Enabling Systems Biology: A Scientific Problem-Solving Environment

M. Singhal, E.G. Stephan, K.R. Klicker, L.L. Trease, G. Chin Jr., D.K. Gracio, and D.A. Payne

Computational Sciences and Mathematics,
Pacific Northwest National Laboratory, Richland, Washington
{Mudita.Singhal, Eric.Stephan, Kyle.Klicker, Lynn.Trease, George.Chin, Debbie.Gracio, Debbie.Payne}@pnl.gov

Abstract. Biologists today are striving to solve multidisciplinary, complex systems biology questions. To successfully address these questions, software tools must be created to allow scientists to capture data and information, to share this information, and to analyze the data as elements of a complete system. At Pacific Northwest National Laboratory, we are creating the Computational Cell Environment, a biology-centered collaborative problem-solving environment with the goal of providing data retrieval, management, and analysis through all aspects of biological study. A horizontal prototype called SysBioPSE, demonstrates this vision. Our initial work is centered on developing the Distributed Data Management and Analysis subsystem, which is a specific tool for retrieving data from multiple heterogeneous data stores, providing storage facilities that support pedigree tracking and data and information analysis under a common user interface. With time, many such individual subsystems will be developed and integrated to fulfill the Computational Cell Environment vision.

1 Introduction

The genomics revolution is beginning to produce data at a rate that will soon rival that of high-energy physics. Biologists and informaticists are faced with critical issues of how to best organize this information and share it so that it will provide the greatest value to the greatest number of researchers. Major advances are being realized in the biological sciences by applying a systems approach. However, there are many obstacles that prevent scientists from using the vast amounts of data that are generated and made available on a daily basis. Systems biologists need access to distributed data repositories. Once the data are collected, tools are needed to integrate the heterogeneous data sets. Informatics-based analysis tools are then needed to analyze the data and help draw conclusions. Next, the researcher needs access to tools to easily share the data and knowledge with other researchers. The traditional approaches to answering biological questions are focused on the examination of a single gene or protein, whereas systems biology focuses on studying the complex interactions of biological systems integrating across multiple streams of data from various sources. The Computational Cell Environment (CCE) addresses each of these

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3037, pp. 540–547, 2004.

needs, using advanced and emerging technologies to develop a flexible, problem-solving environment (PSE) for systems biology.

CCE is a scientific PSE. Scientific PSEs are complex computing systems that integrate the activities necessary to accomplish high-level domain tasks [1] [2]. Just like most other PSEs, the goal of CCE is to provide access to data stores, data pedigree tracking, analysis tools, domain-specific workflow support, a user environment, experiment management, and data transformation and filtering capabilities. CCE is complex because of the goal to access data from multiple distributed heterogeneous data stores while analyzing those data via multiple distributed heterogeneous applications. CCE is being developed as a laboratory-directed research and development project for the Biomolecular Sciences Initiative at the Pacific Northwest National Laboratory (PNNL).

2 SysBioPSE Prototype

The spectrum of biological study covers DNA, genes, RNA, proteins, phosphorylation, intra-cellular and extra-cellular networks, tissue, organs, and community interaction. Any individual researcher will likely focus on one of these areas but must draw on research or experimental results from a different area. For instance, a researcher studying intracellular networks must draw on the experimental results from gene expression and/or protein abundance. Once the network is defined, the researcher will likely bring in published research on the proteins of interest. Cross-cutting software analysis tools that can link multiple domains of biological research exist only on a limited scale. CCE has been developed to help link the multiple disciplines within the field of biology.

One of CCE's initial goals was to derive a vision of how biologists and bioinformaticists would conduct computational and collaborative research through a scientific PSE. A horizontal user-interface prototype called SysBioPSE (Biology PSE) was developed to convey the user-centered features and capabilities of an ideal PSE for biological sciences. SysBioPSE was driven directly from the requirements and design criteria emerging from interactions with biologists and bioinformaticists.

User requirements were gathered using a human computer interaction technique known as participatory analysis [3] [4]. Participatory analysis strives to understand the tasks and activities of a user by including the user in the process. A series of interviews with different classes of biologists, including biochemists, molecular biologists, cell biologists, and bioinformaticists, were conducted. We worked with the scientists to understand their work processes, methodology of performing research, the resources that they use, the mechanisms in which they use data and analysis tools to hypothesize, and the way in which they collaborate with each other. Requirements for an overall system were developed from this process.

Based on user requirements, SysBioPSE (Fig. 1) was developed, using the software design tool Director from Macromedia, which was then evaluated and refined across various biological domains. The key feature described in SysBioPSE is the capability to access and integrate various kinds of computational resources such as data sets, applications, and people. An important objective of SysBioPSE is to allow biologists to represent and organize their computational resources in ways that are logical and meaningful to them. To this end, SysBioPSE provides three different

mechanisms or paradigms for working with computational resources. First, a project view allows biologists to define related sets of applications, datasets, and team members and to organize these within specific workspaces called “workbenches.” Second, a concept view allows biologists to relate resources to specific theoretical or experimental

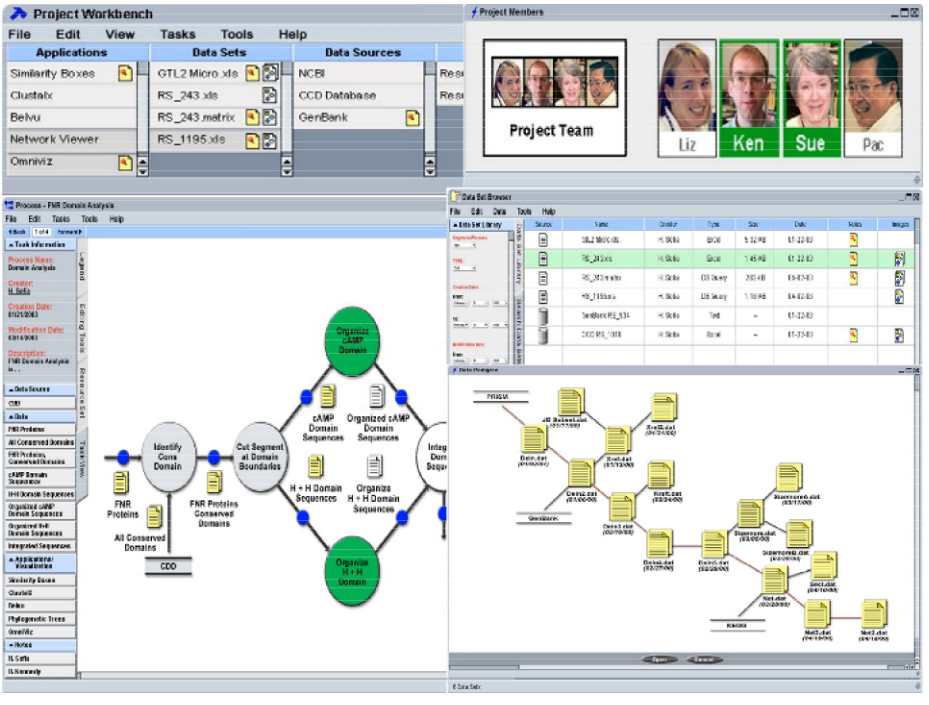

Fig. 1. SysBioPSE Prototype

concepts. In this view, biologists sketch concept diagrams in a free-form manner and link resources directly to visual concept representations. Third, a process view allows biologists to attach resources to specific steps of an experiment or research process. The three paradigms enforce the idea that computational resources are applied by biologists, not in an isolated manner, but in very specific scientific contexts. SysBioPSE captures these scientific contexts such that biologists may manage and apply computational resources in ways that are consistent with their scientific conceptions and perceptions.

3 Distributed Data Management and Analysis Subsystem

In addition to SysBioPSE, the CCE project also developed a functional subsystem aimed at implementing a narrow set of capabilities from the much larger SysBioPSE vision that may be immediately applied by biologists in their research endeavors. The

Distributed Data Management and Analysis (DDMA) system was designed to allow biologists to extract data from various external scientific databases and combine them into an integrated dataset. The development of this system required the CCE project to focus heavily on system architecture and implementation strategies and issues.

3.1 System Architecture

DDMA was developed as a client-server application based on the CCE architecture. The CCE architecture supports independently designed and developed components combined to deliver scalable solutions and reach a large community of researchers from many biology-related scientific domains. Accomplishing this goal requires that the components be designed for loose coupling and operation in a heterogeneous and rapidly evolving service-based environment. To achieve this, an open data-management architecture is devised.

An overall extensible environment is facilitated by defining a multi-tiered, structured architecture that provides a component-based plug-in capability for data resources, collaborative technologies, and analytic tools. Figure 2 shows the CCE architecture. This architecture is designed such that a few basic components are built and deployed initially to a user base, while new functionality is continually added through the development of new components.

The current system provides access to the proteomic data produced at PNNL stored in a system called the Proteomics Research Information Storage and Management (PRISM) System [5] as well as data from the National Center for Biotechnology Information (NCBI)[6], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[7], microarray data, and delimited data. CCE provides easy access to several external data-analysis tools, including PubMed[8], the Conserved Domain Architecture Retrieval Tool (CDart)[9], and the Basic Local Alignment Search Tool (BLAST)[10], all from NCBI; and Cytoscape[11] from the Institute for Systems Biology, as well as analysis tools that are being developed internally at PNNL. Work is in progress to provide customized access to data sources and applications enabling data access from any data source and data analysis from any application.

CCE is implemented as a three-tier architecture system and is based on the Java 2 Enterprise Edition (J2EE) architecture [12]. J2EE is a platform that manages infrastructure and supports web services to enable the development of secure, robust, and interoperable applications. Several of the reasons for adopting a J2EE architecture include fast searching and querying, a better component architecture implemented through JavaBeans (an industry standard), better and more robust security features, transaction management, and most notably, a production quality system.

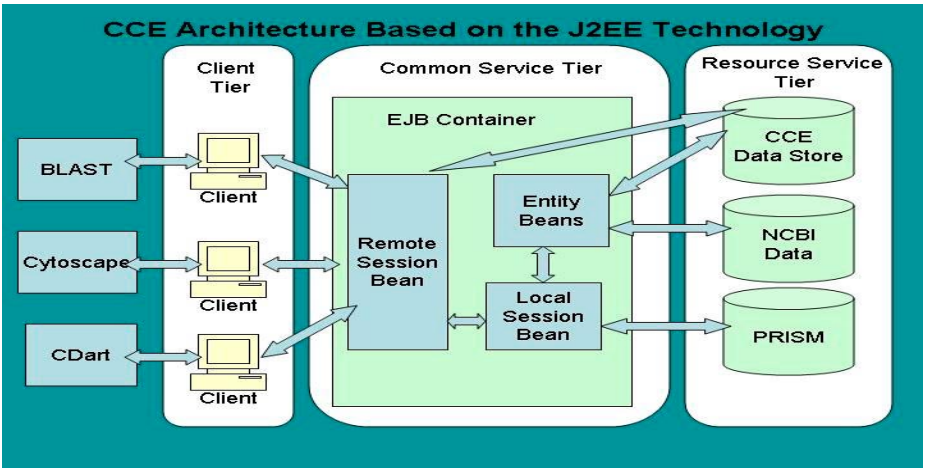

Fig. 2. CCE Architecture

3.2 User Interface

The DDMA user interface (Fig. 3) is centered on a project hierarchy containing projects and the datasets based on authentication. Navigation functions are provided to help users traverse large hierarchies. Once a user has identified the dataset of interest, it can be selected for viewing and analysis. When creating a new dataset, the user has the option of selecting an interface to one of the CCE-supported data sources, such PRISM[5], KEGG[7], and NCBI[6], or creating his/her own custom connection to a data source. Micro-array data as well as data from delimited files can be imported. The connection to the data source is seamless. The user can export the data in spreadsheet format. The system has the capability to handle very large datasets without long delays through a paging technique that is also seamless to the user. Also, large datasets can be navigated seamlessly through a visual panning technique. DDMA also incorporates a flexible, user-driven merge technology to integrate data from different data sources.

4 Summary and Future Work

SysBioPSE and the DDMA system comprise an effective, two-prong strategy for researching and developing PSEs for biological sciences. SysBioPSE provides an overall vision and high-level design of a powerful scientific user environment, while DDMA provides a functional tool that biologists may immediately apply to satisfy the critical data-integration needs of their current research. Through these two systems, CCE addresses four key user requirements:

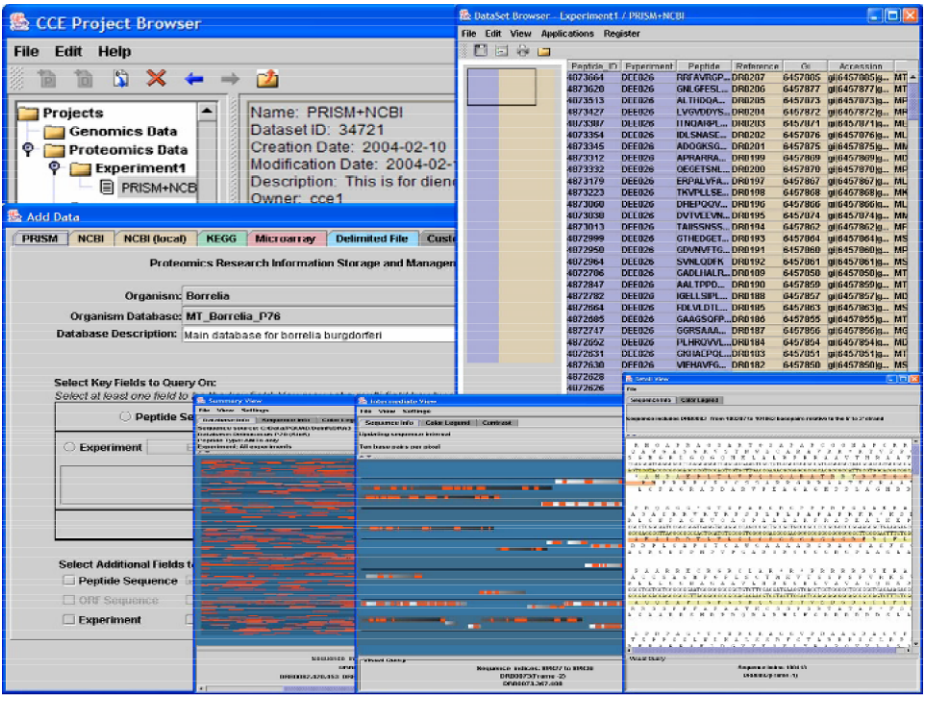

Fig. 3. DDMA User Interface

1) Data Storage and Retrieval: This supports access to terabytes of genomic and proteomic data produced by experimental and computational biologists. The linking and management of distributed, heterogeneous data sources is transparent to users. The data-management system works with and conforms to evolving and *de facto* standards within the biology community. In addition, the data-management system is responsible for storing and retrieving analysis data, images, annotations, tools, and reference information. Subscription capabilities allow users to securely place and retrieve data into and out of the data store. Software security mechanisms are deployed to verify the authentication of users. In the future, CCE will be extended to manage permissions and privileges to the data of users and groups. Encryption capabilities will be added to the system to ensure the secure transmission of data. A mechanism to plug-in data-conversion capabilities and tools, which allow researchers to apply and exchange data in standard scientific data formats, will be provided.

2) Data and Information Analysis: These capabilities include access to a suite of standard data-analysis and visualization tools. Access to these analytical tools allows researchers to examine, investigate, and navigate data to identify and analyze portions that are relevant to their specific research needs. Data-integration capabilities allow researchers to compare and merge different kinds of data, such as proteomic and micro-array data. A flexible environment allows researchers to work with and link combinations of analysis and visualization tools in useful and novel ways.

3) *Common User Interface*: This integrates and delivers data and information management, analysis, and sharing capabilities. The community data portal provides user-interaction layers on top of the various data-management capabilities, providing, at the highest level, interaction with the data-management capabilities through a common user interface. Through the use of profiles, the user interface and its access to data-management functions are customizable to address the specific demands of individuals.

4) Collaborative Data and Information Sharing: This is largely supported in CCE by common access to data, tools, and other resources. However, in the future, CCE will provide facilities beyond shared access to share data in the context of collaborative experimentation and analysis. Through real-time collaboration tools, such as audio/video conferencing, text chat, electronic whiteboards, and application sharing tools, remote researchers may spontaneously collaborate on the application and analysis of data. Through asynchronous, records-based mechanisms, such as mailing lists, bulletin boards, and electronic laboratory notebooks, researchers may collaborate on data over extended periods of time, involve larger numbers of researchers, and maintain a running history of the collaboration and analysis.

5 Conclusions

Today's scientific problems are complex and can benefit from a multidisciplinary environment that can assimilate remarkably diverse data repositories and computational bioinformatics tools that integrate across multiple domains. CCE has proven the possibility of accessing multiple, distributed heterogeneous data stores through one unifying interface. This foundation for a biological PSE will become more comprehensive with time. Moreover, the CCE architecture could easily be adopted in new PSEs supporting other domains.

The goal of CCE is very broad and shall realistically take many years to accomplish. The system has been designed initially to access a subset of the data and tools of interest and is targeted to a small user group. Over time, CCE will expand to include new user groups that will define the additional sets of data and analysis tools to be added to the system. Today, CCE, through DDMA, provides a single visual environment, portions of the data pedigree tracking, tools for data collection and basic data fusion, and several bioinformatics tools for data analysis, generally focused in the area of protein family and protein complex research.

The CCE is a PSE that will not only provide biologists and bioinformaticists greater ability and increased flexibility in manipulating, collecting and analyzing their data, and solving their problems, but will also transform the way this research is carried out and in turn enhance the science of systems biology.

References

1. Gallopoulos SE, Houstis, and JR Rice. 1994. "Problem-Solving Environments for Computational Science."*IEEE Computational Science and Engineering*, Summer: 11-23.
2. Rice JR, and RF Boisvert. 1996. "From Scientific Software Libraries to Problem-Solving Environments."*IEEE Computational Science & Engineering*, Fall: 44-53.
3. Chin G Jr., KL Schuchardt, JD Myers, and DK Gracio. 2000. "Participatory Workflow Analysis: Unveiling Scientific Research Processes with Physical Scientists." In Proceedings of Participatory Design Conference (PDC 2000), pp 30-39. Nov. 28-Dec. 1, New York, NY. CPSR, Palo Alto, CA.
4. Schuler D, and Namioka (eds). 1993. *Participatory Design: Principles and Practices*. Erlbaum Associates, Hillsdale, NJ.
5. PRISM: `http://www.sysbio.org/technologies/comp.stm`
6. NCBI: `http://www.ncbi.nlm.nih.gov/`
7. KEGG: `http://www.genome.ad.jp/kegg/`
8. PubMed: `http://www.ncbi.nlm.nih.gov/PubMed/`
9. `http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=cdd&uid=16610`
10. BLAST: `http://www.ncbi.nlm.nih.gov/BLAST/`
11. Cytoscape: `http://www.cytoscape.org/`
12. Shannon BM, V Hapner, E Matena, Pelegri-Llopart, JS Davidson, and L Cable. 2000. *Java™ 2 Platform, Enterprise Edition: Platform and Component Specifications*. Addison-Wesley Professional. Available URL: http://java.sun.com/j2se