Viewpoint ■

The Interactions Between Clinical Informatics and Bioinformatics:

A Case Study

RUSS B. ALTMAN, MD, PHD

Abstract For the past decade, Stanford Medical Informatics has combined clinical informatics and bioinformatics research and training in an explicit way. The interest in applying informatics techniques to both clinical problems and problems in basic science can be traced to the Dendral project in the 1960s. Having bioinformatics and clinical informatics in the same academic unit is still somewhat unusual and can lead to clashes of clinical and basic science cultures. Nevertheless, the benefits of this organization have recently become clear, as the landscape of academic medicine in the next decades has begun to emerge. The author provides examples of technology transfer between clinical informatics and bioinformatics that illustrate how they complement each other.

■ **J Am Med Inform Assoc.** 2000;7:439–443.

The Stanford Medical Informatics (SMI) laboratory was created in the mid-1980s as part of a reorganized Stanford Knowledge Systems Laboratory (KSL). The KSL was a federation of investigators interested in applying artificial intelligence methods to real-world problems in engineering, science, and medicine. The core competency of the SMI centered on the research interests of Dr. Edward Shortliffe, who had well-established research programs in the medical applications of expert systems (such as the MYCIN system[1]) and the creation of an electronic infrastructure to support such systems. The initial focus of the laboratory was clearly on the clinical applications of artificial intelligence.

The KSL itself was an outgrowth of the Stanford Heuristic Programming Project, which was formed around the DENDRAL program for interpreting mass spectroscopy data using a production rule formalism.[2] The excitement about production rule systems led to the application and refinement of these ideas into rule-based expert systems and to MYCIN, a system for diagnosing infectious diseases and recommending treatment.[1] At the same time, other projects arose at the Heuristic Programming Project that focused more on the support of basic science. The MOLGEN I project studied planning techniques for molecular biological experiments,[3,4] the MOLGEN II project modeled the process of scientific discovery in molecular biology,[5,6] and the PROTEAN project looked at interpreting NMR experimental information about three-dimensional structures using constraint satisfaction techniques.[7]

These projects did not reside explicitly in the SMI laboratory at the time of its formation, but they were staffed by close collaborators and students in the KSL, so the opportunities in basic science were always clear. Most important, there was an early recognition that the basic methodological approaches taken in the two disciplines could often be transferred to produce novel contributions.

By the late 1980s, it became clear that there was an impending explosion of data that would become available in molecular biology and that the SMI was in a good position to apply, to basic biology, the research strategies it had been using in clinical medicine. The National Library of Medicine encouraged

Affiliation of the author: Stanford University, Stanford, California.

Correspondence and reprints: Russ B. Altman, MD, PhD, Stanford Medical Informatics, Stanford University, 251 Campus Drive, MSOB X-215, Stanford, CA 94305-5479;
e-mail: ⟨russ.altman@stanford.edu⟩.

This work was supported in part by NIH grants LM-05652, GM-61374, and LM-06422 from the National Library of Medicine; by grant DBI-9600637 from the National Science Foundation; and by the Burroughs Wellcome Foundation and Sun Microsystems.

Received for publication: 5/11/00; accepted for publication: 5/31/00.

Figure 1 The Stanford Medical Informatics logo stresses the combination of clinical medicine, basic biology, and computer science. Variations are found at http://smi-web.stanford.edu/logos/index.html.

Stanford (and other institutions at which it funded training programs) to seriously consider expanding the training mission to include the application of computational technologies to basic biology. Starting in 1992, therefore, faculty in what eventually came to be called "bioinformatics" were hired at SMI, and students with interests in this area were recruited.

The philosophy behind the addition of bioinformatics to the SMI research and training mix was not that a separate track should be introduced but that the existing training and research experience in the SMI and KSL could be adapted in an evolutionary manner. In the case of training, the five elements of the medical informatics training program (core informatics, domain biology, computer science, probability/statistics/decision science, and ethical/legal/social issues) were very well suited to training in both bioinformatics and clinical informatics, so the training program expanded quite naturally.[8] Some projects focused on the medical domain, others on the biological domain. The problems in the two domains are clearly different, but the methodologies that are used to approach them are shared and form the basis of a cohesive program.

Thus, the SMI houses both bioinformatics and clinical informatics efforts at Stanford.[1] The new laboratory logo (Figure 1) combines the traditional medical caduceus with a DNA double helix to stress the interactions between the two fields. There are also active programs in bioinformatics research in the Departments of Biochemistry, Structural Biology, Pathology, Genetics, Mathematics, and Computer Science, and work in clinical informatics is being done in the Departments of Medicine, Computer Science, Anesthesiology, and Pathology. The focus on methodology, and the management of the informatics training program (which sends students to many of the other departments to work on projects), are what make the SMI a focal point for efforts in these areas.

The pursuit of bioinformatics and clinical informatics together is not without some difficulties. Practitioners in clinical medicine and basic science do not instantly understand the distinction between the scientific goals of their domains and the transferability of methodologies across the two domains. They sometimes question whether informatics investigators are really devoted to the solution of scientific problems or are simply enamored of computational methodologies of unclear significance. It is therefore imperative that informatics investigators (and their students) be able to work collaboratively with physicians and scientists in a manner that makes it clear that the creation of excellent, well-validated methods for solving problems in these domains is the paramount goal. Usually, the particular skills of informatics students and the quality of their contributions to the research efforts they join significantly allay these concerns. Students in informatics have a much richer understanding of computer science, statistics, and information technology than do students in collaborating disciplines, and they think more like engineers in creating solutions.

Technology Transfer Between Subdisciplines

One of the most compelling reasons for housing clinical informatics and bioinformatics under the same roof is the opportunities this provides for rapid technology transfer between the two subdisciplines. Because there are still barriers between the two fields (bioinformatics professionals may not even know the clinical informaticians at their institutions), there are also opportunities for accelerating progress in both fields by knowing the problems and literature in each. Even in the days of the Heuristic Programming Project, ideas from the DENDRAL project on mass spectroscopy were applied to the MYCIN project on diagnosing disease. Success with MYCIN inspired some to pursue planning and discovery efforts in the MOLGEN framework. Two particularly illustrative areas of technology transfer that have had significant impact are the development of Bayesian methods for reasoning and the development of frameworks for knowledge representation and acquisition.

Probabilistic Methods in Clinical Informatics Applied to Biological Structure

In the mid 1980s, the SMI experienced a "probabilistic revolution," during which many projects in medical diagnosis began to focus on the use of probability theory as an alternative to other uncertainty calculi that had been proposed. (As an aside, the interest in these methods stemmed, in large part, from student exposure to courses in probabilistic decision making, which created a cadre of converts to this mode of thinking.) In particular, a great interest developed in the use of Bayesian belief networks as an organizing paradigm for understanding diagnosis (compute the most likely diagnosis), treatment (compute the highest utility), knowledge acquisition (acquire the best numbers), and other elements of medical reasoning. The work that resulted made an impact on medical informatics and launched a number of successful careers in the application of these technologies in medicine and beyond.[9–12]

The furor over probabilistic methods in clinical informatics spilled over in many ways to the work that was going on in bioinformatics. First, the protean project (which focused on computing three-dimensional molecular structures from sparse and noisy data) moved to adopt a probabilistic representation of structure in which an ill-defined "prior" structure was updated with uncertain data to compute the most likely a posteriori structural estimate. The impetus for this change in representation (from an initial formulation as a discrete constraint satisfaction or combinatorial optimization problem[13]) can be traced directly to the success of probabilistic, Bayesian methods in the clinical side of the laboratory.[14,15]

Second, in a more direct application of the clinical informatics experience, Bayesian methods were applied to problems in sequence analysis. In particular, dependencies between positions in biological sequences were modeled as dependencies in a Bayesian net, and good performance was demonstrated in the ability of such models to characterize and recognize patterns in biological sequence.[16] Again, abandoning the traditional assumption of the independence of columns in a multiple alignment could be traced directly to the work in clinical diagnosis. Essentially, this work was diagnosing an alpha helix based on probabilistic evidence.

Knowledge Representation and Structured Knowledge Acquisition

A second area in which methodologies from clinical informatics have inspired efforts and contributions in bioinformatics is that of knowledge representation. The experiments in supporting structured representations of clinical medicine—in particular, the representation of clinical treatment protocols (in the Oncocin project[17])—led the clinical informatics researchers at the SMI to focus attention on the creation of general-purpose frame representation systems that allowed strict ontological modeling of domains, and used the resulting ontologies to automatically create templates for the structured acquisition of knowledge. The PROTÉGÉ project[18] continues to be a source of software for structured data modeling and knowledge acquisition. PROTÉGÉ has recently been adopted by the World Wide Web Consortium (http://www.w3.org/RDF/) as a recommended platform for authoring RDF documents, and meetings of PROTÉGÉ user groups (including health care organizations) interested in robust methods for data modeling are held annually.

Around 1995, an SMI bioinformatics effort to create three-dimensional structural models from combinations of experimental data (as reported in the literature) was focusing on the structure of the bacterial ribosome, a critical cellular molecular ensemble.[19,20] The algorithms used to create the three-dimensional models worked well, but it was a struggle to collect all the relevant data from the literature.

Aware of the work on ontological modeling on the clinical side, we proposed the creation of an ontology of biological objects as well as the experiments that supply the data used in modeling. As a test bed, we proposed to gather the published literature about the ribosome (having identified approximately 200 papers that contain reports of primary data) and create a set of structured representations for the data. The frame-based system that was chosen for RiboWEB and the organization of the system were closely related to those of the PROTÉGÉ system. In addition, a long history of creation and maintenance of controlled terminologies in clinical medicine helped inform the creation of a controlled vocabulary for describing structural biological experiments.[19]

The RiboWEB system now serves as a test bed for the use of structured representations of biological data. RiboWEB makes it possible to compare a new piece of data with all the relevant data reported previously,[21] it allows inconsistencies in the literature to be identified,[22] and it allows three-dimensional models to be built on the basis of a selection and interpretation of a subset of the available information.[20] Once again, there was a technology transfer of informatics techniques from the clinical to the biological domain.

Blurring the Boundary Between Clinical Informatics and Bioinformatics

The success of the RiboWEB project was responsible, in part, for the newest efforts in the SMI to create a pharmacogenetics knowledge base (PharmGKB, available at http://pharmgkb.stanford.edu/) that combines a diverse array of data, from genomic data to cellular/molecular phenotype and clinical phenotype. The data modeling and knowledge base infrastructure developed at SMI form the basis for this project, which aims to provide scientists nationwide with integrated access to data over the Web. Data will be acquired in structured form and will be distributed using the same kinds of structured representations that were developed for RiboWEB.

The PharmGKB project is the first project at SMI that blurs the distinction between bioinformatics and clinical informatics. The knowledge base will hold genomic sequences, multiple alignments, and structural information, which is the stuff of bioinformatics. At the same time, it will store clinical patient records of diseases, medications, side effects, and laboratory results, which are the traditional material of clinical informatics. Similarly, the analytic capabilities will be based on algorithms introduced in both fields. Clearly, for projects such as this, there may not be a useful distinction between clinical informatics and bioinformatics, and in many ways the strategy of combining the training and research environments for these two subjects is vindicated.

As the biological world enters the post-genome age, the interplay between basic biological data (sequences, structures, pathways, and genetic networks) and clinical information systems is, clearly, critical. Genes and structures are useful only in the context of the functions and phenotypes that they produce, and so we look toward continuing interaction and, perhaps, unification of the two fields.

One of the ultimate goals of both bioinformatics and clinical informatics is to have robust computational models of physiology that will enable us to model, store, retrieve, and analyze the effects, on patients, of disease, medications, and the environment. Bioinformatics approaches these models from the bottom up, while clinical informatics approaches them from the top down. The other critical technologies of clinical informatics, including knowledge representation, data mining, automated diagnosis, and information retrieval, can be viewed as technologies supporting this goal.

The training program curriculum at SMI has recently been updated and generalized to be appropriate for both clinical informatics and bioinformatics. This process was surprisingly simple. We now require 1) instead of medical physiology, any biology or physiology of interest to the student; 2) as before, significant coursework in probability, statistics, or decision analysis, or a combination of these; 3) a substantial amount of core computer science; and 4) nontechnical courses in the ethical, social, legal, or business aspects of the field. We have also generalized our introductory core informatics courses to introduce the principles of data representation and algorithms in biomedicine that make it a challenging field.

The resulting generalized curriculum has been tested for both bioinformatics and clinical informatics students and appears to strike an appropriate balance between domain knowledge and methodologic knowledge. Recognizing the success of this generalized curriculum, we have changed the name of the training program from Medical Information Sciences (the name of the degree program as it was defined in 1982) to Biomedical Informatics.

The Need for Informatics to Co-exist with Other Areas of Biocomputation

As we look to the future of informatics, it becomes clear that informatics is one aspect of a larger area of endeavor, which can be loosely called "biomedical computation." At a recent Stanford faculty retreat, we found that most biomedical computation efforts fall into one of six affinity groups. (The summary abstract book is available in the archives, at http://bits.stanford.edu/.) These groups are divided roughly into those whose investigators think mostly (but not exclusively) about computing directly with and about physical systems and those whose investigators think mostly about information acquisition, storage, retrieval, and management. Both use significant computation skills and require a strong understanding of computer science.

The affinity groups are:

- Image acquisition and analysis (physical systems)
- Structural biology and genetics bioinformatics (physical systems)
- Biomechanical modeling for macroscopic systems (physical systems)
- Computer-assisted interventions and robotics (physical systems)
- Data modeling, statistics, and informatics (informatics)
- Networked and computer-enabled education (informatics)

There was a broad consensus at this retreat that an umbrella organization for these affinity groups makes sense, to support joint teaching, research, and shared infrastructure and to provide a focus for the application of advanced computational techniques to problems in biomedicine. The creation of such an organization is a work in progress that promises to solidify and extend Stanford's commitment to these areas.

References ■

1. Buchanan B, Shortliffe E (eds). Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming. Menlo Park, Calif: Addison-Wesley, 1984.
2. Lindsay R, Buchanan B, Feigenbaum E, Lederberg J. Applications of Artificial Intelligence for Organic Chemistry. New York: McGraw-Hill, 1980.
3. Stefik M. Planning and meta-planning (MOLGEN: Part 2). Artif Intell. 1981;16(2):141–69.
4. Stefik M. Planning with constraints (MOLGEN: Part 1). Artif Intell. 1981;16(2):111–40.
5. Karp P. Design methods for scientific hypothesis formation and their application to molecular biology. Machine Learning. 1993;12:89–116.
6. Friedland P. Knowledge-based Experiment Design in Molecular Genetics. Palo Alto, Calif: Stanford University, 1979. Report CSD-79-771.
7. Duncan B, Buchanan B, Hayes-Roth B, et al. PROTEAN: A new method for deriving solution structures of proteins. Bull Magn Res. 1987;8(3/4):111–9.
8. Altman R. A curriculum for bioinformatics: the time is ripe. Bioinformatics. 1998;14(8):549–50.
9. Shwe MA, Middleton B, Heckerman DE, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base, part I: the probabilistic model and inference algorithms. Methods Inf Med. 1991;30(4):241–55.
10. Middleton B, Shwe MA, Heckerman DE, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR, part II: evaluation of diagnostic performance. Methods Inf Med. 1991;30(4):256–67.
11. Herskovits EH, Cooper GF. Algorithms for Bayesian belief-network precomputation. Methods Inf Med. 1991;30(2):81–9.
12. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems, part I: the Pathfinder project. Methods Inf Med. 1992;31(2):90–105.
13. Brinkley JF, Altman RB, Duncan BS, Buchanan BG, Jardetzky O. Heuristic refinement method for the derivation of protein solution. J Chem Inf Comput Sci. 1988;28(4):194–210.
14. Altman RB. Probabilistic structure calculations: a three-dimensional tRNA structure. Ismb. 1993;1:12–20.
15. Altman R, Jardetzky O. New strategies for the determination of macromolecular structure in solution. J Biochem. 1986;100(6):1403–23.
16. Klingler TM, Brutlag DL. Discovering structural correlations in alpha-helices. Protein Sci. 1994;3(10):1847–57.
17. Hickam DH, Shortliffe EH, Bischoff MB, Scott AC, Jacobs CD. The treatment advice of a computer-based cancer chemotherapy protocol. Ann Intern Med. 1985;103(6,pt 1):928–36.
18. Tu S, Eriksson H, Gennari J, Shahar Y, Musen M. Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTÉGÉ-II to protocol-based decision support. Artif Intell Med. 1995;7:257–89.
19. Altman R, Bada M, Chai X, Carillo M, Chen R, Abernethy N. RiboWEB: an ontology-based system for collaborative molecular biology. IEEE Intell Systems Applications. 1999; 14(5):68–76.
20. Chen RO, Felciano R, Altman RB. RiboWEB: linking structural computations to a knowledge base of published experimental data. Ismb. 1997;5:84–7.
21. Altman RB, Abernethy NF, Chen RO. Standardized representations of the literature: combining diverse sources of ribosomal data. Ismb. 1997;5:15–24.
22. Chen RO, Altman RB. Automated diagnosis of data-model conflicts using metadata. J Am Med Inform Assoc. 1999;6(5): 374–92.