Teaching Bioinformatics in a Mathematics Department

Steven Deckelman[1]

University of Wisconsin-Stout

Name of Institution: University of Wisconsin-Stout	
Size	about 9,000 students
Institution Type	Primarily undergraduate institution
Student Demographic	Majors in the Applied Mathematics and Computer Science Program with a Bioinformatics Concentration
Department Structure	Mathematics, Statistics, and Computer Science are in the same department.

Abstract

Undergraduate programs in bioinformatics are usually in a life science department, a computer science department, or both. A bioinformatics program in a mathematics department can include much more mathematics. We describe the discipline of bioinformatics, connections between bioinformatics and the mathematical sciences, undergraduate research possibilities, and resources for mathematicians who want to learn more about bioinformatics or develop a bioinformatics curriculum. Our concentration in bioinformatics in the Department of Mathematics, Statistics and Computer Science at the University of Wisconsin-Stout, and part of a BS degree in Applied Mathematics and Computer Science, will be described as an example.

Course Structure

- Weeks per term:16 week semester
- Classes per week/type/length: Three fifty-five minutes classes each week.
- Labs per week/length: One fifty-five minute lab per week in bioinformatics capstone. Required science courses will have science labs.
- Average class size: 20 Students; 10 or fewer students the bioinformatics capstone.
- Enrollment requirements: Students should be in the bioinformatics concentration within the Applied Mathematics and Computer Science major.
- Faculty/dept per class, TAs: Taught by a single faculty member.
- Next course: Students in bioinformatics may want to take related courses in biology.
- Web pages: UW Stout, 2012a; UW Stout, 2012b.

[1] deckelmans@uwstout.edu

What is Bioinformatics?

The term *informatics* has sometimes been used to describe fields in which massive amounts of data are generated and need analysis, interpretation, and application. For example, medical informatics refers to ways of organizing, using, and making sense out of the massive amount of medical research and patient record data that are being generated. Bioinformatics may be described as informatics that arise from the huge amounts of data generated by technologies for genomic sequencing. It is an area that draws on biology, chemistry, and physics, as well as mathematics, computer science, and statistics. There is no universally accepted definition of bioinformatics, and practitioners of different fields emphasize different aspects of it. Biologists may view bioinformatics as a collection of software to be used as tools in life science research, while computer scientists may regard it as a subspecialty of computer science. A mathematician exploring the subject will find it to be rich in mathematical structure. Bioinformatics is all of these things. Different conceptions of the field are emphasized to different degrees depending on the department in which the undergraduate bioinformatics curriculum resides. Bioinformatics is interdisciplinary, which makes its definition hard to nail down. It is sometimes likened to discrete computational biology, as in matching DNA sequences, as contrasted with the partial differential equations, energy equations, and approximate algorithms used to analyze protein folding. In this article, we use a broad definition that includes health informatics, computational biology, and algebraic statistics.

Much of the mathematical content of bioinformatics falls within the realm of traditional discrete mathematics. Indeed, bioinformatics could be thought of as applications of discrete mathematics in a particular biological context. Graph theory provides a natural framework for formulating fundamental problems such as biological sequence alignment. The problem of globally aligning two DNA sequences can be viewed as finding a longest path in a weighted directed graph (Jones, 2004). Tools such as generating functions arise naturally in the biological problem of restriction mapping, where the objective is to reconstruct a genomic sequence from a sequence of fragments cleaved by restriction enzymes. The notion of permutation is fundamental in the study of genome rearrangements in which organismal genomes are shuffled by evolutionary processes. Bioinformatics also uses statistics. Cluster analysis is a statistical technique that is useful in analysis of gene expression data from DNA arrays and in molecular phylogenetics---the reconstruction of evolutionary family trees. Probability (Waterman, 1995) arises in connection with hidden Markov models, a machine learning tool used in gene hunting (Durbin, 1998). Other ideas such as NP-completeness and complexity analysis are useful in understanding the nature and limitations of computational approaches to solve the problems of bioinformatics. The field has been developing rapidly and extending far beyond what mathematicians would recognize as traditional discrete mathematics. Algebraic statistics is a new subdiscipline that uses ideas of computational algebra and algebraic geometry to the study of statistical problems (Pachter, 2005). The statistical models of bioinformatics are viewed as algebraic varieties (zero sets of polynomial equations in several variables. This leads to

applications such as Gröbner bases to bioinformatics. The related field of biomedical informatics has seen application of mathematical ideas from number theory and quantum mechanics (Robson, 2007). Systems biology is a field which straddles bioinformatics (as in functional genomics) and mathematical biology. It is concerned with the study of biological information networks (metabolic pathways, biochemical networks, gene regulatory networks) and uses mathematical modeling techniques involving calculus, discrete dynamical systems, and differential equations (Alon, 2007). Additional resources on these ideas and their possible inclusion in the undergraduate curriculum will be given below.

Institutional Profile

The University of Wisconsin-Stout is a four-year comprehensive institution in the University of Wisconsin system located about an hour from Minneapolis-St. Paul. There are about 9000 students, mostly undergraduates. The Department of Mathematics, Statistics, and Computer Science has about twenty-five staff members, representing the three discipline areas. There are approximately 130 applied mathematics and computer science majors with about twenty graduates per year. There is also infrastructure to support bioinformatics. The university has recently been designated as a Polytechnic Institution, defined as one having a mission that focuses on STEM disciplines and emphasizes the use of scientific theory and research to solve real-world problems and contribute to the economy and society. There is interest in and support for biotechnology research and curriculum development. We have a biology faculty active in research who use bioinformatics tools. Indeed, the proposal to create an undergraduate program in bioinformatics came from our biology faculty. The Minneapolis-St. Paul area has a growing biomedical technology industry, and the campus is about two hours away from the Mayo Clinic in Rochester, Minnesota, and one hour away from the Marshfield Clinic in Marshfield, Wisconsin, both of which have active biomedical research divisions that make use of bioinformatics. The Marshfield Clinic has close research and exchange ties with the University of Wisconsin.

Our Program at the University of Wisconsin-Stout

Our program offers a BS degree in Applied Mathematics and Computer Science with a concentration in bioinformatics. The concentration combines mathematics, computer science, statistics, biology, and chemistry. In many respects the program is like a double Math-CS major. In mathematics, students take course work up to and including the traditional junior-senior level curriculum of modern algebra and real analysis, including a year of calculus-based statistics. In computer science, students take courses in computer organization, data structures and database systems, and upper division electives. The courses are described on our websites (UW Stout, 2012c; UW Stout, 2012d).

Mathematics Courses in the Bioinformatics Curriculum

Bioinformatics students take the same mathematics core courses as all majors, including three semesters of calculus, one semester of linear algebra, and a foundations course in mathematical language and proofs. Graph theory and numerical analysis are recommended electives. Although they are not required, our bioinformatics students typically take courses in real analysis and modern algebra to fill out their programs. These courses reflect our view that our bioinformatics students are still mathematics majors, albeit with a bioinformatics emphasis. There are additional required courses in computer science, statistics, biology, and chemistry. A two-semester senior capstone course in bioinformatics (see below) rounds out the curriculum.

Bioinformatics Content in the Curriculum

Because our degree program is in the mathematical sciences, albeit with an applied slant, part of the philosophy of the program has been to teach bioinformatics as a subject at the intersection of mathematics, statistics, and computer science. But since bioinformatics is a fundamental tool in the life sciences, we also teach students how these tools are used by scientists. We want our graduates to understand how to use bioinformatics tools as biologists do and to understand the mathematical ideas on which these tools are based. Resource constraints prevent us from developing very many new courses. Instead, our approach has been to develop a small number of bioinformatics courses, while putting additional bioinformatics content in existing courses in biology and chemistry as well as in computer science and mathematics.

In our computer science Data Structures class, Dr. Terry Mason (Mason, 2010) includes a module that shows students how to build a map data structure that translates a gene sequence into a protein sequence. Projects such as this expose students to the nucleotide representation of DNA and how it maps to RNA and eventually to protein. In our Database System course, Dr. Mason has students work on design and implementation of a biological database to collect information for various agents such as chemicals and nucleotide sequences. This work is part of a project in collaboration with the biology department to build web tools to run computations and data mining on their collected data. In our Advanced Biochemistry course, Dr. Marcia Miller-Rodeberg (Miller-Rodeberg, 2010) introduces students to the use of bioinformatics through examination of protein structures and the effects of mutations on protein function, specifically focusing on metabolic diseases. Protein sequencing and homology are discussed in relation to evolutionary trees, the evolution of new enzymes, etc. Students are also required to do a research paper and talk on a disease caused by mutations in an enzyme, with bioinformatics used to analyze structure and genetic relations. Biologist Steve Nold (Nold, 2010) integrates research questions into the laboratory and has his students create clone libraries and sequence DNA for bacteria in the Red Cedar watershed, sinkhole microbes, and grasshoppers. They take those sequences to GenBank, the National Center for Biotechnology Information Sequence Database, (NCBI, 2012a) for identifications using utilities such as Blast (Basic Local Alignment Search Tool, NCBI, 2012b).

The Bioinformatics Capstone

Our capstone course is a two-semester sequence. The first course (MSCS 492), which is taught by a mathematician or computer scientist, focuses on mathematical and computational aspects of bioinformatics, while the second (MSCS 493), taught by a life scientist, focuses on application in the life sciences. The prerequisite for MCSC 492 is senior standing in the bioinformatics major. This usually means that students have taken the full slate of courses mathematics majors typically take, from calculus to real analysis and modern algebra.

Topics in MSCS 492 include bioinformatics algorithms and programming in PERL, complexity and analysis of algorithms, proofs of correctness, software tools, and languages (Perl, Matlab etc.). It also provides an introduction to the classical problems of bioinformatics (motif finding, genome rearrangements, sequence alignments, gene prediction, phylogenetics), and some of the classical approaches to these problems (dynamic programming, combinatorial and statistical methods, hidden Markov models, graph theoretic formulations). We use the text by Jones and Pevzner, *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004. Students also have a senior project on which they have to give a talk. Some of the projects have dealt with computational methods for protein structural analysis (software tools for obtaining, visualizing, and performing *in silico* analysis on protein models). Other projects have dealt with experimental methods for obtaining biological data such as gel electrophoresis, DNA probing and sequencing. Other possibilities for a project include reading a research paper in the field such as "Bounds For Sorting by Prefix Reversal," an undergraduate researcher paper by Bill Gates from his Harvard days under Christos Papadimitriou.

The second course, MSCS 493, is called Bioinformatics Practicum. In it, biologist Michael Pickart has students work on projects with bioinformatics content. For example, one student focused on improving access to multiple-species sequence data for comparative genomics. The student examined existing genomics data portals at resource sites such as the National Human Genome Research Institute (NHGRI), National Center for Biotechnology Information, National Institutes of Health Intramural Sequencing Center, and UC Santa Cruz Genome Browser to determine the existing methods for mapping of human genes to ENCODE (The Encyclopedia of DNA Elements) regions. The ENCODE consortium was launched by NHGRI. Its goal is to produce a comprehensive catalog of about 1% of the structural and functional components encoded in the genomes of humans and other species. As the number of fully sequenced genomes grows, it becomes easier for researchers to use comparative genomic data; however, our student found that current access to this data was not straightforward, limiting our students' ability utilize this resource. To assist student genomic research of ENCODE data, a student project explored ways of restricting the data for human and zebrafish genes to facilitate mapping to ENCODE regions. Students mapped protein sequences of genes involved in pigment development from the zebrafish genome to the human genome. Zebrafish are a model

organism, a non-human species whose genomic structure gives insight into the human genome due to the common evolutionary descent of all species.

Special Curriculum Challenges (Theory vs. Real World)

Teaching bioinformatics at the undergraduate level involves curriculum challenges. Depending on the clientele, we may have to look for simple ways to teach mathematical ideas to biology students with minimal mathematics background. Likewise, when teaching mathematics students, there is the challenge of making the material biologically meaningful by connecting the theory to real world applications. An advantage to housing a bioinformatics program in a mathematics department is that the curriculum can be developed for students with a solid grounding in mathematics. One way of bridging the gap between theory and practice is to use real world biological sequence data. The National Center for Biotechnology Information website (NCBI, 2012c) contains links to large public databases of DNA and protein data. The European Bioinformatics Institute (EBI, 2012a) is another good resource. A student can gain a sense of the applicability of bioinformatics by downloading a genome and analyzing it. The curriculum experience for students can be enhanced through either team teaching with biologists or developing paired biology courses that complement the mathematical bioinformatics. In a mathematics class, for example, students can be taught the mathematical basis of cluster analysis, and to understand the nature of the data furnished by DNA microarrays and how to analyze these data. In a paired biology or biotechnology class, students can gain valuable wet lab experience using DNA arrays. Likewise, students can learn the ideas of hidden Markov models in a mathematics class while applying the technique to gene finding in a biology laboratory. This approach is currently under development at the University of Wisconsin-Stout.

Undergraduate Research Possibilities in Bioinformatics

Bioinformatics is rife with opportunities for interdisciplinary undergraduate research. Because bioinformatics is a relatively new field, there is a wealth of open problems (Pevzner, 2000), many of which are understandable by undergraduates. At the same time, the life sciences are generating large data sets in need of analysis. Student projects can involve studies of mathematical and algorithmic ideas for solving problems or software development for the analysis of available biological sequence data. Partnering with biologists who make use of genomics in their research creates an opportunity for interdisciplinary student projects with interdisciplinary mentoring. Finally, there are a growing number of REU and similar opportunities for students in various biomedical, industrial, educational, and research settings. The Rochester Institute of Technology maintains a list of these (RIT, 2012a).

Some Process Issues

Originally the suggestion for a bioinformatics concentration at UW-Stout came from the biology faculty. With an undergraduate major of applied mathematics and computer science, the Department of Mathematics, Statistics, and Computer Science already had the necessary

infrastructure for a bioinformatics program with a strong mathematics component. With the exception of the capstone course and a lower level introduction to bioinformatics course still under development, the program was put together with minimal additional resource requirements by using existing courses. Thus the administrative hurdles that can arise with new programs were never an issue. Once the program was up and running, advising of majors has been done jointly by mathematics and biology faculty who meet with students together.

Resources for Retooling in Bioinformatics and Bioinformatics Curriculum Development

How much biology does a mathematician need to know to get involved with bioinformatics? While it helps to have some exposure to biology, it's not necessary to be a credentialed biologist or to have wet lab skills to get involved in bioinformatics. Indeed, most mathematicians have minimal or no real training in biology and most biologists lack advanced mathematics or computational training; this makes interdisciplinary collaboration especially fruitful. It is helpful to know enough about molecular biology to be able to understand, at least conceptually, the nature of data produced by biological experiments and the questions biologists are interested in. At the undergraduate level, much of the biology needed for bioinformatics can be distilled down into what is often referred to as the central dogma of molecular biology: "DNA makes RNA makes proteins." The molecular biology text by Alberts, Bray, et. al. (Alberts, et al 1994) gives a readable introduction to the subject. The MAA publication *Math & Bio 2010* (Steen, 2005) includes a list of bioinformatics resources, as does Claudia Neuhauser and Kristine Fowlers' *Recommended Resources in Mathematical Biology* (Fowler, 2004). Bioinformatics may also be regarded as a subspecialty of mathematical biology. The Mathematical Biosciences Institute at the Ohio State University offers visiting and sabbatical opportunities for mathematics faculty. The institute also has a wealth of curriculum-relevant resources on its website (MBI, 2012a), and offers regular tutorial and workshops in all areas of mathematical biology including bioinformatics. The BioQUEST Curriculum Consortium is specifically aimed at integrating undergraduate mathematics and biology curriculum (Bioquest, 2012a). It offers workshops and a series of publications and maintains the BEDROCK (Bioinformatics Education Dissemination: Reaching Out, Connecting, and Knitting-together) project, a repository of curriculum resources that support "an inquiry-based approach in which students explore and analyze actual data in a way that recreates the experience of conducting research." Both the Mathematical Association of America and the Society for Industrial and Applied Mathematics have interest groups in mathematical biology that include bioinformatics, and there is also a Society for Mathematical Biology. Curriculum development resources are now available online at such sites as Pavel Pevzner's page, (Bioalgorithms, 2012a), and Lior Pachter's homepage, (Pachter, 2012a).

Summary

Bioinformatics employs computational methods and technologies to analyze biological sequence data. Its methods have origins in computational biology, a field at the junction of

mathematics and computer science in which the underlying problems are mathematical in nature. To acquire an understanding of the field and its tools, students need to understand the underlying mathematical problems, which draw on discrete mathematics, theoretical computer science, probability and statistics, algebra, and algebraic geometry. Mathematicians are ideally suited to contribute to bioinformatics curriculum development and the education of the next generation of life scientists.

References

Alberts, Bray, et al. (1994). *Molecular Biology of the Cell.* New York: Taylor and Francis.

Alon, U. (2007). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Boca Raton: Chapman Hall/CRC.

Bioalgorithms, 2012a, http://www.bioalgorithms.info

Bioquest, 2012a, http://www.bioquest.org

Durbin, et al. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

EBI, 2012a, http://www.ebi.ac.uk/2can/genomes/genomes.html

Fowler, K. (2004). *Using the Mathematics Literature.* New York: Marcel Dekker.

Gates, W. and Papadimitriou, C. (1978). Bounds For Sorting By Prefix Reversal, *Discrete Mathematics* (27).

Mason, T.(2010). Private Communication.

MBI, 2012a, http://www.mbi.osu.edu

Miller-Rodeberg, M. (2010). Private Communication.

NCBI, 2012a, http://www.ncbi.nlm.nih.gov/genbank

NCBI, 2012b, http://blast.ncbi.nlm.nih.gov/Blast.cgi

NCBI, 2012c, http://www.ncbi.nlm.nih.gov

Nold, S. (2010) Private Communication.

Pacheter, 2012a, http://mathematics.berkeley.edu/~lpachter

Pachter, Sturmfels. (2005). *Algebraic Statistics for Computational Biology.* New York: Cambridge University Press.

Pevzner, P. (2000). *Computational Molecular Biology: An Algorithmic Approach*. Cambridge Massachusetts: MIT Press.

Pevzner and Jones. (2004). *An Introduction to Bioinformatics Algorithms.* Cambridge Massachusetts: MIT Press.

RIT, 2012a, http://www.rit.edu/~gtfsbi/Symp/bioinformatics.htm

Robson, B. (2007). The New Physician as Unwitting Quantum Mechanic: Is Adapting Dirac's Inference System Best Practice for Personalized Medicine, Genomics, and Proteomics? *Journal of Proteome Research* , 3114-3126.

Pickart, Michael. (2010) Personal Communication.

Steen, L. (2005). *Mathematics & Bio 2010: Linking Undergraduate Disciplines.* Washington D.C.: The Mathematical Association of America.

Tymann, P. E. (2005). Computer Science and Bioinformatics. In L. Steen, *Mathematics & Bio 2010, Linking Undergraduate Disciplines* (pp. 75-82). Washington D.C.: The Mathematical Association of America.

UW Stout, 2012a, http://www3.uwstout.edu/programs/bsamcs/index.cfm

UW Stout, 2012b, http://www3.uwstout.edu/programs/bsamcs/upload/c_bsamcs_bio.pdf

UW Stout, 2012c, http://www3.uwstout.edu/programs/bsamcs/index.

UW Stout, 2012d, http://www3.uwstout.edu/programs/bsamcs/upload/c_bsamcs_bio.pdf

Waterman, M. (1995). *Introduction to Computational Biology: Maps, Sequences, and Genomes*. Boca Raton: Chapman & Hall/CRC.