Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 67–72
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

An information technology emphasis in biomedical informatics education

Michael D. Kane *, Jeffrey L. Brewer

Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907-1421, USA

Received 5 December 2005
Available online 9 March 2006

Abstract

Unprecedented growth in the interdisciplinary domain of biomedical informatics reflects the recent advancements in genomic sequence availability, high-content biotechnology screening systems, as well as the expectations of computational biology to command a leading role in drug discovery and disease characterization. These forces have moved much of life sciences research almost completely into the computational domain. Importantly, educational training in biomedical informatics has been limited to students enrolled in the life sciences curricula, yet much of the skills needed to succeed in biomedical informatics involve or augment training in information technology curricula. This manuscript describes the methods and rationale for training students enrolled in information technology curricula in the field of biomedical informatics, which augments the existing information technology curriculum and provides training on specific subjects in Biomedical Informatics not emphasized in bioinformatics courses offered in life science programs, and does not require prerequisite courses in the life sciences.

Keywords: Biomedical informatics; Bioinformatics; Information technology; Education; Interdisciplinary

1. Introduction

The growth of the biotechnology industry in recent years is unprecedented, and advancements in molecular modeling, disease characterization, pharmaceutical discovery, clinical healthcare, forensics, and agriculture fundamentally impact economic and social issues worldwide. Research and discovery activities in the life sciences were once limited to a single gene or protein, but the development of computational and information systems (integrated with biotechnology) has facilitated a shift to high-throughput screening (thousands of samples per day) and high-content detection systems (thousands of data points per sample), and the supporting information system represents the enabling factor in these endeavors. The National Center for Biotechnology Information (NCBI) maintains a growing collection of databases housing genetic sequence and protein structure/activity data (among other biological data sets), which are currently growing at an exponential rate. On February 15, 2005, the NCBI published its 146th release of GenBank, a flat-file database of gene sequences, which contains 42,734,478 gene sequences [1]. In addition, genomic research enjoyed a landmark accomplishment with the completion and publication of the human genome in February, 2001 [2,3]. This represents one of two fundamental advancements in the last 10 years that have moved genomic research almost completely into the computational domain, and dedicated information systems now represent the enabling factor for life sciences research.

As mentioned, the human genome has been completely sequenced and published [2,3], yet the boon of high-throughput gene sequencing is not limited to the human species. Many other genomes have been completely sequenced representing a breadth of mammalian, agricultural, viral and bacterial organisms, and this fundamental advancement in genomic data availability offers scientists

* Corresponding author. Fax: +1 765 496 1212.
E-mail address: mdkane@purdue.edu (M.D. Kane).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.02.006

the informational basis of living systems. The other fundamental breakthrough in biotechnology can best be described as "high-content" genomic screening, where the integrity (single nucleotide polymorphisms) and activity (gene expression profiling) of every gene in a known genome can be detected from a single biological sample. This is accomplished using the DNA microarray technology platform, which can detect tens of thousands of genes using a small functionalized system the size of a postage stamp [4–6]. The information flowing from genomic laboratories using DNA microarray technology constitutes hundreds of thousands of gene-specific measurements each day. The overall impact of this revolutionary technology depends upon an integrated information system to analyze and store the data, as well as computational systems to design each of these gene-specific detections (hybridizations). The Food and Drug Administration (FDA) has recently published guidelines for the development of biotechnology methods, such as the DNA microarray platform, for use in genome-based prognostics and diagnostics in humans [7], which marks the beginning of a new era in healthcare that utilizes a patient's genome sequence to enable "personalized medicine."

With this explosion of molecular data and biotechnology capabilities, the pharmaceutical, biotechnology, and healthcare industries are dependent on professionals with information systems and technology skills to fully exploit these resources and translate molecular and cellular data into new genetic and therapeutic discoveries, as well as develop new biotechnologies to impact human health. Similarly, information technology professionals must be trained in the utilization of biomedical data structures and capable of developing and integrating information systems with biotechnology systems to meet this challenge.

Given that bioinformatics has evolved to meet the objectives of life sciences research, courses and curricula that offer training in bioinformatics, or biomedical informatics [8], have primarily originated in life science programs where traditional life sciences training in genetics, genomics and proteomics is augmented with training in sequence alignments, genomic data analysis and protein modeling. This training largely involves the use of existing software applications dedicated to queries in gene and protein sequence databases. Yet this training is not appropriate for students enrolled in information technology programs that lack what is considered prerequisite knowledge in biology and genetics. However, students enrolled in information technology programs are developing skills that are directly applicable to biomedical informatics such as data formats, database structure and development, applications development, and systems design. Given that both fields of study are important to training in biomedical informatics; interdisciplinary training must be developed that develop skills relevant for both groups of students.

It is important to mention that the terminology utilized for training information technology students is described herein as "biomedical informatics" since the learning objective of this training is to apply information sciences and technology skills to all subdisciplines within this domain, and aspects from all defined subdisplines are included in the courses described. It is recognized that "bioinformatics" is a subdiscipline of "biomedical informatics" [8], and the primary emphasis of subject matter in the course series falls within the realm of bioinformatics. This larger emphasis on bioinformatics is intentional simply to provide the necessary background in areas most lacking in existing information technology education, since students have little or no background in life sciences. Furthermore, the author's institution does not harbor a medical school or clinical research environment, and information technology students are much more likely to apply their skills to independent or course-directed projects in the basic life sciences during their undergraduate and graduate education. This approach addresses the lack of knowledge of biology that is typical of information technology students and provides the fundamental aspects of biotechnology, genomics and statistics, which augment skills in information technology, ultimately to design, develop and implement discovery support information systems, while providing exposure to ongoing research projects across campus for demonstration purposes and student project activities.

The authors are involved in a campus-wide graduate-level interdisciplinary Computational Life Sciences program that involves a distinct (and growing) list of academic departments including Computer and Information Technology, Biological Sciences, Statistics, Chemical Engineering, Computer Sciences, Agronomy, and Electrical & Computer Engineering. This program offers graduate students in any department a breadth of training opportunities, and has been developed to encourage students to develop skills outside their formal discipline. In support, the authors have developed two 1-credit courses specifically for students in the life sciences to study: (1) Biomedical Systems Architectures and (2) Biomedical Systems Analysis and Design. Furthermore, the authors have developed two courses to allow information technology students to develop skills in biomedical informatics. The focus of this manuscript is limited to the development and rationale of a formal education program in biomedical informatics specifically for students majoring in information technology programs.

2. Student audience

One of the fundamental decisions to consider when designing an interdisciplinary program, in this case the integration of life sciences with information technology, is how to most effectively augment one discipline with the relevant content from the other. In the present case, does an educational program intend to develop courses that augment a life sciences curriculum with information technology training, or intend to develop courses in the life sciences for students in an information technology curriculum? This fundamental distinction is debatable, is

certainly not exclusive, and often reflects the strengths and needs of a given academic setting. Presently both approaches exist in academia, and are evolving to meet the strengths of the respective student audiences. Specifically, there are many components of an information technology curriculum that are not required for life science students to become proficient in bioinformatics. Similarly there are many components of a life sciences curriculum that are not necessary for information technology students to be proficient in biomedical informatics. For example, students in an information technology program will develop skills in the bioinformatics domain by understanding the data types (e.g., gene sequences), search query methods and algorithms, existing databases, development and utilization of distributed architectures, as well as processes that are evolving to directly integrate with common biotechnology methods (e.g., DNA microarray data collection and analysis, automated DNA sequencer data collection and analysis).

The distinction between life scientists developing skills in information technology versus information technologists developing skills in life sciences is best articulated as bioinformatics tool users and bioinformatics tool builders, respectively. This distinction captures the rationale for content in courses developed for information technology students, which are seen as both hardware and software tool builders. Notably, the common ground between 'tool builders' and 'tool users' is applications development, which is a difficult subject to address completely in bioinformatics given the breadth of programming languages utilized (e.g., C++, java, perl, VB, etc.) as well as the emergence of new application development tools (e.g., bioperl, XML, etc.), much of which already exists in information technology curricula. Furthermore, the development, implementation and utilization of bioinformatics systems require that operational and architectural specifications be developed, which form the basis for a development plan. This rationale reflects the professional activities in the pharmaceutical, biotechnology and biomedical industry where an interdisciplinary team is tasked with developing or updating the capabilities of systems in this arena. In support, a task force within the American College of Medical Informatics (ACMI) published their recommendations regarding the objectives of biomedical informatics training, which included the adoption of interdisciplinary curricula, and challenged educators to provide training in diverse fields of study (Table 1) [8]. Educators and professionals considering training issues and objectives in biomedical informatics are encouraged to read this task force report from the ACMI [8].

Furthermore, the creation of interdisciplinary curricula as described in this paper reflects the proposed Information Technology (IT) model curriculum outlined in the Association for Computing Machinery (ACM) Special Interest Group on Information Technology Education (SIGITE) curriculum committee report dated April 2005 [9]. As you can see from Table 1, the skills identified by the ACMI report can be directly linked to learning outcomes addressed in the proposed model IT curriculum. This intersection of educational objectives provides the rationale for IT curricula strongly supporting a Biomedical Informatics program.

3. Course development

The development of a course series in biomedical informatics for students in our information technology program spans the upper-undergraduate and graduate level. The rationalization for this level of training involves the assumption that students will *apply* skills developed in their undergraduate information technology program to the bioinformatics domain. We have developed a 1-credit introductory course that is offered to both graduate and undergraduate students, which provides prerequisite content for the more intensive graduate-level coursework. Importantly, students do not require any prerequisite life sciences coursework for the course series described herein. Certainly previous coursework in biology benefits the students (and typically reflects a given student's interest in this area), however we have not identified an existing course in the life sciences (to serve as a prerequisite course) that provides training in, for example DNA sequence, without much more emphasis placed on molecular aspects of genetics that are not requisite knowledge for the biomedical informatics course series. As an example, our students do not need to be capable of diagramming the structures of nucleotide and amino acids, the building blocks of DNA and proteins, to be proficient in sequence alignment queries in DNA and protein databases.

The overarching goal of this training is to provide information technology students the ability to support team-based bioinformatics systems development and integration. The team-based aspect is paramount as information technology students are not likely to initiate and lead life sciences-based research efforts (e.g., genomics, proteomics, cheminformatics, metabolomics, etc.), which involve interdisciplinary teams of biologists, statisticians, and information systems specialists. Under this interdisciplinary team model, information systems specialists conduct *requirements discovery* by being engaged in perpetual dialogue with biologists to understand the immediate and long-term objectives of the project, and therefore need to understand the methods and data structures common, and under development, in this domain. For example, using genomic tools to research the genetic basis of cancer involves a clinical understanding of cancer and effective therapies, as well as an information system to manage and query patient-specific health-related data, DNA microarray data (raw data images, gene-specific data, single nucleotide polymorphism data, etc.), sequence databases, and protein databases. Importantly, the first clinical manifestation of personalized medicine is upon us [7], where genomic methods will be employed to identify common single nucleotide polymorphisms (SNPs) in specific drug metabolizing enzymes for

Table 1
Core skills from fields related to information and computational sciences

American College of Medical Informatics (ACMI) Report [8]	Special Interest Group on Information Technology Education (SIGITE) Document [9]
Algorithms	Section PF4: Algorithms and problem solving
Basic mathematics, including calculus	Section 8.1.1: Mathematics
Cognitive/human factors and interfaces	Section HCI: Human computer interaction
Data structures	Section PF1: Fundamental data structures
Database design	Section IM2–IM6: Database query languages, data organization architecture, data modeling, managing the database environment, and special-purpose databases
Evaluation/research methods	Section 8.1.2: Scientific method
Information retrieval	Section IM: Information management
Knowledge representation	Section IM: Information management
Modeling	Section IM4: Data modeling and throughout the entire curriculum
Networking/architecture	Section NET: Networking and IAS (information assurance and security)
Ontology/vocabulary	Section 8.1.3: Familiarity with application domains
Probability/statistics	Section ITF6: Applications of math and statistics to IT
Programming languages	Section PF: Programming fundamentals and WS Web systems and technologies
Simulation	Section 6.2: Experiential learning
Software engineering	Section SIA: System integration and architecture

the purposes of identifying persons that have a deficiency in metabolizing (clearing) certain drugs. Although much of the hype regarding this subject surrounds the impact of the human genome sequence and genomic biotechnology, the implementation of this genetic screening system in the healthcare industry requires the development of a supporting, scalable information system integrated with multiple testing biotechnologies, databases, and interfaces for lab testing staff, health care providers, insurance carriers, and the general public, not to mention the relevant information security issues in patient-specific data transfer and storage. Importantly, this system must be scalable to accommodate a growing number of patients, as well as the inevitable growth of genetic-based tests/information for each patient.

3.1. Course 1: introduction to computational life sciences

The first course in this series is entitled ''Introduction to Computational Life Sciences.'' This course introduces students to the central dogmas of cell biology, or more specifically how information flows through a cell system. This includes DNA (nucleotide sequence and gene structures) as an information repository, messenger RNA as the genetic information retrieved and required for a specific cell system, and protein as the functional entity and outcome of genetic information utilization (in molecular terms; transcription and translation). As students grasp the information contained in the primary sequence of gene and proteins, the methods used to analyze molecular sequences and search for sequence similarity across and between genomes and proteomes is studied. This proceeds into the relationship between protein sequence and function, as well as how pharmacological agents bind to, and disrupt protein function, thereby introducing the students to processes in virtual drug screening. The integration with high-throughput and high-content biotechnology platforms is studied to offer students a ''state of the art'' view of biotechnology, which includes class sessions in a genomics laboratory (DNA microarray), proteomics laboratory (mass spectrometry) and atomic force microscopy laboratory (cellular imaging). Finally, the course concludes with a survey of processes essential to public health and clinical medicine, including emerging areas such as teleradiology and genetic-based personalized medicine.

The learning outcomes for the course include an understanding of nucleotides and amino acids as building blocks for gene and proteins, respectively. How genetic information is ''transcribed'' into an active template, ultimately ''translated'' into functional proteins. The methods used to represent nucleotides and amino acids in the digital realm, and how to access the enormous (publicly available) gene and protein databases available through client-server architectures. The algorithms (Smith–Waterman [10], Needleman–Wunsch [11]) used to efficiently carry out pair-wise sequence alignment queries (e.g., BLAST) and multiple sequence alignments (e.g., CLUSTAL). The methods and rationale for constructing cladograms to measure paralogous and orthologous relationships based on gene and protein sequence similarity. The genetic basis of single nucleotide polymorphisms (SNPs), existing SNP databases, and the role of SNPs in personalized medicine and forensics. The rationale and utilization of DNA design applications supporting molecular biology methods (e.g., PCR primer design). The rationale and utilization of 3-dimensional protein rendering applications. All aspects of DNA microarray technology including DNA probe design (gene-specific sequence analysis), sample labeling and hybridization, microarray imaging and image analysis, data analysis, and gene-specific bioinformatics (k-means clustering, hierarchical clustering, similarity clustering, and sammon mapping). The emphasis on DNA microarray technology reflects the emerging utilization of this assay platform in both basic and clinical research. In addition the course includes on-site demonstrations of

"high-content" biotechnology applications, the existing data acquisition and management systems, as well as discussions with users regarding the short-comings of the existing data acquisition and management systems. Currently, we demonstrate DNA microarray technology (gene-specific data involving thousands of genes), mass spectrometry (protein analysis) and atomic force microscopy (3-dimensional detection and rendering of cellular structures).

3.2. Course 2: biomedical informatics: systems analysis and design

The second course in this series is entitled "Biomedical Informatics: Systems Analysis and Design." This course applies the skills in systems analysis and design (SAD) to the biomedical informatics domain. More specifically, the analysis and development of systems in this domain involves identifying the computational, architectural and performance specifications of a given system or subsystem to be developed and implemented to utilize genomic and/or proteomic data in basic and applied research (e.g., health care, pharmaceutical research, veterinary sciences, agriculture, biodefense, environmental sciences, etc). Students are exposed to a survey of information systems that support academic, governmental and industrial information systems recently developed to manage genomic and proteomic data. These systems are dissected to reveal the rationale for their (often distributed) architectures, data storage requirements, and data processing capabilities. Furthermore, the limitations of these systems are discussed as we explore the next generation of biomedical informatics systems. Students are assessed on their ability to perform in a team environment and design dedicated systems for the management of genomic and proteomic data. For example, students are designing systems to support genome-based personalized medicine, where patient-specific genetic information is derived using DNA microarray technology, screened for parameters associated with genetic risk of disease, generate categorical responses and reports based on results (e.g., warning of heightened risk of adverse drug response), integration with corollary health data, securely stored and accessed, and contribute to an anonymous database of human genetic profiles.

The objective of the course content and team-projects is to allow students to understand how genomic and proteomic data is utilized in a research setting, a cost-benefit analysis of these systems during the systems development life cycle, the integration with existing biotechnology platforms and data formats, as well as the financial implications for these systems in health care and pharmaceutical research. In addition, students are expected to read and report on assigned research papers within the bioinformatics domain. The learning outcomes for this course involve the development of knowledge of bioinformatics systems performance objectives, the application of systems analysis and design methods to the development of biomedical information systems, the ability to evaluate existing biomedical information systems for limitations in scalability and integration with other systems, and the ability to assemble specifications and design an information system to utilize genomic and/or proteomic data.

4. Programmatic analysis

It is difficult to exhaustively evaluate the educational foundation of bioinformatics training for the simple fact that this interdisciplinary field is very new, and new undergraduate and graduate programs are emerging each year. It is a fair generalization that these emerging biomedical informatics programs have been structured to offer cross-training in both molecular biology (life sciences) and information/computer technology, with an inference that any given student can excel in both arenas. More specifically, it is agreed that students within the primary area of life sciences can be trained in various aspects of information technology and computer science, yet are not expected to develop the depth of training that is expected of IT professionals and computer scientists. Certainly the contrary is true that information/computer technology students are not expected to garner the depth of training that is expected of molecular biologists. Given the disparate, yet overlapping nature of these professional domains, biomedical informatics struggles to define itself as a distinct domain, particularly when viewed from an educational/training perspective. It is the authors' view that as the arena of biomedical informatics continues to evolve, the primary influence on its evolution (and what is expected of a "biomedical informatician") is derived from the information technology arena and the storage, retrieval and processing capabilities of modern information systems. More specifically, information technology professionals that command a basic understanding of the data types and research objectives of life sciences research will be tapped to develop the systems to support the clinical implementation of genomic, genotyping and other high-content assay platforms. Therefore the preferred approach to training in biomedical informatics is to consider genomics and proteomics as application domains for information system development students who are familiar with the dynamic state of information technology capabilities and are trained to design, develop, and implement information systems that work within existing data types (e.g., FASTA sequence file formats) and can accommodate emerging data types from biotechnology advancements, evolving systems modeling and development languages, and the continuing evolution of hardware and computational capabilities. It is this perspective that forms the foundation of our emerging program.

Given the fact that new academic biomedical informatics programs are constantly emerging, the authors have chosen to compare their course development strategy to their peer institutions in information technology education. The peer institution group includes five universities that co-founded the Special Interest Group on Information

Technology Education (SIGITE) organization, which include Purdue University, Rochester Institute of Technology (RIT), Brigham Young University (BYU), Georgia Southern University (GSU), and the University of Pennsylvania (PENN). All of these institutions have embarked on the development of interdisciplinary training in biomedical informatics, with the exception of GSU, as of June, 2005.

Perhaps the most distinct aspect of the course series described in this manuscript, and the intended student audience, is the absence of prerequisite coursework in the life sciences. More specifically, the development of a course series that makes no inferences to expected training in molecular biology or biochemistry. As described above, the students derive an understanding of these areas as they relate to data formats and storage, integration with existing biotechnology applications (e.g., DNA sequencing, microarray analysis) and client-server system architectures developed to support routine sequence analysis (e.g., BLAST). A comparison to our peer institutions must be qualified with the fact that their programs are primarily housed in the life sciences arena and are more developed as formal programs. A simple assessment reveals this trend as a balance of life science courses versus computer/information technology (IT) courses in each of the peer institution's highest degree programs in bioinformatics; MS degree at RIT (4 IT courses, 8 non-IT courses), BS degree at BYU (6 IT courses, 14 non-IT courses), and MS degree at PENN (2 IT courses, 6 non-IT courses). It is important to note that this simple description of programmatic requirements is limited to the minimum requirements reported in published documents, and additional course hours may be required but the specific courses are chosen at the student's discretion. Purdue University has initiated a similar interdisciplinary graduate program to offer cross-training for students pursuing diverse graduate degrees, which is distinct from the emerging course series described in this manuscript. Fundamentally the author's efforts reflect a need for more involvement from the IT side, ultimately developing and supporting the need for impact of biomedical informatics in the clinical arena.

5. Conclusion and future directions

The future of this course series will involve the incorporation of ongoing research projects both on campus and available through outside partnerships. The rationale for this "advanced" course is to allow students to function within an interdisciplinary team development environment, contribute to the project, and present their contribution at a national conference. Students at this level will have demonstrated knowledge of this domain both through coursework and contributions to a specific project.

The employment prospects for graduates in biomedical informatics are inviting, and unprecedented growth for this interdisciplinary domain reflects the recent advancements in genomic sequence availability, high-content biotechnology screening systems, as well as the expectations of computational biology to command a leading role in drug discovery and disease characterization. In 2004, the California Employment Development Department reported the top two occupations leading employment growth were in this interdisciplinary field of study. The forecasts included 99% job growth for a master's degree-level Bioinformatics Specialist, as well as 59% job growth for a bachelor's degree-level Scientific Programmer Analyst, over the next 5 years [12].

References

[1] http://www.ncbi.nlm.nih.gov/—data obtained on April 5, 2005 from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.
[2] Venter JC et al. The sequence of the human genome. Science 2001;291(5507):1304–51.
[3] Lander ES et al. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860–921.
[4] Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999;21(Suppl 1):33–7.
[5] Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. Nat Genet 1999;21(Suppl 1):48–9.
[6] Heller MJ. DNA microarray technology: devices, systems, and applications. Annu Rev Biomed Eng 2002;4:129–53. Epub 2002.
[7] Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document: Drug Metabolizing Enzyme Genotyping System. Available from: <http://www.fda.gov/cdrh/oivd/guidance/1551.pdf>.
[8] Friedman CP et al. Training the next generation of informaticians: The impact of "BISTI" and bioinformatics—A report from the American College of Medical Informatics. J Am Med Informatics Assoc 2004;11(3):167–72.
[9] Association for Computing Machinery Special Interest Group for Information Technology Computing Curricula. Retrieved on June 1, 2005 from www.acm.org/education/curricula.html.
[10] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–7.
[11] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53.
[12] Heintz R. Biotech's winning formula for steady job growth. California Job J 2004.