Published in final edited form as:
Computer (Long Beach Calif). 2008 November ; 41(11): 58–66. doi:10.1109/MC.2008.459.

e-Science, caGrid, and Translational Biomedical Research

Joel Saltz[Director],
Center for Comprehensive Informatics at Emory University

Tahsin Kurc[Senior researcher and chief software architect],
Center for Comprehensive Informatics at Emory University

Shannon Hastings[Codirector],
Software Research Institute at the Ohio State University

Stephen Langella[Codirector],
Software Research Institute at the Ohio State University

Scott Oster[Codirector],
Software Research Institute at the Ohio State University

David Ervin[Research specialist],
Software Research Institute at the Ohio State University

Ashish Sharma[Assistant professor],
Department of Biomedical Informatics at the Ohio State University

Tony Pan[Senior research specialist],
Department of Biomedical Informatics at the Ohio State University

Metin Gurcan[Assistant professor],
Department of Biomedical Informatics at the Ohio State University

Justin Permar[Technical manager],
Software Research Institute at the Ohio State University

Renato Ferreira[Associate professor],
Universidade Federal de Minas Gerais

Philip Payne[Assistant professor],
Department of Biomedical Informatics at the Ohio State University

Umit Catalyurek[Associate professor],
Department of Biomedical Informatics at the Ohio State University

Enrico Caserta[Postdoctoral researcher],
Ohio State University

Gustavo Leone[Associate professor],
Department of Molecular Virology, Immunology, and Medical Genetics at the Ohio State University

Michael C. Ostrowski[Professor and chair],
Department of Molecular and Cellular Biochemistry and the Comprehensive Cancer Center at the Ohio State University

Ravi Madduri[Senior research scientist],
Argonne National Laboratory

Ian Foster[Argonne Distinguished Fellow],
Argonne National Laboratory and the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago

Subhashree Madhavan[Associate director],
Life Science Informatics at the National Cancer Institute. She conducted her PhD research at the Uniformed Services University for the health sciences and received an MS in information systems management at the University of Maryland

Kenneth H. Buetow[Associate director],
Bioinformatics and information technology and the director of the NCI Center for Bioinformatics at the National Cancer Institute

Krishnakant Shanbhag[Director], and
Core infrastructure engineering, NCI Center for Biomedical Informatics and Information Technology at the National Cancer Institute

Eliot Siegel[Professor]
University of Maryland School of Medicine

Joel Saltz: jhsaltz@emory.edu; Tahsin Kurc: tkurc@emory.edu; Shannon Hastings: shannon.hastings@osumc.edu; Stephen Langella: stephen.langella@osumc.edu; Scott Oster: oster@bmi.osu.edu; David Ervin: ervin@bmi.osu.edu; Ashish Sharma: ashish@bmi.osu.edu; Tony Pan: tpan@bmi.osu.edu; Metin Gurcan: gurcan@bmi.osu.edu; Justin Permar: jpermar@bmi.osu.edu; Renato Ferreira: renato@dcc.ufmg.br; Philip Payne: philip.payne@osumc.edu; Umit Catalyurek: umit@bmi.osu.edu; Enrico Caserta: enrico.caserta@osumc.edu; Gustavo Leone: gustavo.leone@osumc.edu; Michael C. Ostrowski: michael.ostrowski@osumc.edu; Ravi Madduri: madduri@mcs.anl.gov; Ian Foster: foster@ci.uchicago.edu; Subhashree Madhavan: madhavas@mail.nih.gov; Kenneth H. Buetow: buetowk@mail.nih.gov; Krishnakant Shanbhag: shanbhak@mail.nih.gov; Eliot Siegel: esiegel@umaryland.edu

Abstract

Translational research projects target a wide variety of diseases, test many different kinds of biomedical hypotheses, and employ a large assortment of experimental methodologies. Diverse data, complex execution environments, and demanding security and reliability requirements make the implementation of these projects extremely challenging and require novel e-Science technologies.

Researchers are harnessing dramatic advances in many areas of biomedical technology to better understand the causes of disease and direct disease treatment. Our work explores the critical role e-Science plays in enabling translational biomedical research—the process of developing and applying basic science knowledge and techniques to enable new ways of diagnosing and staging, treating, or preventing diseases, as well as the adoption of best practices in the community.

Translational research projects are diverse in nature. They target a wide variety of diseases, test many different kinds of biomedical hypotheses, and employ a large assortment of experimental methodologies. Diverse data, complex execution environments, and demanding security and reliability requirements, among many factors, make the implementation of these projects extremely challenging and require novel e-Science technologies.

We use pattern templates to identify the requirements that different groups of translational research projects impose on e-Science platforms. The seminal work of Christopher Alexander on designing languages used to capture and describe important and common aspects of architectural design[1] motivates our notion of using pattern templates. Work on software design patterns, in which developers apply somewhat analogous principles to software design, also guides our notion of pattern templates.

In the present context, pattern templates abstract common components and characteristics found in various research categories. While the specifics of the approach employed in a particular project will differ from those of the approaches in other projects, the main

principles and processes can be grouped into several common pattern types. We employ pattern templates to classify and describe these common patterns and to capture design requirements, best practices, and constraints in broad families of applications.

By identifying common requirements, capturing best practices, and recommending strategies, pattern templates motivate the architectural characteristics of e-Science platforms that can enable and simplify the implementation, deployment, evaluation, and management of complex translational research projects.

Our work with a variety of translational research projects has led us to identify several "pattern templates" and, based on analysis of their requirements, to develop tools that facilitate their implementation and deployment. Table 1 shows several examples of pattern templates in translational research and the primary characteristics of studies that can be modeled by the respective pattern templates. We note that different components of a single translational research study can be modeled by more than one pattern template.

We describe three of these templates and present how two e-Science tools, caGrid and caIntegrator, can be used to support these templates. *caGrid*, a service-oriented, model-driven Grid software infrastructure,[2] provides an integral component of the cancer Biomedical Informatics Grid (*caBIG*) program (https://cabig.nci.nih.gov); caGrid supplies the core infrastructure for federating data and analytical resources and applications deployed at different institutions within the caBIG environment. We designed the infrastructure and implemented it as a general-purpose middleware system that can support other biomedical and non-biomedical application domains. *caIntegrator* offers a novel translational informatics platform (http://caintegrator-info.nci.nih.gov) that lets researchers and bioinformaticians access, analyze, and integrate clinical and experimental data across multiple clinical trials and studies.

SAMPLE PATTERN TEMPLATES

Examining the first three pattern templates in Table 1— prospective, multiscale, and integrative—shows how present cancer research projects can be modeled.

Prospective clinical research

The prospective template involves studies in which researchers systematically follow a group of patients over time. Researchers design these prospective studies to understand risk factors for the development or progression of disease, to assess the effects of various treatments, and to perform quality control in disease classification and treatment.

Clinical studies that rely on biomedical imaging as both an indicator of disease progression and an assessment of the treatments' effects are examples of prospective studies. These studies can also capture other metrics, such as treatment reports and outcome data. In studies conducted by cooperative groups, researchers obtain imaging and clinical data from patients at multiple institutions. The interpretation of radiology, radiation treatment, and pathology information plays a crucial role in these studies for reproducible disease classification and assessment of treatment response.

In radiation oncology, for example, digital imagery helps define the tumor volume, outlines areas that receive radiation, and excludes the portions of healthy tissue that must be spared. There is, however, a high interobserver variability among image data reviewers. One strategy increasingly used to reduce this variability and consequently improve protocol compliance focuses on a central review of imaging objects. With this approach, multiple

expert radiologists at different institutions review image and clinical data, with an independent adjudicator incorporating these reviews into a consensus assessment.

The diversity of data management systems and their incompatibility present a major challenge in implementing central review. In addition, the prospective template has a huge semantic scope that encompasses a vast span of possible diseases, treatments, symptoms, and radiology and pathology findings in imaging-based studies.

Multiscale investigations

The multiscale template models the studies that attempt to measure, quantify, and in some cases simulate biomedical phenomena in a way that takes into account multiple spatial and temporal scales. They thus attempt to understand the interplay between anatomical structures, physiology, and systems biology in disease processes—the development of tumors or metastases, for example. This pattern template involves formulation, execution, and analysis of large-scale coordinated experiment sets that address multiple types of microscopy data and high-throughput biomolecular data such as genetic and epigenetic data.

Study of the tumor microenvironment provides an example of the multiscale template. Cancer development occurs in both space and time, and cancers are composed of multiple different interacting cell types. The genetics, epigenetics, regulation, protein expression, signaling, growth, and blood vessel recruitment all take place in time and space.

The tumor microenvironment consists of different cell types, including fibroblasts, glial cells, vascular and immune cells, and the extracellular matrix (ECM) that holds them together. Researchers normally investigate cellular signal interchange in such an environment for a harmonic development and subsequent maintenance of the various tissues' elaborated cellular organization. Mounting evidence indicates that alteration of intercellular signaling could be responsible for malignant progression of a developing cancer. Recent investigations have also revealed the important role played by fibroblasts during cancer's initiation and progression.

An ongoing project at the Ohio State University investigates how the heterogeneous microenvironment surrounding the tumor stroma can influence the four main phases of tumor progression over time: normal, hyperplasia, adenoma, and invasive carcinoma. In this project, genomics and molecular information extracted through a laser-capture microenvironment combines with a microanatomic structure analyzed through confocal microscopy.

The study analyzes "function" as a parameter of tumor progression. Datasets are semantically complex because they encode ways in which morphology interrelates over time with genetics, genomics, and protein expression. Researchers analyze high-resolution microscopy images obtained from tissues and process the analysis results as annotations on image regions of interest, which represent cells, cell types, and the cells' spatial characteristics. Researchers then combine and enrich this information with molecular information acquired through laser capture microdissection analysis. Image annotations and molecular information can be associated with concepts defined in one or more ontologies that represent the domain knowledge. Raw and annotated image and molecular data can then be organized into a semantic knowledge base.

This knowledge base represents three-dimensional models of the tumor microenvironment that a cancer biologist can, for example, use to ask morphological and biochemical functionality questions to understand how mutations of the tumor suppressor gene PTEN in fibro-blasts promote breast cancer progression and metastasis. Researchers can use this

information to develop novel therapeutic strategies that target fibroblasts to stop growth, preventing recurrence or metastasis.

Deep integrative clinical analyses

The integrative template involves patient-related studies attempting to predict and explain patient response to treatment protocols by collecting and analyzing comprehensive sets of high-throughput molecular data, image data, and clinical data. Researchers can use the integrative template to model large-scale projects such as the Glioma Molecular Diagnostic Initiative (GMDI; http://bethesdatrials.cancer.gov/neuro_oncology/nci02c0140/default.aspx), the Cancer Genome Atlas project (TCGA; http://cancergenome.nih.gov), the Cancer Genetic Markers of Susceptibility (CGEMS; http://cgems.cancer.gov) project, and the I-SPY breast cancer trial (http://ncicb.nci.nih.gov/tools/translation_research/ispy).

GMDI, funded by the National Cancer Institute (NCI), studies gene expression and genomic data from patients afflicted with tumors (gliomas). Researchers will follow these patients through the natural history and treatment phases of their illness. The GMDI seeks to develop a molecular classification schema that is both clinically and biologically meaningful and to explore gene expression profiles to determine the patients' responsiveness and correlate it with discrete chromosomal abnormalities. TCGA is a large-scale community resource project cofunded by the NCI and National Human Genome Research Institute.

The glioblastoma multiform (GBM) component of TCGA is systematically studying the patterns of epigenomic, genomic, and transcriptomic aberrations in 500 primary GBMs, accompanied by targeted resequencing of 3,000 genes. This work seeks to achieve a better understanding of molecular subtypes, identify key interrelationships among recurring molecular alterations, and define major glioma genes that are rational targets for therapeutic and diagnostic development. GMDI and TCGA generate copious data (~1 terabyte of data from 224 patients in TCGA) in the areas of genomics, proteomics, clinical record, and biospecimen datasets at disparate locations and in varied formats. These datasets must be integrated in semantically meaningful ways to extract new knowledge that can impact therapies.

The InterSPORE (http://spores.nci.nih.gov) multicenter clinical I-SPY trial monitors women undergoing neoadjuvant chemotherapy for breast cancer. A collaboration of physicians, researchers, and cancer cooperative groups, the study uses molecular and imaging characteristics (obtained from magnetic resonance images) to identify those patients likely to respond to novel therapeutic agents, which could then be tested in the neoadjuvant setting. The CGEMS project conducts *genome-wide association studies* (GWASs) with follow-up replication studies to identify common, inherited gene variations that either increase or decrease cancer risks.

As GWAS technology becomes increasingly efficient, researchers are challenged to meaningfully integrate and transform the wealth of genetic association data from thousands of individuals into better strategies for diagnosing, treating, and even preventing disease. Both I-SPY and CGEMS provide query and analysis of associations between genetic variations—phenotypic changes from the heterogeneous collection of data types and databases.

Pattern templates' software infrastructure requirements

The discovery, analysis, and integration of heterogeneous information resources commonly occur in translational research pattern templates. In many research studies, however, distributed data sources are fragmented and noninteroperable. Datasets vary in size, type,

and format and are managed by different system types. Many of the cooperative groups that conduct image-based prospective studies use different database systems at data collection sites, which cannot easily interface with each other and with image archival systems.

The need for semantic integration of heterogeneous data types and data sources also arises in translational research patterns. The multiscale template involves synthesis of information from multiple scales of biological and functional data. The integrative template requires integration of different types of high-throughput molecular data. In large-scale, multi-institution studies, naming schemes, taxonomies, and metadata used to represent the data's structure and content are heterogeneous and managed in silos; any two databases can define data that contains the same content with completely different structural and semantic information.

To support the classes of studies presented here, e-Science information systems must enable syntactic and semantic interoperability and integration of heterogeneous and potentially distributed data and analytical resources. *Syntactic* interoperability enables programmatic access to a system's functionality, while *semantic* interoperability refers to systems' ability to reproducibly and consistently exchange and reason upon conceptual knowledge types and to correctly and unambiguously use resources. The sample pattern templates provide particularly strong drivers for standardization and a support services infrastructure, such as gridwide metadata management, management of data models' structure and semantics, and advertisement services to facilitate discovery and interoperability of services containing data relevant to a study.

Federated query support is critical for realizing the pattern templates we describe. Studies in the prospective and integrative templates involve querying and assimilating information associated with multiple groups of subjects from multiple data sources, comparing and correlating the information about the subject under study with this information, and classifying the analysis results.

Workflow requirements also arise from all three templates. Multiscale template information systems, for example, should support analysis of data by a series of simple and complex image analysis operations expressed as a data analysis workflow. Workflows in the sample multiscale template project might include operations such as correction of various data acquisition artifacts, filtering operations, segmentation, registration, feature detection, feature classification, and interactive image annotation.

Many pattern templates rely on protecting sensitive data and intellectual property. The prospective template in particular has strong requirements for authentication and controlled access to data because prospective clinical research studies capture, reference, and manage patient-related information. While security concerns are less stringent in the other translational research pattern templates, the intellectual property of scientists and proprietary resources must be protected. Researchers need comprehensive software support for authentication, authorization, access control, and management of trust relationships within and across institutions.

CAGRID AND CAINTEGRATOR

Biomedical research pattern templates motivated the designs for two e-Science platforms: a service-oriented, model-driven grid software infrastructure and a novel translational informatics platform that the NCI Center for Bioinformatics is developing (http://caintegrator-info.nci.nih.gov).

caGrid

Service-oriented architectures (SOAs) and model-driven architectures (MDAs) have gained popularity in recent years as frameworks for developing interoperable systems. SOAs encapsulate standards, such as the Web Services Resource Framework, that provide for common interface syntax, communication and service invocation protocols, and services' core capabilities. MDAs promote a software design approach based on platform-independent models and the metadata used to describe them.

caGrid leverages grid services technologies and grid systems, including the Globus Toolkit[3] and Mobius,[4] and tools developed by the NCI such as the caCORE infrastructure.[5] The system exposes analytical and data resources to the environment as grid services, hosted at different sites and interacting with clients through grid service protocols. caGrid services are standard resource framework services,[6] and any specification-compliant client can access these services.

The driving principle behind caGrid's design focuses on enabling and supporting syntactic and semantic interoperability among resources through a federation of such resources. caGrid's design draws from the MDA paradigm. caGrid leverages and supports the concepts of controlled vocabularies, strongly typed services, common data elements, published information models, and rich service metadata.

caGrid service and client APIs provide an object-oriented view of the encapsulated data or analytical resource. Developers using caGrid build their object-oriented view of each resource on common data elements and registered domain models, expressing these models as object classes with attributes consisting of common data elements and class relationships. Developers annotate these attributes with terms from controlled vocabularies, registering common data elements in the Cancer Data Standards Repository (caDSR).[5,7]

The definitions of caGrid data elements draw from the vocabulary registered in the Enterprise Vocabulary Services (EVS) repository.5,7 In the caGrid environment, a client and a service conceptually exchange objects from the domain model exposed by the service. The system serializes any object transferred over the grid into an XML document that adheres to an XML schema published in the Mobius Global Mode 1 Exchange (GME) service.4

Thus, the system makes the object's XML structure available to any client or resource in the environment. The system defines the properties and semantics of data types in caDSR and EVS, then stores the structure of their XML materialization in Mobius GME. In this respect, caGrid services are strongly typed, consuming and producing objects with classes and XML representations that conform to well-defined, publicly accessible models.

The caGrid core infrastructure consists of several elements: tools for developing application services, querying data sources, and composing services into workflows; a suite of coordination services; and a runtime environment for service deployment, execution, and invocation. The coordination services provide support for such common operations as metadata management, workflow execution, federated query processing, and security. Figure 1 shows the caGrid infrastructure's core components.

To protect researchers' intellectual property and ensure the safety and privacy of patient-related information, security plays an essential role in the environment's successful deployment. caGrid thus provides a comprehensive set of services to support secure and controlled access to resources based on policies set forth by resource owners.[8] These services enable gridwide management of user credentials, provide support for grouping

users into virtual organizations for role-based access control, and allow management of trust fabric in the grid.

The caGrid infrastructure supports federated querying of multiple data services to enable distributed aggregation and joins on object classes and object associations defined in domain object models. The caGrid workflow service supports the execution and monitoring of workflows expressed in the Business Process Execution Language (BPEL; www.ibm.com/developerworks/library/specification/ws-bpel). In a recent effort, researchers are integrating the workflow service support with the Taverna Workflow Management System[9] to support advanced user interfaces and higher-level mechanisms for workflow composition, execution, and monitoring—all features critical to end users' wider adoption of caGrid's workflow support.

caIntegrator

A novel translational informatics platform being developed by the NCI Center for Bioinformatics, caIntegrator lets researchers and bioinformaticians access, analyze, and integrate clinical and experimental data across multiple clinical trials and studies. The caIntegrator framework provides cohesive access to a variety of data types, such as microarray-based gene expression, SNPs, and clinical trials data. This framework offers a paradigm for rapid sharing of information and accelerates the process of analyzing results from various biomedical studies, with the ultimate goal of rapidly changing routine patient care.

caIntegrator provides a common set of application programming interfaces and specification objects that define the clinical genomic analysis services. These interfaces and objects act as templates for the caIntegrator-based translational research applications, which extend and implement these interfaces and specification objects. An application's user interface communicates with its caIntegrator-based middle-tier services via domain as well as business objects. A generic, real-time analytical service implemented in caIntegrator currently supports simple statistical analysis within an application or the import of data into third-party analytical tools. The caIntegrator data system consists of a star schema database that contains the clinical and annotation data as dimensions, precalculated gene expression copy-number data as facts, and tables for user provisioning data. A generic interface in caIntegrator allows visualization of genomic data—for example, copy-number, scatter, and ideogram plots.

The system exposes community-provided data and analytical resources to clients as caGrid data and analytical services. Each service provides access to the functionality of its back-end system via well-defined service interfaces and data models registered in the caDSR, EVS, and GME infrastructure. A service advertises itself to the environment by registering service metadata, which includes information about the service's host location, the service provider's contact information, which data models the service exposes, and so on.

A service provider can use security services to enforce authentication and access control policies to restrict access to a service. caGrid-enabled clients can discover services using metadata registered in the metadata services (the Index Service) and submit federated queries across multiple services or execute workflows involving multiple data and analytical services.

TRANSLATIONAL RESEARCH PATTERN TEMPLATES

Researchers have employed caGrid and caIntegrator in the implementation of applications, tools, and infrastructure support in several biomedical research projects. caIntegrator has

been motivated by and used in the implementation of the I-SPY trial and the CGEMS projects, among other studies. caGrid provides the core grid architecture for the caBIG program, and the community employs it for implementation of grid-enabled, interoperable services, tools, and applications such as a distributed image analysis and review application.[10] It also provides the core grid middleware components, along with components from the Biomedical Informatics Research Network,[11] for the CardioVascular Research Grid (CVRG; http://cvrgrid.org).

Researchers can use tools such as caGrid and caIntegrator to develop translational research pattern templates. The integration of data from different data types and databases is a common requirement in pattern templates. Researchers can use caIntegrator to create a data warehouse of molecular, tissue, and clinical data collected in the laboratory, then analyze and integrate information from these data types. caIntegrator provides analysis services and integrated data models for several molecular and clinical data types. Using caIntegrator security mechanisms, researchers can restrict access to datasets by different collaborators.

caGrid's SOA and MDA architecture and core elements address the requirements for interoperability, remote access, security, resource discovery, and federated query in a multi-institutional, distributed environment. Researchers can use the caGrid infrastructure to create multiple instances of the caIntegrator framework, as well as other data management systems used in a study. They can wrap the data management and analytical service interfaces of a caIntegrator instance as caGrid data and analytical services. Similarly, researchers can set up different databases as caGrid data services and use published data models and schemas to expose the data. A client application can interact with these databases via well-defined interfaces and data models without needing to know how the data is stored in the respective database systems. Figure 1 shows a deployment in which multiple caIntegrator instances and other data and analytical resources are wrapped as caGrid services and can be accessed by client applications.

Researchers can use caGrid's security infrastructure to secure services so that only scientists participating in a study can access sensitive information and invoke analysis methods. A cooperative group, for example, could expose its data collection repositories as secure caGrid services, allowing clinical-trials collection sites to securely upload data directly without needing to install any special software. Using the caGrid infrastructure would let both researchers and reviewers doing central review in prospective studies securely connect to the grid service, retrieve their review work orders, and submit their findings directly. This approach removes the need for a trial coordinator to explicitly send the review objects and provide participants with review software.

Using the federated query mechanism in caGrid, a reviewer or researcher can compose and execute a query across multiple data services to, for example, collect molecular and imaging data on a group of patients, based on specified clinical characteristics. The query's results can then be processed through an analysis workflow involving analytical services at multiple locations.

The efficiency with which researchers can integrate, disseminate, and analyze clinical, imaging, molecular, and tissue data both within and across functional domains represents a critical factor in advancing translational research. e-Science platforms building on SOA and MDA paradigms have tremendous potential to address the requirements arising from common pattern templates in translational research. The caGrid system integrates MDA and SOA methods, with an emphasis on common data elements and controlled vocabularies.

While caGrid provides a comprehensive suite of core services and tools, room for improvement remains in several areas. For example, interoperability between e-Science

platforms developed using different architecture designs and by different communities is critical to future national- and international-scale biomedical research and biomedical informatics efforts. The Biomedical Informatics Research Network[11] and myGrid[12] provide two prominent examples of infrastructures developed by different communities for grid-enabled biomedical research.

Developing support for integration of caGrid with such infrastructures and enterprise systems must involve multiple axes of interoperability. Shared domain semantics that manifest as published information models, common and controlled terminologies, and standards for data-type bindings are clearly key to interoperability. Another axis involves harmonizing security and policies. Moreover, developers need tools and services to enable efficient mappings between different messaging standards, controlled vocabularies, and data types associated with many communities, and between different messaging and resource invocation protocols.

Acknowledgments

We thank Laura Esserman and Nola Hylton at the University of California, San Francisco, the PIs for the I-SPY trial project; Howard Fine at the NCI, the PI for the GMDI and Rembrandt projects; and Stephen Chanock, NCI, the PI for the CGEMS (GWAS) project. Our work was supported in part by the NCI caGrid Developer grant 79077CBS10, the State of Ohio BRTT Program grants ODOD AGMT TECH 04-049 and BRTT02-0003, the NHLBI R24 HL085343 grant, the NIH U54 CA113001 and R01 LM009239 grants, and NSF grants CNS-0403342 and CNS-0615155; by the NCI and NIH under contract no. N01-CO-12400; by the US Department of Energy under contract DE-AC02-06CH11357; and by Children's Neuroblastoma Cancer Foundation. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

References

1. Alexander, C. A Pattern Language: Towns, Buildings, Construction, (Center for Environmental Structure Series). Oxford Univ. Press; 1977.
2. Oster S, et al. caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. J Am Medical Informatics Assoc 2008;15:138–149.
3. Foster I. Globus Toolkit Version 4: Software for Service-Oriented Systems. J Computational Science and Technology 2006;21:523–530.
4. Hastings, S., et al. Distributed Data Management and Integration: The Mobius Project. Proc. Global Grid Forum 11 (GGF 11), Semantic Grid Applications Workshop, Global Grid Forum; 2004. p. 20-38.
5. Covitz PA, et al. caCORE: A Common Infrastructure for Cancer Informatics. Bioinformatics 2003;19:2404–2412. [PubMed: 14668224]
6. Foster I, et al. Modeling and Managing State in Distributed Systems: The Role of OGSI and WSRF. Proc IEEE 2005;93:604–612.
7. Phillips J, et al. The caCORE Software Development Kit: Streamlining Construction of Interoperable Biomedical Information Services. BMC Medical Informatics and Decision Making 6(2):2006.
8. Langella, S., et al. Proc 2007 Am Medical Informatics Assoc (AMIA) Ann Symp. Am. Medical Informatics Assoc; 2007. The Cancer Biomedical Informatics Grid (caBIG) Security Infrastructure; p. 433-437.
9. Hull D, et al. Taverna: A Tool for Building and Running Workflows of Services. Nucleic Acids Research 2006;34(Web Server issue):729–732.
10. Gurcan M, et al. GridImage: A Novel Use of Grid Computing to Support Interactive Human and Computer-Assisted Detection Decision Support. J Digital Imaging 2007;20:160–171.
11. Grethe JS, et al. Biomedical Informatics Research Network: Building a National Collaboratory to Hasten the Derivation of New Understanding and Treatment of Disease. Student Health Technology Information 2005;112:100–109.

12. Stevens RD, Robinson A, Goble CA. myGrid: Personalised Bioinformatics on the Information Grid. Bioinformatics 2003;19:302–304.

Figure 1.
caGrid infrastructure core components. The infrastructure consists of three elements: tools for developing application services, querying data sources, and composing services into workflows; a suite of coordination services; and a runtime environment for service deployment, execution, and invocation.

Table 1

Examples of pattern templates that arise in translational research studies.

Template	Characteristics of translation research studies
Prospective clinical research study (prospective template)	Studies in which a group of patients are systematically tracked over time. These studies explain risk factors for the development and progression of disease or assess the effects of various treatments.
Multiscale investigations that encompass genomics, epigenetics, (micro)-anatomic structure and function (multiscale template)	Studies that attempt to measure, quantify, and—in many cases—simulate biomedical phenomena to take into account multiple spatial and temporal scales. These studies involve formulation, execution, and analysis of large-scale coordinated sets of experiments involving multiple types of microscopy imaging and high-throughput genetic, genomic, and epigenetic analyses. The studies also seek to understand the interplay between anatomical structures, physiology, and systems biology in disease processes—for example, the development of tumors or metastases.
Deep integrative clinical analyses (integrative template)	Studies that attempt to predict and explain patient response to treatment protocols by collecting and analyzing comprehensive sets of high-throughput molecular data.
Secondary data analysis	Studies that carry out analysis of already curated data to gain new insights. Data is analyzed using new methods and integrated or correlated in new ways.
Coordinated system-level attack on a focused problem	Studies that carry out a closely coordinated set of experiments in which results are integrated to answer a set of biomedical questions. Typically these studies must analyze and integrate information from multiple complementary experiments that yield different but closely interrelated data types such as gene expression, image, SNP, and clinical data.
Adaptive image-guided intervention	Studies that involve interactive use of image data for therapy and surgery planning and for performing surgery. Data can come from different types of image modalities. Image data is analyzed iteratively and interactively to create a treatment plan, adjust drug dosage, and so on. Researchers use time-dependent images and simulation in soft real time to modify plans.