BIOINFORMATICS: A NEW FIELD IN ENGINEERING EDUCATION

Richard Hughey and Kevin Karplus

Abstract— *Bioinformatics is a new engineering field poorly served by traditional engineering curricula. Bioinformatics can be defined in several ways, but the emphasis is always on the use of computer and statistical methods to understand biological data, such as the voluminous data produced by high-throughput biological experimentation including gene sequencing and gene chips. As the demand has outpaced the supply of bioinformaticians, the UCSC School of Engineering is establishing undergraduate and graduate degrees in bioinformatics. Although many schools have or are proposing graduate programs in bioinformatics, few are creating undergraduate programs. In this paper, we explore the blend of mathematics, engineering, science, and bioinformatics topics and courses needed for an undergraduate degree in this new field.*

Index Terms—*Bioinformatics, computational biology, undergraduate degree program*

Introduction

Bioinformatics is a new engineering field poorly served by traditional engineering curricula. Bioinformatics can be defined in several ways, but the emphasis is always on the use of computer and statistical methods to understand biological data, such as the voluminous data produced by high-throughput biological experimentation including gene sequencing and gene chips.

The field's rapid growth is spurred by the vast potential for new understanding that can lead to new treatments, new drugs, new crops, and the general expansion of knowledge. The public and private efforts to determine and annotate the 3.1 billion nucleotides of the human genome is only the most visible creation of computerized data.

Bioinformatics is broadly defined as the application of computer techniques to biological data. This field has arisen in parallel with the development of automated high-throughput methods of biological and biochemical discovery that yield a variety of forms of experimental data, such as DNA sequences, gene expression patterns, and chemical structures. Bioinformatics encompasses everything from data storage and retrieval to the identification and presentation of features within data, such as finding genes within DNA sequence, finding similarities between sequences, structural predictions and correlation between sequence variation and clinical data.

As the demand has outpaced the supply of bioinformaticians, the UCSC School of Engineering is establishing undergraduate and graduate degrees in bioinformatics. Although many schools have or are proposing graduate programs in bioinformatics, few are creating undergraduate programs. In forming our graduate program, we realized a fundamental problem: where would we find qualified graduate students? This quickly lead to the development of an undergraduate curriculum that we believe will serve students heading both to graduate school and to industry.

We are, of course, joining a wave of activity in the development of academic programs in bioinformatics [1], [2]. Although most programs are at the graduate level, there are a few undergraduate programs as well. More details can be found at the web site of the International Society for Computational Biology (ISCB), the area's 4-year-old professional society (`http://www.iscb.org`) [3].

In the following sections, we discuss the history of bioinformatics at UCSC and the new undergraduate bioinformatics curriculum. Included are both background courses and bioinformatics course, including descriptions of two new courses, Introduction to Bioinformatics and Bioethics.

Bioinformatics at UCSC

UC Santa Cruz is well known in the field of bioinformatics for pioneering work on applications of hidden Markov models (HMMs) and stochastic context-free grammars to biological sequence data and for the development of software tools for sequence analysis [4], [5], [6], [7]. Our efforts include participation in the international Human Genome Project [8], development of the gene-finding software chosen to annotate the human and Drosophila (fly) genomes [9], success in international protein structure prediction contests [10], [11], and creation of a SIMD programmable accelerator for sequence-analysis [12]. Since the introduction of our first bioinformatics course in 1996, we have trained many students with majors in Computer Science, Computer Engineering, Mathematics, Molecular, Cell and Developmental (MCD) Biology, Chemistry, and Biochemistry. These students are in very high demand in industry and academia. Based on our experience, we are now attempting to create distinctive BS, MS, and PhD programs of study in bioinformatics. These curricula emphasize the foundations of basic molecular biology and biochemistry coupled with substantial training in mathematics and computer engineering and science.

Department of Computer Engineering, Jack Baskin School of Engineering, University of California, Santa Cruz, CA 95064, {rph,karplus}@soc.ucsc.edu

Foundation Material

Bioinformatics, as with the classical engineering disciplines, is the application of mathematics (e.g., probability and statistics), science (e.g., biochemistry), and a core set of problem solving methods (e.g., computer programming and database design) to find solutions to critical problems. Bioinformatics systems include multi-layered software, hardware, and experimental solutions that bring together a variety of tools and methods to analyze immense quantities of data.

In describing the curriculum, it is important to know that UCSC is on a quarter system with 5-credit courses (3.5 hours of lecture per week for 10 weeks). A total of 180 credits, including major and general education requirements, are needed to graduate. The B.S. in Bioinformatics major requirements include 129 credits in 23 courses, slightly less than our computer engineering curriculum.

Core Math

The mathematics requirement is similar to that of most majors within the UCSC Baskin School of Engineering (presently Computer Engineering, Computer Science, Electrical Engineering, and Information Systems Management). Students complete 4 quarters of calculus, including a 2-quarter multivariable sequence, a combined linear algebra and differential equations course, discrete mathematics, and statistics. Students receive additional statistics training in the bioinformatics courses and, possibly, in electives.

Core Science

The core science requirement is expected to be the most difficult hurdle in the degree program for computer-oriented students. The tools that the bioinformatician use are primarily drawn from mathematics and engineering. However, a thorough understanding of the underlying biochemistry is critical to employing the tools effectively and interacting with laboratory scientists. Therefore, the science requirement includes two quarters of college chemistry, one year (or an accelerated two quarters) of organic chemistry, one quarter of cell biology, and, most important, the first of the 3-course sequence in biochemistry.

This first biochemistry course studies the basics of DNA, RNA, and proteins. Unlike the chemistry courses leading up to it, there is no associated laboratory. Students may complete additional courses in the biochemistry sequence (discussing enzymes, regulation, metabolism, biochemistry laboratory, and other topics) as electives in the bachelor's degree program.

Core Engineering

The engineering requirements focus on developing programming and algorithm skills. They include two quarters of introductory programming, a programming and proof course in abstract data types, database systems, technical writing, and an advanced programming elective. Because of the broad scope and extensive requirements of this major, we did not include the computer organization course required of other engineering majors.

The database systems class is presently offered by the computer science department for computer science and information systems management majors. It is likely that soon after the Bioinformatics program has started, we will construct a bioinformatics-oriented database class. This potential new class will focus on the structure, contents, and use of biological sequence databases.

The technical writing course, long a staple in our computer engineering major, serves the dual purpose of formal discipline-specific writing instruction and fulfillment of the campus writing requirements.

The goal of the advanced programming elective is to provide additional programming experience. Any programming-intensive course will be accepted; the recommended electives include object-oriented programming, compilers, and applied graph theory. The compiler class may be the most appropriate because of its dual emphasis on programming and formal languages, a core area in bioinformatics.

Bioinformatics Study

The program includes three bioinformatics courses—Introduction to Bioinformatics, Bioethics, and the graduate Bioinformatics I course—and two advanced electives. The exceptionally strong math, science, and engineering training of bioinformatics majors leaves little doubt that they will be able to excel in the graduate bioinformatics course. Initially, in fact, we expect the undergraduates to frequently outperform the graduate students, who may not have as broad understanding of the three foundation areas.

Introduction to Bioinformatics

The Introduction to Bioinformatics course is the only course being introduced specifically for this major; the other two bioinformatics courses were developed for our graduate program.

Introduction to Bioinformatics (Table I) has prerequisites of probability and statistics, the second quarter of introductory programming, and the first quarter of organic chemistry. While one could make a strong case for requiring the abstract data types and the biochemistry courses, including them in the prerequisite chain would prevent most undergraduates from taking this course until their senior year (or later). Instead, we are hoping that many students will take this class in their junior year, making them ideal candidates for summer research assistantships.

We are also including a laboratory section with this course. One of the fundamental tasks of the bioinformatician is connecting of the different tools into a data analysis pipeline. Thus, the scripting language Perl is perhaps the most used programming language in bioinformatics. Approximately half of

TABLE I
CORE AND OPTIONAL MATERIAL FOR INTRODUCTION TO BIOINFORMATICS, IN 70-MINUTE LECTURES

	Core topics
1	The fundamental dogma of biology (DNA→RNA→protein), DNA bases, RNA bases, Amino acids.
2	DNA sequence databases on the web: genomic, unfinished genomic, EST, clustered EST. Specific databases include Genbank, NCBI's NR, dbEST, Unigene, STACK, EGAD, and others. Error rates in different DNA databases due to sequencing error.
2	Protein sequence and structure databases on the web: Swiss-Prot, TrEMBL, NCBI's NR, PDB, FSSP, SCOP, CATH. Visualization tools: RASMOL or CHIME.
3	Pairwise alignment: substitution matrices and dynamic programming, Needleman-Wunsch and Smith-Waterman algorithms.
3	Probabilistic models: Bayesian statistics, definition of E-value, null models.
2	Fast search: BLAST algorithm and interpreting BLAST results. BLAST options. Visualization tool: NCBI's pictorial BLAST results.
4	Multiple sequence alignment: clustal, HMMs, and regularizers. Information and relative entropy. Visualization tools: belvu (or equivalent) and sequence logos.
	Core topics with potential for expansion
1	Protein structure prediction: HMMs, threading, secondary structure prediction, ab initio prediction.
1	DNA chips: how they work, design of oligos, extracting data from images.
1	mRNA expression analysis: clustering of profiles.
1	Genefinding: HMM structure, signals and content sensors, using EST and homology data, using cross-genome similarity.
	Example optional topics
4	RNA structure prediction and alignment: energy minimization (Zucker), correlated mutation, stochastic context-free grammars.
6	Phylogeny: parsimony, max likelihood, evolutionary models.
2	DNA assembly.
2	X-ray crystallography—how models are created.
3	NMR spectroscopy—how models are created.
2	2D gel proteomics.
2	Mass spec proteomics.
3	Finding binding sites in DNA regulatory regions.

References for these topics may be found on the web or in bioinformatics texts

the laboratory time will be spent using Perl, and the other half using WWW and command-line tools.

There are presently no books that match our course design. Many appropriate sections are in Durbin, Eddy, Krogh, and Mitchison's *Biological Sequence Analysis* [13], used in the graduate Bioinformatics I class. This book is quite advanced mathematically and includes too much detail for an advanced undergraduate class. We expect to develop most of the course material ourselves, possibly leading to a new undergraduate text on bioinformatics. Interested readers can find information on many bioinformatics texts at the ISCB web site mentioned above.

Bioethics

This is a joint undergraduate and graduate course, with additional work at the graduate level. The course has weekly topics with visiting distinguished lecturers from our campus, other UC campuses, other universities, national labs, and local (Santa Cruz and Santa Clara county) industry. Topics include general scientific ethics, biological research ethics, the ethical implications of the Human Genome Project, and governmental and research policy issues. Course work is primarily reading and writing.

TABLE II
WEEKLY TOPICS IN THE GRADUATE COURSE BIOINFORMATICS I

week 1	Introduction to molecular biology and biosequence databases.
week 2	Probability/statistical significance, Bayesian statistical methods.
week 3	Pairwise alignment, dynamic programming, multiple alignment.
week 4	Pairwise alignment, dynamic programming, multiple alignment.
week 5	Profiles, hidden Markov models.
week 6	Forward-backward algorithm and expectation maximization
week 7	Dirichlet mixtures and sequence weighting
week 8	Neural nets, nearest-neighbors, secondary-structure prediction.
week 9	Profile-based protein structure prediction and protein threading
week 10	Fragment packing

Bioinformatics I

Undergraduate Bioinformatics majors have the choice of completing our graduate Bioinformatics I course or a senior thesis. Bioinformatics I has a focus on core statistical methods, including the mathematics behind hidden Markov models and Dirichlet Mixture distributions, that form the foundation of the UCSC approach to bioinformatics. Fewer topics are covered in far more detail than Introduction to Bioinformatics (Table II). The primary evaluation criteria for Bioinformatics I is a quarter-long project in bioinformatics.

Several undergraduates have already completed this course (without the benefit of Introduction to Bioinformatics), many also taking part in Research Experiences for Undergraduates, funded by the National Science Foundation.

Additional information on this course can be found at links from the UCSC Computational Biology web page (`http://www.soe.ucsc.edu/research/compbio/`)

Science, Math, and Engineering Electives

The degree program includes two relatively free electives. These may be filled with any pair of courses that make sense, such as additional biochemistry courses, graduate bioinformatics courses, probability and statistics courses, biology and chemistry courses, and other electives in computer engineering and science.

Graduate Bioinformatics Study

With the assistance of the Alfred P. Sloan Foundation, we have also created a curriculum for the professional masters degree in bioinformatics (http://www.sciencemasters.org) [14]. There are six required graduate courses: Advanced Molecular Biology, Protein Structure, Stochastic Models, Bioinformatics I, Bioinformatics II, and Bioethics. Students must also take three graduate electives and three 2-unit seminar courses. The seminars typically alternate between invited guest speakers and student research presentations.

We expect the graduate program to be approved for the 2002 academic year, one year after the undergraduate program, because of the more extensive review process. Once both programs are approved, we shall propose a combined BS/MS option in bioinformatics that will allow undergraduate bioinformatics students to apply Bioinformatics I and Bioethics toward both the graduate and undergraduate degrees.

Conclusions

The undergraduate bioinformatics major is a rigorous major combining courses from many departments (Table III; the second and third quarters of biochemistry are electives). We expect the program to be far more effective than double majors or major and minor combinations at integrating the fundamental knowledge and techniques required by the bioinformatician.

As we gain experience with the program, we expect to refine several of the requirements. As mentioned above, a bioinformatics-centered database course is a particularly attractive possibility. We may also replace the fourth quarter of calculus with a course in Bayesian statistical methods. One of our colleagues in Molecular, Cellular, and Developmental Biology is developing a course in biotechnology for the bioinformatician: a course that would expose students to the full chain of laboratory work required to, for example, sequence DNA. We will continue working with our colleagues in all associated departments to introduce and revise courses for the discipline.

The development of multi-disciplinary majors can be a difficult and time-consuming task. Coordination and discussion is required between many departments and faculty committees. However, as bioinformatics and other multi-disciplinary programs are developed, this path will become easier here and elsewhere. We hope that others will consider introducing bioinformatics degrees at all levels to better train our scientists for the new biology.

Acknowledgments

The authors thank David Haussler, Director of the UCSC Center for Biomolecular Science and Engineering, for starting bioinformatics at UCSC, Ann Pace, Assistant Direc-

TABLE III
PLAN OF STUDY WITH COURSES IN APPLIED MATH AND STATISTICS (AMS), BIOCHEMISTRY (BCH), BIOLOGY (BIO), BIOMOLECULAR ENGINEERING (BME), CHEMISTRY (CH), COMPUTER ENGINEERING (CE), COMPUTER SCIENCE (CS), AND GENERAL EDUCATION.

Year	Fall		Winter		Spring	
1		General education		General education		General education
	CS	Programming 1	CS	Programming 2	CE	Discrete Math
	M	Calculus 1	M	Calculus 2	M	Calculus 3
2	CH	Chemistry 1/L	CH	Chemistry 2/L	BIO	Cell Biology
		General education		General education	CS	Abstract Data Types
	M	Calculus 4	AMS	ODE and Linear Algebra		General education
3	CH	Organic Chem 1/L	CH	Organic Chem 2/L		General education
	CE	Technical writing		General education	CE	Stochastic Methods
	CS	Databases	CS	Compilers		General education
4	BCH	Biochemistry 1	BCH	Biochemistry 2	BCH	Biochemistry 3
	BME	Intro to Bioinformatics	BME	Bioinformatics I		General education
		General education	CH	Bioethics		

tor of the Center for contributing some of the introductory text, David Deamer, Professor of Chemistry, for developing the bioethics course, and Eric Rice for valuable comments. Our work in bioinformatics is supported in part by NSF grants EIA9905322 and DBI9808007, DOE grant DE-FG03-99ER62849, and a grant from the Alfred P. Sloan Foundation.

References

[1] Russ B. Altman, "A curriculum for bioinformatics: The time is ripe," *Bioinformatics*, vol. 14, no. 7, pp. 549–550, 1998.

[2] Andy Brass, "Bioinformatics education — a UK perspective," *Bioinformatics*, vol. 16, no. 2, pp. 77–78, 2000.

[3] Chris Rawlings and Larry Hunter, "Creating a professional society for bioinformatics — the International Society for Computational Biology (ISCB)," *Bioinformatics*, vol. 14, no. 6, pp. 471, 1998.

[4] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *JMB*, vol. 235, pp. 1501–1531, Feb. 1994.

[5] Yasubumi Sakakibara, Michael Brown, Richard Hughey, I. Saira Mian, Kimmen Sjölander, Rebecca C. Underwood, and David Haussler, "Stochastic context-free grammars for tRNA modeling," *NAR*, vol. 22, pp. 5112–5120, 1994.

[6] Richard Hughey and Anders Krogh, "Hidden Markov models for sequence analysis: Extension and analysis of the basic method," *CABIOS*, vol. 12, no. 2, pp. 95–107, 1996, Information on obtaining SAM is available at `http://www.cse.ucsc.edu/research/compbio/sam.html`.

[7] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia, "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods," *JMB*, vol. 284, no. 4, pp. 1201–1210, 1998, Paper available at `http://www.mrc-lmb.cam.ac.uk/genomes/jong/assess_paper/assess_paperNov.html`.

[8] "The Human Genome," *Nature*, 15 Feb. 2001.

[9] D. Kulp, D. Haussler, M.G. Reese, and F. Eeckman, "A generalized hidden Markov model for the recognition of human genes in DNA," in *ISMB-96*, St. Louis, June 1996, pp. 134–142, AAAI Press, http://www.cse.ucsc.edu/~dkulp/cgi-bin/genie.

[10] Kevin Karplus, Kimmen Sjölander, Christian Barrett, Melissa Cline, David Haussler, Richard Hughey, Liisa Holm, and Chris Sander, "Predicting protein structure using hidden Markov models," *Proteins: Structure, Function, and Genetics*, vol. Suppl. 1, pp. 134–139, 1997.

[11] Kevin Karplus, Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, and Richard Hughey, "Predicting protein structure using only sequence information," *Proteins: Structure, Function, and Genetics*, vol. Supplement 3, no. 1, pp. 121–125, 1999.

[12] J. D. Hirschberg, D. Dahle, K. Karplus, D. Speck, and R. Hughey, "Kestrel: A programmable array for sequence analysis," *Journal of VLSI Signal Processing*, vol. 19, pp. 115–126, 1998.

[13] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[14] A. J. S. Rayl, "Designer degrees or academic alchemy?," *The Scientist*, vol. 14, no. 7, pp. 41, 3 Apr. 2000.