Development of a Database Course for Bioinformatics

Zhengxin Chen

College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182-0500, USA

Abstract

Biology databases (bioDBs) play an ever-increasing role in the post-genome era for bioinformatics education. In order to handle the increasing demand for bioDBs in bioinformatics education, since Spring 2005, the College of Information Science and Technology (IST) at University of Nebraska at Omaha (UNO) has offered a course entitled “Database Search and Pattern Discovery” for seniors of our undergraduate Bioinformatics program (which is also available for graduate students majored in Computer Science program) on an annual base. This paper summarizes considerations behind development of this course, describes how the course is structured, reports our findings and observations, and raises issues of urgent tasks. The main objective of this paper is to exchange our thoughts of bioinformatics course development and solicit feedbacks and suggestions from colleagues working in bioinformatics education and related fields.

Keywords: Bioinformatics; Databases; Information retrieval; Pattern discovery; Machine learning; Text mining

1. Introduction

Since Spring 2005, the College of Information Science and Technology (IST) at University of Nebraska at Omaha (UNO) has offered a course entitled “Database Search and Pattern Discovery” for seniors of our undergraduate Bioinformatics program (which is also available for graduate students majored in Computer Science programs) on an annual base. Overall 47 students (including 21 graduate students, most of them were in Computer Science program, along with several PhD students in Information Technology) have taken the course. We believe now we are at a crucial moment to evaluate what we have done and to receive critical comments from colleagues in bioinformatics community. Although there are websites (such as http://www.nslij-genetics.org/bioinfotraining/) provide information about bioinformatics curriculum development worldwide, the detailed descriptions of individual course developments are hard to come by. Therefore, we hope we can use ICCS as an opportunity to exchange thoughts with colleagues in this field. This paper summarizes considerations behind development of this course, describes how the course is structured, reports our findings and observations, and raises issues of urgent tasks. The main objective of this paper is to exchange our thoughts of bioinformatics database course development and solicit feedbacks and suggestions from colleagues working in bioinformatics education and related fields.
A glance of UNO’s bioinformatics curriculum is available at UNO http://bioinformatics.ist.unomaha.edu/. In addition to other requirements (such as humanities, etc.), required courses in bioinformatics include 3-credit hour courses for each of the following: Intro to Bioinformatics (freshmen level), Bioinformatics Programming (junior level), Applied Bioinformatics (senior level) and two senior level courses: Bioinformatics Algorithms, and Database Search & Pattern Discovery in Bioinformatics. In addition, before graduation, each student should take 1-credit hour Seminar (Colloquium) in Bioinformatics and 3-credit hour Senior Project in Bioinformatics (arranged in two consecutive semesters).

In order to explain our main considerations behind the course development, we start with a general remark on role of database education in bioinformatics program. Bioinformatics, or even the entire biology, can be considered as an information science, and bioinformatics is database-centred. It is unthinkable for bioinformatics education to bypass databases at all. So what is the place of a database course in a bioinformatics education program? Some may argue that the basics of database management systems can be covered in a couple of sessions in fundamental courses such as Introduction to Bioinformatics or Applied Bioinformatics, along with the introduction of most popular biological databases. Throughout the bioinformatics program, students are gradually exposed to different biological databases in more depth, as well as additional aspects of database systems (such as XML). Since basics of databases knowledge can be absorbed into various courses during the entire span of bioinformatics curriculum, a separate database course seem to be unnecessary. For those students who are interested in more depth of database management systems (DBMS) and machine learning/data mining, they can take elective courses which are widely available in our college or other colleges of our university. However, the opinion presented above is questionable, because according to this opinion, students can only get scattered fragments of the whole picture of DBMS from various courses. Since the future of bioinformatics is closely tied to databases, a dedicated database course in bioinformatics curriculum at both undergraduate and graduate levels is a recommended choice, particularly at the current post-genome era, when huge amount of biological data need to be consolidated in various databases and effectively searched and analyzed.

A database course for bioinformatics should be a bioinformatics-centered database course. But what consists of such a course? We should keep in mind that for many biologists, the meaning of the term “database” is much broader than what people in IT (information technology) can usually think of; because a “database” in biology may loosely refer to many different things, such as a data source containing only flat files. In fact, instead of “standard” way in IT field of querying databases using SQL or other commercial languages, traditionally search on biology databases (bioDBs) is based on keywords (such as accession numbers or sequences), which resembles more to information retrieval (IR) on unstructured documents than database queries. Therefore, a database course offered to bioinformatics students should be based on the broader sense of databases. In other words, databases discussed in bioinformatics context is not just the traditional databases discussed in database community (such as relational databases), but also other forms of data storage involving unstructured and semistructured data, so long as they are useful in biology and bioinformatics. In addition, loose use of certain terms such as “data mining” (sometimes it simply refers to advanced search) and “text mining” (sometimes it simply refers to information retrieval) as observed in various fields of life science may cause even more confusions, if not treated appropriately in a bioDB course. All of these issues, along with many other aspects, must be carefully considered when we develop a database course for bioinformatics.

2. Overview of database course development

2.1. Basic principles

Contents of a course for bioDBs are determined by two basic topics: How to build databases for bioinformatics and how to use databases for bioinformatics. From this consideration we have developed three basic principles for course development:

• **Principle 1: Effective storage and search of the stored biology data.** Note that we stick to with the broad-sensed meaning of bioDB as stated before. This principle helps us resolving a dilemma which involves the limited time and completeness and accuracy of select materials, because it tells us the necessity of tradeoffs
between accuracy and completeness from a rigid database perspective and the usefulness for bioinformatics in selection of course materials.

- **Principle 2: Integrative bioinformatics.** The complexity of biological data is not only demonstrated in various forms of data (such as structured or semistructured), but also in the form of integrated use of data, which is part of the considerations behind integrative bioinformatics. The database course should cover basics of data integration based on important concepts such as ontology.

- **Principle 3: Effective use of stored data.** Effective use of the data contains two huge topics: On one hand, there are some issues intrinsic to the nature of biology (such as gene discovery), where many methods have already been developed by researchers and practitioners in biology and bioinformatics community. On the other hand, there are also many cases where basic ideas of machine learning/data mining (such as clustering, association) and text mining are “imported” from IT field and applied to various problems in bioinformatics research. A database course in bioinformatics must reflect this dual aspect for effective use of stored data.

2.2. Overview of main topics

As the title of this course “Database Search and Pattern Discovery” suggests, there are two main parts of this course: database part and pattern discovery part (which is largely related to machine learning or data mining). The main topics of the course contain the following:

- Basics of biodatabases (bioDB) and bioDB search
- Basics of database management systems (DBMS)
- Basics of information retrieval (IR) and text mining and applications in bioinformatics
- Basics of biodata integration and data warehousing
- Basics of pattern/gene/protein discovery
- Basics of machine learning/data mining applications in bioinformatics

3. Course modules
We have identified following basic modules, each takes 1 or two weeks. Each module may also have its own options or extensions.

Module 0: In order to make this course somewhat self-contained, we designed this module to introduce minimum background of biology (and terminology). Although it is optional, it is critical to those CS students who do not meet basic background in biology. A website containing a tutorial was developed in early stage of course development.

Due to the complex meaning of the term “database” (as explained above), we have designed two modules to cover basics of database materials: Module 1 covers narrow sense DBMS, while Module 2 covers other aspects related to biological databases (such as information retrieval).

Module 1: DBMS basics and application interface, which covers conceptual design (focusing on entity-relationship or ER modeling), SQL, WAMP, XML and others. Options and extensions include Enhanced ER, relational algebra, bioinformatics extension/variation of XML, etc., so long as there have been related development pertaining to bioDB design or applications.

Module 2: Basics of information retrieval (IR), text mining, and ontologies. Note there are three different concepts, but they are related. Including both of them in the same module is mainly due to convenience of handling lecturing time.

Module 3: Based on our understanding of various aspects of databases covered in two previous modules, in this module we examine the actual biological databases, with a focus on search and alignments (both theory and practice). We also allocate sufficient time to practical issues such as understand data format, search result evaluation, etc.

Module 4: Integration of Biodata and data warehousing. In fact, these are two different topics, but they are related. For biodata integration, we review how integrated use of databases have been carried out in the practice of bioinformatics, and study effective ways of supporting future data integration. In addition, data warehousing provides an alternative approach for database integration and analysis, and is an important part of bioDB education.

The second part of this course is on pattern discovery, which is closely related to machine learning (or machine learning). According to [12] (Chap. 2), there are two major tasks for machine learning in bioinformatics, i.e., exploration of existing data mining tools for biodata analysis and development of new, efficient and scalable methods stimulated by recent progress in data mining. Since there are issues intrinsic to biology (i.e., no parallel from other science disciplines) and there are general principles and methods of machine learning or data mining from which bioinformatics can borrow important hindsight (such as classification, clustering or association), we have designed two separate modules, mainly from a pedagogical perspective:

Module 5: Gene and protein prediction. This is an example of taking care of intrinsic biological problems as mentioned above. After a brief review of the prediction problem itself, we introduce basics of Hidden Markov Models (HMM), Bayesian belief Networks (BBN) and support vector machines (SVM) and related materials from selected research papers to illustrate how these methods can be used effectively. (There may be some overlap with other bioinformatics courses such as the algorithm course. But the emphasis of this course is more on applications using existing tools or software.)

Module 6: Machine learning/data mining methods for bioinformatics. We introduce basic idea of machine learning/data mining, including classification, clustering and association, and discuss their applications in bioinformatics through research papers.

Module 7: Case studies. This is a flexible part, because the instructor can select any useful materials as fit. For example, due to the eminent threat of avian flu to public health around 2007, for a couple of years, we have arranged related materials into the lecture, and encouraged students conducting term projects around the theme of surveillance and control of avian flu. In general, we feel a kind of flexibility is needed in offering a database course in bioinformatics.

4. Course assignments

There are three types of exercises:

- Written assignments: Such as designing a simple database using ER diagram on biological objects chosen by students themselves.
• Programming assignments, such as building a simple database and perform queries using SQL, WAMP or XML, etc. These assignments may pave the way for students working on their term projects in the later stage of the semester.
• Hands-on practice, including search of public databases such as GenBank and PDB and study the important features of these databases, explore PubMed website, and use software or tools, such as HMMER or SVM to analyze sequence data. Based on skills learned from these exercises students can further use these tools in their own term project.

5. Course projects

5.1. Why term projects

Since this course is almost the last course in the bioinformatics curriculum, it is an ideal place for students to practice their knowledge of what they have learned in the bioinformatics curriculum, with all kinds of research activities around databases. Students can start building basic research skills which can be further practiced and improved in senior projects in bioinformatics.

5.2. Project topic selection

In general, there are two ways of picking a topic for course project. In order to encourage students to start conducting their own research, at the beginning of the semester, we have asked students to identify topics or biological objects they are interested by themselves. An alternative way is for the instructor to suggest a particular research area so that students can identify subtopic within that area. For example, twice during the course offering, students were given opportunities to work on issues related to surveillance and control of avian flu (a funded research project conducted at UNO at that time), a critical issue related to public health.

5.3. How to conduct student projects

Term projects are arranged at the second half of the semester (although the students are informed about this at the very beginning of the semester and various preparation steps are taken at the first half of the semester). There are several stages involved in each term project, including project proposal, progress report (weekly or biweekly), final report, and final presentation. Each project is evaluated based on performance of all stages (not just the final result). In addition, each student is responsible for the basic idea of projects conducted by other students, so that all term projects as a whole, serve as an important part of required course materials.

5.4. Sample student projects

Below is a list of student projects in 2009 and 2010. Note that since most students were undergraduate students who were at very early stage of conducting their own research or academic exploration, in some cases the titles may be inflated or did not appropriate reflect the actual contents (even this author had tried to guide students to make appropriate titles). Nevertheless in general these topics demonstrate a wide spectrum of student interests.

• An Oncogene Database for Querying Different Types of Cancers and their Responsible Genes
• A Working Database for Atopic Dermatitis Research
• Constructing a human disease symptom database using disease ontology
• Creating a Small Avian Influenza Virus Ontology
• Distinguishing influenza host specificity and classifying proteins using SVM
• Avian flu analysis: Revealing similarity of sequence structures for H5N1 using HMM
• Exploring genetic evolution of influenza H3N2 using bioinformatics tools
• A database to collect sequence word frequency for HMM training and maintenance
• Building a database for analysis of mutation patterns in Human Genome
6. Findings and observations

We have learned a lot from our practices. As an example, below we report our three findings and observations.

(1) Rather than viewing the broad-sensed concept of database as a sacrifice of accuracy of this term traditionally used in IT context (which usually refers to storage and retrieval of structured data only), we view this “compromise” as a welcome development for integrated treatment of traditional databases and IR, because keyword search and measures of retrieval effectiveness offer good opportunities to examine users’ information needs of different data sources in a uniformed way. This is fully compatible with recent surge of interests of keyword search in combined database/IR/XML platforms ([4]).

(2) As well noted before, there is a significant difference between data mining in general and data mining methods applicable for bioinformatics, namely, the need for mining methods potential adaptable to noisy biodata: Not all the frequent pattern analysis methods can be readily adopted for the analysis of complex biodata, because many frequent pattern analysis methods are trying to discover “perfect” patterns, whereas most biodata patterns contain a substantial amount of noise or faults. For example, a DNA sequential pattern usually allows a nontrival number of insertions, deletions, and mutations. Research in bioinformatics raises new challenges and stimulates new research in machine learning/data mining and text mining. As such, study and education of databases in bioinformatics context contributes to the development of IT field as well.

(3) Another interesting observation is about graduate students in IT. Although the course was designed for such students along with undergrad students in bioinformatics, and each time we have included additional materials to help such students catch up their background, understandably, we still have some concerns of these students during the entire semester. Although such concerns are not groundless and at the end of semester these students may still have not a strong biology background, the good news is that, in the past, almost all of the graduate students passed the course. In fact, their IT background actually helps producing several research papers in bioinformatics (co-authored with faculty members). In addition, a number of graduate students continued on to complete masters’ thesis in bioinformatics.
7. Urgent tasks

As indicated at the beginning of this paper, the main reason of this paper is not only to report what we have done, but more importantly, to receive advices, feedbacks and helps from bioinformatics education community so that we can improve the database education in bioinformatics curriculum. In our view, there are a number of urgent issues need to be addressed in a timely fashion:

(1) A consensus of relatively well-defined scope of this course (and handle overlap with other courses in the curriculum) among bioinformatics educators. Development of one or more recommended course templates (at least course contents) or syllabus will be very helpful. [Preferable a consensus of the overall Bioinformatics curriculum is also needed, but this is not in the scope of this paper.]

(2) Related to issue (1), good textbooks covering both databases and machine learning aspects of bioinformatics are in urgent demand. Of course, there is no lack of good books for introduction for bioinformatics (such as [1,10]) and books providing comprehensive coverage of databases for bioinformatics (e.g., [6]), biological database integration (such as [11]), or biological data modeling (such as [3]), but these books do not offer sufficient coverage on machine learning. On the other hand, there are monographs providing in-depth coverage of various aspects of machine learning for bioinformatics (such as [2,5,7,9]), or edited volumes dedicated to advanced research work in data mining (such as [8,12]), none of them are suitable as textbooks in undergraduate bioinformatics curriculum (even materials have been taken from these books and several useful online sources for our course development).

(3) Development of a rich pool of exercises or assignments, or even a test pool. This pool will become a valuable resource for imposing a kind of academic standard for the database course in bioinformatics.

8. Conclusion

Biology databases (bioDBs) play an ever-increasing role in the post-genome era for bioinformatics education. In this paper, we discussed considerations behind the development of a database course for bioinformatics and described course contents and activities, such as course modules and term projects, as well as some of our observations of conducting this course. We have also identified urgent tasks. We believe that critical comments and suggestions from our colleagues in bioinformatics education community will help us improve the developed course, and in general, a vivid discussion among colleagues will significantly enhance the status of bioinformatics education as a whole.

Acknowledgements

In the past decade, INBRE has continued supported bioinformatics programs at University of Nebraska at Omaha (UNO), including the development of all courses in the bioinformatics undergraduate program. The author thanks all the students who had taken this course at UNO since 2005. Useful comments and advice for this from several faculty members at UNO, particularly Prof. Guoqing Lu of Biology Department and Prof. Dhundiy Kiran Bastola of IST are highly appreciated. We also thank valuable advice from Dr. Mark Pauley for developing a website which provided necessary background of biology for students who were not majored in biology and bioinformatics.

References