CyberInfrastructure for the Life Sciences (CILS)[1]

FINAL VERSION FOR BIO ADVISORY COMMITTEE REVEW – as of 14 June 2013

Executive Summary

The National Science Foundation Directorate for Biological Science (BIO) has a long tradition of effective investments in cyberinfrastructure ("computational systems, data and information management, advanced instruments, visualization environments, and people, all linked together by software and advanced networks to improve scholarly productivity and enable knowledge breakthroughs and discoveries not otherwise possible"). These investments in cyberinfrastructure tools (components of cyberinfrastructure including software and databases) have led to many transformative discoveries in the biological sciences. BIO investments in cyberinfrastructure can be mapped along axes of scale (small to large), scope (specific to general), and stage in the CI development lifecycle (supporting new innovation to sustaining established and mature tools). BIO core program awards are likely to fall close to the origin (circumscribed in scale, scope, and focused on new innovation rather than sustaining established tools). Awards that are larger, more general, and more focused on sustaining established tools are likely to be supported through co-funding with other Directorates. BIO investments in cyberinfrastructure in FY12 exceeded $90M, with investments well aligned with five areas identified as funding priorities by BIO:

- *Understanding the Brain*
- *Understanding Biological Diversity*
- *Interactions of the Earth, its Climate and the Biosphere*
- *Phenomics: Genotype to Phenotype*
- *Synthetic Biology*

The next step in the long history of BIO's strategic initiatives in cyberinfrastructure is called CILS – CyberInfrastructure for the Life Sciences (CILS). CILS is a strategic framework for BIO cyberinfrastructure investments and co-funding with other NSF Directorates, developed by the NSF BIO Directorate in consultation with the NSF BIO Advisory Committee and other subunits of NSF. CILS is a plan that will guide BIO investments in cyberinfrastructure (CI) for the next several years. It will also inform NSF BIO's engagement in the NSF-wide CyberInfrastructure Framework for 21st Century Science and Engineering (CIF21). Strategic goals for CILS are:

1) **Ensure the preservation of important biological data.** Ensure that important biological data are preserved in interoperable formats. Ensure that biologists have tools for preservation of important data and tools to analyze and draw insight from those data. Promote the setting of community-specific data standards by biological research communities, with a goal of technical and biological interoperability. (This goal, and the goal that follows, will aid the BIO Directorate and NSF as a whole in fulfilling NSF goals for providing public access to data and to research results).

2) **Develop the necessary tools to use and analyze biological data including:**
   -Visualization and knowledge representation tools
   -New tools to meet currently unmet needs and future needs
   Ensure that the biological research community of today and tomorrow has the tools needed to analyze, understand, and make use of biological data to create biological understanding, support or correct theories, and enable new discoveries. Ensure that the biological research community has access to visualization and knowledge representation tools. Deploy, deliver,

support, and, where needed, aid the creation of tools for automated feature detection in data sets so large that they are beyond the cognitive capabilities of humans. Make use of new consumer electronics tools and developments in collaboration science to implement and support rich, interactive remote collaboration within and among biological research teams, and foster efforts to create collaborative and interactive virtual communities Develop new tools to meet current needs of the biological research community – while also look ahead to the future to begin now developing the tools to meet future needs that the biological research community will face in 5 to 10 years.

3) **Facilitate development of a national cyberinfrastructure for biology (hardware, software, and people).** Facilitate development of a national cyberinfrastructure (hardware, software, and people) for biology that enables new discoveries and more efficient research. Such a BIO-centric national CI must include NSF-funded cyberinfrastructure, commercially provided CI, and CI implemented at universities and colleges with funding from sources other than the federal government. This must include interoperability among the facilities and services provided by the NSF and other federal agencies; by individual campuses, labs, and researchers; by commercial "cloud" services; and by volunteer computing virtual organizations. It must also enable the use of such facilities with robust, reliable, open software.

4) **Educate and train the new and the current generations of biologists to be capable of and comfortable with using the most advanced cyberinfrastructure.** Establish clear and attractive career plans for the essential laboratory and cyberinfrastructure professionals who will support 21st-century biological research. Develop comprehensive education and workforce programs to create communities of biologists who view advanced cyberinfrastructure as a routine tool for use in biological research.

NSF BIO investments in cyberinfrastructure, and to the greatest extent possible NSF investments in cyberinfrastructure relevant to biology, should be based on careful portfolio analysis. Investment must be balanced among areas termed in this plan "new" (funding initiatives led by BIO and perhaps co-funded by other NSF Directorates); "glue" (maintenance over time of major BIO-funded centers and establishments, and co-funding of inter- and multidisciplinary research of relevance to the biological sciences); and continuation of core BIO programs.

History suggests that the specifics of high priority biological research challenges may change on a timeline that is shorter than the NSF can effect changes in the cyberinfrastructure it funds and influences. However, BIO's five priority areas represent broad themes of interest to humankind for decades or more. Investing in cyberinfrastructure tools that support these priority areas with the best possible set of overall – and where possible integrated – solutions is the surest path to BIO and the national research community being prepared to meet specific research challenges as they arise. Implementation of a cyberinfrastructure suitable to support cutting-edge biological research must also include considerations of sustainability over time and support for users of that cyberinfrastructure – the practicing biologists. A focus on tools (e.g. software, data sources, and algorithms) in BIO's funding activities should provide the best return on investment over time as the nature of the US open research cyberinfrastructure changes (in particular, as use of cloud computing resources grows). Adoption of new CI tools by large BIO-funded centers is a mechanism though which CILS-related activities may be implemented to rapidly and effectively influence the activities of the biological research community as a whole. This approach will aid research where there is consensus within the community on important research questions and research progress is limited by the availability of cyberinfrastructure tools. This approach also enables BIO to be responsive to new, emerging major questions in biological research, because as consensus around such major questions emerges so will consensus on needs for new tools.

Carefully planned investment by BIO in cyberinfrastructure can enable new discoveries never before possible. This possibility is a result of the current rapid rate of development of biological knowledge, which stems from the ongoing development of new instruments for biological research and CI tools to support the use of such instruments. Communication and careful portfolio management across infrastructure investments within the BIO Directorate of the NSF and coordination between BIO and other units of the NSF are needed to capitalize on the rich and varied opportunities that now exist. CILS will be implemented upon the foundation of the NSF Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21). Coordination of cyberinfrastructure investments across the BIO Directorate and across NSF as a whole will be of great value to NSF BIO in its pursuit of its strategic goals, and this in turn will benefit the US biological research community. This will aid the US scientific community overall and support US innovation, discovery, and global competitiveness.

Introduction

The purpose of this document is to set out the next steps in the long history of BIO strategic initiatives: CILS – CyberInfrastructure for the Life Sciences – provides a strategic framework to guide BIO cyberinfrastructure investments and co-funding with other NSF Directorates for the next several years. This CILS plan was developed by the NSF BIO Directorate in consultation with the NSF BIO Advisory Committee and other subunits of NSF, and was approved by the NSF BIO Advisory Committee at its June 2013 meeting. CILS will advance and support the five priority areas for BIO-funded research and inform NSF BIO's involvement in engagement with the rest of NSF through the NSF-wide Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21).

BIO's Tradition of CyberInfrastructure Investments

The BIO Directorate of the National Science Foundation was supporting cyberinfrastructure for many years before the term "cyberinfrastructure" came into common use. Cyberinfrastructure is "computational systems, data and information management, advanced instruments, visualization environments, and people, all linked together by software and advanced networks to improve scholarly productivity and enable knowledge breakthroughs and discoveries not otherwise possible" [1]. NSF investment in what we now call cyberinfrastructure supporting biological research dates back to at least 1972 when CHRYSNET enabled network-based access to the Protein Data Bank [2]. The creation of open, reusable cyberinfrastructure by and for the biological research community dates back to at least 1985 and Lipman and Pearson's [3] original paper about FASTA. The NSF Biological Sciences Directorate (BIO) has maintained a well-formulated strategy for cyberinfrastructure (CI) investments since 1987 when what is now called cyberinfrastructure began appearing regularly as components of BIO grant awards.

BIO generally invests in specific tools or sets of tools, rather than building large and self-contained cyberinfrastructure systems (as does the Computer and Information Science and Engineering Directorate of NSF). Tools may be a particular piece of software, a database, an algorithm, or other component of cyberinfrastructure that aids biological research or education. As illustrated in Figure 1, the set of BIO investments in cyberinfrastructure and cyberinfrastructure tools can be well understood by considering the dimensions of scope, scale, and stage in the CI development lifecycle.

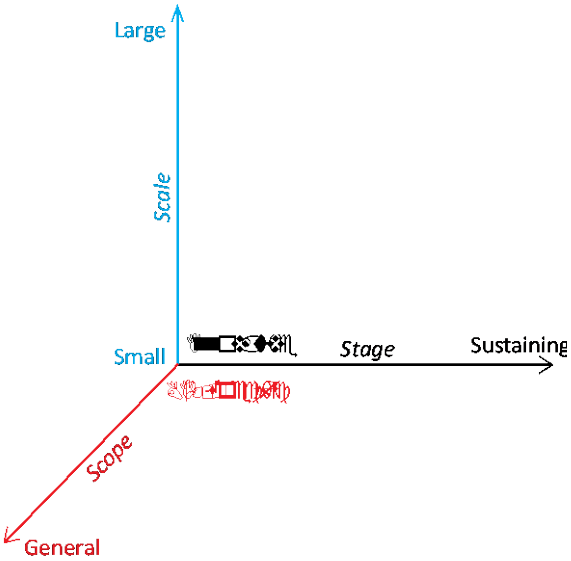

Figure 1. BIO's CI investment space. Every BIO CI investment can be defined by a combination of scale, scope, and stage in the CI development lifecycle.

- **Scope** defines the specificity of the need addressed: Is it a CI investment to solve a specific biological research question or is it a general tool to that addresses general scientific needs?
- **Scale** defines the size of the activity: Does it include modest resources to develop targeted tools or does it employ substantial resources to develop a persistent CI ecosystem?
- **Stage in the CI life cycle** defines the maturity of the project: Is it an innovative and novel idea with high risk and transformative potential or is it a vetted, established, mature resource whose impact is measured through the science it consistently enables?

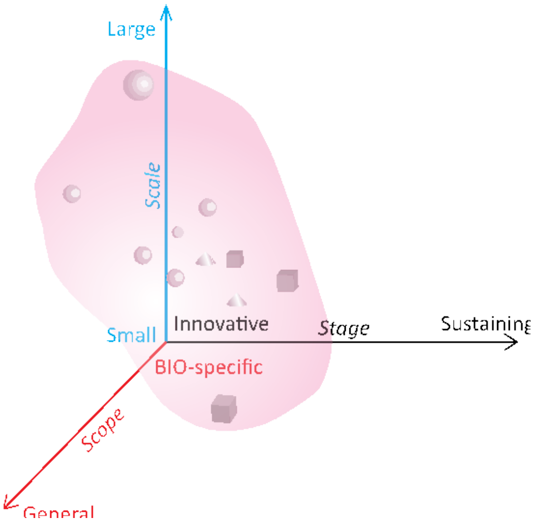

Figure 2. An example of how BIO's CI investment strategy is implemented. The ADBC initiative includes a new center-scale activity (HUB; large sphere), new coordinated data capture activities (TCNs; smaller spheres), investments in tool development through ABI (cubes), and investments in targeted digitization through CSBR (cones). The cloud surrounding the dots represents multiscale integration accomplished through such mechanisms as award conditions and supplements.

Understanding the CI needs to be met, and knowing which area of the investment space is most appropriate for meeting those needs, is fundamental to BIO's CI investment approach. Activities supported by BIO core programs typically fall near the origin – specifically, restricted in scope, and high in risk and innovation. Generalized or cross-foundational activities typically fall farther away. Every CI investment made by the NSF BIO Directorate can be mapped to these axes, even as program names change and programs come and go over time.

BIO's implementation of the Advancing the Digitization of Biodiversity Collections (ADBC) initiative demonstrates the value of planning and placing investments in cyberinfrastructure on the three axes of scope, scale, and stage in the CI lifecycle. In the case of ADBC, this model was purposefully applied to plan investments within the ADBC program. Through a series of workshops, the biodiversity community recognized a need for a centralized repository of collections-based data and identified several related CI requirements, including an organizing infrastructure, increased data capture and annotation, and novel tool and workflow development. Using these CI challenges as drivers, and considering the investment space in Figure 1, BIO determined that the needs of the biodiversity collections community would best be met by a mixture of investments including a center-like resource for CI integration, increased data capture and annotation, and directed tool development. BIO then considered mechanisms within its current investment portfolio, and concluded that some of the activities could be achieved through existing programs (e.g., Advances in BioInformatics and Collections in Support of Biological Research). New investments were also required to create a center-like organizing entity (the ADBC Home Uniting Biocollections - HUB) and to facilitate large-scale data capture (the ADBC Thematic Collection Networks - TCNs). These components within the ADBC program are depicted on the axes of specificity, scale, and stage in the CI lifecycle in Figure 2. Finally, program directors and senior managers actively engaged in "gluing" these diverse digitization-related activities together through a number of mechanisms, including: award letter conditions, supplemental funding, co-funding, and EAGER awards.

This strategy of planning, portfolio analysis, and purposeful investment across the axes of scale, scope, and stage in the CI development life cycle can be applied to all aspects of the BIO investments in cyberinfrastructure: planning individual initiatives, coordinating efforts with other Directorates to address BIO grand challenge questions, assessing BIO's portfolio balance and performance, and planning for future budget years. This approach is consistent with the NSF strategic plan. That plan sets out the structure organizing the NSF's activities in four core strategic goals – discovery, learning, research infrastructure, and stewardship [4]. This is shown diagrammatically in Figure 3, taken from that document.

For CILS to be of most value, it must be implemented in ways that

- are based on focused efforts that identify realistic, specific, and measurable CI needs;
- incorporate thoughtful categorization of those needs within the framework of BIO's CI investment space and incorporate identification of the most appropriate mechanisms for meeting those needs;
- integrate related investments through partnerships, proactive portfolio management, and open communication among all BIO stakeholders; and
- can change and adapt to changing technological and financial environments, such as increased use of cloud computing resources and cyberinfrastructure funded by sources other than the NSF.

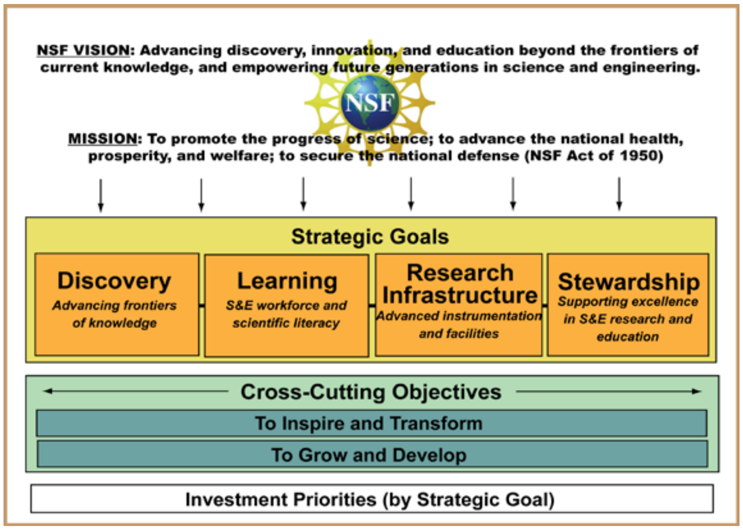
NSF VISION: Advancing discovery, innovation, and education beyond the frontiers of current knowledge, and empowering future generations in science and engineering.
NSF
MISSION: To promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense (NSF Act of 1950)
Strategic Goals
Discovery
Advancing frontiers of knowledge
Learning
S&E workforce and scientific literacy
Research Infrastructure
Advanced instrumentation and facilities
Stewardship
Supporting excellence in S&E research and education
Cross-Cutting Objectives
To Inspire and Transform
To Grow and Develop
Investment Priorities (by Strategic Goal)

Figure 3. NSF strategic goals and cross-cutting objectives.

History suggests that the specifics of high priority biological research challenges may change on a timeline that is shorter than the NSF can effect changes in the cyberinfrastructure it funds and influences. However, BIO's five priority areas represent broad themes of interest to humankind for decades or more. Investing in cyberinfrastructure tools that support these priority areas with the best possible set of overall – and where possible integrated – solutions is the surest path to BIO and the national research community being prepared to meet specific research challenges as they arise. History also suggests that the computational and storage resources within the US open research cyberinfrastructure may change relatively rapidly. In the future, for example, the use of cloud computing resources is likely to grow. By focusing on investment in tools (e.g. software, data sources, and algorithms), BIO can have the greatest impact on the research community over time, adapting to changes in underlying hardware technology while making the national cyberinfrastructure more easily usable by biologists.

BIO research priorities and NSF CIF21

There are tremendous new opportunities for breakthroughs in the biological sciences. The National Academy of Sciences 2009 report "A new biology for the 21st Century" [5] asks, "What are the implications for the life sciences research culture of a newly integrated approach to biology?" and offers the answer,

> The essence of the New Biology . . . is integration— re-integration of the many sub-disciplines of biology, and the integration into biology of physicists, chemists, computer scientists, engineers, and mathematicians to create a research community with the capacity to tackle a broad range of scientific and societal problems.

A related National Academies of Science report on "Research at the Intersection of the Physical and Life Sciences" [6] formed the basis for five research priorities of the BIO Directorate identified in the 2013 budget request:

- *Understanding The Brain*
- *Understanding Biological Diversity*
- *Interactions of the Earth, its Climate and the Biosphere*
- *Phenomics: Genotype to Phenotype*
- *Synthetic Biology*

These initiatives represent broad, durable research areas and priorities for the BIO Directorate of NSF. Discovery in each of these five areas is advanced and enabled by interdisciplinary partnerships with other NSF directorates and across disciplines. In particular, new developments in computational and data-enabled science and engineering (CDS&E) and advances in cyberinfrastructure create important new opportunities for rapid research advances now and in the very near future. Indeed, there is now the real potential to enable a revolution in integrative theories of and knowledge about biological processes that could rival or surpass the evolutionary synthesis of the 20th century [7] with profound practical impacts for the way we live and interact with our global environment.

NSF research, development, and investment in cyberinfrastructure are guided and facilitated by the Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) [8]. CIF21 represents an interdisciplinary and crosscutting NSF-wide effort to create cyberinfrastructure – infrastructure for knowledge discovery, management, and preservation. CIF21 is a portfolio of activities that integrate cyber resources and enable new research opportunities in all science and engineering fields by leveraging ongoing investments and using common approaches and components wherever feasible. CIF21 is organized into seven themes, as depicted in Figure 4. NSF principles for CIF21 include:

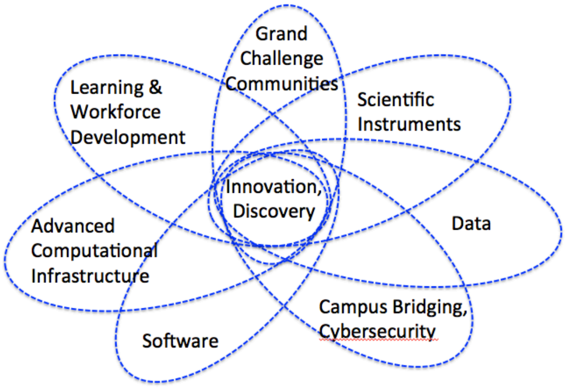
Grand Challenge Communities
Learning & Workforce Development
Scientific Instruments
Innovation, Discovery
Advanced Computational Infrastructure
Data
Campus Bridging, Cybersecurity
Software

Figure 4. Theme areas of NSF CIF21 initiative.

- CIF21 will build a national infrastructure for science and engineering
- CIF21 will leverage common methods, approaches, and applications with a focus on interoperability
- CIF21 will catalyze other CI investments across NSF
- CIF21 will be a vehicle for coordinating efforts and programs
- CIF21 will be a "force multiplier" across NSF
- CIF21 will be based upon a governance model involving every Directorate and Office

Strategic planning for CIF21 began with broad community input in the form of a set of taskforce reports [9]. NSF staff, working with community input and advice, are working to transform the recommendations from these reports into a set NSF strategic and tactical plans, including plans for Advanced Computing Infrastructure [10]; Software for Science, Engineering, and Education [11]; and this document. This CILS document will guide NSF BIO's engagement in CIF21, other emerging new research initiatives related to cyberinfrastructure (e.g. BIGDATA). This document will also inform BIO as it engages in multidisciplinary efforts with other BIO directorates where scientific opportunities and priorities will lead to engagement in the cyberinfrastructure development and implementation. (GEO, MPS, and SBR are among the other NSF Directorates where there are clear scientific opportunities that may create collaboration in cyberinfrastructure implementation).

CILS Strategic Goals

The fundamental strategy of CILS is to build upon CIF21 and BIO's existing programs and investments. In some areas, BIO will lead in identifying needs and supporting solutions specific to the biological research community. In many cases, BIO will partner with other NSF Offices and Directorates to address needs particular to the biological research community. In so doing, BIO will support and influence NSF-wide activities such as CIF21 and OneNSF. The following goals are based on identification of constraints that presently limit our advance toward solutions of Grand Challenge problems and on the presently divergent trajectories of emerging needs in the biological research community and the new cyberinfrastructure that can meet those needs.

1) **Ensure the preservation of important biological data.** Ensure that important biological data are preserved in interoperable formats. Ensure that biologists have tools to preserve important data and analyze and draw insight from those data. Implement easy-

to-use tools to store, annotate, curate, retrieve, and use those data in order to preserve information over time and to enable their use across multi- and interdisciplinary research, potentially revealing important emergent features of biological systems. These data will be of value now and generations from now. Promote the setting of community-specific data standards by biological research communities, with a goal of technical and biological interoperability [12]. Since it is impossible to permanently archive every datum collected, community input must also inform priorities for data archiving in ways that are sustainable in the long term. Data management tools must allow for massive data sets and for integration of heterogeneous data sets across taxa, time, and geographic scales, integrating with earth sciences data as needed. (This goal, and the goal that follows, will aid the BIO Directorate and NSF as a whole in fulfilling NSF goals for providing public access to data and to research results).

Rationale: Biology has always been largely about data and converting data into information and insight. Because biology has a strong historical element, biological data are often unique and often impossible to reproduce or reconstitute if lost. Integrating heterogeneous data sources to address models and analyses at increasing breadths of disciplinary, taxonomic, and geographic scales is a growing challenge as the number, diversity, and size of those data sources continue to expand. This goal, and the two that follow, will also help NSF and the BIO Directorate fulfill their goals relative to public access to research results [13].

2) **Develop the necessary tools to use and analyze biological data including:**
-Visualization and knowledge representation tools
-New tools to meet currently unmet needs and future needs

Visualization and knowledge-representation tools: With important biological data preserved on an ongoing basis and made available to the biological research community, ensure that the biological research community of today and of tomorrow has the tools needed to analyze, understand, and use biological data to create biological understanding, support or correct theories, and enable new discoveries. Ensure that the biological research community has access to visualization and knowledge-representation tools, including 3D visualization tools that will help biologists discover and convey new insights from biological data. Deploy, deliver, support, and where needed aid the creation of tools for automated feature detection in data sets so large that they are beyond human cognitive capabilities. Make use of HD and 3D consumer-electronics tools and new developments in collaboration science to implement and support rich, interactive remote collaboration within and among biological research teams, operating and interacting across geographic boundaries, across boundaries between subdisciplines within the biological sciences, and across boundaries between biology and other scientific disciplines. In so doing, foster efforts within communities to create collaborative and interactive virtual communities.

Rationale: Data sets now generated in biological experiments exceed human cognitive capabilities. Techniques such as immersive visualization and automated feature detection have long been recognized as useful tools in such situations. However, until now, high-quality visualization environments were often very expensive. Software for automated feature detection and display remains largely experimental and has been neither well developed for nor widely used by the biology research community, but advances in capabilities and integration with visualization environments could make such tools very useful to the biological research community. 3D and HD flat panel displays on the consumer-electronics are now easily added to any research office or lab (and easily within reach of many primarily educational institutions). New consumer-electronics capabilities also open up new opportunities for technology-mediated collaboration across distance. Biological research has strong traditions of benefitting from specialists, generalists, and integration across subdisciplines and with other scientific disciplines. The scale and interdisciplinary scope of new research initiatives place new demands on the technologies supporting distributed collaboration.

New tools to meet currently unmet and future needs: While developing new tools to meet current needs of the biological research community, also look ahead to the future to begin now developing the tools to meet future needs that the biological research community will face in 5 to 10 years. NSF BIO will collaborate with other NSF Directorates and other agencies so that the needs of the biological research community inform research and development in computer science and computational and data-enabled science & engineering (CDS&E).

Rationale: Current needs require attention, but the NSF and the NSF-funded research community, including computer scientists, must also consider the scale and complexity of the computational demands of cutting edge biological research in 5 or 10 years.

3) **Facilitate development of a national cyberinfrastructure for biology (hardware, software, and people).** Facilitate development of a national cyberinfrastructure (hardware, software, and people) for biology that enables new discoveries and more efficient research. Such a BIO-centric national CI must include NSF-funded cyberinfrastructure, commercially provided CI, and CI implemented at universities and colleges with funding from sources other than the federal government. This must include interoperability among the facilities and services provided by the NSF and other federal agencies; by individual campuses, labs, and researchers; by commercial “cloud” services; and by volunteer computing virtual organizations. It must also enable the use of such facilities with robust, reliable, open software. The particular cyberinfrastructure implementations and services should be focused on and shaped by needs in the five BIO grand challenge priority areas. The implementation of new usable CI tools within the BIO community should be fostered within the community by early adoption and promulgation from BIO-funded major centers, and the network of such centers should provide support and assistance for the community of NSF-funded practicing biologists.

   Rationale: Current general NSF-funded CI is often implemented in ways that represent the work patterns of physics, astronomy, and weather and climate modeling. CI implementations that are broadly useful to biologists must take into account biologists’ application needs and working styles.

4) **Educate and train the new and the current generations of biologists to be capable of and comfortable with using the most advanced cyberinfrastructure.** Foster development and training of biologists who are comfortable using the most advanced cyberinfrastructure in existence. Establish clear and attractive career plans for the essential laboratory and cyberinfrastructure professionals who will support 21st century biological research. Develop comprehensive education and workforce programs for biologists who view the most advanced cyberinfrastructure in existence as routine tools

for use in biological research; these programs should focus on educating the current generation of biologists who are thought leaders in their fields, as well as new generations of cyber-savvy biologists [12]. Engage the Directorates for Computer and Information Science and Engineering (CISE) and Education & Human Resources (EHR) to foster new and current generation computer and computational scientists who have a deep understanding of and appreciation for the biological sciences. Recognizing that the scientists of tomorrow are often the children of the US lay public of today, help the US populace as a whole understand basic biological concepts and appreciate the value of biological research through use of online information, social media, and public education efforts. Improve appreciation for biological research and at the same time garner the aid of the US populace in biological research through citizen science projects.

Rationale: Biological research will, in the future, increasingly depend upon cyberinfrastructure, making it essential to develop a strong workforce of biologists who are digital and IT experts, as well as computational scientists who have a deep understanding of biology. Furthermore, public sector funding in basic research must create and maintain clear, high-quality career opportunities if this critical niche in the 21st century workforce is to attract and retain high-quality talent. Without such talent and strong investment in CI professionals supporting biological research, it will be impossible to replicate research and analyze data at the increasing rates of production that new technologies are making possible. NSF's efforts to develop a broad societal appreciation of science in general and the biological sciences in particular are essential to the future of biological research. A US lay populace that is perhaps skeptical but willing to study and appreciate scientific research is essential so that the children of today are encouraged at home to be the scientists and cyberinfrastructure experts of tomorrow. Citizen science can be of great value in many areas of biological research, as well as a tremendous way to help the lay public learn about biology. From a better-educated lay public, we may also hope to more easily foster the development of a 21st century workforce. To attract talent from the full richness of US society, outreach and education efforts must be disseminated in ways that reach and are interesting to young people of all abilities and all racial, ethnic, and societal backgrounds.

These goals support the integrative science at the core of the "new biology" vision cited above, and will advance BIO activities in the five identified areas of strategic focus. Achievement of these goals will create new tools for citizen science, developing lay appreciation for biology as a science, and will feed into the development of a 21st century workforce.

These strategic goals should be implemented in ways that take advantage of the current patterns of investment by the NSF BIO Directorate and should focus first on serving the needs of large research centers and collaborations funded by NSF BIO. NSF-funded biological research in the US is strongly influenced and organized over time by these major centers, including field stations and environmental/ecological monitoring projects. These centers create an organizational structure within the NSF-funded biological research community and represent both major sources of data and major community efforts, supported by significant NSF investments. These centers also often represent a combination of best practices and bodies that formally or in practice set standards for the biological research community as a whole. These large centers thus form a focus area of the most critical needs for new CI capabilities, and a mechanism though which CILS-related activities may be implemented so as to rapidly and

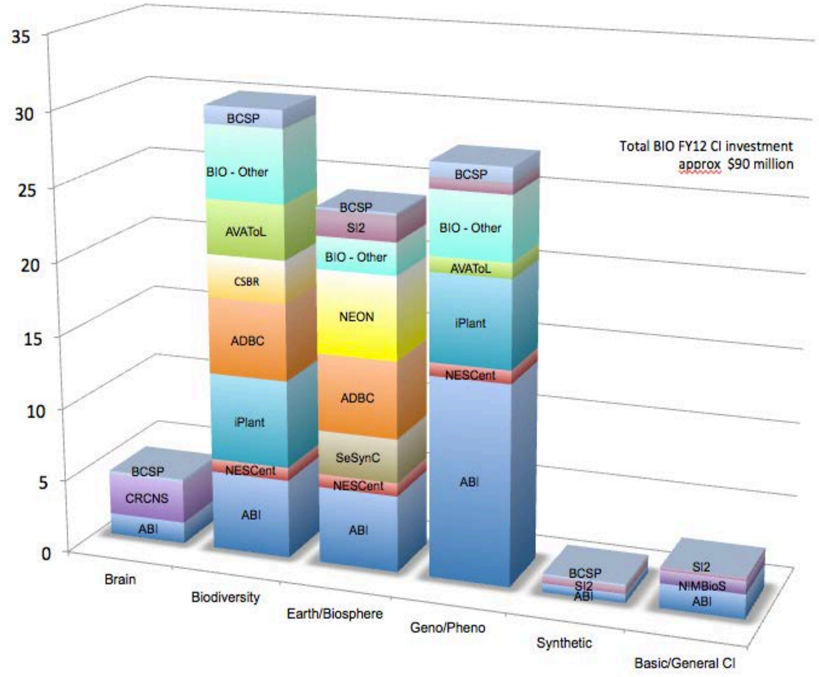

Figure 5. NSF BIO investment in CyberInfrastructure for the Life Sciences in FY12, organized by BIO's five areas of strategic investment in grand challenge problems and a sixth area of general cyberinfrastructure.

effectively influence the activities of the biological research community as a whole.[i] A focus on centers or center-like initiatives is also a practical approach to BIOs five research priority themes. It seems unlikely, for example, that we can come to a full understanding of the relationship between genome and phenome or any of the other four BIO priority theme areas within the next very few years. However, centers or center-like groups may be able to identify particular areas of opportunity within one of these themes where a breakthrough is nearly in our grasp, and the concerted application of theoretical development, experimentation, and tools can lead to major new discoveries.

Current and Planned Patterns of Investment

In FY12, the NSF BIO Directorate investment exceeded $90M in cyberinfrastructure. Current patterns of investment in cyberinfrastructure by the NSF BIO Directorate are well aligned with strategic topical areas for NSF BIO-funded research, as shown in Figure **5**.

Implementation of CILS should be informed by careful portfolio management supporting projects grouped into three categories of investment:

- **"New."** "New" refers to new funding initiatives led by BIO and perhaps co-funded by other NSF Directorates. BIO proposes specifically to fund Software Institutes – a new activity with mid- to large-scale investments creating BIO-specific integration layers and tools, informed by the needs of communities of biologists. Such integration features and tools would address major research initiatives either in BIO or shared jointly between BIO and other units. The proposed scale is proportional to the SI2 program – with 2–3

[i] This view and the CILS plan overall are consistent with the results of a workshop of BIO center principal investigators (PIs) held in summer of 2012.

concurrent centers active in any year plus support for incubation of new potential centers. Projects such as iPlant [14] and iDigBio [15] could be incorporated into this network. Other activities funded through mechanisms such as the Software Infrastructure for Sustained Innovation (SI2) program could also be incorporated, operating on timelines consistent with the SI2 program.

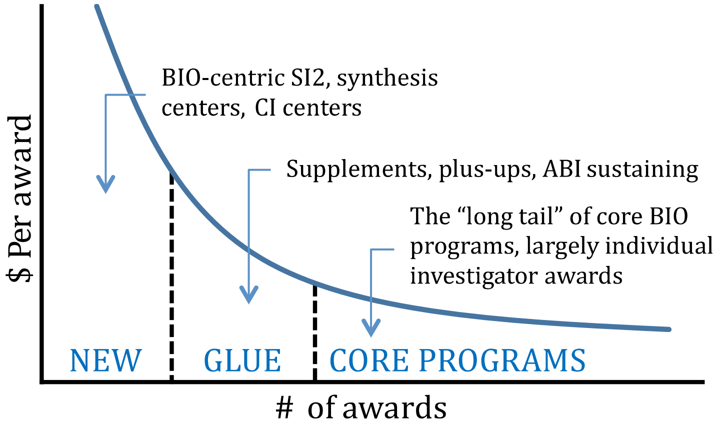

Figure 6. NSF BIO patterns of investment to implement CILS.

- **"Glue."** "Glue" refers to the major sustaining and basic infrastructure funding that BIO provides for the maintenance over time of major BIO-funded centers and the occasional establishment of new centers as appropriate over time. Glue also refers to strategies of funding the maturation and hardening of computer science innovations into usable biological research tools, the funding of communities of biologists to develop standards and priorities for development of software tools, biological cyberinfrastructure, and data integration, and co-funding of inter- and multidisciplinary research of relevance to the biological sciences. Mechanisms such as Research Coordination Networks could also be used to facilitate such activities. The award conditions and supplements that tie together the elements of the ADBC program described earlier are an example of "gluing" activities already in action to organize and deliver BIO cyberinfrastructure.
- **Core programs**. The core biological programs – individual awards – are and will remain a critical aspect of the overall BIO investment. These awards constitute the long tail of science and innovation and fund innovation and development of new, diverse, high-quality science-driven CI.

The relationship between the size of awards and the number of awards in these three categories is depicted in Figure 6. BIO's traditional focus on discovery is clearly identified in this graph in the large number of innovative, and smaller (in amount of funding) awards. This diagram also indicates more purposeful investment in a smaller number of larger (in amount of funding) awards.

The funding of cyberinfrastructure *as infrastructure* should take into account explicitly the accrued, aggregate impact of investments over time in biological cyberinfrastructure, so that the infrastructure thus funded aligns with the strategic priorities of BIO and NSF and with the needs of the NSF-funded US biological research community. BIO's approaches to CILS will incorporate service models already in use and proved by other NSF-funded infrastructure projects. BIO will work to provide a biological and BIO-CI consulting service that will accelerate BIO-funded research in the US generally. BIO will also engage with other Offices and Directorates within the NSF to support BIO's topical priorities and implementation of CILS. In particular, this document will inform NSF BIO's involvement in activities such as CIF21, multidisciplinary funding activities with other NSF directorates, and BIO's approach to emerging initiatives such as BIGDATA. Such agency-wide coordination will maximize the effectiveness of the NSF overall in fostering discovery and innovation by the US research community, will maximize the effectiveness of NSF-wide initiatives (most particularly CIF21), and will help NSF foster innovations that lead to better quality of life in the US and better US global competitiveness.

References

[1] Stewart, C.A., S. Simms, B. Plale, M. Link, D. Hancock, and G. Fox. (2010, October). What is Cyberinfrastructure? In: Proceedings of SIGUCCS 2010. (Norfolk, VA), 24–27. http://portal.acm.org/citation.cfm?doid=1878335.1878347

[2] Meyer, E.F. (1997). The First Years of the Protein Data Bank. Protein Science 6 (7): 1591–1597. doi:10.1002/pro.5560060724. PMC 2143743. PMID 9232661.

[3] Lipman, D.J., and W.R. Pearson. (1985). Rapid and Sensitive Protein Similarity Searches. Science 227 (4693): 1435–41

[4] National Science Foundation. (2012, September). Investing in America's Future: Strategic Plan FY 2006–2011. September 2006. Available from: http://www.nsf.gov/pubs/2006/nsf0648/nsf0648.jsp

[5] Committee on a New Biology for the 21st Century. (2009). Ensuring the United States Leads the Coming Biology Revolution. A new biology for the 21st century. National Academies Press.

[6] Committee on Research at the Intersection of the Physical and Life Sciences, National Research Council of the National Academies. (2010). Research at the Intersection of the Physical and Life Sciences. The National Academies Press. Washington, D.C. http://www.nap.edu/openbook.php?record_id=12809

[7] Mayr, E., and W.B. Provine (eds). (1998). The Evolutionary Synthesis. Harvard University Press.

[8] National Science Foundation. (2012). Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504730

[9] National Science Foundation Office of Cyberinfrastructure. (2011). ACCI Taskforces. http://www.nsf.gov/od/oci/taskforces/

[10] National Science Foundation. (2012). Advanced Computing Infrastructure. www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf

[11] National Science Foundation Office of Cyberinfrastructure. (2012). A Vision and Strategy for Software for Science, Engineering, and Education: Cyberinfrastructure Framework for the 21st Century. http://www.nsf.gov/pubs/2012/nsf12113/nsf12113.pdf

[12] Almeida, J.S., B.M. Tyler, C. Brewer, H. Hoekstra, G. Montelione, and J. Onuchic. (2012, April–August). Report of BIO Advisory Committee Data Working Group.

[13] National Science Foundation. (2013). National Science Foundation Collaborates with Federal Partners to Plan for Comprehensive Public Access to Research Results. http://www.nsf.gov/news/news_summ.jsp?cntn_id=127043&org=NSF&from=news

[14] iPlant. (2013). Home page. http://www.iplantcollaborative.org/

[15] iDigBio. (2013). Home page. https://www.idigbio.org/

[1] *This document is a draft so far internal to NSF, to be shared with the NSF BIO Advisory Committee for review with a request for the AC's approval at the June 27 2013 meeting of the BIO AC.. It represents the collected effort of Judy Verbeke, Peter McCartney, Anne Maglia, Reed Beaman, Julie Dickerson, Ann Haake, and Craig Stewart, and incorporates suggestions from many colleagues within the NSF (particularly colleagues in CISE). It is informed by the considerable wealth of existing NSF documents, advisory committee reports, and community-based guidance including workshop reports and peer-reviewed technical publications. This draft has been reviewed by the BIO Advisory Committee Chair B.M. Tyler and has been judged consistent with the current BIO Advisory Committee Data report. Any errors are the sole responsibility of the compiler, Craig Stewart. Public release of this document is anticipated after approval by the BIO Advisory Committee.*