Higher Education Studies; Vol. 2, No. 3; 2012
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

Development of a Course Sequence for an Interdisciplinary Curriculum

Muhammad Ali[1]

[1] College of Business & Information Science, Tuskegee University, Tuskegee, Alabama, USA

Correspondence: Muhammad Ali, College of Business & Information Science, Tuskegee University, Tuskegee, Alabama, USA. Tel: 1-334-727-8159. E-mail: alim@mytu.tuskegee.edu

Received: April 14, 2012 Accepted: April 24, 2012 Online Published: July 18, 2012

doi:10.5539/hes.v2n3p1 URL: http://dx.doi.org/10.5539/hes.v2n3p1

Abstract

Interdisciplinary curriculum development is challenging in the sense that materials from more than one discipline have to be integrated in a seamless manner. A faculty member has to develop expertise in multiple disciplines in order to teach an interdisciplinary course, or the course has to be team-taught. Both approaches are difficult to implement. There are administrative issues, such as proportional posting of expenditures across departmental budgets for the courses taught collaboratively, or courses with students from multiple departments. This paper describes the development and teaching of a sequence of bioinformatics related interdisciplinary courses for incorporation into undergraduate biology curricula. Three courses were developed with collaboration between the Departments of Biology and Computer Science at Tuskegee University. Each course contains contents from different subjects, traditionally considered to be virtually independent of each other. The courses have contents from biology, computer science, statistics, mathematics and biochemistry. The first two courses, Introduction to Bioscience Computing and Biological Algorithms & Data Structures, cover the computing and computer science fundamentals necessary for the informed use of bioinformatics tools. The third is an introductory course in bioinformatics. The focus was on teaching the effective use of bioinformatics tools, as compared to development of bioinformatics tools which is more relevant at the graduate level. Administrative issues encountered are also discussed. This work was supported by a NSF HBCU-UP grant.

Keywords: bioinformatics, biocomputing, biological algorithms, interdisciplinary curriculum development

1. Introduction

The past several centuries have been characterized by differentiation of knowledge into different fields and further subfields. Each discipline developed in virtual isolation from others. The experiences of last half of the twentieth century highlighted the need for integrating knowledge from the erstwhile isolated fields of knowledge (Snow, 1990) (Boyer, 1991). A prominent example of integration that has taken place involves the field of biology. This paper discusses interdisciplinary curricula development in the context of modern biology. Over the past forty years biology has changed radically, and has become highly interdisciplinary in nature (Cooper, 2007) (Maloney, 2010). The departments of Computer Science and Biology, Tuskegee University collaborated to develop introductory bioinformatics related courses for the biology curriculum, with an ultimate goal of introducing a computational genomics track in the Department of Biology. Bioinformatics is an interdisciplinary field that focuses on the discovery of biological knowledge using computational techniques. The paper first discusses the overall considerations for the contents of the sequence. This is followed by presentation of rationale for the selection of computing environment and contents for each course. A few earlier studies exist focusing on this issue (Burhans, DeJongh, Doom, & LeBlanc, 2004) (Kumar, Shumba, Ramamurthy, & D'Antonio, 2005) (Cohen, 2003). This paper focuses on the development of undergraduate courses that equip undergraduates with the tools they would use in the industry, while laying a solid foundation for in-depth graduate studies. The paper also focuses on the successful integration of these courses into existing undergraduate curricula that are constrained by credit hour requirements.

2. The Contents of the Course Sequence

Graduate level bioinformatics curriculum has the capacity for detailed coverage of computer science, mathematics, statistics and biology topics, and therefore the early bioinformatics curricula were offered at the graduate level. With the transition from the “Modern Synthesis” phase to the “New Biology” phase in the last third of the twentieth century, molecular biology has become increasingly important (Rose, 2007). The amount of sequence data collected for various organisms is such that it is not possible to study them and compare them manually. In 2008 the number of sequence records had reached almost 100 million. By 2011 the number had increased to over 135 million (NCBI, 2008) (NCBI, 2011). Computational means to analyze the biological data and solve biological problems has become essential. It has therefore become evident that some core bioinformatics knowledge needs to be imparted to undergraduate biology students. Present-day undergraduate biology curricula have very limited mathematical, statistical and computer science contents that are foundational knowledge for bioinformatics. This has called for a major shift in undergraduate biology curriculum for the 21st century.

The goal of the effort reported in this paper was to incorporate the essential bioinformatics related knowledge into an undergraduate biology curriculum in Tuskegee University, a small minority institution with fewer resources as compared to larger universities.

The in-depth treatment of bioinformatics that is possible in a graduate level bioinformatics curriculum is neither possible nor relevant to an undergraduate biology curriculum. This is because there is already shortage of available credit hours, and many of the graduates may not later follow a graduate bioinformatics curriculum. The focus needs to be on the basic uses of bioinformatics tools, rather than the development of these tools.

A survey of graduate level bioinformatics curricula shows considerable variation in content. There are three ends to this spectrum as shown in Figure 1. The central triple intersection area ABC represents bioinformatics relevant knowledge. Curricula positioned towards the end-point C have high mathematics and statistics content (Dept of Mathematics & Statistics, Hunter College, CUNY, 2012). Similarly, curricula positioned towards the end-point B have high computer science content (Department of Computer Science, New Jersey Institute of Technology, 2012), while those positioned towards the end-point A have high biology content (Dept of Biological Sciences, Carnegie Mellon Unversitiy, 2012). Such variations in content take place mainly based on which department is offering a curriculum. Courses developed for a bioinformatics curriculum are influenced by the positioning of the curriculum with reference to Figure 1. For an undergraduate biology curriculum, it is more appropriate to position it towards end-point A.

An undergraduate biology curriculum should provide sufficient foundation in mathematics, statistics and computer science knowledge, while keeping the primary focus on biological applications.

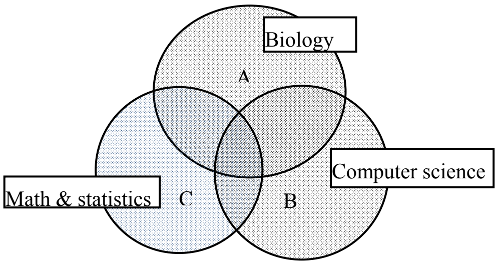

Figure 1. Interdisciplinary nature of bioinformatics

Bioinformatics curricula also vary depending on whether the curriculum is targeted at bioinformatics tools developers, or the users of the bioinformatics tools. Almost all graduate level bioinformatics curricula have a high programming and algorithmic content, and thus are targeted at tool development in addition to developing proficiency in the use of bioinformatics tools. On the other hand most students graduating from a biology undergraduate program are not likely to become bioinformatics tool developers but are likely to be called upon to use bioinformatics tools across many biology/medical related jobs. Careers such as Clinical Geneticist, Medical Geneticist, Genetic Counselor, Genetics Laboratory Research Assistant and Genetics Laboratory Technician would require knowledge and use of bioinformatics tools (Collins, 2011). In the case of undergraduate biology students, the bioinformatics courses should be targeted at the use of bioinformatics tools rather than tool development. Consequently in-depth treatment of mathematics, statistics and computer science contents was avoided. The courses contain enough computer science, mathematics and statistics content that are sufficient for the students to use the bioinformatics tools effectively. For example, a mostly qualitative treatment is presented pertaining to the development of substitution scoring matrices, rather than a rigorous mathematical treatment.

3. Development of the Course Sequence

The inclusion of multi-disciplinary knowledge in undergraduate biology curricula becomes difficult because of the competing need to include several new biology topics. The number of bioinformatics foundational courses needs to be kept to a minimum. A good understanding of basic probability and statistics is essential to the understanding and effective use of bioinformatics tools. An interdisciplinary course BIOL 0202 Mathematics, Computers and Biosciences was developed with collaboration between the Mathematics and Biology department, to address this deficiency. A sequence of three bioinformatics foundational courses was developed with collaboration between Biology and Computer Science departments, to be offered after BIOL 0202. This sequence is designed to communicate enough bioinformatics knowledge to undergraduate biology students to teach them the effective use of bioinformatics tools and to be able to perform rudimentary programming that can help them to handle tasks peculiar to the problems they are solving and not supported by standard bioinformatics tools. These courses were also designed so they could be offered to students who are majoring in computer science and are interested in acquiring bioinformatics knowledge. The three courses are given in Table 1. The first and third courses have additional two hour computing lab components associated with them.

Table 1. Sequence of interdisciplinary courses

Course Number	Title
BIOL/CSCI 0366	Introduction to Biosciences Computing
BIOL/CSCI 0367	Biological Algorithms & Data Structures
BIOL/CSCI 0368	Introduction to Bioinformatics

4. Introduction to Bioscience Computing Course

4.1 Selection of Computing Environment

Although the focus on tool development in undergraduate curricula should be minimal, there is nevertheless the need for some programming skills.

As a consequence of the increase in the computing needs in all fields over the past several decades, the computing tools at the undergraduate level steadily got upgraded steadily through slide rules, simple calculators, scientific programming calculators and on to graphing calculators. Now the computing needs have increased to a level that a typical graphing calculator is proving to be inadequate. As an example, simulation and modeling is becoming increasingly common in almost all fields of study, and cannot be supported by a graphing calculator. Students now need to be familiar with more sophisticated computing tools, especially those students who are studying in the science/engineering fields. In fact, we even need programming skills for effective use of common productivity tools such as spreadsheets and word processors. The more powerful functions of common productivity tools can be utilized only by writing macros. Such productivity tools are used by those working in technical as well as non-technical fields. The use of computing tools and the need for programming skills is now becoming increasingly important even in fields that are traditionally considered to be non-technical.

Graduate level bioinformatics curricula include programming courses in high-level languages such as Java, Biopython and BioPerl. Programming skills in such languages are essential for development of bioinformatics tools. Since only a few of the undergraduate biology students may continue into jobs or graduate programs that require programming skills in such languages, the teaching of programming skills in such languages is not advisable.

The next more powerful option compared to Biopython, C++, etc. is a computing environment such as MATLAB, Mathematica, R, etc. These environments are based on a very high level programming language. Programming in a very high level programming language such as R is easier as compared to programming in languages such as C++, and therefore allows more focus on problem solving.

R is a powerful general purpose computing environment, with a very high level programming language support. It is open source and therefore remains freely available to all students during and after leaving school. An extensive range of bioinformatics modules exist for the R environment and more are being developed in the open source environment, and therefore a wide variety of bioinformatics tasks can be undertaken in the R environment. For biologists, R meets the requirement as a general-purpose computing tool, a biosciences computing tool, and especially as a general-purpose molecular bioinformatics computing tool. Due to these relative advantages as compared to proprietary languages such as MATLAB, it was decided to use R package as the computing environment for this course.

4.2 Selection of Contents of the Course

This is an interdisciplinary course, containing contents from biology, computer science, mathematics and statistics. The contents of the course were determined on the basis of:

a) The essential knowledge required for subsequent courses in the sequence.

b) Statistical skills required beyond what was covered in BIOL 0202

c) Developing the basic programming skills required to deal with unique computing tasks that may arise in the course of working with bioinformatics tools. Tasks such as sorting data, merging data, deleting selected portions from a data set, etc. require some basic programming skills.

It is very common for bioinformatics work to be done on Linux machines, and it was considered important to expose the students to the Linux environment. To achieve this, some basic Linux commands along with text editing were included in the course. Concepts typically used in bioinformatics work, including redirection, piping and shell scripting were also included.

Undergraduate biology students not only need to understand the computing relevant to molecular bioinformatics, they also need to know computing relevant to research activities typical for the biology discipline. This includes presentation of data acquired from lab experiments, by means of graphs. Many biology research activities involve the application of the scientific method, and students need to have a basic understanding of statistical techniques for data analysis.

Markov chain theory is fundamental to many bioinformatics techniques. While it is possible to use the tools without a good understanding of this topic, an informed and focused application of relevant bioinformatics tools will be facilitated by understanding the basics of Markov chain theory and its applications.

The following is a list of topics included in this course:

a) Introduction to Linux environment; basic Linux commands, text editing, redirection; piping; simple shell scripts

b) Introduction to R system; installation of R

c) Programming in R

d) Working with datasets in R

e) Basic data analysis and graphing

f) Probability theory

g) Univariate and multivariate probability distribution

h) Bayesian data analysis

i) Markov chain theory

5. Biological Algorithms & Data Structures Course

5.1 Selection of Computing Environment

It was decided to keep the computing environment the same as for BIOL/CSCI 0366 to avoid expending time and effort in yet another programming language that may not be of much use to the students after the course had ended. It was also decided to keep programming to a minimum, and teach algorithmic techniques using pseudo-code, keeping in mind the goal of producing informed users of bioinformatics tools, rather than developers of these tools. Students are asked to program some algorithms in R to demonstrate the feasibility of using R environment for bioinformatics work.

5.2 Selection of Contents of the Course

This is an interdisciplinary course with contents mainly from computer science. It has contents from statistics and biology. The initial topics in the course follow that of a typical algorithm and data structures course, starting with an introduction to algorithm analysis and complexity. Rigorous proofs are avoided, and focus is given to explaining the behavior of the algorithms. Basic molecular biology topics are covered, that are essential for a proper understanding of biological algorithms. Those algorithms are emphasized that are more relevant to bioinformatics:

a) Exhaustive search algorithms

b) Greedy algorithms

c) Dynamic programming algorithms

d) Divide-and-conquer algorithms

e) Sequence alignment algorithms

These topics are covered with reference to typical problems encountered in bioinformatics, especially sequence alignment.

In-depth study of Needleman-Wunch and Smith-Waterman algorithms are included for sequence analysis. The probability theory needed to understand these algorithms were covered in the preceding course. Simple phylogenetic tree generation techniques are included so that the students understand phylogenetic trees generation taught in the next course in the sequence, for solving biological problems. As an example of bioinformatics specific data structures, Biostrings defined in association with the Bioconductor package are introduced.

6. Introduction Bioinformatics Course

6.1 Selection of Computing Environment

Alignment of a query sequence against large datasets is required frequently. These datasets are increasing exponentially in size, due to the rapid increase in the rate of sequencing. Now, computing requirements for even an introductory bioinformatics class require High Performance Computing. Even powerful workstations are inadequate.

Several different bioinformatics tools are required throughout the course. Keeping the datasets and the tools updated is a difficult task. Additionally, these computing resources will be available only on the campus, due to restrictions on off-campus access to university computing resources. The initial and recurring system administration costs for the necessary resources, is significant. For a department with no significant research in Bioinformatics, the recurring cost of maintaining and administering a HPC resource is difficult to justify for the purpose of supporting an undergraduate Bioinformatics course in a small institution.

A much better option for an undergraduate course at a small institution is to use web-based interfaces to Bioinformatics tools hosted on HPC resources installed at NCBI, EMBL and other organizations. These tools access large datasets. The tools are of high-quality and definitely adequate enough for an undergraduate curriculum. The tools as well as the datasets are frequently updated. The use of these web-based tools has the advantage of allowing students free access to the tools and datasets from anywhere, anytime. The resources remain available after graduation, and therefore the students' links to the subject material remains largely intact.

Web-based tools do not allow for automation of tasks and flexibility of processing possible with in-house HPC resources. Such automation and flexibility is necessary at the graduate level courses. However it is possible to conduct an undergraduate level course using web-based resources only. This is a necessary compromise to offset the high costs of in-house HPC resources.

6.2 Selection of Contents of the Course

This is an interdisciplinary course with contents from biology, computer science, mathematics and statistics. As stated earlier, it is desired that these courses should be designed such that they can be taken by computer science students also. These students do not normally take courses in biochemistry. It was decided to include some biochemistry topics that are directly relevant to nucleotide and amino acid sequences, so that the students develop a reasonable understanding of the tructures that are masked by a string of nucleotide alphabets. The central dogma of Genetics is briefly covered. Covering these topics is found to be beneficial to biology students as a refresher. Before starting on next major topic, a brief introduction to sequence alignment was done to give a feel of the most important activity in the field of molecular bioinformatics.

The next major topic is familiarization with the major NCBI and major European sequence databases, and methods to access them. The main resource used for teaching purposes is the NCBI databases, and the NCBI Entrez search engine for accessing these databases. While a large amount of sequence data are in the public domain, most publically available tools for processing these data are not user-friendly at this point in time, including those at NCBI. Consequently teaching students to use these tools effectively takes several lecture hours. It needs to be said that frequent updating of the features supported by these tools, while desirable, is also a factor that negatively impacts sustainability of consistent user-friendly interfaces for these tools. Hopefully these interfaces will improve in the next few years.

A powerful feature of these tools is the extensive cross-linking of related information and tools. A researcher can reduce his research time by using this feature effectively. It is important to make the students aware of this facility.

Another powerful feature in these search tools is the search query building facility. With the sequence and literature data increasing at a very high rate, the need to get very high quality search results, by significantly reducing false positives and false negatives is becoming crucial. The search query building feature needs considerable improvement as far as user-friendliness is concerned. The students find it difficult to understand and master this feature. Nevertheless, it is an important feature for obtaining good quality search results. Teaching students how to build effective search queries takes several lecture and lab hours.

The next topic covered was the search for appropriate scientific papers in PubMed. A good attempt has been made to standardize the user-interface across several tools available at NCBI, and in that sense the user-interface to PubMed is similar to that for searching sequence databases. However the lack of user-friendliness experienced with sequence search tools is also experienced while searching PubMed.

The lack of user-friendly interfaces may not be a handicap for a frequent user of these tools, but it is problematic when trying to teach a reasonable level of search skills amongst undergraduate students in a short period of time. Significant amount of classroom practice sessions was included in the course to develop these skills, in addition to the lab sessions. However sustainability of search skills remained an issue.

Another problem is lack of good-quality guidelines for using these tools. The documentation lags the frequent tool upgrades, and cause confusion when consulted, especially for students who need unambiguous guidelines to the latest versions of the tools. Detailed lecture slides/notes are required to bridge the gap and guide the students effectively. Lecture slides/notes need to be updated on a semester to semester basis, to keep up with changes in the search tools.

The next major topics are pair wise sequence alignment and the search of NCBI databases for sequences similar to a query sequence, using pair wise sequence alignment as the search technique. At the undergraduate level there is a need to teach students how to do sequence alignment, without in-depth statistical treatment. Statistical treatments are viable in a graduate program. At present most undergraduate biology students do not have a good grounding in the relevant statistics that form the basis for the alignment tools. The courses discussed earlier are meant to provide the requisite statistical background, but for a period of time there will be students in the class without the necessary statistical background. Realistically speaking, it will take several years for the notion to take root that mathematics and statistics have become an integral part of biological sciences. The sub-topics to be covered include alignment algorithms, the development of scoring matrices used in these algorithms and the adjustment of alignment parameters of NCBI BLAST tools. The NCBI sequence alignment tools do not allow the same level of flexibility for choosing search options as available in the command line versions, but it is more than adequate for an undergraduate course.

The problems with documentation are similar to those for the sequence search tools.

Multiple sequence alignment are also taught using web-based tools available at NCBI and EBI. Simple phylogenetic analysis is practiced using PHYLIP package for genetic phylogenetic trees and NCBI taxonomy database tools for species trees.

7. Administrative Issues

Interdisciplinary courses crossing several science disciplines are relatively new. Interdisciplinary teaching requires institutional level support and changes to departmental structure (Holley, 2009). Budgetary accounting for interdisciplinary faculty teaching load is generally not well-addressed (MacKinnon, Rifkin, Hine, & Barnard, 2010). While teaching the discussed sequence of courses, it transpired that the majority of students were from biology, while the instructor was from the computer science department. This lead to a discussion as to whether the course teaching expenses should be shown against the computer science or biology department budget. A procedure is yet to be determined for the allocation of expenditures proportional to the number of students, and reflecting them in the participating departments' budgets. One course in Introduction to Bioinformatics was team taught with faculty from biology and computer science departments. A procedure is yet to be evolved to show proportional expenditure of time and effort for the concerned faculty members, against departmental budgets.

Another issue is to estimate the workload taken on by a faculty member who is team teaching an interdisciplinary course with faculty members of other departments. These problems need to be resolved at the institutional level, and cannot be done at the departmental level. Such issues need to be streamlined for success of interdisciplinary teaching and research.

8. Conclusion

Interdisciplinary courses are becoming increasingly necessary in current undergraduate curricula. Modern biology is an excellent example of a field that has become irreversibly interdisciplinary. An accomplished biologist has to develop reasonable level of expertise in several other fields. Inclusion of bioinformatics foundational knowledge in an undergraduate biology curriculum with computational genomics track is a challenging task. In the course sequence discussed in this paper, important bioinformatics concepts are taught, with the introduction of only the essential mathematics, statistics and statistics material necessary to provide clear concepts. The focus of the course sequence is on the use of bioinformatics tools, rather than the development of bioinformatics tools.

References

Boyer, E. L. (1991). The scholarship of teaching from scholarship reconsidered: Priorities of the professoriate. *College Teaching*, *39*(1), 11-14.

Burhans, D. T., DeJongh, M., Doom, T. E., & LeBlanc, M. (2004). Bioinformatics in the undergraduate curriculum: Opportunities for computer science educators. *Proceedings of the 35th SIGCSE technical symposium on Computer science education (SIGCSE' 04)* (pp. 229-230). New York: ACM.

Cohen, J. (2003). Guidelines for Establishing Undergraduate Bioinformatics Courses. *Journal of Science Education and Technology*, *12*(4), 449. http://dx.doi.org/10.1023/B:JOST.0000006304.01183.ba

Collins, D. (2011). *Human Genetic Careers*. Retrieved November 6, 2011 from The University of Kansas Medical Center: http://www.kumc.edu/gec/prof/career.html

Cooper, G. (2007). Design and Implementation of an Undergraduate Bioinformatics Curriculum in an Online Environment. *37th annual frontiers in education conference—global engineering: Knowledge without borders, opportunities without passports*, (pp. F2D-7 to F2D-11).

Department of Computer Science, New Jersey Institute of Technology. (2012). *NJIT: Computer Science: MS in Bioinformatics (MS BioInf)*. Retrieved February 25, 2012 from New Jersey Institue of Technology web site: http://cs.njit.edu/academics/graduate/ms-bioinf/index.php

Dept of Biological Sciences, Carnegie Mellon Unversitiy. (2012). *Computational Biology & Bioinformatics-Dept of Biological Sciences-Carnegie Mellon University*. Retrieved February 25, 2012 from Carnegie Mellon University web site: http://www.cmu.edu/bio/research/compbio.html

Dept of Mathematics & Statistics, Hunter College, CUNY. (2012). *Mathematics and Statistics Programs*. Retrieved February 25, 2012 from City University of New York web site: http://math.hunter.cuny.edu/graduate.shtml

Holley, K. A. (2009). Understanding Interdisciplinary Challenges and Opportunities in Higher Education. *ASHE Higher Education Report*, *35*(2), 131.

Kumar, A. N., Shumba, R. K., Ramamurthy, B., & D'Antonio, L. (2005). Emerging areas in computer science education. *Proceedings of the 36th SIGSE technical symposium on Computer science education (SIGSE' 05)*. St. Louis, Missouri USA: ACM, New York, NY, USA.

MacKinnon, P., Rifkin, W. D., Hine, D., & Barnard, R. (2010). Complexity and Mastery in Shaping Interdisciplinarity. In M. Davies, & M. Devlin (Eds.), *Interdisciplinary Higher Education: Perspectives and Practicalities* (1st ed., Vol. 5, pp. 29-54). Bingley, UK: Emerald Group Publishing Ltd. http://dx.doi.org/10.1108/S1479-3628(2010)0000005005

Maloney, M. P. (2010). Bioinformatics and the Undergraduate Curriculum. *CBE- Life Sciences Education*, *9*, 172-174. http://dx.doi.org/10.1187/cbe.10-03-0038

NCBI. (2011, November 7). *GenBank Home*. Retrieved April 2, 2012 from www.ncbi.nlm.nih.gov: http://www.ncbi.nlm.nih.gov/genbank/

NCBI. (2008). *GenBank Statistics*. Retrieved April 2, 2012 from www.ncbi.nlm.nih.gov: http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/

Rose, M. R. (2007). The new biology: Beyond the Modern Synthesis. *Biology-Direct*, *2*(1), 30. http://dx.doi.org/10.1186/1745-6150-2-30

Snow, C. P. (1990). Two Cultures. *Leonardo*, *23*(2/3), 169-173. http://dx.doi.org/10.2307/1578601