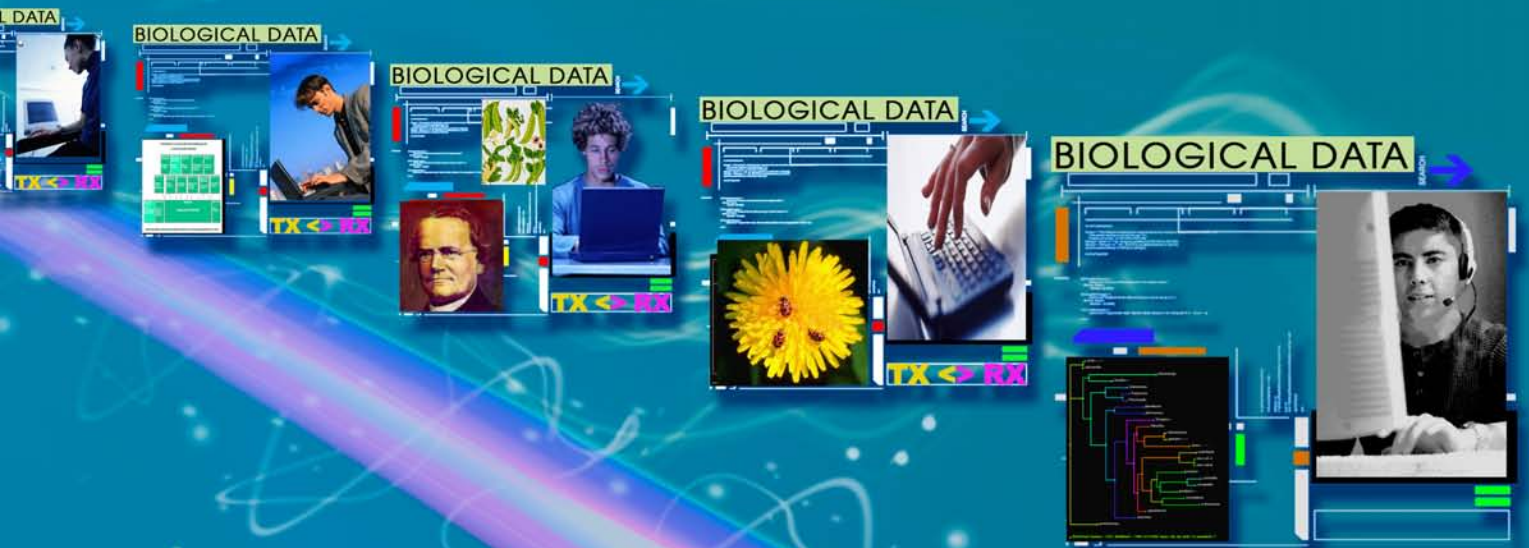


BUILDING A CYBERINFRASTRUCTURE FOR THE BIOLOGICAL SCIENCES (CIBIO)

JULY 14-15, 2003



Building a Cyberinfrastructure for the Biological Sciences (CIBIO) A BIO Advisory Committee Workshop

Report of a Workshop on Cyberinfrastructure to Develop
Recommendations for the Directorate for Biological
Sciences of the National Science Foundation

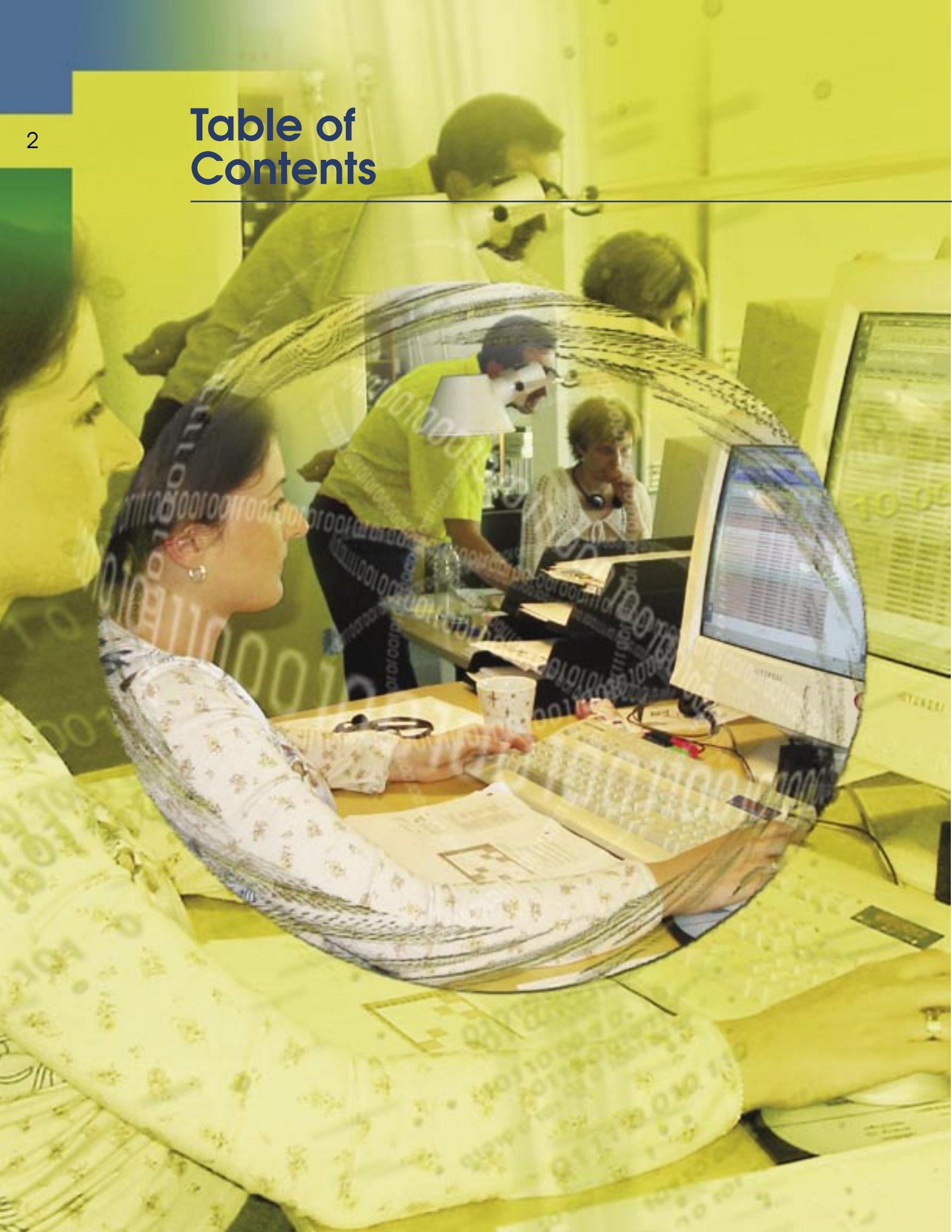
Workshop Dates: July 14-15, 2003

Convenor:

John Wooley
University of California, San Diego

jwooley@ucsd.edu

Table of Contents



Recommendations	5
The Rationale for the Workshop	7
The Workshop Report	11
Introduction: The Context for Cyberinfrastructure	11
The Unique Case for Including Biology	12
Intrinsic Aspects of the Biological Sciences	13
Multiplying Exponentials through an Extensive Partnership	14
The Essence of the Objectives for NSF BIO	14
Resource Requirements and Initial Stages of Implementation	15
Education and Training	17
Coordination and Collaborations	18
Immediate Steps for BIO in Preparing for C I Activities	18
Appendices	
I. Material Provided to Workshop Participants	21
II. Bioadvisory Committee Members as of July 2003	24
III. Central Questions for CIBIO Workshop	24
IV. Workshop Participants	25
V. References	26
VI. Schedule	27

Workshop Recommendations: A Cyberinfrastructure View

*Envisioning and Empowering Successes for
21st Century Biological Sciences*



■ The time has arrived to create a comprehensive cyber-infrastructure (CI)—the pervasive applications of all domains of scientific computing and information technology—for research and education and to do so for all of the sciences. The National Science Foundation (NSF) will lead the way, and the Directorate for Biological Sciences (BIO) should play an integral role.

■ Establishing a CI will be increasingly vital for biology. Overall, creating and sustaining a CI is as relevant and necessary for the biological sciences as for any field of science or intellectual endeavor and is likely to have a large impact.

■ In funding the advances that led to today's opportunity, BIO has made numerous ad hoc contributions and now can integrate its efforts to build the complete platforms needed for 21st Century biology. Doing so will accelerate progress.

■ The world of computing and information technology has become fully applicable to the wide range of cutting edge themes and highly complex characteristics of biological research. Both the biological sciences and the computer and information sciences have seen sustained, remarkable advances. Uniting the continuing advances in both areas, through building a CI for the biological sciences (CIBIO), could lead to discoveries not yet even imagined.

■ BIO must work within the administrative structure of NSF for implementation of a comprehensive CIBIO. The process will involve major internal NSF partnerships in addition to external partnerships with other agencies and should be fully international in scope.

■ The goals for BIO must include providing support for the complete spectrum of CIBIO, including:

- **People and Training,**
- **Instrumentation,**
- **Collaborations,**
- **Advanced Computing and Networking,**
- **Databases and Knowledge Management, and**
- **Analytical Methods (Modeling and Simulation).**

■ An implementation meeting, as well as briefings at the meetings of professional societies, will be required to enable the participation of all the biological sciences.

■ The biology community must decide how best it can interact with the computer and information science community, where and when to intersect with computational sciences and technologies, how and when the biological sciences should contribute to infrastructure projects, and how NSF BIO should partner administratively. This discussion can begin at the implementation meeting and continue at the meetings of professional societies, BIO review panel meetings, workshops and other ongoing meetings.

■ For NSF to underestimate the importance of CIBIO, or fail to provide fuel for the entire journey, would severely retard progress and would be very damaging for the entire biological sciences community, both national and international.

Essential Actions for the NSF Directorate for the Biological Sciences:

■ **Invest in People:** The next generation will be the ones who complete the journey; as always, people are our best resource.

■ **Ensure Science Pull, Technology Push:** Science should always drive decisions but be fueled by advances in technology.

■ **Stay the Course:** Partial or inadequate implementation will severely retard progress.

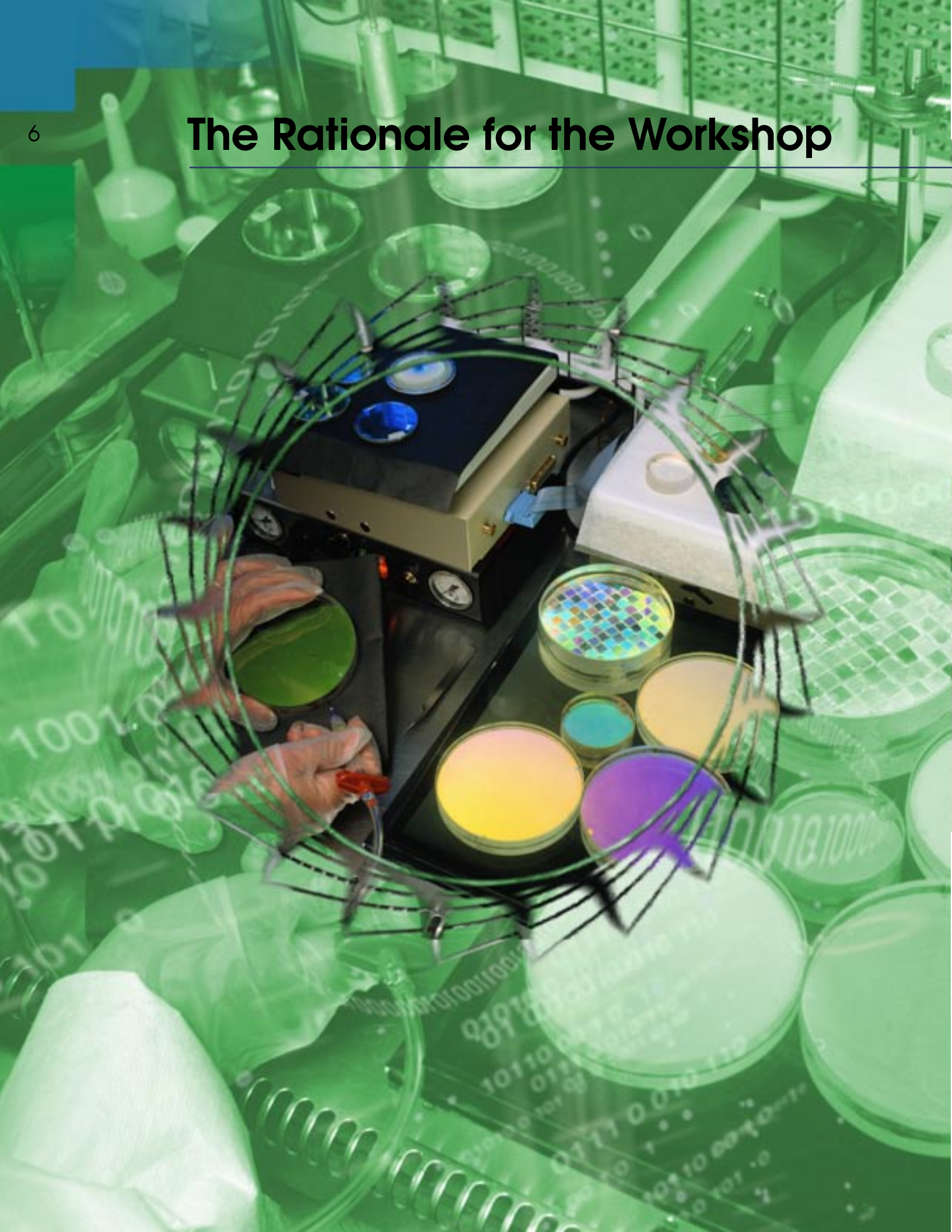
■ **Prepare for the Data Deluge:** The era of high-throughput and high information content data to address the complexity of biology has already placed the nature and rapid growth of experimental data at the forefront of challenges.

■ **Enable Science Targets of Opportunity:** Pilot projects and early successes are needed to establish a path.

■ **Select and Direct the Technology Contributions:** The challenges for the biological sciences are distinct. The entire community should be engaged in the decisions about CIBIO.

■ **Establish National and International Partnerships:** Collaborations within NSF, among the nation's science agencies and with private and international partners will be essential. The infrastructure will be universal, and universal access will be an essential enabler of progress.

The Rationale for the Workshop



The core of the long-term partnership between the National Science Foundation (NSF) and the scientific community is a shared commitment to the value of well-organized, integrated, synthesized information or knowledge, which contributes to the advancement of basic science and the use of science for public benefit. The current opportunities in the biological sciences to advance our fundamental understanding of life and to apply that understanding to societal needs—ranging from the environment to personal well being—are simply extraordinary.

We have entered an era characterized by data-intensive research observations. Collecting, managing and, in particular, connecting data from various modalities and on multiple scales—from molecules to ecosystems—is essential for turning that data into usable information. Each discipline within the biological sciences increasingly requires the tools of information technology to delve into its findings, to connect experimental observations and modeling, and to contribute to a deeper understanding. The central aspect of this challenge is why the 21st century has become widely recognized as the “Century of Biology”. The very complexity of biology means that the information technology challenges are as significant and exceptional as those in any other area of research.

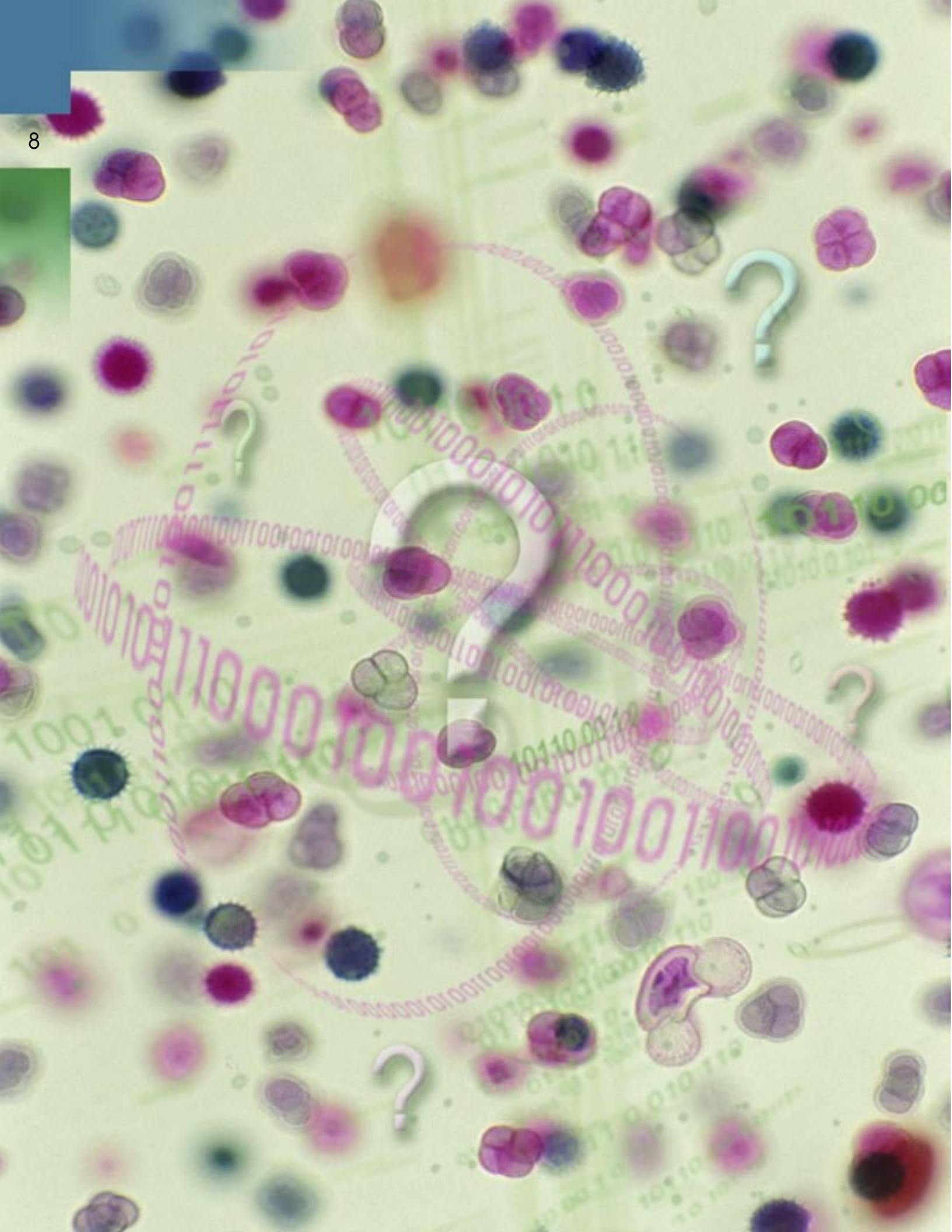
Especially over the past decade, information technology and other aspects of computing have played an increasingly important, pervasive, obvious and essential role within all of the sciences, in education and across many other elements of society and the economy. The extraordinary impact of computing and information technology, from transaction databases and wired commerce to the Internet, is apparent already in society, in which the transition to an Internet-based economy has happened in a fraction of the time it took the telephone and television to penetrate the nation’s homes. All disciplines of science and engineering face challenges inherent in studying more complex phenomena. They are deluged in information generated by contemporary instrumentation and experimental observation and need to establish data mining tools to probe these massive data sets.

The various NSF directorates have held workshops and planning processes to ascertain the opportunities and challenges and have fully endorsed the value of advanced information technology and scientific computing.

In response to the changing landscape—both the vision, or pull, from scientific objectives and the empowerment, or push, from the innovations in technology—NSF has com-

mitted to enabling a new level of creativity and innovation through building a cyberinfrastructure (CI)—the pervasive applications of all domains of scientific computing and information technology for research, education and society. The Directorate for Biological Sciences (BIO) Advisory Committee, in working with other members of the scientific community and with BIO staff to sharpen the vision for 21st Century Biology, recognized it should explore the needs of the community for CI, ascertain what specifics are required and suggest what NSF BIO should do. We recognize that, by consolidating its CI activities, BIO will be able to have a profound impact on the biology research community.

BIO is well positioned, given its history, to play a major role in this transformation of science. From the early 1980s, NSF has taken a leadership role in bringing the tools of scientific computing to the biology community. In 1984, as the High Performance Computing program began, the BIO Advisory Committee held a workshop at Airlie House to evaluate if and how biologists should use supercomputers. The answer to the first question was a resounding “yes”. The answer to the second led the instrumentation program to undertake entirely new directions and led BIO to begin what has become a long-term partnership with the NSF Directorate for Computer and Information Science and Engineering (CISE). Similarly, NSF funded the creation of the Protein Databank (PDB) decades ago, when few could see its ultimate impact. Today, the PDB, the repository for the architectures of macromolecules, has become the singularly most important information resource for structural biologists. Analyses of plant and microbial genomes also rely on computational tools, as do the ecosystem research teams involved in the Long Term Ecological Research (LTER) program. Indeed, BIO, over a decade ago at a particularly prescient point in time, began the first government programs to fund computational biology and bioinformatics (then known as database activities). In the following decade, BIO strengthened this interdisciplinary effort across all its own programs and has partnered with CISE on many of its recent initiatives. In sum, BIO already has a foundation upon which to build a CIBIO and is certainly positioned to take a leadership role in its implementation.



The revolution in the realm of computer science and information technology, which has been driven by both the academic research community and the commercial sector, coincides with the revolution in the biological sciences, and the two have now become ideal partners. Of particular importance in achieving the vision of NSF BIO is that the growing world of CI promises increasing participation by the entire spectrum of scientists supported by NSF, ranging from those at minority serving institutions to research intensive universities. When a comprehensive CI is

in place, the community of biologists in the broadest sense can participate more fully in the power and joy of discovery and the impact of its consequences.

To produce a broad vision, NSF BIO, the BIO Advisory Committee and other members of the community held a small, community workshop July 14-15, 2003, in Alexandria, Virginia. The workshop addressed the following key questions and issues:

Why is a cyberinfrastructure for biology so important?

What difference will it make?

What is its administrative scope?

What is its scientific scope?

Where are we now? What are successful examples?

Where do we have to go? What are the opportunities and challenges?

What resources and level of management do we need to get there?

What should BIO do?

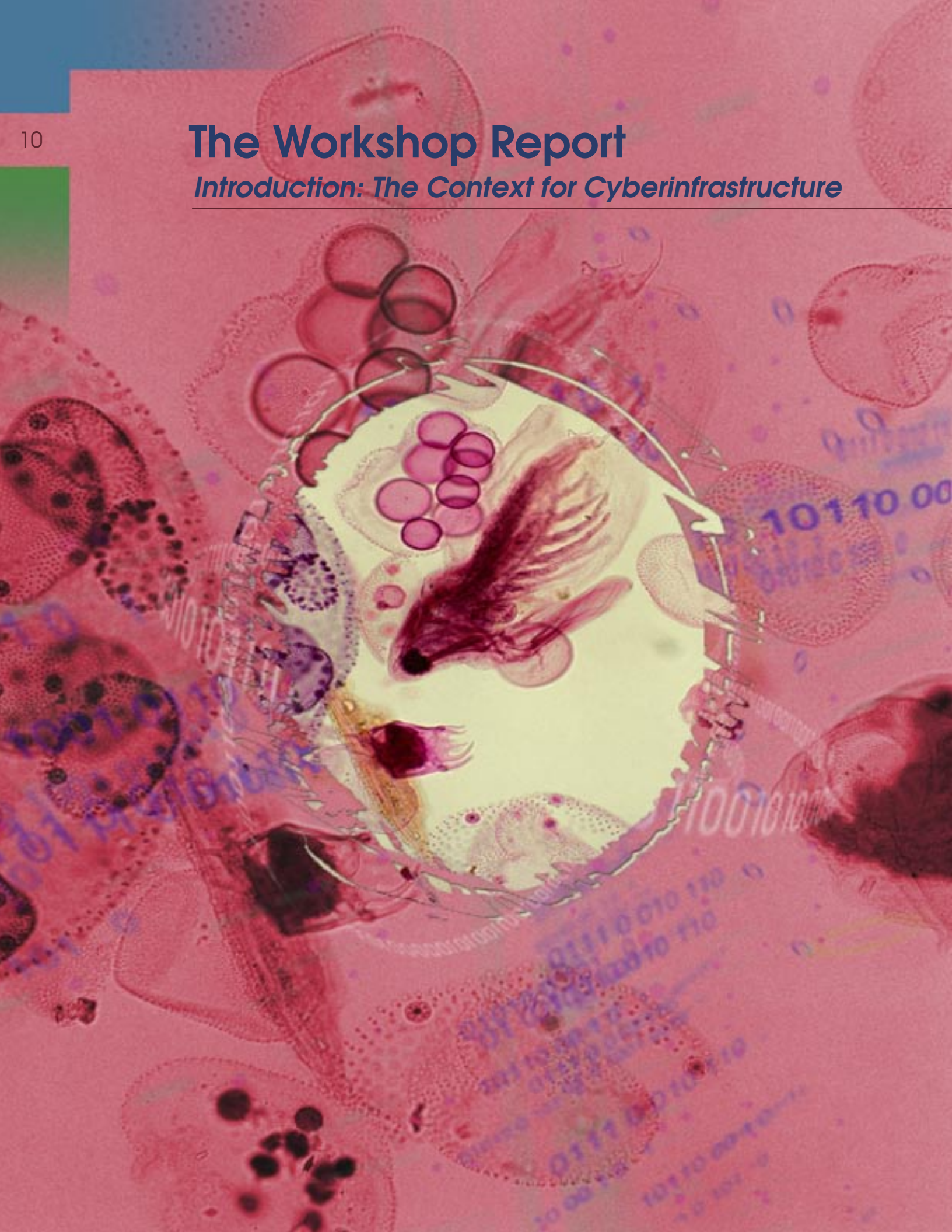
Who are internal NSF partners and external government agency partners?

What are the international considerations and impacts?

This workshop report summarizes the answers to these questions and provides advice to help establish the role of NSF BIO in building a CIBIO. The report recommends some initial actions that could provide a basis for a substantial implementation of a CI and identifies issues that should be addressed in greater detail through an implementation workshop and other planning activities.

The Workshop Report

Introduction: The Context for Cyberinfrastructure



The biological sciences are at a critical junction in their history, having advanced over several decades through the tremendous successes of “reductionist” experimentation, which carefully focused on simple systems, model organisms and biological abstractions, and models. Today, as the direct consequence of such extraordinary and even unanticipated successes, a new era of “synthesis” pervades thinking about the future of biological research, from macromolecules to ecosystems. **To enable this synthesis, biological scientists must collect, organize, analyze and comprehend unprecedented volumes of highly heterogeneous, hierarchical information obtained by different means or modalities, with different standards, widely varying kinds (types) of data and over vast scales of time, space and organizational complexity.**

The National Science Foundation (NSF) recently introduced the term “cyberinfrastructure” (CI) to describe the integrated, ubiquitous and increasingly pervasive application of scientific computing and information technology approaches, which are already changing both science and society. For example, a pervasive infrastructure arising from scientific computing and information technology will provide the circumstances and platforms to enable robust, widely distributed research teams or laboratories, user-friendly interfaces to fully integrate information from multi-component systems, and the software and hardware for advanced simulation and modeling projects that are directly and tightly coupled with experimental studies and provide interactive, iterative capacities to refine our knowledge. Such approaches are already essential requirements for many features of contemporary scientific research.

A CI will do many things, but among the most important is providing the means to establish (1) the tools for capturing,

storing and managing data; (2) the tools for organizing, finding and analyzing the data to obtain usable information; (3) the connection of experimental and theoretical analyses and their interplay with simulations and complex models based on that information; and (4) the integration of disparate aspects of that information to provide a synthesis, a knowledge repository for further considerations. The reception of the concept of CI as a maturing, philosophical and practical perspective—that is, on the profound revolution provided through today’s integration of continuing advances in scientific computing and information technology—has been truly remarkable, with the entire worldwide community of scientists joining the dialogue.

People, and their ideas and tools, are at the heart of CI. Building a fully effective CI for science and society will require educating a new generation. The technologies and the effort itself will generate new training environments and open novel options for enriching understanding of science for both technical features and for community relationships. After full implementation including the training of a new cadre of scientists, a comprehensive CI for any community will address (1) the provision of routine, remote access to instrumentation, computing cycles, data storage and other specialized resources; (2) the facilitation of novel collaborations that allow the maturation of new, robust, interdisciplinary and multidisciplinary research team efforts among the most appropriate individuals at widely separated institutions; (3) the powerful and ready access to major digital knowledge resources and other libraries of distilled information, including the scientific literature; (4) platforms or vehicles for the integration of information from multiple, highly diverse and distributed sources; (5) new training environments; and (6) other features essential to contemporary research.

The Unique Case for Including Biology

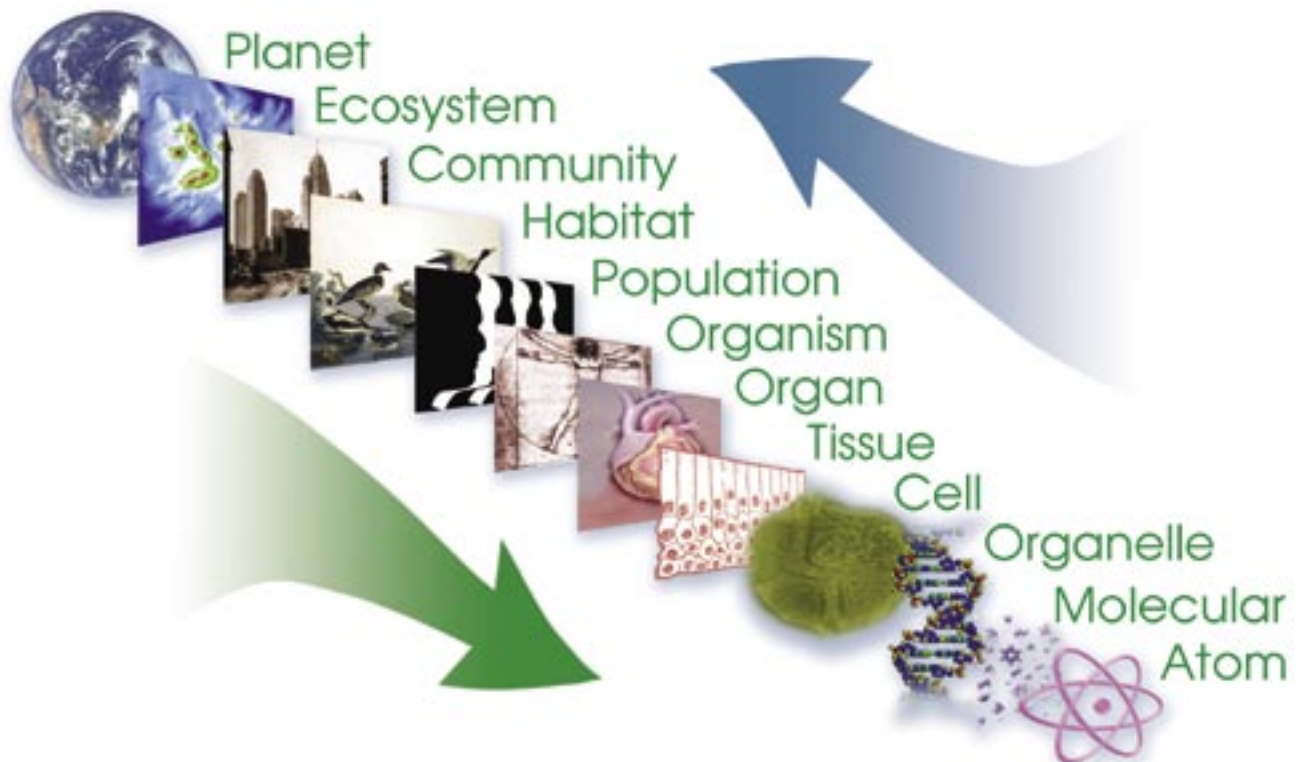
The history of the biological community and their federal sponsor, NSF's Directorate for Biological Sciences (BIO), is especially rich in prescience and sustained commitment to CI. **BIO invested early and in numerous ways** in the advancing world of information technology that is now leading to a comprehensive CI. The original investments, for example, ranged from ecology and the Long Term Ecological Research (LTER) Network to structural biology and the Protein Databank (PDB). In partnership with the NSF Directorate for Computer and Information Science and Engineering (CISE), BIO also invested in a wide range of high performance scientific computing opportunities, such as biophysical and neuroscience modeling, telescience or remote access to specialized instrumentation, and the requisite visualization, networking and database tools.

Today, there is an extraordinary opportunity for BIO to consolidate those activities and thereby build a compelling, integrated program that could only arise at NSF.

Specifically, building a CI for the biological sciences (CIBIO) requires an interface to all of the quantitative sciences as well as to computer science and engineering, and this can only happen at NSF. Already, we have seen examples—such as the National Institutes of Health (NIH) National Center for Research Resources (NCRR) cyberinfrastructure prototype and the Biomedical Information Research Network (BIRN)—in which mission agencies have recognized NSF's contribution and begun to establish CI activities to meet their needs. Numerous other examples will follow to meet the goals of those missions. Indeed, BIO's previous and continuing investments will catalyze revolutionary change, not merely incremental improvements, around the world.

CI is ideally suited for the cottage industry that is biology, due to the revolution in grid services, data integration and modern information technology. This revolution can now be coupled with the advent of a biological research

Cyberinfrastructure Enabled BioScience Research



Multidisciplinary Multidimensional Information-driven Education-oriented Internationally engaged

approach focused at a systems level that is integrative, synthetic and predictive, or what BIO calls “**21st Century Biology**”. The vantage point gained by looking at research issues in biology from a synthetic point of view, including the characterization of interacting processes and the integration of informatics, simulation and experimental analysis, represents the central engine powering the entire discipline.

Not only does 21st Century Biology absolutely require a strong CI, but also, due to its inherent complexity and the core requirement for advanced information technology, biology, more than any other scientific discipline, will drive the future of CI for all areas of science. **NSF BIO must engage fully with CISE in setting the course for CI, in establishing an architectural plan describing the specific needs of the biological sciences, in assembling the parts and building a full blown, highly empowering CI for the entire biological sciences community.**

Complementing the compelling scientific case for building the CI required for the biological sciences, there are very favorable administrative considerations in the context of NSF. Notably, the implementation by NSF of a CI in incremental fashion and tailored to each discipline’s needs offers a special opportunity—a perfect fit—for the biological sciences.

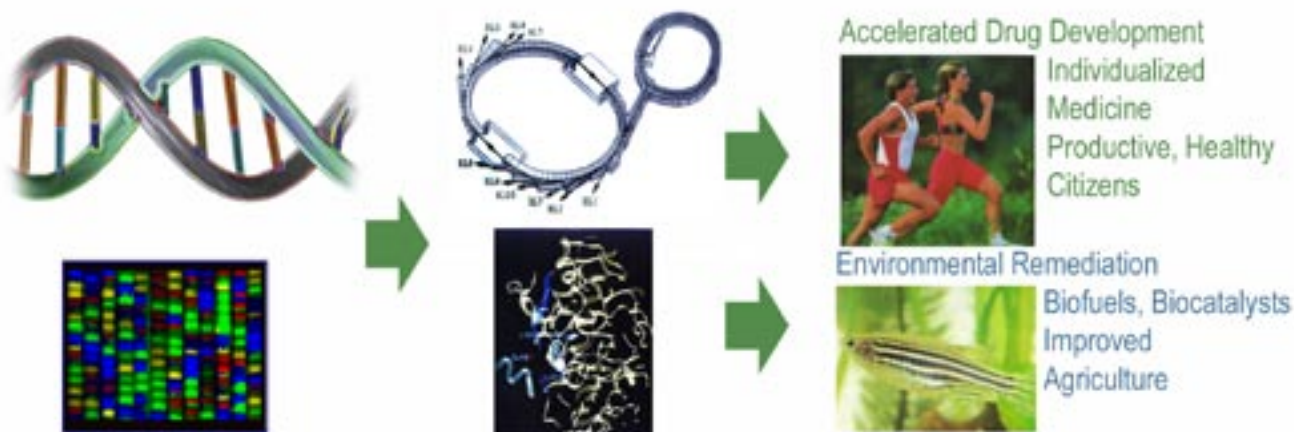
In terms of CI management, access and resources, **NSF should assign the utmost priority to BIO, the only organization positioned to lead the response of the biological sciences community.**

Intrinsic Aspects of the Biological Sciences

As our understanding of living systems increases, the inherent complexity of biology has become very obvious, almost a daunting challenge. Indeed, biology encompasses more than 20 orders of time, more than 10 orders of space and a hierarchical organizational space of enormous variety. **As calculus has been the language of the physical sciences, information technology (informatics) is becoming the language of the biological sciences.** Although biological scientists have already typically managed data sets up to the limit set by each generation’s computing parameters (cycles, storage, bandwidth), the singular nature of observations, the individuality of organisms, the typical lack of simplifying symmetries, the lack of redundancy in time and space, and the depth of detail and of intrinsic features distinguish biological data, rather than their sheer volume.

The biological sciences, in settings around the world, will remain dominated by widely distributed, individual or small

PDB & Genome enabled Biology -- Using Structure to Understand Function



DNA **Sequence** Implies **Structure** Implies **Function**

DNA Sequence Provides
Protein Sequence

Synchrotron Facilities Provide
3-D Protein Structure

Basis for 21st Century Medicine, Sustainable Development: Enhanced U.S. Competitiveness, Environmental Quality

A CYBERINFRASTRUCTURE FOR BIO IS NEEDED TO EXTRACT IMPLICIT GENOME INFORMATION

team research efforts, rather than moving to a particular focus on centralized, community facilities, as has happened for some sciences. The consequences of reaching out to the broadest range of the best performers—wherever they are—is, as a consequence, particularly important. **As telecommunication networks advance, biologists around the entire world will be able to explore and contribute to 21st Century Biology.**

At the molecular level, for example, cybertools developed to extract implicit genome information will allow biologists to understand how genes are regulated, how DNA sequences dictate protein structure and function and how genetic networks function in cellular development, differentiation, health and disease. A CIBIO must integrate the expertise and approaches of scientific computing and information technology with experimental studies at all levels, such as, on molecular machines, gene regulatory circuits, metabolic pathways, signaling networks, microbial ecology and microbial cell projects, population biology, phylogenies and ecosystems.

Multiplying Exponentials through an Extensive Partnership

An extraordinary frontier is emerging at the interface of the fields of biological science and computer and information science and engineering as a consequence of the parallel, fully comparable revolutions in these fields. Both communities, and their federal counterparts BIO and CISE, can facilitate the research agenda of the other. Twenty-first century biology absolutely requires, through the domain of scientific computing, all of the insight, expertise, methodology and technology of advanced information technology arising from the output of computer science and engineering and its vigorous interconnection with experimental research. **Indeed, only the biological sciences, over the past several decades, have seen remarkable, sustained, revolutionary increases in knowledge, understanding and applicability similar to those in the computer and information sciences.**

The Essence of the Objectives for NSF BIO

Today, the exponential increases in these two domains make them ideal partners, and the dynamics of the twin revolutions underpin the potential for unprecedented impact by building a CI for the entire biological sciences field. To build on these successes with the implementation of a CIBIO, the following

actions are essential for NSF BIO:

- Invest in People
- Ensure Science Pull, Technology Push
- Stay the Course
- Prepare for the Data Deluge
- Enable Science Targets of Opportunity
- Select and Direct Technology Contributions
- Establish National and International Partnerships

The most obvious features of 21st Century Biology are the increasing rate of data flow and, simultaneously, the highly complex nature of the data, whether obtained through conventional or automated means. Responding to this enormous challenge requires that biological scientists be able to organize that data into usable information, analyze the information to create insight and knowledge, and synthesize disparate elements of our knowledge to create a deep understanding. A passive role will not suffice when the vitality of the entire biology enterprise is involved. In other words, BIO must provide the vision for a CIBIO and not rely on technology drivers and circumstantial access. Education and the investment in people will, of necessity, include retraining, lowering the barrier for entry by senior faculty and by those from other disciplines, developing programs at all academic levels, and training and constructing stable career paths for future principle investigators and for academic professionals.

Once involved, BIO will have made a major commitment to the community and must have an effective long-range plan to sustain their efforts. The changing relationship of CISE to its high performance computing centers and the introduction of a CI process across NSF places a significant obligation on BIO to structure and maintain the role of the biological science community in the development and utilization of the scientific computing and information technology applied to biology.

Because not all subdisciplines can be simultaneously provided with a CI by BIO, selected pilot projects and areas of high impact should be the first points of effort. Strategic partnerships, discussed below, may well be needed to facilitate and accelerate implementation. The complete implementation of a CIBIO will depend on the initial choices paying off in easily demonstrated ways. Thus, **the early pilots should be selected not just for their long-term scientific contribution but also for their ability to contribute significantly in the near term**, even though many aspects of a comprehensive CIBIO will take years to develop fully,

and the impact will continue to accelerate the science for the foreseeable future.

All research communities should interoperate and work through and with CISE to absorb as many of the computational contributions from other fields as possible, rather than encouraging reinvention. Nonetheless, **BIO must also choose its own technology course**, not passively accept whatever (hardware, software, middleware) is delivered for the needs of other science domains.

The entire scientific community should be involved, even those with fewer resources and alternatives than those available within the biomolecular computing community. Scientists can now facilitate each others' progress in extraordinary ways, and, to optimize 21st Century Biology, the biological sciences need to be interconnected to the other NSF domains. **For NSF BIO to underestimate the importance of a CIBIO or to fail to provide fuel for the journey would be very damaging, perhaps catastrophic, for the community.**

CI promises to be as pervasive and central an influence as any societal revolution. Given its breadth and the potential long-term impact, several considerations are very important. First, working within partnerships and working

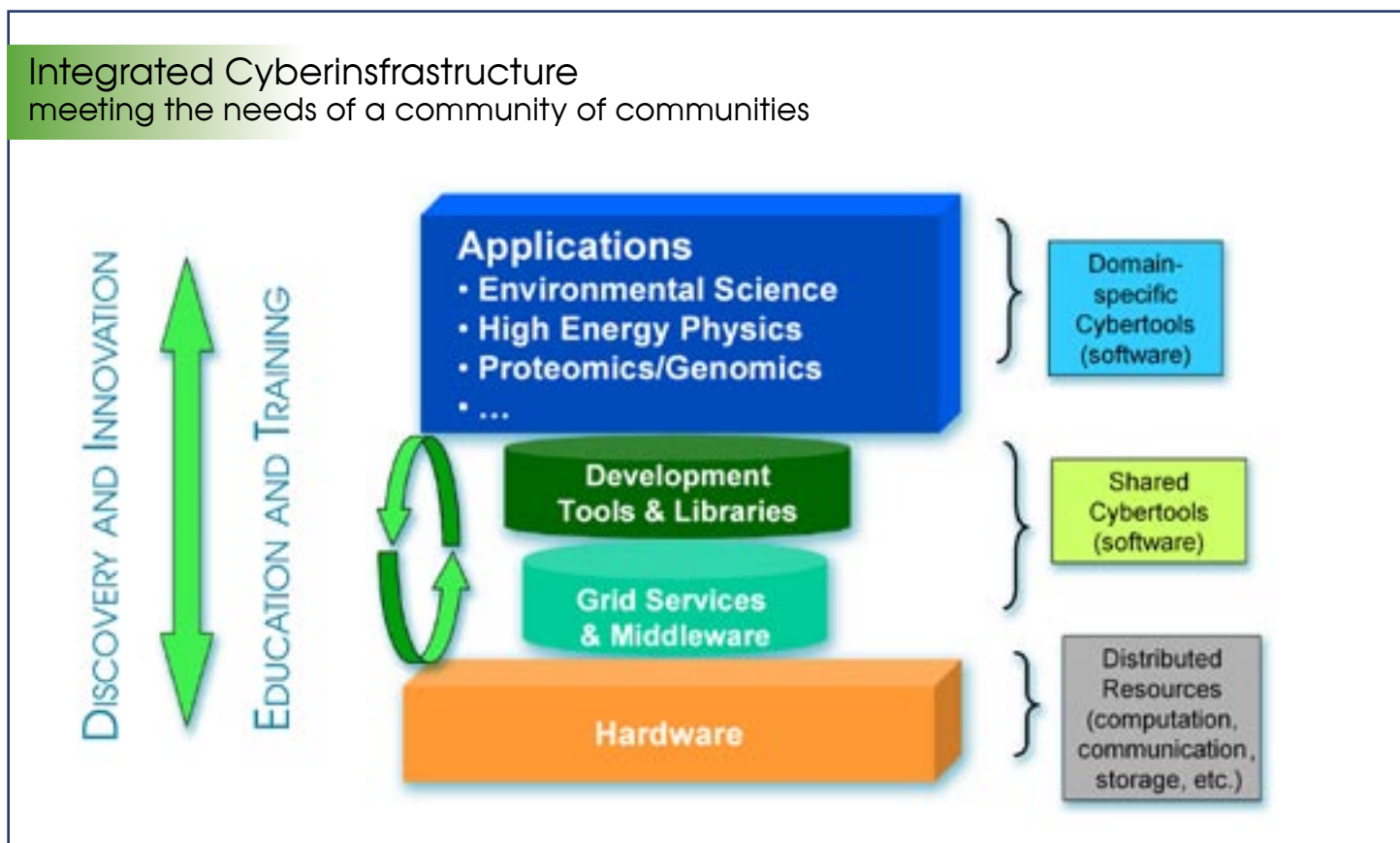
in a global context is obvious and imperative on a scientific basis. Second, these interconnections are equally obvious and imperative on a practical and administrative basis. Third, the cost of full implementation of a comprehensive CI in which the biological sciences, supported by NSF, benefit from cyber-rich environments, such as those piloted by NCRR and BIRN, and will be large as would be expected for a program of such incredible significance and applicability. The administrative scale at which NSF and BIO prepare and sustain this process will have to be well beyond any previous efforts, beyond even the Science and Technology Centers or extant Major Research Equipment programs.

Resource Requirements and Initial Stages of Implementation

Doubling the BIO budget would be justified simply to underwrite the key steps in a comprehensive CIBIO.

Although this will be a decades-long effort, NSF BIO needs to implement the beginnings of a comprehensive CI as soon as possible.

Funding increases will also be needed for the core experimental programs and their projects to permit them to exploit fully the growing CI and to build the requisite collaborations for a synthetic understanding of biology, which requires



computational expertise and the deep involvement of information technology. **In the biological sciences, database activities, modeling/simulations and theory must always be connected to experimental efforts. A balanced expansion of the portfolio will be important.** Beyond this increase in investment within BIO, major partnerships with CISE and with the other sciences will be required. The impact of these collaborations should not be underestimated, but neither should the requirement for greatly enhanced, stable funding.

BIO is already engaged in a series of extraordinary opportunities, in creating a larger scale for shared, collaborative research efforts, through activities like the Frontiers in Integrative Biological Research (FIBR) and the Plant Genome programs and the recently initiated National Ecological Observatory Network (NEON), while sustaining investment in microbial projects and the LTER network. These larger scale projects particularly require CI, with costs of comparable magnitude to the projections for the experimental research component.

BIO will have to (1) build up its own core activities at this interface (e.g., the funding for bioinformatics, biological knowledge resources, computational biology tools and collaborations on simulations/modeling) that allow it to partner with other directorates within NSF; (2) choose test beds for full implementation of CI and establish paths toward deep integration of CI into all of its communities; and (3) set a

leadership role for other agencies around the entire world, including other U.S. mission agencies. Of course, only through a decades long commitment combined with flexible, agile, engaged and proactive interactions with the entire research community and other stakeholders—i.e., with other sources of funding for infrastructure and research newly enabled by a CI—will the effort be a complete success.

Several types of early actions are needed. Implementing these requirements will be the responsibility of BIO and must be in place for effective collaborations on research frontiers with the other domains (directorates) of NSF.

The first implementation steps should be to expand BIO's extant database activities and computational modeling/simulation studies, which need central attention. Despite years of investment, many challenges remain for databases in the life sciences, both in research and implementation. For example, simulation studies could contribute considerably more across all of the biological sciences. Accelerating the introduction and expansion of tools and of the conceptual approaches provided through testing models—a prominent feature of research in the physical sciences—will require continued programmatic emphasis and commitment.

Many biologists trained in more traditional ways are just starting to recognize the opportunities made available by these new technologies, which makes a renewed and invigorated focus on training in the quantitative sciences

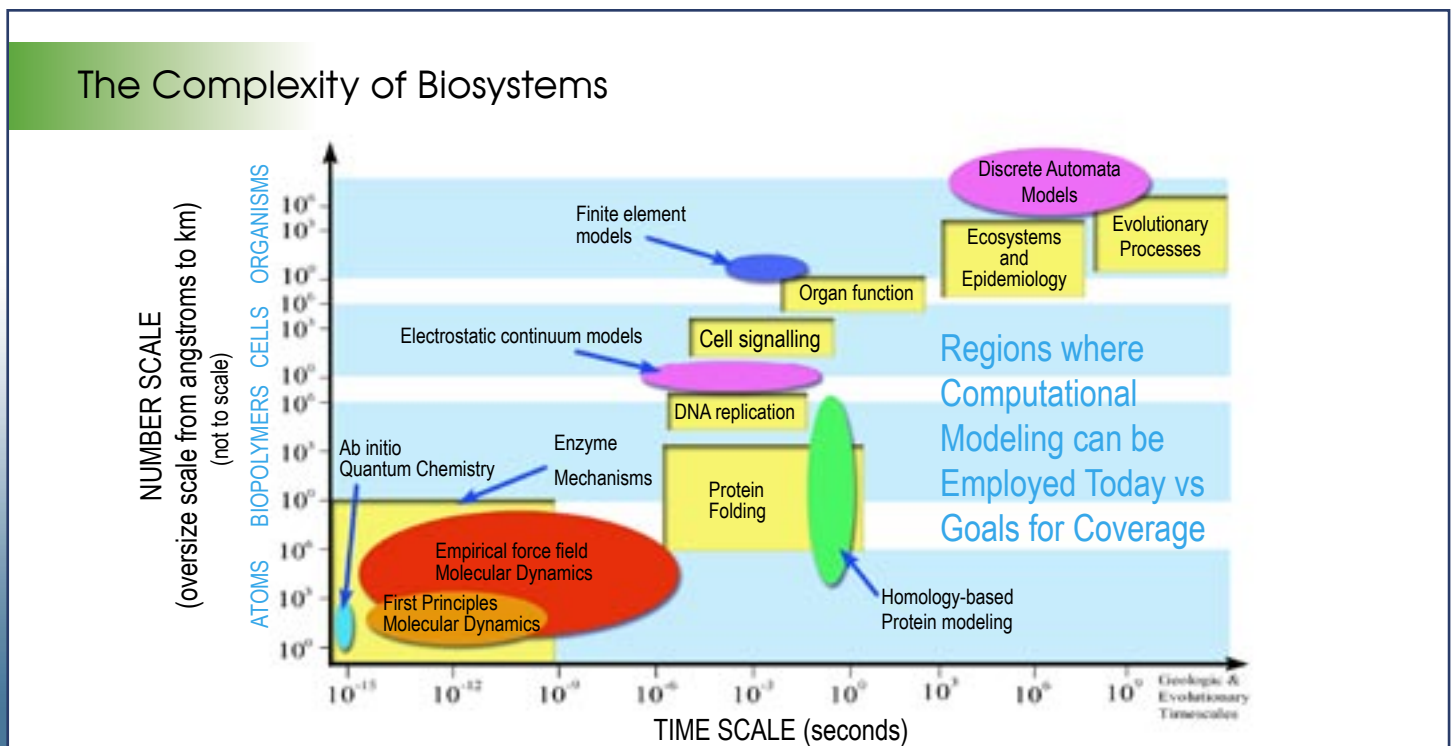


fig 4

essential for 21st Century Biology. The further development of collaborations between and among experimentalists and computational scientists should be encouraged, but the full implementation of these opportunities will require the training of a new generation of translators, of biologists who are able to understand and speak the language of the quantitative scientists well enough to choose the best collaborators and to build bridges with more traditionally trained experimentalists. Many basic requirements involve academic professionals and the use of well-documented approaches within computer and information science. Interdisciplinary training should be restored as a separate, defined program within BIO.

As noted above, the enabling and transformational impact of CI justifies—and for full implementation would require—a doubling of the NSF BIO budget, but it will also require that BIO lead a much larger effort, marshalling resources from other federal agencies and around the world to provide adequate funding and to ensure full participation by the international life sciences community. Consequently, other key, early actions are **to establish a long-range plan for sustained funding and to engage the biological sciences community** in a dialogue to ascertain implementation priorities as well as to prepare the biological scientists from around the nation to participate fully. The dialogue should begin as an open meeting that is highly interactive and inclusive in all ways. A major venue will be needed to explore all options, dig deeply into implementation features for subdisciplines and into national and international partnerships, and provide for the archiving of discussions and recommendations.

Important administrative features include the review and funding of infrastructure and establishment over time of a balance across the biology subdisciplines. Central coordination, needed for effective selection of pilot projects and coordinated efforts, will ensure balance and accelerate penetration of the benefits of modern information technology to every one of BIO's disciplines. All categories of infrastructure are increasingly important for scientific research, but CI will be particularly valuable for the biological sciences. What will be critical is to recognize that infrastructure cannot be treated the same as individual research. One cannot review infrastructure requests and plans against individual research proposals, and separate, centralized review and oversight will be needed. This situation arises because investment in infrastructure benefits all but has differences in time frame, budgets and staffing (more academic professionals). To make informed, equitable and effective judgments on behalf

of the community, a CIBIO simply can not be simultaneously considered with individual projects. At the same time, robust, rigorous peer review is essential to establish the best opportunities. Competition is also important; overlapping efforts will need to be initiated in many cases, and then the best project will ultimately be identified.

Education and Training

The educational challenges are themselves vast and will require an expansion of existing programs and, possibly, the creation of new ones. CI will dramatically alter how education is conducted—the means for training and transferring knowledge—and its full implementation and utilization will require a new cadre of scientists able to understand both the computing and biology fields, to recognize important biological problems, to recognize what computational tools are required, and capable of being a translator or communicator between more traditionally-trained biologists and their collaborators, computational scientists who will be just as traditionally-trained. These requirements are universal; that is, BIO should work with NSF's International Programs (INT) and with international agencies to encourage innovation and sustain the excitement beyond national boundaries. The technology itself will change all levels of education, and BIO should coordinate with the Directorate for Education and Human Resources (EHR) and the other research directorates. A simple example—beyond the graphical nature of knowledge representation and interactive media as a teaching vehicle—is remote learning. Interactive education and collaboration is already at work on the Internet, such as in Nobel scientists answering the queries of far away students and in the ready access with routine tools to the world's information and knowledge store.

18 Coordination and Collaborations

In numerous cases, NSF will be able to initiate an activity, but not have to plan for long-term, expanding support because another agency will ultimately adopt or extend some aspects to meet its own needs. The other agency may even sustain some or much of the original activity. But some research problems, such as ecology, plant science, phylogeny and the tree of life, and the evolution of multicellularity and of developmental processes, among others, are research domains that NSF will always own in the federal context. Besides applying CI to these areas, the overall catalysis of biology by CI will remain an NSF BIO role for the foreseeable future. NSF must ensure that once the CIBIO is put in place, **the funding plan and priority level for resource allocation must be in place to sustain the efforts** and, in particular, intellectual roadmaps linked to budgetary requirements must be developed in order to ensure that first the pilot efforts and then full implementation (each after peer review and selection of the best activities) can be funded and maintained stably in order to deliver on the promise to the community.

Immediate Steps for BIO in preparing for CI Activities

Continue to take risks. NSF BIO has the capacity to be far more adventuresome than the federal government's mission agencies. The plunge is essential for 21st Century Biology, and the adventure will benefit not only fundamental biology but also the applied biology supported by the U.S. mission agencies.

Consolidate existing activities and ensure success through adequate, sustained funding of the best activities. A continued investment in the innovative research already funded by BIO is, of course, necessary. The investment needs to be reviewed in context, not in individual programs, as discussed elsewhere in this report. Infrastructure needs its own review process. In addition, a very serious, central issue that urgently must be addressed before bringing an implementation plan to the community is what are the commitments for sustaining the infrastructure projects over a 15-year time frame. Who actually runs a given project needs to be determined by peer review, and investigators and teams will compete, and naturally some projects will

21st Century BIO-Cyberinfrastructure



Changing How Science Is Done

Providing the Tools to Swim in the Rapid Current of Data

change ownership. However, it will take a number of years to establish the infrastructure, and, once established, the CIBIO must stay in place to sustain the full development of 21st Century Biology.

A CIBIO will, of course, require attention and support for the indefinite future, but no matter what implementations are established at the start, there will be a need—in addition to routine community involvement and active dialogue on the best mechanisms to drive research—to establish plans for a sunset review of the projects as a whole and a careful full assessment of mechanisms, progress and impacts. Furthermore, new plans should be developed and evolve as technology and biological understanding advance.

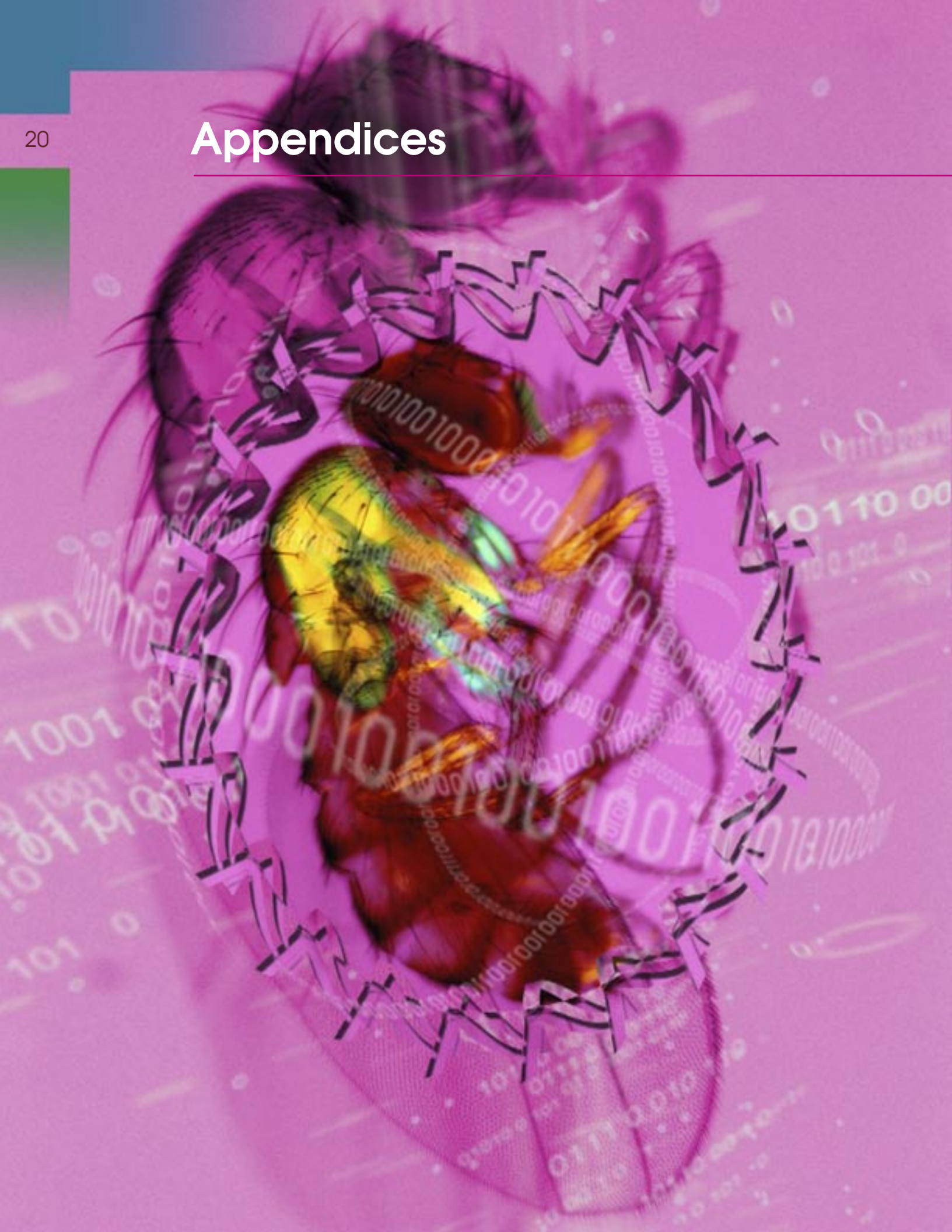
In consideration of the engineering research centers and the science and technology centers, we suggest that this broad overview occur in time to set a revised agenda in a 15-year time frame. The time frame is chosen to emphasize the deep and central **significance of stability in putting together productive teams** at multiple institutions across many disciplines. Without stability, it will be difficult to develop and clearly impossible to sustain an efficient, effective CIBIO. One vehicle to begin discussion about the necessary long-term plan would be for the BIO leadership to put together **a set of milestones or a roadmap with projections for the lifetime of CIBIO**, first of developmental and then of maintenance efforts and share that long range plan with the BIO Advisory Committee and ultimately, through numerous society meetings, with the broad community.

Prepare a five-year plan as the first step to place the biological sciences on the path toward a successful CI; that is, identify the special role that NSF BIO will play, and empower the community to look forward effectively. An implementation workshop would be an effective vehicle to review this plan, which could be developed through a leadership retreat coupled with post-panel long range planning discussions and the output from the satellite meetings held at meetings of professional societies. BIO must also make sure the plan is international.

Anticipate the accelerated pace that will characterize CI for all of the sciences. The growth path will be so rapid, the requirements so extensive and the costs so significant that a broad swath of the biological sciences research community needs to be involved in planning the implementation phase. To ensure the best ideas can compete, the entire community, more generally, needs to be forewarned and encouraged to prepare.

Obtain input from the broader biological sciences community. Given the extraordinary importance of a CIBIO, its creation will require considerable initial and ongoing input from the entire research community. One mechanism, which could prove particularly useful and cost-effective, would be to include a talk, a few talks, or a round-table discussion at subgroup meetings, i.e., at the satellite meetings typically held before biological science professional society annual meetings. BIO program officers could participate in a session in which a given discipline examines its needs and well-established investigators in the field address their vision for CI. Such discussions will also be important to ensure full participation and the consideration of the opportunities for all of the various programs and scientific domains within BIO. Thus, biological society meetings could serve to help NSF gather information and plan how to implement the funding for a CIBIO. As BIO moves toward an implementation phase, NSF could release a call for proposals aimed at scientific societies to encourage them to identify key needs and requirements and make sure that the CI activities are inclusive and bring in new people. In addition, a larger meeting to consider specific implementation across all of BIO will be very important.

Appendices



Appendix I

Material provided to workshop participants in advance of the workshop.

July 1, 2003

Subject: Building a Cyberinfrastructure for the Biological Sciences
A Workshop convened by the BIO Advisory Committee July 14-15, 2003

Dear Colleague,

Thank you for agreeing to participate in this important workshop to provide advice to the Biological Sciences Directorate of the National Science Foundation. This memo provides an overview of the workshop expectations, topics and plans.

At the interfaces with the computing and information sciences, a major revolution is taking place in science and society. This revolution, now termed cyberinfrastructure, involves the pervasive use of all of the domains of computing and information technology to facilitate research and education and the Nation's economy. NSF has made a major commitment to explore how best to implement cyberinfrastructure for the Nation. At the same time, the biological sciences are undergoing a continuing scientific revolution. Progress in the biological sciences, as our research has become data rich, will increasingly depend on the tools from computing, and biology will provide numerous research challenges for the computer and information sciences and engineering. The Advisory Committee for the Biological Sciences Directorate of NSF, or the BIOAC, and its subcommittee, 21st Century Biology, are evaluating the potential for combining the exponential advances in computing and biology. Your participation in the upcoming NSF-sponsored workshop will be very valuable for our deliberations and will lead to a report to advise the BIO Directorate.

The Advisory Committee commends NSF's Directorate for Biological Sciences (BIO) for taking such a proactive role in building Biology 21, the Biological Sciences of the 21st Century. Genome-enabled science, systems biology, high throughput biology, genomes to life, and biocomplexity in the environment, are among the many visible signs of the future. To facilitate tomorrow's advances in biology built upon the revolution in the life sciences, the biological sciences

community will increasingly depend on access to and use of information through computing technology. Biology 21, most notably, will be entirely built upon a foundation of information technology.

The world of information technology, from computing to storage to communication, is moving toward the establishment of a cyberinfrastructure for all science domains; the NSF Director, Rita Colwell, has announced this goal to be essential for the National Science Foundation, to define science and engineering research and education for this Century. As the BIO Advisory Committee works with the staff of the BIO Directorate and with other members of the scientific community to sharpen the vision for 21st Century biology, we must explore the needs of the biological science community for cyberinfrastructure, what specifics are required, and what NSF BIO should do.

The BIO Directorate is well positioned, given its history, to play a major role in this transformation of science. In 1984, as the High Performance Computing program began, BIO, as you will recall, held an Airlie House Workshop to evaluate if and if, how, biologists could use supercomputers. The answer to the first question was a responding yes, and the answer to the second, given the support from you and Jim Brown, led the Instrumentation Program to undertake entirely new directions. The singularly most important information resource for the next decade is the repository for the architectures of macromolecules, the Protein Databank (PDB). NSF started that database decades ago, when few could see its ultimate impact. The revision, updating and expansion of the PDB to serve the entire community of biologists,

as NSF has driven and enabled over the past decade, will turn out to be essential for understanding the mechanisms by which the cell's supramolecular machines work. With the confluence of PDB and the High Performance Computing and Information Technology, BIO began the first programs in government to fund computational biology and bioinformatics (then, database activities). In sum, BIO already has a foundation upon which to build a cyberinfrastructure for the biological sciences, and it is natural and even imperative that the Directorate for the Biological Sciences take a leadership role in its implementation.

The core of the partnership between NSF and the community over the past decades, I believe the core of that partnership concerns a shared commitment to the value of well-organized, integrated, synthesized information, or knowledge. Knowledge is enabling in the deepest sense for both science and society. The opportunities in the biological sciences today to advance the fundamental understanding of life and to apply that understanding to societal needs, ranging from the environment to personal well being, is simply extraordinary. A central aspect of biology today is that we have entered an era characterized by data intensive research observations. Collecting, managing, and in particular, connecting data from various modalities and on multiple scales of biological systems, from molecules to ecosystems, is essential to turn that data into information. Each biological science discipline now requires the tools of information technology to probe that information, interconnect experimental observations and modeling, and contribute to an enriched understanding or knowledge. The central aspect of the challenge is what also makes the 21st Century, the Century of the biological sciences; that is, the very complexity of biology means that the information technology challenges for achieving wisdom, acumen, for basic and applied life science research are at a level and scale at least as significant, and often more so, than other research areas.

The revolution in the computer science and information technology world, driven by the academic research community and the commercial sector, has happened at the same time as the revolution in the biological sciences, and the two are not ideal partners for each other. The frontier between the biology and computing is very exciting and essential for progress in life science research. The development of the computational grid services model, from data and information grids to compute grids to communication grids, will be especially enabling to the biological science community. Grid and cluster computing brings what were supercomputer-level resources to the entire biological sci-

ences community. Linking knowledge resources together with readily exploited portals is equally essential. Cyberinfrastructure promises democracy in action for biology, with the entire spectrum of basic scientists supported by NSF BIO, ranging from minority serving institutions to research intensive universities, participating. Even K-12 education will be facilitated by a cyberinfrastructure for biology. Our entire community, community of biologists in the broadest sense, will participate more fully in the power and joy of discovery and the impact of its consequences. The world wide web promoted a new kind of dialogue in the life sciences, in which everyone can access the same information and a high school student can send out a question concerning some specific aspect of biology and hours later a Nobel Laureate from another continent will send back the answer. The expansion of today's information technologies to create a cyberinfrastructure for biology will accelerate that access, that openness and inclusiveness.

NSF must build a cyberinfrastructure for biology to ensure this vision of democratic access. As early as a decade ago, a molecular biology Nobel Laureate proposed that access to global data would be critical for driving biology as well as for individuals sustaining competitiveness (Gilbert, 1992). Then, coupling biology and computing seemed oxymoronic to many; today, that partnership is inherently obvious and must be a central feature of NSF BIO activities.

Cyberinfrastructure for biology is important to establish the science drivers, pulling on the commercial technology to address our needs. There will continue to be a major technology push arising from the academic and industrial sectors of the computer and information sciences and engineering, and that technology push will interconnect with science pulls from across the entire domain of scientific investigation. From geochemistry to astronomy, from engineering to oceanography, the scientific communities supported by NSF recognize the opportunities from cyberinfrastructure and are rapidly addressing their own needs, through workshops and white papers. While biology writ large and BIO specifically, will be able to utilize these visions and advances, BIO must establish the path for the life sciences, to ensure our needs are met and because our community absolutely requires this infrastructure and will drive it further.

Within NSF, all Directorates have some inherent interest in partnering with BIO. CISE, MPS, ENG, OCE are obvious, and HER will have numerous intersections with BIO as well. While NSF is the only agency and BIO the only potential leader for the overall effort, nationally and globally, there

are many external partners as well. The USGS is already directly involved in biodiversity informatics and will be an important partner. NIH has recognized the importance of computational biology and bioinformatics, and will build programs for its specific needs. Many of which will naturally extend the cyberinfrastructure of basic biology to health care research. The Department of Agriculture, many of the State government programs, and numerous environmental entities outside government are also natural partners.

The nature of the maturation of the biological sciences as we implement 21st Century Biology underpins the expectations for cyberinfrastructure. Beyond the great success of reductionist approaches of the past five decades, biology is moving into an environment to consolidate these gains through information integration. The entry of biology into discovery and synthetic analysis, that is, genome-enabled biology and systems biology as well as the hardening of many biological research tools into high throughput pipelines, serves also to drive the need for cyberinfrastructure. Biodiversity and biocomplexity in the environment are two areas in BIO's domain for which active scientists and policy makers have already begun to think about the cyberinfrastructure needs, needs that have even been recognized at the White House level though PCAST. The most prominent example of an early application of cyberinfrastructure in biology is BIRN (Figures 3,4), which serves to link remote

neuroscience data to accelerate new discovery. Besides cognitive neuroscience and basic neurobiology, BIRN can be generalized for most NSF communities, and is obvious for LTER, NEON, and the continued growth in environmental biology. Similarly, there are already hundreds of molecular biology databases. Connecting them to research conducted on higher levels of biological organization is important for 21st Century Biology.

NSF BIO needs to carry out two levels of response to these incredible opportunities. For 21st Century Biology, it is essential that the BIO Advisory Committee engage to facilitate a small work group or planning session, to provide key first steps, milestones, and near and longer term objectives. Then, as the other NSF communities have done, NSF BIO, through its AC and other members of the community, should establish a small, public workshop, leading to white papers, a web presence, and an NSF BIO AC report. The planning session should be held this summer, not later than August, and the workshop in 2004.

A list of the questions to be addressed at this workshop, along with a list of attendees, is attached. (In this executive report, this is given below in Appendix III.)

Appendix II

Bio Advisory Committee Members as of July, 2003

Dr. Thomas E. Brady
Dean, College of Sciences
University of Texas, El Paso
El Paso, Texas 79968

Dr. Vicki L. Chandler
Department of Plant Sciences
University of Arizona
Tucson, Arizona 85721-0036

Dr. James P. Collins, Chair
Department of Biology
Arizona State University
Tempe, Arizona 85287

Dr. Burt D. Ensley
President and Director
NuCycle Therapy, Inc.
Hillside, New Jersey 07205

Dr. Claire M. Fraser
President and Director
The Institute for Genomic Research
Rockville, Maryland 20850

Dr. Mary Lou Guerinot
Department of Biological Sciences
Dartmouth College
Hanover, New Hampshire 03755

Dr. Lynn Jelinski
President
Sunshine Consultants, International
411 Walnut Street, #1425
Green Cove Springs, Florida 32043

Dr. Leonard Krishtalka
Director
Natural History Museum and
Biodiversity Research Center
The University of Kansas
Lawrence, Kansas 66045-2454

Dr. George L. Liggins
President and CEO
Bacton Assay Systems, Inc.
San Marcos, California 92069-1773

Dr. Cassandra Manuelito-Kerkvliet
President
Dine College
Tsaile (Navajo Nation), Arizona 86556

Dr. Jerry Melillo
Marine Biological Laboratory
Ecosystems Center
Woods Hole, Massachusetts, 02543

Dr. Norine Noonan
Dean, School of Sciences
and Mathematics
College of Charleston
Charleston, South Carolina 29424

Dr. Susan G. Stafford
Dean, College of Natural Resources
University of Minnesota, Twin Cities
St. Paul, Minnesota 55108

Dr. Larry N. Vanderhoef
Chancellor
University of California, Davis
Davis, California 95616-8558

Dr. John Wooley
Associate Vice Chancellor, Research
University of California, San Diego
La Jolla, California 92093-0043

Appendix III

July 2003 Central Workshop Questions; The Context for a CI BIO

Questions were grouped in order to focus and facilitate the breakout group discussions. Each breakout group and attendee considered all questions.

- A:** Why is cyberinfrastructure for biology so important? What difference will it make?
What is its scientific scope?
Where are we now? What are successful examples? Difficulties?
Where do we have to go? Opportunities? Challenges?
- B:** What is its technology scope? Data Intensive Bioscience, Knowledge Management?
What are the educational requirements and opportunities?
- C:** What is its administrative scope?
What do we need to get there? Funds? Management?
Balance between NSF, Institutions; Cost Sharing vs Stable Adequate Funding
What should BIO itself do now? Who will be natural collaborators?
Internal Agency Partners, External Agency Partners
Not for Profits; NGOs; International Implications
- D:** What further meetings or actions are needed in the near term? What are the first steps for BIO?

Appendix IV

Workshop Participants

Nancy Amato

Texas A & M University
amato@cs.tamu.edu
(979) 862-2275

Peter Arzberger

San Diego Supercomputer
Centerparzberger@sdsc.edu
(858) 534-5079

Paul Barber

Boston University
pbarber@bu.edu
(508) 289-7685

James Beach

University Kansas
beach@ku.edu
(785) 864-4645

Helen Berman

Rutgers University
berman@rcsb.rutgers.edu
(732) 445-4667

Emery Brown

Massachusetts General
brown@srlb.mgh.harvard.edu
(617) 726-8786

Brenda Claiborne

University of Texas San Antonio
Bclaiborne@utsa.edu
(210) 458-5487

Mike Colvin

Lawrence Livermore National Lab &
UC-Merced
Colvin2@llnl.gov
(925) 423-9177

Mark Ellisman

University of California San Diego
mhellisman@ucsd.edu
(858) 534-2251

Deborah Estrin

University of California LA
destrin@cs.ucla.edu
(310) 206-3923

Stephanie Forrest

University of New Mexico
forrest@cs.unm.edu
(505) 277-7104

Claire Fraser

The Institute for Genomics Research
cmfraser@tigr.org
(301) 838-3504

Paul Gilna

Los Alamos National Lab
pgil@lanl.gov
(505) 667-3114

Teresa Head-Gordon

University of California at Berkley
TLHead-Gordon@lbl.gov
(505) 667-3114

Gwen Jacobs

Montana State University
gwen@nervana.montana.edu
(406) 994-7334

Leonard Krishtalka

Kansas State University
krishtalk@ku.edu
(785) 864-4540

Michael Levitt

Stanford University
michael.levitt@stanford.edu
(650) 723-6800

Robert MacLeod

University of Utah
macleod@cvrti.utah.edu
(801) 587-9511

William K. Michener

Long Term Ecological Research Network
Office
wmichene@lternet.edu
(505) 272-7831

Margaret Palmer

University of Maryland
mp3@umail.umd.edu
(301) 405-3795

Phil Papadopoulos

San Diego Supercomputer Center
phil@sdsc.edu
(858) 822-2628

Jay Snoddy

University of Tennessee Oak Ridge Na-
tional Labs
snoddyj@ornl.gov
(865) 974-3466

Susan Stafford

University of Minnesota Twin Cities
stafford@umn.edu
(612) 624-1234

Lincoln Stein

Cold Spring Harbor Laboratories
lstein@cshl.org
(516) 367-8380

Russell M. Taylor II

University of North Carolina
taylorr@cs.unc.edu
(919) 962-1701

Tandy Warnow

University of Texas Austin
tandy@cs.utexas.edu
(512) 471-9724

Ross T. Whitaker

University of Utah
whitaker@cs.utah.edu
(801) 587-9549

John Wooley

University of California San Diego
jwooley@ucsd.edu
(858) 822-3630

References for CIBIO Report

There is an enormous wealth of material available—both online and in print—addressing the possibilities for 21st century science within the framework of a modern cyberinfrastructure. This literature covers a wide range of topics and disciplines. Some of the most pertinent references can be accessed by visiting CIBIO on the web at <http://research.calit2.net/cibio/>.

A key reference for building CIBIO and the context under which the efforts began is “Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure” (January 2003), which is available online from the NSF Directorate for Computer and Information Science and Engineering (CISE) at <http://www.cise.nsf.gov/sci/reports/toc.cfm>.

Appendix VI

Schedule

14 July 2003

- 9:30 AM Introduction, Welcome, Charge by Mary Clutter, Assistant Director, BIO and John Wooley, Workshop Chair, BIO AC
- 9:40 Presentation on Cyberinfrastructure (CI) as viewed by CISE by Deborah Crawford, Deputy Assistant Director, CISE
- 10:10 Review of Initial Requirements and Overview of CIBIO
- 10:20 Break
- 10:40 Examples and Models for CI / Examples from Biological Sciences
- 11:40 Overview of Grid, CI Technologies, Issues
- 12:10 Working Lunch - Overviews, General Group Discussion and Assignments for Breakout Groups
- 1:10 Breakout Group A – Science Assignments; Science Drive, Pull; Why CIBIO
- 2:40 Break
- 3:00 Breakout Group B – Application Assignments; Generic Infrastructure; Technology Push
- 4:30 Presentations from Breakout Group Chairs - Review Initial Contributions
- 5:10 Break
- 5:30 Round Table Discussion
- 6:30 Dinner
- 9:00 PM Brief Meeting of Writing Group/Steering Committee, Review Outcomes, Plans for Day 2

15 July 2003

- 9:00 – 11:15 Reform Breakout Groups - During Breakouts, Complete First Draft Major Writing Contributions, especially CIBIO and General Enablers – Connections to CISE & Technology
- 9:00 Reform Breakout Group A

28

10:00	Break
10:15	Reform Breakout Group B
11:15	Review Output for consensus
12:15	Working Lunch – Continued Discussion on Outcomes, Implications, Options for Future Meetings, Other Strategies
1:15	Final Draft of Each Section Prepared - Outline Key Points/Prepare Overview
2:30	Overview Presented to AD, BIO
3:15	Break
3:30 – 6:30	Integrated Draft Prepared

Assignments

Chair, Breakout Group A: Gwen Jacobs
Lead Scribe: Susan Stafford

Chair, Breakout Group B: Jim Beach
Lead Scribe: Paul Gilna

Workshop Chair and Editor-in-Chief: John Wooley
NSF Liaison: Judy Verbeke
Other Writing Team Members: Gwen Jacobs, Susan Stafford

Workshop Administrative Support: Amanda Voight
Follow-up Administrative Support: Joy Gorback

Web Design and Implementation: Kareem Elbayer

Printed Publication Implementation: Courtney Smoot
Copyediting: Sarah Zielinski
Design and Illustration: James Caras

