Topic: **Databases and Bioinformatics Tools**

A WIKI ANNOTATION SYSTEM FOR BIOLOGICAL DATABASES

SA Silveira[1], CH Silveira[2], W Meira Junior[3]

[1]*Federal University of Minas Gerais - Bioinformatics PhD Program*
[2]*Federal University of Itajuba - Department of Mathematics and Computer Science*
[3]*Federal University of Minas Gerais - Department of Computer Science*

Data provenance, also found in the literature as data lineage, is the process of tracing the origins and recording of data and its movement between databases or the descriptions of the origins of a piece of data and the process by which it arrived in a database. Data provenance is currently an issue of importance to scientific databases, because information about the origin of the data is closely related to quality, validation, reliability of data and the reproducibility of an experiment.

Among the sciences, the field of Molecular Biology is possibly one of the most sophisticated consumers of modern database technology and has generated a wealth of new database issues. A substantial fraction of research in genetics is conducted in dry laboratories using *in silico* experiments - analysis of data in the available databases. Moreover, biological databases are highly dynamic, heterogeneous and have huge amount of data. Thus, data provenance is an issue even more important for this type of database.

In this work the proposal is, first, create a computer system capable of associate each piece of data in a relational database with annotation, allowing queries over data and over annotation, which can be information about the origin of data or even additional information about data. Many databases have a rigid structure, so this system will permit annotate data without the necessity of change the database schema. There are some similar approaches in the literature to address the problem of data provenance.

For our initial tests, we built an experimental relational version of PDB data using the DBLoader tool provided by RCSB PDB. This database was fed whit data from SCOP, specifically the globin family and currently we are moleling a database schema to store annotation related to each piece of data in our relational version of PDB. It is important to emphasize that the system is not specific to PDB database, and can be used in databases of proteins, genes, among others.

Second, we want to provide this system (in conjunction with a database) through a web interface, creating a wiki-based web resource with data and annotations, sharing information and allowing users to add and modify annotations on the data, similarly to what is done in Wikipedia. The interface should be as simple and intuitive as possible, which can be achieved using MediaWiki, the open-software wiki package used by Wikipedia. This will offer users an interface with which they are, in general, familiar. A mechanism for access control will be implemented to ensure that data are not changed indiscriminately. We can find in literature some interesting proposals that allow sharing and editing of data.

This work is included in a broader context of data integration, where there is a project in which we want to use data provenance to create strategies for data integration.