Editorial Comments JAMIA

Bioinformatics and Clinical Informatics:
The Imperative to Collaborate

In this issue, Perry Miller and Russ Altman review the experiences at Yale and Stanford that have led to a convergence and cross-pollination between clinical informatics and bioinformatics at those institutions. A related convergence is revealed by a MEDLINE search for the string ''informatics'' in the last five months. Of 346 publications, 175 were in the area of genomics and not clinical applications. Concurrently, there has been much discussion within the informatics community about the dual nature of the research agenda (and, not coincidentally, the funding opportunities) as it pertains to clinical applications and fundamental biological research.[1,2] Informal discussions with investigators in bioinformatics and clinical informatics, however, are tinged with concern that these two disciplines in biomedical informatics will diverge or at least that the two investigator communities are not collaborating sufficiently.

It may, therefore, be timely to briefly review several major categories in which these two strains of biomedical informatics share common methodological and policy challenges. Moreover, as suggested by this overview, the success of bioinformatics and clinical informatics will depend on joint successes in resolving their mutual challenges. The categories outlined here are by no means intended to exhaustively cover the areas of commonality but are intended to provide a useful reference point for discussions on this topic of increasing relevance to the readership of *JAMIA*.

Standard Data Models

In less than a decade, the Human Genome project (HGP)[3] has generated a large amount of biological data that is likely eventually to lead to a qualitative change in the way in which clinical medicine (diagnostics, prognostics, and therapeutics) is practiced. A central intellectual and technologic asset to this effort has been GenBank[4] and related genomic and protein databases (e.g., the SWISS-PROT,[5] Exon-Intron,[6] and IMGT databases[7]). Their standardized data models have allowed research laboratories throughout the world to rapidly populate them with the very latest information. In turn, these databases are freely available throughout the world via the Internet and have seeded, accelerated, and inspired thousands of research projects.

In contrast, there are few, if any, consequential shared national clinical databases. Specifically, patient data in one information system can only rarely be transferred to another to expedite patient care. This, despite decades of research and development of clinical record systems.

This marked contrast is deceptive. The HGP has benefited from the elegant simplicity of the genetic code. In essence, at the level of primary structure, the genetic information coded by any organism is simply a sequence of characters drawn from a very limited alphabet. Consequently, there are only a very few items that GenBank requires be submitted for an entry to be a valid (and useful) component of its database. The clinical care of human beings is far more complex, requiring at the minimum a detailed record of the history of multiple clinical interventions and outcomes, relevant life history, and clinical measurements that span several modalities, from serum chemistry to brain imaging. It is not surprising that the data model required to capture all this information is extremely complex, as is evidenced by the Health Level 7 Ref-

erence Information Model.[8,9] It is a remarkable tribute to the persistence of the individuals involved in these standardization efforts, that they have been able to arrive at a reasonably adequate standardized representation of not only the many descriptors but much of the process and business of clinical care.

As the HGP moves from the acquisition of raw genomics data to the biological function of the discovered genes and their clinical importance, the bioinformatics community will have to address very similar complexities. That is, the clinical annotation of genomic data sets, particularly for human beings, will essentially provide the equivalent, if not identical, challenge of the creation of a comprehensive medical record.

Even prior to encompassing the entirety of clinical annotation, the genomics community has faltered in developing shared and standardized data models where the simplicity of the genome no longer dominates. For example, there are several competing technologies for the massively parallel measurement of gene expression using microarrays. Some of these arrays use two probes per gene and are constructed using robotic spotting techniques. Others are constructed with oligonucleotides using photolithographic techniques.[10] Although all these techniques measure gene expression, a widely adopted standard to represent the results across all microarray technologies has yet to emerge. The GATC[11] proposal, for example, is a possible candidate for such a data model, but its usage is currently spotty and controversial. To clinical informaticians, this will be all too reminiscent of the challenge of creating a shared data model for laboratory results.

Standardized Vocabularies

The lack of widely accepted standardized vocabularies for clinical care has greatly hampered the development of automated decision support tools and clinical research databases. The impossibility of guaranteeing that a serum sodium or systolic blood pressure has the same code or term throughout our hospital system is troublesome. Fortunately, several efforts in the private and public realm (e.g., LOINC[12]) are addressing this issue. The National Library of Medicine has invested large resources to enable these different vocabularies to be interoperable, at least at a basic level.[13]

The same problems are not unknown in bioinformatics. Even at this early stage of the HGP, DNA sequences that were previously not known to be part of the same gene have different names and are joined in only some databases (with varying levels of confidence).

As the HGP ventures into diverse areas of bioscience (as well as into the clinical area), vocabulary issues are also important. Indeed, the lack of a standardized vocabulary already arises in genomics as well, in annotations. For example, despite the fact that the basic data element of GenBank is the sequence (which has an easily standardized representation), there are diverse annotations that are very nonstandardized right now.

Errors

A recent report by the Institute of Medicine[14] highlights the immense mortality and morbidity due to medical errors. Clinical informaticians (e.g., Bates et al.[15] and Kuperman et al.,[16,17]) have been instrumental in demonstrating how automated systems can be used to reduce this error rate. These industrial processing and quality improvement techniques are not without relevance to the HGP. It is well known that the mouse genome database has been contaminated with entries of the rat genome and that the specification of 5′ to 3′ polarity of a gene sequence has been found to be inverted.[18–20] And these are only some of the known errors in a very large effort.

These sources of error can be reduced, as they have in many industries, by the application of increased process automation and automated interception of human error before it becomes consequential. The architectures of clinical order entry systems, designed for complex clinical enterprises to prevent erroneous and dangerous clinical behavior, can inform the design of genome sequencing and expression profiling processes to prevent the kind of errors we are already finding in genomic databases.

Noise

It is well known that the noise in clinical measurements leads to erroneous decision making. The archetypal example is in the intensive care unit, where multiple physiologic monitors each has its own alarm module. Because of the noisy nature of the biological signals that are monitored, the alarms are ignored or switched off because of their high false-positive rate.[21] The noisy nature of the monitored signals thus has a significant impact on the provision of care and the decision-making ability of care providers who are working under conditions of uncertainty and data overload.

Similar noise considerations arise in genomics. For example, with gene microarrays, we can measure the

expression of tens of thousands of genes at a time. There are several sources of noise in these measurements: within a microarray, across microarrays, and from the intrinsic variability of the biological systems being measured. Yet in 1999, several reports, which appeared in scientific journals of the first rank, included changes in expression so small as to be indistinguishable from noise. Such changes are, in essence, a false-positive result. These false positives are potentially extremely costly. A biological researcher might decide to invest several months investigating a gene's regulation because a microarray experiment showed it to be increased or decreased under a particular set of conditions.

In clinical informatics there is a rich literature of the techniques that can be used to identify false positives and reduce noise (e.g., filtering, signal fusion[22–26]). Many of these techniques are transferable to the genomic domain.

Privacy

In 1997, the Institute of Medicine[27] reported significant lacunae in both technology and policy in protecting confidential patient data. Among the problems of greatest concern that were emphasized by this report were the relatively unrestricted access by third parties to these data for secondary uses and the inadequacy of the anonymization process (in both practice and theory[28,29]). Subsequently, the clinical informatics community has developed several model confidentiality policies[30] and cryptographic identification systems.[31]

As the fruits of the HGP are translated first into clinical research protocols and then into clinical practice, personally identifiable genomic data will find their way into some form of information system. The challenges posed to the security and privacy of such data will dwarf any encountered to date with conventional clinical data. The reasons are twofold: First, genomic information is likely to be much more predictive of current and future health status than most clinical measurements. Second, with very few exceptions, an individual's genome is uniquely identifying. This identifiability is much more reliable, persistent, and specific than typically cited identifiers, including a person's name, social security number, date of birth, and address.

At the very least, the architects of information systems storing genetic data should learn from all the mistakes of and designs developed for the security architectures and privacy policies of conventional clinical information systems. Conversely, the extreme concerns posed by the storage of personal genetic data is likely to generate new policies and security architectures that will enhance the confidentiality of clinical information systems. Moreover, when personal genetic data becomes incorporated into routine medical practice, the safeguards for the confidentiality of the medical record will be crucial to the confidentiality of the genetic data referenced there.

Costs of Acquiring Data

"Getting the data in" has often been cited[32] by authorities in clinical informatics as being among the most difficult challenges in successfully deploying clinical information systems. In particular, the costs of acquiring detailed and structured data from the clinical care process have been daunting. Voice and handwriting recognition information systems have not been broadly adopted, because of a variety of performance and usability issues. The cost and practicability issues will continue to present obstacles to clinical information system utility and deployment until better solutions are arrived at. In contrast, the HGP has managed to achieve significant economies of scale in sequencing technology. Gene microarrays alone have dropped in cost by a factor of two in just the last year.

Here again, once genomic investigators attempt to bridge the gulf from purely genomic data sets to phenotypically (i.e., clinically) annotated data sets, they will be confronted with the same challenges of clinically oriented, codified data acquisition. The questions of which user interfaces are the most cost efficient, reliable, and generalizable to multiple clinical domains are among the implementation and design challenges that they will face. Although they have yet to arrive at definitively successful answers, clinical informaticians have already completed several decades worth of engineering and ethnographic studies[33] addressing the very same questions.

Extracting Knowledge from Data

The first rough draft of the human genome was reported to have been completed in May of this year.[34] It is likely that a complete, high-quality human DNA reference sequence will be available by 2003. Yet the function of the vast majority of the genes in the human genome will be unknown. The minority of genes with documented function are likely to have many more functions and interactions that are unknown.

Consequently, one of the primary applications of information technologies in genomics is the application of machine learning techniques to determine how genes are functionally interdependent and how these interdependencies are reflected in the biological and clinical behavior of the system in which they oper-

ate.[3,35] Many of these machine learning techniques were previously applied to the task of extracting knowledge from clinical databases, and some were even developed first in the clinical domain.[36–44] To be sure, the genomic era has challenged these machine learning techniques to the extreme, because of the high dimensionality of data sets (e.g., tens of thousands of measurements per experiment) and the relatively few cases and experiments from which investigators are attempting to glean knowledge.

Summary

Without being exhaustive, this brief review suggests the multiple points of commonality between the genomic and clinical strands of the biomedical informatics research agenda. It also suggests that the training of investigators in informatics should include a set of core competencies that at least cover these common points. In this fashion, the joint research agenda might be well served to the mutual benefit of biomedical science and clinical care. The Stanford Medical Informatics educational program, described in this issue, illustrates this benefit.

ISAAC S. KOHANE, MD, PHD

References ■

1. Rindfleisch TC, Brutlag DL. Directions for clinical research and genomic research into the next decade: implications for informatics. J Am Med Inform Assoc. 1998;5(5):404–11.
2. Lynch C. Medical libraries, bioinformatics, and networked information: a coming convergence? Bull Med Libr Assoc. 1999;87(4):408–14.
3. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New goals for the U.S. Human Genome Project: 1998–2003. Science. 1998;282(5389):682–9.
4. Benson DA, Boguski MS, Lipman DJ, et al. GenBank. Nucleic Acids Res. 1999;27(1):12–7.
5. Bairoch A, Apweiler R. The swiss-prot protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28(1):45–8.
6. Saxonov S, Daizadeh I, Fedorov A, Gilbert W. EID: the Exon-Intron database: an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res. 2000; 28(1):185–90.
7. Ruiz M, Giudicelli V, Ginestoux C, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 2000;28(1):219–21.
8. Quinn J. An HL7 (Health Level Seven) overview. J Ahima. 1999;70(7):32–4.
9. Beeler GW Jr. On the rim: the making of HL7's Reference Information Model. MD Comput. 1999;16(6):27–9.
10. Ramsay G. DNA chips: state of the art. Nat Biotechnol. 1998; 16(1):40–4.
11. Consortium G. GATC Specifications. 1998. Available at: http://www.gatconsortium.org/specifications.html.
12. Huff SM, Rocha RA, McDonald CJ, et al. Development of the Logical Observation Identifiers Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5(3): 276–92.
13. Lindberg D, Humphreys B. The Unified Medical Language System (UMLS) and computer-based patient records. In: Ball M, Collen M (eds). Aspects of the Computer-based Patient Record. New York: Springer-Verlag, 1992:165–75.
14. Kohn LT, Corrigan JM, Donaldson MS (eds). To Err Is Human: Building a Safer Health System. Washington, DC: National Academy Press, 2000.
15. Bates DW, O'Neil AC, et al. Potential identifiability and preventability of adverse events using information systems. J Am Med Inform Assoc. 1994;1(5):404–11.
16. Kuperman GJ, Teich JM, Bates DW, et al. Detecting alerts, notifying the physician, and offering action items: a comprehensive alerting system. Proc AMIA Annu Fall Symp. 1996:704–8.
17. Kuperman GJ, Sussman A, Schneider LI, Fiskio JM, Bates DW. Towards improving the accuracy of the clinical database: allowing outpatients to review their computerized data. Proc AMIA Annu Symp. 1998:220–4.
18. Seluja GA, Farmer A, McLeod M, Harger C, Schad PA. Establishing a method of vector contamination identification in database sequences. Bioinformatics. 1999;15(2):106–10.
19. Lamperti ED, Kittelberger JM, Smith TF, Villa-Komaroff L. Corruption of genomic databases with anomalous sequence. Nucleic Acids Res. 1992;20(11):2741–7.
20. Savakis C, Doelz R. Contamination of cDNA sequences in databases [letter]. Science. 1993;259(5102):1677–8.
21. Tsien CL, Fackler JC. Poor prognosis for existing monitors in the intensive care unit. Crit Care Med. 1997;25(4):614–9.
22. Uckun S, Dawant BM, Lindstrom DP. Model-based reasoning in intensive care monitoring: the YAQ approach. Artif Intell Med. 1993;5(1):31–48.
23. Sittig DF, Gardner RM, Pace NL, Morris AH, Beck E. Computerized management of patient care in a complex, controlled clinical trial in the intensive care unit. Comput Methods Programs Biomed. 1989;30:77–84.
24. Hayes-Roth B, Washington R, et al. Guardian: a prototype intelligent agent for intensive-care monitoring. Artif Intell Med. 1992;4(2):165–85.
25. Gardner RM, Hawley WL, East TD, Oniki TA, Young H-FW. Real-time data acquisition: experience with the Medical Information Bus (MIB). Proc Symp Comput Appl Med Care. 1991:813–7.
26. Clemmer T, Gardner R. Medical informatics in the intensive care unit: state of the art 1991. Int J Clin Monit Comput. 1991–92;8(4):237–50.
27. National Research Council. Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. For the Record: Protecting Electronic Health Information. Washington, DC: National Academy Press, 1997.
28. Sweeney L. Guaranteeing anonymity when sharing medical

Affiliation of the author: Harvard University, Cambridge, Massachusetts.

Correspondence and reprints: Isaac S. Kohane, MD, PhD, Children's Hospital and Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115;
e-mail: ⟨isaac_kohane@harvard.edu⟩.

Received for publication: 6/9/00; accepted for publication: 6/12/00.

data: the Datafly system. Proc AMIA Annu Fall Symp. 1997: 51–5.
29. Sweeney L. Replacing personally-identifying information in medical records: the SCRUB system. Proc AMIA Annu Fall Symp. 1996:333–7.
30. Rind DM, Kohane IS, Szolovits P, Safran C, Chueh HC, Barnett GO. Maintaining the confidentiality of medical records shared over the Internet and World Wide Web. Ann Intern Med. 1997;127(2):138–41.
31. Kohane IS, Dong H, Szolovits P. Health information identification and de-identification toolkit. Proc AMIA Annu Symp. 1998:356–60.
32. McDonald CJ, Tierney WM, Overhage JM, Martin DK, Wilson GA. The Regenstrief Medical Record System: 20 years of experience in hospitals, clinics and neighborhood health centers. MD Comput. 1992;9:206–17.
33. Coiera E, Tombs V. Communication behaviours in a hospital setting: an observational study. BMJ. 1998;316(7132):673–6.
34. Wade N. Assembling of the Genome Is at Hand. New York Times. April 7, 2000.
35. Rastan S, Beeley LJ. Functional genomics: going forwards from the databases. Curr Opin Genet Dev. 1997;7(6):777–83.
36. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research: an introduction to Bayesian methods in health technology assessment. BMJ. 1999;319(7208): 508–12.
37. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems, part I: the Pathfinder project. Methods Inf Med. 1992;31(2):90–105.
38. Andreassen S, Benn J, Hovorka R, Olesen KG, Carson ER. A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study. Comput Methods Programs Biomed. 1994;41:153–65.
39. Avent RK, Charlton JD. A critical review of trend-detection methodologies for biomedical monitoring systems. Crit Rev Biomed Eng. 1990;17(6):621–59.
40. Bellazzi R, Siviero C, Stefanelli M, de Nicolao G. Adaptive controllers for intelligent monitoring. Artif Intell Med. 1995; 7(6):515–40.
41. Berger MP, Gelfand RA, Miller PL. Combining statistical, rule-based and physiologic model-based methods to assist in the management of diabetes mellitus. Comp Biomed Res. 1990;23:346–57.
42. Berzuini C, Bellazzi R, Quaglini S, Spiegelhalter DJ. Bayesian networks for patient monitoring. Artif Intell Med. 1992; 4:243–60.
43. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score. Crit Care Med. 1988;16(11):11106.
44. Butte A, Kohane I. Unsupervised knowledge discovery in medical databases using relevance networks. AMIA Annu Symp. 1999:711–5.

■ **J Am Med Inform Assoc.** 2000;7:512–516.

■