Term Project: GenCluE

This project consists of two parts and together with the exam it is the wrap-up of the course Datamodellering en –verwerking 8C020. The project is related to the material which is given in chapters 1 to 6 in the lecture notes. The project is done in groups of at most two students. The solutions for **Part A** have to be delivered on paper no later than **Thursday, November 1, 2007, 18:00 hours**. After the evaluation the solutions will be returned during one of the lectures. The final version of the solutions to Part A and the solutions to Part B, which should consist of a report and an Access database, have to be delivered no later than **Tuesday, November 20, 2007, 18:00 hours**. The report should be delivered on paper and the Access database file on floppy-disk (diskette) or CD-ROM.

Part A

The Gene Clustering and Expression (GenCluE) group that consists of researchers from the Eindhoven and Maastricht Universities investigates gene expression using gene clustering approach. Your task is to write a computer application based on a relational database that automates (part of the) work of this group. The information given below should help you define the user requirements.

To each gene there corresponds a sequence of nucleotides (a string consisting only of the letters A, C, T, G). (For simplicity assume that this sequence is at most 30 letters long). As a result of experiments and via an automated classification the genes are distributed in disjoint clusters. One of the main research questions is to check if the genes that belong to the same cluster also regulate similar processes in the cell. To this end a gene is assigned a set of keywords that characterize its function. For example, in this way it is possible to check if there are words that occur only in the descriptions of genes of one cluster. For each gene there are several URLs to entries about the gene in different on-line databases. Further, for each gene there are possible references to scientific publications where one can find additional information about it. There are different kinds of publications (books, journals, proceedings, Ph.D. theses, etc.) Another possible way to extract meaningful biological information from expression patterns is to observe the expression behaviour of known gene families. There are 26 such families. Of interests are also proteins that are encoded by the nucleotide sequences that are associated to the genes. For those proteins, besides the basic data (e.g. name, amino acid structure), it is convenient to have links (URLs) to public databases from which one can extract further information (e.g., 3D structure of the protein).

The questions given below lead to an Entity-Relationship model, a relational model and some SQL queries.

1. The above description of the GenCluE might be incomplete and ambiguous. Give in a natural language (Dutch, English) a better description for the user requirements. You may fill the missing information and resolve the imprecise formulations yourself.
2. Give an example of a reference (scientific publication) with the associated fields (attributes) that should be stored in the database. (Note: No actual publication is required – you may invent yourself the title and the needed data for the publication.)
3. Make an Entity-Relationship Diagram which represents the genes and the corresponding attributes. Give an explanation for your design decisions.
4. Based on the user requirements from 1 make an Entity-Relationship diagram for GenCluE. The E-R diagram from item 3 is part of the latter. Give the primary keys and the cardinalities of the relationships. Give also a textual explanation of the diagram.

The database should also contain data about the members of the group who are divided into researchers and supporting stuff. Also records are kept for each experiment which usually

contain information about the time when it is performed, the targeted genes, and the participating group members.

5. Extend the E-R model such that these new requirements are included in the model.
6. Identify possible multivalued attributes and transform the E-R model in order to avoid them.
7. Specify for at least three different attributes with different domains the appropriate domain constraints.
8. Based on the E-R model from item 6 make a relational database model of GenCluE.

Given the relational database model from item 8 it is possible to formulate queries with which one could get some relevant information from the database.

9. For a successful research it is important that one has an overview of the related work. Write an SQL query that lists all articles written between 2002 and 2007 that are related to genes from the gene family that contains in its description a given keyword *w* (e.g. atherosclerosis).
10. Often genes that have similar sequences also have a similar functionality. As it was already mentioned above, one can draw some conclusions about the gene functions based on the description (keywords) of the genes. Write an SQL query that given a nucleotide sequence *s* (say "ATTAGTGCC") prints all the keywords that occur in the description of the genes that contain the sequence *s*. Then write a query that gives the URLs for all proteins that are encoded by the genes that contain the sequence *s*.
11. Given a gene cluster *c* (e.g. Collagen) and a member *m* (e.g. Jansen) write an SQL query that lists all the experiments performed in the last five months such that *m* has *not* been involved in those experiments and they are related to genes from cluster *c*.

Part B

Based on the relational database model and possibly the queries form Part A, we can now implement the GenCluE in Microsoft Access. To this end we use tables, queries, forms, reports and a switchboard.

12. Design in Access all tables of the relational database model (item 8 above) and give the relationships between the tables.
13. Design a form for adding genes to the database.
14. Design a form for adding group members to the database.
15. Design a form for adding scientific publications to the database.
16. Design queries (using QBE) that are equivalent (i.e., produces the same result as) the queries from item 9 above.
17. Design queries (using QBE) that are equivalent to the queries from 10.
18. Design a report that prints the members of the group who are researchers.
19. Design a query equivalent to the one from 11 and the corresponding report.
20. Make an original start screen for your application using a switchboard.
21. Design some additional forms, queries, and reports that you think could be useful for GenCluE.
22. Find on the web some gene databases (e.g., GenBank, KEGG) and compare them briefly to your database.

Your report should contain the names of the queries, reports, and forms, and a brief discussion for each of the solutions to the questions which are listed above. Also, please enter some data (not too much!) into the database such that the implementation, i.e., the queries, forms, and reports from above, can be tested.

Good luck!