CTWatch QUARTERLY

ISSN 1555-9874

Volume 2 Number 3 **August 2006**

Trends and Tools in Bioinformatics and Computational Biology

GUEST EDITOR **RICK STEVENS**

INTRODUCTION

1 **Trends in Cyberinfrastructure for Bioinformatics and Computational Biology**
Rick Stevens
Associate Laboratory Director, Computing and Life Sciences – Argonne National Laboratory
Professor, Computer Science Department – University of Chicago

FEATURE ARTICLES

6 **National Biomedical Computation Resource (NBCR): Developing End-to-End Cyberinfrastructure for Multiscale Modeling in Biomedical Research**
Wilfred W. Li, University of California, San Diego (UCSD), San Diego Supercomputer Center (SDSC)
Nathan Baker, Washington University in Saint Louis
Kim Baldridge, UCSD, SDSC
J. Andrew McCammon, UCSD
Mark H. Ellisman, UCSD, Center for Research In Biological Systems (CRBS)
Amarnath Gupta, UCSD, SDSC
Michael Holst, UCSD
Andrew D. McCulloch, UCSD
Anushka Michailova, UCSD
Phil Papadopoulos, UCSD, SDSC
Art Olson, The Scripps Research Institute (TSRI)
Michel Sanner, TSRI
Peter W. Arzberger, California Institute for Telecommunications and Information Technology (Calit2), CRBS, UCSD

18 **Specifications for the Next-Generation Computational Biology Infrastructure**
Eric Jakobsson, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

20 **Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade**
Folker Meyer, Argonne National Laboratory

23 **Computing and the Age of Biology**
Natalia Maltsev, Argonne National Laboratory

AVAILABLE ON-LINE AT
http://www.ctwatch.org/quarterly/

Cyberinfrastructure Technology Watch
http://www.ctwatch.org/

INTRODUCTION

Trends in Cyberinfrastructure for Bioinformatics and Computational Biology

Rick Stevens
Associate Laboratory Director, Computing and Life Sciences – Argonne National Laboratory
Professor, Computer Science Department – The University of Chicago

In this issue you will find a number of articles outlining the current trends and tool development strategies in the bioinformatics and computational biology community. It is by necessity an incomplete survey of today's thinking about the directions of our field. In addition to the four submitted articles, I've enclosed my thoughts on a few of the questions likely to be of interest to CTWatch readers.

What is the most important trend today in biology research?

Probably the most important trend in modern biology is the increasing availability of high-throughput (HT) data. The earliest forms of HT were genome sequences, and to a lesser degree, protein sequences, however now many forms of biological data are available via automated or semi-automated experimental systems. This data includes gene expression data, protein expression, metabolomics, mass spec data, imaging of all sorts, protein structures and the results of mutagenesis and screening experiments conducted in parallel. So an increasing quantity and diversity of data are major trends. To gain biological meaning from this data it is required that this data be integrated (finding and constructing correspondences between elements) and that it be curated (checked for errors, linked to the literature and previous results and organized). The challenges in producing high-quality, integrated datasets are immense and long term.

The second trend is the general acceleration of the pace of asking those questions that can be answered by computation and by HT experiments. Using the computer, a researcher can be 10 or 100 times more efficient than by using wet lab experiments alone. Bioinformatics can identify the critical experiments necessary to address a specific question of interest. Thus the biologist that is able to leverage bioinformatics is in a fundamentally different performance regime that those that can't.

The third trend is the beginnings of simulation and modeling technologies that will eventually lead to predictive biological theory. Today, simulation and modeling applied at the whole cell level is suggestive of what is to come, the ability to predict an organisms phenotype computationally from just a genome and environmental conditions. That capability is probably five years away for microbial organisms and 10 to 20 years away for complex eukaryotes (such as the mouse and human).

What is the role of cyberinfrastructure in biological research?

As I noted above, modern biology will become increasingly coupled to modern computing environments. This means that rates of progress of some (but not all) biological investigations will become rate limited by the pace of cyberinfrastructure development. Certainly, it will make it much easier for the biologist to gain access to both data and computing resources (perhaps without them knowing it) once cyberinfrastructure is more developed and in place. Today, we have early signs of how some groups will use access to large-scale computing to support communities by developing gateways or portals that provide access to integrated databases and computing capabilities behind a web-based user interface. But, that is just the beginning. It is possible to imagine that, in the future, laboratories will be directly linked to data archives and to each other, so that experimental results will flow from HT instruments directly to databases which will be coupled to computational tools for automatically integrating the new data and performing quality control checks in real-time (not that dissimilar from how high-energy physics and astronomy work today). In field research, cyberinfrastructure can not only connect

researchers to their databases and tools while they are in the field, but it will enable the development of automated instruments that will continue working in the field after the scientists and graduate students have returned home.

What are some notable accomplishments in applying CI to biology research?

There are a handful of systems that have fundamentally changed how biologists work. The most important has been the system developed by the National Center for Biotechnology Information[1] including Entrez, which is a search engine (google like) that supports searching across many types of biological data. There are similar systems like this in Europe[2] and Japan.[3] These systems and systems like them have provided the global community access to sequence data (starting out as outgrowths from genome and protein sequence databases) and more recently to publications, annotations, linkage maps, expression data, phylogeny data, metabolic pathways, regulatory and signally data, compounds and molecular structures. Search techniques have expanded from keywords to computed properties (sequence similarity, and more generally "associations") that enable one to find connections between biological or chemical entities. While these systems have enormous user bases and require considerable computing capabilities for indexing and integration, they are essentially client/server in nature, and the computing that an end user can request is closely controlled.

[1] http://www.ncbi.nlm.nih.gov/

[2] http://www.ebi.ac.uk/

[3] http://www.genome.jp/

Approximately a decade ago a number of groups began to produce more flexible tools that support a more unstructured workflow, enabling the user to construct their own mini-environment to pursue computational approaches to problems. One of the first such systems was the Biology Workbench developed at the University of Illinois and now hosted at the University of California, San Diego.[4] Other systems were developed to provide access to a specific type of data (e.g. microbial genomes) in well engineering data integrations. These systems are often associated with teams of curators. Three are particularly important: the Institute for Genomic Research's Comprehensive Microbial Resource;[5] the SEED, an annotation system developed by the Fellowship for the Interpretation of Genomes at the University of Chicago;[6] and the DOE's Joint Genome Institute's Integrated Microbial Genomes resource.[7] These systems provide the user with an integrated view of hundreds of genomes and provide a rich environment for discovery.

[4] http://workbench.sdsc.edu/

[5] http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi

[6] http://theseed.uchicago.edu/FIG/index.cgi

[7] http://img.jgi.doe.gov/cgi-bin/pub/main.cgi

Are there some good road mapping documents available?

In the past couple of years there have been several worthwhile road-mapping documents written by the community. These reports in general attempt to identify the trends in the field and provide some structure for understanding directions. The first is a report from the NSF committee for building a cyberinfrastructure for the biological sciences;[8] the second is the National Academy of Sciences Report on computing and biology.[9] The third report is more oriented towards systems biology and is a program roadmap document developed by the DOE for their Genomes to Life program,[10] which contains a section on computing and infrastructure to support the building out of systems biology, focused on microbial organisms, energy, and the environment. All three documents are worth reading to gain an understanding of where the field is going.

[8] http://research.calit2.net/cibio/archived/CIBIO_Overview_Report.pdf

[9] http://darwin.nap.edu/books/030909612X/html/R1.html

[10] http://doegenomestolife.org/roadmap/index.shtml

Are grids really being used that much for real biology?

There are several national and international projects developing grid infrastructures for biological research. Many of these projects are loosely affiliated by sharing services and technology, and all are working towards a vision of a BioGrid. Several are worth looking at.

The TeraGrid is sponsoring two Science Gateways for Biology; one developed by RENCI, Biology and Biomedical Science Gateway Renaissance Computing Institute, UNC;[11] and one developed by the University of Chicago.[12] Both of the TeraGrid gateways are aimed at enabling communities to leverage the TeraGrid computing and data resources without a need for obtaining a dedicated allocation of resources. They are examples of an emerging concept of "community allocations," which are aimed at lowering the adoption barrier to cyberinfrastructure. The Open Science Grid also hosts biological applications such as the GADU virtual organization.[13] In Europe, one of the most well developed Life Sciences grid projects is MyGrid.[14] MyGrid is developing a comprehensive set of web services based tools and services.

[11] http://www.tgbioportal.org/

[12] http://lsgw.mcs.anl.gov/about

[13] http://compbio.mcs.anl.gov/gaduvo/gaduvo.cgi

[14] http://www.mygrid.org.uk/

There is a lot of talk about web services as a future direction for the Internet. How will web services impact biology?

Web services are the key to providing the ability for groups around the world to collaborate on building new tools that leverage each other's data and computational services without prior coordination. Early web services deployments in life science suffered from poor implementations, poor performance and lack of high-quality data. More recent efforts are dramatically improving. The KEGG group in Japan recently published a comprehensive set of web services[15] for accessing their data, which have proven to be robust and of moderate performance. I've used these services routinely for the last year and find them simple, yet useful. As web services interfaces become more common, it will be possible for many groups to build applications that leverage the major data sources. This is one of the most important trends, but it is still far from being generally demonstrated.

[15] http://www.genome.jp/kegg/soap/

Are there systems in use today that leverage web services?

One system worth exploring is Taverna.[16] Taverna is a collaboration between the European Bioinformatics Institute (EBI), IT Innovation, the School of Computer Science at the University of Newcastle, Newcastle Centre for Life, School of Computer Science at the University of Manchester, and the Nottingham University Mixed Reality Lab. Additional development effort has come from the Biomoby project, Seqhound, Biomart and various individuals across the planet. Development is coordinated through the facilities provided by SourceForge.net and predominantly driven by the requirements of biologists in the UK life science community. Taverna enables end users to compose bioinformatics web services in a graphical environment for computing novel workflows.

[16] http://taverna.sourceforge.net/

Is industry developing any new cyberinfrastructure tools that might alter the landscape in biology?

It is likely that several of the commercial search engine companies (e.g. Google and Microsoft) will explore the issue of coupling biological searches of open literature and databases with computational services with access to commercial tools and databases. These tools will most likely be emerging examples of coupling commercial tools (web services infrastructure, indexing and search technologies) with the best of the open science literature.

What about petascale computing?

Large-scale computational methods can address fundamental biological problems:

- The origins, function, structure, and evolutionary history of genes and genomes ⇒ *large-scale sequence analysis, sequence-based phylogenic analysis*

By studying the details of individual gene history and protein families, we can begin to understand the factors that influence molecular evolution, refine our strategies for building

large-scale databases of protein structures, and lay the foundation for understanding the role of horizontal gene transfer in evolution.

- The structure, function, dynamics, and evolution (SFDE) of proteins and protein complexes ⇒ *large-scale molecular dynamics*

Proteins are the building blocks for biological processes. Using modeling and simulation, we can begin to understand how proteins work, how they evolve to optimize their functions, how complexes are formed and function, and how we can modify proteins to alter their functions.

- Predictive protein engineering ⇒ *large-scale molecular dynamics and electronic structure*

Many processes of interest to the biological community are mediated by proteins, ranging from bio-catalysis of potential fuel stocks to the production of rare and unique compounds to the detoxification of organic waste products. Large-scale modeling and simulation can be used to attack the problem of rational protein design, whose solution may have long-term impact on our ability to address, in an environmentally sound manner, a wide variety of energy and environmental problems.

- The SFDE of metabolic, regulatory, and signaling networks ⇒ *graph-theoretic and network analysis methods and stochastic modeling and analysis techniques*

Understanding the function of gene regulation is one of the major challenges of 21st century biology. By employing a variety of mathematical techniques coupled with large-scale computing resources, researchers are beginning to understand how to reconstruct regulatory networks, map these networks from one organism to another, and ultimately develop predictive models that will shed light on development and disease.

- The SFDE of DNA, RNA, and translation and transcription machinery in the cell ⇒ *large-scale molecular dynamics and stochastic modeling*

The standard dogma of molecular biology relates the transcription of DNA to messenger RNA, which is then translated to produce proteins. This is the foundation of the information-processing operation in all living organisms. The molecular complexes that mediate these processes are some of the most complex nanomachines in existence. Via large-scale modeling and simulation of protein/RNA complexes such as the ribosome and the splisosome, we will improve our understanding of these fundamental processes of life.

- The SFDE of membranes, protein and ion channels, cell walls, and internal and external cellular structures ⇒ *large-scale molecular dynamics and mesoscale structural modeling*

Membranes are the means that nature uses for partitioning biological functions and supporting complexes of proteins that are responsible for supporting the cell's ability to interact with its neighbors and the environment. Large-scale modeling is the means by which we can understand the formation, function, and dynamics of these complex molecular structures.

- Whole-genome scale metabolic modeling ⇒ *linear-programming and optimization*

With the number of completed genome sequences reaching 1,000 in the next few years, we are on the verge of a new class of biological problem; reconstructing the function of entire genomes and building models that enable the prediction of phenotypes from the genotype. With petascale modeling it will become feasible to

quickly produce a whole genome scale model for a new sequenced organism and begin to understand the organism's lifestyle prior to culturing the organism.

- Population, community and ecosystem modeling ⇒ *numerical solution of PDEs, ODEs, and SODEs*

Large-scale computing is making it feasible to model ecosystems by aggregating models of individuals. With petascale computing capabilities, this technique can begin to be applied to natural environments such as soils and to artificial environments such as bioreactors, in order to understand the interactions between different types of organisms and their ability to cooperatively metabolize compounds important for carbon cycling.

The following table gives examples of high-impact problems that could be addressed in the next two to three years on an open access petascale platform and that leverage the methods have already been ported to the IBM BG/L platform.

Biology Problem Area	@ 360 TF/s	@1000 TF/s	@ 5000 TF/s
Determining the detailed evolutionary history of each protein family ⇒ *This will enable rational planning for structural biology initiatives and will provide a foundation for assessing protein function and diversity*	3,000 hours to build reference database	300 hours to build reference database	60 hours to build reference database
Determining the frequency and detailed nature of horizontal gene transfers in prokaryotes ⇒ *This will shed light on the molecular and genetic mechanisms of evolution by means other than direct "Darwinian" descent and will contribute to our understanding of the acquisition of virulence and drug resistance in pathogens and the means by which prokaryotes adapt to the environment*	1,000 hours to study 200 gene families	1,000 hours to study 2000 gene families	1,000 hours to study 10,000 gene families
Automated construction of core metabolic models for all the sequenced DOE genomes ⇒ *This will enable dramatic acceleration of the promise of the GTL program and the use of microbial systems to address DOE mission needs in energy, environment, and science*	One hour per organism, 100 hours per metagenome	10 organisms per hour, 10 hours per metagenome	50 organisms per hour, two hours per metagenome
Predict essential genes for all known sequenced micro-organisms ⇒ *This will enable a broader class of genes and gene products to be targeted for potential drugs and to predict culturability conditions for environmental microbes*	300 hours for 1,000 organisms, 10 hours to predict culturability per organism	30 hours for 1,000 organisms, one hour to predict culturability per organism	30 hours for 5,000 organisms
Computational screening all known microbial drug targets against the public and private databases of chemical compounds to identify potential new inhibitors and potential drugs ⇒ *The resulting database would be a major national biological research resource that would have a dramatic impact on worldwide health research and fundamental science of microbiology*	2 M ligands per day per target (1 year to screen all microbial targets)	20 M ligands per day per target (~1 month to screen all microbial targets)	1 machine year to screen all known human drug targets
Model and simulate the precise cellulose degradation and ethanol and butanol biosynthesis pathways at the protein/ligand level to identify opportunities for molecular optimization ⇒ *This would result in a set of model systems to be further developed for optimization of the production of biofuels*	Simulate in detail the directed evolution of individual enzymes	Simulate the co-evolution and optimzation of a degradation or biosynthesis pathway of up five enzymes	Simulate the optimization of a complete cellulose to ethanol or butanol production system of over a dozen enzymatic steps
Model and simulate the replication of DNA to understand the origin of and the repair mechanisms of genetic mutations ⇒ *This would result in dramatic progress in the fundamental understanding of how nature manages mutations and understanding which molecular factors determine the broad range of organism susceptibility to radiation and other mutagens*	30 ns simulation of DNA polymerase	10 ensembles of different DNA repair enzymes	Complete polymerase mediated base pair addition step
Model and simulate the process of DNA transcription and protein translation and assembly ⇒ *This would enable us to move forward on understanding post-transcription and post-translation modification and epi-genetic regulation of protein synthesis*	Validate current understanding of ribosomal function	Explore splisosome function and the evolution of intron/exon functions	Model the complete coupled processes of DNA transcription to protein translation including regulatory processes
Model and simulate the interlinked metabolisms of microbial communities ⇒ *This project is relevant to understanding the biogeochemical cycles of extreme, natural and disturbed environments and will lead to the development of strategies for the production of bio-fuels and the development of new bio-engineered processes based on exploiting communities rather than individual organisms*	20 organisms in a linked metabolic network	100 organisms in a linked metabolic network	200 organisms in a linked metabolic network
In silico prediction of mutations and activity, conformational changes, active site alterations	One enzyme	Five-enzyme pathway	Eight enzyme pathway optimization

National Biomedical Computation Resource (NBCR): Developing End-to-End Cyberinfrastructure for Multiscale Modeling in Biomedical Research

Wilfred W. Li
University of California, San Diego (UCSD), San Diego Supercomputer Center (SDSC)
Nathan Baker
Washington University in Saint Louis
Kim Baldridge
UCSD, SDSC
J. Andrew McCammon
UCSD
Mark H. Ellisman
UCSD, Center for Research In Biological Systems (CRBS)
Amarnath Gupta
UCSD, SDSC
Michael Holst
UCSD
Andrew D. McCulloch
UCSD
Anushka Michailova
UCSD
Phil Papadopoulos
UCSD, SDSC
Art Olson
The Scripps Research Institute (TSRI)
Michel Sanner
TSRI
Peter W. Arzberger
California Institute for Telecommunications and Information Technology (Calit2), CRBS, UCSD

Abstract – Begun in 1994, the mission of the National Biomedical Computation Resource (NBCR) is to conduct, catalyze and enable multiscale biomedical research by harnessing advanced computation and data cyberinfrastructure through multidiscipline and multi-institutional integrative research and development activities. Here we report the more recent research and technology advances in building cyberinfrastructure for multiscale modeling activities.

The development of the cyberinfrastructure is driven by multiscale modeling applications, which focus on scientific research ranging in biological scale from the subatomic, to molecular, cellular, tissue to organ level. Application examples include quantum mechanics modeling with GAMESS, calculation of protein electrostatic potentials with APBS and the Finite Element Toolkit (FEtk); protein-ligand docking studies with AutoDock; cardiac systems biology and physiology modeling with Continuity; and molecular visualizations using PMV and visual workflow programming in Vision. Real use cases are used to demonstrate how these multiscale applications may be made available transparently on the grid to researchers in the biomedicine and translational research arena, through integrative projects ranging from the understanding of the detailed mechanism of HIV protease and integrase action, to neuromuscular junction research in myopathy, to heart arrhythmia and failure, and to emerging public health threats, as well as through collaborative projects with other research teams across the world.

The adoption of service oriented architecture enables the development of highly reusable software components and efficiently leverages the international grid development activities. We describe an end to end prototype environment, exemplified by the adoption of key components of the Telescience project, that allows existing applications to run transparently on the grid, taking advantage of open source software that provides the following:

- a portal interface using GridSphere,
- transparent GSI authentication using GAMA,
- a web service wrapper using Opal,
- a metascheduler using CSF4,
- a virtual filesystem using Gfarm, and
- a grid-enabled cluster environment using Rocks.

Solutions to complex problems may be developed using workflow tools that coordinate different interoperable services. In addition, we also describe the development of ontology and semantic mediation tools such as Pathsys and OntoQuest for data integration and interoperability, which may be efficiently coupled with the application services provided to the biomedical community.

1. Introduction

With the increasing availability of genome sequencing data, it is becoming apparent that knowing the parts is only a prerequisite to understand the big picture. While genomics and proteomics efforts are producing data at an increasing rate, the data are more of a descriptive nature rather than functional integration and interactions of the parts.[1] Traditionally, most

[1] McCulloch, A.D., Paternostro G. Cardiac systems biology. Ann N Y Acad Sci 1047: 283-295. 2005.

research and modeling activities have focused on a particular system level such as proteins, cells, tissues, organs, organ systems, up to the level of populations. Multiscale modeling, across the length scale from nanometers for molecules to meters for human bodies, as well as across the time scale from nano-seconds for molecular interactions to the length of human life, is crucial to the development of simulation systems for better understanding of human physiology and predictive capabilities for disease prevention and treatment.[2]

Multiscale modeling studies derive mathematical models of structure-function relations at one scale and link to the level below through appropriate parameters. These models need to be based on widely adopted modeling standards, with necessary software tools for developing, visualizing and linking the models.[3] Multiscale modeling requires constant cross validation and feedback from experiments and models. Often the experiments provide the data for development and validation of the models, and the models can in turn provide predictions of behavior or require additional experiments which may lead to new discoveries.[1] Models may impose *a priori* physical constraints and represent complex processes, and provide quantitative predictions that may be verified experimentally.[4] Systems modeling are data-limited when mechanistic models are to be built, because experiments may be slow and difficult to validate. On the other hand, models across scales are compute-limited due to the "tyranny of scale." Molecular dynamics simulations are often limited to a scale 5 to 6 orders of magnitude smaller than the time necessary for a real event to complete. In terms of an extreme case of computational challenge, the panel on Simulation Based Engineering Science (SBES)[5] noted that in the turbulence-flow modeling, the "tyranny of scale" prevents a solution for many generations, even with Moore's law holding true.

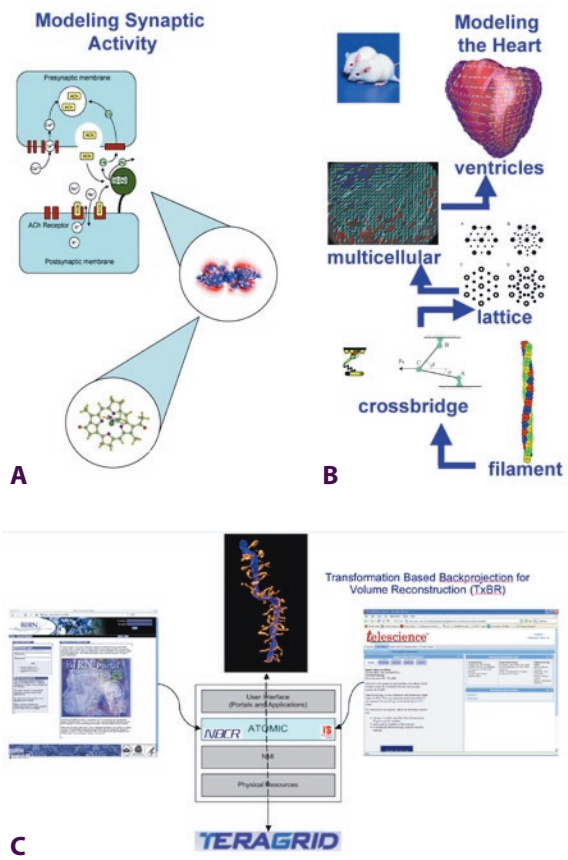

Figure 1. Multiscale modeling of A) the neuromuscular junction, B) physiology of the heart, C) electron microscopy of cellular structures are the main scientific drivers for NBCR cyberinfrastructure development efforts in support of translational research

In recognition of the importance and the severe obstacles to experiments that are cross-scale in space, time and state, there have been a number of workshops held and panel recommendations made. The multiscale modeling consortium or Interagency Modeling and Analysis Group (IMAG),[6] along with the participation of a number of federal agencies that the National Institutes of Health (NIH), National Science Foundation (NSF), National Aviation and Space Agency (NASA), Depart of Energy (DOE), Department of Defense (DoD) and the United States Department of Agriculture (USDA), aims to promote the development and exchange of tools, models, data and standards for the MultiScale Modeling (MSM) community. The NSF

[2] Hunter P.J., Borg, T.K. Integration from proteins to organs: the Physiome Project. Nat Rev Mol Cell Biol 4: 237-243. 2003.

[3] Hunter P., Nielsen P. A strategy for integrative computational physiology. Physiology (Bethesda) 20: 316-325. 2005.

[1] McCulloch, A.D., Paternostro G. Cardiac systems biology. Ann N Y Acad Sci 1047: 283-295. 2005.

[4] Hunter P.J., Li W.W., McCulloch A.D., Noble D. Multi-scale Modeling Standards, Tools Databases for the Physiome Project. Computer: In Press. 2006.

[5] Oden J.T., Belytschko T., Fish J., Hughes T.J., Johnson C., et al. Revolutionizing Engineering Science through Simulation. 2006.

[6] IMAG: Interagency Opportunities in Multi-Scale Modeling in Biomedical, Biological, and Behavioral Systems - http://www.nsf.gov/pubs/2004/nsf04607/nsf04607.htm. 2006.

blue ribbon panel[5] recognizes that SBES applied to the multiscale study of biological systems and clinical medicine, or simulation based medicine, may bring us closer to the realization of P4 medicine (predictive, preventative, personalized, and participatory). The June 2005 PITAC report on computational science[7] and the 2005 National Research Council report,[8] both specifically recommended increased and sustained support for infrastructure development to meet the computational challenges ahead.

The National Biomedical Computation Resource (NBCR),[9] a national center supported by the National Center for Research Resources (NCRR),[10] has a mission to "conduct, catalyze and enable multiscale biomedical research" by harnessing advanced computation and data cyberinfrastructure through multidiscipline and multi-institutional integrative research and development activities (Figure 1). NBCR has co-hosted several workshops and conferences that engage researchers and students from the MSM community.[11] [12] [13] It has been our experience that tools developed with one application in mind tend to be narrow and less flexible, let alone interoperable, with other software. Therefore, we have been focusing on enabling a set of pathfinder examples across scales, through team work with world class scientists who are members of NBCR, and through collaborative projects with members of the broader biomedical and MSM communities.

2. Scientific Drivers and Tools

Understanding the workings of cells, tissues, organs, or entire organisms requires researchers to pull together information from multiple physical scales and across multiple temporal scales. We highlight activities within our collective and collaborative experiences at different or cross scales. These scientific examples are from long running complex multiscale research areas, which drive the development of our technology integration and infrastructure development efforts.

2.1. Continuity 6

The need for integrative analysis in cardiac physiology and pathophysiology is readily appreciated. Common heart diseases are multi-factorial, multi-genic, and linked to other systemic disorders such as diabetes, hypertension, or thyroid disease. The coupling between the L-type Calcium channels (LCCs, also known as dihydropyridine receptors, or DHPRs) and ryanodine receptors (RyRs) are important in the excitation-contraction (E-C) of cardiac myocytes. The influx of calcium releases the calcium store in the Sarcoplasmic reticulum (SR), a phenomenon known as the calcium induced calcium release (CICR). In fact, the latest ionic models of cardiac myocytes include more than 20 ionic fluxes and 40 ordinary differential equations.[1] Computational methods and ionic models for cardiac electromechanics at different scales have also been developed and are available in the software package Continuity. Figure 2 shows an example of how Continuity is used to help develop dual pacemaker systems that are helping to save people's lives today.[14] Continuity is used by a number of researchers in the field of cardiac biomechanics, and receives regular acknowledgment in peer-reviewed publications.[15] [16] Continuity 6 is continuously being improved to support larger scale simulations, for example using the MYMPI package,[17] a standards-driven MPI libraries for Python developed by NBCR to improve the parallel computation efficiency.

In ventricular myocytes, the dyadic cleft is a periplasmic space that spans about 10 nm between the voltage-gated LCCs/DHPRs on the transverse tubule (TT) membrane, and RyRs on the SR. Within the dyadic cleft, the small reaction volume and the exceedingly low number of reactant molecules means that the reaction system is better described by stochastic behavior,

[5] Oden J.T., Belytschko T., Fish J., Hughes T.J., Johnson C., et al. Revolutionizing Engineering Science through Simulation. 2006.

[7] PITAC: Computational Science: Ensuring America's Competitiveness. http://www.nitrd.gov/pitac/reports/index.html. 2005.

[8] NRCNA: Catalyzing Inquiry at the Interface of Computing and Biology. http://www.doegenomestolife.org/pubs/NRCComputingandBiology/index.shtml. 2005.

[9] NBCR: National Biomedical Computation Resource - http://nbcr.net/. 2005.

[10] NCRR: National Center for Research Resources - http://www.ncrr.nih.gov. 2006.

[11] MCMBR: Multiscale Computational Modeling for Biomedical Research - http://nbcr.net/physiome. 2004.

[12] McCulloch, A.D., Arzberger, P.W., Hunter, P. Computational Physiology: From Genome to Physiome. The Physiologist 49: 94-95. 2005.

[13] DCBC: Data and Collaboratories in the Biomedical Community. - http://nbcr.net/Collaboratories/index.html. 2002.

[1] McCulloch, A.D., Paternostro G. Cardiac systems biology. Ann N Y Acad Sci 1047: 283-295. 2005.

[14] Usyk, T.P., McCulloch, A.D. Relationship between regional shortening and asynchronous electrical activation in a three-dimensional model of ventricular electromechanics. J Cardiovasc Electrophysiol 14: S196-202. 2003.

[15] Trayanova, N. In pursuit of the elusive culprit. Heart Rhythm 2: 729-730. 2005.

[16] Xie, F., Qu, Z., Yang, J., Baher, A., Weiss, J.N., et al. A simulation study of the effects of cardiac anatomy in ventricular fibrillation. J Clin Invest 113: 686-693. 2004.

[17] Kaiser, T.H., Brieger, L., Healy, S.N. MYMPI - MPI Programming in Python. International Conference on Parallel and Distributed Processing Techniques and Applications; Las Vegas. In Press. 2006.

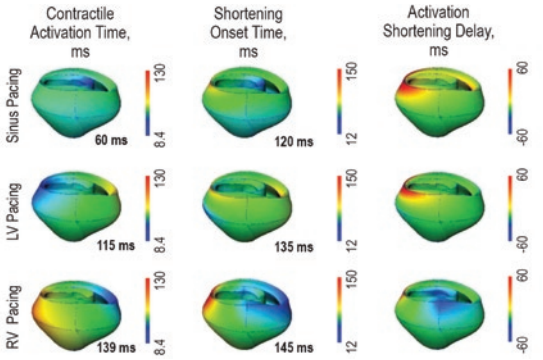

Figure 2. Effects of left and right ventricular pacing compared with normal sinus rhythm on the temporal sequences of electrical activation (left column) and mechanical shortening (middle column) in a three-dimensional model of the canine heart. Activation-shortening delays (right column) are heterogeneous, even during normal sinus rhythm. Simulations rendered using Continuity developed by NBCR.

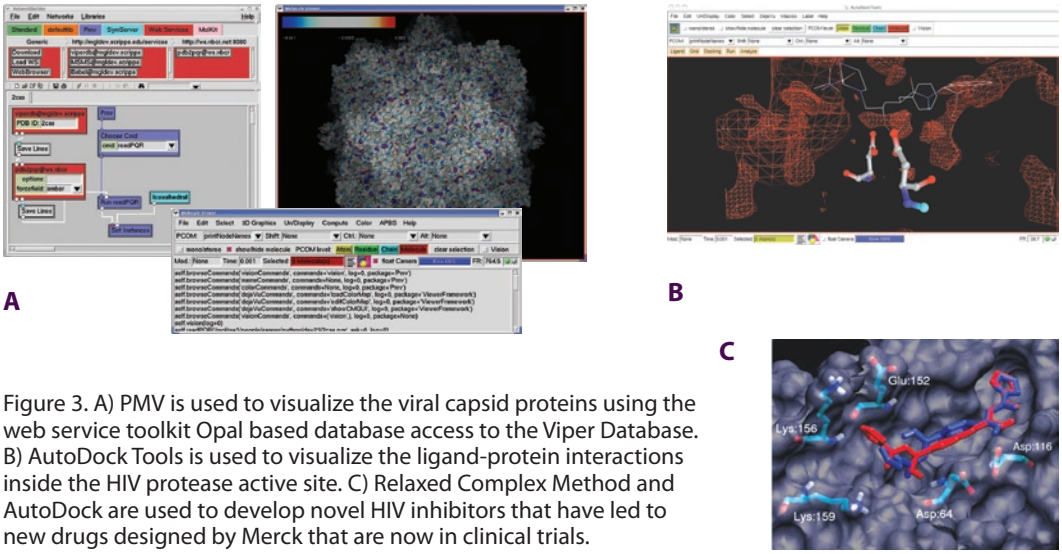

Figure 3. A) PMV is used to visualize the viral capsid proteins using the web service toolkit Opal based database access to the Viper Database. B) AutoDock Tools is used to visualize the ligand-protein interactions inside the HIV protease active site. C) Relaxed Complex Method and AutoDock are used to develop novel HIV inhibitors that have led to new drugs designed by Merck that are now in clinical trials.

rather than continuous, deterministic reaction-diffusion "partial differential equations."[18] [19] This system is one of the focal points of NBCR research at the molecular scale, and for cross scale integrations, with the development of highly realistic 3D models based on electron tomography from the National Center for Microcopy and Imaging Research (NCMIR).

2.2. PMV, Vision, AutoDock Tools & AutoDock

The Python Molecular Viewer (PMV) is a component based software package (Figure 3) written in Python and contains an accompanying visual programming tool called Vision.[20] PMV is among the first molecular visualization software packages that takes advantage of grid services, using the web service toolkit Opal developed by NBCR, to access remote databases and computational resources (Figure 3A). Many key packages from the PMV/Vision framework have been reused in Continuity 6, the multiscale modeling platform for cardiac electrophysiology and biomechanics.

In addition, AutoDock Tools (ADT)[21] has been developed as a module inside PMV for the popular molecular docking software package AutoDock. AutoDock is a world famous software package and has been used in developing inhibitors for many important diseases.[22] [23] The FightAids@Home project has been using AutoDock to screen for HIV inhibitors and is now running on the World Community Grid, an IBM philanthropic activity.[24] ADT greatly simplifies the preparation and post analysis procedures of AutoDock (Figure 3B).

[18] Koh. X., Srinivasan, B., Ching, H.S., Levchenko, A. A 3D Monte Carlo Analysis of the Role of Dyadic Space Geometry in Spark Generation. Biophys J 90: 1999-2014. 2006.

[19] Soeller, C., Cannell, M.B. Analysing cardiac excitation-contraction coupling with mathematical models of local control. Prog Biophys Mol Biol 85: 141-162. 2004.

[20] Sanner, M.F., Stolz, M., Burkhard, P., Kong, X.-P., Min, G, et al., (Eds.). Visualizing Nature at Work from the Nano to the Macro Scale.John Wiley & sons, Ltd. p. 7-11. 2005.

[21] Osterberg, F., Morris, G.M., Sanner, M.F., Olson, A.J., Goodsell, D.S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 46: 34-40. 2002.

[22] Goodsell, D.S., Morris, G.M., Olson, A.J. Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 9: 1-5. 1996.

[23] Morris, G.M., Goodsell, D.S., Huey, R., Olson, A.J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 10: 293-304. 1996.

[24] WCG: World Community Grid: FightAIDS@Home. http://www.worldcommunitygrid.org/projects_showcase/viewFaahFaq.do?shortName=faah. 2005.

2.3. Relaxed Complex Method and its Application in HIV Integrase Inhibitor Studies

The use of computer models to screen for small molecules that will bind to and alter the activity of macromolecules has long been complicated by "induced fit" effects: the macromolecule may undergo shape changes during the formation of the complex with the small molecule. A new approach has been developed in which a variety of conformations of the target macromolecule are generated by molecular dynamics simulation or other methods, to generate structures that might include one or more that are representative of that of the "relaxed complex" with the small molecule.[25] To select the most relevant macromolecular conformations, rapid screening is done using AutoDock and ADT. The most stable complex structures from this screening are then subject to more refined analysis to yield the most probable structure of the complex and an associated estimate of the strength of binding (Figure 3C). More recently, applications of this method to an antiviral target for treatment of HIV infections[26] has aided scientists at Merck & Co. in developing a new class of drugs that are in Phase III clinical trials.[27]

[25] Lin, J.H., Perryman, A.L., Schames, J.R., McCammon, J.A. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. Biopolymers 68: 47-62. 2003.

[26] Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., et al. Discovery of a novel binding trench in HIV integrase. J Med Chem 47: 1879-1881. 2004.

[27] NIH: Clinical Trials Database - http://clinicaltrials.gov/ct/show/NCT00293254?order=1. 2006.

2.4. Finite Element Tool Kit (FEtk) & Adaptive Poisson-Boltzmann Solver (APBS)

The FEtk[28] is an evolving collection of parallel adaptive, multilevel finite element software libraries and supports tools for solving coupled systems of partial differential equations (PDE) and integral equations (IE).[29] The numerical libraries are written in an object-oriented form of ANSI-C (Figure 4A). Left (Top/Bottom): Potential contours of the electrostatic potential around a biomolecule, computed adaptively using FEtk, and a closeup of the adapted part of the simplex mesh. Right (Top/Bottom): Isosurfaces projected onto a cutting plane through two black holes in an astrophysics problem. FEtk was used to compute the initial bending of space and time around two massive black holes, which involved adaptively solving a coupled nonlinear elliptic PDE.

[28] FEtk: Finite Element Toolkit - http://www.fetk.org. 2005.

[29] Holst, M.J. Adaptive Numerical Treatment of Elliptic Systems on Manifolds. Advances in Computational Mathematics 15: 139-191. 2001.

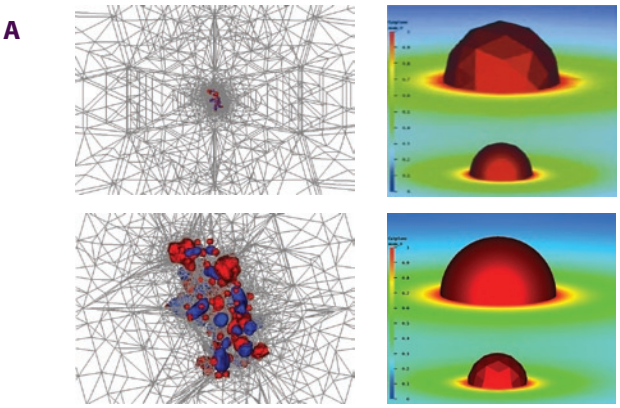

Figure 4. A) Application examples of FEtk for solving the electrostatic potentials of biomolecules or rendering of isosurfaces of black holes in astrophysics problems. B) The electrostatic binding energy of PKA and balanol is visualized using PMV after remote distributed APBS calculations using NBCR strongly typed web service.

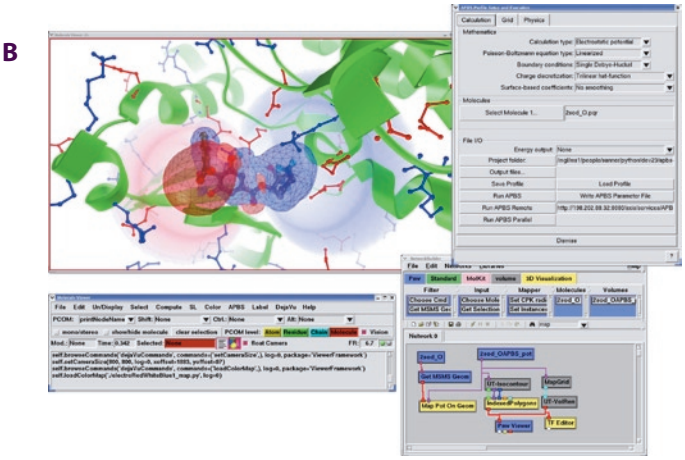

Figure 4. B) The electrostatic binding energy of PKA and balanol is visualized using PMV after remote distributed APBS calculations using NBCR strongly typed web service.

The primary FEtk ANSI-C software libraries include MALOC (Minimal Abstraction Layer for Object-oriented C) for portability, SG (Socket Graphics) for networked OpenGL-based visualization, and MC (Manifold Code) for adaptively solving coupled systems of nonlinear PDE on two- and three-manifold domains. A 2D MATLAB-based prototyping tool called MCLite is also available for fast development of MC-based software. A related package, PMG (Parallel

Algebraic Multigrid) is designed to numerically solve scalar nonlinear PDE problems such as the Poisson-Boltzmann Equation (PBE) on a class of regular domains using algorithms which have optimal or near-optimal space and time complexity.

APBS is a software package for the numerical solution of the PBE, one of the most popular continuum models for describing electrostatic interactions between molecular solutes in salty, aqueous media.[30] Continuum electrostatics plays an important role in several areas of biomolecular simulation, including:

- simulation of diffusional processes to determine ligand-protein and protein-protein binding kinetics;
- implicit solvent molecular dynamics of biomolecules;
- solvation and binding energy calculations to determine ligand-protein and protein-protein equilibrium binding constants and aid in rational drug design;
- and biomolecular titration studies.

APBS was designed to efficiently evaluate electrostatic properties for such simulations for a wide range of length scales to enable the investigation of molecules with tens to millions of atoms. APBS uses FEtk and PMG to solve the Poisson-Boltzmann equation numerically.

APBS is available to users as standalone applications or inside PMV for seamless access (Figure 4B), available in CHARMM or AMBER through the NBCR iAPBS interface,[31] or as web services using the NBCR Opal toolkit.[32] In this example, the electrostatic binding energy for the protein kinase A (PKA) complex with the inhibitor balanol was computed using APBS and displayed in PMV. The PKA complex is shown as a ribbon diagram with side chains or positively (blue) and negatively (red) charged amino acids displayed as Sticks and Balls. The inhibitor molecule is displayed using a fat Sticks and Balls representation and colored by atom types. The electrostatic binding energy is visualized by direct volume rendering with the shown transfer function and using two isosurfaces.

A new software component from NBCR that allows the efficient modeling of diffusion events across the neuromuscular junction in a steady state or time dependent manner called the Smoluchowski equation solver (SMOL) is also based on the FEtk and has been used in conjunction with APBS in a study of the tetrameric complex acetylcholine receptors.[33]

3. Computational and data cyberinfrastructure to support multiscale modeling

In parallel to efforts focusing attention on the needs and the benefits of multiscale modeling, tremendous national and international investments have already been made to develop and deploy a cyberinfrastructure that will revolutionize the conduct of science.[34] This cyberinfrastructure consists of distributed computational, data storage, observational, and visualization resources, including human resources, connected by a network infrastructure and a software layer (middleware), that will "bring access of resources (at one end) to researchers (at another) and allow researchers to conduct team science as part of normal conduct of science,[35] in an end-to-end cyberinfrastructure." Cyberinfrastructure, and grid, are often used interchangeably.[36]

[30] Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98: 10037-10041. 2001.

[31] iAPBS: iAPBS Interface - http://nbcr.net/forum/viewforum.php?f=17. 2006.

[32] Krishnan, S., Stearn, B., Bhatia, K., Baldridge, K., Li, W.W., et al. Opal: Simple Web Service Wrappers for Scientific Applications. International Conference for Web Services. In Press. 2006.

[33] Zhang, D., Suen, J., Zhang, Y., Song, Y., Radic, Z., et al. Tetrameric mouse acetylcholinesterase: continuum diffusion rate calculations by solving the steady-state Smoluchowski equation using finite element methods. Biophys J 88: 1659-1665. 2005.

[34] Atkins, D., Kroegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., et al. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrasturcuture. Arlington, VA: National Science Foundation. 2003.

[35] Hey, T., Trefethen, A.E. Cyberinfrastructure for e-Science. Science 308: 817-821. 2005.

[36] Graham, S.L., Snir, M., Patterson, C.A. Getting Up to Speed: The Future of Supercomputing: The National Academies Press. 2004.

National Biomedical Computation Resource (NBCR):
Developing End-to-End Cyberinfrastructure for Multiscale Modeling in Biomedical Research

The necessary grid infrastructure to support the multiscale modeling community remains to be defined through an iterative process with ongoing interactions between scientists and infrastructure developers. There has been significant progress in the development of the networks and physical resources that are the fabric of the grid.[37] However, the middleware layer, which connects the fabric with the users and applications, is still in a state of flux. To increase the usability and decrease the cost of entry of the grid, new programming models or application execution environments are being developed, and sometimes these are referred to as grid application-level tools.[38] These tools are designed to be built on top of the grid software infrastructure. They are generic, easy to use and shield the users from changes in underlying architecture.

An "end-to-end" cyberinfrastructure for multiscale modeling needs to have the capability to handle the representative data commonly encountered in MSM, by representative users who need to accomplish tasks that are representative of the nature of MSM research and development. The scientific drivers for multiscale modeling at NBCR are diverse and span the scale from subatomic electron charge density flows in development of better photodynamic therapies,[39] to cellular and whole heart physiology modeling,[1] with information integration technology providing the mediator necessary for interoperability. All these projects drive the development and integration effort we are leading to provide the cyberinfrastructure necessary and required for the research objectives to be met. In addition, NBCR works closely with other projects at UCSD, such as NCMIR, BIRN,[40] CAMERA,[41] Optiputer,[42] JCSG,[43] and establishes important collaborations with national projects such as TeraGrid,[44] Open Science Grid,[45] as well as international projects such as PRAGMA[46] and OMII.[47]

The development of infrastructure that can support diverse applications using distributed physical resources and remain easy to use and scalable has ushered in the service-oriented architecture as the dominant modus operandi for reasons eloquently stated in "Service-oriented Science."[48] To enable scientific applications on the grid, many different approaches have been adopted.[49] We have taken a minimalist approach, which is to select the most stable components, achieve the greatest leverage, and develop smart glues that are reusable components, within the service oriented approach (Figure 5). We'll highlight some key components developed by or with critical contributions from NBCR and then discuss how they are used to support multiscale modeling efforts at NBCR and how they are available to the MSM community in general.

Upper middleware services are those that make the development of distributed applications significantly easier, with support for higher levels of abstraction and standardization. For example, NBCR Opal based application web services provide job management and scheduling features based on the Globus toolkit. An application developer may begin using the grid quickly with the basic knowledge of web service development, as shown by the use cases in PMV, My WorkSphere, and Gemstone user environments. Lower middleware services are those that have stabilized over the years and serve as the foundation for the development of more sophisticated and transparent modes of access. However, as often dictated by performance requirements, a user application may access lower layers directly. This is much less desirable unless the integration is based on the service oriented architecture.

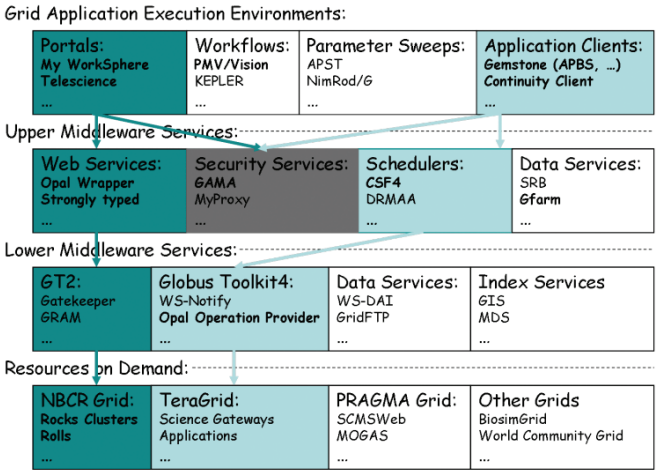

Figure 5. Selected Components of NBCR Software Service Stack. Colored blocks and arrows indicate possible routes for distributed job execution. The different green shades show possible routes to access physical computational resources, and the arrows indicate direction of workflow. The grey shaded region indicates shared security services among the different workflows.

[37] Foster, I., Kesselman, C. (Eds.). The Grid 2: Blueprint for a New Computing Infrastructure. 2 ed. San Francisco: Morgan Kaufmann Publishers, Inc. 2004.

[38] Bal, H., Casanova, H., Dongarra, J., Matsuoka, S. Application-Level Tools. In: Foster, I., Kesselman, C. (Eds.). The Grid 2. 2 ed. Amsterdam: Elsevier. 2004.

[39] Yerushalmi, R., Scherz, A., Baldridge, K.K. Direct experimental evaluation of charge scheme performance by a molecular charge-meter. J Am Chem Soc 126: 5897-5905. 2004.

[40] BIRN: Biomedical Informatics Research Network - http://www.nbirn.net. 2005.

[41] CAMERA: Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis - http://camera.calit2.net/. 2006.

[42] OptIPuter: A Powerful Distributed Cyberinfrastructure to Support Data-Intensive Scientific Research and Collaboration - http://www.optiputer.net. 2005.

[43] JCSG: Joint Center for Structural Genomics - http://www.jcsg.org. 2005.

[44] TeraGrid: http://www.teragrid.org. 2004.

[45] OSG: Open Science Grid - http://www.opensciencegrid.org/. 2006.

[46] PRAGMA: Pacific Rim Applications and Grid Middleware Assembly - http://www.pragma-grid.net/. 2004.

[47] OMII: Open Middleware Infrastructure Institute - http://www.omii.ac.uk/. 2006.

[48] Foster, I. Service-oriented science. Science 308: 814-817. 2005.

[49] Abramson, D., Lynch, A., Takemaya, H., Tanimura, Y., Date, S., et al. Deploying Scientific Applications to the PRAGMA Grid testbed: Strategies and Lessons. Sixth IEEE International Symposium on Cluster Computing and the Grid, 2006 CCGrid 06; Singapore. 1: 241-248. 2006.

Below we describe some key tools and technology actively developed or co-developed by NBCR to support our user communities. For a more comprehensive description of all the tools and software, please see the NBCR download site.[50]

[50] NBCR: NBCR Software and Tools Description and Download Site - http://nbcr.net/tools.php. 2006.

3.1. Opal – a simple yet powerful web service wrapper

Opal is developed by NBCR as a Java based toolkit that automatically wraps any legacy applications with a Web services layer that is fully integrated with Grid Security Infrastructure (GSI) based security, cluster support, and data management.[32] [51] The advantage of using OPAL is that the application may be launched using any Web service client, because the WSDL defines using the standard protocol for how the service may be accessed, and it provides the basic HTTP access to results, as well as any metadata that describes how the results may be handled (Figure 6A). This 'deploy once use by many' feature of a Web service is a key ingredient for achieving interoperability. Because OPAL manages the interaction with the grid architecture, the grid is transparent to the user. In addition, workflow tools like Kepler[52] may easily compose web services based workflows through a common interface (Figure 6B).

[32] Krishnan, S., Stearn, B., Bhatia, K., Baldridge, K., Li, W.W., et al. Opal: Simple Web Service Wrappers for Scientific Applications. International Conference for Web Services. In Press. 2006.

[51] Li, W.W., Krishnan, S., Mueller, K., Ichikawa, K., Date, S., et al. Building cyberinfrastructure for bioinformatics using service oriented architecture. Sixth IEEE International Symposium on Cluster Computing and the Grid Workshops Singapore. 2: 39-46. 2006.

[52] KEPLER: The Kepler project - http://kepler-project.org/. 2006.

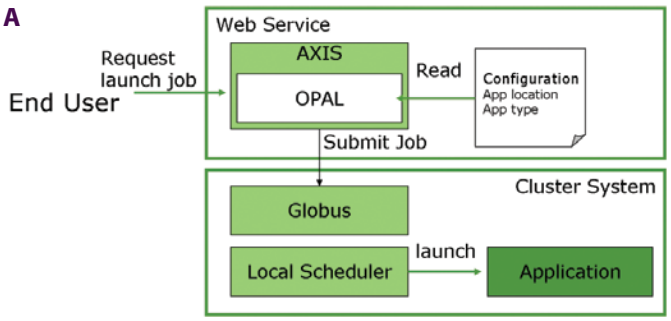

Figure 6. A) Opal allows rapid deployment of applications as web services using user provided configuration options.

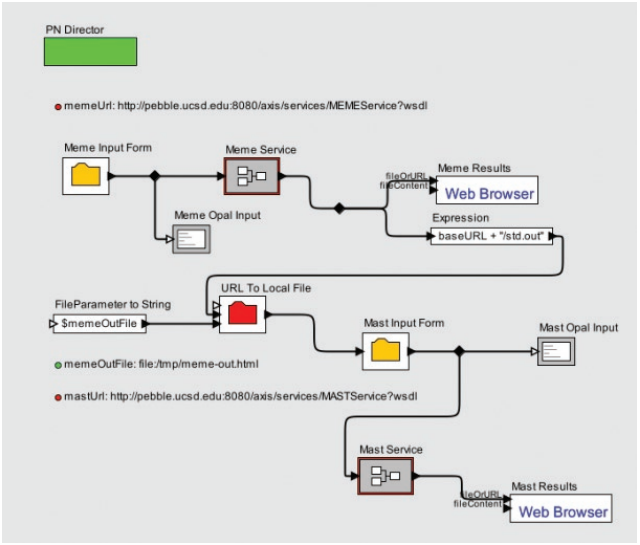

Figure 6. B) Workflow programs such as Kepler may orchestrate Opal based web services with reusable actors.

3.2. Grid Account Management Architecture (GAMA)

GAMA, co-developed by NBCR,[53] is a GSI-based security service that manages X.509 user credentials on behalf of users and supports SOAP-based applications (Figure 7A). The server component leverages existing software packages such as CAS, MyProxy, and CACL with the GAMA version 2 supporting other CA packages such as NAREGI. A portlet component provides the administrative interface to the server. As security is a sensitive and critical issue in the production use of the cyberinfrastructure, GAMA allows any (Figure 7) organization to create their own certificate authority, manage their user certificates, secure the SOAP communications using HTTPS and mutual authentication, and integrate seamlessly with portals and rich clients. For example, GAMA is used by GridSphere based portals,[54] Gemstone[55] and PMV/Vision.

[53] Bhatia, K., Chandra, S., Mueller, K. GAMA: Grid Account Management Architecture. 1st IEEE International Conference on e- Science and Grid Computing; Melbourne, Australiapp. In Press. 2006.

[54] Gridsphere: GridSphere - http://www.gridsphere.org. 2004.

[55] Baldridge, K., Bhatia, K., Greenberg, J.P., Stearn, B., Mock, S., et al. GEMSTONE: Grid Enabled Molecular Science Through Online Networked Environments. Life Sciences Grid Workshop; Singapore. World Scientific Press. 2005.

3.3. Application TO Middleware Interface Component (ATOMIC)

The ATOMIC bundle (Figure 7B), developed for the Telescience Project by NBCR, provides a fabric for seamless interoperability among user interfaces (web portals and applications) and externally addressable grid resources (instruments and computers) inside a grid portal environment.[56] There is a unified programming API for basic scientific research, which liberates the users and programmers from having to learn about the grid. TeleAuth is a specialized version of GAMA. TeleWrap provides seamless access to remote data transparently and has been used successfully in the Telescience portal and EMWorkspace. TeleRun provides a higher level of abstraction than Opal-provided services and may further shield developers from the details of grid execution environment. ATOMIC also provides persistent session information during a particular Telescience session, so that all user portlets have access to the same session information regardless of the specific portlet being used.

[56] Lin, A.W., Dai, L., Ung, K., Peltier, S., Ellisman, M.H. The Telescience Project: Applications to Middleware Interaction Components. The 18th IEEE International Symposium on Computer-Based Medical Systems; Dublin, Ireland:. 543-548. 2005.

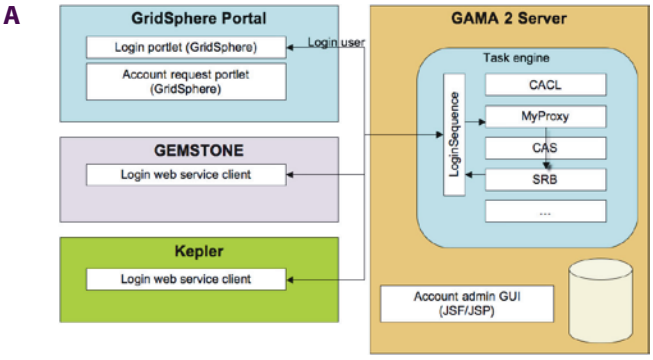

Figure 7. A) GAMA provides web service based access for GSI and proxy management.

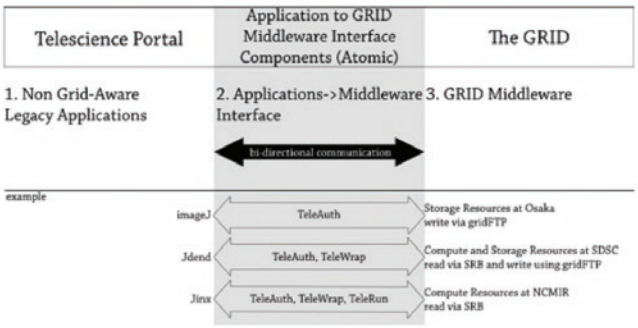

Figure 7. B) ATOMIC leverages GAMA for TeleAuth and adds transparent data access and job execution for the Telescience portal environment.

3.4. MYMPI

The development of MYMPI[57] is an effort of NBCR, driven by the requirement of Continuity, to improve interprocess communications between Python and FORTRAN.[17] It is a Python module used with a normal Python interpreter. MyMPI is an implementation following the standard specifications of MPI, supporting more than 30 of the commonly used MPI calls. The syntax of the calls matches the syntax of C and FORTRAN calls closely to support mixed Python with the other languages. MYMPI will also allow access from Python to other application packages such as the FEtk.[58]

3.5. Gemstone, My Worksphere as Portals to Grid Services

Gemstone[59] is developed with major support from the NSF National Middleware Initiative, in close collaboration with NBCR, especially in the interaction with strongly typed and Opal based web services (Figure 8 A).For example, Gemstone provides an interface to GAMESS[60] and APBS strongly typed web services,[30] both developed by NBCR. It also supports the GARNET visualization web service from the NMI (NSF Middleware Initiative) project. Opal based web service for PDB2PQR,[61] a utility package for APBS calculations, is also accessible from Gemstone as web service based workflows. Gemstone utilizes the open source Mozilla engine and uses XML User Interface Markup Language (XUL) to describe the user interface. My WorkSphere

[57] MYMPI: My MPI - http://nbcr.net/forum/viewforum.php?f=16. 2006.

[17] Kaiser, T.H., Brieger, L., Healy, S.N. MYMPI - MPI Programming in Python. International Conference on Parallel and Distributed Processing Techniques and Applications; Las Vegas. In Press. 2006.

[58] FEtk: Finite Element Toolkit - http://www.fetk.org. 2006.

[59] Baldridge, K.K., Bhatia, K., Greenberg, J.P., Stearn, B., Mock, S., et al. Grid-Enabled Molecular Science through Online Networked Environments. Life Sciences Grid Workshop; Singapore. World Scientific Press. 2005.

[30] Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98: 10037-10041. 2001.

[60] Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., et al. General atomic and molecular electronic structure system. Journal of Computational Chemistry 14: 1347-1363. 1993.

[61] Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 32: W665-667. 2004.

is a prototype GridSphere based portal environment, which leverages portlets that are JSR168 standard compliant. As an example, we have deployed MEME[62] as a portlet using Opal, as part of a generalized Cyberinfrastructure for bioinformatics applications,[51] with Gfarm[63] and CSF4[64] as the other key middleware components.

3.6. Rocks and Rock'n Roll

In an effort to make cyberinfrastructure more readily available to scientists and engineers, it is necessary to not only develop different middleware to support legacy applications, but also to make the different software packages easy to deploy into existing infrastructure. Rocks cluster environment toolkit[65] (Figure 8B) has proven to be invaluable for NBCR to build the basic infrastructure, deploy our software stack, and make our infrastructure replicable by others. NBCR has contributed critically to the development of the Condor roll[66] (a roll is a mechanism, similar to the Red Hat RPM, though fully automated, for building reproducible cluster and grid environments). Other rolls available from NBCR include the APBS, MEME, GAMA, AutoDock, and PMV. Additional rolls for SMOL[67] and FEtk will be available soon (Figure 8 B).

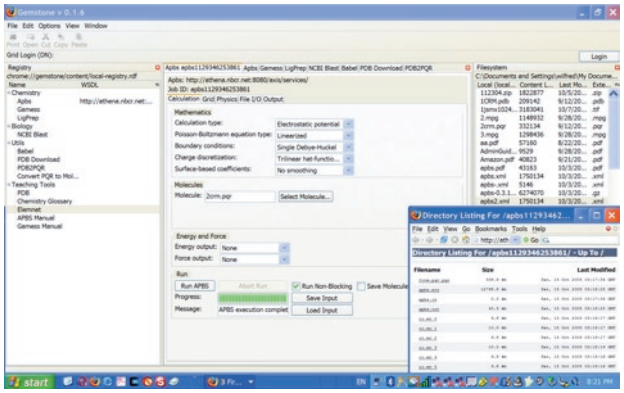

Figure 8. A) Gemstone offers a light weight mode of access to distributed resources using Opal based web services or strongly typed web services.

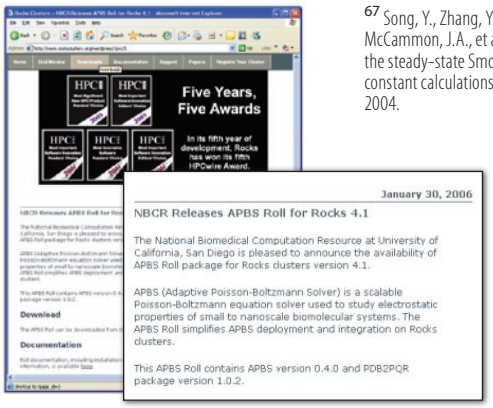

Figure 8. B) Rocks cluster environment software has provided a replicable infrastructure for easy deployment and production use.

3.7. Networks, pathways, ontologies and graph query engines

PathSys is a general-purpose, scalable, graph data warehouse of biological information, complete with a graph manipulation and a query language, a storage mechanism, and a generic data-importing mechanism through schema-mapping. The client software, named BiologicalNetworks, supports the navigation and analyses of molecular networks (Figure 9B). As systems biology integrates information from many data sources, distributed or integrated, there is an increasing demand for managing, querying and automated reasoning using ontology concepts and description logic. OntoQuest is developed to provide the extended mapping schemes for storing OWL (web ontology language) ontologies into backend databases (Figure 9A). It is aimed to guide a user to explore ontological datasets and eventually make non-preconceived, impromptu discoveries.

Both PathSys and OntoQuest take advantages a new semantic-aware RDF (resource description framework) algebra, which supports the inference of complex relationships rep-

[62] Bailey, T.L., Williams, N., Misleh, C., Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369-373. 2006.

[51] Li, W.W., Krishnan, S., Mueller, K., Ichikawa, K., Date, S., et al. Building cyberinfrastructure for bioinformatics using service oriented architecture. Sixth IEEE International Symposium on Cluster Computing and the Grid Workshops Singapore. 2: 39-46. 2006.

[63] Li, W.W., Arzberger, P.W., Yeo, C.L., Ang, L., Tatebe, O., et al. Proteome Analysis using iGAP in Gfarm. The Second International Life Science Grid Workshop 2005; Grid Asia 2005, Singapore. World Scientific Press. 2005.

[64] Wei, X., Li, W.W., Tatebe, O., Xu, G., Hu, L., et al. Integrating Local Job Scheduler - LSF with Gfarm. Lecture Notes in Computer Science 3758: 197. 2005.

[65] ROCKS: Rocks Cluster Distribution - http://www.rocksclusters.org. 2005.

[66] Litzkow, M., Livny, M., Mutka, M. Condor - a hunter of idle workstations. Proceedings of the 8th International Conference of Distributed Computing Systems; June 1988pp. 104-111. 1988.

[67] Song, Y., Zhang, Y., Shen, T., Bajaj, C.L., McCammon, J.A., et al. Finite element solution of the steady-state Smoluchowski equation for rate constant calculations. Biophys J 86: 2017-2029. 2004.

resented in ontological hierarchies. OntoQuest is being developed using the Cell Centered Database (CCDB) and neuroscience ontology, whereas PathSys is driven by research in yeast signaling pathways. The technology is extensible to the data integration between distributed data and computation services described with appropriate semantics and ontology terms (Figure 9). The technology being developed under OntoQuest may also be used to provide additional semantic annotations to Opal based web services to improve automated service discovery and utilization.

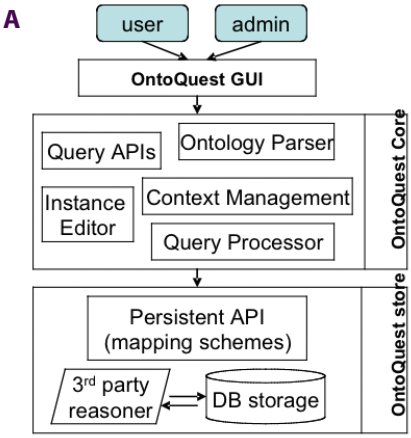

Figure 9. A) OntoQuest architecture: ontology based query engine and database environment.

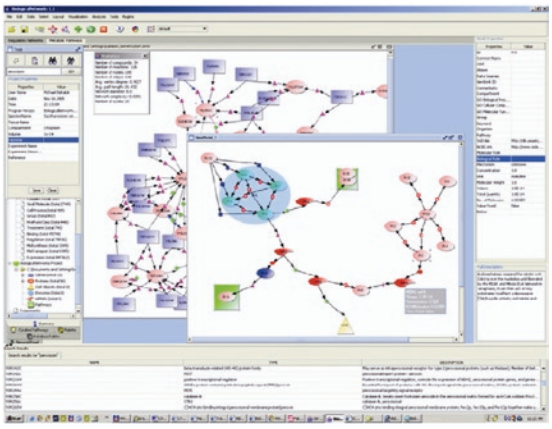

Figure 9. B) BiologicalNetworks is a client interface to the PathSys database warehouse, with support for many xml data formats.

4. Summary & Discussion

There are a plethora of tools and innovative approaches to the development of cyberinfrastructure in order to support multiscale modeling activities. However, better and more robust approaches will always come out of close collaborations between computer scientists and biomedical researchers, as well as other field specialists. The interactions will educate all the groups to be fully aware of the requirements and challenges of the state of the art technology, and make routine use of the grid possible today. In addition, the development of new tools that support applications in different fields and through international collaborations greatly reduces the collective cost for global computational grids. The service oriented approach is gaining momentum and greatly facilitates the development of a knowledge based global economy.

4.1. Collaborations needed to build robust grid computing platform

There are often conflicting requirements between biomedical research and grid communities, with the former accustomed to a "one experiment at a time" approach, and the latter desiring systems that handle large number of jobs simultaneously. The reality is that current grid computing systems are still difficult to learn, only relatively stable, and limited by the technologies available in hardware, software, and programming languages, as noted in "A strategy for integrative computational physiology."[3] It is a challenge to both communities to design better software and use them effectively. With these caveats in mind, one may attempt to select from the available tools and build a robust platform to make routine use of the grid possible, through close collaborations with developers of all components involved where possible. By

[3] Hunter P., Nielsen P. A strategy for integrative computational physiology. Physiology (Bethesda) 20: 316-325. 2005.

working with different applications and addressing the common needs and individual requirements, reusable components may be isolated without sacrificing the customized environments demanded by users. The structure of resources supported by the National Center for Research Resources, such as NBCR and NCMIR, requires collaborative projects to guide the development of the tools, as noted above.

4.2. The Iterative cycle of science and technology development

The number of tools showed in Figure 5, along with those represented only as "…", offer many possible combinations to build end to end problem solving environments, often with overlapping features. For every tool listed, there are solid alternatives under different use cases. As shown in the usage scenarios, the ultimate choice of tools depends on the specific problems to be solved and the target audience of the designed environment. As demonstrated by the number of tools that are using Opal based services, the service oriented approach provides the flexibility in customized front-end tools with transparent access to underlying distributed computation resources. The demands of multiscale modeling applications will drive the development of the computation and data cyberinfrastructure, which in turn enables simulation based biology and medicine. The ensuing challenges in the search for solutions of more difficult problems will further this iterative developmental cycle of technology and science, with continuous and immediate impact to the health and well being of the public.

4.3. Future directions

In developing collaborations and ultimately a user base to conduct research, having a persistent infrastructure is required. By this we mean access to tools and software for researchers to experiment with. The National Centers for Research Resources provide that type of infrastructure for the community. In the case of NBCR, access to the tools and services discussed in this article can be found at the NBCR website.[9]

[9] NBCR: National Biomedical Computation Resource - http://nbcr.net. 2005.

While the experiences and developments discussed above have been informed by our local activities, the concepts of cyberinfrastructure are not bound by national borders. Examples of large scale projects and smaller team projects have been commented on in prior CTWatch issues (e.g., February 2006). Furthermore, it is essential that we actively participate in the international arena, both to ensure the various CI efforts can interoperate (e.g., when we want to use unique resources in other countries) and to reduce reproduction of tools that already exist.

Another exciting development that will link biomedical research with the environment across various scales is the newly funded CAMERA project. CAMERA will provide a community resource that links genetic information from a metagenomics effort, the Global Ocean Survey, with environmental factors such location, temperature and chemistry of the sampled environment.

The critical role played by NBCR to bridge the multiscale modeling research community and cyberinfrastructure is an important one, and will require collaborative efforts from other projects to expedite the translational/medical potential of multiscale biomedical and cyberinfrastructure research.

Acknowledgment
The authors are all members of the National Biomedical Computation Resource, which is supported by the NIH National Centers for Research Resources award P41-RR08605. Major support for APBS is provided by NIH grant GM069702 to NAB.

Specifications for the Next-Generation Computational Biology Infrastructure

Eric Jakobsson
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

All leading edge research in biology now utilizes computation, as a result of the development of useful tools for data gathering, data management, analysis, and simulation of biological systems. While there is still much to be done to improve these tools, there is also a completely new frontier to be attacked. The new initiatives to be undertaken will require much more interaction between applications scientists and cyberinfrastructure architects than has previously been the case. The single word that provides a common thread for the new initiatives needed in the next few years is *Integration*, specifically

- **Integration** of time and length scales of description.
- **Integration** of informatics, dynamics, and physical-based approaches.
- **Integration** of heterogenous data forms.
- **Integration** of basic science with engineering design.
- **Integration** of algorithmic development with computing architecture design.

Integration of time and length scales of description: Biological systems display important dynamics on time scales ranging from femtoseconds and faster (eg., interactions with electromagnetic radiation) to billions of years (evolution), and distance scales ranging from single atoms to the entire biosphere. Events at all time and length scales are linked to each other. For the most extreme example, the emergence of the photosynthetic reaction center (a protein that couples absorption of photons with synthesis of other biological molecules) over a billion years ago produced as a by-product a major change in the composition of the atmosphere (an increase in oxygen) that profoundly altered the course of biological evolution from that time on. Yet the vast majority of the computational tools that we use to understand biology are specialized to a particular narrow range of size and distance scales. We badly need computing environments that will facilitate analysis and simulation across time and length scales, so we may achieve a quantitative understanding of how these scales link to each other.

Integration of informatics, dynamics, and physics-based approaches: There are three core foundations of computational biology: a) Information-based approaches, exemplified by sequence-based informatics and correlational analysis of systems biology data, b) Physics-based approaches, based on biological data analysis and simulation founded in physical and chemical theory, and c) Approaches based on dynamical analysis and simulation, notably exemplified by successful dynamics models in neuroscience, ecology, and viral-immune system interactions. Typically these approaches are developed by different communities of computational biologists and pursued largely independently of each other. There is great synergy, however, in the three approaches when they are integrated in pursuing solutions to major biological problems. This can be seen notably in molecular and cellular neuroscience. Understanding of the entire field is largely organized around the dynamical systems model first put forth by Hodgkin and Huxley, which also had an underpinning of continuum physical chemistry and electrical engineering theory. Extension of the systems and continuum understanding to the molecular level depended on using informatics means to identify crystallizable versions of the membrane proteins underlying excitability. Physics-based computing has been essential to interpreting the structural data and to understand the relationship between the structures and the function of the excitability proteins. All areas of biology need a comparable synergy between the different types of computing. As a corollary, we need to train computational biologists who can use, and participate in developing, all three types of approach.

Integration of Heterogenous Data Forms: The types of data that are relevant to any particular biological problem are quite varied, including literature reports, sequence data, microarray data, proteomics data, a wide array of spectroscopies, diffraction data, time series of dynamical systems, simulation results, and many more. There is a major need for an integrated infrastructure that can enable the researcher to search, visualize, analyze, and make models based on all of the relevant data to any particular biological problem. The Biology Workbench[1] is a notable example of such integration in the specific domain of sequence data. This approach needs to be extended to much more varied and complex data forms.

[1] Biology Workbench - http://workbench.sdsc.edu/

Integration of Basic Science with Engineering Design: Biology is different from other basic sciences such as chemistry and physics, in the sense that adaptation for function is an integral part of all biological phenomena. Physical and chemical phenomena have only one type of cause; i.e., the underlying laws of physics. Biological phenomena have two types of cause: 1) the underlying laws of physics, and 2) the imperatives of evolution, which select the actualities of biology out of all the possibilities that one could imagine for how living systems are organized and function. In this sense, biological systems are like engineered systems - purpose contributes along with the laws of physics to define the nature of both biological and human-engineered systems. Elaborate and sophisticated computer aided design (CAD) systems have been developed to guide the creation of human-engineered devices, materials, and systems. The principles of CAD systems (optimization, network analysis, multiscale modeling, etc.) need to be incorporated into our computational approaches to understanding biology. A direct target for such cross-fertilization of biology and engineering is in nanotechnology, where we seek to engineer devices and processes that are on the size scale of biomolecular complexes. The Network for Computational Nanotechnology[2] is a notable cyberinfrastructure project in nanotechnology, with a versatile computational environment delivered through its Nanohub web site.[3]

[2] Network for Computational Nanotechnology - http://www.ncn.purdue.edu/

[3] nanoHUB - http://www.nanohub.org/

Integration of algorithmic development with computing architecture. The different types of biological computing have vastly different patterns of computer utilization. Some applications are very CPU-intensive, some require large amounts of memory, some must access enormous data stores, some are much more readily parallelizable than others, and there are highly varied requirements for bandwidth between hard drive, memory, and processor. We need much more extensive mutual tuning of computer architecture to applications software, to be able to do more with existing and projected computational resources. A remarkable instance of such tuning is the molecular simulation code Blue Matter, written specifically to exploit the architecture of the IBM Blue Gene supercomputer. The Blue Matter-Blue Gene combination has done biomolecular dynamics on a hitherto unprecedented scale and is directly enabling fundamentally new discoveries.[4]

[4] see Grossfield, A., Feller, S. E., Pitman, M.C. A role for direct interactions in the modulation of rhodopsin by -3 polyunsaturated lipids. PNAS 103: 4888-4893. 2006.

There is finally one more critical issue with respect to the development of a suitable cyber-infrastructure for biology. Our society is not training nearly enough prospective workers in the area of computational biology, nor enough quantitative biology researchers in general, to make progress in biological computing commensurate with the increased availability and power of computing resources. We need focused training at both the undergraduate and graduate levels to produce a generation of computational biologists who will be capable of integrating the physics, systems, and informatics approaches to biological computing, and to produce a generation of biologists who will be able to use computational tools in the service of quantitative design perspectives in understanding living systems. The need for such training is well articulated in the National Academy of Sciences report BIO 2010.[5] The first university to unreservedly embrace the BIO 2010 recommendations by fully integrating computing into all levels of its biology curriculum is the University of California at Merced.[6]

[5] http://www.nap.edu/books/0309085357/html/

[6] http://biology.ucmerced.edu/

Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade

Folker Meyer
Argonne National Laboratory

As genomes grow faster than Moore's law, biology will provide numerous cyber challenges in the next 10-15 years.

Biology is an area in science that is using more and more computational resources as it is turning into a data driven discipline. Most notably, the emergence of genome and post-genome technology has made vast amounts of data available, demanding analysis. Hundreds of bacterial (more precisely prokaryotic) genomes are available today and have already proven to be a very valuable tool for many applications. A prominent example is the *reconstruction of the metabolic pathways*[1] of several bacterial organisms. The analysis of the rising number of genomes is already an application of cyber technologies[2] and to some extent limited by the available cyber resources. As more data is becoming available, this trend is likely to continue.

An important factor in this equation is the fact that the number of available complete genomic sequences is doubling almost every 12-months[3] at the current state of technology. Whereas according to Moore's law, available compute cycles double only every 18 month. The analysis of genomic sequences requires serious computational effort: most analysis techniques require binary comparison of genomes or the genes within genomes. Since the number of binary comparisons grows as the square of the number of sequences involved, the computational overhead of the sequence comparisons alone will become staggering. Whether we are trying to reconstruct the evolutionary history of a set of proteins, trying to characterize the shape they fold into, or attempting to determine correspondences between genes in distinct genomes, we are often using these binary operations, and the cost is rapidly climbing.

Today, traditional research teams in bioinformatics either totally rely on resources provided by institutions like the National Center for Biotechology Information (NCBI)[4] for sequence analysis purposes[5] or build up their own local resources. The NCBI provides services including comprehensive sequence databases and online sequence comparison via a browser interface. Researchers possessing private compute resources have the advantage of running algorithms of their choosing on the machines however, to keep up with the data flood, they either have to accept long waiting times or continue to invest in cluster resources to fulfill their growing sequence analysis needs.

However as the number of sequences available grows, the number of algorithms available for their analysis also increases. So today, numerous bioinformatics techniques exist or are being developed that use considerably more computational power and are yielding different results for sequence comparison than the traditionally used BLAST algorithm. Over the last five years, most notably the influx of machine learning techniques has led to an increased consumption of compute cycles in computational biology.

When researchers began to use Markov Models to search for sequence similarities not visible with BLAST and also began building databases of common sequence motifs represented as Hidden-Markov-Models (e.g. HMMer or InterPro), the CPU requirements were increased dramatically. While a BLAST search against the NCBIs comprehensive, non-redundant collection of known proteins can be run in a matter of minutes either locally or on NCBI's BLAST-server for several hundred query sequences (remember a single genome contains thousands of genes),

[1] http://en.wikipedia.org/wiki/Metabolic_network_reconstruction_and_simulation

[2] Catlett, C., Beckman, P., Skow, D., Foster, I. Creating and Operating National-Scale Cyberinfrastructure Services. CTWatch Quarterly, 2(2), May 2006. http://www.ctwatch.org/quarterly/articles/2006/05/creating-and-operating-national-scale-cyberinfrastructure-services/

[3] GOLD: Genomes Online Database - http://www.genomesonline.org/

[4] National Center for Biotechnology Information - http://www.ncbi.nlm.nih.gov/

[5] http://www.ncbi.nlm.nih.gov/BLAST/

no resource exists that allows querying several hundred (let alone thousand) proteins for protein motifs using the European Bioinformatics Institutes' (EBI) InterPro tool.

Today, few resources exist outside TeraGrid[6] that could provide the computational power needed to run a comprehensive protein motif search for more than a few complete bacterial genomes. Only a massive, high-performance computing resource like TeraGrid can provide the CPU-hours that will be required for this and other future challenges stemming from the increasing amount of sequence data.

[6] TeraGrid - http://www.teragrid.org/

In Figure 1, the reason for developing new algorithms and looking for more computation power becomes apparent: we cannot generate annotations fast enough as the speed of sequencing accelerates. So applying new bioinformatics techniques as well as high throughput computing provides a much-needed means of reducing the growing gap between the number of sequences and annotations. Today we are clearly limited in our ability to generate annotations fast enough.

This limitation is currently of interest to people working in basic science, with the advent of more and more complete genomic sequences for crop-plants, pathogens and ultimately individual human beings. The demand for precise and fast bioinformatics analysis of genomes, not only from bacteria but also plant and human, is going to grow fast.

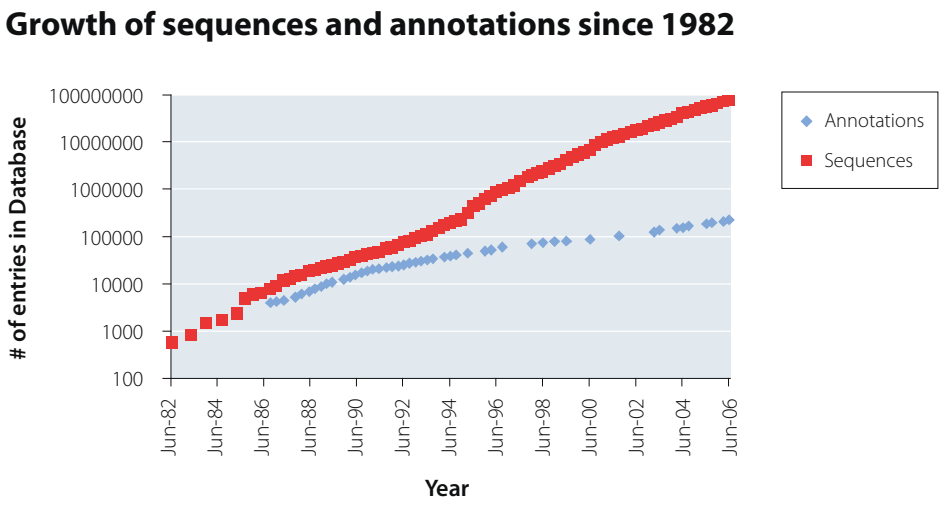

Figure 1: Using a logarithmic scale, the growth of sequence databases and annotations. Numbers are taken from the respective database release notes of the databases Genbank (NCBI) and Swissprot (EBI).

As daunting as our limited ability to generate annotations seems, we have so far only discussed a fraction of the challenges posed by biology. Annotations cover only the static components of the genome. They are a description of the gene load.

Ever since we have learned that the human genome contains relatively few genes (estimates are changing but all are below 50,000) it has become clear that the dynamics of gene expression and regulation thereof hold the key to understanding the organisms in question.

**Genome Sequencing vs. Moore's Law:
Cyber Challenges for the Next Decade**

As long as we are unable to fully enumerate, let alone describe, the functional elements in the respective genomes, we are a long way from understanding the full complexity hidden in the static and the dynamic components of the genome. Cyber Technologies will play a key role in furthering our understanding in understanding the data that we are currently amassing. While currently important insights into the respective organism's lifestyles can be obtained from studying the dynamic components of life (gene expression and regulation), we are at the beginning of another data deluge. The NCBI presents, as part of their training material, a comparison of the growth of sequence and gene expression data,[7] highlighting the fact that both are growing dramatically.

[7] http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Expression/exp45.html

The analysis of the growing volume of gene expression data becoming available from the various post-genomics technologies will present an even greater challenge than the annotation problem we are faced with right now. A single gene expression experiment can generate data for thousands of genes at a time, while gene expression studies have the potential of helping us understand annotation much better. Initially, we are faced with more data that not only needs integration with the annotations but also exceeds the annotations in volume and complexity.

While we are currently faced with the problem of generating annotations for the sequences we are producing, the next steps are already well defined, and it is clear that there is serious need for computational support of biology that in turn will require large-scale computation. Biology is in the middle of a paradigm shift towards becoming a fully data driven science.

Computing and the "Age of Biology"

Natalia Maltsev
Argonne National Laboratory

The past decade has completely changed the face of biology. The image of a biologist passionately chasing butterflies in the wilderness of an Amazon rainforest or losing sight spending hours staring in a microscope has been substituted by pictures of factory-like sequencing facilities and high-throughput automated experimental complexes. The technology has changed the entire fabric of biology from a science of lonely enthusiasts to a data-intensive science of large projects involving teams of specialists in various branches of life sciences spread between multiple institutions.The new generation of biology is tightly interlinked with the progress in computer science. Indeed, in order to exploit the enormous scientific value of biological data for understanding living systems, the information must be integrated, analyzed, graphically displayed, and modeled computationally in a timely fashion. The development of computational models of an organism's functionality is essential for progress in medicine, biotechnology, and bioremediation. Such models allow predicting functions of the genes in newly sequenced genomes and existence of particular metabolic pathways and physiological features. Conjectures developed during computational analysis of genomes provide invaluable aid to researchers in planning experiments and save an enormous amount of time and resources required for elucidation of an organism's biochemical and physiological characteristics. Essential for fulfilling this task is the development of high-throughput computational environments that integrate (i) large amounts of genomic and experimental data, (ii) comprehensive tools and algorithms for knowledge discovery and data mining, and (iii) comprehensive user interfaces that provide tools for easy access, navigation, visualization, and annotation of biological information. To achieve such an integrated environment, computational biologists face four primary challenges:

1. *Exponential growth of the biological data in the databases requires scalable public computational resources.* In the past 10 years the amount of data in genomic databases has doubled or tripled each year. For example, the current 151.0 release of the largest genomic database, GenBank, contains 56 billion bases, from 52 million sequences; and the rate of growth of GenBank is expected to increase dramatically as the cost of sequencing of new genomes drops. To date, 394 genomes have been completely sequenced and 1644 are on various levels of completion. However, the development of highly integrated and scalable bioinformatics systems for interpreting newly sequenced genomes is a time and resource-consuming task. While large sequencing or bioinformatics centers have the resources needed, a significant number of institutions are working on a small number of genome projects. Indeed, out of 1174 on-going sequencing projects, 66% of institutions have only one sequencing project, and over 87% have four or fewer.[1] The development of large, public computational systems that provide computing resources, integration of data, automated analyses, and the capability for expert-driven refinement of these analyses may significantly benefit the field of genomics. Such systems will provide state-of-the art computational support to the smaller genomics groups and will allow them to concentrate on biological scientific questions rather than on the development of bioinformatics environments.

2. *Complexity of biological data.* Increasingly, biological models are utilizing information from different branches of life sciences; genomics, physiology, biochemistry, biophysics, proteomics, and many more. The development of such models is a task of unprecedented computational, semantic, and algorithmic complexity. It requires integration of various classes of biological information as well as similar classes of data from different resources. Such large-scale integration presents a number of computer science challenges because of the large volume

and complexity of data, distributed character of this information residing in different databases, shortfalls of current biological ontologies, and generally poor naming conventions for biological objects (e.g., a large number of synonyms describing the same object or notion, or nonunique names for describing different objects). More than 100 groups (e.g., GO, BioPAX, W3C; see the Open Biomedical Ontologies website for a partial list of such efforts) are developing ontologies for various branches of biological sciences. For example, in order to satisfy diverse scientific communities, the ontological description of glycolysis has evolved into an enormously complex data structure integrating various classes of data, historical conventions of the communities, and links to the ontologies of other sciences (e.g., chemistry, biophysics, taxonomy). However, the development of most of these ontologies still depends on social consensus between the scientific communities—a task that seems to be of insurmountable social and scientific complexity. The development of new tools and algorithms for mining and clustering of existing scientific notions and terms may provide significant assistance to this process.

3. *Algorithm development.* The most popular bioinformatics tools (e.g., BLAST, FASTA) perform pairwise comparisons of the query sequence with all sequences in a specified database. Such a computationally intensive approach at a time of exponentially growing amounts of sequence data will inevitably lead to the emergence of an N-squared problem. Bioinformatics will significantly benefit from the development of a new generation of algorithms that will allow efficient data mining and identification of complex multidimensional patterns involving various classes of data. Visualization of multifarious information is another essential need of high-throughput biology: it allows for reducing the complexity of biological knowledge and developing much-needed overviews.

4. *Development of collaborative environments.* A typical biological project involves data sources and users distributed among various institutions. Such projects require a mature infrastructure that allows seamless integration, analysis, storage, and delivery of information to a distributed community of users. Warehousing of the data and its analysis by the researchers residing in one location will not be sufficient for the needs of biology in the future. Essential for the success of large biological projects is further development of collaborative environments that will allow the scientists residing in different locations and sometimes even on different continents to analyze, discuss, annotate, and view the data. Access Grid conferencing, shared interfaces, Web services, and other collaborative tools will allow groups to identify, discuss, and solve scientific problems efficiently.

The 21st century is considered to be the "age of biology." Advances in genomic research will establish cures or therapies for numerous diseases that were considered to be incurable, and future genetically engineered bioproducts will contribute significantly to solving the global hunger problem. Such progress will, to a large extent, be driven by the formulation of new computational approaches for analysis of biological data and the timely transfer of technologies developed in other disciplines (e.g., physics, linguistics). Computing and the biological sciences will become intimately intertwined, opening new possibilities and cause unprecedented changes to life as we know it.

CTWatch QUARTERLY

ISSN 1555-9874 Volume 2 Number 3 **August 2006**

Trends and Tools in Bioinformatics and Computational Biology

GUEST EDITOR: **RICK STEVENS**

AVAILABLE ON-LINE:
www.ctwatch.org/quarterly/

E-MAIL CTWatch QUARTERLY:
quarterly@ctwatch.org

CTWATCH IS A COLLABORATIVE EFFORT

http://icl.cs.utk.edu/

http://www.ncsa.uiuc.edu/

http://www.sdsc.edu/

CTWATCH IS A PUBLICATION OF THE CYBERINFRASTRUCTURE PARTNERSHIP

www.ci-partnership.org

SPONSORED BY