COURSE DESCRIPTION

Goals:

a. Students will learn where and how to obtain information from public biological databases and how to assess the quality of that information.
b. Students will learn what a data model is and how to design, implement, populate and query relational databases to support their own research, which often requires the integration of public data with new data, and merging data from multiple sources having different formats.

Methods:

This course is taught as a Research Methods course, in which students learn practical research skills, including the theory of the relational model and the practice of creating a well-constructed database within a relational database management system. Students will build the skill set needed to create a research database to support scientific inquiry; students also learn methods required to access, retrieve and integrate data from diverse public databases as well as query that database to uncover new information.
Background:

There are dozens of large public repositories of biological data and thousands of smaller databases. These data sources use a variety of designs, including hierarchical and relational schemas. Regardless of the underlying model, these are often housed within relational database management systems. Successful data retrieval is very dependent on implementation of the model as a schema and in how data is exposed to users. For some databases screen scraping is the only mechanism for retrieving the data, while access to a schema, SQL and APIs permits more effective use. Serialization methods are important to maintain structure during data transfer, we will discuss XML and XML Schema as examples of ways to accomplish data transfer without loss of organization.

- Because the generation of biological data is very complex and our understanding is not complete, there are many valid data representations. Throughout the course we will consider the availability of representation standards and their proper use, taking as our exemplar the Generic Model Organism Databases as a standardized but extensible model, with the associated toolkits.
- In the framework of the relational database management system the language used for creating and retrieving data is SQL, to which students will be introduced. As students become familiar with SQL and other programming tools they will be expected to create a project database in which they demonstrate their abilities to locate, acquire, integrate and interpret data from public repositories.
- Particular topics will include formats and schemas in important biological data repositories, as databases (GEO/SRA, Expression Atlas, PDB), or knowledge bases (KEGG, PharmGKB).
- We will also discuss biological ontologies (such as SO and GO) and how their deployment affects database design, integration and queries.

PRE/CO-REQUISITES

- The lecture and lab are co-requisites.
- Prerequisite: It is assumed that students have some background in cell and molecular biology, such that they understand the basic biological units being modeled and how they relate to one another. No prior programming or database skills are required.
- Pre-requisite: Admission to graduate standing and undergraduate training in Computer Science or one of the Life Sciences
LEARNING OBJECTIVES

Having successfully completed this course, the student will be able to explain and apply:

- The relational model and Codd’s rules; normalization.
- Standard conventions for communicating relational data models.
- Install and use a freely available RDBMS.
- Load data into an RDBMS using several methods and the verify outcome.
- Use SQL for data retrieval from public databases, perform data transformation and load data into a personal database.
- Interpret and use BioOntologies and XML; explain foundations of interoperability.
- Data Cubes and other non-normalized systems are used for establishing Knowledge Bases.
- Create a research database of his/her own design; implement and demonstrate its use.
- Present a working version of the research database in the context of a scientific problem that it solves.
- Write a paper describing the project database in a journal format.
- Research ethics with a specific emphasis on how the interoperability of disparate databases can lead to unintentional invasion of privacy.

INSTRUCTIONAL METHODS

The course is divided between lecture and lab, with more lectures at the beginning.

- Concepts are presented primarily as lectures elucidating important concepts presented in detail in reading assignments, with presentation of factual material in a standard lecture style, and discussion of alternative approaches presented in readings.
- Labs will include explanation of methods by the instructor, demonstrations of methods using software applications by the instructor or TA, followed by supervised student activities in which they carry out tasks using the methods.
Expected Class Work:

Reading
- Papers, blogs, links to demonstrations, etc. about data models, schemas, management systems, integration methods and new approaches will be assigned weekly.

Problem Sets
There will be 5 problem sets, which are aimed at having you practice specific technical skills.
- Use a design tool to create a conceptual model of a data model
- Install SQLite and a Browser, use documentation to understand the application
- Answer questions about Relational Algebra
- Practice writing SQL queries on a class database
- Practice interpreting XML and ontologies

Homework
There will be 5 homework assignments, which help you use skills to build up the components necessary to create your project database.
- Critical Evaluation of a Database
- Identifying and mapping key data elements in GEO and SRA files
- Conceptual Design of your Project Database – iteration 1
- Normalization of your Project Data Model
- Creation and Testing of your Project Schema

Exams
There will be one midterm (given during class) and one final exam (turned in during the scheduled final period) covering the lecture content. Questions are multiple choice or short answer with one or two using SQL as the DML or DDL problems to solve.

Student Projects:
Construct a database that supports scientific research using genomic data, in an area decided each year by the professor.
**GRADING**

Students will be evaluated on their ability to synthesize lecture and lab material effectively, including their ability to answer factual questions regarding lecture and reading material on exams, their ability to practice specific skills in problem sets, their ability to reason about novel problems based on that material in homework assignments, and their ability to put into practice concepts and skills to create a project, which must be documented and presented live to the class. PhD students are required to undertake more complex projects, integrating additional information into the schema and carrying out more complex queries for their presentations.

- Homework 20% (5)
- Problem Sets 20% (5)
- Midterm exam 15% (Comprehensive)
- Final exam 15% (Comprehensive)
- Project Presentation (live) 15%
- Final Project Report (documentation) 15%

**TENTATIVE SCHEDULE (WEEKLY, OR TOPICS)**

<table>
<thead>
<tr>
<th>Date (Week)</th>
<th>Subject/Topic</th>
<th>Assignments Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Introduction:</strong> Scientific Data and Databases, Standards, Exchange methods, Ontologies</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Genomics Experiments and Data types; Methods and Measurements</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data Models and DBMS</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Entity Relational Modeling</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>The relational model and Normal Forms</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SQL and SQLite</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Generic Model Organism Databases and CHADO</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Extract/Transform/Load and Advanced Queries</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Ontologies: Gene Ont. and Sequence Ont.</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Resource Description Framework and OBO</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Serialization methods for structures data transfer</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>NoSQL databases, Data Cubes and Knowledge Bases</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Responsible Conduct of Research</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Presentations of Projects</td>
<td></td>
</tr>
</tbody>
</table>
Note on textbooks: I do not require a text, as there are none that use scientific applications as the teaching paradigm. There is an on-line text that I can recommend:


and I will assign content from specific Web sites as well as a fair number of journal papers. For examples of current scientific databases I strongly recommend looking at recent editions of the Nucleic Acids Research database edition that appears annually, as well as on-line documentation of the RDBMS, tools and specific databases that we discuss in class. The basics of the relational model and relational schemas and SQL are covered in almost any beginning database text – I will provide likely chapter titles to guide your reading.

Good information may be found in the following texts, available in the library, relevant chapters will be cited for each lecture:


Note on Lab Applications:
You will be required to install and learn to use
- SQLite v 3.8.7.1, a single-file RDBMS
- Database browsers SQLite Manager v 0.8.1 (a Firefox plugin which is light-weight) and DB Browser for SQLite v 3.4.0, which has many more features
- MySQL Workbench v 6.2.3, which is a data-modeling tool.

We will demonstrate the use of these tools during the lab section of the course, but you are also expected to install and use them on whatever computer you are using to carry out your project.
POLICIES AND PROCEDURES

ACADEMIC INTEGRITY

All students are required to read and abide by the Code of Student Academic Integrity. Violations of the Code of Student Academic Integrity, including plagiarism, will result in disciplinary action as provided in the Code. Definitions and examples of plagiarism are set forth in the Code. The Code is available from the Dean of Students Office or online at: http://www.legal.uncc.edu/policies/ps-105.html. A set of links to various resources on plagiarism and how to avoid it is available at the UNCC Library website: http://library.uncc.edu/display/?dept=instruction&format=open&page=920.

ATTENDANCE

I do not take attendance. I also don’t go over lecture material for students who miss lectures (barring illness), office hours are for those who have questions about a specific lecture point or homework and practice problems or for help with the project. Note that there is no official textbook that carries the burden of delivering the bulk of instruction so that might weigh into the attendance decision.

OTHER

Students are encouraged to use computers during the lecture and lab periods for note-taking and other course-related activities. However, the use of cell phones, beepers or other communication devices, and using class laptops for email, social media, etc. is disruptive to your concentration and your neighbors – hence these uses are prohibited during class. Except for emergencies (for which the instructor has been alerted in advance), those using such devices for those purposes must leave the classroom for the remainder of the class period.