Stroke of GENEous: A Tool for Teaching Bioinformatics to Information Systems Majors

Rahul Tikekar
Department of Computer Science
Southern Oregon University
Ashland, OR 97520
TikekarR@sou.edu

Abstract: A tool for teaching bioinformatics concepts to information systems majors is described. Biological data are available from numerous sources and a good knowledge of biology is needed to understand much of these data. As the subject of bioinformatics gains popularity among computer and information science course offerings, it will become essential for computer science and information systems majors to understand and appreciate basic biological concepts. The tool described in this paper involves the class working as a group on a project to design and develop an online database of interesting genes, proteins, and disorders. Students learn the complexity of life by searching for and finding data to populate the homegrown database. The project is highly extensible thereby making it possible for future classes to extend the database developed in preceding classes.

Keywords: Bioinformatics education, database design, DNA, Genes

1. Introduction and Motivation

The field of information systems (IS) differs from conventional computer science in that the focus of IS is to apply computing solutions to current problems while that of computer science is to improve existing techniques and to design new technologies and algorithms. Hence IS majors are trained to apply computer science techniques to current information technology (IT) needs. Some examples of such IT needs are: designing and administering web and multimedia applications, building and administering computer networks, and designing and administering databases systems. As bioinformatics applies computing solutions to biology, it is important for IS majors to understand the field of bioinformatics and the opportunities that it provides.

Immense biological data is available today generated as a result of many government and privately sponsored projects. The task of compiling and analyzing this data requires designing computational algorithms, databases, tools, etc. As a result there are several opportunities available to qualified individuals and the field of bioinformatics is gaining in prominence. Many universities offer opportunities to study bioinformatics by means of newly created bioinformatics majors or stand-alone courses; some examples are listed in [1-5]. This paper describes a tool used in one such course offered to mostly information systems majors.

One of the ways IS majors can serve the field of bioinformatics is by helping in the design, development, and maintenance of various databases. A major component of an introductory course in bioinformatics, beyond understanding the central dogma of molecular biology, is browsing through databases like GenBank and SWISS PROT. To understand and appreciate the entries in these databases requires a good background in biology – indeed they serve to help researchers in the field and not people with an interest in it. So it is no wonder that a typical IS student can get lost while looking for information in the many databases available today, not only because of the complexity of the data (trying to understand the description of a gene, for example) but also because of the amount of it. Similar attempts to help in the understanding of these concepts are described in [8, 9].

As an example consider the case of the COMT gene in humans. The objective is to understand the gene and the protein for which it codes. A search of Entrez Gene (www.ncbi.nlm.nih.gov/Entrez) for COMT produces over 20 results in two tabs: Genes Genomes and SNP GeneView. Next, a search of COMT in SWISS-PROT produces over 290 entries. Choosing one of the entries (COMT_HUMAN) produces a long listing of the protein's properties and the associated cross listings. Selecting one such cross listed entries from GenBank produces a report that includes the

DNA sequence that specifies part of the protein, or sometimes, the protein itself. While the entries provide a wealth of information to the trained scientist, it is very challenging for an IS student to understand the information that is contained in the various reports.

1.1 Pedagogical Motivation

Numerous studies have been performed that support the notion of active and collaborative learning as a better model of teaching than the conventional model of lecturing. These are explained and summarized in [6] and an example applied to computer science is presented in [7]. This formed the author's secondary motivation: to use concepts of active and collaborative learning to teach the concepts of bioinformatics. By getting the students to collaborate on a common project, one that involved concepts that were new, yet fascinating, it is possible for them to learn the subject better. It is also very likely to make the class more interesting. The computer science department's computer lab contains 20 workstations running Windows and Linux with access to the Internet. As the size of the class was small – 15 students – it was feasible to adopt the active learning approach as described in [7].

2. Instructional Approach

At its heart the concept of bioinformatics is simple – to apply computing techniques to problems in biology. The author discovered that getting this point across can be greatly enhanced by taking an analogy from the realm with which all IS majors are familiar: business. Business informatics, known more popularly throughout the world as eCommerce, is the branch of knowledge where computing techniques are applied to business processes. Students understand how the Internet has revolutionized businesses – from buying online to online auctions to online banking, the Internet and sophisticated computer algorithms and databases have made business processes more efficient and convenient.

To understand and appreciate a similar revolution affecting biology, however, is more challenging. This is primarily due to the fact that in the case of eCommerce most people are already familiar with the different business processes of buying, selling, banking, auctions, etc. Concepts of Biology, however, are not something that one uses on a regular basis and understanding biology related to genes, proteins, and diseases requires more than just a passing reference to these concepts. Indeed, students take courses in Chemistry, Physics, and Biology before they can appreciate the issues involved. Hence teaching bioinformatics to IS majors requires a major redesign of the teaching approach.

The approach adopted by the author involved using tools that IS majors understood and with which they were familiar: databases and web development. In order to promote active and collaborative learning, all the students were asked to participate in a class project – to build a database of genes that people could query, via a web application, and obtain information that was presented in lay terms. The first four weeks were spent in lecturing on the concepts of genetics – DNA, genes, chromosomes, proteins, etc. The students understood the central dogma of molecular biology. Lab time was used to search the different biology databases like ENTREZ and SWISS-PROT, though the students did not quite understand all of the information provided by the databases. It was now time to start designing the database.

The class was divided into teams of two. In addition to the teams, one student who through a stroke of providence was a professional software developer, acted as the project manager. Each team would be responsible for a task (like getting information about a gene) and would report to the project manager. The task assigned to each team would depend on the status of the project – at the design stage, for example, the task would involve getting information about various databases available and the data they provided. At the start of every class there would be a "meeting" to discuss the progress and plan the strategy and task allocations for the next day or week. This meeting time allowed the teams to suggest ideas and promote interaction. The title of the database – Stroke of GENEous – was a result of one such meeting.

3. Database and Application Design

In order to design and build the database, it was necessary to decide on the type of information that will be in the database. The author suggested that the database should be centered on the gene as most people know genes. The objective of the project would be to implement a database through which lay people could navigate and learn about genes. The Internet would be the most convenient platform on which to implement an application – the students were comfortable with building web database applications.

In order to design the conceptual database, an accurate albeit simplified model of the central dogma was adopted:

- Human cells contain DNA that consists of nucleotides containing bases (ACTG).
- This DNA is arranged in a set of 23 pairs of chromosomes.

- Genes are sequences of nucleotides; that is genes may not span chromosomes.
- Genes have names (e.g., COMT) and usually serve a function (like protein coding).
- Most genes encode for proteins – the gene itself is a coding sequence.
- Proteins are composed of amino acids.
- Proteins are expressed in tissues.
- Mutations in genes can result in disorders.

The resulting ER Diagram of the conceptual database design is given below (Figure 1).

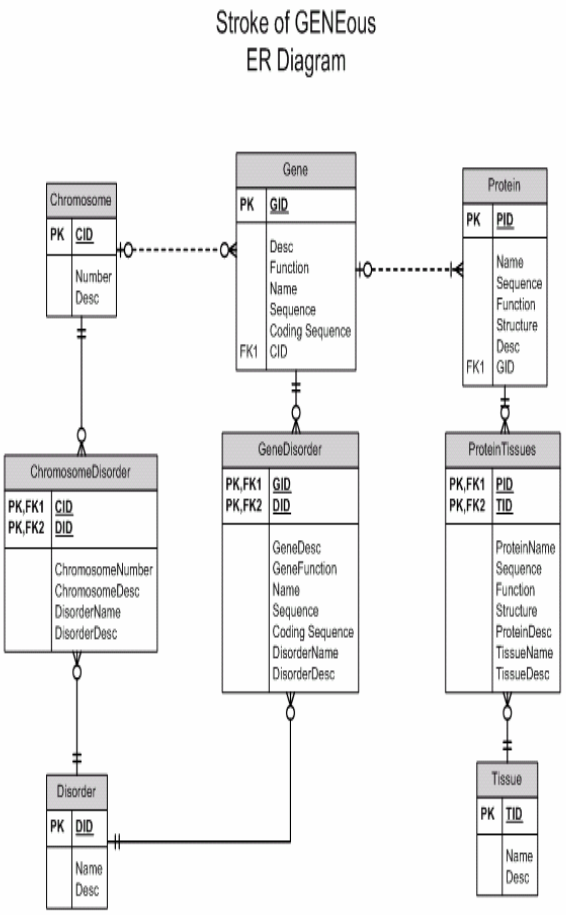

Figure 1. Conceptual Design of the Database

The database design has captured the following properties of the model:

- Mutations in genes can be responsible for many disorders; disorders can be associated with chromosomes as well.
- A gene may code for multiple proteins – to account for alternate splicing. However this model does not capture the alternate splicing sequences nor does it capture protein isotopes. This extension can be incorporated into future classes.
- A protein may be expressed in many tissues and one tissue may utilize multiple proteins. Each protein is assigned a function (e.g., transport, enzymatic, receptor, etc.)

4. Implementation

The application was implemented as a web application using the ASP.NET development platform while Microsoft Access was used as the backend database to store the data. The class was divided into teams of 2 students. Each team was assigned the task of gathering information relating to three genes. In order to get this information, the students had to consult the many online databases. One team was given the additional responsibility for designing and building the application (front-end and back-end). As per the original vision of the project, the 3-D structure of a protein was to have been in the database; however this part was not implemented since visualizing and understanding the 3D structure of a protein was covered toward the end of the term.

Since Microsoft Access stores the database as a single mdb file, it is possible to circulate this file among the different teams. As one team gathered sufficient information to populate the database, they would download the database file, enter the data into the database, and then upload the file to the server. A text file on the server was used to indicate if the database file was being used by another team. This prevented another team from downloading the database file if another team was working on it. During this term no particular strategy was adopted to identify "interesting" genes. Each team decided to choose a gene that they found to be interesting. At the beginning of each class a quick check ensured that two teams did not pick the same gene.

Two screen shots showing the execution of the application are shown below. Figure 2 shows the first page of a list of genes. Clicking on a gene (CFTR) leads to a screen which provides more information on the gene and also provides its coding sequence (figure 3). As can be seen the application is not complete and the database is missing vital pieces of linking information. Also the descriptions entered into the database are not in lay terms – something a novice may comprehend. The goal of future courses is to fill these missing pieces and build a more complete application.

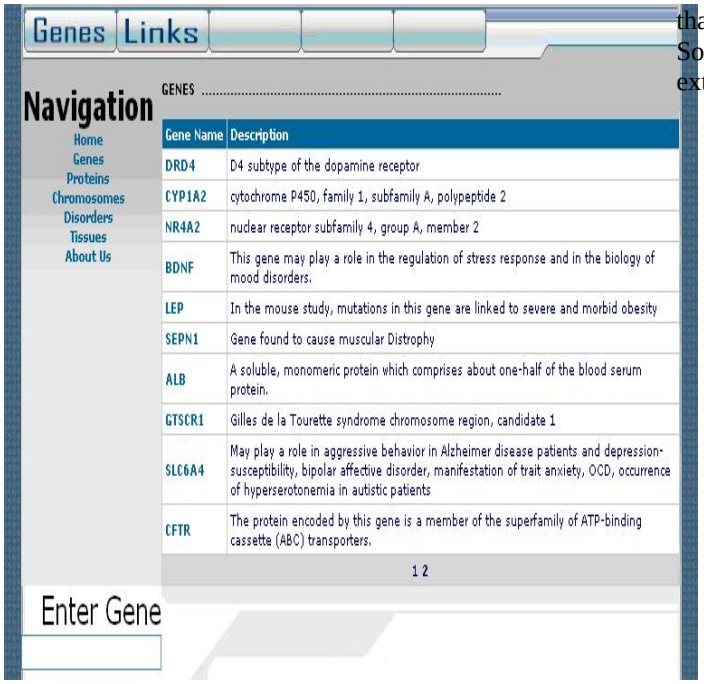

Figure 2. A List of Genes

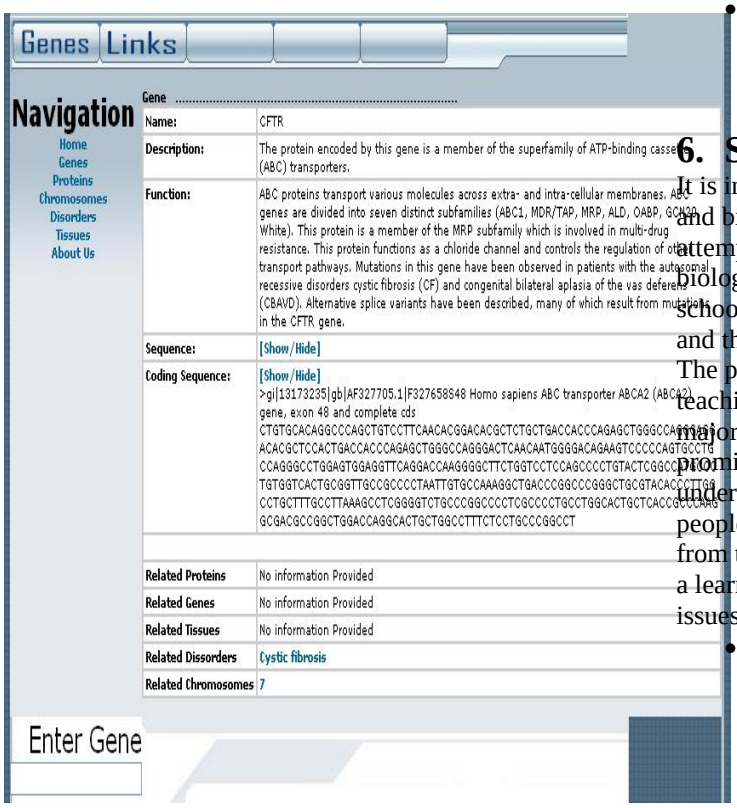

Figure 3. Details of the CFTR Gene

5. Extensibility

One of the features of this project is extensibility – its ability to grow in several ways. At the simplest level, the database can be populated with more data. Since the domain of biology is large, there is plenty of data that can be found in order to populate the database. Some of the other ways in which the project may be extended are listed:

- The database schema could be enhanced to support additional relationships and entities. For example, one might add more information relating to chromosomes or additional species could be incorporated into the database.
- Educational material can be provided to better explain some of the concepts. Small introductory lessons and links may be added that can provide information to someone desiring to learn more.
- Annotations may be made to cite references from where information in the database was obtained so future projects might use the same references.
- Visualization can be a very useful and powerful pedagogical method. Students with knowledge of graphics design and multimedia tools can provide useful animations of biological processes, like that in [10].
- Advanced users could be provided the ability to use tools like BLAST from within the application.

6. Summary and Conclusions

It is impossible to teach the importance of biology and bioinformatics in a single course. This course attempted to re-introduce the students to the world of biology that most had studied during their high school years and to show them of the advances made and the opportunities available for the enterprising. The project described played a very useful role in teaching biology and bioinformatics to students majoring in information systems. The project shows promise of growing into a very useful tool for understanding genes, proteins, and disorders for lay people. Students enjoyed the course as evidenced from the evaluations. The project can also be used as a learning tool in future classes. There are several issues that will require resolution:

- In order for the application to provide meaningful information to the user, there needs to be more data in the database – this will provide useful links from one part of the database to another. For example, with sufficient data it will be possible to browse a gene, find the protein for which it codes, look at the tissues where the protein is expressed, and then find other proteins that are expressed in the same tissue – this will lead back to other genes.
- The nucleotide sequence for many genes is quite large and it may be difficult to

incorporate it into a database like MS Access since it increases the size of the database significantly. Also, a better software tool might be useful in presenting this sequence.

- There are several versions of the coding sequence as found on GenBank; figuring out which one to incorporate into the database can be a challenge.
- Providing an overview of a gene in lay terms can be a challenging task and would require consultation with a biology staff member.
- Classifying genes into phenotypic groups like eye color, stomach disorders, etc. can provide greater search flexibility – this would also require consultation with members of the biology department.

7. References

1. The University of California, Santa Cruz, Bioinformatics Major, *http://admissions.ucsc.edu/discover/majors/Bioinformatics.cfm*
2. University of New South Wales, Bioinformatics Major, *http://www.cse.unsw.edu.au/undergrad/programs/BINFA13647.html*
3. Wellesley College, Bioinformatics Major, *http://www.wellesley.edu/CS/courses/CS-BiSc303/themajor.html*
4. Johns Hopkins School of Medicine, *http://pevsnerlab.kennedykrieger.org/bioinfo_BCMB2005_course.htm*
5. Department of Computing, Imperial College, London, *http://www.doc.ic.ac.uk/~sgc/teaching/341/*
6. Faust, J. & Paulson, D. "Active Learning in the College Classroom", *Journal on Excellence in College Teaching*, 9(2), 3-24, 1998
7. Walker, H.M., "Collaborative Learning: A Case Study for CS1 at Grinnell College and UT-Austin", *SIGCSE Bulletin*, Vol. 29, No. 1, March 1997, pp. 209-213
8. Almeida, C.A. et al., "Using Bioinformatic (sic) Software to Understand the Central Dogma of Biology", *Bioscene*, Vol. 29(2), May 2003.
9. Benz, S. et al., "Genomics Research and the Liberal Arts: Building a Database for Exploring Your Favorite Set of Genes (favGene v2.0)", *Transformations-Liberal Arts in the Digital Age*, v2(1), May 2004.
10. *www.JohnKyrk.com*