Chapter XVII
Molecular Structure Determination on the Grid

Russ Miller
Hauptman-Woodward Medical Research Institute, USA
SUNY-Buffalo, USA

Charles M. Weeks
Hauptman-Woodward Medical Research Institute, USA

ABSTRACT

Grids represent an emerging technology that allows geographically- and organizationally-distributed resources (e.g., compute systems, data repositories, sensors, imaging systems, and so forth) to be linked in a fashion that is transparent to the user. A state-of-the-art grid presents a ubiquitous set of resources to a user. The New York State Grid (NYS Grid) is an integrated computational and data grid that provides access to users from around the world to a wide variety of resources. This Grid can be accessed via a Web portal, where the users have access to their data sets and applications, but do not need to be made aware of the details of the data storage or computational devices that are specifically employed in solving their problems. Grid-enabled versions of the SnB and BnP programs, which implement the Shake-and-Bake method of molecular structure (SnB) and substructure (BnP) determination, respectively, have been deployed on NYS Grid. Further, through the Grid Portal, SnB has been run simultaneously on all computational resources on NYS Grid as well as on more than 1100 of the over 3000 processors available through the Open Science Grid.

1. INTRODUCTION

The Grid is a rapidly emerging and expanding technology that allows geographically-distributed resources that extend across administrative boundaries to be linked together in a transparent fashion (www.gridcomputing.com/; www.globus.org/; Berman *et al*., 2003; Foster & Kesselmann,

1999). These resources include compute systems, data storage devices, sensors, imaging systems, visualization devices, and a wide variety of Internet-ready instruments. The concept and terminology of the Grid is borrowed from the electrical grid, where utility companies have the ability to share and move resources (electricity) in a fashion that is transparent to the consumer. With rare exception, the view taken by a consumer is that they are able to plug a piece of equipment into a power outlet in order to obtain electricity and do not need to know, and in fact, do not want to know, the details pertaining to the manner in which electricity makes its way to the outlet to provide power to their device. Similarly, the power of both computational grids (*i.e.*, seamlessly connecting compute systems and their local storage) and data grids (*i.e.*, seamlessly connecting large storage systems) lies not only in the aggregate computing power, data storage, and network bandwidth that can readily be brought to bear on a particular problem, but also on its ease of use.

Numerous government-sponsored reports state that grid computing is a key to 21st century discovery by providing seamless access to the high-end computational infrastructure that is required for revolutionary advances in contemporary science and engineering. In fact, National Science Foundation Director Arden Bement stated that "*leadership in cyberinfrastructure may determine America's continued ability to innovate – and thus our ability to compete successfully in the global arena.*"

Grids are now a viable solution to certain computationally- and data-intensive computing problems for reasons that include the following.

- Users can access many grids through a Web portal from virtually anywhere in the world. That is, a user only needs an account on a Grid administrative server in order to use a grid. This is similar to how someone uses a search engine or large e-business system. One only needs access to a gateway and not specifically to each individual server that the company has configured to be able to handle the requests/queries/business. For most grids, a user needs access to a Grid Portal, but does not need to be logged in to a site that hosts some Grid resource, does not need to be logged in to a computer that is on a Grid, and does not need to install any additional software on their Web-accessible system (workstation, cellular phone, laptop, etc.) in order to be able to use a Grid.
- The Internet is mature and able to serve as the fundamental infrastructure for network-based computing. In fact, network bandwidth, which has doubled approximately every 12 months over the past couple of decades, has increased to the point of being able to provide efficient and reliable services for the vast majority of Grid applications.
- Storage capacity, which has been doubling approximately every 9 months during the past decade, has now reached commodity levels, where one can purchase a terabyte of disk for roughly the same price as a high-end PC.
- Many instruments are Internet-aware.
- Clusters, supercomputers, storage and visualization devices are becoming more mainstream in terms of their ability to host scientific applications.

As grid computing initiatives move forward, issues of interoperability, security, performance, management, and privacy need to be carefully considered. In fact, security is concerned with various issues relating to authentication in order to insure application and data integrity. Grid initiatives are also generating best practice scheduling and resource management documents, protocols, and API specifications to enable interoperability. Several layers of security, data encryption, and certificate authorities already exist in grid-enabling toolkits such as Globus Toolkit (www.globus.org/toolkit/).

In this chapter, we present the *New York State Grid* (*NYS Grid*), which represents a major effort by our research group. The design, implementation, and support of this Grid has provided a wide variety of opportunities for institutions in New York State in terms of science, engineering, scholarship, education, outreach, and training. Indeed, the New York State Grid consists of a heterogeneous set of resources from institutions throughout New York State. These managed resources are available in a simple and seamless fashion to users worldwide. One of the major features of our grid is that it integrates a computational grid with a data grid so that a user may deploy computationally-intensive applications that read or write large data files in a very simple fashion. In particular, we designed this grid so that users do not need to know where data files are physically stored or where an application is physically deployed. In this chapter, we will discus the use of an active Web portal for deploying applications, dynamic resource allocation so that clusters and networks of workstations can be scheduled to provide resources on demand, and a scalable and dynamic scheduling system, to name a few.

In Section 2, we present background material as it pertains to both cyberinfrastructure, in general, and grid computing, in particular. Further, in section 2, we introduce and give an overview of the New York State Grid and a key application in molecular structure determination that we have ported to this Grid. In Section 3, we discuss details of the web interface and the underlying technology that allows for the *Shake-and-Bake* method of molecular structure determination to be ported to the New York State Grid. We focus on the Web Portal and its ease of use in manipulating key parameters to the *Shake-and-Bake* routine. In Section 4, we discuss future trends in molecular structure determination and how it dovetails with the maturing of cyberinfrastructure. Section 5 serves as the conclusion.

2. BACKGROUND

2.1 Cyberinfrastructure and Grid Computing

Cyberinfrastructure sits at the core of modern simulation and modelling, providing entirely new methods of investigation to scientists, engineers, and scholars in their attempt to address previously unsolvable problems or investigate problems in previously unexplored domains. In fact, new insights that are derived from advances in cyberinfrastructure might well lead to new "grand challenges" as well as providing solutions of previously stated "grand challenge problems." Specifically, the development of necessary software, algorithms, portals, and interfaces will enable research and scholarship by freeing end-users from dealing with the complexity of various computing environments. This is critical to extending the reach of high-end computing, storage, networking, and visualization to the general user community. In fact, globally dispersed and interconnected resources, coupled with advances in middleware and computational methodologies, may well enable new ways of thinking about research and enhance traditional means of discovery to advance knowledge. It might also provide a means for a non-professional scientist or scholar, the so-called "citizen-scientist" to drive scientific discovery. (CDI, 2007)

While cyberinfrastructure, or e-science as it is sometimes called, has existed for decades, the term "cyberinfrastructure" became commonplace when introduced in the "Atkins Report" (Atkins, et. al., 2003), which was produced by a National Science Foundation blue-ribbon panel. This report, which radically changed the vision and funding directions at NSF, recommended that existing barriers to the rapid evolution of high performance be removed so that such infrastructure would be easy to use by scientists, engineers, scholars, and non-professional citizen-scientists. The term cyberinfrastructure was introduced to

describe a research environment that supports advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services through a network of geographically distributed resources, including people, compute systems, storage devices, visualization systems, sensors, and other Internet ready devices. More specifically, the goal of cyberinfrastructure was to advance the technology so that geographically-distributed high-end systems could be used in a seamless fashion and integrated in a ubiquitous fashion into scientific discovery and innovation.

While standard definitions for both "cyberinfrastructure" and "grid" do not exist, Yafchak & Trauner (2008) propose a simple definition that states that "a grid consists of shared heterogeneous computing and data resources networked across administrative boundaries." While we would like to see other Internet-ready instruments included in such a definition, it should be noted that a grid can refer to a set of workstations distributed throughout a company, where not all workstations are under control of the same IT manager, or it can refer to a set of supercomputers located around the world in different organizations, or some combinations thereof, including other Internet-ready devices such as data storage devices, visualization engines, and sensors.

As stated in the Introduction, grids are starting to move out of the research laboratory and into early-adopter production systems. Numerous grid projects have been initiated, many of which are discussed in Yafchak & **Trauner (2008)** [*e.g.,* GriPhyN (www.griphyn.org/), PPDG (www.ppdg.net/), EGEE (www.eu-egee.org/), EU DataGrid (eu-datagrid.web.cern.ch/eu-datagrid/), NASA's Information Power Grid (IPG) (www.gloriad.org/gloriad/projects/project000053.html), TeraGrid (www.teragrid.org/), Open Science Grid (www.opensciencegrid.org/), and caBIG (https://cabig.nci.nih.gov/), to name a few]. However, the construction of a real general-purpose grid is in its infancy since a true grid requires coordinated resource sharing and problem solving in a dynamic, multi-institutional scenario using standard, open, general-purpose protocols and interfaces that deliver a high quality of service. The immediate focus of grid deployment continues to be on the difficult issue of developing high quality middleware (www.nsf-middleware.org/).

2.2 New York State Grid

In order to support research, scholarship, education, and community outreach, Miller's Cyberinfrastructure Laboratory (MCIL; www.cse.buffalo.edu/faculty/miller/CI) at SUNY-Buffalo and the Hauptman-Woodward Institute focuses on the integration of research in disciplinary domains with research in enabling technologies and interfaces. The goal of MCIL is to provide systems that allow users to transparently collect, manage, organize, analyze, and visualize data without having to worry about details such as where the data is physically stored, processed, rendered, and so forth. This ease of use and high availability of data and information processing tools is expected to stimulate revolutions in science, engineering, and beyond.

MCIL provided the design, development, and deployment of the *New York State Grid* (*NYS Grid*), which includes resources from institutions throughout New York State and is available in a simple and seamless fashion to users worldwide. NYS Grid contains a heterogeneous set of resources and utilizes general-purpose IP networks (Green & Miller, 2003, 2004a-c). A major feature of NYS Grid is that it integrates a computational grid with a data grid so that a user may deploy computationally-intensive applications that read or write large data files in a very simple fashion. In particular, NYS Grid was designed so that a user does not need to know where data files are physically stored or where an application is physically deployed, while providing the user with easy access to their files (uploading, downloading,

editing, viewing, etc.). The core infrastructure for NYS Grid includes the installation of standard grid middleware, the use of an active Web portal for deploying applications, dynamic resource allocation so that clusters and networks of workstations can be scheduled to provide resources on demand, and a scalable and dynamic scheduling system, to name a few.

Several key packages were used in the implementation of NYS Grid and other packages have been identified in order to allow for the anticipated expansion of the system. The Globus Toolkit (www.globus.org/toolkit/) provides APIs and tools using the Java SDK to simplify the development of OGSI-compliant services and clients. It supplies database services and Monitoring & Discovery System index services implemented in Java (www.globus.org/toolkit/mds/), GRAM service implemented in C with a Java wrapper (www.globus.org/toolkit/docs/2.4/gram/), GridFTP services implemented in C (www.globus.org/grid_software/data/gridftp.php), and a full set of Globus Toolkit components. The recently proposed Web Service-Resource Framework provides the concepts and interfaces developed by the OGSI specification exploiting the Web services architecture (www.globus.org/wsrf/).

NYS Grid is the current incarnation of an MCIL-led grid that progressed from a Buffalo-based grid (ACDC-Grid) to a persistent, hardened, and heterogeneous Western New York Grid (WNY Grid), before being enhanced, expanded, and deployed throughout New York State. This series of Grids was largely funded by the National Science Foundation through ITR, MRI, and CRI grants. NYS Grid currently supports a variety of applications and users from NYS Grid institutions and the Open Science Grid. In addition, a grass-roots cyberinfrastructure initiative in New York State has been granted access to NYS Grid. NYS Grid also serves as a gateway to the Open Science Grid, TeraGrid, MCEER (mceer.buffalo.edu/), and NEES (it.nees.org/ and nees.buffalo.edu/), to name a few.

2.3 Molecular Structure Determination by Direct Methods

No microscope is powerful enough to let us observe biological molecules and their interactions with each other. Instead, structural biologists use tools provided by the science of X-ray crystallography. In a crystallographic experiment, a single crystal of a purified substance such as a protein is irradiated, and the incident radiation is scattered in many directions to produce a diffraction pattern. A complex mathematical analysis of the diffraction data (a process known as "solving" the structure) is then undertaken to determine the shape and atomic arrangement of the molecules comprising the crystal. Once the structure has been solved, molecular models can be constructed and examined for insight into how the protein molecules function, what might be happening when disease occurs, and what compounds might be designed as drugs to modify activity.

The pathway from diffraction data to atomic coordinates is not straightforward. Atomic coordinates can be computed readily from complex-valued quantities known as structure factors, but only the magnitudes of the structure factors can be measured in the diffraction experiment. The phases of the structure factors are lost, and this gives rise to the so-called "phase problem" of X-ray crystallography. However, in most cases, mathematical techniques known as direct methods provide a way to solve the phase problem. Direct methods use probabilistic relationships among the phases to derive the values of individual phases from the measured amplitudes. Conventional direct methods, implemented in computer programs prior to 1992 (*e.g.*, MULTAN (Main *et al*., 1980), SHELXS (Sheldrick, 1990), SAYTAN (Debaerdemaeker *et al.*, 1985) and SIR (Burla *et al.*, 1989)), provided computationally efficient solutions for structures containing less than approximately 100 unique non-hydrogen atoms. However, few larger structures (*i.e.*, structures with more than 200 unique equal atoms) have ever been solved

using these programs. The computer-intensive direct methods *Shake-and-Bake* approach to molecular structure determination, developed by the authors (Weeks *et al*., 1993; Miller *et al*., 1993) and described in detail in the next section, provided the breakthrough needed to achieve automated direct-methods solutions for much larger structures.

2.4 *Shake-and-Bake*

Shake-and-Bake is a powerful algorithmic formulation of direct methods that, given accurate diffraction data to 1.2Å or better resolution, has made possible the *ab initio* phasing of complete crystal structures containing as many as ~2000 independent non-H atoms (Frazão, *et al*., 1999). It has also been used to determine the anomalously scattering substructures of selenomethionyl-substituted proteins containing as many as 160 selenium sites using 3-4Å data (Von Delft *et al*., 2003). *Shake-and-Bake* belongs to the class of phasing methods known as 'multisolution' procedures (Germain & Woolfson, 1968) in which multiple sets of trial phases are generated in the hope that one or more of the resultant combinations will lead to a solution. Solutions, if they occur, are identified on the basis of the value of a suitable figure of merit, such as the minimal function (see below) or the crystallographic *R* value. Since each of the sets of trial phases can be processed independently, the *Shake-and-Bake* algorithm can be easily adapted to a coarse-grained parallel processing approach and implemented on a computational grid.

The distinctive feature of *Shake-and-Bake* is the repeated and unconditional cyclical alternation of reciprocal-space phase refinement with a complementary real-space process that seeks to improve phases by imposing constraints through a physically meaningful interpretation of the electron density (Miller, *et al*., 1993; Weeks, *et al*., 1994). First, a random number generator is used to assign initial coordinates to the atoms comprising the trial structures, and structure-factor calculations are performed to generate the corresponding sets of starting trial phases. Then, phases are refined either by the tangent formula (Karle & Hauptman, 1956) or by constrained minimization of the so-called minimal function (DeTitta, *et al*., 1994) using the parameter-shift algorithm (Bhuiya & Stanley, 1963). Following Fourier transformation to real space and computation of an electron-density map, peak picking is used to impose the atomicity constraint. Peaks are located on the map, an appropriate number of the largest of these maxima are assumed to be atoms, and the cycle is completed by using inverse Fourier transformation (in the form of a structure-factor calculation) to generate phases for another round of refinement. The entire process is repeated for a predetermined (by the user) number of cycles. The success rate of this process (*i.e*., the percentage of trial structures that converge to solution) depends on data quality and the size of the structure.

As mentioned, the *Shake-and-Bake* procedure has been implemented in a computer program, *SnB*, in a manner convenient for both protein substructures and complete structures (Miller *et al*., 1994; Weeks & Miller, 1999; Rappleye *et al*., 2002). (The *Shake-and-Bake* algorithm has also been implemented independently in the program *SHELXD* (Schneider & Sheldrick, 2002)). The *SnB* graphical user interface (GUI), written in Java, controls not only the main phasing program but also the *DREAR* program suite (Blessing & Smith, 1999) that computes the normalized structure-factor magnitudes ($|E|$) required for direct-methods calculations. In addition, the two-step process of substructure determination and protein phasing has been combined in the program *BnP* (Weeks *et al*., 2002), which provides a common interface for *SnB* and components of the *PHASES* suite (Furey & Swaminathan, 1997). Thus, *BnP* provides an automated pathway from processed intensities to an unambiguous protein electron-density map. This pathway includes

SnB substructure determination, heavy-atom site validation, enantiomorph determination, substructure and protein phase refinement, and solvent flattening.

The repetitive shuttling of trial structures between real and reciprocal space gives the *Shake-and-Bake* algorithm its power, but the need to perform two Fourier transformations in each cycle yields a computationally-intensive procedure. In fact, the running time for *SnB* or *BnP* varies widely - from just a few seconds to many hours for large structures or structures with diffraction data of marginal quality that typically require a large number of trial structures to be examined before a solution is found. In such cases, the ability to increase throughput by processing many trial structures simultaneously on a cluster or a computational grid is invaluable.

3. STRUCTURE DETERMINATION ON THE GRID

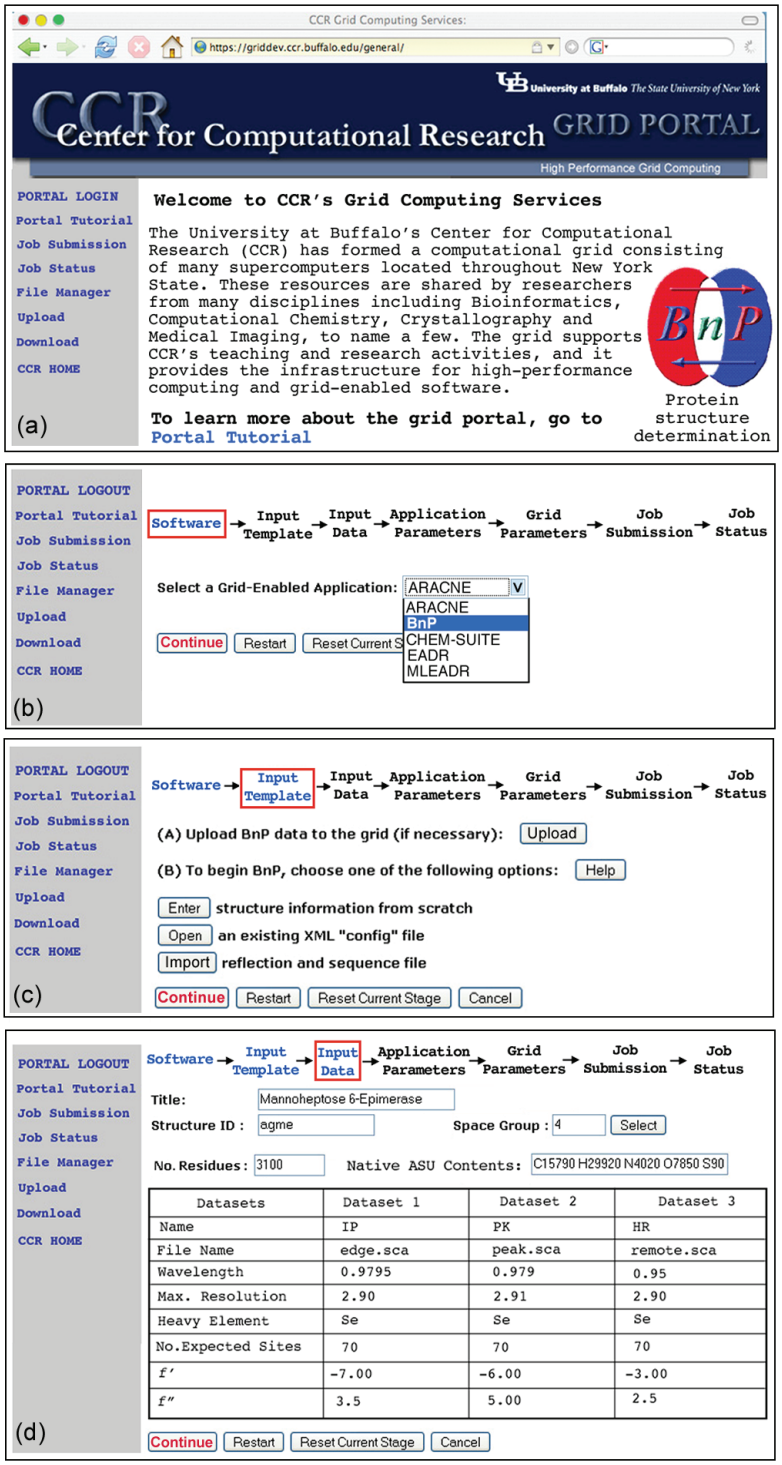
NYS Grid may be accessed via a browser through a web portal, as shown in Fig. 1a, or through a standard command-line submission. Versions of the *SnB* and *BnP* programs capable of parallel processing have been equipped with web-compatible PHP interfaces and incorporated into the web portal for ease of use using Grid-enabling Application Templates (Green & Miller, 2004c). The traditional stand alone Java versions and the new grid-enabled PHP versions are compatible in the sense that files created in both versions are interchangeable and can be uploaded to and downloaded from the grid as needed. As mentioned previously, *SnB* has been run on NYS Grid and Open Science Grid through the Web Portal, as well as through command-line submission, to compute hundreds of thousands of trial structures using an aggregate of nearly 100,000 CPUs.

BnP combines the *Shake-and-Bake* substructure determination step, which can easily take advantage of a parallel computing environment, with protein phasing calculations that are applied to a successful *Shake-and-Bake* substructure trial. The approach that has been taken is to spread the processing of trial substructures among a large number of computational nodes. In most cases, it is possible for *BnP* to determine when a substructure solution has occurred and to move on to the next step. Jobs that complete the first step without identifying a solution can (optionally) be terminated. Since jobs on different machines are independent once they have been spawned, multiple solutions are possible - an advantage if any jobs produce false solutions or solutions of marginal quality. It was decided that this execution scenario maximized the probability that at least one solution would be found for a difficult structure in a modest amount of time.

Figure 1 illustrates the use of the grid-enabled version of *BnP* to phase the protein mannoheptose 6-epimerase (Deacon *et al*., 2000) after first solving the 70-Se substructure. Fig. 1a shows the Grid portal as it initially appears to a user. After login, the user is presented with screens representing the various stages of the GAT, as indicated by the workflow defined at the top of the screen. In each case, the current stage (1b-1g) is indicated by the red rectangle. To move to the next stage, the user must click the "Continue" button at the bottom of the screen. Several "Help" buttons are available along the way.

First, the *BnP* software is selected (Fig. 1b). At the next stage (Fig. 1c), the user is reminded to upload input data files to the grid (not illustrated) and then instructed to choose how additional required information is to be supplied. If "Enter" (or "Continue") is selected, a blank version of screen 1d will appear, and the information requested about the structure (*e.g.*, space group) and its datasets must be entered manually. If "Import" is chosen, most or all of the information will be extracted from the headers of the input data files, and "Open" will restore information saved from a previous job.

Figure 1. Job submission, monitoring, and examination of results using the grid portal. (a) Login to the SUNY-Buffalo grid portal. (b) Select BnP as the software application to be used. (c) Select execution options and upload files from the home computer. (d) Input additional information about the structure.



The next step is to supply the values for the parameters required to execute the *BnP* application (Fig. 1e). Experience has shown that it is sometimes important to vary the values of certain key parameters when performing a *Shake-and-Bake* calculation in order to ensure that a solution will be found. These parameters include the space group (specified by giving its number in *International Tables*), the maximum resolution of the diffraction data to be used for the substructure determination, the maximum number of expected heavy-atom or anomalously scattering sites, and the type of normalized difference data to be used for *Shake-and-Bake* (*e.g.*, peak wavelength anomalous differences or maximum dispersive differences). The interface computes the number of different parameter combinations that have been specified, and the user must then indicate the number of different substructure trials that are to be processed on the grid for each set of parameter values. The final choice the user has to make at this stage is to decide if the default values of other *BnP* parameters (*e.g.*, the numbers of phases, peaks, and refinement cycles) - that are to be fixed for all jobs - should be changed (option not illustrated).

Parameters such as the computational resource (*e.g.*, a particular computer cluster) on which the jobs are to be run and the maximum time allowed per job are specified at the "Grid Parameters" stage (Fig. 1f). The user can also specify if jobs are to be terminated after the substructure determination (*Shake-and-Bake*) step if no solution has been detected automatically based on the figure-of-merit calculation. Otherwise, the job will continue using the trial substructure with the best figure of merit even though the program does not regard it as a solution. Making the decision to terminate frees resources for other users, but there is a small risk that an actual solution might be missed.

The next screen (Fig. 1g) permits the user to review all parameter choices before the job(s) are submitted. The progress of the jobs can then be monitored from the "Job Status" screen (Fig 1h). A drilldown button (shown in red) provides access to more information about an individual job or multiple job set, and Fig. 1i shows details about the status of the multiple job (number 44267) set up in step 1e and selected for viewing in step 1h. The values of the variable parameters for each job are given in the table, which indicates that the jobs were grouped in pairs (differing by the random number seed). In the case illustrated, all jobs that used the peak anomalous difference data gave recognizable solutions and proceeded to perform all *BnP* tasks up to and including protein phasing and solvent flattening (note indications that stage 5 was completed). The other jobs did not find solutions and were killed (terminated) after the *Shake-and-Bake* stage (#2) had processed all 500 trial structures assigned to them. If any jobs had failed because of *BnP* errors or for other reasons, that fact would have been indicated in the status column.

Drilling down one level further (indicated here by the red symbol on the line for job 1.1) provides access to the results for a single job (Fig. 1j). Those results include, for example, a *Shake-and-Bake* histogram (Fig. 1k) of minimal function values for the 89 trials that were processed in job 1.1 before the program automatically detected that a probable solution had been found. Finally, clicking the "Download results" button (on Fig. 1j) causes the screen shown in Fig. 1l to be displayed with options for downloading the files resulting from the *BnP* calculations to the home computer.

4. FUTURE TRENDS

The future of molecular structure determination lies in an automated workflow that will progress from crystal growth through the determination of a refined structure. In the near term, such an effort will rely on a workflow where the instrument used to measure diffraction data, data storage devices,

Figure 1 (continued). (e) Assign values to program parameters that will be varied in different jobs. Some additional parameters (constant for all jobs submitted simultaneously) are accessible from a pop-up window. (f) Supply information controlling job execution on the grid. (g) Review all parameters and start the jobs. (h) Check whether the jobs are still running.

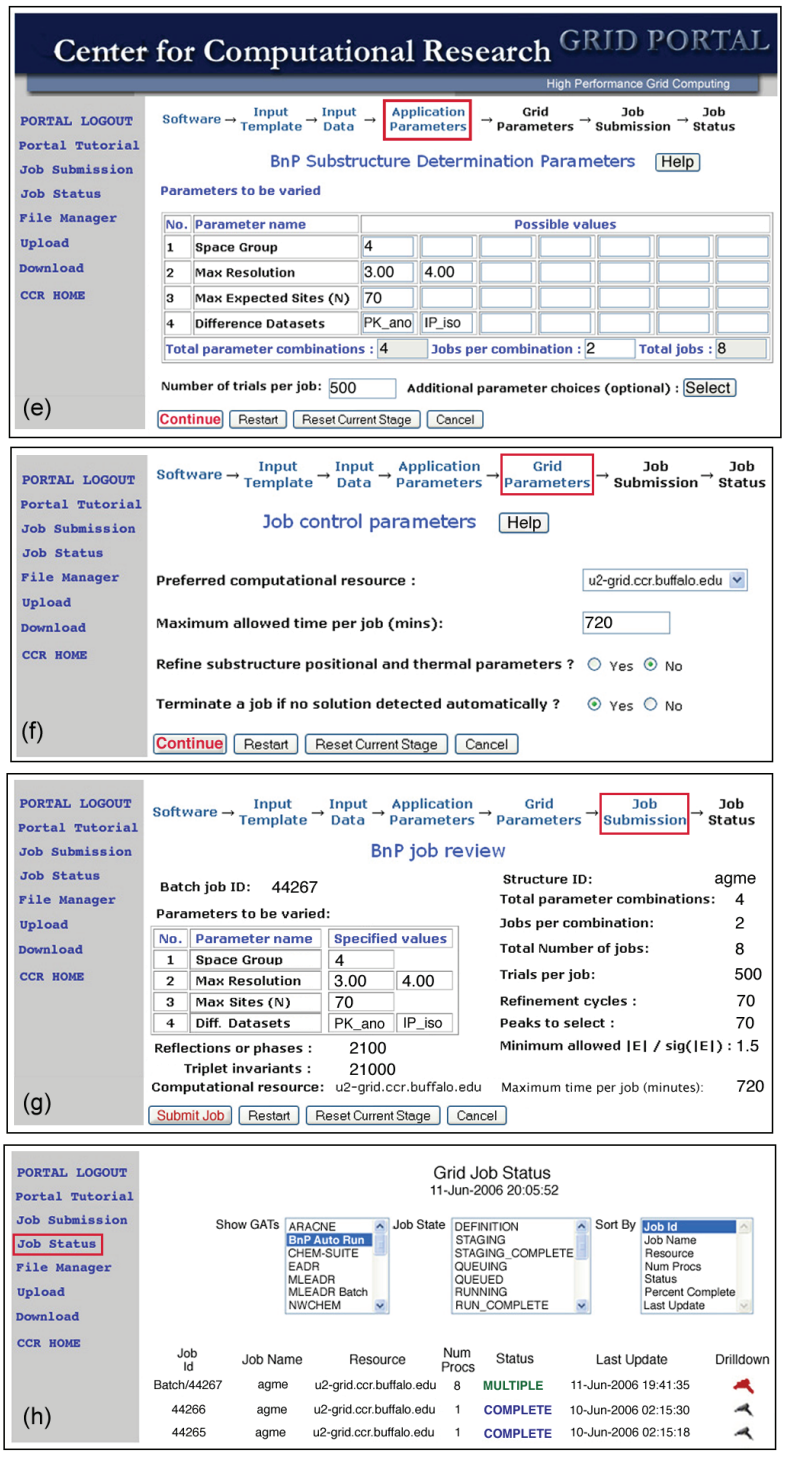
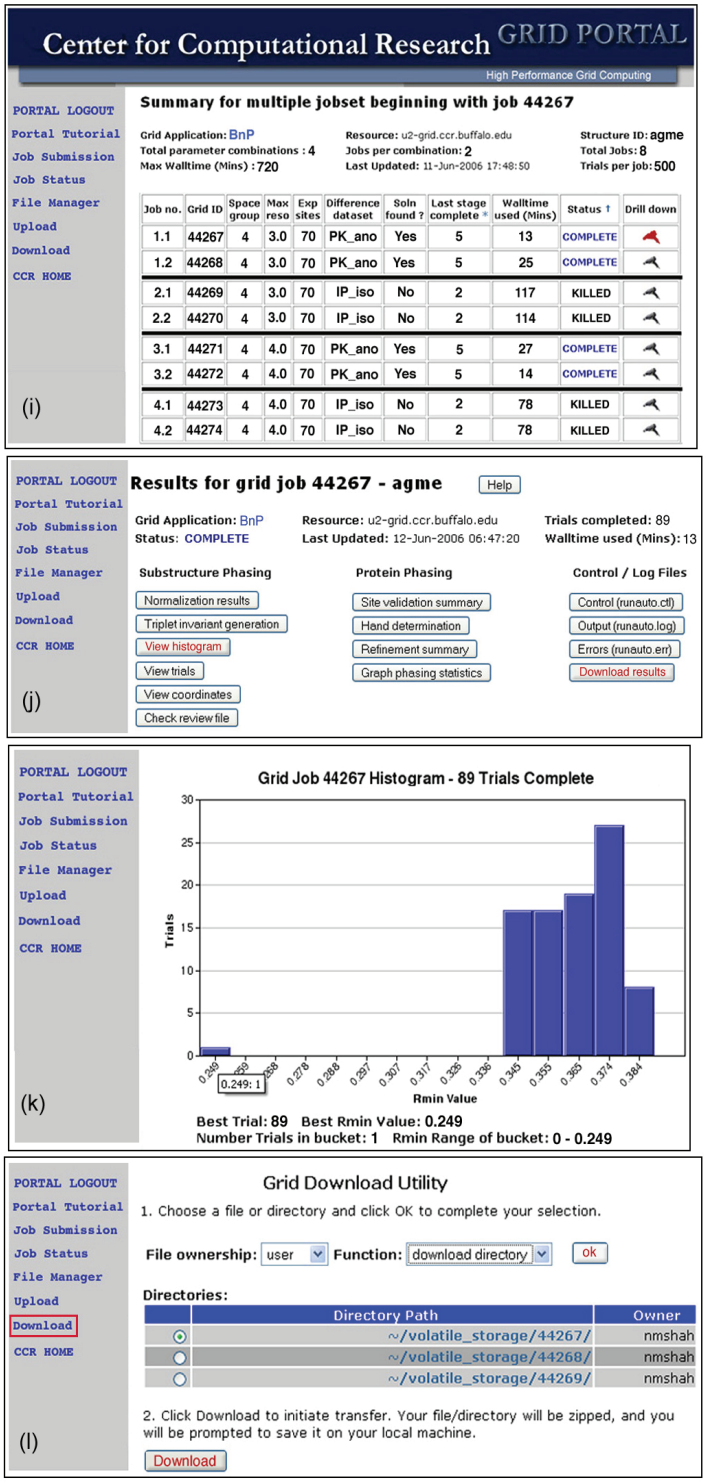


Figure 1 (continued). (i) Drilldown to check the execution status of individual components of a multiple job set. (j) Drilldown to inspect the results of an individual job. (k) Figure-of-merit histogram for the selected job. (l) Download the output files to the home computer.



compute systems, collaborative visualization systems, and final data repositories are all on a grid and are able to be harnessed into a solution strategy. Currently, many of the individual pieces exist. However, the design, development, and deployment of an entire workflow will require a significant effort. In particular, many imaging systems are on the Internet. The Protein Data Bank (PDB; http://www.wwpdb.org/, Berman, Henrick, & Nakamura, 2003), which serves as a repository for 3-D structural data of proteins and nucleic acids, typically determined through X-ray crystallography or NMR spectroscopy methodologies, is freely available via the Internet and is also available through teragrid. The *SnB* and *BnP* programs have been ported to several grids and a wide variety of compute platforms. Some of the post-processing and refinement routines are Internet accessible. Collaborative visualization facilities for molecular structures have been developed in Buffalo. So, a project similar to NIH's cancer Biomedical Informatics Grid, caBIG (https://cabig.nci.nih.gov/), which is dedicated to accelerating research discoveries and improving patient outcome by linking researchers, patients, and physicians throughout the cancer community, should be sufficient to put such a black-box system for molecular structure determination in place. Note that the molecular structure determination effort would be a much smaller effort in terms of funding, equipment, and personnel than the caBIG project.

5. CONCLUSION

NYS Grid is a state-of-the-art, integrated computational and data grid. Using Grid-enabling Application Templates, the *SnB* and *BnP* programs have been adapted to this grid in a straightforward fashion. The web-based Grid portal provides convenient access to these programs for users of the grid. Users do not need to know details of where their files are maintained, nor where the computations are performed, although details are always available for those who wish to be informed. The grid-enabled versions allow for a much shorter wall time to solution for large and difficult structures.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Nos. CNS-0454114, CCF-0204918, EIA-0101244, ANI-0124929, ACI-9721373, and NIH Grant No. EB002057. The authors would like to thank members of Dr. Miller's Cyberinfrastructure Laboratory (including Steve Gallo, Jason Rappleye, Cathy Ruby, Jon Bednasz, Tony Kew, Sam Guercio, Adam Koniak, Martins Innus, Dori Macchioni, Amin Ghadersohi, and Cynthia Cornelius) for their contributions to the efforts described in this paper. Much of this research was performed while Dr. Miller was Director of the Center for Computational Research at SUNY-Buffalo. The authors would also like to thank Herb Hauptman, George DeTitta, and Billy Furey for their contributions to this work.

REFERENCES

Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H. Klein, M.L., Messerschmitt, D.G., et al. (n.d.). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report to the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* Retrieved on January 15, 2008, from http://www.communitytechnology.org/nsf_ci_report/report.pdf.

Berman, F., Hey, A. J. G & Fox, G. C. (Eds.), (2003). *Grid Computing: Making the Global Infrastructure a Reality.* New York: John Wiley.

Berman, H.M., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data

Bank. *Nature Structural Biology, 10* (12), 980.

Bhuiya, A. K. & Stanley, E. (1963). The refinement of atomic parameters by direct calculation of the minimum residual, *Acta Crystallogr, 16*, 981-984.

Blessing, R. H. & Smith, G. D. (1999). Difference Structure Factor Normalization for Determining Heavy-Atom or Anomalous Scattering Substructures, *J. Appl. Cryst. 32*, 664-670.

Burla, M.C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). SIR88 - a direct-methods program for the automatic solution of crystal structures. *J. Appl. Cryst. 22*, 389-393.

CDI (2007). *Cyber-Enabled Discovery and Innovation (CDI)*, *Computer and Information Science and Engineering Directorate, National Science Foundation*. Paper presented at the National Science Foundation.

Chang, C.-S., DeTitta, G., Hauptman, H., Jones, R., Miller, R., Thuman, P., & Weeks, C. (1993). Solving the phase problem of x-ray crystallography on parallel machines. In R.F. Sincovec, D.E. Keyes, M.R. Leuze, L.R. Petzold, and D.A. Reed, (Eds.) *Proceedings of The Sixth SIAM Conference on Parallel Processing for Scientific Computing*, (pp. 304-307).

Chang, C.-S., DeTitta, G.T., Miller, R. & Weeks, C.M. (1994). On the application of parallel genetic algorithms in x-ray crystallography. *Proceedings of the 1994 Scalable High Performance Computing Conference*, (pp. 796-802). New York: IEEE Computer Society Press.

Deacon, A. M., Ni, Y. S., Coleman, W. G. Jr. & Ealick, S. E. (2000). The crystal structure of ADP-L-glycero-D-mannoheptose 6-epimerase: catalysis with a twist. *Structure*, *8*, 453-462.

Debaerdemaeker, T., Tate, C. & Woolfson, M.M. (1985). On the application of phase relationships to complex structures (XXIV: The Sayre tangent formula). *Acta Cryst.*A41), 286-290.

DeTitta, G., Hauptman, H., Miller, R., Pagels, M. Sabin, T., Thuman, P., & Weeks, C. (1991). Parallel solutions to the phase problem in x-ray crystallography: an update. In *Proceedings of The Sixth Distributed Memory Computing Conference*, (pp. 587-594). New York: IEEE Computer Society Press.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). Structure solution by minimal function phase refinement and Fourier filtering, (I. Theoretical basis). *Acta Crystallogr.* A50, 203-210.

Foster, I. & Kesselmann, C. (Eds.) (1999). *The Grid: Blueprint for a New Computing Intrastructure*. San Francisco, CA: Morgan Kaufmann Publishers.

Frazão, C., Sieker, L., Sheldrick, G. M., Lamzin, V., LeGall, J. & Carrondo, M. A. (1999). Ab initio structure solution of a dimeric cytochrome c3 from Desulfovibrio gigas containing disulfide bridges. *J. Biol. Inorg. Chem. 4*, 162-165.

Furey, W. & Swaminathan, S. (1997). PHASES-95: A Program Package for the Processing and Analysis of Diffraction Data from Macromolecules. *Meth. Enzymol. 277*, 590-620.

Germain, G. & Woolfson, M. M. (1968). On the application of phase relationships to complex structures, *Acta Crystallogr. B24*, 91-96.

Green, M. L. & Miller, R. (2003). Grid computing in Buffalo. *Annals of the European Academy of Sciences*, 191-218.

Green, M. L. & Miller, R. (2004a). Molecular structure determination on a computational & data grid. *Parallel Computing*, *30*, 1001-1017.

Green, M. L. & Miller, R. (2004b). Evolutionary molecular structure determination using grid-enabled data mining. *Parallel Computing*, *30*, 1057-1071.

Green, M. L. & Miller, R. (2004c). A client-server prototype for application grid-enabling template design. *Parallel Processing Letters*, *14*, 241-253.

Karle, J. & Hauptman, H. A. (1956). A theory of phase determination for the four types of non-centrosymmetric space groups 1*P*222, 2*P*22, 3$P_1$2, 3$P_2$2. *Acta Crystallogr.*, *9*, 635-651.

Main, P., Fiske, S.J., Hull, S.E., Lessinger, L., Germain, G., Declercq, J.P. & Woolfson, M.M. (1980). *MULTAN80: a system of computer programs for the automatic solution of crystal structures from x-ray diffraction data* (Universities of York and Louvain).

Miller, R., Bednasz, J.J., Chiu, K., Gallo, S.M., Govindaraju, M., Lewis, M., Ruby, C., & Weeks, C.M. (2008). Grid-based research, development, and deployment in New York State, In *Proceedings of the International Parallel and Distributed Processing Symposium,* in press.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). On the application of the minimal principle to solve unknown structures, *Science,* 259, 1430–1433.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *SnB*: Crystal structure determination via *Shake-and-Bake. J. Appl. Cryst.. 27*, 613–621.

Rappleye, J., Innus, M., Weeks, C. M. & Miller, R. (2002). *SnB* v2.2: An Example of Crystallographic Multiprocessing. *J. Appl. Cryst., 35*, 374-376.

Schneider, T. R. & Sheldrick, G. M. (2002). Substructure solution with *SHELXD*. *Acta Crystallogr., D58*, 1772-1779.

Sheldrick, G.M. (1990). Phase annealing in SHELX-90: direct methods for larger structures. *Acta Cryst*. *A46*, 467–473.

Von Delft, F., Inoue, T., Saldanha, S. A., Ottenhof, H. H., Schmitzberger, F., Birch, L. M., et al. (2003). Structure of E. coli Ketopantoate Hydroxymethyl Transferase Complexed with Ketopantoate and Mg2+, Solved by Locating 160 Selenomethionine Sites. *Structure*, *11*, 985-996.

Weeks, C. M., Blessing, R. H., Miller, R., Mungee, R., Potter, S. A., Rappleye, J., Smith, G. D., Xu, H. & Furey, W. (2002). Towards automated protein structure determination: *BnP*, the *SnB*-PHASES interface. *Z. Kristallogr., 217*, 686-693.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). Structure solution by minimal function phase refinement and Fourier filtering (II. Implementation and applications). *Acta Crystallogr.*, *A50*, 210-220.

Weeks, C.M., DeTitta, G.T., Miller, R. & Hauptman, H.A. (1993). Applications of the minimal principle to peptide structures. *Acta Cryst., D49*, 179-181.

Weeks, C. M. & Miller, R. (1999). The design and implementation of *SnB* v2.0. *J. Appl. Cryst., 32*, 120-124.

Yafchak, M.F. & Trauner, M., (Eds.) (2008). *The Grid Technology Cookbook*. Retrieved January 15, 2008, from http://www.sura.org/cookbook/gtcb/index.php?topic=12&mlevel=1&parent=0

KEY TERMS AND DEFINITIONS

Cyberinfrastructure: Provides for the transparent and ubiquitous application of technologies central to contemporary science and engineering, including high-end computing, networking, and visualization, data warehouses, science gateways, and virtual organizations, to name a few. It is a comprehensive phenomenon that involves creation, dissemination, preservation, and application of knowledge.

Diffraction: In very simple terms, diffraction is the bending or spreading of waves as they

pass through an obstruction or gap. The gaps or distances between molecules in a crystal are such that X-rays (with a wavelength of ~10^6 mm) are diffracted by crystals. The X-rays scattered in a diffraction experiment produce a distinctive pattern that is related to the atomic arrangement within the crystal that was irradiated. Diffraction patterns can be recorded on photographic film or a suitable electronic recording device.

Direct Methods: A mathematical approach that makes it possible to glean phase information from the diffraction magnitudes. For example, the fact that molecules consist of atoms, and that atoms are small, discrete points relative to the spaces between them, creates certain constraints. Since there are many more reflections in a diffraction pattern than there are independent atoms in the corresponding crystal, the phase problem is overdetermined, and the existence of relationships among the measured magnitudes is implied. Certain linear combinations of three phases have been identified as relationships useful for determining unknown structures, and direct methods use probabilistic techniques to exploit these relationships. Herbert Hauptman and Jerome Karle won a Nobel prize in 1985 for their work in developing direct methods.

Grid Computing: Computational efforts involving computing, networking, storage, or visualization that involves geographically-distributed and independently-operated resources that are linked together in a transparent fashion.

GUI: A graphical user interface allows people to interact with a computer and computer-controlled devices. Instead of offering only text menus or requiring typed commands, graphical icons, visual indicators or special graphical elements are presented. Often the icons are used in conjunction with text, labels or text navigation to fully represent the information and actions available to a user. The actions are usually performed through direct manipulation of the graphical elements.

Molecular Structure Determination: The study of the three-dimensional architecture of molecules and the arrangement of their component atoms in space.

The Phase Problem: A diffraction pattern consists of a set of spots called *reflections* that result from capturing the image of an X-ray beam as it is scattered by the atoms in a crystal. Each reflection has a *magnitude*, which is experimentally accessible, and a *phase*, which is not. The inability to measure phase angles experimentally is known as the *phase problem*. Together, the magnitude and phase of a reflection constitute a quantity known as a *structure factor*. The set of structure factors provides the *reciprocal-space* representation of the structure, and the set of atomic coordinates provides the *real-space* representation. A Fourier transformation of the complex-valued structure factors leads directly to the real-valued electron density, which, after suitable interpretation, reveals the atomic positions and describes the molecular architecture of the crystalline material responsible for the scattering. Therefore, a solution to the phase problem requires an algorithm that will recover phase information that cannot be directly measured in the diffraction experiment.

Shake-and-Bake: A powerful algorithmic formulation of direct methods that, given complete and accurate diffraction data, has made possible the *ab initio* phasing of crystal structures containing as many as ~2000 independent non-hydrogen atoms. The distinctive feature of this algorithm is the cyclical alternation of phase refinement with the imposition of atomicity constraints.

SnB and BnP: *SnB* is a computer program developed in Buffalo (at the Hauptman-Woodward Medical Research Institute and SUNY-Buffalo) that provides an efficient implementation of the *Shake-and-Bake* method of molecular structure determination. This program has been distributed worldwide from the website http://www.hwi.buf-

falo.edu/SnB/ and is available for workstations, networks of workstations, clusters, and grids. *BnP* is a computer program that combines the direct-methods program *SnB* with components of Bill Furey's (University of Pittsburgh) PHASES suite. *BnP* targets the determination of large protein structures. First, the *SnB* component is used to find a substructure consisting of heavier atoms, and then the substructure is used as a starting point for phasing the complete protein.

X-Ray Crystallography: The scientific method most commonly used to determine molecular structures. Individual molecules cannot be seen under a light microscope because the wavelength of visible light is larger than the molecular size. However, crystals are made up of an array of many ($\sim 10^{11}$-10^{12}) identical, regularly-spaced molecules, and the regular spacing allows a technique called X-ray diffraction to be used to "see" the molecules that comprise the crystal.

APPENDIX

In this section, we provide a timeline and history with regard to MCIL, the ACDC-Grid, the WNY Grid, and NYS Grid. We also discuss the evolution and status of a grass-roots cyberinfrastructure initiative within New York State.

Timeline (MCIL and NYS Grid)[1]. In the early 1990s, in an effort to collaborate with scientists at the Hauptman-Woodward Institute in their effort to solve the phase problem of crystallography (Weeks, *et al.*, 1993), Miller's research group began to use commercial parallel computers, including shared- and distributed-memory machines from Thinking Machines Corporation, Encore, SGI, nCube, Intel, Alliant, and Sequent, to name a few (DeTitta, *et al.*, 1991; Chang, *et al.*, 1993; Chang, *et al.*, 1994). In addition, Miller's group used a laboratory of Sun workstations and ran their *Shake-and-Bake* method of structure determination using RPC (remote procedure call) to employ a master/worker solution of *Shake-and-Bake*. In 1998, with funding from the Department of Energy, Miller's group worked with Ian Foster's group at Argonne National Laboratory in order to use Globus to port *Shake-and-Bake* to early grids and clusters in an effort to solve larger molecular structures. In 1999, with funding from the National Science Foundation (NSF), Miller's group initiated a Buffalo-based grid research project that included the Hauptman-Woodward Institute, SUNY-Buffalo, and several Buffalo-area colleges, representing the genesis of a Western New York Grid.

In 2001, an NSF MRI grant funded a significant storage system that was incorporated into this existing grid, serving as a data repository for *Shake-and-Bake* results. This work led to an NSF ITR grant, funded in 2002, which focused on the deployment and efficient implementation of *Shake-and-Bake* on clusters and grids. In fact, this ITR grant also funded critical aspects of the design, development, deployment, and hardening of the aforementioned Buffalo-based grid (ACDC-Grid) and Western New York Grid (WNY Grid). In addition, these funds led to the expansion of WNY Grid to include institutions throughout New York State, which resulted in the establishment of the New York State Grid (NYS Grid) in 2004. The number of sites and variety of resources available on NYS Grid has grown substantially since 2004 and now includes a heterogeneous set of compute and storage systems throughout New York State. The institutions include academic and non-profit organizations, though NYS Grid is not restricted to such institutions. The NYS Grid has been used extensively by the *Shake-and-Bake* team and numerous other users from Buffalo and the Open Science Grid.

A virtual organization, called GRASE (Grid Resource for Advanced Science & Engineering), was established by Miller's CI Laboratory in the early 2000's in order to support general science and engineering users on Open Science Grid and NYS Grid.

Given the success of the Buffalo-based ACDC-Grid, the WNY Grid, and the establishment of the NYS Grid, the NSF provided CRI funds in order to provide significant resources to the core sites in Western New York (SUNY-Buffalo, Niagara University, Hauptman-Woodward, and SUNY-Geneseo). More details, including publications, presentations, and the current status of these grids and their associated projects, as described earlier in this paper, are available at www.cse.buffalo.edu/ faculty/miller/CI/.

NYS Grid, Miller's CI Laboratory, and the Grass-Roots NYS CI Initiative. Beginning in 2000, SUNY-Buffalo incorporated requests from Miller in terms of high-end computing in its annual requests for funding to the State of New York. Miller's group also responded to requests from elected officials in the State of New York for a vision of high-end computing to support science and engineering in the state. Initial requests focused on the establishment of a NYS Supercomputing Center. These requests morphed into a request to establish NYS Technology Park that would include a New York State High-End

Computing Center, including Computing, Storage, Networking, and Visualization for NYS. Beginning in 2004, the requests focused on establishing an all encompassing New York State Cyberinstitute that would include funding to enhance and expand the New York State Grid that MCIL had established. In 2006, the Senior Vice Provost of SUNY-Buffalo supported and made announcements confirming the establishment of a Cyberinstitute of the State of New York (CSNY).

Subsequently, in July of 2006, a group gathered in Ithaca, New York, to discuss the possibility of initiating a state-wide effort in cyberinfrastructure. At this meeting, entitled a "New York State Workshop on Data-Driven Science and Cyberinfrastructure," Miller presented CSNY, which by then was being established within SUNY-Buffalo's Bioinformatics Center (BCOEBLS). Per the announcement of CSNY by SUNY-Buffalo's Senior Vice Provost, and with his approval, it was disclosed that CSNY would include i) the Center for Computational Research, ii) faculty working in computational science and engineering, iii) faculty working on fundamental problems in cyberinfrastructure, as well as iv) enabling staff, including programmers, GUI designers, and personnel focused on integrating middleware with applications. Miller also presented an overview of the Center for Computational Research and on-going efforts in MCIL.

At the end of the meeting, the attendees decided to schedule a meeting to continue to explore possibilities within NYS. In addition, the membership asked Miller's permission (granted) for *NYS Grid to serve as the underlying grid architecture for this potential state-wide CI initiative* so that this new initiative could focus on higher-level, non-grid, issues. Following the meeting, MCIL personnel worked with sites to educate system administrators on deploying and maintaining a node on a grid and obtaining a GRASE certificate.

In September of 2006, Miller gave a presentation at the second exploratory NYS Cyberinfrastructure meeting. In this talk, Miller discussed the status of NYS Grid and related information (www.cse.buffalo.edu/faculty/miller/talks.shtml). At the end of the September meeting, an inaugural board was voted on by the membership. It was agreed by the membership that the board would have a term of 1 year and would be required to a) propose a set of by-laws to be voted on by the membership, b) present a mission statement and vision for the initiative, and c) provide a status report of activities. Each initial board member was in charge of a working group: i) resource provider group, ii) user group, iii) technical working group, iv) communications group, v) education, outreach and training group, vi) funding group, and infrastructure group. Subsequent to this meeting, the group established a web site and chose the name NYSGrid for this organization, causing significant confusion nationally between this grass-roots effort and MCIL's well-established New York State Grid (NYS Grid).

In January of 2007, at the third Workshop, Miller and members of MCIL presented the status of NYS Grid to the membership, along with details and demonstrations. Critical results of this meeting included the recognition that NYS Grid was stable and serving numerous users from outside of New York State, that high-end users from New York State required assistance in order to move them onto a Grid, and that education and outreach to faculty, students, and staff throughout New York State was required.

Subsequent to this meeting, the board of this grass-roots initiative, which decided to call itself NYSGrid (www.nysgrid.org/), decided that this initiative was now an initiative of NYSERNet[2], that the board was now a steering committee within NYSERNet, to add new members to the steering committee, to create a position of "program manager" for one of the members of the steering committee, and to remove the executive director. They also received permission from OSG to create a VO called NYSGRID, informed all users of NYS Grid that they were being moved by NYSGrid from GRASE to NYSGRID, and stated that this NYSERNet initiative entitled NYSGrid was now in charge of NYS Grid. In addition, and still

without input from the membership, the steering committee/board sent a proposal to the State of New York requesting funds for several member institutions of the steering committee/board. Needless to say, many of these actions created concern and confusion within New York State.

ENDNOTES

[1] Information in this and the next section is available in more detail in (Miller, *et. al.*, 2008).

[2] NYSERNet is a non-profit organization that focuses on networking concerns throughout New York State.