DANIEL R. MASYS, M.D.
Associate Editor

Medical Informatics

Structural Informatics and Its Applications in Medicine and Biology

JAMES F. BRINKLEY, M.D., Ph.D.

The amount of information generated in all fields of science, particularly medicine and biology, is exponentially increasing. This trend, plus rapid advances in computer hardware, has led to the emergence of the field of informatics and, within the medical arena, medical informatics. Blois and Shortliffe define medical informatics as "the rapidly developing scientific field that deals with the storage, retrieval, and optimal use of biomedical information, data, and knowledge for problem solving and decision making."[1] Although most medical informatics research has concentrated on clinical applications (e.g., medical records, hospital information systems, radiology systems, and pharmacy systems),[1,2] there is nothing in the definition to limit the field to clinical medicine. In fact, the definition implies that medical informatics covers the entire spectrum of information within a medical center, including basic science as well as clinical medicine. This broad interpretation is reasonable because clinical decision-making relies on information at the basic science level.[3]

An important type of basic medical information is structural data and knowledge describing objects at the gross anatomic, cellular, and molecular levels. Accordingly, the National Library of Medicine has recognized the emergence of a subdiscipline of medical informatics, called "structural informatics," that deals with the ". . . computerized representation of biomedical structural data, and its linkage to related text and numeric data."[4] I will expand on this definition and show how structural informatics research applies to problems in medicine and biology.

The Structural Approach in Biology

Implicit in the word "structure" is not only the concept of elementary units or parts, but also the interdependence and relationships of those parts to form a whole. It can thus be argued that modern science has adopted a structural approach to understanding the natural world, in which parts are defined and the interactions among them are explored. In medicine, much of the progress in our ability to understand and treat disease can be seen to be a result of the structural approach. At one time, diseases were thought to be due to mysterious vitalistic forces, but once biologists began to dissect the body, they were able to observe parts (organs and cells) and their interrelationships, and were able to form theories (such as the cellular basis of life) that now provide the foundation for modern basic and clinical science. Our continued probing of ever finer structural levels is leading to an increasingly sophisticated understanding of structure-function relationships, and the current pre-eminence of molecular biology is simply a logical extension of this progression.

As the level of structural analysis become finer, there is a corresponding increase in the amount of structural information. At the gross level, the number of facts is small enough that a single human being can comprehend most or all of them and relate them to a coherent whole. However, at the cellular and molecular levels, where most current research is performed, there are simply too many facts for one person to comprehend. And if the human genome project is successful, the number of facts (nucleotide sequences) will increase by orders of magnitude.[5]

At least one eminent biologist, Walter Gilbert, laments that this constant reductionism has led to a certain dissatisfaction in biology, a feeling that there has been an exponential increase in facts without the corresponding theories to explain them.[6] Accumulation of data has become the endpoint, rather than deeper understanding. This observation has led Gilbert to suggest that a paradigm shift is beginning to occur in biology, that although the accumulation of facts at all levels will continue, the current purely descriptive paradigm will gradually be replaced by a theoretical paradigm that only turns to experimental methods in order to test models or hypotheses. Thus, using the example of the human genome project, the critical issue will be how to synthesize the individual pieces of sequence information into a coherent and useful body of knowledge. Because of the vast amount of data required for such synthesis, the starting point for biological research will, of necessity, be networked databases and knowledge bases of structural information.

Others have also recognized the importance of computerized information sources to the basic sciences, and have proposed developing a computer-based matrix of biological knowledge.[7] Such an information resource could lead to better methods for sharing information, and for inferring patterns and theories from disparate sets of facts at many levels of the structural hierarchy. The utility of such a matrix would be due to the computer's capacity to simultaneously handle large amounts of data, and thereby to find analogous patterns of organization in highly diverse areas. For example," . . . researchers have found similar molecules in protozoans, bacteria and yeast. And a signalling molecule in yeast mating has also been found in the sex hormones of higher organisms—which belies the assumption that endocrine glands, a much later development than yeast, evolved their own signaling molecules."[7]

The availability of a highly interconnected network of biological information would allow computers to search for these common patterns, and to present them to researchers and clinicians in ways that would facilitate synthesis and integration into experimentally testable theories. The challenge is how to actually build this matrix, how to represent structural data and knowledge, and how to make this information widely available to both humans and computers. Many of the ongoing research activities involved in meeting these challenges would comprise the field of structural informatics.

The Goals of Structural Informatics

The fundamental view of structural informatics is that emergent properties of complex systems (including the human organism) arise from coherent interac-

Examples of Research Problems in Structural Informatics

- Storing and retrieving gene sequence data in large databanks
- Gene sequence comparison methods
- Determination of protein structure from experimental data derived from the techniques of X-ray crystallography or NMR spectroscopy, or from theoretical constraints of protein folding
- Development of visual databases to manage images, and making them available for researches and clinicians
- Analysis (as opposed to generation) of images, particularly when models of anatomy are used to guide the process
- Methods for representing the expected shapes and ranges of variation of individual structural objects, the relationships between them, and their spatial decomposition into parts
- Methods for naming structural objects, and for placing them in symbolic hierarchies outlining subdivisions, subparts, and functions
- Methods for presenting structural data and knowledge to the user in a manner that facilitates synthesis: graphics, scientific visualization, hypermedia, and virtual reality
- Methods for distributing data and knowledge in linked databases and knowledge bases that are accessible over the computer network, to computer programs as well as to humans

tions among many parts; that all objects, whether humans or atoms, do not exist in isolation but rather within ecosystems comprised of other interacting objects; and that there is no single active entity that controls the behavior of an object or system. The fundamental goal, therefore, is to provide an information framework within which the multitude of facts gained from reductionist approaches can be integrated into models of complex interacting systems. These models will consist of hierarchical networks of interacting objects, where each object is itself a network of interacting objects. The objects will be highly linked, distributed throughout the worldwide computer network, and accessible at all times.

A secondary goal is to make this framework accessible to humans in ways that let multiple viewpoints be expressed while promoting the development of consensus, and facilitate the synthesis of large numbers of facts into new models and theories that represent new knowledge in science.

It is clear at the outset that, because of limitations in computer resources and the human mind, it is impossible to precisely model objective reality at all levels of detail, much less to present the information in its entirety to the human user. For example, since no object or organism exists in isolation, a complete model of its structure and function would require that the position and velocity of every atom in the universe be known. Heisenberg has shown this to be impossible, but even if it were not the computer required to implement such a model would have to be at least as large as the universe itself. Thus, we will always be forced to choose which aspects of objective reality to include in our models, depending on the uses to which the models will be put. That is, a numeric model of heart function using fractal geometry might be appropriately accessed by computer programs for simulating the heart, but the results of this simulation might better be presented to the human user as a graphic model showing three-dimensional animated displays of heart motion. The research issues of structural informatics arise as a result of this tradeoff between the desire to precisely model objective reality and practical limitations of computer and human resources.

Research Problems in Structural Informatics

The information structures previously described are already being developed, and will continue to be developed, whether or not a formal field called "structural informatics" ever exists. The primary reason for defining a new field is that it may allow cross-fertilization among various researchers as they discover common methods for dealing with problems.

Within biology, it may be useful to describe a set of research problems that exemplify structural informatics research and therefore provide, by example, concrete starting points about the nature of the field. The problems to be considered deal with information about physical structure, since physical structure provides a useful framework for understanding function in biology. Information about the physical structure of the body falls into two major categories: spatial and symbolic.[8,9]

The spatial category is concerned with the structure of objects in space. Within this category, objects can be considered according to their levels of organization: primary structure (for example, linear gene sequences specifying protein amino acid sequences), secondary structure (protein alpha helices and beta sheets), tertiary structure (the three-dimensional folding of proteins), quaternary structure (protein complexes), and higher levels of organization (organelles, cells, tissues, organs, and the entire organism).

The symbolic category is concerned with the names of objects, taxonomic hierarchies, descriptions of what the objects do, how they develop, and what can go wrong with them. The spatial category roughly corresponds to the images in an anatomy or molecular biology textbook, whereas the symbolic category corresponds to the textual descriptions. These categories are somewhat arbitrary, however, in that spatial information can be described symbolically as well as numerically.

Research problems can also be classified along the spectrum from data to knowledge. Problems at the data end of the spectrum deal with information about individual objects (a single protein, a single cell, or a single patient), whereas problems at the knowledge end deal with information about classes of objects represented as models (all globular proteins, all T-cells, all patients with AIDS). Nearer the data end are methods for determining structure, and methods for storing and retrieving structural data. Nearer the knowledge end are methods for building models that capture knowledge about structure, methods for determining how structural models interact to produce changes in structure, methods for

storing and retrieving models, methods for displaying the models and data to the human user, and methods for distributing the models and data in the computer network.

Career Paths and Training in Structural Informatics

The research problems described in the previous sections are inherently interdisciplinary, requiring expertise in both computers and biology. As the information crisis continues to worsen, it is likely that workers with knowledge of both these areas will be in great demand, both in academia and in industry.

Academic research in structural informatics will initially take place within traditional departments. Within the medical school these departments might include anatomy, biological structure (also called structural biology), molecular biology, biochemistry, radiology, radiation oncology, and surgery. On the technical side the departments might include computer science, electrical engineering, and bioengineering. As medical informatics departments become established, structural informatics will also be very suitable as a focal area within these departments.

Industrial positions will become available in areas such as medical imaging and biotechnology. Medical imaging companies have, until now, been concerned mostly with image generation. However, there are now so many digital images available that the companies are looking for ways to manage, analyze, and display the images. Similarly, biotechnology companies have perfected the techniques for cloning virtually any gene. The pertinent question now is, which gene should they clone, or which amino acid modification should they make to produce a desired protein structural change? Structural informatics techniques of protein structure determination, gene sequencing, and management of molecular databases should be in demand as these problems become more pronounced. Because imaging and biotechnology are currently two of the fastest-growing biomedical industries, the industrial prospects for workers trained in structural informatics should be very promising.

Students of structural informatics will need to learn aspects of both biology and information science. The basic core courses can be similar, or even identical to the parent field of medical informatics; electives can provide the structural dimension. A basic set of core requirements in computer science might consist of programming, data structures, simple computer architecture, databases, computer networks, and basic artificial intelligence techniques, with emphasis on knowledge representation and qualitative modelling. On the biological side, emphasis should be placed on basic medical science, particularly with one or two anatomically based courses such as anatomy, histology, cell biology, biochemistry, and molecular structure. Other required courses might consist of basic math through calculus, linear algebra, and statistics.

In addition to these basic courses, students could take electives depending on their individual research interests. These might include computer graphics, scientific visualization, virtual reality, hypermedia, mathematical modelling, crystallography, sequencing techniques, NMR spectroscopy, and medical image analysis. These courses could also be supplemented by research seminars that would help clarify the field.

Broader Implications

There is nothing in the name "structural informatics" that necessarily restricts it to biology or even to physical structure. One of the main reasons for defining such a field is the observation that patterns of organization repeat themselves throughout nature. Thus, it may be that methods for representing structures, as networks of interacting substructures, will have implications outside of biology as well. For example, hierarchical networks could be defined below the molecular level to the chemical and atomic level, leading to applications of structural informatics in chemistry and physics. Similarly, such networks could be extended to larger ecosystems involving interactions between humans and the environment, so may prove useful for environmental and social studies as well.

The structural approach in science has been both a blessing and a curse. Most of our technological and medical advances have arisen because of our insatiable desire to take things apart and see how they work, but the sheer number of parts has now become so great that it is difficult to put them together again. Frustration with this situation has led some to abandon the structural approach entirely. But the structural approach does not only imply reductionism; rather it implies a balance between taking things apart and putting them back together. The difficulty is that much of science, and particularly biology has become imbalanced, putting too much emphasis on taking things apart, but not enough on fitting them back together. It is not that there is lack of desire to put things together, it is just that recreating the whole is now more difficult because of the larger number of parts. Structural informatics has as its goal the development of computer-based tools that will help us put things back together. To do this we must recognize that information is an important entity worthy of study in itself, and that by understanding the nature of information, we can organize it so as to regain the wholeness of science without throwing away the parts.

This research was supported in part by National Library of Medicine grant LM04925, the Murdock Foundation Charitable Trust, and the University of Washington School of Medicine. The author thanks John Prothero, Cornelius Rosse, and Sheila Lukehart (all of the University of Washington) for valuable discussions concerning this paper.

Dr. Brinkley is research assistant professor, Department of Biological Structure, University of Washington, Seattle.

References

1. Shortliffe, E. H., Perreault, L. E., Wiederhold, G., and Fagan., L. eds. *Medical Informatics: Computer Applications in Health Care.* Menlo Park, California: Addison-Wesley, 1990.
2. Greens, R. A., and Shortliffe, E. H. Medical Informatics: An Emerging Academic Discipline and Institutional Priority. *JAMA* **263**(1990):1114–1120.
3. Blois, M. S. Medicine and the Nature of Vertical Reasoning. *N. Engl. J. Med.* **318** (1988):847–851.
4. National Library of Medicine. *Electronic Imaging: National Library of Medicine Long Range Plan*, Washington, D.C.: U.S. Department of Health and Human Services, April 1990.
5. Pearson, M. L., and Soll, D. The Human Genome Project: A Paradigm for Information Management in the Life Sciences. *FASEB Journal* **5**(1991):35–39.
6. Gilbert, W. Towards a Paradigm Shift in Biology. *Nature* **349**(1991):99.
7. Holden, C. An Omnifarious Data Bank for Biology. *Science* **228**(1985):1412–1413.
8. Brinkley, J. F., Prothero, J. S., Prothero, J. W., and Rosse, C. A Framework for the Design of Knowledge-based Systems in Structural Biology. In *Proceedings of the 13th Annual Symposium on Computer Application in Medical Care*, pp. 61–65. Washington, D.C., IEEE Computer Society Press, 1989.
9. Rosse, C., Brinkley, J. F., and Prothero, J. S. Structural Informatics: The Representation of Anatomical Knowledge in Computer Readable Form. Paper presented at the American Medical Informatics Association, 1st Annual Educational and Research Conference, Snowbird, Utah, June 1990.