The Biomedical Information Science and Technology Initiative

Prepared by the Working Group on Biomedical Computing
Advisory Committee to the Director
National Institutes of Health
June 3, 1999

CHARGE TO THE WORKING GROUP ON BIOMEDICAL COMPUTING

The biomedical community is increasingly taking advantage of the power of computing, both to manage and analyze data, and to model biological processes. The working group should investigate the needs of NIH-supported investigators for computing resources, including hardware, software, networking, algorithms, and training. It should take into account efforts to create a national information infrastructure, and look at working with other agencies (particularly NSF and DOE) to ensure that the research needs of the NIH-funded community are met.

It should also investigate the impediments biologists face in utilizing high-end computing, such as a paucity of researchers with cross-disciplinary skills. The panel should consider both today's unmet needs and the growing requirements over the next five years (a reasonable horizon for extrapolating the advances in the rapidly changing fields of computing and computational biology).

The result of deliberations should be a report to the NIH Director, which will be presented to the Advisory Committee to the Director. The report should include recommendations for NIH actions to support the growing needs of NIH-funded investigators for biomedical computing.

EXECUTIVE SUMMARY

In science and technology in the latter half of the 20th century, two fields have stood out for their speed of progress and their effect on society: biomedicine and computation. The charge of this Working Group is to assess the challenges and opportunities presented to the National Institutes of Health by the convergence of those two disciplines.

The principal obstacle impeding effective health care is lack of new knowledge, and the principal mission of the NIH is to overcome this obstacle. At this point the impact of computer technology is so extensive it is no longer possible to think about that mission without computers.

Increasingly, researchers spend less time in their "wet labs" gathering data and more time on computation. As a consequence, more researchers find themselves working in teams to harness the new technologies. A broad segment of the biomedical research community perceives a shortfall of suitably educated people who are competent to support those teams. The problem is not just a shortage of computationally sophisticated associates, however. What is needed is a higher level of competence in mathematics and computer science among biologists themselves. While that trend will surely come of its own, it is in the interest of the NIH to accelerate the process. Digital methodologies — not just digital technology — are the hallmark of tomorrow's biomedicine. The NIH therefore must find ways to discover, encourage, train, and support the new kinds of scientists needed for tomorrow's science.

To make optimal use of information technology, biomedical researchers need, first of all, the expertise to marry information technology to biology in a productive way. New hardware and software will be needed, together with deepened support and collaboration from experts in allied fields. Inevitably, those needs will grow as biology moves increasingly from a bench-based to a computer-based science, as models replace some experiments and complement others, as lone researchers are supplemented by interdisciplinary teams. The overarching need is for an intellectual fusion of biomedicine and information technology.

Invariably, scientists learn best by doing rigorous science. Indeed, the NIH mission is to do science, including teaching and learning. Socially meritorious goals of improving human health and preventing, detecting, diagnosing, and treating disease and disability are achieved most effectively when pursued within the overall context of rigorous science. This report and its recommendations focus, therefore, on science — both for its insights and as a path toward building an educated interdisciplinary workforce. The centerpiece of our recommendations is the proposal to inaugurate National Programs of Excellence in Biomedical Computing. It is in the context of those National Programs that the best opportunities can be created for doing and learning at the interfaces among biology, mathematics, and computation. With such new and innovative programs in place, scientists will absorb biomedical computing in due course, while supporting the mission of the NIH.

Recommendation #1:

The NIH should establish between five and twenty National Programs of Excellence in Biomedical Computing devoted to all facets of this emerging discipline, from the basic research to the tools to do the work. It is the expectation that those National Programs will play a major role in educating biomedical-computation researchers.

National Programs of Excellence in Biomedical Computing would advance research in particular areas of biomedicine, focusing on those in which computation is becoming increasingly essential. They would be funded in part through a new program, and in part through research grants from one or more of the Institutes that make up the NIH. The academic or research institutions at which the National Programs would be housed would be expected to contribute to the programs — and teaching would be an essential contribution.

National Programs could range in size. At a modest level, three to five researchers in complementary disciplines might receive $1.5 million a year to undertake the exploration of a single problem. Larger National Programs might bring together several problems and several technologies, perhaps in association with more than one institution or Institute, for up to $8 million a year. The NIH will determine the number and scope based on the applications and the grant process.

One important goal of the National Programs will be to develop and integrate the use of computational tools to meet the important challenges of biomedical research. These Programs are in keeping with the conclusions of the President's Information Technology Advisory Committee (PITAC) report in that it focuses on basic information technology research in the pursuit of insight into the issues facing biomedical research. Concurrently, the National Programs will create homes for interdisciplinary teams, and those teams will establish nurturing environments for exploration and education. In establishing National Programs, the NIH will send a powerful message, both in academe and within the NIH community itself, about the importance of computation and the value of interdisciplinary research.

Strong action by the NIH is required because the existing biomedical research and teaching structures of the universities and research institutions of this country inadequately value interdisciplinary efforts generally, and computation in particular. Few grant programs and fewer academic departments foster the kind of interdisciplinary work required to meet biomedical challenges, let alone educate students about them. National Programs specifically would include formal and informal instruction from the undergraduate through post-graduate levels, and incorporate a range of opportunities for scholars and researchers to participate.

Recommendation #2:

To make the growing body of biological data available in a form suitable for study and use, the NIH should establish a new program directed toward the principles and practice of information storage, curation, analysis, and retrieval (ISCAR).

The information that biomedical researchers are amassing in profuse quantities today-from the Human Genome Project, clinical trials, statistics, population genetics, and imaging and visualization research — creates enormous digital repositories of information. The scale of those databases swamps all the information collected before. They encompass multigigabyte, multivariate functions of three dimensions, of wavelength, and of time. The infrastructure needed to make them available is phenomenal: A single biomedical laboratory could produce up to 100 terabytes of information a year — about the same as the information in one million encyclopedias. In order to be useful, the data must be indexed and stored, and the challenges for data analysis and abstraction are formidable.

The creation and development of advanced databases and database technologies (methods for storing, retrieving, and analyzing biomedical data) is becoming more important in all biomedical fields. The emerging technology of bioinformatics helps researchers gather and standardize data from basic research and computer modeling, and combine and manipulate databases to tease out the knowledge they contain. The goal is a system of interoperable databases that will make available the fruits of the increased productivity enabled by computation.

That is particularly true in clinical research: As more information from clinical trials becomes available, the need for standardization and interoperability of clinical databases will increase dramatically. Coordinating knowledge gained from clinical trial data with new insights from genetic research could appreciably advance knowledge about the treatment of disease. A system of interoperable databases would allow clinical researchers to track any finding back to its basic science roots; conversely, a research scientist might track forward to postulate from hypotheses through potential applications based on innovative uses of existing data.

As the amount of data grows, the tools to compare and manipulate the data become more important. These tools form software bridges between databases that will allow researchers to link disparate information sources.

The NIH has been a leader in establishing databases of valuable information and making them available for study. Now it must organize and expand database resources internally and externally. Currently the agency uses contracts, grants, and cooperative agreements in bioinformatics, but no program focuses specifically on database development. Both the collection of the information, and the creation of the tools for storage, management, and access are increasingly important. Therefore, the NIH needs a program that will rally new and important bioinformatics efforts and build this vital part of the biomedical infrastructure.

Recommendation #3:

The NIH should provide additional resources and incentives for basic research (through R01 grants) to provide adequate support for those who are inventing, refining, and applying the tools of biomedical computing.

Biomedical scientists know best what they need, and they often need to take advantage of computational opportunities. However, in evaluating research grants and programs, reviewers and staff sometimes have been reluctant to provide support for computation and computational infrastructure at the level required. The computational infrastructure, of course, includes not only the hardware but also the people with the expertise to make good use of the hardware. It is time for the NIH to recognize the importance of both the tools and those who build them. In order to do that, the NIH needs to ensure that R01 grants may be used for biomedical computation. That is particularly important for grants that support environments rich in teaching potential as well as research excellence. Researchers who work with students should have the resources that will allow them to set an example of the use of biomedical computing.

As with any special emphasis or targeted funding, evaluation at three years is recommended.

Recommendation #4:

The NIH should foster a scalable national computer infrastructure. To assure that biomedical researchers can gain access to the computing resources they need beyond their desktops, the NIH should provide financial resources to increase computing capacity, both local and remote. The purpose of this recommendation is to establish a balanced infrastructure for all computational needs.

Biology is becoming increasingly complex and computation is becoming increasingly sophisticated. Today's biomedical computing needs resources that go beyond desktop computers to local clusters of processors, to mid-level facilities, and to the most powerful computers at national centers. Many biomedical researchers cannot do their work on their desktop computers alone. They need varying amounts of computing power at different times, and those resources should be made available. The infrastructure must be better balanced for a dynamic range of computational needs.

Powerful computers alone are not enough. The entire support system must be in place. Even researchers who can do their work on small clusters need access to the expertise to set up and manage those clusters, and need support from programmers who can write or adapt the necessary software. As the computing-power needs increase, so do the support needs.

The NIH should support facilities with mid-level computers where new algorithms and applications can be developed specifically for biological problems. The biomedical expertise at those facilities would support researchers seeking to adapt and apply the best computer technology to their work. For some applications, mid-level facilities could offer smaller versions of scalable systems that exist at the national supercomputer centers. Researchers might use those resources to test and develop code or design before moving to national supercomputer centers, or — in appropriate cases — to do their work on more powerful computational resources than they have in their laboratories. Mid-level facilities could be created through National Programs that focus on supercomputing science, or the resources could be made available through cooperative agreements with existing extramural centers as well as at intramural centers.

NIH scientists have long taken advantage of the national supercomputer centers run by the National Science Foundation and the Department of Energy for high-level computing. Because the number of biomedical researchers who can profit from using those facilities is increasing, the NIH should take a strong leadership position and help support the national supercomputer centers. Such NIH support would provide a welcome opportunity for a partnership between NSF and the NIH as the future of science unfolds in the 21st century.

CONCLUSION

The NIH can make a powerful contribution to the development of tomorrow's biomedical research community by increasing efforts to promote and support computational biology today. With the appropriate support in place, interdisciplinary research teams will coalesce for National Programs of Excellence in Biomedical Computing and ISCAR efforts. The natural byproduct of their emphasis on biomedical research will be a new generation of researchers who are skillful with computing, and who will have helped to create the computational tools they need to meet tomorrow's challenges.

As biomedical research becomes more computationally intensive, the Biomedical Information Science and Technology Initiative (BISTI) is essential if the NIH is to fulfill its mandate. This Initiative will be the means by which new techniques are developed, new knowledge is discovered, new research communities are created, and new ideas are disseminated to the institutions and people who can use them to solve the mysteries of life and health.

PREFACE

Methods that dramatically expand biological data also demand new modes of analysis and new ways to ask scientific questions. - Harold Varmus, NIH Director

Only the most rudimentary elements of biomedicine and computation were known in 1950. Development of the essential ideas and the technologies to implement them began with the discovery of the DNA structure and the construction of the first practical digital computers. Although there are intellectual connections between the two fields — DNA encodes the program for life — biomedicine and computation have advanced largely independently. And both have advanced with a rapidity that is unprecedented in history. The functional capacity of computing machines has doubled every 18 months, in accord with the prediction encapsulated in Moore's Law. At the same time, the increase in known genomic sequences — information relevant to our own genetic endowment — is being submitted to GenBank at a rate of more than 5,000 sequences (over 2 million nucleotides) per day. Computation has already transformed industrial society; a comparable biotechnological transformation is clearly on the horizon. Yet only in the last few years has it become clear that those two exponentially growing areas are now actually converging.

That convergence is already obvious in modern medicine. Medical diagnosis has been revolutionized by a suite of modern clinical-imaging methods including computed-axial tomography (CAT scans), magnetic resonance imaging (MRI), and ultrasonography. Each of them is fundamentally a computational method. The rate of their development has generally been limited by the availability of affordable computation capacity; the physical methods and concepts were waiting for computation to catch up. In the basic science of genomics, the acquisition and analysis of genomic DNA sequence has computation at its heart. Without highly capable computers, algorithms, and software, DNA sequences would have little practical value, even if we could determine them without computation. Another obvious example of the convergence is protein structure determination: The rise of crystallographic and magnetic resonance methods is bound to Moore's Law. Today even the rudimentary visualization of a protein structure requires a computer with functional capacity unknown in 1960, unaffordable in 1980, and routinely available as a commodity today.

On the horizon are developments that will require and generate more data than science is currently prepared to utilize or assimilate. For instance, nanotechnology machines that function like minuscule test tubes and minuscule pumps will allow investigators to deliver suitable dosages of medicine responding to biological signals, and capture cellular-level information about disease. The chemist's pharmacological intuition is fast being replaced by high-throughput screenings, delivered at the rate of 50,000 or more tests a week, to track the exact effect of any drug or chemical substituent. Those advances are contingent on advances in computation.

The dominant trend in biomedical science and medical practice, as in every realm of science, is the increasing value and usage of computers. Computers in our laboratories are becoming as necessary and ubiquitous as laboratory instruments. The complexity of today's problems demand that the research scientist now spends less time doing experiments and more time figuring out what they mean. The data so painstakingly extracted in past years are now, through progress in biomedicine, produced in such volumes as to require computers just to record them. The scientist spends more and more time using the computer to record, analyze, compare, and display their data to extract knowledge. Libraries are being taken over by computers as well, and clinical practice is becoming increasingly computerized-not even considering electronic patient records and billing.

Despite all those well-known realities, the convergence of computation and biomedicine is poorly reflected in our universities and schools of medicine. Biomedical computing is not a recognized discipline, and despite the extraordinary demand for people with good education in both biomedicine and computing, only a few cross-disciplinary training programs exist. Recognition of the convergence of biomedicine and computing is also quite limited among the agencies that fund biomedical and computation research. This Working Group was established to offer recommendations to remedy that situation at the National Institutes of Health.

MEETING THE POTENTIAL

Science rides on insight, that flash of understanding that suddenly gives a researcher a new way to explain a phenomenon. Insight itself comes from the hard work of gathering bits of information and ordering them, taking pieces of the puzzle and rearranging them until a new picture emerges. The process might be simple if the puzzle had a fixed number of pieces; in biology, hundreds of new puzzle pieces are added every day. To keep up with that flood of data, and to help order it, biomedical researchers are increasingly using computation. Computers are becoming puzzle-assembling tools.

But the computers, algorithms, and software, let alone the support infrastructure, are not keeping up with the exponentially rising tide of data in biomedical research. There is a consensus that much of the delay is in the lack of computational expertise in the clinics and the biomedical laboratories. Biomedical researchers need to know better how to use the powerful technology that both informs and advances their work, but the time spent developing that expertise should not come at the expense of time spent focusing on basic scientific problems. Today's researchers need the option to work closely with colleagues who know the computing part of biomedical computing as well as the investigators who know the biomedical part. It is an inevitable (and welcome) mark of research progress and success that the problem space has grown too large to be tackled predominantly by lone researchers. A team might be able to turn data into databases, turn intuition into algorithms, turn processes into computer programs. It is a rare and unlikely individual, today or in the future, who can do all of those things solo at the state of the art.

For those reasons the primary recommendation of this Working Group is the establishment of National Programs of Excellence in Biomedical Computing.

With National Programs of Excellence bringing together interdisciplinary teams, researchers will be able to harness the power of tomorrow's computers by collaborating to develop mathematical models, write software, and adapt systems. Team members can cooperate on algorithm development, software development, database development, and system development. They can make computers useful research tools, from high-performance systems in biomedical laboratories to ultra-high-performance systems in national centers. Such teams can help biomedical research move to a new horizon where new paradigms, ideas, and techniques can emerge. Biomedicine needs human power to utilize the computer power.

For many biologists, however, that human power is not available, making it hard for them to use even the tools now available to them. Many are bemoaning the lack of the human resources they need to use the computational resources that could be so helpful. The situation in biomedical research is the same as the situation in other research specialties: It now takes a cadre of experts. Just as every surgeon requires a team of nurses, medical technicians, and anesthesiologists, a computational biologist requires a team of software engineers, computer technicians, and biomedically trained algorithmists to do the best work. The focus of the National Programs of Excellence in Biomedical Computing will be research; the subtext will be the opportunity to bring together related specialties and train a new generation of researchers whose skills cross-disciplinary boundaries.

The National Programs might focus on one or more of the following areas of biomedical computing: biology, medicine, algorithms, software, database research, or devices (e.g., image capture). The spectrum of research will be from the fundamental level of scientific discovery to usable tools to do science, all of which are vital to tomorrow's biomedical research.

A Program of Excellence might be cross disciplinary or focused entirely on biology or medicine; it might be cross-institutional or at a single institution, or it may stand alone; it might pinpoint a single problem in the field, or several. The distinguishing features would include:

- A range of work, from fundamental discoveries to useful tools in biomedical computing.
- A plan for disseminating the results of the research-and-development effort, so that others can take advantage of the data that is produced, the tools that are created, and the science that is discovered.
- A full menu of education, ranging from formal undergraduate and graduate programs to courses and seminars for students and working researchers, visiting-scientist programs, "total-immersion" programs, one-week or two-week accelerated-training programs, and other innovative programs to help spread the knowledge gleaned in the course of research. That training would underline the scientific effort within the Program.

National Programs of Excellence in Biomedical Computing will answer the question of who will do computation tomorrow by educating students at all levels, with an emphasis on bachelor's and master's students to fill today's critical need for people with cross-disciplinary knowledge. Programs may be housed at a university or they may be freestanding and link to several universities; they will provide some new faculty positions and integrate and coordinate existing resources. They will offer short courses to biomedical researchers, and encourage visiting scientists.

THE NATIONAL PROGRAMS OF EXCELLENCE

Computation is becoming an enabling technology for biomedicine; some of the most exciting, challenging, and hardest problems posed to computing and computational scientists are emerging from the biomedical field.

Examples of the scope of the problems (and the cognate opportunities) abound:

SURGERY

Advanced medical-imaging systems give surgeons a better view of a patient's anatomy than they can see in front of them on the operating table. With computers that create three-dimensional models of real-time MRI scans, and programs that incorporate that model into a video of the operation in progress, surgeons can more precisely cut and suture, knowing both the extent of a tumor and its relationship to adjacent healthy tissue.

In other work, researchers are exploring the use of computer models to help surgeons decide whether to recommend surgery for stenosis, the narrowing of an artery. MRIs measure the flow of blood around a blockage, but they cannot measure the pressure on artery walls. Working together, surgeons, experimentalists and theoreticians, are building mathematical models of the pressure in the artery based on fluid dynamics.

Other researchers are exploring a computer-based virtual-reality interface with tactical feedback that would allow remote control of micro-surgical tools. Although that work is still in its early stages, it might eventually allow surgeons to perform microscopic surgery with minimal invasion, checking their progress and effectiveness with remote sensing devices, and thus reducing trauma and promoting healing.

A National Program devoted to the application of computing to surgery would concentrate the skills and knowledge of a range of experts on developing the hardware and software tools that are needed to bring computing into the operating room. It would also educate and train the physicians, bioengineers, programmers, and technicians who will develop and apply the new computer-based surgical techniques.

CLINICAL PRACTICE

In the not-too-distant future, clinicians will be able to match reconstructed images of a tumor deeply hidden in the body with a genetic characterization of a tumor, correlating the tumor's growth and metastatic involvement (the microcosm of the disease) with the patient's clinical response (the macrocosm of the disease). Imaging technologies might automate tissue-pathology analysis, leading to greater diagnostic accuracy.

Such work requires basic science research to amass the baseline data that allows that kind of exciting application of computationally based clinical medicine. A National Center focused on clinical practice could coordinate that kind of research and its direct application to human health.

It is worth noting that fundamental discovery is the foundation for such advances in medicine, but because of the diversity of diseases as complex as cancer, the ultimate impact of a discovery on the treatment of human disease almost always requires studies in human populations, that is, clinical studies.

Weaknesses in computing support for clinical research — quality assurance, varying capabilities for electronic data capture, connectivity on the Internet, security and privacy of data, and high-speed communication between laboratories, to name a few — pose enormously expensive problems. This Working Group has not attempted to deal directly with those problems, but recommends that when NIH Institutes fund clinical research they be sensitive to the need for computing, connectivity, and high-speed links between laboratories studying the bases of disease.

NEUROBIOLOGY

Neurobiologists working on the brain's ability to process information are limited not by their ideas, but by the tools to create realistic models of brain function. Until recently, neurobiologists have been able to record only the activity of single cells; new technological advances allow them to record from hundreds or even thousands of cells at the same time. With that breakthrough, the focus has turned to creating the techniques that will allow monitoring and analysis of the large numbers of neurons involved in specific behaviors. The data and the computational power are available; neurobiologists need to address the bigger issue of manipulating their data. A neurobiology Program of Excellence could bring together expertise to apply the latest data-management tools to the study of how the brain controls motor movements or how it forms memories.

MEDICAL GENETICS

Geneticists are running analyses of large numbers of subjects against the enormous amounts of data now being released about the human genome, utilizing the data from hundreds of subjects and their family members to map disease genes within a region of 30-40 megabases of DNA — more than 100 megabytes of information on each person. Those analyses can take as long as six months on routine laboratory computers. To gain the advantage of a two-day turnaround on a supercomputer, geneticists must adapt their programs to the more powerful systems. Good research should not be hurried, but delaying progress because software is not available could delay the discovery of new findings, new treatments and new cures.

CLINICAL TRIALS

Much of the information that comes out of clinical trials is statistical in nature. While some statisticians have been involved in helping to interpret those results, with the vast amounts of data now being generated, the issues are becoming more interesting to statisticians as data problems. The statistical community is only now beginning to realize that it may have much to contribute. A National Program directed towards the display and understanding of high-dimensional data from clinical trials would involve statisticians, physicians, and computer scientists in the attempt to deal with specialized data of enormous complexity and size.

Such a National Program would not be strictly computational. From the statistician's perspective, some problems that are labeled computational are really problems of the analysis of complex data. That analysis requires computational support, to be sure, but the challenge is to create appropriate analytical tools, whether algorithmic or programmatic. That is certainly the case with genetic-array data on tumor cells, or pattern-recognition problems in some image reconstruction — the kinds of problems that engage clinicians as well.

RATIONAL DRUG DESIGN AT THE CELLULAR LEVEL

Biological chemists attempting to model entire cells are waiting for the data to catch up to the technology. When the human genome has been fully sequenced, with all the genes identified, biological chemists hope they can test their theories of drug activities on computer models of cells. While researchers know a great deal about drugs that simply inhibit enzymes, they are much less sure about drugs that have subtle effects on cellular function. Researchers might possibly chart the effect of drugs on genes themselves when they can model an entire cell. Microarrays and complex genomic databases might be used to help biological chemists identify drug side effects with minimal human or animal testing. Sophisticated, linked databases of chemical substances, gene and protein structures, and reaction pathways and physiological effects will be required to make that kind of drug design possible. It is part of the idea behind National Programs of Excellence to find ways to coordinate those disparate kinds of data.

CELL BIOLOGY

Why do some cells die, and others grow uncontrolled? In cells, what is aging, and what is cancer? Cell biologists believe the answer lies in the way proteins assemble in the cell. There, function seems to follow form: The shape of proteins determines what they will do. The secret of protein assemblies seems to be in the ability of adjacent proteins to pass enough information to reach a corporate consensus on action. To correlate the arrangement of the proteins with their functions, researchers need high-resolution images of protein structures, and they need to compare structures across functions. That is not a trivial task. It takes hundreds of thousands, maybe millions of cross-sections of cell structures captured by microscopy (electron, light, MRI microscopy) to create a clear picture of the structure. That work is impossible without computational tools to collect, process and interpret the data to help understand how biological systems works. A National Program might give researchers the computational equivalent of heavy machinery that they need to plow into such data-massive science. By bringing together the machinery; the people who know how to collect, curate, and manipulate that data; and the scientists who are familiar with cell biology, the NIH could move researchers forward in understanding the life cycle of the cell, and the diseases that affect it.

A COMMON FOUNDATION

Sequencing the genome, image reconstruction, the simulation of biological systems and other emerging areas have all led to increased opportunity for understanding biology while illuminating the alarming gap between the need for computation in biology and the skills and resources available to meet that need. Much of what needs to be done in this new approach to biology will have to be done by people who are currently either not drawn into biology, have little standing in biology, or whose career opportunities are better in industry or in other scholarly disciplines. The NIH should act to increase the number of people who are trained in both biology and computation, and dignify that expertise within the biomedical research community.

At the same time, the NIH needs to insure that computer power is available. While most biomedical researchers have the desktop systems they need, they do not have up-to-date local clusters, they do not have sufficient access to regional computing centers, and they do not have a viable plan for using national computing centers — particularly those that promise teraflop computers by the next century.

Biomedical computing is on a cusp. Its growth is inevitable, but the timetable is still unknown. A small push by the NIH could result in great changes in a short time. If the NIH does not act, change could take another five, ten, or twenty years.

Workforce Development

From the Principal Investigators who understand how to use computers to solve biomedical problems to the people who keep the computers running, there is a shortfall of trained, educated, competent people. The NIH needs a program of workforce development for biomedical computing that encompasses every level, from the technician to the Ph.D. The National Programs of Excellence in Biomedical Computing would provide a structure for developing expertise among biomedical researchers in using computational tools.

Today the disciplines of computer science and biology are often too far apart to help one another. A computer-science student often stops studying other sciences after freshman biology or chemistry; a biology student, even one knowledgeable about computers, may not ever have had formal computer-science classes. Biomedical computing needs a better — and more attractive — meld of those disciplines. Today computer-science students have little incentive to learn about biomedicine. The barrier is not just the rigorous demands of computer science, it is also the relative rewards: The $50,000 to $80,000 a year that professional programmers earn makes the compensation associated with many research positions in biology laughable. This situation is even more risible when one includes the reality that staff positions on NIH research grants are guaranteed for no longer than the grant award.

In the future, many biomedical scientists will have to be well educated in both biology and computer science. One-sided education will not work. The Department of Biological Structure at the University of Washington offers one of the few programs in biomedical computing. The computer-science side incorporates programming, data structures, simple computer architecture, databases, computer networks, basic artificial intelligence, knowledge representation, and qualitative modeling. On the biology side, the program emphasizes basic medical science with courses such as anatomy, histology, cell biology, biochemistry or molecular structure. Other courses provide the quantitative basis for the broad spectrum of biology, from basic mathematics through calculus, differential equations, linear algebra, and statistics.

Such cross-discipline education should be supported by the NIH grant system. Awards should be competitive with those for computer-science and physics education. Establishing such programs will not alone create an academic infrastructure for biomedical computing; research grants are needed to make a fundamental difference in academe. Grants to faculty members are more likely to change the focus of a Ph.D. program than any change in the job market for graduates.

Strong action by the NIH is required because the existing biomedical research and teaching structures of the universities and research institutions of this country inadequately value interdisciplinary efforts generally, and computation in particular. Few grant programs and fewer academic departments foster the kind of interdisciplinary work required to address biomedical challenges fully, let alone educate students about them. National Programs of Excellence would specifically include formal and informal instruction from the undergraduate through post-graduate levels, and incorporate a range of opportunities for scholars and researchers to participate.

Software Development

Biomedical computing needs software tools to take advantage of the hardware. Often that software is cobbled together by graduate students with little programming knowledge, for use by those whose expectations are bound by the immediate problem. The application may be used once, then abandoned when the problem is solved, the graduate student moves on, or the technology changes. The publication goes out, but the tools remain in the laboratory.

That system worked for years only because computing had not yet become an important tool for biologists. Now that biomedical research is more dependent on computers, the discipline cannot afford to waste the effort to produce one-off software that is used once and discarded. Software can be shared if it is correctly conceived, properly developed, and effectively promulgated. Such a process offers two benefits: Needed software will be made available, and time spent reinventing the same processes in one laboratory after another will be freed for basic research.

One important element in the system is the creation of software-development groups: software and computer engineers who can take laboratory-based software and "harden" it-standardizing it for more general use, testing it under various conditions, documenting it, supporting it, and upgrading it as technology changes. Currently the NIH generally does not support such efforts; grants from the NIH are typically available only to develop a working model, a prototype. Completing that software and distributing it is not possible under today's funding programs. It is a generally accepted rule in the software business that producing a working prototype is only 20% of the cost of making a commercial product. NIH funding mechanisms finance only that first 20%. Where software has shown itself to be valuable to a range of researchers in biomedical computing, the NIH needs to find ways to support its full development. That might be done through public-private agreements between research centers and industry, or through direct NIH funding.

Algorithms

The need for numerical computation continues to challenge the most advanced computers, so the design and application of new algorithms continue to be of major importance. Good algorithms make computers more effective. Algorithms are the mathematical expression of information in a specialized environment. They are the bridge between data and understanding.

Discovering algorithms that advance scientific knowledge requires a thorough grounding in computer science and mathematics, as well as a keen understanding of the particular problem domain. In biology, algorithm development is now done only by the most knowledgeable computational biologists, a small fraction of the Ph.D.s in the field. Yet algorithms encapsulate the hypotheses that drive science, and their development should be an integral part of biomedical-computing research. More expertise is clearly needed as biological data increase and more computational power becomes available. To put complicated biological applications on tomorrow's teraflop machines will require teams of people working for several years. Without new algorithms and software, the power of such computers will be wasted, and mid-level machines will flounder in a sea of data. Algorithm development, the process by which researchers harness computing power, is as necessary in biomedical computing as computer power. The NIH should put resources into algorithm research if it is to advance biomedical research.

However, those with a bent for mathematics and computer science and the tenacity to seek a Ph.D. now see little reward in biomedical computing. There are few academic positions in that field; research grants tend to support the biological and not the computational aspects of their work; and their salaries are based on standards in biology, not computer science. A Ph.D. in computer science or mathematics carries more prestige, offers more job options, and guarantees more money than a Ph.D. in biology. If the NIH does not act to make biomedical research more attractive to those who are knowledgeable in computational areas, as biology increasingly becomes an information science, there will not be enough people who can create algorithms for biomedical research.

Databases

Biomedical computing is entering an age where creative exploration of huge amounts of data will lay the foundation of hypotheses. Much work must still be done to collect data and to create the tools to analyze it. Bioinformatics, which provides the tools to extract and combine knowledge from isolated data, gives us new ways to think about the vast amounts of information now available. It is changing the way biologists do science. Analyzing biological, physical, and chemical data is new — mathematical biology has done that for more than a century — but because the advent of extensive databases and the tools to manipulate them gives researchers the ability to tease knowledge about living systems from complex biological data using modern computational tools. In large part because of the tools of bioinformatics, biology is becoming a data-driven science.

Researchers use bioinformatics tools to create models that help them understand data in large problem spaces — from whole systems to whole organisms. That new understanding of the data helps them form hypotheses about biological systems. Scientists whose research once encompassed a single gene or a single protein are using bioinformatics to study integrated functions among tens of thousands of genes. In a now-classic example of the changes wrought by bioinformatics, a team of scientists discovered previously unknown sets of interrelationships when they did a standard fibroblast experiment on thousands of genes instead of the handful of genes that had been studied previously. They found a system far more complex than anyone had imagined. As biomedical researchers develop ways of dealing with large data sets, they can make leaps in understand those more-complex systems.

The Human Genome Project will require tools that can handle information on three billion base pairs — DNA units. The HGP, when it is completed early in the next century, will give biology the equivalent of a periodic table of the elements for human systems. Tomorrow's researchers will be Keplers to the Tycho de Brahes who are today sequencing the human genome. But with three billion base pairs and 100,000 genes in the human genome that could be involved in disease, biomedicine needs better techniques to store and identify genes and gene groups, and better methods to analyze them.

The study of the techniques and methods that allow researchers to collect and process data toward the understanding of the life and death of organisms is the essence of bioinformatics. It incorporates database creation, curation, and access. Some of the specific problems bioinformatics researchers are facing include:

- Standards. Terminology, syntax, and semantics need to be defined and agreed upon to allow integration of data.

- Curation. Database submissions need to be checked and cross-referenced to avoid the transitive propagation of error.
- Interoperability. Data should be as consistent as possible across databases so that researchers can compare and contrast it. For instance, three genomic databases (those concerned with the genomes of yeast, flies, and mice) are jointly producing a genetic ontology so that every biological process and function common to all three organisms can be referred to with the same words. Where databases are not consistent in schema, researchers need tools that will make transparent the querying and analysis across databases.

The database issue is in part a computational issue. To store and manipulate databases that have answers to biomedical questions hidden in thousands or hundreds of thousands of data points requires a level of sophisticated manipulation that grows more difficult as the volume of data grows. Moreover, the information needs to be presented in a format that humans can use: Reducing ten million data points to ten thousand still presents more information than a human mind can encompass. Writing the software that will turn those data points into models is a conceptual challenge.

Database issues are also systems issues. Biomedical researchers increasingly need databases of images and software as well as databases of numeric data. Those databases need to be housed on computers powerful enough to manipulate all the data quickly for many researchers at the same time.

Finally, there are research and policy issues. When are specialized databases appropriate, and how is that decided? How long should they be maintained, and by whom? What standards should apply? How should they be interconnected?

The Information Storage, Curation, Analysis, and Retrieval program this Working Group has proposed would give the NIH a way to support and advance databases and database development directly, either through grants or by establishing National Programs of Excellence focused on the special problems of data and its use. It would allow the NIH to reward proposals for research aimed at gathering and testing data, not just for research intended to test hypotheses.

Infrastructure

To deal with increasing amounts of biomedical data, the research community needs access to scalable computing systems. The need for computation is growing in bioinformatics analysis as well as in molecular dynamics and bioengineering simulations. The need is growing exponentially as the data from imaging and sequencing balloon and the use of computational simulations snowballs. Computational facilities are vital as biologists tackle more and more complex problems.

Researchers who five years ago spent little time on computers report that they now spend 90% of their research time in front of their monitors. Much of that change is because of the development of important biomedical databases such as those at the National Center for Biotechnology Information. Investigators have come to depend on those databases in their work. A study late last year showed that usage is increasing at 10% to 15% a month. In 1991 there were 195 searches a day. By 1994 that had increased to 5,000 a day. Last year there were 600,000 a day. At that rate, the NCBI databases will be used more than 25 million times a day by 2002. During the same period, the amount of determined DNA sequence had increased from 71 to 217 to 2,008 million base pairs. Sequencing the human genome (three billion base pairs) is expected to be completed sometime shortly after the turn of the century.

Those large databases require that researchers have available both the hardware and the software to manipulate them, either remotely or — when the application is unique — on their desktops. They also need to handle large datasets such as those used for imaging or simulations. A 3-D image that has a resolution of 1024 by 1024 by 1024 pixels contains at least a gigabyte of data. At least eight gigabytes of data are required for an image that is 2048 by 2048 by 2048, and clinical researchers and clinicians are demanding resolution beyond what the technology can offer today.

Biologists report problems finding funds for infrastructure support to maintain the computational resources in their laboratories: network routers, file servers, printers, and other facilities that are shared among many grantees. A great need is for people with the expertise to manage those systems and tailor them for biomedical uses. Those problems are exacerbated by the rapidly growing demand for local computer clusters where researchers can quickly turn around computational problems.

Some researchers have had to find novel ways to get the computational resources they need. One team used a major corporation's computers at night and on weekends to do its protein-folding analyses. In all, they used three times the computational resources that had been awarded for all their research projects for a year. Because the computing resources were made available, they were able to try new computational experiments, with good results. Unfortunately, such public-private partnerships are hard to put together, and so most research teams make do with inadequate equipment and power.

The unrelenting pressure on computational technology is evident in the increase in the usage of the nation's high-performance computing centers. At the National Science Foundation's supercomputer centers, for instance, out of the 50 largest projects in fiscal 1998, biomolecular computation consumed more resources than any other scientific discipline. That year the supercomputing cycles doubled, yet two-thirds of the requested cycles were turned down because of lack of sufficient resources. According to the NSF, 12% of all investigators who use their supercomputer centers are in biology, and they account for 25% of all cycles — an increase of 54% from fiscal 1997 to fiscal 1998. The biologists who used the NSF supercomputers used large amounts of time, not just a few hours, suggesting that for less-intensive applications researchers were able to find mid-size facilities to meet their needs. The pool of researchers changed, too: An analysis of the projects shows a 40% turnover in users. Together those facts suggest that supercomputers are broadly needed — and used — across biological disciplines.

For most supercomputer users, the access to computing cycles is only one of the benefits provided by a supercomputer center. The strength of the National Science Foundation's supercomputer centers is as much in their support staff as in their hardware, and in the collegial interactions among supercomputer users. The opportunity to discuss problems and solutions is an important part of the centers' gestalt. Most biomedical-computing researchers who use supercomputers have no colleagues doing similar work at their own institutions; today it is only at the supercomputer centers that they find colleagues — many of them in fields like physics, chemistry, and mathematics — with whom they can discuss their approaches. (National Programs, as they are developed, will also offer opportunities for biomedical researchers to work alongside colleagues in computer-rich environments, building new communities around common interests.)

The current levels of computing bring a variety of computational-biology problems within reach. However, to systematically study those systems — to really explore phase space, to understand not only how it works, but how those systems can be manipulated — requires computation at adequate resolution for sufficiently long periods of time, and also requires large numbers of related computations. For the biomedical promise of computation to be realized, tera-scale computing must become routine.

As more powerful computing becomes routinely available, more researchers will use it because the increased computing power will open up opportunities that did not previously exist, and biomedical researchers will move to exploit those opportunities. For that reason, any attempt to predict future needs based on current usage will result in a substantial underestimate.

IMPLEMENTATION

Because of the importance of this initiative across the NIH, and because of the basic emphasis on scientific research as a means to train scientists across disciplines and provide the tools for their work in the 21st Century, funding for the four parts of the Biomedical Information Science and Technology Initiative might be shared among the Institutes. National Centers of Excellence in Biomedical Computing, in particular, are good candidates for shared funding. Their basic educational purpose should encourage institutions to provide support for National Programs associated with their campuses.

To help the reviewers and staff who will be awarding grants under this initiative, this Working Group suggests the following review criteria for National Programs of Excellence in Biomedical Computing:

- Value to the biomedical community: Will the programs provide significant advances in the selected areas of research? Will the research provide foundations or infrastructure for other research? Will the research advance human health directly or indirectly?
- Cross-disciplinary focus: While the National Programs are not required to be formally multidisciplinary, does the program take advantage of the conjoining of biomedicine and computation?
- Research results: Does the Program incorporate both fundamental discovery and the development of useful tools? Is there a viable plan for developing, refining, and applying those tools that includes contributions from software engineers and computer scientists or other appropriate collaborators? Is there evidence of widespread usefulness of those tools, with publications and patents that document usage of or pressing need for those tools?
- Dissemination of software, hardware, algorithms, or databases: Is there a plan for making tangible, useful output available to other researchers?
- A new approach: Does the National Program bring in new ideas and new personnel and resources, or is it an aggregate of existing facilities?
- Fiscal responsibility: Especially in a virtual or cross-institutional program, is there a well-defined sharing of responsibilities among the institutions so that there is a clear principle under which to assign funds (and overhead) on an annual basis?
- Training plans: Is there a full menu of education, ranging from formal undergraduate and graduate programs to courses and seminars for students and working researchers? Are there visiting scientist programs, "total immersion" programs, one to two week accelerated-training programs or other innovative programs to help spread the knowledge? How many students, post-docs, and working researchers are trained, and what is their placement after that training?
- Success indicators: Does the National Program educate people and forge tools in the process of doing basic research?

CONCLUSION

The National Programs of Excellence in Biomedical Computing and the teams they bring together are important because biomedical computing needs cross-disciplinary expertise. The result of those Programs will be individuals with broad knowledge that can be applied to biomedical issues — knowledge that incorporates the strengths of biology, computer science, and mathematics. In the short term, biomedicine will benefit from the team approach. In the long term, there will be individual biomedical researchers who can apply much of the expertise that biomedical computing needs. The Biomedical Information Science and Technology Initiative (BISTI), and particularly its National Programs of Excellence in Biomedical Computing, is a bootstrapping approach to that next level.

The Initiative will presage smaller changes, as well. NIH study sections may come to expect that a fair proportion of biomedical research will need computational resources, and may even suggest that researchers include provision of those resources in their grant applications. In academe, there inevitably will be some restructuring of academic departments of biology and biomedicine, and tenure and promotion decisions at universities may depend as much on computational achievements as on traditional biomedical research. Both changes will improve biomedical research. Biomedical computing offers promise of profound advances in understanding and improving human health. Its advent is assured: Biomedical researchers are increasingly using computers to collect, store, access, and explore new data about the human condition, and that ripple of change will soon be a tidal wave. However, although it is inevitable, the promulgation of this critical enabling technology could face delays of five to ten years without action by the NIH. These recommendations are intended to shape the socio-technical aspects of biomedical computing to realize more quickly the anticipated benefits.

Advisory Committee to the Director, NIH Working Group on Biomedical Computing

Co-Chairs:

David Botstein, Ph.D.
Professor and Chair
Department of Genetics
Stanford University Medical School
Stanford, CA 94305

Larry Smarr, Ph.D.
Director, National Center for
Supercomputing Applications
University of Illinois
Champaign, IL 61820

Biomedical Instrumentation, Imaging
David A. Agard, Ph.D.
Professor
Department of Biochemistry and Biophysics
University of California at San Francisco
San Francisco, CA 94143

Molecular Modeling and Simulation
Michael Levitt, Ph.D.
Chairman
Department of Structural Biology
Stanford University Medical Center
Stanford, CA 94305

Clinical Trials
David Harrington, Ph.D.
Professor
Department of Biostatistics
Dana-Farber Cancer Institute
Harvard University
Boston, MA 02115

Digital Library
David J. Lipman, M.D.
Director
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894

Neurosciences
Gwen Ann Jacobs, Ph.D.
Co-Director
Center for Computational Biology
Montana State University at Bozeman
Bozeman, MT 59715

Heart Modeling
Charles S. Peskin, Ph.D.
Professor
Department of Mathematics
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Surgical Decision Support
Christopher R. Johnson, Ph.D.
Director
Center for Scientific Computing and Imaging
University of Utah
Salt Lake City, UT 84112

Industry
Arthur Levinson, Ph.D.
President and Chief Executive Officer
Genentech, Inc.
So. San Francisco, CA 94080-4990

Algorithms
Hamilton O. Smith, M.D
Investigator
The Institute for Genomic Research
Rockville, MD 20850

Information-Based Biology [Bioinformatics]
Shankar Subramaniam, Ph.D.
Professor
Departments of Biochemistry, Biophysics and
Physiology
University of Illinois at Urbana-Champaign
Senior Research Scientist
National Center for Supercomputing Applications
Urbana, IL 61801

Proteins
George Rose, Ph.D.
Professor
Department of Biophysics and
Biophysical Chemistry
Johns Hopkins University School of Medicine
Baltimore, MD 21205-2196

Genomics
Gerald M. Rubin, Ph.D.
Department of Molecular and Cell Biology
University of California at Berkeley
Berkeley, CA 94720-3200

Population Genetics
M. Anne Spence, Ph.D.
Professor, Genetics
Department of Pediatrics
University of California Irvine Medical Center
Orange, CA 92868-3298

Liaison Members of the Working Group:

Robert R. Borchers, Ph.D.
Director
Division of Advanced Computational
Infrastructure and Research
National Science Foundation
Arlington, VA 22230

Alan S. Graeff
Chief Information Officer, NIH
Bethesda, MD 20892

Mary E. Clutter, Ph.D.
Assistant Director for Biological Sciences
National Science Foundation
Arlington, VA 22230

Michael L. Knotek, Ph.D.
Program Advisor for Science and Technology
Office of the Secretary of Energy
U.S. Department of Energy
Washington, DC 20585

Margaret L. Simmons, Ph.D.
Associate Director, Program Development
National Partnership for Advanced
Computational Infrastructure
San Diego Supercomputer Center
University of California at San Diego
La Jolla, CA 92093

John Toole
Deputy Director
National Center for Supercomputing Applications
University of Illinois
Champaign, IL 61820