The CardioVascular Research Grid (CVRG) Project

Raimond L. Winslow, Ph.D.[1], Joel Saltz, M.D./Ph.D.[2], Ian Foster, Ph.D.[3], John J. Carr, M.D.[4], Yaorong Ge, Ph.D.[4], Michael I. Miller, Ph.D.[1], Laurent Younes, Ph.D.[1], Donald Geman, Ph.D.[1], Stephen Granite, MBA[1], Tahsin Kurc, Ph.D.[2], Andrew Post, MD./Ph.D.[2], Ravi Madduri[3], Tilak Ratnanather, D. Phil[1], Jennie Larkin, Ph.D.[5], Siamak Ardekani, MD./Ph.D.[1], Timothy Brown[1], Anthony Kolasny[1], Kyle Reynolds[1], Michael Shipway[1]

[1]The Johns Hopkins University, Baltimore, MD; [2]Emory University, Atlanta, GA; [3]The University of Chicago, Chicago, IL; [4]Wake Forest University, Winston-Salem, NC; [5]NHLBI, NIH, Bethesda, MD

Abstract

The CardioVascular Research Grid (CVRG) Project's goal is to facilitate research on heart disease through open-source informatics and data analysis tools, making it easier for researchers to manage, share, and analyze complex data collected in cardiovascular studies. Inefficient mechanisms for sharing and analyzing data hamper large studies and make it hard to combine data across them. Through work with "Driving Biomedical Projects", the CVRG team learns about the research community's informatics needs, and develops tools to meet those needs. The CVRG has deployed: 1) genetic, genomic, proteomic, electrocardiographic, imaging, and clinical data storage systems; 2) easy to use interfaces to query, retrieve, and analyze the data; and 3) novel tools for statistical analysis of study data and for analysis of heart shape and motion. As an NHLBI-funded resource, the CVRG provides informatics tools that help researchers focus on their science rather than on the details of data representation and management.

Introduction

Large-scale, multi-institutional studies of cardiovascular disease play an important role in biomedical research and the mission of the National Heart, Lung, and Blood Institute (NHLBI). These projects have the potential to increase our fundamental understanding of disease mechanisms across hierarchical levels of biological organization, and to enable discovery of biological markers correlating with different disease states and inter-individual differences in risk.

However, several challenges must be addressed before cardiovascular researchers can fully capitalize on these studies' emerging data. First, the current data storage methods used in these studies make it difficult or impossible to search data interactively (studies store data in file systems as text files, spreadsheets, plain images, etc.). As we enter an era in which it will be increasingly common to acquire multi-scale data (genetic, genomic, proteomic, structural and functional imaging, clinical, electrocardiographic (ECG)) from large cohorts, researchers must be able to search for, select, and analyze specific subsets of data that meet complex search criteria. Second, data must be shared so that others may use it to discover new knowledge[1]. However, it is necessary to provide explicit definitions of the variables collected in a study and the collection methods applied (semantic description of the data). This is seldom done in cardiovascular research, and as a result, research teams cannot easily understand and re-use data collected by others. This is especially important in genome-wide association analyses where meta-analysis is proving to be particularly important. Third, a data-coordination center (DCC) functions as a central data repository in many studies. However, DCCs typically do not provide tools for interactive access to and exploration of study data sets. Instead, investigators must request access to data, the DCC reviews the request, and if granted, the DCC sends data to the investigators on a storage medium or makes it available via FTP. This is a slow and cumbersome process. All of these barriers limit the extent to which cardiovascular data may be disseminated, analyzed, and re-used. In a very real way, this decreases the value of the data[1].

Increasingly, the creation of "bio-grids" facilitates cross-institutional scientific collaboration and data sharing. The Biomedical Informatics Research Network[2] provides a national infrastructure for sharing and analyzing image data on human and animal models of brain disease. The Cancer Biomedical Informatics Grid[3] provides a national infrastructure for sharing data and tools in cancer research. The technology developed and used in these projects addresses some of the problems discussed above. Cardiovascular research, however, is unique in terms of the many different types of data that must be

collected, shared and analyzed to help reveal the mechanisms of heart disease. The purpose of this article is to describe the CardioVascular Research Grid (CVRG) Project. The CVRG Project addresses the issues outlined above by developing open-source tools for the management, semantic description, and analysis of cardiovascular data, and making these tools available for community use. CVRG goals are threefold: a) to listen to the informatics needs of the cardiovascular research community; b) to develop and make freely available well tested data management and analysis tools that address these needs; and c) to help researchers use these tools, freeing them from the myriad details of managing data, and allowing them to focus on their scientific research.

Methods

The most important principle guiding the CVRG is that the needs of the cardiovascular research community drive all development. To discover these needs, the CVRG team is recruiting Driving Biomedical Projects (DBPs). These are large-scale, (inter-) national research projects that use the CVRG infrastructure, and that provide the CVRG team with research scenarios that guide further technology development and refinement. Starting with one DBP, the CVRG has expanded to eight DBPs over the past year. These projects are listed in Table 1 and described on the CVRG website. This table also shows the CVRG resources that are in use.

The CVRG team designed tools for managing and analyzing data in a highly modular fashion. Each module, referred to as a service, performs a single data management or analysis function (e.g., one service is used to store ECG data, while another service computes properties of ECG waveforms). The isolation of specific data processing operations into services simplifies the task of building and adding new tools. Services interact over a communications network so that data may be passed from one service to another to perform a sequence of data storage, access, and analysis operations. The communications network and the set of services are referred to as a "grid". We are using technologies developed in the caGrid Project to develop these services[4].

Two types of services are available on the CVRG - data services and analytic services. Data services store data, and output data in response to queries. Analytic services accept input data of a particular format, process that data, and then output results in a particular format. For example, an analytic service may receive ECG data, calculate average QT interval

Driving Biological Project	CVRG Tools
Coronary Artery Disease Risk in Young Adults (Echo, MR, CT, and ECG Reading Centers)	ECG data and analysis services, XNAT, OpenClinica data service
Genome-Wide Association Analysis in Essential Hypertension Study (FEHGAS)	SNP data service, caGWAS
Hypertrophic Cardiomyopathy Consortium	SNP data service, ECG data and analysis services, XNAT, OpenClinica data service, LDDMM and PTA analytic services
Jackson Heart Study (MR, CT, and ECG Reading Centers)	ECG data and analysis service, XNAT
Multi-Ethnic Study of Artherosclerosis (MR, CT, and ECG Reading Centers)	ECG data and analysis services, XNAT, caGWAS, OpenClinica data service
Pediatric Heart Network (Echo Reading Center)	XNAT
Prospective Observational Study of the ICD (PROSE-ICD)	SNP data service, ECG data and analysis services, XNAT, OpenClinica data service, LDDMM and PTA analytic services
Study of pathophysiology of cardiovascular disease and CV disparities (sites: Emory, Grady)	SNP and gene expression data services, ECG data and analysis services, OpenClinica and clinical data services
Minority Health Genomics and Translational Research Biorepository Database (MH-GRID) (sites: Morehouse School of Medicine, Grady, Jackson Hinds Clinic, and Kaiser)	Clinical data services, biospecimen data service, SNP data services, laboratory data service

Table 1: Current DBPs (left) and CVRG tools used (right). *The final two DBPs are in the start-up phase, plan to use CVRG tools for data management and service federation, and add enhancements and new data services to the CVRG system.*

and its variability, and return output values displayed to the user and/or stored for subsequent use. Workflows sequentially coordinate the action of data and analytic services to perform particular data management and processing functions.

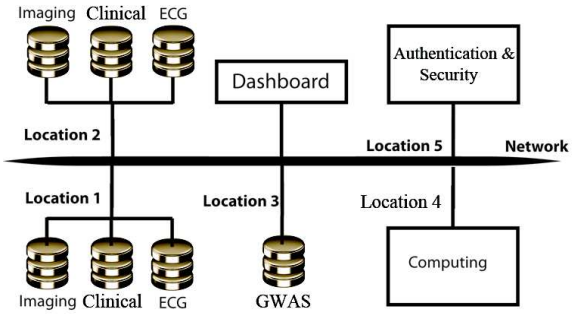

Figure 1: A hypothetical deployment of the CVRG software.

Figure 1 shows a hypothetical deployment of CVRG tools to support collaborative research in which image, ECG, genome-wide association study (GWAS) and clinical data are acquired. This example illustrates how data may be collected and managed at different sites (e.g., study field centers), and how the CVRG Dashboard may be used as a central portal to query data, retrieve data sets of interest, and send these data to computing resources for analysis.

In addition to the distributed model described above, CVRG tools may be installed at a single site. The advantage of this is that it provides a way of quickly establishing a single data repository at which users may access the full complement of data and analysis tools via the CVRG Dashboard. This is often preferred when investigative sites lack the local technical expertise and computing resources needed to operate CVRG tools.

Results

ECG Data and Analysis:

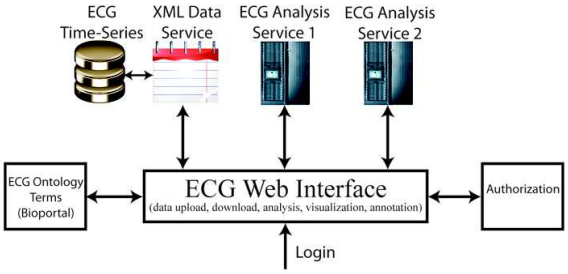

Figure 2: An illustration of the CVRG ECG data management and analysis workflow.

The CVRG team has developed a set of novel services for managing and analyzing ECG data (Fig. 2). Users log into the CVRG Dashboard to upload ECG time-series data into the ECG data service. The Dashboard stores ECG time series in a file system, and stores metadata about data collection protocols and values computed from the data using analysis algorithms in an XML data service.

Users may retrieve data, visualize, and annotate it within the web interface. Annotations are made by interactive search of an ECG ontology developed in the CVRG Project. The National Center for Biomedical Ontologies (NCBO)[5] manages the ontology using Bioportal. Annotation terms are retrieved via REST calls to Bioportal. Annotated data may be stored back into the data service for retrieval and further use. Users may also select data sets for analysis, and send them to one of two data analysis services (Fig. 2). The first is a set of algorithms developed by Berger et al[6] which compute several different ECG properties. The second is an automated algorithm developed by Chesnokov[7] for QT-interval analysis. On completion, the interface displays the results, which may be downloaded in Excel spreadsheet format and/or stored back into the data service for subsequent use.

Image Data Management and Analysis:

The CVRG uses the eXtensible Neuroimaging Archive Toolkit (XNAT)[8] for managing processed DICOM image data and information relating to heart shape and motion analyses[9-12]. The CVRG also uses XNAT for automated quality control of the incoming image data, through the use of its pipeline engine and extensions to its DICOM processing software. Cardiac image data stored in DICOM format are retrieved, reformatted, and stored in XNAT for further processing. Successive heart image sections obtained using either CT or MR imaging are segmented using the Seg3D software[13], yielding full three-dimensional volume reconstruction of populations of imaged hearts.

The CVRG team has developed two novel analysis methods for analyzing heart shape and motion. The first uses the Large Deformation Diffeomorphic Metric Mapping (LDDMM) algorithm[9,10]. This algorithm computes a common coordinate system in which to represent the shape of each volume-reconstructed heart at a given time point. Upon representation of multiple heart volumes in this common coordinate system, average shape and its variation may be calculated and studied over populations of hearts. The second workflow makes use of the Parallel Transport Algorithm[11]. This algorithm computes a common coordinate system in which to represent the deformation of volume

reconstructed hearts over the cardiac cycle. The CVRG team has used this algorithm to determine features of heart deformation that also support the discrimination of ischemic versus non-ischemic cardiomyopathy on the basis of heart deformation[12]. These algorithms are available as services on the CVRG.

Clinical Information and Integrative Management:

The CVRG uses OpenClinica[14], an existing open-source, clinical trials software package, to manage patient data and case report forms. OpenClinica provides tools for clinicians to perform both observational and longitudinal studies, maintaining a list of subjects separate from the studies. Upon study creation, the study coordinator can select the subject cohorts from the subject list, assigning unique study ids to each cohort. The subject list itself can expand as studies progress, and subjects can be assigned to multiple studies maintained in OpenClinica. Case report forms can be developed using a template provided by the OpenClinica developers.

The CVRG team is developing integrative clinical data management by combining OpenClinica, the Informatics for Integrating Biology and the Bedside (i2b2) system[15,16] and the Analytical Information Warehouse (AIW). The AIW development provides an infrastructure and methods for integrating and analyzing various types of clinical observations, therapeutic response and outcomes data, clinical reports, and omics and imaging data. The AIW development supports generation of data subsets having a specified schema, which can be analyzed with predictive tools, or used for query, visualization and analysis with external software systems such as i2b2.

Queries:

One of the most important capabilities requested by users of the CVRG is the ability to execute complex queries across different types of data. For example, design of the classifiers discussed above requires that subsets of data be retrieved for analysis. Figure 3 illustrates execution of a query generated on behalf of the Multi-Ethnic Study of Atherosclerosis MR Reading Center research group – "Determine subject IDs for all male patients ages 56 – 76 imaged using MR, and with isolated ST-T wave abnormalities". The query mediator currently in use on the CVRG first decomposes this query into parts. The mediator executes the query part "find all subjects imaged using MR" on the image data service. The mediator executes query part "find all male subjects ages 56 – 76" on the clinical data service. The mediator

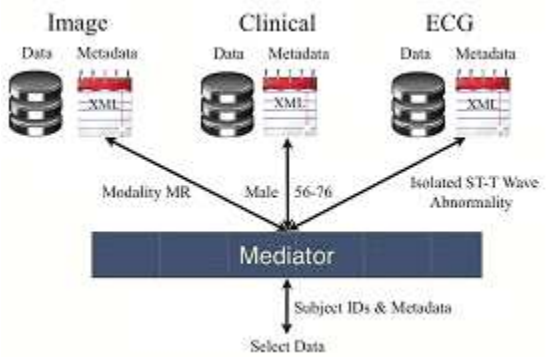

Figure 3: An illustration showing how federated queries are performed.

executes query part "find all subjects with isolated ST-T wave abnormalities" on the ECG data service. Data is linked across the data services by means of subject IDs. Upon execution of the query parts, the results are returned (subject ID and user-specified metadata) to the mediator. The mediator then determines the subject IDs that are common across the three query parts, and a list of these Subject IDs is displayed to the user on the CVRG Dashboard. Users may then select, for each or all subjects, the data to be downloaded. The user may retrieve and save the data on the local file system for further analysis.

Web and Portal Interfaces:

Using technologies developed by Google, Inc., we are able to quickly develop interfaces to the CVRG data and analytic services, tailoring these interfaces to user needs. A set of interfaces for interacting with different data and tools have been developed and collected together in the CVRG Dashboard. The Dashboard is built using Web 2.0 technologies (the Google Web Toolkit and Visualization API)[17,18] and provides a secure authentication and authorization mechanism.

Security:

The CVRG portal provides secure access to CVRG resources through a single login process. The CVRG uses security tools developed in the Cancer Biomedical Informatics Grid Project to authenticate users and control access to services[19]. Recently, these tools have been extended to support fine-grained control of data access.

Semantic Interoperability:

Semantic interoperability occurs when two or more entities are able to understand and interpret exchanged information properly. In order for both parties to fully understand each other, structured ontologies for describing data must be agreed upon.

This semantic description of data makes it easier for other investigators to re-use existing data for further analyses because study data and calculation results are described in a precise way. The CVRG tools use NCBO existing ontologies, and the CVRG team develops new ones (e.g., for annotating ECG data), to semantically describe data.

Conclusion

The CVRG Project makes a wide range of data and analytic services available, to help the cardiovascular research community in national, collaborative research projects. Advantages of the CVRG approach are: a) data access is rapid, eliminating the need to distribute hardcopies or to perform FTP download of individual data sets; b) data access is secure, is controlled by the individuals who collect the data, and conforms to institutional IRB requirements; c) the CVRG query capability allows heterogeneous data sets to be queried and explored to discover new relationships between data; and d) the CVRG software infrastructure may be used as a centralized data repository and analysis service, or in a fully distributed fashion with each study site retaining management of their data. The CVRG team believes these design features of the CVRG can help accelerate the process of biomedical knowledge discovery for the cardiovascular research community.

Acknowledgements

This work was supported by the National Heart, Lung, and Blood Institute award R24 HL085343. Users with interest in using CVRG tools should contact the project Principal Investigator, Raimond L. Winslow at rwinslow@jhu.edu, or the CVRG Project Manager, Stephen Granite at sgranite@jhu.edu. The CVRG web site is located at http://www.cvrgrid.org and the wiki site is at http://wiki.cvrgrid.org.

References

1. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. PLoS Med 2008;5:e183.
2. Grethe JS, Baru C, Gupta A, et al. Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud Health Technol Inform 2005;112:100-9.
3. Buetow KH, Niederhuber J. Infrastructure for a learning health care system: CaBIG. Health Aff (Millwood) 2009;28:923-4.
4. Oster S, Langella S, Hastings S, et al. caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J Am Med Inform Assoc 2008;15:138-49.
5. Musen M, Shah N, Noy N, et al. BioPortal: Ontologies and Data Resources with the Click of a Mouse. AMIA Annu Symp Proc 2008:1223-4.
6. Berntsen RF, Cheng A, Calkins H, Berger RD. Evaluation of spatiotemporal organization of persistent atrial fibrillation with time- and frequency-domain measures in humans. Europace 2009;11:316-23.
7. Chesnokov YC, Nerukh D, Glen RC. Individually adaptable automatic QT detector. Computers in Cardiology, 2006:337-340.
8. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics 2007;5:11-34.
9. Helm PA, Younes L, Beg MF, et al. Evidence of structural remodeling in the dyssynchronous failing heart. Circ Res 2006;98:125-32.
10. Beg MF, Helm PA, McVeigh E, Miller MI, Winslow RL. Computational cardiac anatomy using MRI. Magn Reson Med 2004;52:1167-74.
11. Younes L, Qiu A, Winslow R, Miller M. Transport of relational structures in groups of diffeomorphisms. J. Math Imag Vision 2008;32:41-56.
12. Ardekani S, Weiss RG, Lardo AC, et al. Computational method for identifying and quantifying shape features of human left ventricular remodeling. Ann Biomed Eng 2009;37:1043-54.
13. Seg3D: Volumetric Image Segmentation and Visualization. Scientific Computing and Imaging Institute (SCI).
14. OpenClinica Home Page. Available at: http://www.openclinica.org.
15. Gainer V, Hackett K, Mendis M, Kuttan R, Pan W, Phillips LC, Chueh HC, Murphy S. Using the i2b2 hive for clinical discovery: an example. Proceddings of AMIA Annual Symposium. 2008/08/13 ed; 2007:959.
16. i2b2: Informatics for Integrating Biology and the Bedside. Available at: https://www.i2b2.org/.
17. Google Web Toolkit Home Page. Available at: http://code.google.com/webtoolkit/.
18. Google Visualization API Home Page. Available at: http://code.google.com/apis/visualization/.
19. Langella S, Hastings S, Oster S, et al. Sharing data and analytical resources securely in a biomedical research Grid environment. J Am Med Inform Assoc 2008;15:363-73.