

A Proposed Undergraduate Bioinformatics Curriculum for Computer Scientists *

Travis Doom¹, Michael Raymer¹, Dan Krane², and Oscar Garcia¹
Departments of ¹Computer Science and Engineering and ²Biological Sciences
Wright State University, Dayton, OH 45435-0001
{travis.doom, michael.raymer, dan.krane, oscar.garcia}@wright.edu

Abstract

Bioinformatics is a new and rapidly evolving discipline that has emerged from the fields of experimental molecular biology and biochemistry, and from the the artificial intelligence, database, and algorithms disciplines of computer science. Largely because of the inherently interdisciplinary nature of bioinformatics research, academia has been slow to respond to strong industry and government demands for trained scientists to develop and apply novel bioinformatics techniques to the rapidly-growing, freely-available repositories of genetic and proteomic data. While some institutions are responding to this demand by establishing graduate programs in bioinformatics, the entrance barriers for these programs are high, largely due to the significant amount of prerequisite knowledge in the disparate fields of biochemistry and computer science required to author sophisticated new approaches to the analysis of bioinformatics data. We present a proposal for an undergraduate-level bioinformatics curriculum in computer science that lowers these barriers.

1 Introduction

Bioinformatics is a new discipline that deals with the research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data [1]. Much bioinformatic research relates to the discovery of the functional relationships between the composition of the *genes* within the context of the *genome* and the structure

and function of the proteins encoded by these genes. Because the interaction of the proteins within an organism determines metabolism, reproduction, form, and health, the implications of bioinformatics studies are far reaching. Recent advances in the experimental techniques of molecular biology have resulted in an explosive growth in the availability of molecular data. As a result, current bioinformatics research is generally focused on the representation, analysis, annotation and mining of large databases of protein and genome sequence information. In the near future, the focus will shift to a functional analysis of the proteins produced by these genes. Bioinformatics techniques promise to provide information that brings enormous power in areas ranging from disease diagnosis and treatment to evolution, agriculture, and environmental science.

There is a high demand for professionals with a background in bioinformatics. The annotation and analysis of the human genome is one of the most complex computational problems currently being studied on a world-wide scale. Computer scientists are needed to analyze, index, represent, model, display, process, mine, and search large biological databases. This need is already extensive and will continue to grow. The genomic database maintained at the National Center for Biotechnology Information (NCBI) currently doubles every 14 months. Industry analysts forecast that the market for genomic information alone (and the technology to use it) will reach an annual US \$2 billion by 2005 [2]. In the January 2001 issue of *The Scientist*, it is reported that the National Institute of General Medical Sciences (NIGMS) is already having difficulty finding people from other disciplines to perform the kind of modeling and data analysis that researchers in the biological sciences now require.

The educational opportunities available to undergraduate students wishing to participate in this exciting enterprise are currently limited [3]. *The development of an undergraduate curriculum in bioinformatics is essential to meeting the future needs of the nation.* The development of a bioin-

*This research was supported in part by a Wright State University research initiation grant and in part by the National Science Foundation under CISE grant #EIA-0122582.

formatics curriculum must be initiated immediately so that students can be a part of the basic research of this emerging field and immediately available to meet the workforce needs of the nation.

2 Graduate program barriers

Graduate programs in bioinformatics are beginning to emerge at several universities, including Wright State University. Entrance requirements for such programs, however, require students with a specific prerequisite program of undergraduate study that is rarely made available to students as part of an organized program.

Graduate bioinformatics programs must currently accept students with undergraduate degrees in either computer science or biology and have sequences of remedial or prerequisite courses designed to complement the knowledge already acquired by the students as undergraduates. Students holding an undergraduate degree in computer science generally need to spend the majority of their first year of graduate study taking focused remedial courses in basic biochemistry, molecular biology, and genetics. Students holding an undergraduate degree in biology generally spend the majority of their first year of graduate study in coursework covering introductory computer science programming, basic data structures, databases, and artificial intelligence.

The second year of a graduate bioinformatics program is generally dominated by pre-existing graduate courses in computer science and biology. From computer science, courses in artificial intelligence, database, pattern recognition, and genetic algorithms are fundamental. From biology, a course sequence providing specialization in genetics, molecular biology, physiology, or ecology is considered highly advantageous. Finally, students from either background would require a course sequence on contemporary algorithms and research techniques in bioinformatics. It is unlikely that this amount of material can be accommodated in a two-year course of study without significant preparation at the undergraduate level.

3 An undergraduate program

Due to the demanding entrance requirements, graduate programs alone may prove inadequate in providing the number of bioinformatics specialists that industry will require, partly because of the amount of the remedial coursework necessary. New undergraduate programs must be developed that incorporate a more specific (and shorter) biology sequence with a more focused computer science foundation. It may be necessary to redesignate some of the traditional core courses in CS, such as digital system design, as electives to allow for an increased base of knowledge in

the contemporary areas of IT knowledge (such as artificial intelligence, knowledge representation, and data-mining).

It falls to four-year programs to provide opportunities and direction to students to meet the market demand for bioinformatics professionals and to better prepare students for entrance into graduate-level bioinformatics programs. Implementing an academic program of study for bioinformatics is, unfortunately, complicated by its inherently inter-disciplinary nature. Programs accredited by the Computer Science Accreditation Board (CSAB or, more recently, CAC) are required to include at least a two year (24 quarter hour) sequence of fundamental "core" computer science material as well as at least one year of math and one year of a laboratory science (typically physics) [4]. Biology programs typically require at least one year of study in basic chemistry. These sophomore-level courses are usually only taken after a year of study in inorganic chemistry. While an appreciation of basic chemistry concepts such as valency and electro-negativity are useful in the study of bioinformatics, we believe that an accelerated training in chemistry is sufficient and would be more accommodating to the demands of an integrated computer sciences and biology curriculum. At the same time, a streamlined exposure to introductory programming, calculus, and biology (in addition to general education) in the first two years of study is also appropriate.

As bioinformaticians must be equally versed in the languages of biology and computer science, this effort will require a fundamentally interdisciplinary approach. Furthermore, basic research in the field of bioinformatics is progressing rapidly. Professionals in fields such as bioinformatics must possess not only a strong grasp of computer science fundamentals, but must also be equally comfortable in the fundamentals of biology and biochemistry to recognize and appreciate the results of their analyses.

3.1 Integration of computer science core material

Classically, computer science has focused on the study of computer hardware and software. A more contemporary view of information technology, however, must recognize that storage, transmission, and distribution of data make up a significant portion of the future demand on the discipline and on future computer professionals. This mandates a program of study emphasizing contemporary topics in databases and networking.

From the discipline of computer science, a bioinformatics professional should have knowledge of: introductory programming, data structures, AI algorithms (search, optimization, list processing, pattern recognition, etc.), databases, formal and comparative languages (complexity, and specialized algorithm topics such as those explained in [5]), modeling, and simulation, probability and statistics,

the WWW, visualization, and human-computer interaction (HCI) issues.

3.2 Integration of biology core material

From the discipline of biology, a bioinformatics professional should have working knowledge of at least one of several life sciences fields, including genetics, environmental biology, et al. Of these many possibilities, we propose to focus on the area of *molecular bioinformatics*. A professional in this field of study should understand genetics, molecular and cellular biology, chemical and physical aspects of flow of genetic information from DNA to proteins, gene expression, replication, recombination, repair, and the experimental tools of molecular biology.

The amount of practical laboratory experience that should be possessed by an undergraduate bioinformatician is a point of debate. The results of DNA sequencing technology (and other *in vitro* and *in vivo* laboratory technologies) are published, annotated, and made available for analysis world-wide. The real problem is in extracting meaning from the glut of available data. Computationally generated results (*in silico* technologies) are becoming more prevalent in the field.

4 A bioinformatics curriculum

We now present a curriculum proposal which is in accordance with CSAB (now CAC) standards [4], yet incorporates specific sequences in chemistry and biology with a more focused computer science foundation. In order to meet our objectives, it was necessary to remove several traditional, but non-essential, topics from the computer science curriculum for this option. Knowledge of calculus-based physics, for instance, is not as important for students preparing for careers in bioinformatics as it is for those interested in digital signal processing. Furthermore, many of the traditional focuses of computer science that are not required CSAB/CAC standards have been made optional to allow for an increased base of knowledge in the contemporary areas of IT knowledge.

To facilitate the implementation of this program, we have introduced only two new courses. *Introduction to Bioinformatics* is a course which will be co-taught by faculty from both the Department of Computer Science and the Department of Biology. This course has a tools-oriented to bioinformatics with an emphasis on data structure in DNA, representation and manipulation of strings in PERL, data searches and pairwise alignments, protein structure prediction and modelling, proteomics, and the use of web-based bioinformatic tools. This first course in bioinformatics is designed not only for students in the bioinformatics program, but as an elective for all biology or computer science

students who wish some exposure to the field. *Algorithms for Bioinformatics* is a capstone course for students in the program which presents a theory-oriented approach to the application of contemporary algorithms to bioinformatics. This course includes graph theory, complexity theory, dynamic programming, formal language theory, and optimization techniques in the context of their application toward solving sequence comparison, fragment assembly, molecular structure prediction, and other computational problems in biology.

Computer Science - Bachelor of Science
Proposed option in Bioinformatics
Wright State University
Total Quarter Credit Hours: 195

I General Education Courses (42 hours)

Area A: Communication (8 hours)

- ENG 101-4 Composition I
- ENG 102-4 Composition II

Area B: Humanities (34 hours)

- Eleven general elective courses

II Departmental Requirements (87 hours)

Area A. Required Computer Science and Engineering Courses (47 hours)

- CS 240-4 Computer Science I
- CS 241-4 Computer Science II
- CS 242-4 Computer Science III
- CEG 255-5 Intro. to Comp. Information Sys.
- CEG 260-4 Digital Computer Hardware
- CEG 320-4 Computer Organization
- CS 400-4 Data Structures and Software Design
- CS 405-4 Intro to Database Management Systems
- CS 409-4 Principles of Artificial Intelligence
- CS 415-3 Social Implications of Computing
- CEG 433-4 Operating Systems
- CS 480-4 Comparative Languages

Area B. Required Biology Courses (29 hours)

- BIO 112-4 Principles of Biology: Cell Biology and Genetics
- BIO 114-4 Organismic Biology
- BIO 115-4 Principles of Biology: Diversity and Ecology
- BIO 210-4 Molecular Biology I
- BIO 211-4 Molecular Genetics I
- BIO 212-4 Cell Biology
- BIO 410-4 Cell-Molecular Biology Laboratory

- BIO 492-1 Senior Seminar

Area C. Required Bioinformatics Courses (8 hours)

- BIO/CS 399-4 Intro to Bioinformatics
- BIO/CS 471-4 Algorithms for Bioinformatics

Area D. Technical Communications (3 hours)

Choose from:

- EGR 335-3 Technical Communications
- BIO 310-3 Issues in Science

III Required Supporting Courses (58 hours)

Area A: Chemistry (33 hours)

- CHM 121-5 Submicroscopic Chemistry
- CHM 122-5 Macroscopic Chemistry
- CHM 123-5 Reaction Dynamics
- CHM 211/215-6 Organic Chemistry I
- CHM 212/216-6 Organic Chemistry II
- CHM 213/217-6 Organic Chemistry III

Area B: Mathematics (25 hours)

- MTH 229-5 Calculus I
- MTH 230-5 Calculus II
- MTH 231-5 Calculus III
- MTH 253-3 Elementary Matrix Algebra
- MTH 257-3 Discrete Mathematics for Computing
- HFE 301-4 Statistics I

IV CS/Bio/MTH Electives (8 hours of 400-level CS/CEG)

Choose from:

- CEG 416-4 Matrix Computations
- CEG 434-4 Concurrent Software Design
- CEG 465-4 Interactive Systems Modeling, Analysis, and Design
- CEG 466-4 Formal Languages
- CEG 476-4 Computer Graphics I
- CEG 477-4 Computer Graphics II
- CS 407-3 Optimization Techniques
- CS 458-3 Applied Graph Theory
- CS 459-3 Combinatorial Tools for Computer Science
- CS 470-4 Systems Simulation

5 Conclusion

Computer science is a path to understanding genomes just as biology helps us in understanding living organisms. It is hard to imagine a more significant area where we must hone our methods of questioning than bioinformatics. The competitive pressure and rewards for progress in bioinformatics are high, and students can use them to prepare themselves to join this sought-after work-force. The creation of an undergraduate bioinformatics option in computer science and engineering is of utmost importance for global health, the economic development of those nations undertaking this path, and the success of our students.

The central argument that we present for an undergraduate bioinformatics option within a Computer Science BS degree can be summarized as follows: (1) The number and chain of prerequisites that must be satisfied in either case requires about two years of course-work because course dependencies are such that they cannot be taken in parallel. (2) This being the case, an assumption of two years of prerequisites, in addition to the two years to obtain the MS degree, implies that it could take eight years of preparation for a student to obtain an MS degree in bioinformatics. (3) The alternative that we propose would lead to a BS degree in four years and an MS degree in the standard six year time frame.

Our proposed curriculum includes, in addition to traditional computer science, biochemistry, and molecular biology components, several courses tailored specifically to meet the needs of an integrated interdisciplinary program. One such course is an undergraduate introduction to bioinformatics algorithms and methods. As this course will serve as a unifying element for the rest of the bioinformatics program, Drs. Krane and Raymer have formalized the proposed course content and are currently preparing and undergraduate bioinformatics textbook to be published by Benjamin-Cummings in December, 2002.

References

- [1] BISTIC Definition Committee, "NIH working definition of bioinformatics and computational biology." <http://grants.nih.gov/grants/bistic/CompuBioDef.pdf>, July 2000.
- [2] S. K. Moore, "Understanding the human genome," *IEEE Spectrum*, pp. 33 – 35, November 2000.
- [3] T. E. Doom and O. N. Garcia, "Bioinformatics: An option in computer science," in *2001 Midwest Artificial Intelligence and Cognitive Science Conference*, March 2001.
- [4] Computing Sciences Accreditation Commission, "Criteria for accrediting programs in computer science in the united states." <http://www.csab.org>, 2000.
- [5] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT Press, 1998.