Cyberinfrastructure for PathoSystems Biology

Bruno W.S. Sobral

Virginia Bioinformatics Institute at Virginia Tech (0477), Blacksburg VA USA 92024
sobral@vt.edu
http://www.vbi.vt.edu

Abstract. The application of new information and biotechnologies to infectious disease research provides an opportunity to design, develop and deploy a comprehensive cyberinfrastructure for life sciences. The application of integrative approaches including theory, wet experimentation, modeling and simulation and the leveraging of a strong comparative, evolutionary framework has spawned pathosystems biology. I will show examples of how cyberinfrastructure is being developed and used to support pathosystems biology.

1 Introduction

The application of modern information technologies and biotechnologies (including genome-scale approaches, systems biology, etc.) in the context of infectious diseases has spawned a new way to augment our understanding of infectious diseases, as well as new opportunities to leverage the knowledge and apply it to the development of countermeasures (surveillance, vaccines, therapeutics, diagnostics, etc.) to help protect the global community from attacks by infectious agents (of plants, animals, and humans). This paper will focus on these concepts in the context of the research and development programs I am responsible for implementing.

1.1 Cyberinfrastructure

The Atkins Report on cyberinfrastructure (CI) recalled how infrastructure in general is taken for granted until it stops functioning [1]. For life scientists, thinking about infrastructure is novel in most cases, although the need and power of infrastructure has been shown to most life scientists through the Human Genome Project. Many have pointed out how infrastructure is complex and expensive and should be built specifically by groups capable of developing infrastructure. CI refers to infrastructure based upon distributed computer, information and communication technology. Furthermore, CI is required for a knowledge economy, and biological knowledge is required to support the needs of infectious disease research and development. CI technologies are the components of computation, storage, and communication; also the software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities, in the case of infectious diseases, microbiology and related bioscience fields (for example those that consider the effects of pathogens on hosts, such as immunology, plant pathology, etc.). Furthermore, there is the layer of enabling hardware, algorithms, software,

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 11 – 27, 2005.

communications, institutions, and personnel. This crucial layer enables specific communities of researchers to innovate and eventually change what they do, how they do it, and who participates. This last layer requires institutions with service-oriented staff and core facilities to provide operational support and services, as well as high-impact applications of CI in relevant areas of science and engineering research and allied education. I believe that infectious diseases provide a high-impact arena in which to develop and deploy CI for life sciences. Infectious disease biology is ready for CI because full deployment of a working system to support public health and biodefense will require grids of computational centers, libraries of digital objects, including software programs and literature, multidisciplinary, well-curated federated collections of scientific data, thousands of online instruments and distributed sensor arrays, convenient software toolkits for resource discovery, modeling, and interactive visualization, and the ability to collaborate with physically distributed teams of people using all of these capabilities, in real-time or quasi-real-time. These are specifically what the Atkins Report characterizes as the vision for CI. Finally, as noted by that report, this "vision requires enduring institutions with highly competent professionals to create and procure robust software, leading-edge hardware, specialized instruments, knowledge management facilities, and appropriate training."

1.2 Pathosystems Biology

Infectious diseases are caused by the interaction of hosts, pathogens, and environmental factors. It is not possible to speak about disease outcomes meaningfully without specifying these factors; thus, a pathogen is not equivalent to a disease and most pathogens are not capable of infecting most organisms (i.e., most organisms are non-hosts of a given pathogen). Therefore, it is common for example in plant pathology to speak of a "pathosystem" when referring to the interaction of hosts, pathogens, and their environments. Some argue that this "disease triangle" (Figure 1) does not apply to animal systems because the environment within the animal is somewhat constant. I would say that even if this is believed to be the case, the epidemiological level clearly involves environmental factors even for animal systems. Systems biology is a relatively new term that can be seen as an extension and modernization of cybernetics [2] Many definitions exist for systems biology, but in my opinion it is characterized by an approach that fully integrates modeling, simulation, theory and wet chemistry experimentation in a unified, multidirectional feedback loop (i.e., theory effects modeling, modeling affects how you design wet chemistry experiments, and so on, in all possible combinations). Taking together the disease triangle as a comparative biological focus area and using a systems biology approach yields the term "pathosystems biology". The comparative aspect is crucial to increase our understanding of pathosystems because evolution re-uses successful components for other needs of the organism (wings may become flippers, for example). At the level of the ongoing molecular arms race that hosts and pathogens engage in, this is well documented [3]. Comparative approaches also may provide crucial benefits because some systems are more tractable to experimentation in the laboratory than others and some of the successful components (of host response or pathogen attack) may be more easily revealed in some systems when compared to others.

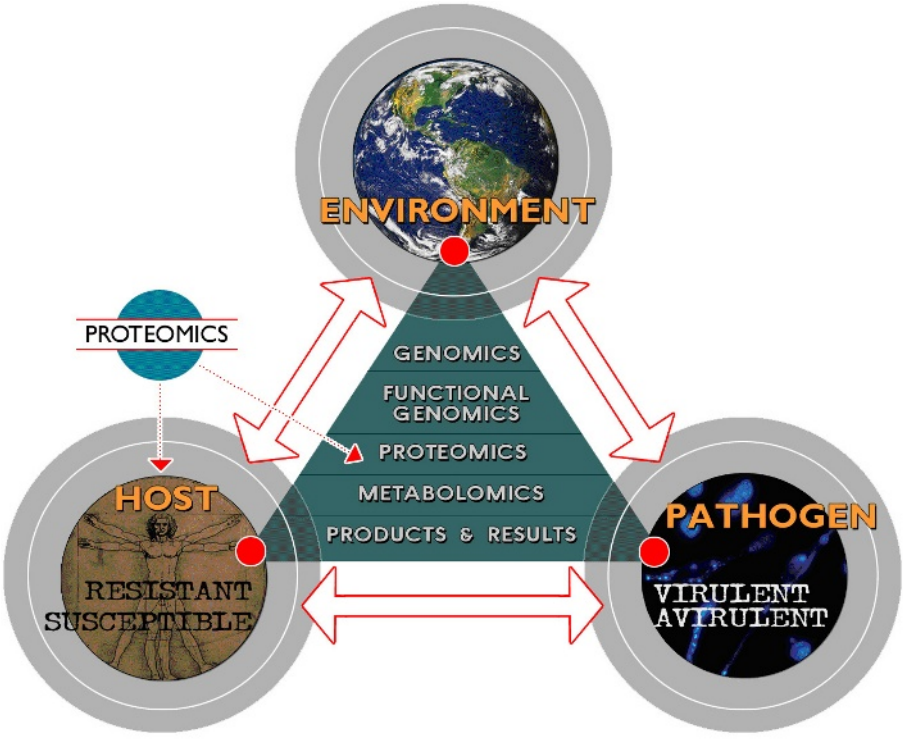

Fig. 1. Host, pathogens, and the environment interact at diverse levels in what is known as the "disease triangle". In the center is a triangle illustrating the use of molecular signatures of DNA, mRNA, proteins and metabolites as but some of the types of data that can be provided through generation, analysis and management of these data, along with the human resources to use the information in pathosystems biology

2 Some Components of Cyberinfrastructure to Achieve Synthesis in Pathosystems Biology

Part of the overall plan for infectious disease monitoring will of necessity reside in data management and analysis capabilities. Coincidentally, as the explosion of types and volume of data occurs, there is an ongoing change in software architectures that support data integration and interoperation. Briefly, in the 1990s, client-server applications changed information systems. Prior to the 90s, mainframes were the norm – these were replaced by client-server architectures with the rise of the PCs. On the software side, vendors released client-server applications, yielding enterprise applications. From the user's perspective, these changes brought end-users into the dialog for the first time. So, IT departments came out to affect all departments in an organization; this is true for scientific organizations as well. Now, there is an evolution from client-server to web-services (see below for characteristics). Web services are enabled because of agreement on standards across a very broad range of

hardware and software organizations, (for example, W3C and webservices.org). From the perspective of the mission that must be accomplished to support pathosystems biology, this technological advance is enabling because there is a huge need for information systems interoperation to support collaboration across organizations and real-time information access and analysis, whether it be for public health or biodefense needs.

The catalyzing force behind web services is the agreement by major software and hardware vendors on standards for communication between computer systems, building off the foundation of the Internet (TCP/IP, http and XML). The Internet removed the communication/information bottleneck for information consumers in the client-server model. Web services promise to relieve the information and communication barriers that limit organizational collaboration (because of barriers caused by proprietary, non-interoperable information systems that were independently developed under client-server models). Despite what we see on TV, trans-federal agency or trans-research institution information system interoperability is largely not possible with current architectures without dramatic investments in integration.

By definition, web services are characterized by being: 1) loosely coupled; 2) self-describing (WSDL[1]); 3) accessed programmatically (SOAP[2]); 4) network distributed; and 5) exchange data using platform, vendor and language-neutral protocols. These characteristics provide: flexibility and ease of reconfiguration (1); the software rather than the user determines how to invoke the service and what results the service will return (2); access via Internet protocols and data formats complying with security measures and policies, such as firewalls, allowing deployment and access across intranets as well as Internet (3); data exchange via vendor, platform and language-neutral protocols, due to broad agreement on standards (4).

There are many resources being funded through diverse federal agencies that could be wrapped to become part of a web-services architecture for pathosystems biology. This could be done by other methods, but non-web-services-based integration efforts have been widely used and are appropriate in some mixture with web services, especially in the initial phases of implementation of novel approaches for life sciences data interoperation. Typical approaches include (Marks 2003):

- Ad hoc custom integration – heavily based on individual skills.
- Data warehouses and data marts – develop high quality products based on snapshots of data (frozen in time) and periodic extraction into a common system.
- Enterprise application integration (EAI) – a replication-based middleware approach, tying key systems together.

The above approaches are powerful but can suffer from well-know problems, even outside the technical scope. These are typically (Marks 2003): 1) the requirement for very significant investments in time and money, reducing funds for other, more stra-

[1] Web Services Description Language.
[2] Simple Object Access Protocol.

tegic activities; 2) poor quality data, caused by the lack of definition of standards in the master resources, thereby causing additional time and money investments in cleaning up the data; 3) limited operational visibility, especially in life sciences since there is little understanding and comprehension by most life scientists of the problem at hand and the cost of enterprise integration for example – this has the very negative effect of spending a lot of time trying to get the integration itself right, rather than focusing on the data analysis (the reason for integration); and 4) lack of flexibility, since the above approaches result in tightly coupled systems with reduced operational flexibility – this is perhaps the most severe problem for life scientists since the technologies and underlying data are evolving very rapidly.

Pathosystems biology requires the utilization of diverse types of data that are acquired through standard processes, frequently in distributed locations. Early responses to natural, accidental, or intentional infectious disease outbreaks will require that this information be easily accessed in real-time or near real-time if we are to respond effectively to outbreaks [4]. In addition, technologies for data production are rapidly evolving, especially with respect to machinery and techniques to collect high-resolution data about molecular constituents of living cells (DNA, mRNA, proteins and metabolites, for example, see Figure 4), which may be used to develop signatures of the presence of pathogens. Technologies (laboratory and IT) are thus evolving much more quickly than institutions. Meanwhile, biological knowledge and expertise is distributed organizationally throughout the country and globe, requiring broad community involvement to meet the challenges of infectious diseases in the 21st century. Finally, excellent legacy systems composed of data and analysis/visualization tools are "out there", requiring information system architectures that leverage "old" and enable rapid deployment of "new". All of this argues for flexible, decentralized, modular information system architectures to suit evolving requirements and rapid response – and this is precisely what web services enable.

Distributed data systems, analysis tools and infectious disease expertise require strong collaboration to be in place if we are to respond to infectious diseases rapidly and effectively. In life sciences, collaboration is becoming the norm rather than the exception, although many biologists are still evolving sociologically to accommodate this situation, especially in academia[3]. The goal of collaboration is to establish, maintain and strengthen connections to achieve common objectives. Many of these connections are people to people connections, and these are likely the most important. Yet we must also increase the people to data content, people to applications, and applications to applications to content to applications connections – and these are the ones that web services can enable. In all cases, though, we must not loose sight of the need to understand the social networks[4] [5]

The Internet alone is insufficient to support the type of organization-to-organization collaboration that is needed for pathosystems biology. This is because

[3] NIH Roadmap can be obtained at http://nihroadmap.nih.gov/; more directly related is a subset of the Roadmap developed by the BECON Symposium on Catalyzing Team Science at http://www.becon.nih.gov/symposium2003.htm

[4] The "Atkins Report" on Revolutionizing Science and Engineering through Cyber-Infrastructure can be accessed at http://www.communitytechnology.org/nsf_ci_report/

there is a lack of standards for integration and automation. In addition, manual web browsing and searching does not scale well when there is a need to know about and access diverse information systems – web services provide registry-based applications that find one another and auto-invoke at run time to create larger applications serving specific needs from components that may be used for other purposes and that may reside in distributed machines. Distributed development of biological data sets and analysis tools has been the hallmark of the development of most bioinformatics[5] and computational biology[6] systems thus far – so another advantage of web services approaches is that they leverage what has already been done without the need to invest large sums of money and time into enterprise integration of such components into brittle systems that cannot easily evolve further.

2.1 Bioinformatics, Computational Biology and Community Standards

Bioinformatics and computational biology have grown over the last twenty or so years and through this growth diverse database systems and analytical tools have been developed and deployed, mostly by single investigators or small groups of investigators working together on specific biological problems. Some community resources, such as GenBank, have become key enablers of research on a global scale. The power of this distributed approach to development is that innovation has blossomed at various levels. The challenge is that there have been relatively few concerted efforts to standardize data formats, thus hindering efforts to integrate disparate data types from diverse data sources. Paradoxically, further synthesis in biology largely depends on the capability to access and jointly analyze disparate data. This is especially true for pathosystems biology, since it must deal with data from many types of organisms (pathogens and their hosts) in diverse environments (from intracellular to ecosystems and social networks).

Yet, there are important efforts to develop and deploy community standards for biological data communication. It is important that these efforts be supported and succeed in developing at least data exchange standards for the sake of interoperability across information systems that matter to microbial forensics, whether in existence, or to be (being) developed. The web services stack builds on and extends the standards of the Internet. At the lowest level, there are network protocols (such as TCP/IP, HTTP, FTP, SMTP). The next level is concerned with the meta language (XML). This is where diverse community-based efforts are providing useful standards. Going from DNA through molecules that permit an assessment of the dynamic response of the organism to perturbations, as well as capabilities for modeling and simulation, we have (not meant to be exhaustive):

---

[5] "Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data." http://www.bisti.nih.gov/

[6] "The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems." http://www.bisti.nih.gov/

- DNA – DAS-ML[7], BSML[8], MSA-ML[9]
- RNA – MAGE-ML[10] (mRNA profiling)
- Proteins – PEDRo[11] (protein profiling) and ProML[12] (protein sequences, structures and families)
- Molecular models – SMBL[13]
- Cellular levels, including metabolism and signal transduction – CellML[14]
- Organ level – AnatML[15]
- Spatially and temporally varying field information using finite elements – FieldML[16]

Furthermore, there is a need to handle data at the level of phenotypes displayed by organisms. In the case of humans, this is typically placed within clinical records. Fortunately, there are efforts in place to handle these data using XML standards, such as the Clinical Data Exchange Standards Consortium[17] (CDISC). Finally, to handle data from molecular responses to perturbations, through phenotypes and into geographic space (required for epidemiological monitoring and global molecular epidemiologies), there is ArcXML[18] and OpenGIS[19].

There is a danger that of fragmentation of standards through a diversity of non-interacting groups building competing XMLs to represent essentially the same data. Avoiding this will take some vigilance and incentives from funding agencies and requirements for machine-readable interfaces to major resources that are built with federal funding. At some point in the future there will be sufficient advantage through achievement of interoperation that implementations that do not conform to those standards will not be competitive or generally useful.

The network protocol and XML layers are fairly stable technologically and therefore can be thought of as enabling at this point. Above this layer lie three crucial layers that are still undergoing some evolution. These are the services communication layer (SOAP), the services description layer (WSDL), and services publishing and discovery (UDDI[20]/OGSA[21]). These three layers are still evolving and web services implementors need to understand the risks associated with evolution away from the

---

[7] Distributed Annotation System Markup Language, http://stein.cshl.org/das/.
[8] Bioinformatic Sequence Markup Language, http://www.bsml.org/.
[9] Multiple Sequence Alignment Markup Language, http://xml.coverpages.org/msaml.html.
[10] Microarray Gene Expression Markup Language. http://www.mged.org/Workgroups/MAGE/mage.html.
[11] http://psidev.sourceforge.net/.
[12] Protein Markup Language, http://www.bioinfo.de/isb/gcb01/talks/hanisch/main.html.
[13] Systems Biology Markup Language, http://sbml.org/index.psp.
[14] http://www.cellml.org/public/about/what_is_cellml.html.
[15] Anatomical Markup Language http://www.physiome.org.nz/anatml/pages/.
[16] http://www.physiome.org.nz/fieldml/pages/.
[17] http://www.cdisc.org/.
[18] http://support.esri.com/.
[19] http://www.opengis.org/.
[20] Universal Description, Discovery and Integration protocol, http://www.uddi.org/about.html.
[21] Open Grid Services Architecture, http://www.uddi.org/about.html.

currently accepted standard. Finally, the most rapidly evolving layers, comprising of still emerging standards, are the business process execution (BPEL4WS[22], WFML[23], WSFL[24], Biztalk, etc.) and additional standards such as WSXL[25].

Another major need is with respect to ongoing curation of data that requires specific biological knowledge, such as much of the microbial data will require. This is especially important because of the distributed nature of biological knowledge in the field. Although funding is limiting in most cases, there are models for supporting distributed curation among specialists.

Sample Cyberinfrastructure for Pathosystems Biology Projects. One model for distributed curation in pathosystems biology has been prototyped on a limited scale in the Pathogen Portal (PathPort[26]) [6]. PathPort project has developed and deployed the Pathogen Information (PathInfo[27]) resource containing data from about 20 of the 50 pathosystems for which acquisition of highly curated data sets referenced from the literature has been requested. One output of the literature curation effort is the Pathogen Information Markup Language or PIML [7], which can now be used further by a distributed community of experts to enter similar data about other pathosystems into a common, machine-readable format. Figure 2 illustrates PIML architecture; figure 3 shows how distributed data acquisition and dissemination is managed in the context of scientific literature and molecular data sets; this is being further developed and deployed under the recently funded Bioinformatics Resource Centers[28] (BRCs) funded by NIAID to develop the capabilities to support genomic data for NIAID category A, B and C pathogens. The goal of the BRCs is to work on the pathogen side of the genomic data management and interoperation issues. To produce, acquire, integrate, manage, analyze and disseminate proteomics data about pathogens, NIAID has recently awarded contracts to establish the Biodefense Proteomics Research Centers[29]. An integral Administrative Resource for Biodefense Proteomic Centers[30] will be responsible for centralized data management for the network.

A number of efforts are now using PathPort's CI (which includes a Core Laboratory[31] and a Core Computational Facility[32] at the Virginia Bioinformatics Institute but they could be anywhere, based on the web-services paradigm). For example, PathPort + Core Computational Facility + Core Laboratory Facility now provide the Bioinfor-

---

[22] Business Process Execution Language for Web Services. http://www-106.ibm.com/developerworks/library/ws-bpel/.
[23] Windows Forms Markup Language, http://windowsforms.net/articles/wfml.aspx.
[24] Web Services Flow Language, http://xml.coverpages.org/wsfl.html.
[25] Web Services Experience Language. http://www-106.ibm.com/developerworks/library/ws-wsxl/.
[26] https://www.vbi.vt.edu/article/articleview/316.
[27] http://staff.vbi.vt.edu/pathport/pathinfo/.
[28] http://brc.vbi.vt.edu/.
[29] http://www.niaid.nih.gov/dmid/genomes/prc/default.htm.
[30] http://www.niaid.nih.gov/dmid/genomes/prc/administrative.htm.
[31] https://www.vbi.vt.edu/article/articleview/87.
[32] https://www.vbi.vt.edu/article/articleview/88.

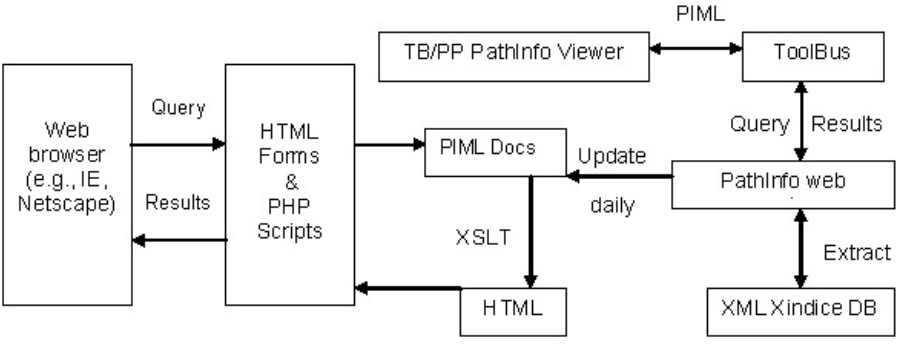

Fig. 2. A web query starts with specifying a particular topic and pathogen(s). Requested pathogen PIML documents are parsed and the results are transformed into HTML by an XSLT script. The PIML documents are updated daily from the Xindice DB via the PathInfo web service. Corresponding viewer is also available via TB/PP system

matics and Genomics Research Core (BGRC[33]) for the Mid-Atlantic Regional Center for Biodefense and Emerging Infectious Diseases (MARCE[34]), funded by NIAID. In this large-multi-institutional, multi-investigator program, part of a national network funded by NIAID this year, the main objective is to develop diagnostics and countermeasures for infectious agents on NIAID category A and B priority lists. The general functional model of the BGRC is illustrated in Figure 4. In the context of MARCE, other CI components, such as the MARCE website[35], supporting external visibility for the project as well as "intranet" functionalities for real-time communication are available as well. These capabilities are meant to support a range of activities, from real-time video conferencing within MARCE and from MARCE to other RCEs as well as interactive tools supporting document preparation, discussion of data, presentations, etc., with the goal of a vibrant, functional CI for pathosystems biology. As different agencies and scientists working on different aspects of infectious diseases use and help evolve the CI, one of the benefits that will come out of the infrastructure, without additional investment, is the possibility of doing joint analyses on data sets that were developed with specific goals in mind but that can be useful to other goals. The success of GenBank in enabling comparative analyses of community sequences because of deposition into a standardized repository is but an example of what can be aspired by the infectious disease CI being developed and deployed.

One of the many reasons for using a web-services, federated approach is the leveraging, with relatively little effort, of key resources being built in the community. It is not possible here to provide an exhaustive review of these, but clearly efforts such as the Microbial Rosetta Stone Database (MRS) project (K.L. Hari, J.A. McNeil, IBIS

[33] http://marce.vbi.vt.edu/cores/bioinformatics_and_genomics_core.
[34] https://www.vbi.vt.edu/article/articleview/426/1/33/.
[35] http://marce.vbi.vt.edu/.

Pharmaceuticals; and J.M. Robertson, FBI; personal communication) are aimed in the right direction. MRS has been motivated by the need to map the landscape of infectious diseases and to assist with microbial forensics needs, specifically. Another interesting resource is Gideon Online[36]. This system has been developed essentially to assist in diagnosing (at the clinical level) infectious agents based on information collected by the clinician and a Bayesian analysis system. It is continually updated and has information on all infectious agents of humans and related mammals, and also a recently released bioterrorism module. It has also been used for training and teaching of physicians. Models could be developed to support further documentation and referencing of the system to the scientific literature and online, real-time update by distributed experts that start to then use the system for data entry to support monitoring.

The PathPort project itself has been federating through web services diverse data sources and analysis tools to support the needs of (currently and primarily) discovery scientists working on developing a more comprehensive knowledge of the mechanisms that infectious agents and their hosts deploy in their interactions (an "arms race"). The client-side interconnect for the federated services, ToolBus (Figure 5), allows users of the system to access and analyze (mostly molecular currently) data of diverse types from diverse sources. The overall architecture of the PathPort system is shown in Figure 6 and the architecture of the client-side interconnect, ToolBus, is shown in Figure 7.

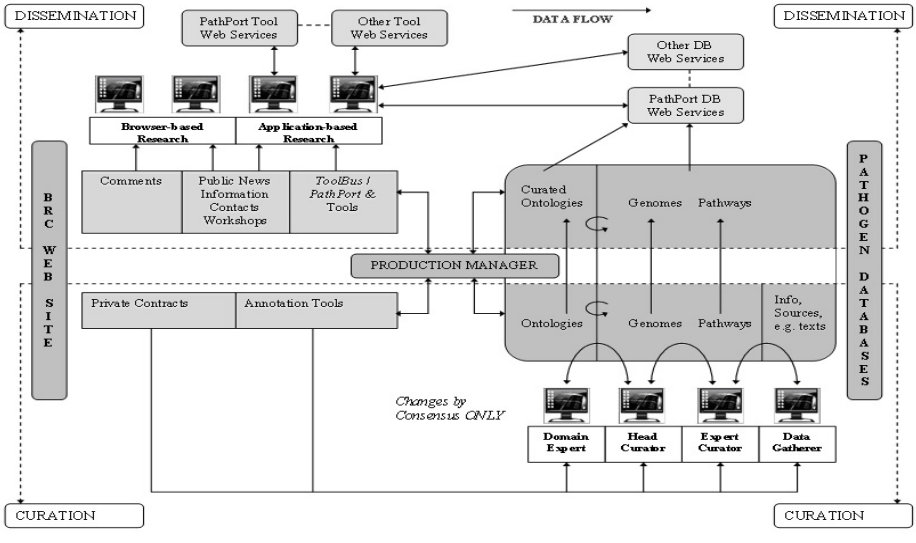

Fig. 3. A model for distributed curation involving subject matter experts throughout the community and showing how many of the (molecular) data types are dealt with, along with the CI needed to ensure that the data are acquired and disseminated appropriately

[36] http://www.gideononline.com/.

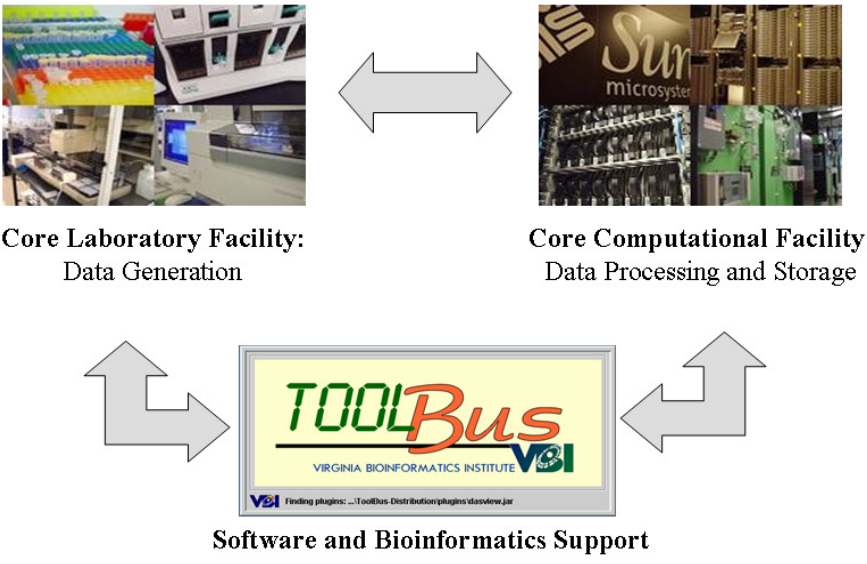

Fig. 4. CI supporting data generation, acquisition, analysis, storage for MARCE. Note that although the physical capacities provided by thee cores happen to reside at the same institution in this case, one can envision a number of such facilities distributed, working under standard operating and quality assurance procedures, and supported by the interoperability middleware such as provided by PathPort's implementation of a web services strategy and ToolBus for the client-side interconnect. Also note that analysts providing training support and where appropriate conducting analyses collaboratively with a distributed set of partners is integral to this model but not shown in the figure

PathPort project is following some of the typical phases explored in web services adoption. These are: 1) integration/interoperation, 2) collaboration and 3) innovation [8]. PathPort project is in the first phase, primarily, and exploring the second phase. The first phase typically involves building wrappers around legacy systems and applications. During this first phase, the project has embraced fast cycles of development and deployment[37] with opportunity for community involved in the rapid cycles of learning. The goal has been to deploy early and often to allow users to react and participate effectively with the software development team. This has resulted in sharing of information across collaborators and mutual learning. During this phase the CI team and its collaborators sometime encounter limits based on immature standards and unprepared IT architectures. With the coming of the second phase, collaboration, we eventually expect a reduction in the levels of human intervention required to support collaboration. Finally, as being experienced in the PathPort project, "external" partners start to increase in their sharing and collaboration thus further driving the

[37] See http://staff.vbi.vt.edu/pathport/scrum/ for information about the SCRUM/SPRINT process being employed to agilize software development.

development/implementation/evolution chain. In the innovation phase, we hope to use the lessons learned from the previous phases to drive entirely new processes and models. New, distributed web-services models tend to be disruptive and thereby enable change. We hope, in good CI form, that there will be a redefinition of how research is conducted across organizational boundaries, something I believe the MARCE project, the NIAID RCE Network, the BRC Network and the Biodefense Proteomic Research Centers[38] can help prototype both within their own networks as well as across networks. This redefinition is sorely needed and enabled by exposing specific operational information system elements for dynamic linking to processes of partners/collaborators. The goal is to have organizations operating as a truly inter-

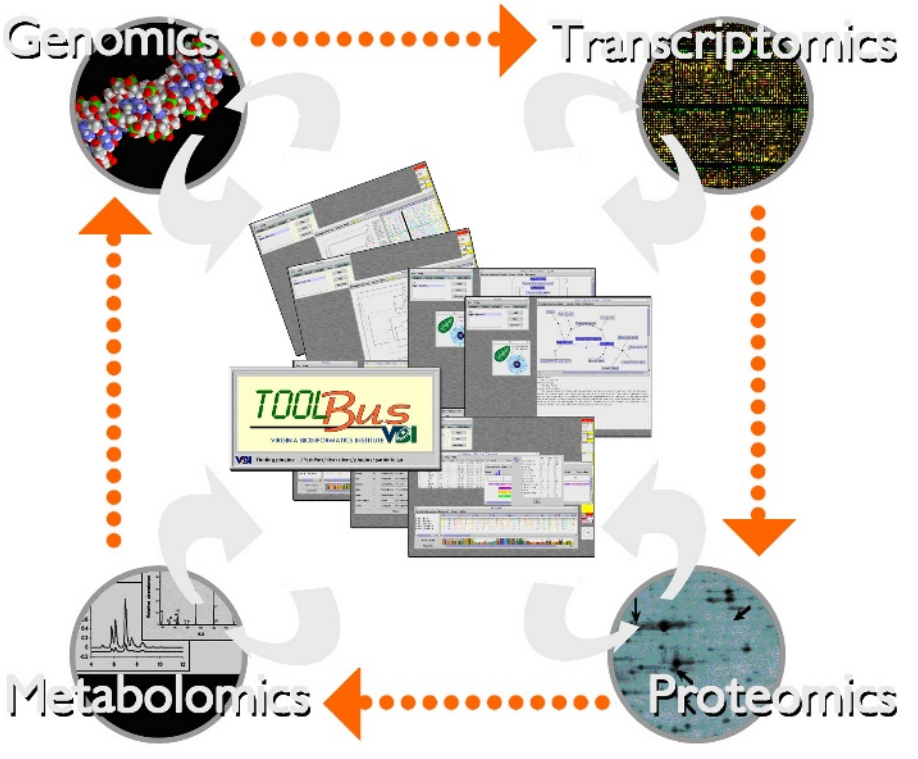

Fig. 5. "-Omics" data provide the opportunity to develop the "parts lists" for pathogens and their hosts (genomics data), along with the contextual "state" data that describe the dynamic molecular responses of living organisms (pathogens and hosts) as they respond to each other in a given environmental condition (transcriptional profiles or transcriptomics data, protein profiles or proteomics data and metabolite profiles or metabolomics data). These data sets not only will allow molecular signatures to be developed, they will also help establish a mechanistic understanding of infectious agents attacking their hosts, thereby enabling development of new countermeasures, such as vaccines and therapeutics

[38] http://www.niaid.nih.gov/dmid/genomes/prc/default.htm.

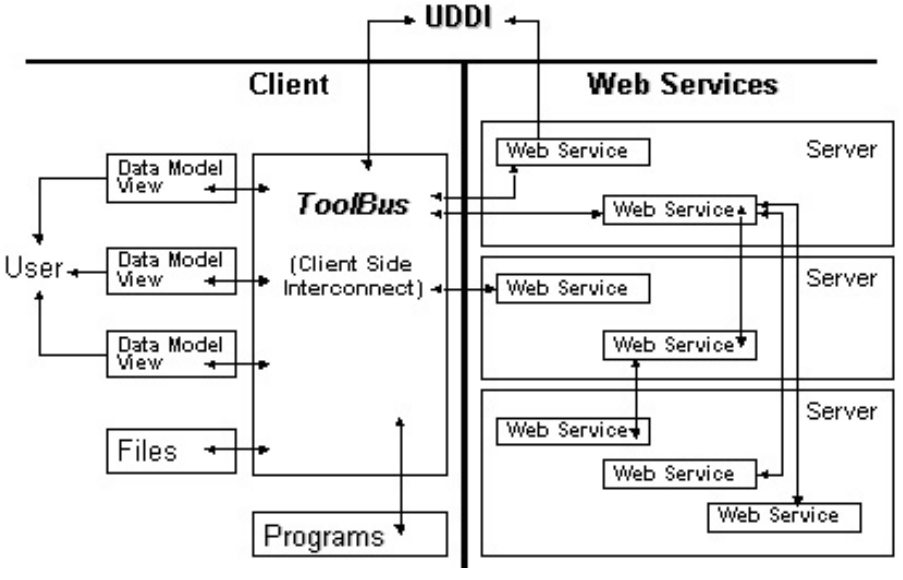

Fig. 6. A simplified view of the architecture employed by PathPort project to allow interoperability of diverse analysis tools and data sources of relevance to infectious diseases. Web services can be either analysis tools, such as BLAST, or a data source, such as GenBank. They can reside anywhere. Local files (available to the local user only), whether programs or data, can be used without making them available to the entire federation if desired

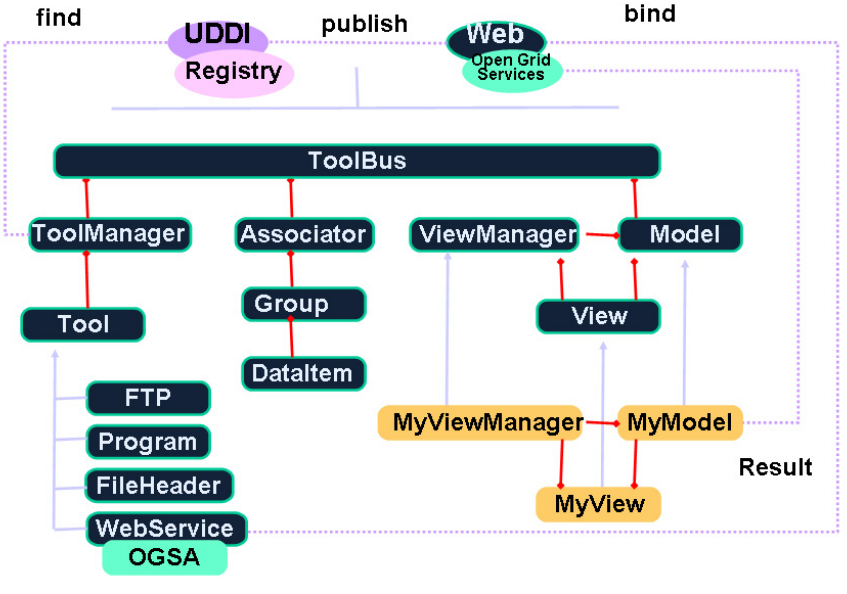

Fig. 7. Architecture of the client-side interconnect, ToolBus, that allows for access of web-services relevant to PathPort project. Note that new data models can be added easily without breaking the system or requiring major re-engineering

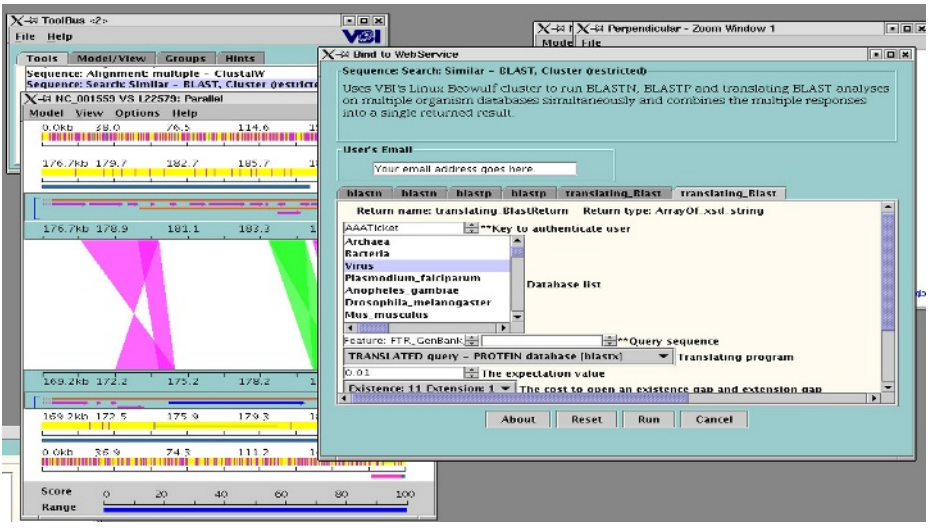

Fig. 8. A simple example of interoperability across previously incompatible systems (DAS Viewer, BLAST, MUMer) built by the community as well as of allowing communication between the visualizer for a given analysis result (in this case comparative genomics between *Vaccinia* and *Variola*) through a drag-and-drop approach into another analysis server (in this case BLAST). A web-services architecture separates the building of visualizers from databases such that new visualizations can be achieved easily. The communication across incompatible systems enables a much faster and more efficient workflow for the human knowledge worker/operator/analyst

connected cyber-ecosystem. The newness of these research networks provides a unique opportunity to develop this from the beginning, if this an objective that is adequately and integrally planned.

One question that frequently arises with infectious disease research and data in our post-9/11 world is security. There are many different levels of need to security. From an IT perspective, web-services can provide security via models being developed and implemented, such as the WS-Security[39] or OASIS WS Security TC[40]. Importantly, again, is to leverage community standards for implementation. Although some of the needs may be national security related, it is important to note that most life sciences companies, such as Pharmaceuticals and biotechs, have very stringent security needs because of Intellectual Property requirements. (This is to say that there are meaningful solutions that can leverage web services and be enabling all the same, based on specific requirements.) The Intel community is already implementing prototypical projects in this direction, noting[41]: "In a network-centric envi-

[39] http://www-106.ibm.com/developerworks/webservices/library/ws-secure/.
[40] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wss.
[41] See http://fcw.com/fcw/articles/2003/0317/web-nces-03-18-03.asp, for example.

ronment, data would be made available as quickly as possible to those who need it across the organization or on the battlefield. Many DoD systems in the field today use a client/server architecture." and "…would create an infrastructure that will enable users to quickly take advantage of DoD and intelligence community networks, eliminating the system-by-system approach"…"The system will enable users to customize the way they search and actually view information in real-time and display previously unavailable combinations of intelligence, surveillance and reconnaissance data. Access based on individual users' security clearances will be built into the design." Thus there is nothing specific about web-services that will not support security as needed.

In the three years of development experience provided by the PathPort project, interoperability across previously incompatible systems, using web-services, has already been implemented and used by scientists (Figures 8 and 9). In the future, as we move toward the innovation phases of development, ideas and concepts that support large-scale simulations of real-world events pertaining to infectious disease outbreaks (Figure 10) will be possible.

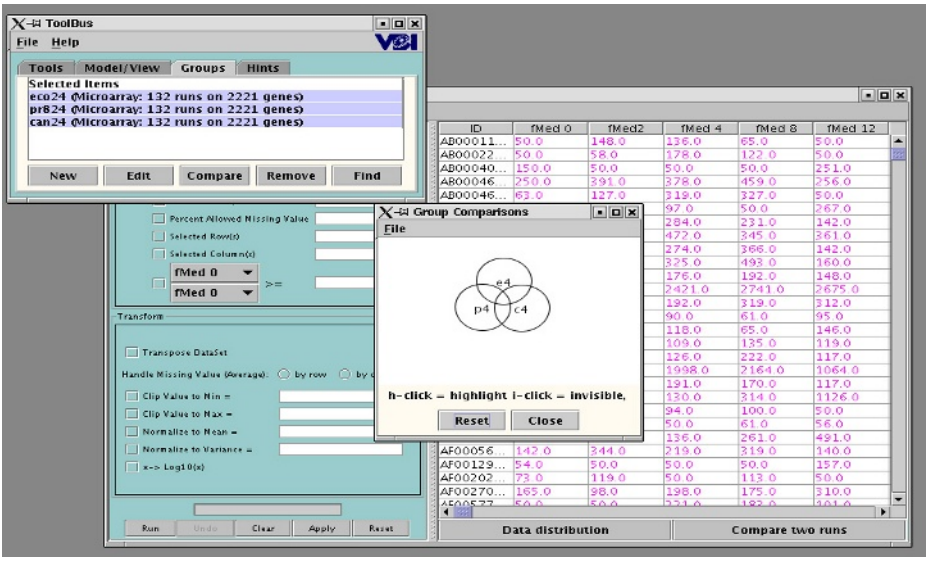

Fig. 9. ToolBus use case ToolBus showing the group suggestor function working on a set of transcriptional profiles. It is important to note that ToolBus and the "group suggestor" capability of the system do not "know" about the type of data being analyzed – although in this example all the data are of one type (mRNA expression levels), any type of data that is available in such an interoperable framework could be analyzed with the group suggestor capabilities (for example, transcriptional profiles and GIS coordinates of people or plants from which the profiles were obtained)

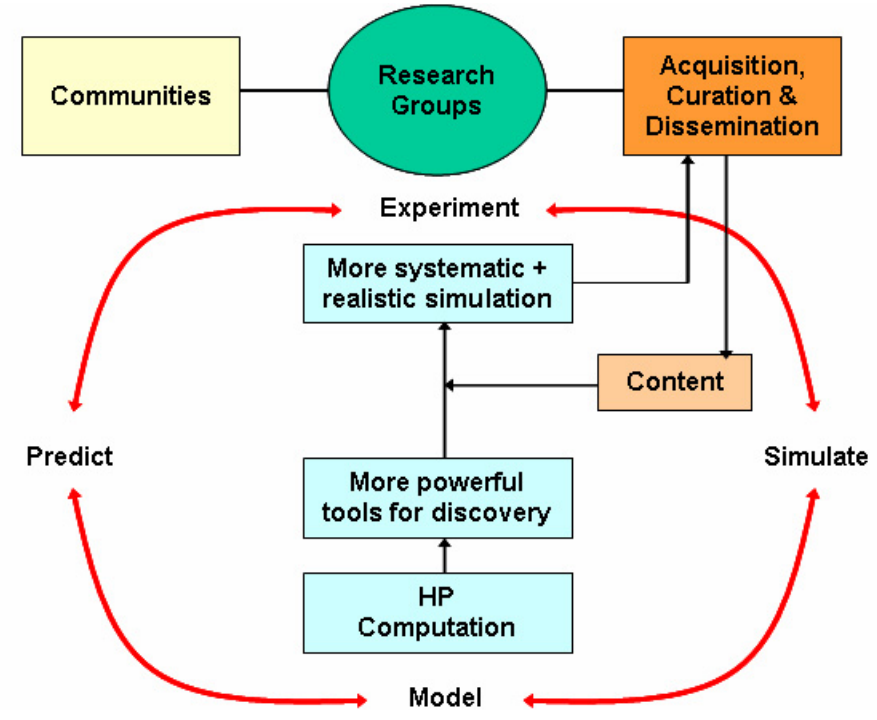

Fig. 10. An illustration of the conceptual, integrative framework for a long-term CI for patho-systems biology

Acknowledgements

I am grateful for funding by the US Department of Defense (contracts DAAD 13-02-C-0018 and W911SR-04-0045), the National Institute of Allergy and Infectious Diseases (contract HHSN266200400035C, HHSN266200400061C and Cooperative Agreement 1 U54 AI057168-01) in support of ToolBus, and PATRIC projects respectively. Special thanks to Dana Eckart, Yongqun He, Ron Kenyon, Dave Sebring and my Cyberinfrastructure Group at VBI for making this work possible. I am also grateful to Darleen Baker for assistance and editorial improvements to this paper. Finally, this work is dedicated to Minnis Ridenour, to whom VBI owes its existence.

References

1. Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerschmitt, D., Messina, P., Ostriker, J., Wright, M.: Revolutionaizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. (2003). http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203
2. Wiener, N.: Cybernetics: or Control and Communication in the Animal and the Machine. 2nd edn.MIT Press (1948) 212

3. Studholme, D. J., Downie, J. A., Preston, G. M.: Protein Domains and Architectural Innovation in Plant-Associated Proteobacteria. BMC Genomics. 6 (2005) 17
4. Eubank, S., Guclu, H., Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling Disease Outbreaks in Realistic Urban Social Networks. Nature. 429 (2004) 180-184.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15141212
5. Ibid
6. Eckart, J.D., Sobral, B.W.: A Life Scientist's Gateway to Distributed Data Management and Computing: The PathPort/ToolBus Framework. Omics. 7 (2003) 79-88.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12831562
7. He, Y., Vines, R. R., Wattam, A. R., Abramochkin, G. V., Dickerman, A. W., Eckart, J. D., Sobral, B. W.: PIML: the Pathogen Information Markup Language. Bioinformatics. (2004)
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15297293
8. Marks, E., Werrell, M. : Executive's Guide to Web Services. John Wiley & Sons, Inc., Hoboken, N.J. (2003)