Focus on JAMIA Intersections with Bioinformatics

*Viewpoint* ■

# Opportunities at the Intersection of Bioinformatics and Health Informatics:

## A Case Study

PERRY L. MILLER, MD, PHD

**Abstract** This paper provides a "viewpoint discussion" based on a presentation made to the 2000 Symposium of the American College of Medical Informatics. It discusses potential opportunities for researchers in health informatics to become involved in the rapidly growing field of bioinformatics, using the activities of the Yale Center for Medical Informatics as a case study. One set of opportunities occurs where bioinformatics research itself intersects with the clinical world. Examples include the correlations between individual genetic variation with clinical risk factors, disease presentation, and differential response to treatment; and the implications of including genetic test results in the patient record, which raises clinical decision support issues as well as legal and ethical issues. A second set of opportunities occurs where bioinformatics research can benefit from the technologic expertise and approaches that informaticians have used extensively in the clinical arena. Examples include database organization and knowledge representation, data mining, and modeling and simulation. Microarray technology is discussed as a specific potential area for collaboration. Related questions concern how best to establish collaborations with bioscientists so that the interests and needs of both sets of researchers can be met in a synergistic fashion, and the most appropriate home for bioinformatics in an academic medical center.

■ **J Am Med Inform Assoc.** 2000;7:431–438.

This paper provides a "viewpoint discussion" based on a presentation made to the American College of Medical Informatics (ACMI) Symposium, which was held in February 2000 at San Marco Island, Florida. The activities of one day of that three-day symposium centered on the intersection of bioinformatics[1] and health informatics. The implicit theme was to help identify opportunities for researchers in health informatics to become involved in the rapidly growing field of bioinformatics. In general, there are two types of such opportunities. One set of opportunities occurs where bioinformatics research itself intersects with the clinical world. The second set of opportunities occurs where bioinformatics research can benefit from the technologic expertise, techniques, and approaches that health informaticians have used extensively in the clinical arena.

Affiliation of the author: Yale University, New Haven, Connecticut.

This work was supported in part by NIH grant G08-LM05583 from the National Library of Medicine.

Correspondence and reprints: Perry L. Miller, MD, PhD, Center for Medical Informatics, Yale University School of Medicine, P.O. Box 208009, New Haven, CT 06520-8009;
e-mail: perry.miller@yale.edu.

Received for publication: 2/16/00; accepted for publication: 5/4/00.

Several related questions were also addressed. One question concerned how best to establish collaborations with bioscientists so that the interests and needs of both sets of researchers can be met in a synergistic fashion. A second question concerned the appropriate home for bioinformatics in an academic medical center.

Since these questions are very broad and complex, they can be difficult to discuss in abstract terms. This paper uses the experience of the Yale Center for Medical Informatics (YCMI) as a case study to structure a discussion of these issues.

## What is Bioinformatics?

The term "bioinformatics" has been used with different scopes and meanings by different groups of researchers. The term could refer to a range of activities:

- *Informatics involving genomics*. A number of researchers in the field of genomics have used the term "bioinformatics" to refer to the applications of informatics within their discipline. In the early 1990s, this work tended to focus on chromosome mapping and sequencing. Now that a number of smaller genomes have been fully sequenced and many genes in the human genome have been at least partially identified, genomic informatics has expanded into exploring the function of these genes, giving rise to fields such as functional genomics and structural genomics. As these trends continue, the distinction between genomic informatics and informatics in support of bioscience as a whole will become less distinct.
- *Informatics involving the biosciences*. Beyond genomics, computers are being used in a wide range of ways to support the biosciences. For example, the national Human Brain Project has coined the term "neuroinformatics" to describe the storage, analysis, and integration of experimental neuroscience data at many levels of bioscience research. If one defines "bioinformatics" as involving the biosciences as a whole, then one question concerns its relationship to the field of computational biology, whose scope is also evolving.
- *Informatics involving bioscience and clinical research*. Since genomic data will increasingly become the subject of a wide range of clinical research, one could define bioinformatics to include this work as well.
- *All biomedical and health informatics*. At the most general level, "bioinformatics" might be defined to include all medical and health informatics in addition to the biosciences. There will certainly be an increasing number of intersections between work in these areas.

In this paper, "bioinformatics" is loosely defined to include the first two areas discussed above—informatics involving the biosciences, including genomics.

## The Spectrum of Biomedical Informatics at the Yale Center

Table 1 outlines the spectrum of biomedical informatics activities at the YCMI. The three areas of research are increasingly likely to intersect in the near future.

### Genomic Informatics

Over the past decade, the YCMI has been involved in a number of projects in support of genomics and genetics. An early project explored the use of parallel computation in biological sequence analysis and genetic linkage analysis, in collaboration with faculty in the Department of Computer Science.[2] Another project provided Internet-based informatics support for the collaborative Genome Center, involving the Albert Einstein College of Medicine and Yale, to map human chromosome 12.[3]

*Table 1* ■

The Spectrum of Biomedical Informatics at the Yale Center for Medical Informatics

| Areas of Research | Purpose |
|---|---|
| Genomic informatics:<br>Yeast Genome Analysis Center<br>Human genome diversity<br>Microarray projects | Working up from the genetic blueprint |
| Neuroinformatics:<br>Informatics in support of olfactory research, including molecular modeling and neuronal modeling | Storing, integrating, and modeling experimental data at many biological levels in the most complex organ system |
| Clinical informatics:<br>Informatics support for clinical research<br>Computer-based clinical decision support<br>Network-based clinical information access<br>Electronic patient record system research and development | Working with the fuzziness of clinical data and disease |

Current YCMI activities in genomic informatics focus on three areas. One long-standing collaboration, with the laboratory of Prof. Kenneth Kidd (Genetics),[4] centers on human genome diversity. A second major collaboration is with Prof. Michael Snyder, director of Yale's Yeast Genome Analysis Center.[5] A recent and rapidly growing set of collaborations involve the support of microarray technology, as discussed further below. All these collaborations have involved the development of databases and informatics tools for internal use at Yale and also for providing public access to the data via the Web.

With regard to integrating genomic informatics with clinical practice, now that many genes and gene fragments have been identified, there are tremendous opportunities to work up from the genetic blueprint to explore gene expression, functional genomics, and structural genomics on a massive scale, including their implications for human disease.

### Clinical Informatics

Clinical informatics and genomic informatics are at far ends of the spectrum shown in Table 1. Here the field of medical and health informatics has long been confronting the development of informatics techniques that deal with the fuzziness of clinical data and disease. At the YCMI, we have been working on many different projects:

- Informatics support for clinical research. One rapidly growing YCMI project involves the development and use of Trial/DB,[6] a Web-accessible database designed to collect data for clinical trials and clinical research. In addition to its use at Yale, Trial/DB is being supported by the National Cancer Institute to serve as the "special studies database" for its national Cancer Genetics Network (CGN). The CGN by definition focuses specifically on clinical studies that have a genetic component.
- Computer-based clinical decision support. Faculty at the YCMI have a long-standing research interest in computer-based clinical decision support, including implementing clinical practice guidelines and providing access to network-based reference information in the context of care for particular patients. We are also working with our medical center to move incrementally toward the electronic patient record. As more and more genetic tests become available, test results will become an important part of the patient record, and there will be many opportunities to use this information to provide clinical decision support.

Thus, we anticipate increasing interplay between the genomic and clinical levels, which will have a major impact on our informatics activities.

### Neuroinformatics

In Table 1, neuroinformatics sits between genomic informatics and clinical informatics and is really a placeholder for informatics activities involving many different tissues and organ systems. Once the genetic blueprint is known, the next step is to determine what it means in a range of different tissues and organ systems, including the kidney, the liver, the gastrointestinal tract, the heart, and the endocrine system, among others. Neuroinformatics focuses on the central nervous system, which is clearly the most complex organ system.

The goal of the field of neuroinformatics is to provide enabling informatics technology that supports neuroscience research at many different levels[7,8] (Table 2). These include the genetic level, the cellular level, the physiologic and pharmacologic levels, and, eventually, the level of behavioral research. The research includes experimental microanatomic studies (e.g., imaging cells) and macroanatomic studies (e.g., imaging brains). Each type of experiment uses different experimental techniques and generates different types of experimental data. Historically, different laboratories have tended to focus on one or two of these levels.

At each of these levels, large amounts of experimental data are being generated in a form that can be stored and analyzed online. To fully understand a neuroscience phenomenon, however, it is ultimately important to gather data at many different levels and analyze all those data in an integrated fashion.

In summary, neuroinformatics in a sense "connects" genomic informatics and clinical informatics in the areas of neuroscience research. As such, it is representative of many other bioscience disciplines that are addressing similar issues in other tissues and organ systems. The three levels shown in Table 1 represent a spectrum of informatics activities that will become increasingly integrated in many different ways.

## Bioinformatics and Health Informatics: Selected Areas of Intersection

Health informaticians have many potential opportunities to become involved in collaborations involving bioinformatics. One set of opportunities occurs where bioinformatics research intersects the clinical world:

- *Clinical correlation of genetic variation*. Genetic variation might be caused by different mutations of a

*Table 2* ■

Examples of Experimental Neuroscience Data at Different Levels of Brain Function

| Levels of Brain Function | Types of Data |
|---|---|
| Behavior | Performance quantification, video monitoring, drug testing |
| Distributed systems | 2-D and 3-D axon tracing between regions, electrophysiologic recordings (spike timing), brain imaging, and 3-D brain maps |
| Specific regions | 2-D and 3-D cytoarchitectonics of layers and functional columns, transmitter-receptor localization, anatomic, physiologic, and metabolic maps |
| Nerve cells | 3-D cell morphology, 3-D functional imaging, electrophysiologic recordings of action-potential firing patterns and membrane currents |
| Neuronal components | 3-D imaging of axon terminals, growth cones, dendrites, dendritic spines, 3-D localization of organelles and synaptic microcircuits |
| Microcircuits | 3-D fine structure and imaging of synaptic patterns, synaptic pharmacology, action-potential firing patterns and synaptic currents, and potentials |
| Organelles | 2-D and 3-D fine structure and molecular composition of synapses, mitochondria, microtubules, etc.; recordings of synaptic currents and potentials |
| Molecules | 3-D molecular models of receptors, channels, enzymes and structural proteins, molecular physiology, and pharmacology of transmitters, modulators, hormones, guidance molecules, growth factors and gene-transcription factors |
| Genes | DNA and protein sequences |

SOURCE: Shepherd et al.[9] Used with permission.

single gene or by mutations of different genes that are related, for example, because they code for different enzymes in a single metabolic pathway.

- Genetic variation may be correlated with different levels of severity of a disease or different presentations of signs and symptoms.[10]
- Patients with different genetic makeups may have different responses to treatment. The new field of pharmacogenetics is exploring the possibility of tailoring treatment of disease to a patient's underlying genetic makeup.
- A patient's genetic makeup may make the patient more susceptible, or relatively resistant, to risk factors associated with a disease.
- A patient's prognosis might differ depending on underlying genetic factors.

■ *Comparison of gene expression in normal and disease states*. Microarray technology is a potentially productive area for collaborations, as discussed in a later section. This technology will be used both in the biosciences and in clinically oriented projects.

■ *Genetic test results as part of the patient record*. We have already mentioned the potential inclusion of genetic test results in a patient medical record and their use in computer-based clinical decision support. Legal and ethical issues that arise from inclusion of this information in the electronic patient record will also need to be addressed.

A second set of opportunities occurs where bioinformatics research can benefit from techniques and approaches that informaticians have used extensively in the clinical arena:

■ *Data mining*. As large and diverse databases of biological data are developed, there will be opportunities to explore many different approaches to data mining, to understand the complex interactions and implications of the data.

■ *Database organization and knowledge representation*. There will also be many opportunities to explore research issues in database design and interoperability; in data querying; in representing knowledge derived from data, which guides the analysis of the data; and in inferencing based on the knowledge. The creation, use, and maintenance of standardized biomedical vocabularies will also be needed. Such vocabularies will include not only standardized sets of terms but also standardized sets of relationships between those terms and standardized sets of attributes describing those terms.

- *Computer modeling of normal and disease processes at many levels*. Modeling is already being used at many different levels to understand biological phenomena. As more and more data become available, there will be opportunities to create computer models, of many different types, that are closely tied to the data. Ideally, experimental data should refine a model, and analysis using the model should suggest further laboratory experiments, in an iterative, cyclic fashion.

## Storing and Analyzing Microarray Data: A Case Study

The microarray, a recently developed technology, offers a wide range of informatics opportunities.[11–14] Yale is currently installing two microarray analyzers, one in the School of Medicine and one on Yale's main campus. These use a technology whereby "DNA chips" measure whether and to what degree different genes are expressed in experimental tissue samples. Each DNA chip can analyze the presence or absence and the approximate level of expression of tens of thousands of genes.

For example, one group of Yale researchers will use this technology to study hematologic disease involving white blood cells (WBCs). They will take WBCs at different stages of cell differentiation and in a single experiment see which of roughly 10,000 genes are expressed in two samples that have been combined (e.g., a normal sample and a cancerous sample at the same stage of differentiation). The test for each gene (really a small DNA fragment that is part of a gene) is seen on the microarray as a single spot in a massive array of spots. The two samples (from normal and abnormal WBCs) will be tagged with fluorescent markers of different colors (red and green), so that each gene can be tested in both samples in a single microarray experiment.

Each experiment will generate 10,000 data points. Each point will have several associated values, reflecting the actual intensity and relative intensity of both fluorescent markers at each of the 10,000 spots. The researchers estimate they will ultimately perform up to five such experiments a day.

In addition, other laboratories will be using the same machine to generate similar numbers of data for many different experiments. It takes roughly 10 minutes to perform each microarray analysis. Slides can be loaded to allow the machine to perform analyses automatically 24 hours a day.

As a result, we see microarray experiments as an exciting opportunity to expand our activities in genomic informatics:

- Huge amounts of data will be generated. In the near future, microarrays will probably be able to analyze all 60,000 to 100,000 human genes in a single experiment.
- These experiments will have major needs for bioinformatics, far beyond the need previously experienced by our bioscience collaborators.
- There are extensive opportunities to explore many different approaches to data mining and analysis.
- There is also a need for people who understand database design issues. For example, as an increasing number of different microarray experiments are performed, it will useful if entity-attribute-value (EAV) technology can be used so that new database tables do not need to be programmed for each experiment.
- The data need to be robustly and flexibly linked to many external databases and software analytic tools, so that their meaning and implications can be fully explored.
- Supercomputer capabilities, including parallel computation, are clearly needed for the performance of all the required analyses.
- Microarray research will be performed both by bioscientists and by clinically oriented researchers.

As a result, a broad, stable infrastructure of informatics staff and faculty will be required to support the high volume of microarray experiments that will soon be performed.

## Establishing Collaboration with Bioscientists: Informatics Support vs. Informatics Research

If health informaticians are to become centrally involved in bioinformatics, they need to establish robust collaborations with bioscientists. In the forging of such collaborations, it is important to understand that informaticians can play two general types of informatics roles—specifically, providing informatics support for bioscience research and performing informatics research that uses the bioscience domain as a context for addressing basic informatics research issues.

In this regard, it is important to point out that bioscientists typically have motivations that are very different from those of the clinician collaborators with whom many health informaticians have worked in the

past. Clinician collaborators are, typically, primarily involved in clinical practice and are looking for additional interesting research projects in which they can become involved. For such collaborators, embarking on a clinical informatics research project allows them to provide their clinical expertise and participate in sophisticated research that relates to their field. If this research results in additional visibility and publications, they have reason to be happy.

Bioscientist collaborators may have a very different motivation. They are already doing research. They are looking to informaticians to provide tools to help them perform their research more effectively. They are, typically, not looking for additional areas for research peripheral to their field, nor do they want to devote their time to such projects. They have more than enough research in their own field to keep them busy, and their time is precious. As a result, they will not be satisfied with the clinical informatics model of serving as domain experts in informatics research projects, even if the projects are in their field. They want help solving their immediate research problems.

As a result, bioscientists typically want informaticians to provide them with informatics support. Conversely, academic informaticians want at some level to be performing informatics research, although they are certainly willing to provide informatics support if this leads to interesting research opportunities.

The YCMI's neuroinformatics experience in the national Human Brain Project (HBP) provides, as a case study, a chance to discuss these issues more concretely. The YCMI's HBP work involves the integration of multidisciplinary sensory data, using the olfactory system as a model system. This HBP work involves both neuroinformatics support and neuroinformatics research.

### Neuroinformatics Support

Our neuroinformatics support activities involve building a variety of databases and tools. In general, we have attempted to build databases that can serve the needs of our collaborating laboratories and also serve as pilot resources for the field as a whole. These databases include ORDB, containing information about olfactory receptors[15]; OdorDB, containing information about odor molecules; NeuronDB, containing information about neuronal cell properties[16]; and ModelDB, containing neuronal models that can be searched, examined, downloaded via the Web, and run locally.

The development of these databases involved a great deal of practical work with our collaborators to fine-tune their design, functionality, and interface so that they can be readily useful to, and usable by, neuroscientists. It is important to emphasize, however, that "just" performing good neuroinformatics support requires that informatics faculty work closely with the neuroscientists to understand the biological problems, to appreciate the needs of the neuroscience researchers, and to develop well-structured solutions to enable neuroscience research.

### Neuroinformatics Research

In developing these tools, we have been able to define interesting neuroinformatics research projects. As discussed in more detail below, however, this did not happen immediately. Our current neuroinformatics research includes:

- *The EAV/CR data model*. The entity-attribute-value (EAV) model has been used in a number of clinical information systems to store clinical data. This data model has the advantage, over strictly relational databases, that a large number of clinical data items can be accommodated without massive numbers of tables and that new data items can be included without restructuring and reprogramming the database. We have extended the EAV data model to include complex data items (classes) as values and to allow relationships between data items to be explicitly represented in the database. We call the resulting data model EAV/CR (entity-attribute-value with classes and relationships) and believe that it is well suited to handling heterogeneous bioscience data.[17] We have implemented an EAV/CR database framework and have migrated the operational versions of all four of our HBP databases (ORDB, OdorDB, NeuronDB, and ModelDB) to the EAV/CR model.[18]
- *Tools to support the iterative modeling process*. Another area of neuroinformatics research that we are currently exploring involves developing database approaches and related tools to support the iterative process of neuronal modeling. These tools will maintain an organized record of the different versions of the model, the input data used to test each version, and the results of running the model with those data.

### The Problem and the Solution

In a neuroscience collaboration, one would like to strike a balance between neuroinformatics support and neuroinformatics research, so that both can be pursued in a synergistic fashion. The problem that we encountered in our HBP work was essentially that of

"the chicken and the egg." At the start of our HBP activities, in particular, there was no critical mass of neuroinformatics support activities to provide a context for neuroinformatics research. In addition, our neuroinformatics support activities were applied and pragmatic, reflecting the real-world needs of our neuroscience collaborators. It was therefore difficult (in retrospect, impossible) to perform neuroinformatics research that was directly tied to our collaborators' immediate research needs.

The problem was that we had not achieved a robust level (a critical mass) of neuroinformatics support activities to provide a context for neuroinformatics research. Once we had achieved a sufficiently robust level of neuroinformatics support (which took about five years), we could then embark on neuroinformatics research that was built on our support and therefore directly tied to our collaborators' research needs. This, in turn, meant that the results of our neuroinformatics research could be folded back to enhance our neuroinformatics support, in a fully synergistic fashion.

For example, as described above, we were able to integrate the operational versions of all four of our HBP databases into our EAV/CR data model. We believe that this provides a strong pilot proof of concept for the EAV/CR model and also provides a robust, flexible database environment for the further development of these and future HBP databases at Yale.

## Finding an Academic Home for Bioinformatics

An important question concerns the most appropriate home for bioinformatics in an academic medical center. One possible academic home is in a bioscience department. To the extent that a particular computational technique is unique to a department, then that department may well be a logical home for researchers who focus on that technique. This would be particularly true if such faculty members need to be fully trained in that department's discipline.

To the extent that bioinformatics faculty members require broad training in informatics issues and have skills that are applicable across many bioscience fields, however, there is logic to basing those faculty members in a broader academic unit containing colleagues who share this informatics background. Two general types of such a unit are:

- An academic bioscience informatics unit comprising faculty trained in informatics focused on the biosciences
- An academic biomedical informatics unit comprising faculty trained in bioscience informatics and faculty trained in clinical and health informatics

A unit of the later type would promote—among all faculty, staff, and students—work at the intersection of clinical and bioscience informatics as well as a broader appreciation of biomedical informatics as a whole. As the current trends in bioinformatics continue, the latter model is likely to become an increasingly logical solution. It is clear, however, that many historical, political, and practical considerations will influence how any individual academic medical center approaches this issue.

*References* ■


1. Altman RB. Bioinformatics. In: Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM (eds). Medical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer-Verlag, in press.
2. Sittig DF, Shifman MA, Nadkarni P, Miller PL. Parallel computation for medicine and biology: experience with Linda at Yale University. Int J Supercomput Appl. 1992;6:147–63.
3. Miller PL, Nadkarni PM, Kidd KK, et al. Internet-based support for biomedical research: a collaborative genome center for human chromosome 12. J Am Med Inform Assoc. 1995; 2:351–64.
4. Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. Alfred: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res. 2000; 28:361–3.
5. Kumar A, Cheung KH, Ross-Macdonald P, Coelho PSR, Miller P, Snyder M. Triples: a database of transposon mutagenesis in S cervisiae. Nucleic Acids Res. 2000;28:81–4.
6. Nadkarni PM, Brandt C, Frawley S, et al. ACT/DB: a client–server database for managing entity-attribute-value clinical trials data. J Am Med Inform Assoc. 1998;5:139–51.
7. Martin JB, Pechura CM (eds). Mapping the Brain and Its Functions: Integrating Enabling Technologies into Neuroscience Research. Washington, DC: National Academy Press, 1991.
8. Koslow SH, Huerta MF (eds). Neuroinformatics: An Overview of the Human Brain Project. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.
9. Shepherd GM, Mirsky JS, Healy MD, et al. The Human Brain Project: neuroinformatics tools for integrating, searching, and modeling multidisciplinary neuroscience data. Trends Neurosci. 1998;21:460–8.
10. Nadkarni PM, Reeders ST, Zhou J. CECIL: a database for storing and retrieving clinical and molecular information on patients with Alport syndrome. Proc 17th Symp Comput Appl Medl Care. 1993:649–53.
11. Bassett DE, Eisen MB, Boguski MS. Gene expression informatics: it's all in your mine. Nature Genetics. 1998(suppl 21):51–5.
12. National Human Genome Research Institute, Division of Intramural Research. Microarray project Web site. Available at: http://www.nhgri.nih.gov/DIR/LCG/15K/HTML. Accessed Jun 28, 2000.

13. Stanford University Department of Biochemistry. Patrick O. Brown Laboratory homepage. Available at: http://cmgm.stanford.edu/pbrown. Accessed Jun 28, 2000.
14. Albert Einstein College of Medicine (AECOM). Functional Genomics Project Web site. Available at: http://sequence.aecom.yu.edu/bioinf/funcgenomic.html. Accessed Jun 28, 2000.
15. Skoufos E, Healy MD, Singer MS, Nadkarni PM, Miller PL, Shepherd GM. Olfactory Receptor Database: a database of the largest eukaryotic gene family. Nucleic Acids Res. 1999; 27:343–5.
16. Mirsky JS, Nadkarni PM, Healy MD, Miller PL, Shepherd GM. Database tools for integrating neuronal data to facilitate construction of neuronal models. J Neurosci Methods. 1998;82:105–21.
17. Nadkarni P, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of heterogeneous scientific data using the EAV/CR representation. J Am Med Inform Assoc. 1999; 6:478–93.
18. Marenco L, Nadkarni P, Skoufos E, Shepherd G, Miller P. Neuronal database integration: the Senselab EAV data model. AMIA Annu Symp. 1999:102–6.