Bioinformatics in support of molecular medicine

Russ B. Altman, MD, PhD
Stanford Medical Informatics
Stanford University, 251 Campus Drive, MSOB X-215
Stanford, CA 94305-5479, altman@smi.stanford.edu

Abstract

Bioinformatics studies two important information flows in modern biology. The first is the flow of *genetic information* from the DNA of an individual organism up to the characteristics of a population of such organisms (with an eventual passage of information back to the genetic pool, as encoded within DNA). The second is the flow of experimental information from observed biological phenomena to models that explain them, and then to new experiments in order to test these models. The discipline of bioinformatics has its roots in a number of activities, including the organization of DNA sequence and protein three-dimensional structural data collections in the 1960's and 1970's. It has become a booming academic and industrial enterprise with the introduction of biological experiments that rapidly produce massive amounts of data (such as the multiple genome sequencing projects, the large scale analysis of gene expression, and the large scale analysis of protein-protein interactions). Basic biological science has always had an impact on clinical medicine (and clinical medical information systems), and is creating a new generation of epidemiologic, diagnostic, prognostic, and treatment modalities. Bioinformatics efforts that appear to be wholly geared towards basic science are likely to become relevant to clinical informatics in the coming decade. For example, DNA sequence information and sequence annotations will appear in the medical chart with increasing frequency. The algorithms developed for research in bioinformatics will soon become part of clinical information systems.

In this paper, I briefly review the intellectual roots of bioinformatics and how the field has evolved in the last few years. Fortunately, a core set of scientific paradigms have provided a focus to the field. Even in this short period, however, there has been a change in the nature of the questions being asked and the types of experiments being attempted. These changes are consistently leading bioinformatics towards problems of clinical relevance. Some molecular biology information systems already have important clinical implications. I will discuss the differences in the culture and approach to science of clinical informatics and bioinformatics, but will argue that the two disciplines share important intellectual challenges which make them very closely allied fields (despite the cultural differences). Finally, I will identify a few areas common to both disciplines where developments in one field may help catalyze faster progress in the other. For example, useful database integration technologies have (arguably) matured more rapidly within bioinformatics than in clinical informatics. At the same time, clinical informatics embraced the idea of controlled terminologies relatively early, and offers lessons to those in bioinformatics attempting similar tasks.

What is Bioinformatics?

Bioinformatics, as a discipline (or subdiscipline) has been recognized for less than 10 years. The first efforts in bioinformatics can be traced back to the early application of computers to biology in the 1950's and 1960's. The early applications of computers to molecular biology were for graphical rendering of three-dimensional molecular structures (1), and the creation of databases of molecular sequence (2) and three-dimensional structure information (3). Bioinformatics can generally be defined as the study of how information technologies are used to solve problems in biology. The precise definition of bioinformatics is a matter of some debate. The most narrow usage of the term refers to the creation and management of biological databases in support of genomic sequences. The most broad usage includes essentially all applications of computers and information sciences to problems in biology. Of course, bioinformatics is based on the fundamental paradigm of molecular biology: genetic information is stored in sequences of DNA bases (a four letter alphabet), which get translated into sequences of protein amino acid building blocks (a twenty letter alphabet). Protein sequences have a remarkable ability to reproducibly fold into a three-dimensional shape, and this shape confers on them the ability to form a variety of critical functions for

life: enzymatic catalysis, structural support, generation of motion, reception of signals between cells, and transduction of forces (light, pressure, shear) into chemical signals, to name a few. The functions of proteins combine within a cell to create a living apparatus. Unicellular organisms (such as bacteria) and multicellular organisms (such as people) then function in their environment to acquire nourishment and reproduce. Over time, the demands of the environment create a pressure for these organisms (or their offspring) to build proteins that are better suited to compete for resources. This leads, in turn, to changes in the DNA code in response to environmental pressures, and through Darwinian natural selection.

Perhaps the best way to define the field is to outline the scope of topics covered in the three principle scientific meetings in the field (the International Conference on Intelligent Systems for Molecular Biology[1], the Pacific Symposium on Biocomputing[2], and the Annual Conference on Computational Biology[3]). These meetings typically include reports of methodologies in the following areas: alignment and analysis of DNA and protein sequence information, three-dimensional alignment and analysis of macromolecular (RNA, DNA and protein) structure, assessment of how small molecules (e.g. potential therapeutic agents) interact with drug targets, integration of heterogeneous biological databases, representation of biological information to facilitate sharing, communication and automated analysis by computers, analysis of networks of interacting gene products, simulation of biological processes ranging from chemical reactions to intercellular communication, and analysis of data created by large scale biological experiments. The biological problems that are addressed include prediction of protein structure and function, design of small molecules to augment or inhibit biological function, analysis of complex genetic phenomena, design of modified macromolecules for medical or industrial uses, and understanding how genetic factors contribute to host susceptibility to disease (and pathogenicity of infectious agents). Computational biology is sometimes used synonymously with bioinformatics, although it tends to be more inclusive of all computational and information science approaches to biology, whereas bioinformatics currently is focused on computational *molecular* biology. This distinction is likely to erode because all of biology clearly has a molecular basis, and so new problems are likely to draw bioinformatics professionals increasingly into other areas of biology (4, 5).

There has been some debate about whether bioinformatics (or computational biology) represents a separate branch of scientific investigation, or is a subdiscipline of biology. This is a question about which many people using computer technologies in biology have a strong opinion. On one hand, bioinformatics professionals typically receive a different training from biologists--there is more emphasis on quantitative fields such as probability, statistics and computer science, and there is less emphasis on training in bench research techniques. These differences in training spill out into differences in the culture of conferences, professional societies and publications. For example, all three major bioinformatics conferences include rigorously peer-reviewed papers collected in printed proceedings (a practice typical of the computer science community), whereas biological conferences rarely contain peer-reviewed work, and do not "count" in standard measures of academic productivity. On the other hand, the problems being attacked in bioinformatics are biological ones, and so it can be considered a field of biology. This simple logic, however, is problematic. Biology is still dominated by experimentalists who do not always acknowledge the importance of computation as a research area within biology. "If you don't do bench experiments, you are not a real biologist" is a frequent refrain. Some biologists consider bioinformatics investigators to be important primarily as providers of biologically-relevant computing services, and do not acknowledge the set of core paradigms that guide bioinformatics research. The independent status of bioinformatics as a discipline will stand on the success of the field in defining a clear, separable research agenda, creating administrative units in academia and industry that support this agenda, and gaining attention and support from public and private funding agencies. To that end, the International Society of Computational Biology (ISCB) has recently been formed[4], and begun to address these issues in an organized manner.

Most senior investigators in bioinformatics have entered the field through a variety of backgrounds, including through traditional training in the biological sciences or computational science/statistics. There is now a small but growing set of training programs to train young scientists in bioinformatics. These programs typically include a

[1] http://www-lbit.iro.umontreal.ca/ISMB98/
[2] http://www.cgl.ucsf.edu/psb/
[3] http://www.mssm.edu/biomath/recomb98.html
[4] http://www.iscb.org/

mix of computer science, probability and statistics, and core biological courses. The key stimulus for the development of the field has been the availability of computer technologies to build large biological databases. In many ways, the pioneering efforts to create the basic databases for DNA sequences (GENBANK (6) and EMBL (7)) and for three-dimensional biological structures (the Protein Data Bank (3)) have created the infrastructure required to catalyze bioinformatics. In a oft-repeated sequence of events, biologists with valuable or voluminous data gathered together to create resources for the storage of this data in a standard format. These collections of data attracted the attention of computer scientists and statisticians (as well as other biologists) who then began to create algorithms for the analysis of these collections, and methods for connecting the databases together. In a very real sense, bioinformatics has illustrated the popular mantra "Build it, and they will come." Simply by organizing the data and making it available, cottage industries developed to support the deposition, retrieval, cross-linking and analysis of the data. As time passed, principles emerged and become part of the core paradigm within the field.

The recent explosive growth in bioinformatics can be traced to one phenomenon: the emergence of the genome sequencing projects, and other large scale data collection efforts in biology. The opportunities for understanding completely the biological processes underlying the normal physiology of both hosts and pathogens have created an expectation for a new generation of medical diagnostics and therapeutics. The pharmaceutical industry has recognized the importance of having people who both understand biology, and have skills in computing with biological data. The technologies catalyzing this growth include the genome sequencing projects (8, 9), as well as new methods for systematically determining which proteins physically interact with one another in a cell by creating all possible pairs of proteins in an experiment that can detect the presence or absence of an interaction (10, 11). Other technologies are being developed that allow a snapshot to be taken of all genes in a cell that are being used (or expressed) within a cell (12). The profile of gene expression over time can be used to understand the sequence of interconnected events that occur during the life of a cell. Obviously, an understanding of how these patterns of expression change in response to external challenges such as infections or physiological stress will be essential for understanding which sets of genes should be targeted for new therapeutic or diagnostic approaches. Similarly, understanding how these expression patterns change in response to internal challenges, such as the development of a cancer by improper regulation of important genes, will be critical for developing approaches to these problems.

Unlike other areas of science and engineering, biological data (including medical data) seems to have prominent features that make the straightforward transfer of technologies difficult. Biological data are sparser and noisier than the typical data from many areas of engineering. Biological objects rarely have straight lines and right angles, and so standard simplifications can become irrelevant. There are rarely reductionist models that can be used to summarize the behavior of the smallest subunits in a biological system (we simply do not yet know the rules by which molecules act). Probabilistic reasoning is critical in looking at biological systems. Finally, the concept of sequences (their representation, storage and analysis) is critical and central in biology, and does not occur in many other areas of information science.

Accomplishments of Bioinformatics

There are two classes of accomplishments that can be attributed to bioinformatics. The first is a set of empirical principles that have become accepted by the field, and create the context which guides both research and training. The second class is the particular artifacts (databases and algorithms) that have been developed, and which are serving biological science. A full exposition of the principles underlying the research agenda within bioinformatics would require an entire textbook, but can be summarized to provide a general sense of the field.

1. The structure of protein molecules is strongly conserved even when evolution makes multiple changes (mutations, deletions, insertions of amino acid building blocks) to the sequence. That is, there are many similar biological structures with very different sequences (less than 10% of amino acid building blocks identical in an alignment of two sequences), but it is very rare to find similar sequences that don't have similar structures (13).

2. Biological sequences can be aligned optimally using algorithms based on dynamic programming. These mathematically optimal alignments are most reliable when the similarity between sequences is high (>30% identical), but become less reliable in the range below 25% (14).

3. The physical interactions between molecules can be approximated with a set of energy equations that capture the basic physical attractions and repulsions between atoms. If carefully modeled, these energy equations are sufficient to model the dynamic behavior of molecules (15, 16).

4. Many (but not all) functions of proteins are encoded in linear segments of protein sequence ("sequence motifs") that can be recognized even between proteins that otherwise have very divergent sequences (17). Other functions are encoded in the three-dimensional arrangement of biophysical and biochemical properties ("3D motifs") that illustrate how different sequences can create similar microenvironments (18).

5. Probabilistic models of sequences, particularly a class of models known as Hidden Markov Models, can sensitively characterize a class of related sequences, and can recognize new members of the class (19).

6. Three-dimensional biological structures can be aligned using geometric algorithms that consider the distances between corresponding atoms. The proper correspondences between atoms can be determined using a combination of geometric and functional considerations (20-22).

7. Some elements of the three-dimensional structure of biological macromolecules can be predicted with 70% confidence from local sequence (continuous stretches of about 20 amino acids) alone (23, 24). The remainder of the information required to construct the 3D structure is contained in the sequence, but requires consideration of long range interactions (not within local stretches of amino acids).

8. The task of evaluating the compatibility of a new sequence to a known structure (can this sequence adopt this known structure?) is considerably easier than predicting the structure of a new sequence *ab initio* (what structure does this sequence adopt?) (25).

9. Variation (especially correlated variation) in both host and pathogen DNA genetic information can be used to understand which proteins are involved in disease, and how they interact (26).

The impact of bioinformatics has also been made with the development of a set of algorithms and databases that are routinely used, some of which were developed by bioinformatics researchers, and others of which have been adopted as important bioinformatics resources. To understand the capabilities of modern bioinformatics tools, let us consider a scenario: imagine a physician seeing a new patient who says that she was told she has "a genetic form of diabetes." The physician is not sure which syndrome the patient may have, and decides to go on the World Wide Web (WWW) to investigate the known genetic syndromes associated with diabetes. The physician could (today!) traverse the WWW to gather information in the following manner.

1. The physician might first go to the National Center for Biotechnology Information (NCBI) web pages.[5] The NCBI maintains a number of databases for biology and molecular medicine, which are integrated within the Entrez server (27).
2. The Online Mendelian Inheritance of Man (OMIM) resource contains a compilation of human genetic disorders, including automatic links to references in the literature and to the involved genes in the genetic databanks (28). A search for the word "diabetes" reveals multiple disorders, including "Diabetes Mellitus, Autosomal Dominant, Type II." The OMIM entry mentions that the primary defect is in the control of a gene encoding the protein glucokinase, and includes links to the database of protein sequences encoding the glucokinase gene.
3. A search for the term "glucokinase" in the protein sequence database, as part of NCBI's Entrez server yields a hit to the human glucokinase gene, including different versions that are created under different physiological conditions. This gene is linked to its underlying DNA sequence, protein sequence, as well as a set of references in the literature.
4. The link to the MEDLINE literature database is followed, in which the original article reporting the association of diabetes with a modification in this gene is provided.
5. The link to the protein sequence database is followed, and the detailed sequence of amino acids for this gene can be found. An algorithm can be run using this sequence to find all related sequences in the protein sequence databases, SWISS-PROT and PIR (29, 30). In addition, a library of motifs that have been associated with various functions is used to search for the occurrence of these motifs in glucokinase. It is found to bind the small molecules glucose and ATP, among others.

[5] http://www.ncbi.nlm.nih.gov/

6. The genetic databank, GENBANK, is then accessed from the protein sequence to see the detailed sequence of DNA bases that encode for the gene. The GENBANK record mentions that one region of the gene can have some inserted DNA that alters regulation of the gene.
7. The protein sequence entry is also linked to an entry in the Protein Data Bank, the database of three-dimensional structure. This entry provides the 3D coordinates of all the atoms in glucokinase as determined by x-ray crystallography, and provides a pictorial summary of the key structural features of the protein. By virtue of structural-similarity metrics, the database also provides links to other related protein structures from other organisms. Algorithms are also available to annotate the key functional sites within the protein, so that the precise location of glucose binding can be identified, as well as where the ATP molecule binds, and other sites of interest on the molecule.

At this point, the physician has used a number of databases to collect information about the disease, and its associated molecular components. Not routinely available today, however, are the obvious next set of tools, to allow the physician to 1) confirm the diagnosis by ordering genetic testing of the patient to sequence the relevant pieces of DNA, 2) select an appropriate treatment based on the manifestations of the disease and the results of the diagnostic genetic tests, 3) discuss with the patient the prognosis of this disease, in terms of anticipated course of untreated disease, of disease treated appropriately, as well as possibly prenatal counseling about the likelihood of transmission of this disease to children. Finally, the physician might ask the patient if she were interested in participating in a clinical study of a set of new treatments for this disease, which are based on an understanding of the genetics.

Current Challenges to Bioinformatics

The discussion in the previous section shows that there has developed a core set of "truths" in the field, which focus effort and provide a set of paradigms upon which individual investigators work. The information about diabetes and the glucokinase gene reveals at once the great array of tools and databases that have become available, but also highlights significant work that remains--in particular the integration of bioinformatics tools with clinical informatics tools concerned with the delivery of care. Together, these two sources of information promise to provide the information necessary for the development and delivery of new therapies, targeted to the particular genetic background of patients. The linkage of clinical medical information to molecular information represents one of the primary challenges for bioinformatics in the next century. As the genome sequencing projects mature and complete, we will have the genetic DNA sequences of both humans and a host of human pathogens, and informatics tools will be necessary to deliver this information in appropriate ways to medical decision makers. The information will impact decisions about diagnosis, prognosis, treatment and epidemiology.

Genomic information is already playing a leading role in the generation of new therapeutics for medicine. The ability to associate particular genes with particular organs, and the ability to associate defects in these genes with disease now allows drug targets to be identified primarily by computational analysis. Instead of the old paradigm of expensive, repetitive screening of candidate drug compounds against targets of interest, we can now imagine a scenario in which DNA sequence information is selectively collected in a patient (or group of related patients) to determine the set of proteins involved in a pathological process. These proteins are analyzed computationally to understand their functional properties and to look for places where their function can be augmented, diminished or modified (depending on the nature of the disease process). Other computational techniques are then used to design small compounds that interact with these proteins, based on principles of structural interactions. Finally, the compounds are analyzed and compared with known medications to assess possibilities for drug-drug interactions and toxicities. The small set of compounds that remain can finally be synthesized and tested in animals. The entire drug discovery process, under this (currently not possible) scenario is done computationally up to the point when the medication is actually synthesized and tested in animals. The expensive, large scale “shot-gun” screening of today is avoided, and can be automated with computational technologies.

The promise of integrating molecular biological information with the processes of delivering improved patient care and accelerating the discovery of useful new therapeutics requires significant progress in a number of areas, and these constitute the primary challenges to bioinformatics.

Simulations from Molecules to Populations

The first major challenge to bioinformatics is the creation of detailed physiological models at a molecular level. In order to understand disease processes and how molecular components contribute to them, it will be necessary to develop techniques for simulating the physiology within a cell. Currently, physiological modeling at the organ level has only been linked successfully with molecular simulations in a few areas. With the genome projects complete, we will have a complete catalog of the molecular players involved in physiology, and our task will be to accurately model the interactions between proteins, DNA, RNA, small molecules and their physiological aqueous milieu. We will depend upon both simulation from first principles of physics, as well as upon reasoning by analogy to other systems where the details have been worked out. The ability to model single cells will then provide the basic data necessary to perform simulations of entire organs and organisms. In turn, these simulations may provide parameter values necessary to simulate populations. These population models will be critical in understanding the spread of disease in a population, as well as the spread of genetic traits, and the spread of resistance to therapy.

The effective linking of biomedical data for "clinical genomics"

The ability to organize and interlink data sources from new biological and medical experiments will require the creation of new paradigms for database and knowledge base organization. The apparent success of the "large science" approach embodied in the genome sequencing projects has affected the way many biologists conceive of their work. There are more new technologies being developed for the large scale collection of data. These massive efforts at comprehensive data collection contrast to the traditional approach in which individual investigators work within a single, highly specialized experimental system trying to elucidate principles that apply more generally. At the same time, this new biological data is offering new ways to stratify patients in clinical trials, and pharmaceutical companies have already created an infrastructure to acquire high quality information about patients during the course of a clinical trial. These two data sources are extremely valuable independently, but may contain valuable information when properly combined. This becomes the challenge of what some have termed "clinical genomics" or the marriage of clinical investigation with genomic science. These data sources must be organized and linked in a manner that allows investigators to "mine" them for new knowledge. Enough biological databases have been designed that we are beginning to understand how to effectively organize these resource. Standard relational database technology is rarely the most effective way to capture the information collected, and new methods for organizing and distributing data are required..

Improved support for biomedical investigation in a data-intensive era.

Bioinformatics does not only study the flow of information from genes to organisms and populations, but also studies the ways in which biomedical investigators use information in their cycles of hypothesis generation and testing. The same data about which we are so excited threatens to confuse and frustrate investigators because of its volume. It therefore becomes critical to develop methods to assist investigators (both clinical and basic science) in the robust analysis of data. These methods include the development of useful paradigms to support biomedical collaboration at a distance. How should scientists interact with resources that store data and computational methods for manipulating this data? How should scientific results be published, made available for others, indexed and effectively communicated. For example, the graphical display of biomedical information can be used to summarize information effectively. Certain domains of biology have developed standard conventions for the display of data (31). Computer technologies are required to capture these conventions and use them for the automated creation of graphics to display the contents of databases or the results of new algorithms. Similarly, as the developers of biocomputing tools make them available to the scientific community, there is a danger that tools will be misapplied and data misinterpreted. There is therefore an urgent need for tools to assist in the diagnosis of problems that arise during data analysis and the creation of scientific models. Since most data analyses require multi-step processing (involving many algorithms with associated parameters), it is important to have tools that provide expert assistance in understanding problems that arise while analyzing data. Traditionally, small biological experimental systems have been simple enough to allow individual investigators to manually track the work of other groups, in order to identify problems with their underlying models and conflicts of the models with the supporting data. The mass of biological data, and the increasing specialization of investigators causes two problems. First, it is impossible to look at

complex computational models and identify their flaws without automated assistance in the form of systematic sensitivity analyses and expert systems for diagnosing errors. Second, it is impossible to be aware of all the possibly relevant work being done in allied fields. Computational methods are required to bring relevant information to the attention of investigators, so that rapid progress can be made, and redundant work can be minimized.

In addition to the incremental improvement of existing algorithms and data repositories, the development of simulation capabilities, the highly linked storage of data, and the development of tools to support the use of these capabilities represent the three major areas into which most current bioinformatics efforts can be classified.

Bioinformatics and clinical informatics

Why should clinical informatics investigators give any consideration to bioinformatics? The emphasis of clinical informatics on the delivery of health care is quite different from the emphasis of bioinformatics on support of basic biological research. There are good reasons, however, why investigators from these two "cousin" disciplines should follow progress in the other discipline. First, molecular information is rapidly impinging upon the traditional concerns of clinical informatics. It is very likely, for example, that sequence information will routinely be stored in the medical record in the next five to ten years. The lessons learned in the creation and maintenance of the biological genetic databases will be useful to those designing the electronic medical record of the future. Similarly, the genetic background of patients is likely to be a critical element of their past medical history, and may effect the way patients are stratified in clinical studies. Finally, the methods for tracking infectious disease and for determining optimal treatment are likely to be driven by considerations of genetic and structural features of the pathogen, which will directly use tools developed within the bioinformatics community. Already, databases have been created to track the HIV sequences of individual patients, and followed as they are exposed to different anti-viral drug regimens (26). As new antibiotics are designed against highly resistant bacteria, it is likely that the indications for their use will be based on molecular sequence and structural models.

The second reason that clinical informatics investigators should track progress in bioinformatics is that the two fields share very similar methodologies of applied computer science, including human-computer interactions, algorithms, probabilistic reasoning, diagnosis, database integration, knowledge representation, information retrieval, analysis of three-dimensional structural models, and studying the organization and structuring of the scientific literature. For every area of clinical informatics investigation, one can find an equivalent area within bioinformatics. As might be expected, however, the two fields have focused on these areas with different priorities, and so we find that certain areas of bioinformatics have matured more quickly that corresponding areas within clinical informatics, and vice-versa. In general, clinical informatics has focused more on the task of expert reasoning in the area of diagnosis and treatment. This is hardly surprising, given the daily tasks of physicians. Bioinformatics has only recently approached the task of emulating the diagnostic reasoning skills of experimental scientists as they design and debug experimental protocols and scientific models. In addition, clinical informatics has long recognized the need for standardized terminologies for clear communication (including, for example, SNOMED, ICD-9, CPT and many other vocabularies developed with particular medical activities in mind) (32-34), whereas the terminology in most molecular biology databases is quite nonstandard and idiosyncratic. The importance of facile and fluid interfaces to information resources is obvious in the setting of time-pressured physicians who will only use systems that fit into their daily routine, and has created a strong impetus for studying the user-interface issues associated with introducing technology into a work environment. Molecular biologists have been more patient with bioinformatics resources, and so the pressure for matching the tool to the task has, until recently, not been as strong within bioinformatics. With the exponential increase in the availability of biological data, however, there is more of a perceived time pressure, and the lessons from clinical informatics should be quite useful within bioinformatics.

At the same time, bioinformatics has made progress in other areas that have moved more slowly in clinical informatics. The general availability of biological data, and the culture of sharing data within biology has lead to a variety of novel database integration technologies which are difficult to develop in clinical informatics because of the need to respect patient privacy. There are several technologies for integrating combinations of relational, object oriented and flat-file databases that are in routine use (35). Bioinformatics has also

benefited from a closer relationship to the contributing parent endeavors of biology and computer science. For this reason, bioinformatics seems to have established a better defined set of core principles that form the basis for the discipline. It is not hard to identify the core data representations and algorithms that are critical for entering this field (as outlined above), and this provides an anchor which allows new investigators to position their work and convey the "well known" problems they are attacking, and what the previous measures of success should be. The heterogeneous nature of clinical informatics has rarely allowed a set of core principles to be elucidated (the use of Bayes' Rule in diagnostic reasoning may be one), and thus it is frustrating to identify the "core principles" of the field. Finally, validation of bioinformatics techniques is somewhat easier than in clinical informatics, where the interventions often involve patient care, where control experiments may be difficult or impossible to run. A bioinformatics-generated hypothesis can often be tested in the laboratory of an interested experimentalist, and the published data sets are large enough to allow routine division into "training" and "test" sets. The high stakes nature of clinical data can often make such validation experiments difficult. At the same time, there are many open areas within bioinformatics and clinical informatics where success in one domain is likely to carry directly over. In particular, efforts to standardize the formats for data exchange, and for declaring distributed computing components are likely to be similar within the two disciplines. The need for methods to analyze, store and retrieve bitmapped images (radiographic studies, electron micrographs, molecular structural ensembles, experimental assays, cell fields, karyotypes, etc...) are shared and could be combined. Finally, the organization of the biomedical literature is of great interest to both areas. The increased use of structured documents for reporting biomedical research is likely to create an infrastructure that will permit the creation of more useful tools for assisting in the cross-referencing of the biomedical literature (36, 37).

At some universities bioinformatics and clinical informatics students both train together. At Stanford, students within the Medical Informatics Training program can undertake dissertation projects in either discipline. The Stanford program in informatics was established primarily for clinical informatics in the early 1980's. In the last 7 years, however, we have increased our presence and focus on bioinformatics. Students in medical informatics are required to take basic courses in both clinical and bioinformatics, including decision making, image analysis and computational molecular biology. The rest of the curriculum remains essentially the same for computer science, probability, statistics, with a slight change in emphasis from human physiology to molecular biology in the "domain biology" course area. Students complete the program eligible to work in either clinical informatics or bioinformatics. Our goal is to prepare them to lead the next generation of medical information scientists who will be faced with the new challenges of providing information support within a clinical enterprise that has been revolutionized by developments in molecular biology.

Acknowledgments

RBA is supported by NIH LM-05652, LM-06422, NSF DBI-9600637 and a grant from IBM.

References

1. Langridge R. Interactive three-dimensional computer graphics in molecular biology. *Fed Proc* 1974;**33**(12):2332-5.
2. Smith TF. The history of the genetic sequence databases. *Genomics* 1990;**6**(4):701-7.
3. Bernstein FC, Koetzle TF, Williams GJ, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 1977;**112**(3):535-42.
4. Gusfield D. *Algorithms on Strings, Trees and Sequences*. New York: Cambridge University Press, 1997.
5. Setubal J, Meidanis J. *Introduction to Computational Molecular Biology*. Boston, MA: PWS Publishing Company, 1997.
6. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF. GenBank. *Nucleic Acids Res* 1998;**26**(1):1-7.
7. Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M, Sterk P. The EMBL nucleotide sequence database. *Nucleic Acids Res* 1998;**26**(1):8-15.
8. Green P. Human Genome Project: data quality [letter; comment]. *Science* 1998;**279**(5354):1115-6.
9. Cantor CR. How will the Human Genome Project improve our quality of life? *Nat Biotechnol* 1998;**16**(3):212-3.
10. Young KH. Yeast two-hybrid: so many interactions, (in) so little time.. *Biol Reprod* 1998;**58**(2):302-11.

11. Frederickson RM. Macromolecular matchmaking: advances in two-hybrid and related technologies. *Curr Opin Biotechnol* 1998;**9**(1):90-6.
12. Marshall A, Hodgson J. DNA chips: an array of possibilities. *Nat Biotechnol* 1998;**16**(1):27-31.
13. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 1980;**136**(3):225-70.
14. Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. *Structure* 1996;**4**(10):1123-7.
15. Elber R. Novel methods for molecular dynamics simulations. *Curr Opin Struct Biol* 1996;**6**(2):232-5.
16. Surles MC, Richardson JS, Richardson DC, Brooks FP, Jr. Sculpting proteins interactively: continual energy minimization embedded in a graphical modeling system. *Protein Sci* 1994;**3**(2):198-210.
17. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1997. *Nucleic Acids Res* 1997;**25**(1):217-21.
18. Wei L, Altman R. Recognizing Protein Binding Sites Using Statistical Descriptions of their 3D Environments. In: Altman R, Dunker, K, Klein, T, Hunter, L, ed. Pac Symp Biocomput. Singapore: World Scientific Publishing, 1998:497-508. vol 3).
19. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.
20. Holm L, Sander C. Mapping the protein universe. *Science* 1996;**273**(5275):595-603.
21. Gerstein M, Levitt M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Ismb* 1996;**4**:59-67.
22. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol* 1996;**266**:617-35.
23. Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. *J Mol Biol* 1997;**268**(1):31-6.
24. Rost B. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* 1996;**266**:525-39.
25. Smith TF, Lo Conte L, Bienkowska J, Gaitatzes C, Rogers RG, Jr., Lathrop R. Current limitations to protein threading approaches. *J Comput Biol* 1997;**4**(3):217-25.
26. Shafer RW, Winters MA, Palmer S, Merigan TC. Multiple concurrent reverse transcriptase and protease mutations and multidrug resistance of HIV-1 isolates from heavily treated patients [see comments]. *Ann Intern Med* 1998;**128**(11):906-11.
27. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 1996;**266**:141-62.
28. Pearson P, Francomano C, Foster P, Bocchini C, Li P, McKusick V. The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res* 1994;**22**(17):3470-3.
29. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res* 1998;**26**(1):38-42.
30. Barker WC, Garavelli JS, Haft DH, et al. The PIR-International Protein Sequence Database. *Nucleic Acids Res* 1998;**26**(1):27-32.
31. Felciano RM, Chen RO, Altman RB. RNA secondary structure as a reusable interface to biological information resources. *Gene* 1997;**190**(2):GC59-70.
32. Cote RA, Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *Jama* 1980;**243**(8):756-62.
33. Slee VN. The International Classification of Diseases: ninth revision (ICD-9) [editorial]. *Ann Intern Med* 1978;**88**(3):424-6.
34. Current procedural terminology (CPT). *Jama* 1970;**212**(5):873-4.
35. Markowitz VM, Ritter O. Characterizing heterogeneous molecular biology database systems. *J Comput Biol* 1995;**2**(4):547-56.
36. Altman RB, Abernethy NF, Chen RO. Standardized representations of the literature: combining diverse sources of ribosomal data. *Ismb* 1997;**5**:15-24.
37. Chen RO, Felciano R, Altman RB. RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Ismb* 1997;**5**:84-7.