New Generation Computing, 22(2004)97-110
Ohmsha, Ltd. and Springer-Verlag

Preface

Life Sciences and Cyberinfrastructure: Dual and Interacting Revolutions that will Drive Future Science

Peter W. ARZBERGER
University of California San Diego
9500 Gilman Dr, mc0043
La Jolla, CA 92093-0043 USA
parzberg@ucsd.edu

Abbas FARAZDEL
IBM Life Sciences
2455 South Road
MS P099, Poughkeepsie, NY 12601 USA
farazdel@us.ibm.com

Akihiko KONAGAYA
RIKEN, Genomic Sciences Center
W-520, 1-7-22 Suehiro, Tsurumi, Yokohama,
Kanagawa, 230-0045 Japan
konagaya@gsc.riken.go.jp

Larry ANG
Bioinformatics Institute
21 Heng Mui Keng Terrace
I2R, Level 3, Singapore 119612

Shinji SHIMOJO
Cybermedia Center and Biogrid, Osaka University
5-1 Mihogakaoka, Ibaraki, Osaka 567-0047 Japan
shimojo@cmc.osaka-u.ac.jp

Rick L. STEVENS
Mathematics and Computer Science Division
Argonne National Laboratory
9700 South Cass Avenue
Argonne, IL 60439, USA
stevens@mcs.anl.gov

Received 21 September 2003
Revised manuscript received 30 November 2003

Abstract Over the past quarter century, two revolutions, one in biomedicine, the other in computing and information technology leading to cyberinfrastructure, have made the largest advances and the most significant impacts on science, technology, and society. The interface between these areas is rich with opportunity for major advances. The Life Sciences Grid Research Group (LSG-RG) of the Global Grid Forum recognized the opportunities and needs to bring the communities together to ensure the cyberinfrastructure will be constructed for the benefit of science. This article gives an overview of the area, the activities of the LSG-RG, and the minisymposium organized by LSG-RG, and introduces the papers in this Special Issue of New Generation Computing.

Keywords: Life Sciences, Grid, Cyberinfrastructure, International Collaborations.

§1 Introduction

Over the past quarter century, two revolutions, one in biomedicine, the other in computing and information technology, have made the largest advances and the most significant impacts on science, technology, and society. These scientific revolutions continue unabated, are actually accelerating, and are creating an exciting frontier at the interface with opportunities for major advances in biological understanding, and in turn, will have a key impact on translational medicine and subsequently clinical practice.

The revolution that continues to characterize the biomedical sciences today - which has been compared with that of physics at the beginning of the past century - began with the discovery of the chemical underpinnings of life and heredity some 50 years ago. In the past decade, our understanding of biology has grown rapidly due to our ability to capture ever more data across the biological spectrum, from atoms to entire populations. Recently, the so-called blueprints for life, the complete DNA sequences of many model organisms (e.g., microbes, mouse) and humans, have been published. Many more will follow, enabling a new era of computational comparative biology, with the ability to exploit model genomes to help us understand the molecular basis for human disease and health. Furthermore, capitalizing on these advances and on new high throughput technologies, structural genomics, also known as the protein structure initiative, aims to determine the 3-D structures of on the order of 10,000 proteins in order to collect the vast majority of all possible domain structures. Based on modeling studies to date, having these structures will drive more accurate predictions of all protein structures, which in turn can be used to produce a better understand of their function. Similarly, advances in imaging technology are starting to reveal anatomical locations of functional activity of the intact nervous systems in the course of sensory perception and cognition. These advances are being driven by information technology.

The revolution in information technology has led to changes in the entire computational infrastructure during the last decade, moving from central-

ized major resources to a distributed computational, data and observation platforms, a cyberinfrastructure, where sequences of operations can be composed into workflows across these resources. This great change is driven by advances in technology: over the last nine years, compute capacity has increased 64-fold; storage, 512-fold; and the leading driver network bandwidth, 4096-fold.[1)] Together, these trends allow for the integration of data, instruments, scientists, networks, software, and high-performance computing, into a cyberinfrastructure (CI).[2)] CI provides universal desktop access, via the Internet, to distributed resources, global collaboration, and the intellectual, analytical, and investigative output of the world's scientific community via e-science initiatives. A main function of cyberinfrastructure will be to knit together the disparate communities, with expertise in basic science and in clinical practice that will be required to generate breakthroughs in our understanding of disease processes and to subsequently develop effective clinical approaches

§2 International Perspective

The vision of cyberinfrastructure is not limited to one discipline, institution, country; it is international. This international perspective is driven by several factors.

Science is an intrinsically global activity with an increasing number of large-scale, international projects and initiatives[*1] and with the idgrowing recognition that certain problems are global in scale and in need of resources to address the, e.g. environment and health. In addition, ***grid is transforming computing and collaborating***.[3~5)] The scientific community has begun to establish standards groups[*2] and nations and regions have made significant investments and created activities.[*3] Furthermore, ***the grid is progressively getting easier to use or deploy***, with more applications being deployed to take advantage of it, but much more needs to be done to make it usable on a daily basis. However, ***interoperable middleware software is needed*** to ensure that the grid is truly global in practice, and will enable researchers to use the grid on a daily basis. In short, the type of problems scientist address increasing take on global proportions. The CI, where information technology (computers, storage, networks) meets applications and scientific instruments, is exploding in both size and scope, and holds the potential to address many global science issues. However, much work needs to be done to make the grid usable.

[*1] Global Biodiversity Information Facility http://www.gbif.org; National Virtual Observatory (NVO) http://www.srl.caltech.edu/nvo; The Data Grid Project http://www.eu-datag.org; Grid Physics Network http://www.griphyn.org/index.php

[*2] Global Grid Forum http://www.globalgridforum.org

[*3] UK Research Councils e-Science Programme http://www.research-councils.ac.uk/escience/; The Data Grid Project http://www.eu-datag.org; NASA Information Power Grid http://www.ipg.nasa.gov; TeraGrid http://www.teragrid.org; Asia Pacific Grid http://www.apgrid.org; Asia Pacific Grid (APGrid) http://www.apgrid.org; Asia Pacific Advanced Network http://www.apan.net; Pacific Rim Application and Grid Middleware Assembly (PRAGMA) http://www.pragma-grid.net

§3 Global Grid Forum

The Global Grid Forum is one organization that is focused on broader grid activities and standards. GGF is a community-initiated forum of more than 5000 individual researchers and practitioners from over 400 organizations in over 50 countries, with a primary objective of promoting and supporting the development, deployment, and implementation of grid technologies and applications via the creation and documentation of best practices - technical specifications, user experiences, and implementation guidelines.

GGF focuses its activities to facilitate and support the creation and development of regional and global grids that will provide to the scientific community, industry, government and the public at large dependable, consistent, pervasive and inexpensive access to high-end computational, data and other resource capabilities; to address architecture, infrastructure, standards and other technical requirements for grids and to facilitate and find solutions to obstacles inhibiting the creation of these grids; and to educate, facilitate the application of grid technologies, and provide a forum for exploration of grid technologies among the scientific community, industry, government and the public.

GGF is organized into areas[*4] where activities are carried out in either ***working groups*** (focused on a particular, specific problem, technology, or opportunity for which they will deliver a document or series of documents in a finite time frame) or ***research groups***, which have longer-term horizons where it may be premature to develop technical specifications or recommendations.

§4 Life Sciences Grid Research Group

The Life Sciences Grid (LSG) Research Group, founded in late 2002, explores issues rat the rich interface of life sciences and cyberinfrastructure technologies. LSG Research Group is the first and to date the only GGF research group focused explicitly on a specific discipline (rather than a technology),The initial charter[*5] is broad by design, including all scales of biological processes, from atoms to populations, as well as issues of handling multiple scales of biological complexity. A breadth of grid technologies are necessary to handle the capture and initial manipulation of data from instruments and the images they produce, to the reduction of those data and deposition into databases, to the incorporation of those data into increasingly realistic simulations of biological processes, and to the integration of resources from a variety of sources, separated by distance, by discipline, by security policy and by language.

From the charter, specific goals of the LSG group include: 1. Identify different solution areas and classify them; 2. Explore possible reference architectures for each solution area; 3. Identify clear examples and diverse use of grid within life sciences; 4. Discuss issues of access to data within life sciences; 5. Discuss state of standards, within subdisciplines and between subdisciplines of

[*4] The seven areas of GGF activities include: Grid Information Services (GIS), Scheduling and Resource Management (SRM), Security (SEC), Grid Performance (GP), Architectures and Frameworks (ARCH), Data, and Applications, Programming Models and User Environments (APE).

[*5] https://forge.gridforum.org/projects/lsg-rg/document/LSG-RG_Charter/en/1

life sciences; 6. Identify how the grid is being challenged by life sciences, and where there is need for activity; and 7. Identify linkages to other GGF research and working groups

§5 LSG Minisymposium

As a first step to engage the LSG research community, a minisymposium[*6] was held in March 2003 specifically address goals 3, 6 and 7.

The workshop reflected the diversity of applications calling for the use of grid technologies, ranging across ***biological scales*** (from quantum chemistry calculations, to molecular and protein sequence and structure analysis and annotation, to systems biology from cells to organs, to the neuroscience of understanding structure and function of the brain, to biodiversity studies of a global nature, to health grids and beyond), across the ***technology components of cyberinfrastructure*** (from larger compute resources or more efficient throughput calculations, to manipulation of data for sharing and analysis, to control of instruments, to finally tools to enhance collaborations) across stages of ***development*** (from early research, to prototype systems to instantiations of testbeds and production systems for the life sciences); and finally across ***scope of project*** (from investigator lead efforts, as well as national, regional and international efforts to marry grid technologies with life science applications). Also discussed were areas of technological challenge, as described below.

5.1 Grid Initiative Examples

The cost of producing data in many instances is very high, thus the ability to share data with other researchers allows for both a more efficient use of the initial investment to produce the data, but also allows for stronger powers of inference. With the maturity of several grid technologies, such as grid middleware (Globus), portal technologies (NPACI GridPort), and the distributed data storage and retrieval environments (Storage Resource Broker), the National Institutes of Health's Nation Center for Research Resources launched a bold and successful program call the ***Biomedical Informatics Research Network(BIRN).***[*7] This program, described by Mark Ellisman, Director of the BIRN Coordinating Center and Professor of Neuroscience at UCSD, is putting in place a scalable infrastructure from the brain research community to share and combine data among researchers at different institutions, and across the research and clinical communities. The development of this infrastructure is driven by three key application communities[*8]: Brain Morphology, Mouse Models of Disease, and Functional Imaging of Schizophrenic Humans. This infrastructure is in place

[*6] The six sessions included an Overview, BioGrid Inititiatives, Portals and Workflows, Global and Regional Initiatives, Protein and Molecular Informatics, Summary of Session and Next Steps. See either http://www.pragma-grid.net/Presentations/GGF7/ggf7_presentations.htm or https://forge.gridforum.org/docman2/ViewCategory.php?group_id=39&category_id=386

[*7] http://www.nbirn.net

[*8] http://www.nbirn.net/Publications/Articles/20030715-NYT-Grid-ME.pdf. The Brain Morphology BIRN involves pooling and processing magnetic resonance imaging data to look for early anatomical and functional precursors of Alzheimer's disease. That knowledge may then be used to tailor drugs to inhibit the onset of the disease.

and is being used by researchers at more than 15 institutions, with plans to expand this to many more. Additionally noteworthy is the challenge to the social interactions between scientists who have not traditionally shared data, a challenge that will be increasingly relevant to other CI activities.

In another overview, Rick Stevens, Director, Mathematics and Computer Science Division, Argonne National Laboratory and Professor of Computer Science at the University of Chicago, described the exciting future of systems biology and how grid technologies will help researchers gain access to data and computing power to solve currently intractable problems.

Additional talks by Arzberger on the National Biomedical Computation Resource,[*9] Nakamura on the BioGrid,[*10] and Breton on HealthGrid[*11] also stressed the challenges of integration of models and data across biological scales as an area ripe for development of grid technologies.

Several multi-national efforts were discussed. ***Global Biodiversity Information Facility (GBIF)***[*12] has the stated goal of making the world's biodiversity information freely available. Much of the current information on biodiversity resides in museum collections, but increasingly those collections will be linked to genetic and other environmental and geographic databases. This global effort requires the grid to link intelligently the information in these distributed databases and to provide the computing infrastructure to analyze the data, with the potential impact of helping to understand the spread of invasive species or identifying habitat for preserving known species based on environmental factors.

Another effort is the ***Pacific Rim Application and Grid Middleware Assembly (PRAGMA)***,[*13] which is an open institution-based organization aimed at establishing sustained collaborations and advancing the use of the grid technologies in applications among a community of investigators around the Pacific Rim. From its inaugural meeting in 2002, it has grown to 20 institutional members. PRAGMA applications (See brochure, see envision article) have demonstrated the value of all key components of the cyberinfrastructure, from the control of remote instruments such as a microscope for neuroscience research in Osaka or wireless sensing devices for environmental research in Taiwan, to distributed computing in areas of quantum chemistry (see Sudholt) and tomography, to controlling, manipulating and administering distributed data files at many sites, as need by the high-energy physics community, to rapid deployment and use of access grid technology to assist health care providers in coping with isolation of patients and doctors in the spring 2003 outbreak of SARS in Taiwan.[*14] PRAGMA is an example of a model for collaboration to demonstrate the value of the grid in applications, to the sharing and testing of middleware, and to the construction of a testbed for applications to use, in which researchers can also begin to understand the aspects of widely share resources

[*9] NBCR, http://nbcr.ucsd.edu
[*10] BioGrid, http://www.biogrid.jp
[*11] HealthGrid,http://www.healthgridorg
[*12] http://www.gbif.org
[*13] http://www.pragma-grid.net
[*14] http://antisars.nchc.gov.tw/, http://www.nchc.org.tw'

5.2 Workshop Session Summaries

Other ***BioGrid Initiatives*** were presented. In the *Biomedical Grid Initiatives in Europe*, V. Breton, provided the status and plans of European Data Grid[*15] in regards to biomedical applications available (in phylogenetics, Parasitology, Medical Data and metadata management), and presented activities such as the HealthGrid.[*16] V. Batra described the *North Carolina BioGRID*: Challenges in Grid Deployment and Application Enablement[*17] which will provide access to, and coordination of, the computing, data storage, and networking infrastructure required for researchers and educators throughout North Carolina to take full advantage of the genomics revolution, where they chose Avaki as the metascheduler for the grid and the deployment of Turbo Blast on the grid. In the talk on ***EUROGRID and GRIP - grid development for biomolecular biology***,[*18] P.Bala (see article this issue by J. Wypychowski *et al.*) gave a brief introduction to Eurogrid and heard about the development of BioGRID, which will develop an access portal for biomolecular modeling resources, to enable chemists and biologists to be able to submit work to HPC facilities. The main task of Bio-GRID is to integrate selected applications with UNICORE infrastructure and provide easy tools for non-experts in high performance computing.

In the session on ***Portals and Workflows***, A. Konagaya reported on the *Web Portal Service on the Open Bioinformatics Grid* (OBIGrid)[*19] giving an overview of OBIGrid, the development of tools for integration of databases, and several applications, such as OBITco(Open BioInfomatics Thermus thermophilus Cyber Outlet), OBIYagns (a cell simulation environment), and OBISgd (Scalable Genome Database) [please see the article in this issue by A. Konagaya *et al.*). In *Web Services and Genome Annotation in GRID*, H. Sugawara and S. Miyazaki reviewed the activities to make the DNA Data Bank of Japan[*20] (DDBJ) available via webservices (XML/SOAP technology) with the goal of integrating diverse data resources. In *A Workflow and Grid Computing Systems for Molecular Simulation*, K.Jeong and S.Hwang described an approach to create workflows in such a way as to use legacy software (without the need to change code), with applications to molecular docking and molecular simulation. In a final series of talks on *High Throughput Proteomics and the Encyclopedia of Life* (EOL),[*21] *iGAP - Integrated Grid-enabled Genome Annotation Pipeline*, M. Miller, W. Li, and A. Shahab provided an overview of an ambitious project (EOL) to catalog the complete proteome of every living species in a flexible, powerful reference system. EOL would take advantage of iGAP, which uses AppLeS Parameter Sweep Template (APST) to provide transparent access to Grid resources and smart scheduling via Grid middleware.

In the ***Global and Regional Initiatives*** session, A. C. Jones, W. A. Gray

[*15] http://eu-datagrid.web.cern.ch/eu-datagrid/ or http://web.datagrid.cnr.it/LearnMore/index.jsp
[*16] http://healthgrid.org
[*17] http://www.ncbiogrid.org/ and http://www.gridtoday.com/02/0916/100378.html
[*18] http://www.eurogrid.org, http://biogrid.icm.edu.pl
[*19] http://www.obigrid.org/
[*20] http://www.ddbj.nig.ac.jp/
[*21] http://eol.sdsc.edu

and H. Saarenmaa described Facilitating Access to Biological Information with a Global Catalogue of Life. The talk introduced the Species 2000 project,[*22] which is to enumerate (in a catalogue) all known species of plants, animals, fungi and microbes on Earth as the baseline dataset for studies of global biodiversity; and SPICE for Species 2002,[*23] which is a tool to build a federated 'registry' of scientific names organised by taxon to allow for interoperation among resources based on different taxonomies and a query on them, and GRAB (Grid And Biodiversity)[*24] which illustrates the grid's potential for collaborative research, discovering & using diverse biodiversity-related databases. In *Biodiversity World: a GRID-based Problem Solving Environment for Global Biodiversity*,[*25] W. A. Gray and others outlined efforts to bring together various databases of different types (taxonomy, climate, geography) to produce analysis such as modeling species distributions against climate change, conservation prioritization and linking evolutionary changes to past climates. These tools would be of use to the Global Biodiversity Information Facility (GBIF). Finally, P. Arzberger described two collaborative, multiple investigator, activities to build tools to enhance the grid, namely the National Biomedical Computation Resource (NBCR),[*26] Pacific Rim Application and Grid Middleware Assembly (PRAGMA)[*27] (described elsewhere in this paper).

In the session on ***Protein and Molecular Informatics***, H. Nakamura (see also this special issue) presented *BioGrid - Grid Applications to Protein Informatics*,[*28] a large, national effort to develop a computer grid technology to meet IT needs specialized in biology and medical science. The project is composed of four cores on basic grid, computational grid, data grid, and remote data collection systems technology. The talk highlighted applications of hybrid quantum mechanical and continuum mechanical interactions, which push computing capability. A. Krishnan described *GridBLAST: High Throughput BLAST* giving an overview of the architecture and results of GridBlast on a grid architecture. I. Ono (see article in this issue) outlined *Grid-oriented Genetic Algorithms for Large Scale Optimization Problems in Bioinformatics* describing a distributed environment to run genetic algorithms with globus. K. Satou described *OBIEnv: An Environment for Typical Computing in Bioinformatics,* an environment designed to run a large number of similar tasks, where the program would automatically divide up the tasks and select machines and monitor the process (the jobs could even be monitored by mobile phones). A final talk of the session by B. Sorbal on *Data and Tool Integration and Interpolation in Life Sciences: Examples and Challenges*, presented ToolPath and ToolBus tools to create integration framework based on the WebServices technology.

In other sessions, C. Goble described a *High level Knowledge-based Grid ser-*

[*22] http://www.sp2000.org/
[*23] http://www.systematics.reading.ac.uk/spice/
[*24] http://www.gridoutreach.org.uk/docs/pilots/grab.htm
[*25] http://www.bdworld.org
[*26] http://nbcr.ucsd.edu
[*27] http://www.pragma-grid.net
[*28] http://www.biogrid.jp

vices for Bioinformatics (MyGRID),[*29] a research project that will extend the Grid framework of distributed computing, producing a virtual laboratory workbench that will serve the life sciences community. All components of this work environment are open source, web services based and OGSA compliant. J. Brooke outlined *REALISTE - Realistic Modeling in Environmental and Life Science Through E-science,*[*30] giving the philosophy of REALISTE to address the issues of complexity and knowledge creation and management. It utilizes work in myGrid and Reality Grid UK e-science Project and also combines with Framework Programme 5 Grid projects EUROGRID, GRIP and DataGrid.

A. Krishnan summarized the status and interim results of a survey conducted on behalf of the Life Sciences Grid Research Group (the final results are one of the papers in this special issue).

5.3 Common Needs of the Life Sciences Grid Community

One of the goals of the meeting was not only to present applications using or needing the grid, but also to draw attention to the unmet needs of the Life Science community. Some of the needs that came from the talk include the following.

Handling ***diverse and heterogeneous data*** is a common need. The life sciences are an information science. In particular, specific needs included the storage (and long-term persistence), querying, finding intelligently new sources of data, and integration of data from different resources and across scales of science. Aspects of these issues were highlighted in presentations on BIRN, OBIGrid, DDJB, Toolpath, Toolbus, and Biodiversity World.

Using the grid for gaining access to large ***compute resources*** was motivated by talks on the BioGrid (the hybrid modeling), and the development of UNICORE. Then, needed to monitor the use of those machines and the processing was illustrated in such projects as OBIGrid and OBIEnv.

Workflows and portal technologies were illustrated in several examples, ranging from iGAP. The issue of standards for interfaces came up. The ***control of instruments*** was illustrated in the example on telescience. The need for improved ***collaborative tools and environments*** was demonstrated by the SARS-Grid, and addressed in talks on REALISTE, PRAGMA and MyGrid. Issues of ***security*** relate to giving access to resources (compute, data, equipment).

Other issues that were raised is the need for ***on-demand computing***, developments of ***semantic web***, and a ***reference architecture*** for life science applications. Finally, issues that came up in several talks revolved around ***data access policies***, in light of national security, privacy (medical grids), and intellectual property rights.

§6 Overview of Papers in this Special Issue of New Generation Computing

It is the intent of the editors of this special issue to present some key

[*29] http://www.mygrid.org.uk
[*30] http://www.esnw.ac.uk/projects.shtml

examples of the use of the grid by the Life Sciences community. It is hoped that by showing, by example, how the grid is being used, it will motivate others in this community to examine the tools that have been developed, to explore the quickly evolving landscape of middleware, and ultimately to test and contribute new tools and experiences to this international effort.

The papers in this special issue reflect one of the following characteristics: Demonstrate concretely the impact of the grid on life sciences to gain insights into biological processes; Have tangible infrastructure that can be used by others; Demonstrate international collaborations in building the grid or in doing applications; and Have tangible components rather than purely prospective in nature.

The ten papers in this issue can be categorized as follows:

Survey of the field: As an activity of the Life Sciences Grid Research Group, Arun Krishnan conducted and reports on *A Survey of Life Sciences Applications on the Grid*. He compiled the results of a survey[*3] and provides an analysis. Among the respondents, "most of the problems being solved [currently] on the grid are either embarrassingly parallel in nature or make use of domain decomposition techniques." Furthermore "a few applications use remote procedure calls (RPC) or RPC like programming paradigms (e.g., Ninf) but for the most part, applications are being written in easy to use scripting languages or ... else by using the grid version of MPI, viz. MPICH-G." But most of Krishnan's survey focused on compute problems. He concludes his analysis by stating that many database access and query applications are being grid-enabled, which will be important for life sciences and lead to broader usage of CI.

International Applications of Grid: The two papers here represent international efforts, driven by the science. In one case, *The Encyclopedia of Life Project: Grid Software and Deployment*, by Wilfred Li *et al.*, described the vision of creating a catalog of the complete proteome of all species, with plans to provide current functional and structural annotations tools. The work involves tracking distributed workflows, and is built upon middleware such as AppLeS (Application-Level Scheduling) Parameter Sweep Template (APST), integrative genome annotation pipeline (iGAP), and more recently the GridSpeed technology. The second paper, *Parameter Scan of an Effective Group Difference Pseudopotential Using Grid Computing*, by Wibke Sudholt *et al.*, is motivated by modeling complex molecular systems, such as those in drug design. The paper describes bringing together the quantum chemistry codes, GAMESS with Nimrod/G a distribution tool that tracks and manages distributed runs of GAMESS on a widely distributed grid. The results give new insight into the science, and demonstrate the power of the grid in completing the calculations in far greater time than in most individual machines. Finally, in both of these cases, these groups have either started or expanded their collaborations via PRAGMA (described earlier).

Regional Grid Activities: Three regional efforts are described that involved development of grids and tools, with a focus on life sciences. In *Life Sciences Grid in EUROGRID and GRIP Projects*, J. Wypychowski *et al.* describe the UNICORE

[*3] http://www.bii.a-star.edu.sg/ggf

software, in the context of EUROGRID, as a framework providing uniform access to a number of resources and applications important for life science, with a focus on computational chemistry and molecular biology. *A Challenge towards Next-Generation Research Infrastructure for Advanced Life Science*, by Haruki Nakamura *et al.*, describes lays out an ambitious framework to tackle a broad set of life science problems, that include challenges of integrating data and models from various scales of biological processes. *The Superstructure towards Open Bioinformatics Grid*, Akihiko Konagaya *et al.*, describes the structure of and applications using this new infrastructure,

Framework for a Specific Class of Problems: *A Grid-Oriented Genetic Algorithm Framework for Bioinformatics*, by Hiroaki Imade *et al.*, describes a framework for enabling researchers of genetic algorithms (GA), an important class of algorithms in the life sciences, to easily develop GA running on the grid.

Short Technical Notes on Applications on Grid: To reflect the diversity of life science applications on the grid, three technical papers have been included. These example include *Biological Structure Determination by EM is Well Suited to Grid Computing*, J. J. Fernandez *et al.*; *MOLECULAR DOCKING: An Example of Grid Enabled Applications*, Habibah A. Wahab *et al.*; *Computational Proteomics on the Grid*, E. Huedo *et al.*. It is important to note that these applications indicate that the grid has moved to many countries and institutes around the world, and applications are built by many researchers. In particular, these examples illustrate the worldwide contribution to the growing development of cyberinfrastructure.

§7 Conclusions

The dual revolutions in biology and biomedical research on the one hand and information technology on the other, in particular in the broader grid and CI technology, have a future full of mutually beneficial opportunity and need. The Life Sciences Grid - Minisymposium and this special edition of New Generation Computing, several concrete examples of grids or grid applications in the life sciences currently in use. Many more potential examples exist. The challenge for this community is to move from potential to reality, across the broad array of grid technologies and in all aspects of life sciences research.

To do that will involve clearly articulating the technological needs and working in multidisciplinary teams. We have seen the need for better collaborative tools (e.g., SARS), the need for broader access to data (e.g., BIRN), the need for access to computing. In addition, as pointed out in several articles, there are other needs of security for data and access to resources, there are needs for being able to integrate data from a variety of sources, there are needs in the area of semantic web and mediation, there is a need for distributed resource management.

Finally, it is critical that our efforts are conducted in an international arena. Issues of metadata standards cannot be solved by researchers in any one country, the demonstration of grid for solving international problems in areas of health and the environment needs international participation, and finally,

from a purely practical perspective, there are not enough resources to allow for many replicated experiments or non-connected infrastructure. Science is an open international effort, the grid community has embraced an open setting of standards via GGF and a philosophy of open source software; so the Life Sciences Research Community should lead the way in demonstrating the value of broad international and collaborative efforts.

Acknowledgements

P. Arzberger wishes to acknowledge the support from grants NBCR (NIH/NCRR P41 RR08605) and PRAGMA (NSF INT 0314015) and would like to express his gratitude to all of the PRAGMA members. A. Farazdel gratefully acknowledges the continual support of the IBM Life Sciences organization. A. Konagaya gives a special thanks to a support from Grant-in-Aid for Scientific REsearch on Genome Information Science from the Ministry of Education, Culture, Sports, Science and Tehnology of Japan and to all of the OBIGrid project members.

References

1) Stix, G., "The Triumph of the Light," *Scientific American, 284*, pp. 80–86, 2001.

2) Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P. and Wright, M. H. (n.d.), "Revolutionizing science and engineering through Cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on Cyberinfrastructure, Draft 1.0. as of November 11, 2002. National Science Foundation, Directorate for Computer and Information Science and Engineering, Advisory Committee for Cyberinfrastructure site: https://worktools.si.umich.edu/workspaces/datkins/001.nsf/Resources/F13F12AA2F8A56C685256BA0005A2CDA/$FILE/CI-DRAFT-1.0-4-19.pdf"

3) Campanelli, M. ed., "What is A Grid?," in *Grid Physics Network Outreach Center*, http://www.aei.mpg.de/ manuela/GridWeb/info/grid.html.

4) Foster, I, C. Kesselman ed., "The Grid: A Blueprint for a New Computing Infrastructure," http://www.mkp.com/grids/, 1998.

5) Wladawsky-Berger, I., "Grids will Transform Computing," *Primeur*, http://www.hoise.com/primeur/02/article/monthly/AE-PR-01-02-5.html (http://www-1.ibm.com/servers/events/pdf/transcript.pdf, address to Kennedy Consulting Summit 29 Nov 2001).

Peter Arzberger, Ph.D.: He is the Director of Life Sciences Initiatives, University of California San Diego; Director of the National Biomedical Computation Resource (http://nbcr.ucsd.edu), funded by the National Center of Research Resource of NIH; and the Chair of the Pacific Rim Application and Grid Middleware Assembly (http://www.pragma-grid.edu), an organization of 20 institutions around the pacific rim whose mission is to establish sustained collaborations and to advance the use of grid technologies in applications. He serves on the US National CODATA Committee and the National Advisory Board of the US Long Term Ecological Research. His hobby is working on Lloyds.

Abbas Farazdel, Ph.D.: He is a Senior Scientist and an IT Solution Strategist in the Advanced Technologies unit at the IBM Life Sciences. Previously, Dr. Farazdel worked at several positions in IBM including Cluster System Strategist; Data Warehousing and Data Mining Solutions Implementation Manager; and High Performance Computing Consultant. Abbas is the co-chair of the Global Grid Forum (GGF) Life Sciences Grids Research Group. He serves on the Scientific Board of the European HealthGrid and the Mid Hudson Technology Council of New York. Abbas received his Ph.D. in Quantum Chemistry and M.Sc. in Computational Physics from the University of Massachusetts concurrently.

Akihiko Konagaya, Dr.Eng.: He is Project Director of Bioinformatics Group, RIKEN Genomic Sciences Center. He received his B.S. and M.S. from Tokyo Institute of Technology in 1978 and 1980 in Informatics Science, and joined NEC Corporation in 1980, Japan Advanced Institute of Science and Technology in 1997, RIKEN GSC in 2003. His research covers wide area from computer architectures to bioinformatics. He has been much involved into the Open Bioinformatics Grid project since 2002.

Larry Ang: As the Project Director in the Bioinformatics Institute (BII), he is in charge of major international collaborative projects on biomedical grids between BII and other research organizations (http://web.bii.a-star.edu.sg/ larry/). In particular, he works actively with bodies such as Pragma where he serves on the Steering Committee. He is also the Secretary of the Life Sciences Grid Research Group of GGF (Global Grid Forum) He serves on the Gelato Federation; Gelato was started by HP Labs and pushes open source software on linux platforms.

Shinji Shimojo, Ph.D.: He received his M.E. and Ph.D. degrees from Osaka University in 1983 and 1986, respectively. He was an Assistant Professor with the Department of Information and Computer Sciences, Faculty of Engineering Science at Osaka University from 1986, and an Associate Professor with Computation Center from 1991 to 1998. During the period, he also worked as a visiting researcher at the University of California, Irvine for a year. He has been a Professor with Cybermedia Center (then Computation Center) at Osaka University since 1998. His current research work is focusing on a wide variety of multimedia applications, peer-to-peer communication networks, ubiquitous network systems and Grid technologies. He is a member of ACM, IEEE and IEICE.

Rick L. Stevens, Ph.D.: He is Professor, University of Chicago; director, Mathematics and Computer Science Division/Argonne National Laboratory; director, ANL/UC Computation Institute; project director for National Science Foundation supported TeraGrid project; head of the Argonne/Chicago Futures Lab. He is interested in the development of innovative tools and techniques that enable computational scientists to solve important large-scale problems effectively on advanced scientific computers. His research focuses on three principal areas: advanced collaboration and visualization environments, high-performance computer architectures (including Grids), and computational problems in life sciences and systems biology. He teaches courses on computer architecture, collaboration technology, virtual reality, parallel computing, and computational science.