DATABASE DESIGN I – 1DL300 – Spring 2015

Assignment 1

Conceptual modeling of the Protein Data Bank (PDB)

This first assignment gives you the opportunity to practice the first part of a relational database design process: conceptual modeling using EER diagrams. Before starting this assignment, you must have learned the basic constructs of the EER model and the corresponding methodology. Now it is time to test your knowledge on a realistic scenario, to prepare you to deal with "the real world".

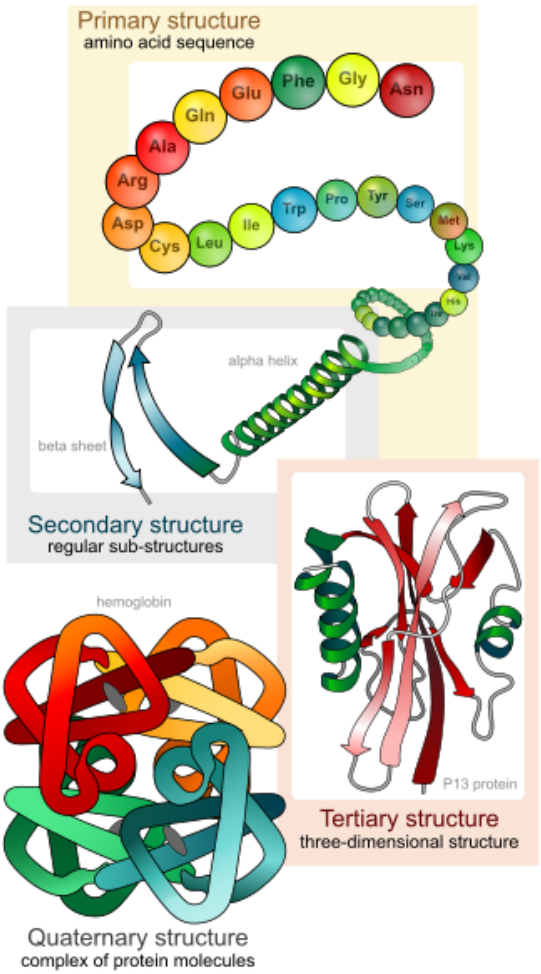
Differently from typical "academic" examples used to learn the EER modeling constructs, real specifications can contain redundant, vague or even inconsistent information, they can refer to a topic you are not familiar with and be long and complex.

This assignment has the following learning objectives:

1) Autonomously apply the theory of database modeling.
2) Extract data modeling requirements from a technical text.
3) Master EER modeling constructs.

Description of the reality to be modeled (also available as a separate text file)

The Protein Data Bank (PDB) is a repository for the three-dimensional structural data of large biological molecules. In this assignment we focus on proteins, even though since its foundation other types of molecules have been included in this data bank. The data are submitted by biologists and biochemists from around the world, and are then freely accessible on the Internet. Your task is to design a relational database for the Protein Data Bank.



Proteins are large biological molecules consisting of one or more long chains of amino acid residues. The protein can be described at different levels of granularity, known as structures. The amino acid sequence forms the so-called **primary structure** of a protein. As an example, a (small) portion of the amino acid sequence of the RIBONUCLEASE A protein found in the pancreas of cattle is:

LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS MET …

where each three-letter code represents an amino acid (LYS = lysine, ALA = alanine, etc.). The amino acids in the sequence are geometrically organized into regularly repeating local structures, forming the **secondary structure** of the protein. Common examples are *alpha helix*, where a sub-sequence of amino acids generates a helical shape, and *beta sheet*. Secondary structures are local, therefore many regions of different secondary structure can be present in the same molecule. The **tertiary structure**, or fold, corresponds to the complete 3D shape of a single protein molecule. This is very important, because the shape of the protein plays a fundamental role in determining its basic function. Finally, multiple protein molecules, usually called *protein subunits*, form the **quaternary structure**, functioning as a single protein complex.

Each entry in the database represents a quaternary structure, which can be composed of one or more entities. We generally refer to a database entry as a protein, also in the case in which it is made of multiple protein molecules. For each protein in the database, we store

the corresponding sequences of amino acids (one sequence for each subunit) and also their secondary structures, that is, sub-sequences of amino acids should be annotated with the type of secondary structure to which they belong. In addition, we want to store the tertiary structure of each sequence as a sequence of atomic positions. More precisely, for each atom composing each amino acid in the sequence we want to store its type (e.g., O or C) and its 3D coordinates, plus one numerical value indicating how stable its spatial location is. We also want to know the length (that is, the number of amino acids) in each sequence and the method used to study the protein, e.g., X-ray crystallography, NMR spectroscopy, or dual polarisation interferometry.

When a protein is submitted to the database, there must be a corresponding research paper or report where the protein structure is first studied and described. Among the authors of the paper, one can be marked as the person submitting the data. If the paper has been published in a journal, we also want to now the details of the journal (ISSN, Name, Publisher), plus the details of the publication (Title, Volume, Issue, Date, Pages). However, the main citation of a protein is not necessarily published in a journal. In addition, we want to store all the PubMed articles referencing this protein. PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a collection of more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. Each PubMed citation contains an article title, the text of the abstract, a set of chemical names, a DOI, the ISSN, Volume and Issue of the journal, pages, publication date, list of authors and a unique PubMedID. Notice that the main citation associated to a protein may not be part of PubMed and not have a PubMedID, but we still want to store it in the database.

For each entry submitted to the database we want to know the title of its main citation. In addition, we also want to know its minor and major version numbers, deposit site, status code, date of last revision, and a set of keywords.

Several textual remarks can be associated to each entry and to each entity composing an entry. Multiple keywords can also be associated to both citations and PubMed entries.

Examination

You must submit the result of this assignment as **one single PDF file per group** via the Student Portal. The report must contain (1) an EER diagram for the PDB database, and (2) a list of *business rules*[1] indicating requirements that cannot be expressed in the diagram, if any. Please consider that all the submitted assignments will be printed and corrected on paper, so indicate group name, group participants with SSNs and draw readable diagrams.

To draw your diagram you must use some drawing/modeling software, and you can choose any software you prefer, on your laptop or on the computers in the labs. The assignment must be submitted before the indicated deadline. Please notice that you are not expected to finish this assignment during the lab: it is expected to take longer time.

Sources

- Protein Data Bank in Europe (http://www.ebi.ac.uk/pdbe)
- Schema of the real PDB MySQL database
- Wikipedia

[1] A business rule is a statement that cannot be modeled using EER constructs but will be needed to be implemented in the final database. Therefore, we need to keep track of these statements. As an example, a business rule may be: "The sum of the salaries of the players of each team cannot exceed € 40 000 000".