WIKI FEATURES AND COMMENTING

Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges

Lincoln D. Stein

Abstract | Biology is an information-driven science. Large-scale data sets from genomics, physiology, population genetics and imaging are driving research at a dizzying rate. Simultaneously, interdisciplinary collaborations among experimental biologists, theorists, statisticians and computer scientists have become the key to making effective use of these data sets. However, too many biologists have trouble accessing and using these electronic data sets and tools effectively. A 'cyberinfrastructure' is a combination of databases, network protocols and computational services that brings people, information and computational tools together to perform science in this information-driven world. This article reviews the components of a biological cyberinfrastructure, discusses current and pending implementations, and notes the many challenges that lie ahead.

WIKI FEATURES AND COMMENTING

The online version of this Review is associated with WIKI pages that allow the content to be edited and updated. There is also a facility to comment on this article. Please visit: http://nrgwiki.nature.com

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. Ontario Institute for Cancer Research, 101 College Street, Toronto, Ontario M5G 1L7, Canada.
e-mail: lstein@cshl.org
doi:10.1038/nrg2414

Twenty years ago, although the computer was a handy gadget to have around when writing a paper or grant, it was certainly not an essential piece of laboratory equipment like an electrophoresis box. But times have changed. The advent of e-mail, web sites and WIKIs has made the personal computer as essential as the telephone for establishing and maintaining scientific collaborations. Furthermore, for many biologists, particularly those in genetics, molecular biology and evolutionary biology, the way they practice science has been fundamentally changed by easy online access to genome sequencing and other large-scale data sets. For these researchers, trying to practice biology without a computer and a broadband network connection would be like trying to do cell biology without a tissue-culture hood.

Yet despite the dramatic changes that the computer has already brought to biology, its potential is far greater. In particular, the computer brings to biological research the ability to create predictive, quantitative models of complex biological processes within a cell or an organ system, or among a community of organisms. However, the tools for doing this are inaccessible to all but a few experimental biologists. Even the more prosaic task of integrating information from different specialties, such as data sets from population biology, genomics and ecology, requires specialized training in mathematics, statistics and software development.

In recognition of the transformative nature of the computer and the internet on biological research, scientific funding agencies are increasingly prioritizing the development and maintenance of something called the 'biological cyberinfrastructure'. For example, the US National Science Foundation (NSF) recently announced a US$50 million 5-year programme to create a Plant Science Cyberinfrastructure Collaborative (PSCIC), an organization that would foster "new conceptual advances through integrative computational thinking [...] to address an evolving array of grand challenge questions in plant science." The National Cancer Institute (NCI) is now 4 years into its Cancer Bioinformatics Grid (caBIG) project, which receives approximately $20 million per year[1]. The European Union's Framework Programme 7 (FP7) research infrastructure programme, which totals 27 million euros per year over 5 years, also includes a substantial component for biology cyberinfrastructure. And the Biomedical Informatics Research Network[2] (BIRN; approximately $14 million per year), established in 2001 by the National Center for Research Resources, has been developed to provide a geographically distributed virtual community using cyberinfrastructure to facilitate data sharing and to foster a biomedical collaborative culture.

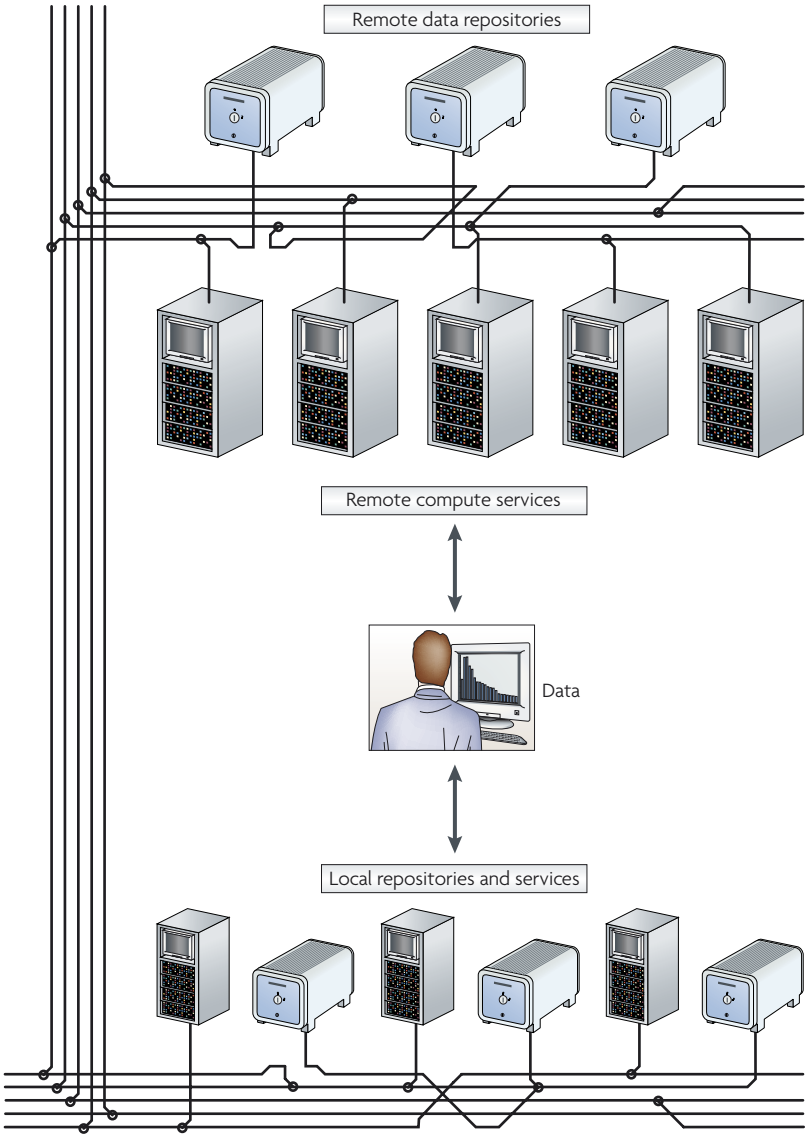

Figure 1 | **The components of a cyberinfrastructure.** A cyberinfrastructure consists of data repositories for storing community data sets and compute services for querying, integrating and analysing that data. Local data sets and compute services can be plugged into the cyberinfrastructure to allow individual researchers and groups of collaborators to work with private and semi-private data sets in the context of the community resources.

However, despite all the talk of biology cyberinfrastructure, it can be a hard to pin down what exactly it is, as the term means different things to different people. To some, cyberinfrastructure is access to raw compute power via a distributed computing grid system. To others, it is access to vast online databases of collected information. Still others think of advanced desktop tools for managing their research. In fact it is all of these things and more. This article will describe the necessary components for a cyberinfrastructure, highlight a few examples of current and planned cyberinfrastructure projects, and sketch out the path towards a world in which biologists have full access to the potential of computational analysis.

WIKI
A popular web page authoring system that allows individuals to collaborate on large communal documents. Wikipedia is the best known example, but there are many tens of thousands of WIKIs in use. The name comes from the Hawaiian word for quick.

The pieces of a cyberinfrastructure

The aims of a scientific cyberinfrastructure are to get both data and the tools needed to understand it into the hands of scientists so that they can run sophisticated computational analyses on that data, and to facilitate the publication and exchange of the knowledge arising from those analyses. The essential components of a cyberinfrastructure are: a data infrastructure comprised of a series of repositories for storing, integrating and retrieving essential information; a computational infrastructure for manipulating and analysing those data sets; a communications infrastructure for interconnecting the computational and data resources; and the human infrastructure for supporting collaboration among researchers (FIG. 1).

The data infrastructure. The data infrastructure is probably the most familiar to readers. The emerging biology cyberinfrastructure already has a mature network of databases. A few familiar examples include PubMed[3], Ensembl[4], and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[5]. New information flows into these databases by automatic acquisition and direct submission, and the information currently contained in them is typically accessed by researchers browsing them via web-based front ends.

The computational infrastructure. The computational side of a cyberinfrastructure might be a less familiar concept. This part of the infrastructure gives researchers access to the hardware and software needed to perform computation-intensive tasks. Examples include using image-analysis software to measure the distribution of nucleur sizes in a set of histological slides, building a phylogenetic tree from a collection of gene sequences, and modelling a network of biochemical reactions using a kinetic simulation package. In biology, most of this work is done locally in the researcher's own laboratory or institution; if the researcher needs more compute power, he buys more central processing units (CPUs). The major exceptions to this are a limited number of genomics tools, such as the BLAST[6] sequence search and alignment algorithm, which require access to such large and unwieldy data sets that the software is usually hosted by the same organization that maintains the data repository.

Relying entirely on local compute resources has drawbacks, however. It can be inefficient: one laboratory is probably not using all its compute resources 100% of the time and so the machines are left idle. A larger drawback is that there is significant overhead for installing, configuring and maintaining computational biology software. This job can keep a postdoctoral researcher or system administrator busy for a long time. Physicists, astronomers and atmospheric scientists long ago figured out how to lessen this problem by relying on shared supercomputer centres for their hardware and personnel needs. This where the idea of 'compute grids' comes in. These are systems in which geographically scattered compute clusters are combined through the internet into a virtual supercomputer centre. When a researcher is not actively using his own cluster, other groups around the world are using its otherwise idle time.

The communication infrastructure. The third essential component of an effective cyberinfrastructure is communication among the pieces. This includes three types of connectivity: low-level, syntactic and semantic. Low-level connectivity is the easiest part, simply requiring the network to have the connectivity and bandwidth needed to transfer data between repositories, computational resources, and the researcher's desktop with acceptable speed. Syntactic and semantic connectivity, by contrast, involve standards for describing data. If the researcher is trying to make new discoveries by integrating data from two different repositories, the representations of the data from the two repositories must be compatible with each other.

Syntactic connectivity is achieved by establishing common formats for organizing data: for example, the GenBank file format is a widely understood way of exchanging formatted nucleotide and protein sequence data. However, even if two data files share the same format, it doesn't mean that they are semantically interoperable. For example, it would be incorrect to assume that a gene expressed in a mouse 'vein' has a similar anatomic expression pattern to a gene expressed in fly wing 'vein' or a plant leaf 'vein'. Semantic interoperability means that the concepts embodied in the data use a common terminology, and is usually achieved with the help of an ontology, which is a formal description of the central data types and concepts in a domain of knowledge. For example, the plant anatomy ontology[7] describes the major anatomic parts of flowering plants and can be used to exchange information about tissue-specific gene expression, mutations that affect plant development, and other anatomical knowledge.

The human infrastructure. The final key part of a cyberinfrastructure is the people who build it, use it and contribute to it. A true research cyberinfrastructure must be part of the sociology of science, and become as integral to the practice of science as publishing and reading papers. This means that a cyberinfrastructure must encourage the electronic sharing of protocols, analysis algorithms and data sets, as well as community curation of core data sets. It also means that a significant fraction of researchers must have the training to be comfortable with the design and development of software systems.

A vision for the biology cyberinfrastructure

The current biology cyberinfrastructure has a strong data infrastructure, a weak to non-existent computational grid, patchy syntactic and semantic connectivity, and a strengthening human infrastructure.

Much of the data is 'out there', but it can be difficult to find and challenging to use effectively once found. An example would be a researcher who wants to interrogate one of the recent yeast two hybrid (Y2H) protein interaction data sets[8] to ask whether proteins that have an unusually large number of pairwise interactions are more likely to be evolutionarily conserved (FIG. 2a). Although this is not a particularly complex question, answering it requires a fair bit of work. Here is one possible path: first the Y2H data set must be translated from its representation as pairs of protein names into pairs of gene names, which requires use of NCBI RefGenes[9] or UniProt[10]. Next the coordinates of the exons of those genes must be looked up on a human genome annotation database, such as Ensembl or the University of California Santa Cruz (UCSC) Genome Browser[11]. After this, the UCSC Genome Browser is used to fetch the PhastCons[12] conservation scores for the 17-way vertebrate genome alignment across the exon coordinates determined in the previous step. Finally, the conservation scores across the exons of the highly interacting genes are integrated to see whether they are significantly more conserved than an average gene.

The main characteristic of this scenario is that much of the work is done locally. The data sets are downloaded, transformed and integrated on the researcher's local computer. Aside from the precomputed genome alignments and associated PhastCons scores, no shared compute resources are used. Owing to data format and naming system incompatibilities, a lot of programming work is involved. Furthermore, the results of the analysis are local. The method used for the analysis and the integrated results remain on the researcher's personal computer until he publishes his findings and makes the additional effort to put the analysed data on a public web site.

A mature biology cyberinfrastructure should make this type of analysis much more straightforward (FIG. 2b). Ideally, the researcher would be able to design the experiment at a high level by describing the data sets he wishes to work with and the relationships he wishes to traverse (protein to gene to exon to conservation score) by using a graphical tool or a high-level description language. The infrastructure would then do the hard work of finding databases, analysis services and compute resources that can satisfy the request, thereby transforming and integrating the data and returning the results. If the researcher desired, he could then easily share the method and results with the research community by pushing a 'publish' button; this information would then become a discoverable service that could be re-used by others. Over time, other members of the research community could add value to this work by commenting on it, linking it to related work, contributing modifications to the method, and submitting new raw and analysed data sets that enrich it.

It will be some time before this vision is a reality. In the meantime, several projects are striving towards this vision, which I discuss below in order of increasing complexity.

Centralized online databases and analysis systems

The simplest and by far the most successful form of cyberinfrastructure is the online database. Readers will already be familiar with such resources, which include the GenBank and European Molecular Biology Laboratory (EMBL) nucleotide databases, the UniProt protein database, the UCSC and Ensembl genome databases, the Protein Data Bank (PDB) protein structure database, the PubMed literature database, the Online Mendelian Inheritance in Man (OMIM) database of genetic diseases, and the various model-organism

Ontology
An enumeration of the concepts used in a particular domain of knowledge, their definitions and the relationships between them.

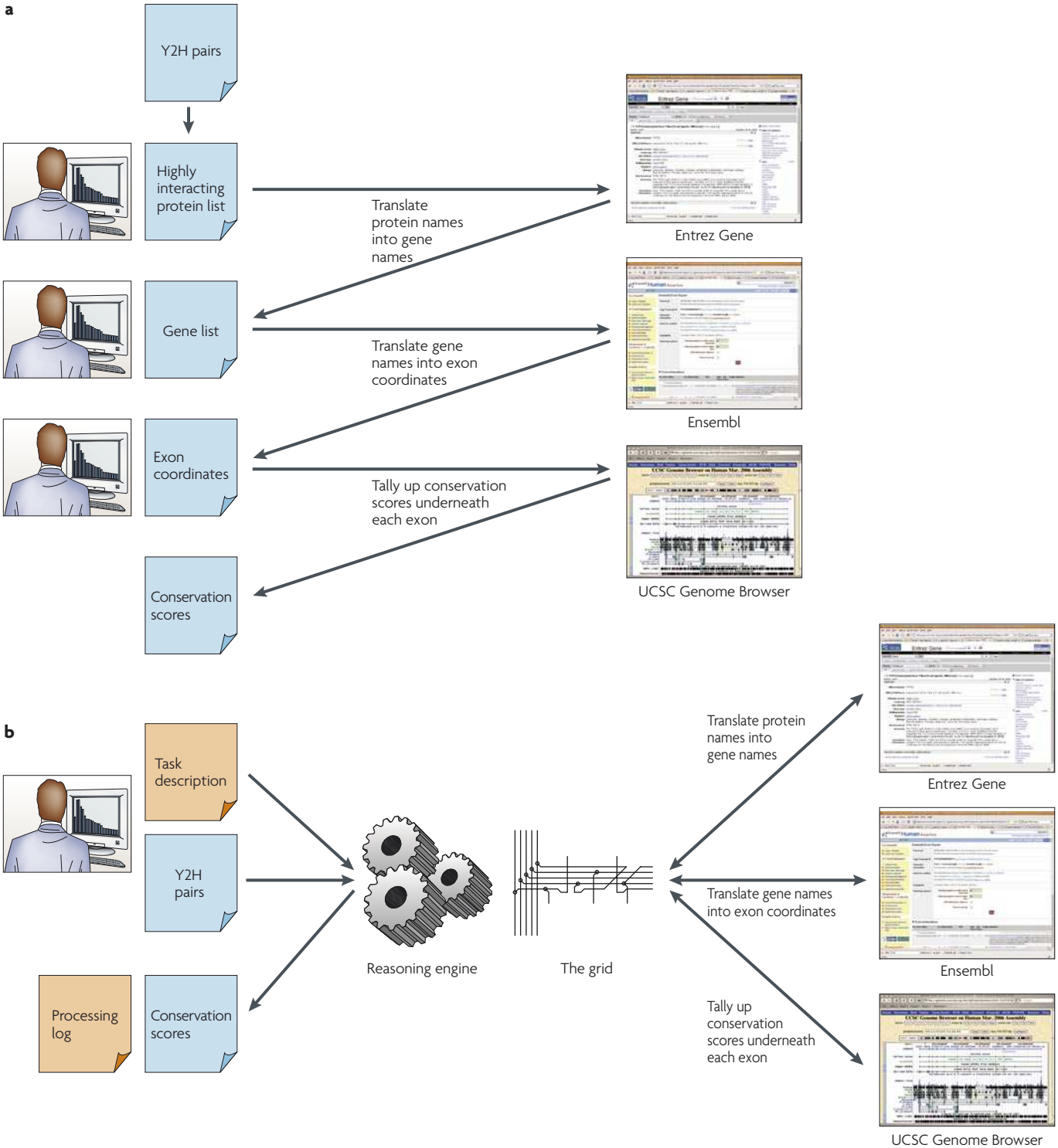

Figure 2 | **The process of bioinformatics research now and in the future.** The researcher is analysing a set of yeast two hybrid (Y2H) protein–protein interactions and wants to know whether the most highly connected set of proteins is more conserved than average. **a** | To address a typical bioinformatics question today, a researcher might have to mine multiple databases and write custom software at each step to reformat and collate the results. **b** | In a fully realized cyberinfrastructure, a researcher describes the task in a high-level language or graphical flowchart, submits the task description and input data to the grid, and the grid performs the desired set of operations without user intervention. The researcher might be assisted by a 'reasoning engine' that proposes all or a portion of the workflow. The results include a description of the steps taken so that the provenance and reliability of the outcome can be determined. Screenshot images are courtesy of Entrez Gene (National Library of Medicine), Ensembl[4] (http://www.ensembl.org/index.html; screenshot URL: http://jul2008.archive.ensembl.org/Homo_sapiens/exonview?db=vega;transcript=otthumt00000050411) and the University of California Santa Cruz (UCSC) Genome Browser[11] (http://genome.ucsc.edu).

databases. These resources aim at a particular problem in biological data management and solve it very well. However, they are restricted in scope. They use centralized resources and are typically managed either by a single research centre or a small number of collaborating groups. They are also heavily data-centric; the computational tools they provide are geared towards facilitating access to the data. For example, each of the sequence databases provides a sequence similarity search service such as BLAST or BLAT[13].

Community annotation hubs

Community annotation hubs are the result of opening up a centralized database or toolkit to direct contribution by the community. A familiar mainstream example is the Flickr photo-sharing site. Someone visiting the Flickr site uses the tools the site provides to upload and organize their digital photograph album, while other users add value to these photographs by adding descriptive tags and ratings. The result is a searchable database of images that is far more richly annotated than a finite group of curators could achieve on their own.

Community annotation systems are growing in popularity in biology as well. An early example is the Entrez GeneRif system[9], which allows researchers to tag gene records with a short description of its function. A richer example can be found in the facility provided by many genome browsers for sharing tracks of genome annotations, either by uploading structured files or by using a web communications protocol such as the Distributed Annotation System (DAS)[14]. This feature allows researchers to attach arbitrary information to the genome, ranging from simple comments to sophisticated computational predictions of small RNA genes and to instantly share their results with others.

The most sophisticated community annotation systems currently used in biology are based on the WIKI concept of an online editable document repository. The myExperiment system, part of the myGrid project[15], is an online repository of bioinformatics protocols. Researchers can search the repository for a protocol that does more or less what they want, customize it, run it and contribute their customized version back to the repository. EcoliWiki is a community annotation site for *Escherichia coli* genes, genome and biology. At its core is a series of gene pages that contain information about the structure, function, regulation and evolution of a gene. Pages were initially generated in an automated fashion, but have since been enhanced by a distributed community of over 400 contributors. Similarly, WikiPathways is a repository of community-editable biological pathways for human and other species. It comes complete with a custom pathway-editing tool that allows users to draw and modify pathway pictures directly on the web.

Although its focus is not on biology, nanoHUB[16], a community site that is based at Purdue University (Indiana), and is devoted to nanotechnology, shows just how effective a research tool a community annotation system can be. This site combines a research community calendar, chat rooms, educational and teaching materials, data sets and research tools in a convenient one-stop location. The site is built around a community-contributed series of nanoscience simulation, analysis and visualization tools, which can be interactively run online on top of a compute farm maintained by Purdue University. These tools can be freely shared, combined in interesting ways, incorporated into online publications, and rated and tagged by community members. The site has 500 active contributors per year and serves requests from 60,000 users per year.

Bioinformatics toolkits

Both centralized databases and community hubs tend to be insular. They stand alone and cannot easily interchange data with one another. Bioinformatics toolkits seek to break down the insularity by providing a common set of tools that can be used among multiple projects and which interoperate with each other.

A good example of a bioinformatics toolkit project is the Generic Model Organism Database (GMOD) project[17], a collaborative endeavour among the model-organism databases FlyBase, WormBase, *Saccharomyces* Genome Database (SGD), The *Arabidopsis* Information Resource (TAIR), EcoCyc, DictyDB, Rat Genome Database (RGD) and Gramene, among others. It is a collection of software applications that interoperate to provide some, but by no means all, of the needs of a typical model-organism research community. GMOD applications include a genome browser, a genome annotation editor, a web site framework, a database query engine and tools for linking genomic features to the literature. The applications interoperate by relying on a common database schema called Chado[18], which provides a shared data model for genomes and their annotations. GMOD sites can share genomic annotations with each other, and with other DAS-enabled sites. Other popular bioinformatics toolkits include: the Cytoscape project[19], a package of tools for visualizing and analysing gene and protein networks; Protégé[20], a toolkit for working with ontologies; GenePattern[21], a framework for analysing gene expression signatures; and Galaxy[22], an interactive web-based system for performing analyses across genomes drawn from a variety of sources.

In contrast to centralized databases, toolkits are built around a distributed model. They are typically installed and run locally, and serve the needs of individual researchers or discrete research communities. The toolkit development process is also typically decentralized; many toolkits are developed by multiple groups using an open-source model. Another characteristic of toolkits are their extensibility. Cytoscape, Protégé and GenePattern are each built around plug-in architectures that allow bioinformatics developers to add new features and functionality. GMOD also uses plug-ins, but additionally can be extended by writing entirely novel applications that are compliant with the way in which the Chado database represents genome data.

Like the centralized resources, toolkits are effective because their scope is limited. They handle a limited set of biological data types, and all parties involved in their development agree on how those data types will be named and organized.

Web service links

Although databases, hubs and toolkits lead to sites that can exchange information about the same classes of biological data, they encounter problems when attempting to link up disparate but related classes of information. For example, a hub devoted to protein structures cannot automatically exchange information with a database of genetic polymorphisms even though the two share something in common (in this case, genomic sequence coordinates). It is also difficult to interconnect very similar resources that were written by different groups, because the groups will have used different technologies and different representations of the data. Web service protocols attempt to solve this problem by providing a common technology for heterogeneous data and compute resources to interoperate. Two web standards, the Web Services Description Language (WSDL) for describing the capabilities of services and the Simple Object Access Protocol (SOAP) for invoking those services, are the pillars on which most service systems are built. Web services are important for bioinformaticians and software developers who are creating a cyberinfrastructure. Researchers will probably never use web services directly, but will rely on the software tools, such as analysis and visualization engines, that run on top of these services.

Globus. The industry heavyweight in the web services arena is the Globus Toolkit[23], a collection of open-source libraries and utilities that provide software developers with the means to announce the availability of a compute or data resource, to discover the existence of that resource, and to invoke it. Globus also handles authentication and authorization so that data providers can restrict sensitive or proprietary data sets to those users who have the proper approvals to access that information. As we discuss later, two of the largest emerging bioinformatics grids, BIRN and caBIG, use the Globus Toolkit technology.

BioMOBY. An alternative web services system that is used in bioinformatics is BioMOBY[24]. Like Globus, it uses SOAP and WSDL to connect heterogeneous services but it is relatively lightweight, meaning that it does not require a large investment of time and effort to install and maintain. BioMOBY is simpler to install, and easier for developers to work with, primarily because it dispenses with most authentication and authorization facilities. BioMOBY is used by several large plant biology community collaborations, most notably the PlaNet Consortium[25].

Ontologies. Although web service systems describe how resources connect to the grid and exchange data, they say nothing about the semantic content of the data. To do anything useful, resources using web services must share common semantics. For example, resources that exchange information about genes must agree, at some level, about what a gene is, how genes are named and what pieces of information can be attached to a gene. This is where biomedical ontologies come in[26]. In the Sequence Ontology (SO), genes are defined and described in a standard, unambiguous way[27]. Web services that use SO to describe genes do not have to worry about meaning being lost in the exchange. All current biomedical cyberinfrastructure efforts use ontologies to a greater or lesser extent. One of the most comprehensive builders of ontologies is the myGrid project. MyGrid contains a large and growing ontology of biological web services that describes many common data types and the way they can be manipulated. For example, the myGrid ontology has an entry for the BLAST sequence search and alignment algorithm that describes how to invoke it and how to interpret its results. The ontologies used in MyGrid are written using a standard format called OWL[28] and whenever possible make use of existing biological ontologies such as the Gene Ontology. MyGrid has recently secured funding to build a fully curated catalogue of biological web services in all their forms.

The most visible software tool used in MyGrid is a workflow management system called Taverna[29]. Taverna is a desktop application that allows researchers to find bioinformatics services, tie them together in a graphical flowchart, and invoke them on his or her own data sets (FIG. 3). Taverna can operate on top of a variety of web service systems including Globus and BioMOBY, as well as the more informal web-based services known as Representational State Transfer (REST). It is rapidly gaining in popularity and is used actively by more than 200 groups around the world. The myGrid web site also features a growing library of user-contributed Taverna workflows, most of which have to do with genome and sequence analysis.

Semantic web

A characteristic of web services is that they make a strong distinction between data and operations on the data. For example, in a typical compute-grid environment a user identifies a service he would like to invoke, formats his or her input data, invokes the service, and unpacks and interprets the results. An alternative approach, called the semantic web, describes everything as data. Under this outlook, there are no services that transform data sets, there are simply pieces of information and the relationships between them. For example, consider a sequence database that relates genes to the proteins that they encode. A web service running on top of this database might provide the grid with a 'Gene to Protein' service, which accepts a gene ID and returns one or more protein IDs. By contrast, a system participating in the semantic web would publish a series of 'Gene' and 'Protein' objects and the relationships among them, in particular the 'Encodes' relation that relates a gene to its protein products. Another feature of the semantic web is that assertions about relationships among objects are not limited to the local database. A research group that works on the prediction of protein folds can publish a series of 'HasFold' assertions that relate named folds to the 'Protein' objects described by another research group. No co-ordination is needed between the researchers that published the relationships between genes and proteins and those that published the relationships between proteins and folds.

Web service
A web-based resource that can be programmatically invoked to perform a database search or a computation, or to provide some other service.

Web Services Description Language
(WSDL). An XML-based language used to describe the nature of SOAP web services.

Simple Object Access Protocol
(SOAP). The dominant messaging protocol for defining and invoking web services.

OWL
A dyslexic acronym for Web Ontology Language. It is an XML-based language used to describe ontologies. A variant of OWL called OWL Description Logics (OWL DL) is particularly suited for creating semantic webs of ontologies that can be traversed by reasoning engines.

Representational State Transfer
(REST). An alternative web services protocol that is sometimes more suitable than SOAP for particular web services.

Semantic web
An interrelated network of ontologies that together describe resources available on the web.

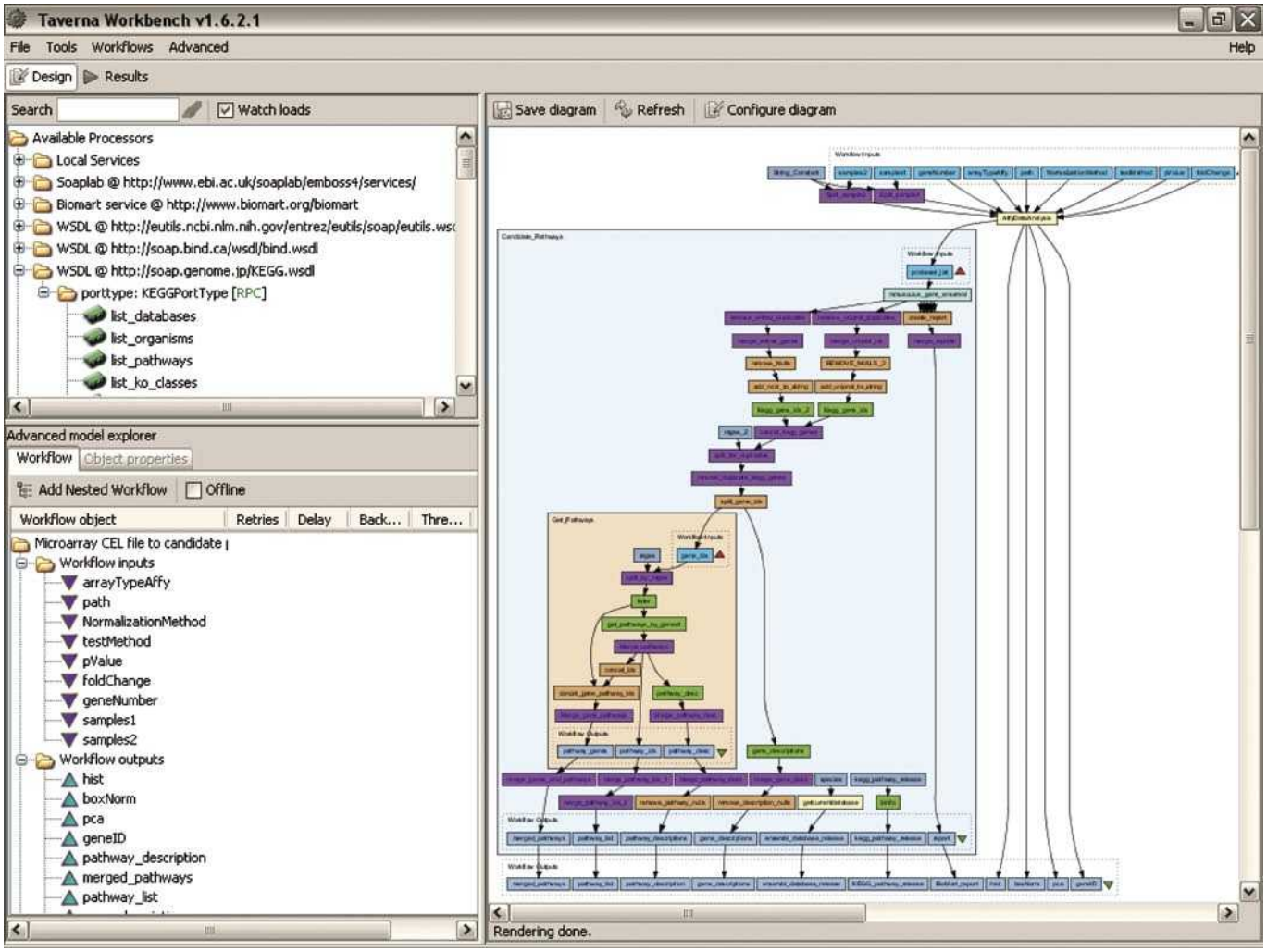

Figure 3 | **The Taverna workflow manager.** The tool lets the user describe each step of a bioinformatics task using a graphical flowchart. The tool then runs each service involved in the task and manages the flow of information from one service to the next. Screenshot is reproduced courtesy of Paul Fisher and the myGrid team.

Under this system, a user who wishes to get information from the grid does not invoke a service, but instead performs a search through semantic web space assisted by something called a reasoning engine, which traverses the appropriate relationships. For example, the researcher who is exploring the evolution of gene families might ask the semantic web to find all genes that encode proteins containing a particular fold. The reasoning engine will determine that it must traverse two data sets to satisfy this request: first it must work backwards from the 'HasFold' relationship that relates folds and proteins, and then follow the 'Encoded' relationship to go from the proteins to the genes that encode them. It then returns the result of this query to the user, who might not even realize that the reasoning engine traversed two distinct web sites to satisfy the request.

Surprisingly, even dynamic computations that seem quintessentially service-like can be represented in the semantic web. For example, the BLAST sequence search and alignment system can be described as a series of relationships between a search sequence, a sequence database, and a set of alignments and their significance scores.

The main advantage of the semantic web over other types of web service is that it greatly reduces the amount of coordination needed among participants. It will work even if not everyone agrees on the same ontologies in advance, or if groups use different subsets of the same ontologies. The disadvantage is that the tools for setting up semantic webs are in the research stage and do not have the industry support enjoyed by Globus and other conventional web-service tools.

SSWAP. To my knowledge, there is currently only one project that aims to bring the pure semantic web to biomedical research. That project is the Simple Semantic Web Architecture and Protocol (SSWAP[30]), led by researchers at the National Center for Genome Resources in Santa Fe, New Mexico. SSWAP uses OWL DL, a form of the OWL ontology language that is particularly suited for making logical connections between data objects (see the SSWAP protocol), and uses a software framework for

discovering and interrogating these connections. SSWAP re-uses existing ontologies such as the Gene Ontology whenever possible, allowing pre-existing databases and services to publish their data sets in SSWAP form. A Discovery Server at the SSWAP web site gathers these published services and submits them to a reasoning engine to build a web-searchable knowledge base.

SSWAP is currently being used in a proof-of-principle project called the Virtual Plant Information Network (VPIN) to share genetic and genomics data among several plant biology databases. Although it is a promising technology, SSWAP is still very much a research system; for the time being, conventional web service architectures dominate.

I now discuss several biological cyberinfrastructure projects that are using web service technologies on a production basis.

The BIRN project

The Biomedical Informatics Research Network, funded by the National Institutes of Health National Center for Research Resources since 2001, supports a growing number of collaborative projects that involve more than 30 universities and 40 research sites. These collaborative groups are primarily centred around the storage and analysis of neuroanatomical, clinical, genomic and behavioural data in humans and in animal models. Much of the data that BIRN works with is structural and functional imaging data, including magnetic resonance imaging (MRI) and functional MRI studies, which require large amounts of storage space and processing power. A distinguishing characteristic of BIRN is that it has developed a robust software installation and deployment system that allows research groups to easily implement a local BIRN end-point, called a BIRN rack. Once a research centre has installed a rack, which costs roughly $20,000, it can host data and contribute compute resources to the BIRN grid. Registered users can access shared BIRN data sets and computational resources from any internet-capable location via a web portal, which provides a collaborative environment for the research scientist, or through a collection of diverse data management, analysis and visualization applications. BIRN also provides free access to published data sets for interested researchers through a data archive called the BIRN Data Repository.

BIRN uses the Globus Toolkit to attach compute and computational resources to the grid and to authenticate authorized users. Semantic integration is achieved by BIRNLex, an ontology that covers multiple aspects of neuroanatomy, species, behavioural and cognitive processes, subject information, experimental practice and design, and provenance information. A large portion of BIRNLex is based on shared community ontologies, such as NeuroNames, Open Bioinformatics Ontologies (OBI), Phenotype and Trait Ontology (PATO) and the NCBI Taxonomy.

The caBIG project

A more complex task was faced by the caBIG project, which was launched in 2003 with the ambitious goal of providing a common information platform to support the diverse clinical and basic research programmes of the US National Cancer Institute's 87 cancer centres. This project had to integrate a highly heterogeneous set of databases and software tools, ranging from workflow systems for managing clinical trials to research tools for genome analysis and annotation.

Like BIRN, caBIG chose the Globus Toolkit as its underlying grid technology, creating a web services network called 'caGrid'. In order to handle the high degree of heterogeneity among the cancer research services they wished to interconnect, the developers of caBIG then had to undertake an extended and painstaking process of unifying the data models used by each of the services. For example, the concept of 'blood pressure' appears in dozens of subtly different ways in the various clinical databases used by the cancer centres. One of the earliest tasks that caBIG performed was to unify all key concepts into a reference vocabulary and set of common data elements (the VCDE), using existing ontologies whenever possible and creating new ones when necessary. To add a new resource to caGrid its developers must ensure that their tool reads and writes data types that are already described by the VCDE. If the VCDE is missing a concept that they need, there is a standard submission and approval process for getting the new concept incorporated into the VCDE.

Currently, caBIG supports over 40 software tools, most of which interoperate with each other at some level. For example, the clinical trial data-collection system, called C3D, stores surgical pathology reports on tumour specimens. These reports can then be read by a text information extraction system called caTIES and converted into a standardized format that describes the type and characteristics of the tumour. It is then possible to associate this histopathological information with gene expression profiles that are captured and stored in the microarray database of caBIG, which is called caArray, and finally analysed for expression signatures that correlate with tumour type or grade using the genePattern tool mentioned earlier. The Taverna workflow management tool described above has also recently been adapted to work with caBIG, allowing researchers to discover and interconnect caBIG data and compute services using an intuitive graphical user interface.

One of the controversial aspects of caBIG has been its top-down management style. caBIG has four domain 'workspaces', three strategic level workspaces, two cross-cutting workspaces, over a dozen special interest groups, and countless *ad hoc* working groups and committees. The day-to-day governance of caBIG is under the direction of management consulting contractor Booze Allen Hamilton (BAH), which juggles an intricate calendar of conference calls, milestones and face-to-face meetings. BAH also tracks a series of certification committees that work to ensure that the software attached to caGrid meets a required set of compatibility standards. This is very different from the working style of most bioinformatics researchers, some of whom have complained of 'culture shock', but the result has been a coherent system that goes a long way towards achieving the vision of a pervasive biology cyberinfrastructure.

According to Mark Adams, the caBIG programme manager at BAH, the challenge that caBIG now faces is lagging adoption by end-users. Because caBIG has been driven in a top-down fashion, it is lacking in 'grass roots' support from clinical researchers and experimentalists. This problem is confounded by the fact that the substantial achievements of caBIG are largely invisible to end-users. There is as yet no 'killer application' that really showcases the abilities of caGrid.

The iPlant Collaborative

The last project I will discuss is the iPlant Collaborative (iPC), a cyberinfrastructure project recently funded by the US NSF. A collaboration between the University of Arizona, Arizona State University, Cold Spring Harbor Laboratory, Purdue University and the University of North Carolina, iPlant is receiving $50 million over 5 years to create a cyberinfrastructure collaborative for the plant sciences that will enable "new conceptual advances through integrative, computational thinking." The project began in February 2008, and so is still in its organizational stages.

In contrast to the technology-driven cyberinfrastructure projects discussed earlier, iPlant focuses more on the human side of the infrastructure than on the technical side. Approximately a third of its budget will be devoted to community building through a series of symposia, workshops and meetings that bring together plant scientists, computer scientists, software engineers and mathematicians to identify and discuss 'grand challenge' problems in plant biology, including such fundamental questions as how genetic diversity in plant populations translates into phenotypic diversity, and how plants perceive and respond to the environment. Participants of these meetings will be asked to identify ambitious but feasible research projects that address some of these grand challenge questions.

The iPC investigators expect that a small number of novel research projects will emerge from these discussions, and that some of the participants will team up to form collaborative research teams to take up some aspect of a grand challenge. The iPC will provide these grand challenge teams with the basic infrastructure needed to support their collaborations, including physical meeting space, mailing lists, videoconferencing systems, web pages, WIKIs, blogs and electronic forums, as well as customized software resources, which the iPC calls 'discovery environments'. These are envisioned to be grid-like collections of data and compute resources that have been organized in a way that is most suitable to the grand challenge team's research needs. For example, a discovery environment geared to a developmental biology project might provide access to anatomy and developmental ontologies, histological image databases, annotated collections of developmental mutants, signalling pathway databases and simulation tools for modelling intercellular communications.

Although they are designed to meet the needs of a specific grand challenge team, the discovery environments are intended to be open to the whole research community. To maximize their generality, discovery environments will have extensive 'mash-up' facilities. The mash-up, a concept that should be familiar from Google Earth, allows disparate data sets to be tied together by a framework. The discovery environment mash-up facilities will encourage researchers to combine their own data sets with public data sets, and with those of their colleagues, with the hope of discerning novel patterns that would otherwise be inapparent. For example, a developmental-biology discovery environment might use an interactive diagram of signalling pathways as its mash-up framework (FIG. 4). One part of the research team might superimpose on top of this framework a set of microarray-derived gene expression patterns from developmentally normal and mutant plants, whereas another could contribute a transcription factor-binding site interaction data set taken from a series of chromatin immunoprecipitation coupled with microarray (ChIP on chip) experiments[31]. The combination of these data sets might point to a hypothesis for the mechanism of action of the mutant, which could then be explored using simulation and modelling tools attached to the discovery environment. The resulting model would be published back to the discovery environment for use by the rest of the team and the broader research community.

Unique among the projects discussed earlier, the iPC has sociologists on staff to monitor the effect the iPC has on patterns of interdisciplinary collaboration and to recommend ways to improve the interactions. It also has a significant education and public-outreach component. Only time, of course, will tell whether this mixture of technical and human infrastructure will live up to its promises.

Where do we go from here?

The biological cyberinfrastructure is slowly emerging, but the exact outlines of what is to come are still unclear. We have an excellent data infrastructure, particularly in the field of genomics, and we are beginning to see the spread and adoption of compute grids, such as the ones created by BIRN, caBIG and MyGrid. The area in which the cyberinfrastructure is weakest is in integration and communication across disciplines. Even though caBIG and BIRN use the same web services technology, they cannot easily talk to each other owing to the different choices that the groups made concerning which ontologies to use to describe services, as well as the directory system used by data providers and users to publish and search for services. Nor do the other cyberinfrastructure components described earlier, such as DAS, BioMOBY, SSWAP or GMOD, interoperate to any meaningful extent.

Thus the most likely map of the biological cyberinfrastructure that is coming in the immediate future is an archipelago of islands; each discipline using a grid that is internally consistent, but effectively isolated from the others. To achieve a future in which information from different disciplines is interoperable we need either to coordinate more tightly, perhaps using a caBIG-like top-down management approach, or adopt a technology that is tolerant of diverse data models, such as the semantic web framework advocated by the SSWAP group. My

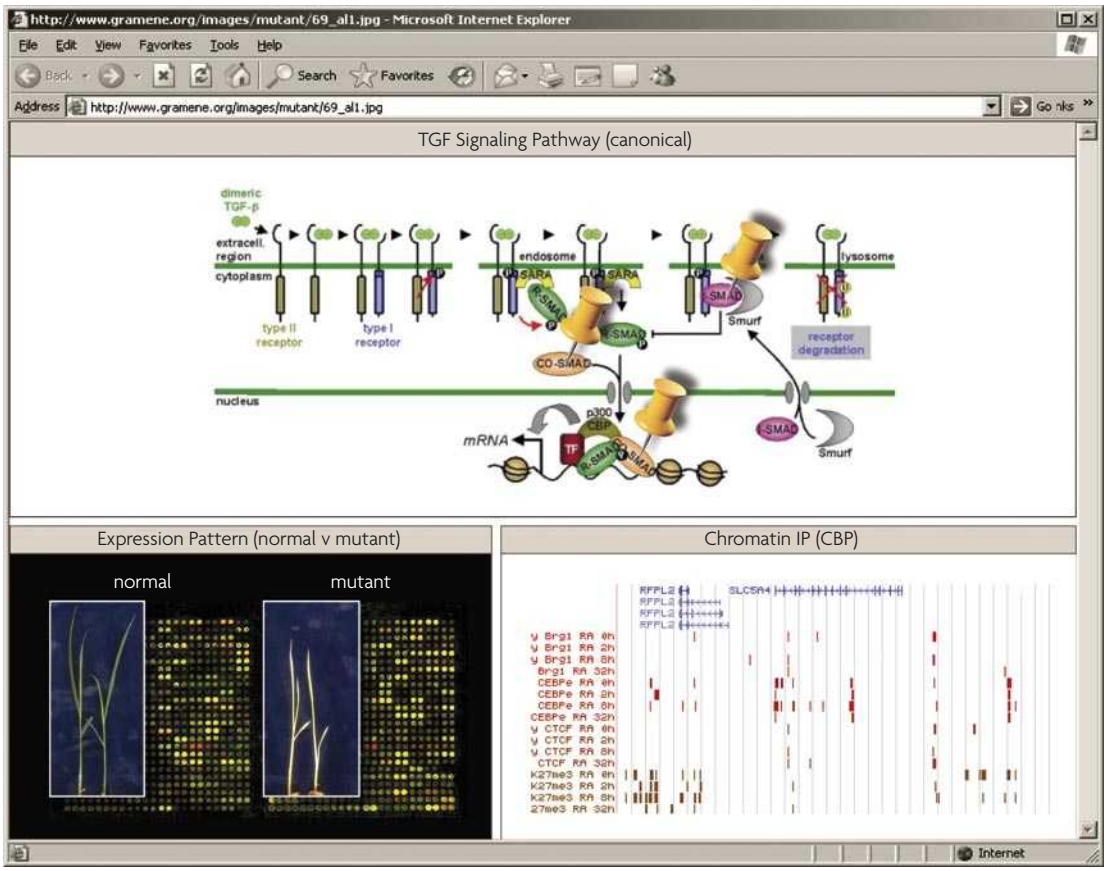

Figure 4 | **A mash-up mock-up.** A 'mash-up' environment allows multiple data sets to be graphically superimposed on each other. In this conceptual example, the researcher is investigating the basis for a mutation in a grass. They have superimposed microarray expression patterns from the normal and mutant plants onto a diagram of a signalling pathway: the pushpins indicate several genes that are significantly downregulated in the mutant. A series of chromatin immunoprecipitation (IP) experiments, superimposed on the same view, shows the binding pattern of one of the transcription factors affected by the mutation. The researcher's microarray data set, and inferences about what it means, can be published to the grid for use by other scientists. The image in the chromatin IP (CBP) panel is reproduced courtesy of the University of California Santa Cruz (UCSC) Genome Browser[11] (http://genome.ucsc.edu).

opinion is that the semantic web framework approach is both more sustainable and more likely to encourage innovation. However, semantic web technology is still immature, and will need at least a few more years of development before it is ready for wide use.

Fortunately, in the short term at least, there is a third path to an effective biology cyberinfrastructure that is immediately available. By using current technology to build systems that encourage the electronic submission of data sets and that facilitate community annotation, collaboration, and the sharing of computational and experimental protocols, we can gain many of the benefits of more sophisticated systems by leveraging the human capacity to make sense of noisy and contradictory information. Semantic integration will occur in the traditional way: many human eyes interpreting what they read and many human hands organizing and reorganizing the information.

Over the longer term we need to merge manual community annotation systems with automatic grid systems, but to do so we must pay attention to the human infrastructure. In order to become active and effective contributors to the cyberinfrastructure, biological researchers will need to become familiar with the basics of computer science, learn to use ontologies to describe their data and protocols unambiguously, and have the skills to put this information in a form that can be readily adapted and re-used by others in the community. This will require changes in the way biology is taught at the undergraduate and graduate levels. The changes will be slow, but they have already begun. Scientific publishers also need to become partners in the development of the biology cyberinfrastructure. Many papers are now accompanied by 'supplementary information' — electronically readable files of raw and interpreted data. However, these supplementary files are usually formatted in an *ad hoc* manner, making it impossible to automatically extract the information in them to combine with data produced by related experiments. Only a limited set of experimental data types, notably sequences and microarrays, follow any standards. How valuable it would be to the community if the raw and

interpreted data from every published experiment were available in machine-readable form. Publishers are well positioned to establish and motivate standards for electronic publication of data, and understand the issues of peer review and provenance better than any other party. A collaboration among publishers, bench scientists, bioinformaticians, computer scientists, community annotation hubs and grid-software developers could be extremely fruitful.

Conclusion

This is an exciting time for biology. The projects that are now in progress or just getting under way point towards a future in which scientific collaborations will be unimpeded by geographic constraints or by limited access to data. Just as it is now inconceivable to do science without access to a personal computer and e-mail, in a decade the cyberinfrastructure will be an absolutely indispensable part of the biological researcher's equipment.

1. caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. *Medinfo* **12**, 330–334 (2007).
2. Martone, M. E., Gupta, A. & Ellisman, M. H. E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nature Neurosci.* **7**, 467–472 (2004).
3. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
4. Flicek, P. *et al.* Ensembl 2008. *Nucleic Acids Res.* **36**, D707–D714 (2008).
5. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
7. Ilic, K. *et al.* The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol.* **143**, 587–599 (2007).
8. Fields, S., Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1989).
9. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26–D31 (2007).
10. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* **36**, D190–D195 (2008).
11. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
12. King, D. C. *et al.* Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.* **15**, 1051–1060 (2005).
13. Kent, W. J. BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
14. Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R. & Stein, L. The distributed annotation system. *BMC Bioinformatics* **2**, 7 (2001).
This paper describes an early biological cyberinfrastructure system that uses a common syntactic protocol to exchange data about genome annotations, but it has the problem of weak semantics.
15. Stevens, R. D., Robinson, A. J. & Goble C. A. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* **19** (Suppl 1), i302–i304 (2003).
16. Qiao, W., McLennan, M., Kennel, R., Ebert D. S., & Klimeck, G. Hub-based simulation and graphics hardware accelerated visualization for nanotechnology applications. *IEEE Trans. Vis. Comput. Graph.* **12**, 1061–1068 (2006).
17. Stein, L. D. *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res.* **12**, 1599–1610 (2002).
18. Mungall, C. J., Emmert D. B. & FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23**, i337–i346 (2007).
This paper describes a cyberinfrastructure approach built on a tightly coupled shared common-data model.
19. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
20. Noy, N. F. *et al.* Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu. Symp. Proc.* **2003**, 953 (2003).
21. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38**, 500–501 (2006).
22. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
23. Sotomayor, B. & Childers, L. *Globus Toolkit 4: Programming Java Services* 1st edn (Morgan Kaufmann, San Fransisco, 2005).
24. Wilkinson, M. D. & Links, M. BioMOBY: an open source biological web services proposal. *Brief Bioinform.* **3**, 331–341 (2002).
25. Wilkinson, M., Schoof, H., Ernst, R. & Haase, D. BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiol.* **138**, 5–17 (2005).
This paper describes a large-scale attempt to integrate multiple resources using web services.
26. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
This is the foundational paper for the Gene Ontology, a system for describing the molecular function of genes in a way that allows gene-based resources to be integrated at the semantic level.
27. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
28. Lacy, L. W. *Owl: Representing Information Using the Web Ontology Language* (Trafford Publishing, Victoria, Canada, 2005).
29. Oinn, T. *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054 (2004).
This paper describes Taverna, an exemplar platform for integrating bioinformatics workflows across loosely coupled sites and technologies that share common semantics.
30. Lord, P. *et al.* Applying semantic web services to bioinformatics: experiences gained, lessons learnt. International Semantic Web Conference 350–364 [online], <http://www.cs.man.ac.uk/~hulld/papers/applying_semantic_web_services_to_bioinformatics.pdf> (2004).
31. Buck, M. J. & Lieb, J. D. ChIP–chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349–360 (2004).

Acknowledgements

I wish to thank the staff of myGrid, BIRN, caBIG, iPlant, EcoliHub and nanoHub for their assistance during the research phase of this Review. I would also like to thank the three anonymous reviewers who took the time to review this article in manuscript stage and to make comments and suggestions. This work was supported in part by a grant from the National Science Foundation Division of Emerging Frontiers (0735191).

Competing interest statement

The author declares competing financial interests: see web version for details.

FURTHER INFORMATION

WIKI features and commenting: http://nrgwiki.nature.com
Lincoln Stein's homepage: http://stein.cshl.org/~lstein
BioMOBY: http://www.biomoby.org
Biomedical Informatics Research Network (BIRN): http://www.nbirn.net
Cancer Bioinformatics Grid (caBIG): https://cabig.nci.nih.gov
EcoliWiki: http://ecoliwiki.net
Ensembl Genome Browser: http://www.ensembl.org/index.html
EU Framework Programme 7 (FP7): http://cordis.europa.eu/fp7/home_en.html
Generic Model Organism Database (GMOD) project: http://gmod.org
Globus Toolkit: http://www.globus.org/toolkit
iPlant Collaborative (iPC): http://iplantcollaborative.org
myExperiment: http://www.myexperiment.org
myGrid: http://www.mygrid.org.uk
NeuroNames: http://braininfo.rprc.washington.edu/nnont.html
NCBI Taxonomy: http://www.ncbi.nlm.nih.gov/Taxonomy
Open Bioinformatics Ontologies (OBI): http://obi.sourceforge.net
Phenotype & Trait Ontology (PATO): http://www.bioontology.org/wiki/index.php/PATO:Main_Page
Simple Object Access Protocol (SOAP): http://www.w3.org/TR/soap
Simple Semantic Web Architecture and Protocol (SSWAP): http://sswap.info
SSWAP protocol: http://sswap.info/protocol.jsp
UCSC Genome Browser: http://genome.ucsc.edu
WikiPathways: http://www.wikipathways.org
Web Services Description Language (WSDL): http://www.w3.org/TR/wsdl
Virtual Plant Information Network (VPIN): http://vpin.ncgr.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF