caBIG™ cancer Biomedical Informatics Grid™

NATIONAL CANCER INSTITUTE

FROM *SCIENCE* VOL 308 6 MAY 2005

CYBERINFRASTRUCTURE:

EMPOWERING A "THIRD WAY" IN BIOMEDICAL RESEARCH

Kenneth H. Buetow

Biomedicine has experienced explosive growth, fueled in parts by the substantial increase of government support, continued development of the biotechnology industry, and the increasing adoption of molecular-based medicine. At its core, it is composed of fiercely independent, innovative, entrepreneurial individuals, organizations, and institutions. The field has developed unprecedented capacity to characterize biologic systems at their most fundamental levels with the use of tools and technologies almost unimaginable a generation ago. Biomedicine is at the precipice of unlocking the very essence of biologic life and enabling a new generation of medicine. Development and deployment of cyberinfrastructure may prove to be on the critical path to obtaining these goals.

The biomedical research community, dynamic and technology driven, shares its information through approaches initiated with Gutenberg's printing press and conceptually recognizable to scientists in the 18th century. Scientific findings are captured, summarized, and shared through manuscripts. The information infrastructure revolution that has transformed business and has had marked impact in other scientific disciplines has had slow uptake in biology and medicine.

Unquestionably, tremendous progress has been made in biomedicine through the application of information technology to this traditional information-sharing process. E-papers and ejournals and indices such as Pubmed all facilitate the sharing of manuscripts. Increasingly, biomedical journals require that primary data be deposited on a publisher's or investigator's Web-accessible site. In some communities, large centralized repository databases have been created for archiving biologic findings. These repositories support information retrieval through evolving current-art information technology [such as file transfer protocol (FTP) sites and Web browser portals]. For example, a recent plug-in for the Firefox Web browser permits researchers to have keyword access to these disparate data resources. However, like the communities that generate them, the infrastructure and information generated in biomedicine are largely disconnected and disjoint. Similarly, biomedical informatics, which I define as the application of information technology and its tools in biomedical disciplines *(1)*, mirrors this structure of the culture it serves: highly heterogeneous in approach, small, independent, dispersed, and fragmented.

Biomedicine at a Crossroads

The current paradigms of information sharing and resource use in biology and medicine are being challenged on several fronts. First, the success of the enterprise means that there has been a marked increase in the number of investigators, organizations, and institutions conducting biomedical research. Tracking the work and providing infrastructure to support the expansion are increasingly difficult. This expansion has resulted in a substantial number of new journals and Web sites. Although current information technology supports ready access, it does not address abstraction, integration, and interpretation of information. The diverse bioinformatics tools generated to consume and evaluate the data rarely interoperate. Commonly, the community demonstrates a willingness to share data and applications, but the number and diversity of components that must be assembled are overwhelming.

The very data generated in modern biomedicine presents a primary challenge to the researcher. Many of the new technologies used in today's research generate large volumes of rapidly expanding and ever-changing data. Although Moore's law and cheap disk space have reduced the impact of this growth, individual scientists and institutions are spending an increasing fraction of their effort and resources simply retrieving and

processing data. Biologic data represents additional challenges. To integrate biologic data, one must traverse multiple orders of magnitude of scale and complexity. Ideally, in biology one would want to move seamlessly between biologic and chemical process, organelle, cell, organ, organ system, individual, family, community, and population. The diversity of data types that are explored in biomedicine is somewhat orthogonal. Technology permits the characterization of genomic, proteomic,metabolomic, image, and other largescale characterizations.

All of the above is further confounded by the organization of biomedicine into research fields and disciplines. Such discipline focus generates an insidious challenge to information integration. Each community speaks its own scientific dialect. This community "speciation" results in reduced flow of information between disciplines, slowing the diffusion of knowledge and critical progress.

Finally, biomedicine's culture is at the nexus of a challenge faced by many other scientific fields: the need for "big" science and team science. The call for big science recognizes that many of the technology approaches required in biology and medicine are expensive, beyond the reach of individual investigators, and increasingly challenging the resource reserves of all but a few institutions. New paradigms are required to support these investigations. The push for team science also recognizes that many problems cross traditional discipline boundaries.

Cyberinfrastructure: A Third Way

A view that the current biomedical research culture is incompatible with team or big science is overly simplistic. It is clear that big science and team science will be necessary to achieve the goals of biology and medicine. However, the small, independent investigator is still the engine of innovative research. Widespread adoption of cyberinfrastructure represents an alternative in which the two approaches can be blended to create virtual team science. In so doing, the organization of biomedicine retains its entrepreneurial independent investigators whose insights and resources can be virtually joined through information technology. Big science contributes large-scale, raw material that feeds the virtual communities. Cyberinfrastructure empowers a reinvention of biomedicine without having to fundamentally change its basic culture or operational characteristics—a third way.

It is one thing to suggest that cyberinfrastructure could transform biomedicine and quite another thing to achieve this transformation. Fortunately, biomedicine can benefit from the long experience of other communities' embrace of informatics infrastructure to guide its approach. To address challenges in biomedicine, it must deliver in several key fronts. First, it must add perceivable value to the enterprise. In order to achieve widespread adoption, users must be motivated to do something different. Traditionally, this means they need to be able to do something they couldn't do without using the technology. Cyberinfrastructure shows great promise in this area because it has the ability to address the challenges of large, complex data sets. However, greater capacity may not be a sufficient driver, as demonstrated by current low penetration. Cyberinfrastructure will also need to enable new capabilities through the integration of communities and their disparate data types.

A primary lesson from other fields is that information technology has its greatest impact when it changes the way work can be performed. This may manifest itself through the apparent elimination of processing steps or the need to duplicate resources locally. Existing technologies permit the sharing and joining of common resources within virtual groups. However, the complex issues and diversity of biologic data still represent a substantial challenge to the creation of automated workflows.

Finally, the infrastructure needs to be easy to use and straightforward to implement. This requirement is more subtle than it might seem. A deeper examination raises the question, easy and straightforward to whom? Looking at the existing Internet and Web provides a useful clarification. End users consuming Internet resources through graphical user interfaces displayed through Web browsers would describe the Internet as easy to use. However, at the level of technical implementation, starting up a network that connects to the Internet and sharing information through a Web server is quite complex and beyond the skill set of an average biomedical researcher. It will be important to understand this dialectic as cyberinfrastructure is deployed across biology and medicine.

Biomedical research has experimented with the use of cyberinfrastructure to address the challenges outlined above for many years. An early example is found in the Cooperative Human Linkage Center (CHLC), a consortia formed early in the 1990s as part of the Human Genome Project

for the purpose of creating genome-wide integrated genetic maps *(2)*. CHLC was a geographically distributed virtual center connecting small specialized laboratories through informatics infrastructure communicating over the Internet (actually NSFnet at the time). It fulfilled a big-science need (creating the genetic map) through team science (each laboratory contributed specialized expertise) integrated virtually through current-art information technology. Each group worked in a context familiar to their specialized skills and the disparate parts were assembled by cyberinfrastructure to create the map. Map construction occurred through a pipelined workflow and used distributed processing over a network of multiuse computers. The raw data, analytic intermediates, and maps were distributed over the Internet through Web servers. The infrastructure to compute the maps was made available to the community through e-mail services. This example provides proof of concept that key aspects of the goals articulated above can be addressed, even with the use of a previous generation of information technology.

Technical Approach

The biology end user really doesn't care what technologies underlie cyberinfrastructure. Moreover, technology may not be the limiting factor in the development and deployment. However, the biomedical end user does provide key requirements that should be taken into consideration when choosing technology.

To facilitate adoption, cyberinfrastructure should be an extension of or interoperate with infrastructure already available to users. Ideally, it should integrate with and/or extend existing World Wide Web applications (supporting enduser needs) and Internet technology stacks (supporting the needs and existing investments of systems administrators where possible). Minimally, there must be a clear path from existing infrastructure to the new cyberinfrastructure.

The cyberinfrastructure vendor, operating system, and hardware should be as agnostic as possible. Users must have the capacity to change all of the above in order to maintain innovation and adjust to changing needs and developing technology. Open source is an oft-suggested solution to this. However, it can also be obtained by open standards and a commitment by those generating closed systems to adhere to these standards and to develop interfaces to communicate to and through them.

Biomedical cyberinfrastructure must also consider access and identity management as primary requirements. Although not unique to biomedicine, protection of human subjects is required, as is the control and tracking of intellectual property and the need to establish academic credit and data provenance.

Many experiments are being implemented to explore alternative technologies that could possibly underlie cyberinfrastructure. These include peer-to-peer technology, Web services, and grid technology. Each has interesting potential. Grid technology has several distinguishing features *(3, 4)*. First, as a consequence of the widespread use of the Globus Toolkit *(5)* in various settings, grid technology is increasingly mature. Grid technology can support virtual communities through sharing of computational resources and data resources. Access and identity control are fundamental components of the architecture. The technology supports deterministic queries across a distributed, common schema. Its fundamental architecture also supports stateful processes important to the concept of workflow. The developing Open Grid Service Architecture– Data Access Integration (OGSA-DAI) framework holds promise for adding semantics to the grid technology so that computable, semantic interoperability may be achieved. Specific database schemas and data representations can be abstracted through a metadata layer. This information can be captured and shared in ontologies and services. This advance shows promise for machine capturing of information from the disparate biomedical communities and integrating of data and information into knowledge.

Grid architecture does have some key limitations. First, despite its developing research maturity, Grid is a distant second in commercial application. Web service architecture is the technology of choice for the vast majority of cyberinfrastructure support installations, in part because of the greater relative simplicity of the architecture. It is a straightforward extension of Internet and Web infrastructure familiar to the vast majority of systems designers and administrators. The broader developer and support base associated with Web services is important to the biomedical community.

Grid technology is not the only architecture with the capacity to address the challenges faced in biomedicine. However, what distinguishes Grid

from, for example,Web services is that the capabilities described above are fundamental to the architecture. Web solutions to the challenges are outside the architecture and as such individually defined in each instance that they are created. The Grid architecture provides a standard framework for their representation and use. Encouragingly, Grid and Web services are converging.

Cyberinfrastructure in Action

As indicated above, the biomedical research community is conducting numerous experiments in developing and deploying cyberinfrastructure. With respect to Grid architecture, many are accessible through an index maintained by the Global Grid Forum (www.gridforum.org).

Many of these demonstration test beds explore the traditional definition of Grid computing in biomedicine, namely the sharing of resources across a virtual community. The range of these applications is impressive. They include molecular docking, protein structure determination, nucleic acid sequence alignment, and biologic feature extraction.

Several "proof-of-concept" test beds are exploring broad aspects of cyberinfrastructure in biomedicine, including the following:

Biomedical Informatics Research Network (BIRN). The BIRN project (www.nbirn.net) has focused on creation of geographically distributed virtual communities through shared resources. Its early work has been addressing the problems associated with new imaging platforms and the need to cross-correlate functional and structural data generated by these platforms. Its challenge is at the heart of cyberinfrastructure: How does one store, manage, curate, access, visualize, and analyze large volumes of data across a virtual community? Imaging projects generate terabytes of data through the use of disparate imaging technologies, all requiring compute-intensive applications to process.

BIRN has approached this problem by creating the virtual community through the distribution of a common, homogeneous, centrally configured hardware rack. This rack comes installed with appropriate software necessary to create the virtual community. The community is connected at high speed through the use of the Internet 2/Abiline backbone. It uses the Grid architecture defined by theGlobus toolkit with numerous extensions, particularly in the areas of brokering storage resources across the community and the use of a metadata catalog.

A series of defined test beds are evaluating and extending the cyberinfrastructure, with a key focus of neuroimaging. Each test bed of defined members is exploring a dimension of the neuroimaging domain, with one centered around brain morphology, another around functional imaging (in schizophrenia), and the last around multiscale models in experimental systems (mouse).

myGrid. ThemyGrid project (www.mygrid.org.uk) takes a different perspective on application of cyberinfrastructure. Its focus is the support of investigator-driven experiments in silico. In myGrid, local and public data can be computationally evaluated to ask and answer questions in biology. It is less focused on resource sharing than BIRN, but rather strives to address issues related to semantic complexity of biologic data and the applications that process that data. It has constructed services that facilitate integration of data and applications. It addresses challenges associated with the

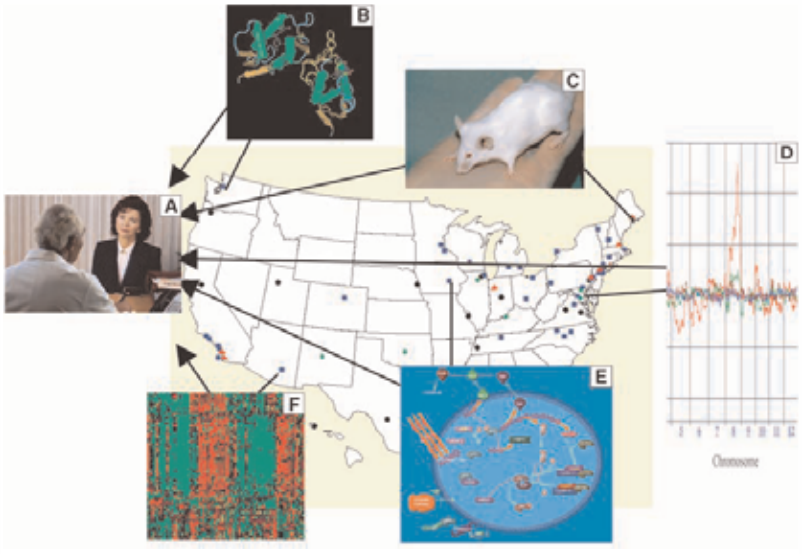

Fig. 1. The caBIG™ aims to integrate diverse biomedical research data so that investigators can consume data, services, and knowledge distributed throughout the research enterprise. For example, a scientist in California **(A)** designs an investigation following a computer modeling hypothesis-generating experiment where agent information from Washington **(B)** is queried in the context of animal model information from Maine **(C)**. Genomic aspects of the experiment use comparative genome hybridization findings generated by colleagues in Maryland **(D)**, which are interpreted in biologic processes from pathway data curated in Iowa **(E)**. These are contrasted to reference expression signatures generated by researchers in Arizona **(F)**.

rapidly evolving nature of biomedical data and issue of data provenance. Particularly interesting is its approach to creating workflows. Within its framework it supports resource discovery and distributed queries.

myGrid is a service-based architecture whose core is Web services and OGSA-DAI. It uses the common Internet and does not require specialized hardware. It accomplishes its semantic interoperability through the use of ontology-based metadata. These metadata describe data, services, and other components of the infrastructure. The environment is open; however, it has the capacity to address the mixed data and service access requirements of researchers.

The myGrid project is exploring the diversity of the domains associated with biomedical cyberstructure. In one test bed, it has explored the circadian rhythms in *Drosophila melanogaster*. In a complementary test bed, it has supported genetic investigations of the human immune disorder Graves disease.

The cancer Biomedical Informatics Grid (caBIG™). The approach of the caBIG™ project (http://caBIG.nci.nih.gov) to cyberinfrastructure is a conceptual hybrid between BIRN and myGrid. Similar to BIRN, its focus is to create a virtual community that shares resources and tackles the key issues of cyberinfrastructure. However, this community is open, spans the vast domain of cancer research, and is attempting to integrate the bench-to-bedside research cycle.

Similar to myGrid, it is an open infrastructure striving to achieve computational semantic interoperability. The caBIG™'s cyberinfrastructure is also a service-based architecture whose core is Web services and OGSA-DAI. It uses the common Internet and does not require specialized hardware. It has constructed services that facilitate integration of data and applications. Within its framework it supports resource discovery and distributed queries.

A key difference between myGrid and caBIG™ is the way they approach semantics and their related services. The caBIG™ cyberinfrastructure uses a common set of services and service registrations for the entire community. The shared caBIG™ semantic services provide biomedical ontologies and vocabularies in common use across biomedicine and cross-mappings between them. These mappings facilitate crossdisciplinary data integration and interpretation. The shared caBIG™ semantic services additionally include common data elements and object-based abstractions of the various research domains they serve. An open community process is used to maintain and extend these semantic resources. The use and registration of this common modeldriven architecture serves as the basis of community-wide service descriptions. The caBIG™ test bed currently supports basic and translational research, clinical trials research, and tissue banking and pathology (Fig. 1). Participation in these groups is open.

Biomedical Cyberstructure and the Future

The above efforts suggest that it is technically feasible to knit the vibrant threads of biomedicine into a rich tapestry. There are still many challenges ahead, both technical and cultural. The differences indicate that there is not a single path joining biomedicine. As each effort reaches maturity it will be important to compare and contrast the lessons learned from their overlapping approaches. For example, how can the community ensure that existing individual, domain, and institution silos are not simply replaced with cybersilos?

Also, although these efforts are provocative, they have not yet crossed the threshold of demonstrated value. Evidence suggests that those in the field of biomedicine are receptive to exploring these alternatives but are still skeptical. Cyberinfrastructure appears to be up to the challenges confronting biomedical research. These are early but exciting times.

References

1. T. Hey, A. E. Trefethen, Science **308**, 817 (2005).
2. J. C. Murray et al., Science **265**, 2049 (1994).
3. I. Foster, Sci. Am. **288**, 78 (April 2003).
4. I. Foster, Science **308**, 814 (2005).
5. I. Foster, C. Kesselman, Int. J. Supercomput. Appl. **11**, 115 (1997).