]. The creation of a full organized information database, that contains reliable information about genes and proteins, responsible for some of the CVDs and ODs, is important. The creating database will allow to carry out a large-scale study of the properties of genes and related proteins, responsible for some of the CVDs and ODs development, that hasn't been conducted before, to create molecular methods of early disease detection, as well as to identify targets for a gene-therapeutic remedy creation. An early diagnosis is especially important for cancer medical treatment, that is, the creation of molecular methods for early diagnosis will greatly increase the chances of a patient medical treatment successful outcome [4],[7].

The novelty of the work is to create a database that contains specialized information about CVD. Existing databases of National Center for Biotechnology Information (NCBI) (www.ncbi.nih.gov), Gene Ontology on proteins of Swiss Institute of Bioinformatics SWISS-PROT (www.geneontology.org), contain information obtained in the course of experimental work. The created database contains information about genes and relevant proteins annotated by hand, which greatly increases the functionality of the base and the data validity level.

II. Logical Database Design

Designing a database schema should solve the problem of minimizing data duplication and simplification of their processing and updating. When database schema is not properly designed anomalies in data modification may occur. They are due to the lack of tools of explicit representation of multiple relations between software objects and underdevelopment of tools for describing integrity constraints at the data model level. A relation normalization is carried out to solve such problems [11],[8].

As part of the relational model the method of E.F.Codd was used, which was developed in order to normalize relations, a mechanism, which allows to convert any relations into the third normal form, proposed by Codd was used [3]. The normalization of relations scheme is done by using the scheme decomposition. Using the method of normalization, the redundancy in the table was decreased, inconsistencies and unnecessary expenditure of disk space problems were solved. Normalization ensured the absence of data loss.

A. *Application of CASE-technologies.*

In connection with the clarity of the database conceptual schemas representation ER-models are widely used in Computer-Aided Software/System Engineering (CASE) systems, supporting the automated design of relational databases.

The creation of modern information systems is a challenge, which requires the use of special techniques and tools. Not surprisingly, that in the recent years among system analysts and developers an interest in the CASE-technologies and CASE-tools, that will enable to maximum systematize and automate all phases of software development, significantly increased [6].

AllFusion Erwin Data Modeler (ERwin) is the leading solution for database modeling, for creating and maintaining databases, data marts and data warehousing, as well as enterprise data resource models. ERwin models visualize data structures to simplify data organization and management, for simplification of complex data interactions, database and deployment environment development technologies. This simplifies and accelerates the process of the database development and its quality and reliability significantly increase.

In this paper, the logical design of databases was created with ERwin CASE tools, the model «Entity-relationship» (Fig. 1) was created.

This scheme gives an intuitive overview of the project and is particularly useful for the idea exchange between users. The next step was to test all the operational use of the organization data, associated with its proceccing, and the exclusion of unnecessary or duplicate data.

III. Database Physical Design

Physical design phase is to link the logical database structure and the physical storage environment in order to achieve the most efficient data distribution, i.e. mapping the logical structure of the database to storage structure. The problems of distribute the stored data in the memory space, the choice of effective methods of access to various components of the "physical" database are solved. The results of this phase are documented in the form of storage schemes in Data Definition Language (DDL). Decisions, accepted at this stage, have a decisive influence on system performance.

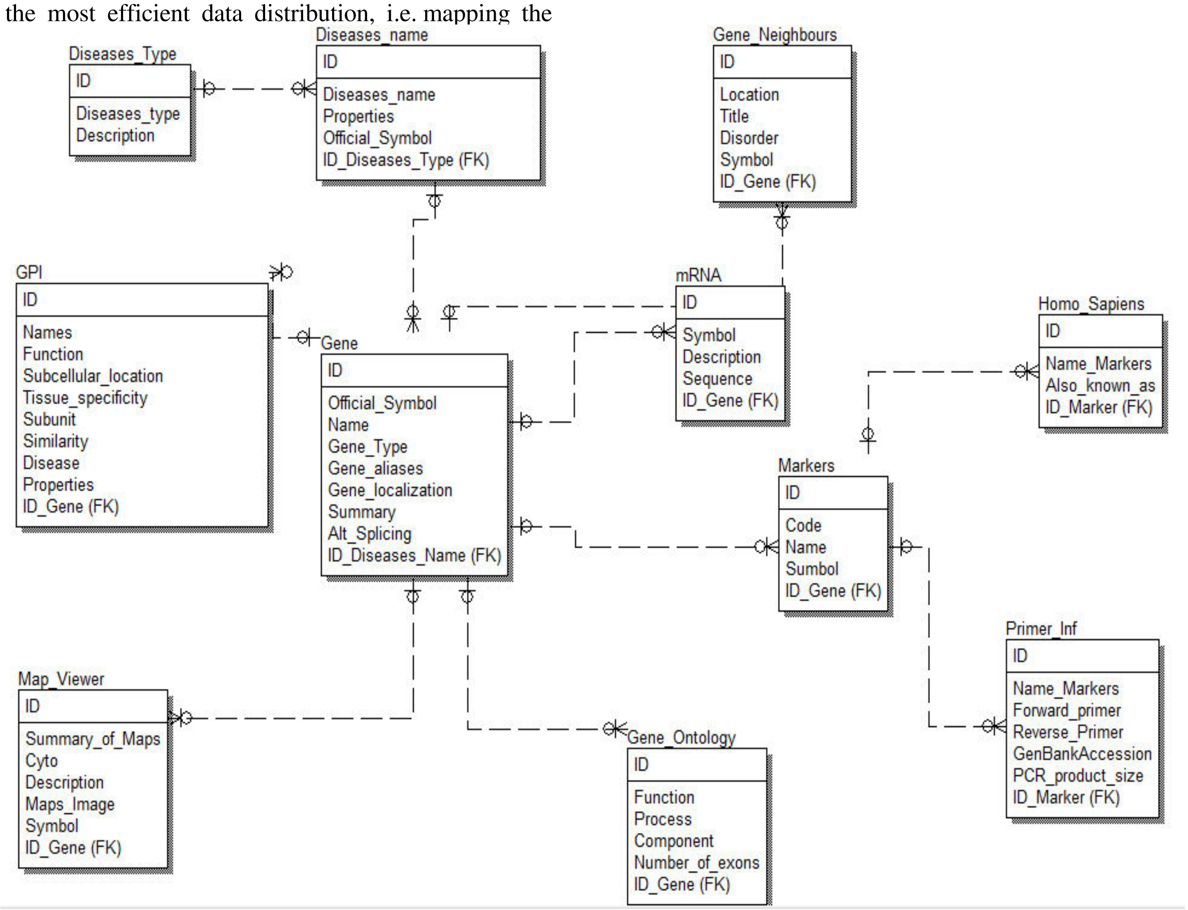


Figure 1. The logical scheme "Entity-Relationship" in ERwin for cardiovascular and oncological diseases.

A. *The physical design of the database*

The physical design of the database is done using the package of professional engineering MS SQL Server. SQL Server database is a relational database that is compatible with SQL (Structured Query Language) with integrated XML support for Internet applications.

Database - is also a place of data storage, but for the most types of data files it does not provide information directly to the user, launches an application that accesses the data base and presents them in an understandable for the user format. Despite the different ways of information in the form of databases, relational database management system (DBMS) is considered to be one of the most effective. In a relational DBMS for efficient data organization mathematical theory, namely the relational algebra, is used.

MS SQL Server has several advantages over other database systems: ease of installation, deployment and operation, as well as scalability, data warehousing and systems integration with other server software. Another factor that influenced on MS SQL Server choice in this paper - it's speed. In relational DBMSs, speed is the time required to execute the query and return the results of processing the user request.

The rapid growth of SQL (Structured Query Language) popularity is one of the most important trends in modern computer industry. Over the past few years, SQL has become the sole language of databases. Today, over a hundred databases support SQL, working on personal computers and mainframes.

B. *Database security*

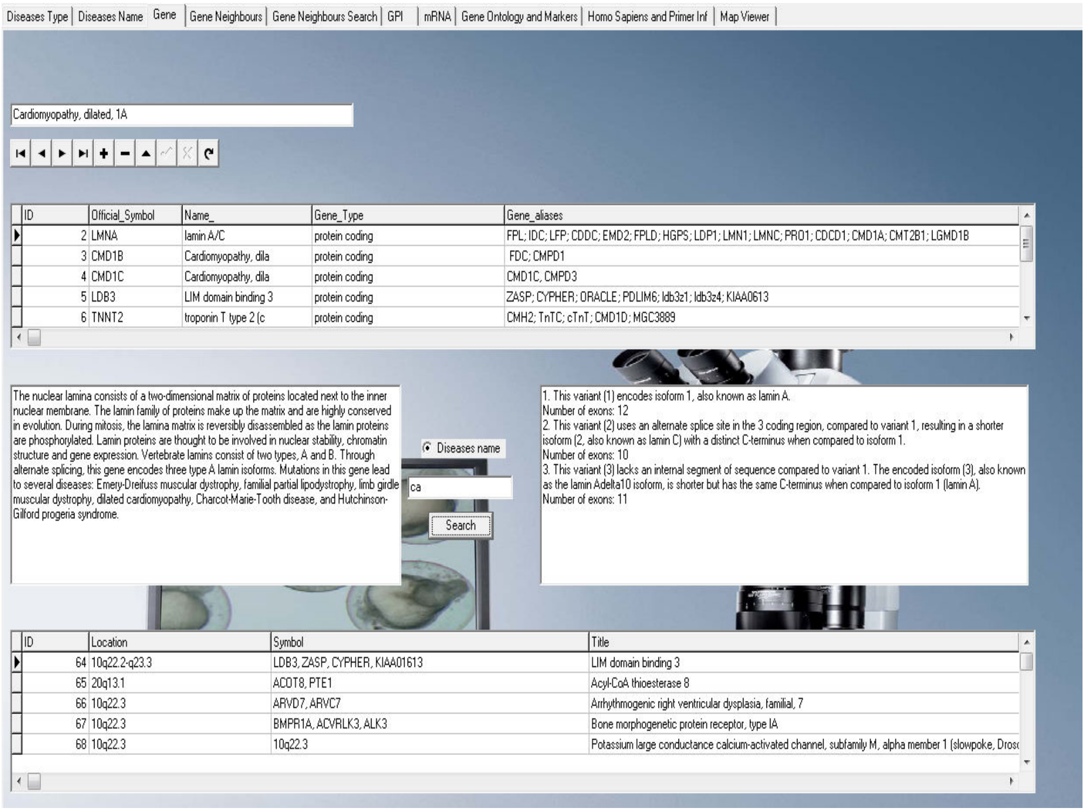
One of the major components of database project is to develop a database security tools. Data privacy has two aspects: security against crashes and unauthorized access. To secure against failures backup strategy is developed. To protect against unauthorized access each user is granted data access only in accordance with his access rights [9].

In this work, carried out by means of SQL Server data security organization is held in standard mode. User accounts are used to control access rights to specific server resources, such as tables and stored procedures. In user records user roles are defined - one or several. User accounts to login into the system as users are created, the user needs enter Logins in the Name field, enter a password in the Password field, select SQL Server Authentication, select database BioInformatics from the Database drop-down list.

C. Client-server application

During the development of distributed information systems the following tasks in the organization of interaction between client and server sides appear: the transfer of personal database to a server for its further shared use as a corporate database; the organization of requests from the client computer to a corporate database, hosted on the server, the development of a client application for remote access to a corporate database from the client computer, the server administration from the client side [5].

In this paper, during the development of client-server application the possibility of work with databases through ActiveX Data Objects (ADO) technology is used [2]. The ADO technology is based on the object model, in which the objects are sets of collections, methods and properties that support databases. The objects of this technology provide the best opportunities for integrating applications with databases.



Cardiomyopathy, dilated, 1A

ID	Official_Symbol	Name_	Gene_Type	Gene_aliases
2	LMNA	lamin A/C	protein coding	FPL; IDC; LFP; CDDC; EMD2; FPLD; HGPS; LDP1; LMN1; LMNC; PRO1; CDCD1; CMD1A; CMT2B1; LGMD1B
3	CMD1B	Cardiomyopathy, dila	protein coding	FDC; CMPD1
4	CMD1C	Cardiomyopathy, dila	protein coding	CMD1C, CMPD3
5	LDB3	LIM domain binding 3	protein coding	ZASP; CYPHER; ORACLE; PDLIM6; ldb3z1; ldb3z4; KIAA0613
6	TNNT2	troponin T type 2 (c	protein coding	CMH2; TnTC; cTnT; CMD1D; MGC3889

The nuclear lamina consists of a two-dimensional matrix of proteins located next to the inner nuclear membrane. The lamin family of proteins make up the matrix and are highly conserved in evolution. During mitosis, the lamina matrix is reversibly disassembled as the lamin proteins are phosphorylated. Lamin proteins are thought to be involved in nuclear stability, chromatin structure and gene expression. Vertebrate lamins consist of two types, A and B. Through alternate splicing, this gene encodes three type A lamin isoforms. Mutations in this gene lead to several diseases: Emery-Dreifuss muscular dystrophy, familial partial lipodystrophy, limb girdle muscular dystrophy, dilated cardiomyopathy, Charcot-Marie-Tooth disease, and Hutchinson-Gilford progeria syndrome.

◉ Diseases name

ca

1. This variant (1) encodes isoform 1, also known as lamin A.
Number of exons: 12
2. This variant (2) uses an alternate splice site in the 3 coding region, compared to variant 1, resulting in a shorter isoform (2, also known as lamin C) with a distinct C-terminus when compared to isoform 1.
Number of exons: 10
3. This variant (3) lacks an internal segment of sequence compared to variant 1. The encoded isoform (3), also known as the lamin Adelta10 isoform, is shorter but has the same C-terminus when compared to isoform 1 (lamin A).
Number of exons: 11

ID	Location	Symbol	Title
64	10q22.2-q23.3	LDB3, ZASP, CYPHER, KIAA01613	LIM domain binding 3
65	20q13.1	ACOT8, PTE1	Acyl-CoA thioesterase 8
66	10q22.3	ARVD7, ARVC7	Arrhythmogenic right ventricular dysplasia, familial, 7
67	10q22.3	BMPR1A, ACVRLK3, ALK3	Bone morphogenetic protein receptor, type IA
68	10q22.3	10q22.3	Potassium large conductance calcium-activated channel, subfamily M, alpha member 1 (slowpoke, Dros

Figure 2. The client side of the database of genes and related proteins, responsible for the development of cardiomyopathy.

D. Client application.

The client application, created in C++ Borland Builder programming environment, is designed for the user to process data, namely: datvretrieval, updating, search [1].

The advantage of data processing proposed in this paper is that for each type of disease the gene localization on the chromosome is specified. Moreover, there is a detailed map of the chromosome area, where the gene is located and the chromosome ideogram. on which you can see the overall structure of the chromosomes and the nature of the bands in the shoulders – the segmentation imade. Also map indicates the total number of genes in this region with their functions description, and the total number of genes on the chromosome. In the protein information graphics and pictures are available.

E. Organization of the search engine.

In large databases it is impossible to perform editing and revising the information without tools for finding the desired record. The easiest way to perform a similar search using a universal method of Locate (), which gives the ability to create insensitive loCasemsensitive and partial key loPartialKey search.

The used search method allows to find relevant information about the genes responsible for disease development, a synonym for the gene-candidate, his marker, map of the chromosome area. To simplify the search, all these characteristics are given in the capital form also. In the descriptions to the protein the list of diseases, for the development of which this protein corresponds, is given. To simplify the search a list of synonyms is given, which avoids confusion when searching for a particular gene or protein.

IV. CONCLUSION

The created database is implemented in the scientific and experimental laboratory of the Kazakh National University named after al-Farabi. Currently, the database is used to conduct a large-scale study of the properties of genes and related proteins, responsible for the development of some CVD and PD, and to create specific targets for a gene - therapeutic drugs. The developed information model is also recommended for use in cardiomyopathy clinic in Almaty for the diagnosis of cardiovascular diseases using molecular markers, contained in the created database.

V. REFERENCES

[1] Arkhangelskiy A.Ya. (2006). Programming in C++Builder 6. Moscow: CJSC «Binom» publishing house

[2] Connoli T., & Begg K. (2003). Databases. Designing, realization and support. Theory and practice. 3rd edition. Moscow: “Williams” publishing house.

[3] Dunayev S. (2006). Access to databases and networking technologies. Practical techniques of modern programming. Moscow: «Dialog» publishing house (Moscow Engineering-Physics Institute).

[4] Komarova F.E. (Ed.) (2004). Diagnosis and treatment of internal diseases. V.1 Diseases of the cardiovascular system, rheumatic diseases. Moscow: «Medicine» press.

[5] Kupcevich Y.E. (Ed) (2008). Programmer almanac. V.1: Microsoft ADO NET, Microsoft SQL Server, access to data from applications. Moscow: «Russian Press».

[6] Maklakov S.V. (2008). Development of an information systems with AllFusing Modeling Suit. Moscow: «Dialog» publishing house (Moscow Engineering-Physics Institute).

[7] Mikheev R.N. (2006). MS SQL Server for administrators. St.Petersburg: «BHV-Petersburg» Press.

[8] Minkin R.B. (2003). Diseases of the cardiovascular system. St.Petersburg: «Acacia» Press.

[9] Miroshnichenko G.A. (2005). Relational databases: practical methods for optimal solutions. St.Petersburg: «BHV-Petersburg» Press.

[10] Organov R.G., & Maslennikova G.Ya. (2006). Mortality from cardiovascular and other chronic non-communicable diseases among the working population in Russia. Cardiovascular therapy and prevention, 3, 4-8.

[11] Sheglova Zh.A., Moldazhanova E.E., Naizabayeva L.K., Boldina G.F., & Turmagambetova A.S. (April, 2007). Logical modeling of information systems for diagnosis of some cardiovascular diseases and cancer. Paper presented at the International Scientific and Practical Conference «Science and Education - 2007», Murmansk, Russia.