Poster Sessions

ID	Title of Abstract
7	Paraffin-Embedded Tissue Archive
8	Microsoft Excel caBIG™ Smart Client Joining the Fight Against Cancer
9	caAdapter (HL7 SDK)
10	Firebird (Federated Investigator Registry)
13	QARC's Digital Image Management System Supporting Cancer Clinical Trials
14	Leveraging caBIG™ Technology for Diseases Other than Cancer
16	Automated Peak Identification in a TOF-MS Spectrum
17	Janus/CRIX Pilot
20	The Mouse/Human Anatomy Ontology Mapping Project
21	The Challenge of Integrating caGrid in the Cancer Center Environment
22	Financial Billing Infrastructure Model for Cancer Centers
24	The Advanced Technology QA Consortium (ATC)
25	Placing Reactome into the caBIG™ Grid
26	A Web-Based Tissue Gene Expression Visualization Tool Developed on the caCORE Cancer Research Informatics Platform
27	CDS (Clinical Data System)
28	eRIC (The electronic Research Information and Compliance)
29	**Docu-MART**
30	caArray: a Standards-Based and caBIG™-Compliant Microarray Data Management System at the National Cancer Institute (NCI) Center for Bioinformatics
32	Grid Enablement of Protein Information Resource (gridPIR)
33	Initial Vocabulary and Common Data Element Compatibility Review Process
35	Cancer Clinical Central Participant Registry (C3PR) Adopter
36	Cancer Models Database 2.0
37	caBIG™ Data (CDE) Standards Development
38	VISDA: A caBIG™ Analytical Tool for Clustering and Beyond
39	RProteomics: An Application for Analyzing Mass Spectrometry Proteomics Data
40	cPath: Pathway Database Software for Systems Biology
42	Implementing a Learning Management System (LMS) to Maximize caBIG™ Training Benefits

ID	Title of Abstract
43	**CTMS Best Practices SIG**
44	Introducing caBIG™ Documentation and Training—Requirements, Processes, and Resources
46	caElmer—The MMHCC Pre-Clinical Trials Laboratory Information management system
52	A Validation Framework for Biological Pathways
53	A Diagnostic Test Suite for the RProteomics Statistical Routines
55	The CAP Cancer Checklist Model and Common Data Elements—An Emerging Standard in Tissue Banking and Pathology Tool
56	Translational Informatics Model for SPECS and SPORES
57	mzHDF for High-Throughput Proteomics Data
58	BreastCancerTrials.org: Feasibility of a caMATCH Tool to Match Breast Cancer Patients to Clinical Trials
59	BRIDG: A Model of the Shared Semantics of Clinical Trials Research
60	Sharing Cancer Microarray Data is Associated with Increased Citation Rate
61	Integration Strategies and Methods for Legacy Systems to be caBIG™ Silver Compatible
63	NCI-60 Dataset Evaluation for the caBIG™ Initiative
64	The Impact of caBIG™ Products on the Cancer Patient's Experience
65	A Grid-Enabled Array Annotator Implementation
66	Model Driven Design Approach to Extending Clinical Annotation Collection at the University of Pittsburgh
67	Model Driven Development of the caBIG™ CTMS Adverse Events Reporting System (caAERS) Prototype
68	Clinical Trials Object Model (CTOM)
70	Clinical Trials Laboratory Data: A Real-World Application of the BRIDG Model
71	The Disease Ontology and Browser
72	Scalable, Federated, caBIG™-Compatible Clinical Trials—Technology in Use
73	Cancer Experimental Protocol Assessment Tool: PADRE
83	The Past and Future of DWD in caBIG™
85	Clinical Research Information Exchange (CRIX) Initiative
86	Regulatory Requirements for Records and Signatures
87	InforSense LLC
88	Nutrition Ontology

Poster Abstracts

7. Paraffin-Embedded Tissue Archive

The Paraffin-Embedded Tissue Archive project is converting Surgical Pathology Reports (SPRs) for paraffin-embedded cancer tissue blocks from microfiche into a caTIES/De-ID-parsed format for uploading into the caTIES database. Through the caTIES query interface these tissue samples will be exposed to the caBIG™ community. To date, microfiche from 1948–1988 of Penn™ Paraffin related SPRs have been scanned and accession numbers manually keyed. An extensive quality assurance (QA) process is now underway. The first half of the QA, an analysis of the quality of the Optical Character Recognition (OCR) of these files is complete. More than 2,500 SPRs were examined; more than 85 percent of these files were deemed satisfactory for caTIES processing. The second half of the QA, determining how well caTIES can code these OCR SPRs, is progressing. A test set has been run through the Phase I version of caTIES. Once the caTIES Phase II beta is released, the same test set will be re-coded for a comparison of coding quality. The database schema will then be populated with all the paraffin files, coded through the caTIES pipeline and a link established between the database record and its corresponding SPR TIFF image for accessibility, only with proper IRB authorization.

AUTHOR(S): Michael Feldman, David Fenstermacher, M.D., Ph.D., Kevin Lux, Tara McSherry, Vishal Nayak, Casey Overby

AFFILIATION: Abramson Cancer Center, University of Pennsylvania

8. Microsoft Excel caBIG™ Smart Client Joining the Fight Against Cancer

For over a decade, Microsoft Office (MS) Excel has been the primary tool, especially in terms of statistical analysis and visualization, used by biomedical scientists for analyzing cancer research data. Scientists traditionally analyzed a limited set of data collected in their labs or made available by select collaborators. Through caBIG™, scientists will have access to data from researchers around the world. How will they analyze it to make meaningful deductions? We are developing extensions to MS Excel for accessing caBIG™ data-services. The proposed caBIG™ extensions are Windows Form Contextual GUIs that will construct appropriate caBIG™ XML queries and reformat the serialized XML responses as cells in the Excel Worksheet. Excel caBIG™ smart clients will be leveraging scientists' intimate familiarity with MS Excel by making caBIG™ data accessible to scientists in an intuitive manner. MS Excel runs on a wide range of Microsoft Windows operating systems including Microsoft Mobile 5.0. This translates into caBIG™ access on a wide range of computing devices including PDAs and mobile phones. This "data intimacy" allows on-the-spot hypothesis testing no matter where the spot is; thus replacing the proverbial "restaurant napkin sketches." http://xl-cabig-client.sourceforge.net/

AUTHOR(S): Katarzyna J. Macura, M.D., Ph.D., Robert T. Macura, M.D., Ph.D.

AFFILIATION: Department of Radiology and Radiological Sciences, Johns Hopkins University; Knowledge Enterprise

9. caAdapter (HL7 SDK)

caAdapter (HL7 SDK) The caAdapter is an open source tool set that facilitates HL7 version 3 message building, parsing, and validation based on specific message definitions. It has the capability to perform vocabulary validation and integrates with NCICB caCORE components. It also provides a mapping tool component that has a drag-and-drop GUI interface for mapping clinical data to HL7 version 3 specifications. This supports HL7 version 3 messaging as part of the caBIG™ solution. The caAdapter will facilitate and support the exchange of all HL7 version 3 messages.

AUTHOR(S): Christo Andonydis, Ph.D., M.S., Sichen Liu, M.S., Eric Chen

AFFILIATION: National Cancer Institute Center for Bioinformatics, National Institutes of Health

10. Firebird (Federated Investigator Registry)

Firebird (Federated Investigator Registry) Firebird is the first module implemented to achieve the vision of the Clinical Research Information Exchange (CRIX) infrastructure. Firebird will automate the submission of the FDA Form 1572 and enable investigators to register online with NCI and other sponsors, including pharmaceutical companies, thus removing paper-based latencies and infrastructure costs and allowing investigators to maintain and manage 1,572 registrations. Firebird will leverage legally enforceable digital signatures compliant with Title 21 Regulations using an Identity Assurance infrastructure, Secure Access for Everyone (SAFE).

AUTHOR(S): Brenda Duggan, Sichen Liu, M.S., Kamal Narang

AFFILIATION: National Cancer Institute Center for Bioinformatics, National Institutes of Health

13. QARC's Digital Image Management System Supporting Cancer Clinical Trials

QARC is an active data management organization with a clinical service and interactive mission. The key objective is to insure the quality of the data and compliance to study for the clinical cooperative groups in order to secure a uniform study population for outcome analysis. Improving study compliance and providing uniform data format promotes the validation of study endpoints. Since QARC interacts with investigators on six continents, the data arrives at QARC in many electronic formats, both with and without proprietary headers, often creating unique issues in opening and archiving data. In order to accommodate the heterogeneity of this data, QARC has developed a sophisticated image management system encompassing a complex patient database, DICOM image transfer software, and a fully-functional PACS archive. The QARC poster will depict the current image management system with specific emphasis on our current project to bring this system onto a Web platform. This Web viewer will allow radiologists and radiation oncologists to view objects in the database remotely for protocol review and resolution of discrepancies in eligibility, staging, and response interpretation.

AUTHOR(S): Nicole Laprise, Richard Hanusik, Fran Laurie, T. J. FitzGerald, M.D.

AFFILIATION: University of Massachusetts Medical School

14. Leveraging caBIG™ Technology for Diseases Other than Cancer

The SLE Biomarkers Consortium is a cross-sector, multi-institutional effort to identify, develop, and validate biomarkers for Systemic Lupus Erythematosus. SLE is a relatively rare (~250,000 patients in the United States) multisystem disorder affecting mainly women of child-bearing age. It has diverse clinical and laboratory manifestations, waxes and wanes over time, and may involve essentially any organ system. The consortium goals are to identify biomarkers of lupus disease activity, to validate them so that they can be used by the FDA as indicators in clinical trials, and to facilitate the development of new safe and effective treatments for SLE. To accomplish these goals, information technology is a key component. Using the caBIG™' EVS functionality, we are able to normalize the vocabulary across multiple research centers in order that we might create standardized data for a community database. This standardized data can then be relationally queried, allowing users to leverage the data from the entire community in a consistent and meaningful manner. Using the caBIG™' caBIO technology, we are able to demonstrate how biological information for things such as SNPs and genes can be retrieved and leveraged by SLE researchers.

AUTHOR(S): Joel Dubbels, Elise Blaese

AFFILIATION: IBM

16. Automated Peak Identification in a TOF-MS Spectrum

The high-throughput capability, and capability of providing a variety of chemical and structural information, have made mass spectrometry a standard tool for proteome and biomarker discovery research using bio-fluid samples such as blood serum and digested protein samples. A bottleneck is that modern mass spectrometers (e.g., MALDI/SELDI),

as a result of improved resolving power, produce very large raw data sets containing hundreds of peaks for these complex biological samples. Hence these spectra must be preprocessed to identify mass peaks from background noise for further analysis. The accuracy of the mass assignment and the alignment among spectra from different subject groups can be another limitation. We have developed an automated peak picking algorithm based on a maximum likelihood approach that effectively and efficiently detects peaks in a TOF-MS spectrum. The algorithm takes into account the underlying noise characteristic and produces maximum likelihood estimates of peak positions and amplitudes. It also simultaneously develops estimates of the uncertainties in each of these quantities. Using the estimated peak positions and their uncertainties, we can align different spectra more accurately than using a few known calibrants. This precise peak summary is crucial for further multivariate analysis and classification, as will be discussed.

AUTHOR(S): Haijian Chen, Eugene R. Tracy, William E. Cooke, Maciek Sasinowski, Dennis M. Manos, Michael W. Trosset, O. John Semmes

AFFILIATION: Department of Physics, College of William and Mary

17. Janus/CRIX Pilot

NCI, FDA, CTIS, and IBM have been conducting an effort to develop a pilot for a clinical trial data facility based on JANUS for use by the FDA for review and analysis and by the NCI to help address a set of the CRIX objectives, including: More efficient submissions based on standards; -Reproducible, custom datasets for analysis; -Reusable tools for analysis and review; -Less data redundancy; -Less time on orientation to data by reviewers; -Less ambiguity in communications of information; -More auditable data; -Less manual, paper processing; -Use of common standards across the entire community (government, industry, academia); and Interoperability with other caBIG™ data sets, tools, capabilities. We have demonstrated an implementation of JANUS under a very flexible set of architectures, including both Oracle and DB2, supporting various ways to access and query the data (JDBC, ODBC, caCORE/CSM). We populated the JANUS repository with sample synthesized human trial data, related to two Iressa/Taxotere oncology studies, furnished in a standard STDM format. We developed a sample application demonstrating the analytical capabilities of the underlying JANUS repository utilizing a caCORE framework to access the data.

AUTHOR(S): Joel Dubbels, Timothy Bishop, Kamal Narang, and Elise Blaese

AFFILIATION: IBM and Capital Technology Information Services, Incorporated

20. The Mouse/Human Anatomy Ontology Mapping Project

Mouse Genome Informatics (MGI) and the National Cancer Institute (NCI) have developed extensive anatomy ontologies for the adult mouse and human to facilitate the standardized description and integration of data for the respective species. The objective of the Mouse/Human Anatomy Ontology Mapping Project is to harmonize and map anatomical descriptors used for the mouse and human. To achieve this goal, we: (1) evaluated ontology tools in terms of their utility in comparing the ontologies and providing mappings, and identified the tools best suited for this work, (2) performed an in-depth comparison of the ontologies using the tools selected, (3) extended and harmonized the existing mouse and human anatomical ontologies, and established mappings between them; and (4) developed procedures for updating and maintaining the Adult Mouse Anatomy (MA) ontology, including the mouse-human mappings, in the NCI Thesaurus. Ultimately, this work will enable closer integration of basic science research and clinical data pertinent to cancer and other human diseases, augmenting the use of the laboratory mouse as a model for pre-clinical cancer research, and accelerating the translation of basic research discoveries into new clinical therapies.

AUTHOR(S): T. F. Hayamizu, S. De Coronado, G. Fragoso, N. Sioutos, C. Hampel, F. Hartel, J. A. Kadin, M. Ringwald

AFFILIATION: Mouse Genome Informatics, The Jackson Laboratory and National Cancer Institute Center for Bioinformatics

21. The Challenge of Integrating caGrid in the Cancer Center Environment

The service-oriented integration approach is quickly changing the way applications can be integrated. In the caBIG™ world, caGrid will provide a preferred environment for delivering services by coordinating independent cancer center resources. However, caGrid also introduces challenges that cancer centers might never have faced before, which may significantly impact their tools and integration strategies. In this poster, we address the issues faced by Kimmel Cancer Center in its participation in several caBIG™ projects. Also, we will describe a toolkit, the MBean Service Toolkit (MST), which we are developing to facilitate our future project integration in the context of service-oriented framework, such as caGrid. We hope our current integration path at Kimmel Cancer Center can be helpful to other cancer centers with their own caGrid adoption plans.

AUTHOR(S): Scott Li, Karl Smalley, Jack London

AFFILIATION: Kimmel Cancer Center

22. Financial Billing Infrastructure Model for Cancer Centers

caBIG™ Clinical Trials Financial Billing Architecture: A Multi-Tiered Model Approach. The CTMS Financial Billing Special Interest Group will present what it has mapped to be the outline for Financial Billing in a Cancer Clinical Trials Environment. A workflow representation for Financial Billing will demonstrate our approach in defining this model and its components. Clinical Trials interactivity with Financial Billing will also be outlined in the workflow component of this model.

AUTHOR(S): Sorena Nadaf, Andrew Winter, Jieping Li, M.B.A., Jill Kuennen, Renee Webb

AFFILIATION: Vanderbilt, Northwestern, Georgetown, Iowa

24. The Advanced Technology QA Consortium (ATC)

The Advanced Technology QA Consortium (ATC), an NCI sponsored organization, combines the efforts of the nation quality assurance (QA) organizations responsible for radiation therapy QA for cooperative group clinical trials. ATC includes the Quality Assurance Review Center (QARC), Radiological Physics Center (RPC), American College of Radiology/Radiation Therapy Oncology Group (RTOG), Image Guided Therapy QA Center (ITC), and the Resource Center for Emerging Technologies (RCET). ATC™ primary objective is to provide a forum to develop infrastructure for uniform data acquisition and credentialing strategies for cooperative group clinical trials. The strategy facilitates digital data exchange and data acquisition between site investigators and QA centers, transparent to cooperative groups. ATC has developed a data acquisition platform with Web-based QA review tools, which has been used in support of 12 advanced-technology RTOG protocols. This platform has been replicated at QARC as a pilot initiative for uniform data acquisition among all cooperative groups for radiotherapy clinical trials. It provides the infrastructure to perform QA review and to be able to adjust objects in a timely manner to maintain compliance with protocol guidelines, thus improving uniformity of the protocol study set. A more recent goal is to make ATC platforms compatible with objectives of caBIG™.

AUTHOR(S): James A. Purdy, Ph.D., Walter R. Bosch, D.Sc., Bill Straube, M.S., John Matthews, D.Sc.

AFFILIATION: Image-Guided Therapy QA Center, Washington University, St. Louis, Missouri

25. Placing Reactome into the caBIG™ Grid

Reactome is a curated database of human biological pathways and reactions. The information in this database is authored by biological experts, maintained by the Reactome editorial staff, and cross-referenced with PubMed, Gene Ontology, UniProt, OMIM and other databases. Reactome provides high values to bench and computational biologists as well as biological students because of its high quality, rich contents and sophisticated data model. The ICR Reactome Database Sharing project (caBIG™-ICR-04-09-01) implemented Reactome as a data feed to caBIG™. To accommodate rich structured pathway and reaction data in the Reactome database, we extended the caBIG™ data model, implemented this extension

in Java and XML Schema. To facilitate semantic exchange, a set of CDEs (common data elements) for this extension has been created and loaded into caDSR. To place Reactome into the caBIG™ grid, a bunch of SOAP based Web services APIs (application programming interfaces), which are described in a standard WSDL file, have been created and deployed in a public Web services server. To provide compatibility with other pathway tools and databases, pathways and reactions can also returned from Web services in the BioPAX level 2 format, an OWL based pathway exchange format. For complete details, please refer to http://www.reactome.org: 8080/caBIOWebApp/docs/services.html.

AUTHOR(S): Guanming Wu

AFFILIATION: Cold Spring Harbor Laboratory

26. A Web-based Tissue Gene Expression Visualization Tool Developed on the caCORE Cancer Research Informatics Platform

The National Cancer Institute has developed an informatics infrastructure called caCORE to facilitate data and application sharing among cancer research institutes. We have evaluated it in our data warehouse project and have established the need for compatibility for tools development, to share data with external organizations, and also as a gateway to access data in the public domain. The gene expression visualization tool we developed is a Web-based application with support of caCORE 3.0 infrastructure. The tool runs on a Java Server Page platform and is designed as a Web component that can be easily embedded in an Internet application for visualizing tissue expression profiles of one or a number of genes. The tool accepts one or a group of gene accession numbers as input and displays their expression levels in various tissues as a graphical matrix. The tool also lists tissues with associated gene accession numbers and expression levels in a separate table allowing users to retrieve pertinent information of genes for annotation. Our effort demonstrates the usefulness of caCORE infrastructure and how caCORE can be used by cancer researchers as a data source and developmental environment for cancer research tool development.

AUTHOR(S): Long Qu2, Hai Hu*, Richard Xiong, Henry Brzeski, Mandy Raab, and Michael N. Liebman
*Corresponding author.

AFFILIATION: 1) Windber Research Institute, Windber, PA. 2) The Children's Hospital of Philadelphia, PA. 3) Boston Medical Center, Boston, MA. 4) Saint Vincent College, Latrobe, PA.

27. CDS (Clinical Data System)

The Clinical Data System (CDS) is an independent and stand alone data submission infrastructure (electronic) at NCICB to serve as the primary data resource for NCI sponsored clinical trials. Data is submitted either as an electronic file submission or via a Web-based interface. The system also provides a mechanism to access data to stakeholders including cancer centers, cooperative groups, and single institutions via a data analysis interface. This interface enables users to view and generate reports about various aspects of the clinical trial process. For each protocol, the lead Group or Institution is responsible for submitting CDS data. Data submitted for protocols are either abbreviated or complete. The Abbreviated CDUS Data Set is limited to protocol administrative and patient demographic information. The Complete CDUS Data Set contains the information found in the Abbreviated CDUS Data Set, patient administrative information (e.g., registering institution code, patient treatment status), treatment information (e.g., agent administered, total dose per course), Adverse Event information (e.g., AE type, grade), and response information (e.g., response observed, date response observed). All data submitted are cumulative.

AUTHOR(S): Suranjan De

AFFILIATION: National Cancer Institute Center for Bioinformatics

28. eRIC (The electronic Research Information and Compliance)

The electronic Research Information and Compliance (eRIC) is a collaborative effort of Capital Technology Information Services, Inc. (CTIS, Inc.), Georgetown University, and MedStar Research Institute. eRIC was designed to build a research and

regulatory portal that supports collaboration and information sharing among the Institutional Review Board (IRB), researchers, sponsors, and the public. In essence, eRIC portal is a virtual one-stop-shop for researchers, IRB members and staff, where these stakeholders can go to obtain research and regulatory information, and to actively share, collaborate or complete administrative and regulatory tasks and processes. The eRIC solution is envisioned to be a model for the integration of human research protection throughout a comprehensive Clinical Trials Research and Management (CTRM) solution by providing enhanced intra and inter communications for those in the clinical research enterprise in order to expand their capabilities and further streamline the operation of the IRBs. eRIC provides a single-source collaboration platform supported by low-cost administrative capabilities that greatly improves the efficiencies of regulatory activities and research practices within an institution. The IRB members and staff and the research community are able to make virtually all of their service selections from a single source Web-based portal. This collaboration platform helps sustain and maintain a decreased workload, paperwork, and redundancy while limiting human errors with diminished liability. By accessing eRIC, IRBs are able to streamline meeting logistics, update and monitor capabilities conforming to all compliances and regulations, while maintaining an audit trial throughout the entire process. eRIC technical architecture confirms to the caBIG™ compliance guidelines and aims at achieving the Silver Level compliance.

AUTHOR(S): Kolaleh Eskandanian, Rachel Kidwiler, Yash Sowale

AFFILIATION: Center for Technology Information Services, Incorporated; Georgetown University; MedStar Research Institute

29. Docu-MART

Docu-MART is a system of software applications utilizing both desktop and Web technologies developed to assist in the authoring, review, and tracking of clinical trial protocol documents. Protocols can be authored using predefined structured protocol representations using XML templates. The application is planned to assist in reducing the administrative burden of protocol development so that investigators can focus on scientific integrity. Docu-MART will help increase the efficiency of clinical trial document development and approval by leveraging the use of standard templates, standard document structure, auto content generation, online reviewing & approval, automated notifications and tracking of the document through its lifecycle. Docu-MART is a collaborative effort between the NCI/ CTEP (National Cancer Institute/ Cancer Therapy Evaluation Program, CALGB (the Cancer and Leukemia Group B), ECOG (Eastern Cooperative Oncology Group), and SWOG (Southwest Oncology Group). The technology is being developed by Capital Technology Information Services, Inc. (CTIS) on behalf of NCI/CTEP. Docu-MART effort has been involved with the Structured Protocol Representation, a SIG under the BRIDG effort of caBIG™ for providing the elements of a protocol. Docu-MART solution can be adapted to caBIG™ requirements and complement the BRIDG activities under CTMS workspace for protocol authoring and tracking.

AUTHOR(S): Sudhir Raju

AFFILIATION/INSTITUTION: Cancer Therapy Evaluation Program

30. caArray: A Standards-Based and caBIG™-Compliant Microarray Data Management System at the National Cancer Institute (NCI) Center for Bioinformatics

caArray database is a standards-based, Web-accessible open source data management system. It has been developed following the caBIG™-compatibility guidelines that highlight the use of controlled vocabularies, common data elements (CDEs), well-documented APIs and UML model. It is an open-development, open-source, and open-access system that seeks to facilitate cross system, cross platform sharing of cancer research data. CDEs for caArray are stored in the NCICB™ publicly accessible metadata repository, caDSR; a connection to the caCore™ Enterprise Vocabulary Services (EVS) will be available. The adoption of caDSR and EVS further ensures cross system interoperability of caArray. caArray features MIAME 1.1 compliant data annotation, controlled vocabularies (MGED ontology), and MAGE-ML import and

export. It currently accepts submission of Affymetrix and GenePix and ImaGene native data files; support to more platforms will be implemented in the near future. In addition to document based data submission, caArray provides application programming interfaces (APIs) for programmatic data access. Several application tools which retrieve data from caArray using this API are currently under development, among these are: caWorkbench, a desktop tool for analysis, annotation and visualization of microarray data and webGenome, an application that allows users to view DNA copy number measurements relative to genome locations and annotated genome features. Both of these tools also connect to the NCICB™ cancer Bioinformatics Infrastructure Objects (caBIO) model, permitting access to a variety of genomic, cancer models, and clinical trials information.

AUTHOR(S): X. Bian, A. Basu, J. Lorenz, J. Zhou, H. Chen, J. Eads, T. Borja, D. Addepalli, S. Settnek, S. Guruswami, S. Madhavan, G. Fragoso, K. Buetow, Ph.D.., P. Covitz, Ph.D., M. Heiskanen

AFFILIATION: National Cancer Institute Center for Bioinformatics

32. Grid Enablement of Protein Information Resource (gridPIR)

The Protein Information Resource (PIR) is an integrated bioinformatics resource that provides protein databases and analysis tools to support genomic and proteomic research. As a participant in the Integrative Cancer Research (ICR) workspace of caBIG™, PIR developed gridPIR in collaboration with the adopter from the Biomedical Informatics Facility (BMIF) for the Abramson Cancer Center at the University of Pennsylvania. gridPIR is one of four reference projects for demonstrating how a caBIG™ Silver compliant data source can be discovered and consumed on caGrid. gridPIR was developed using a Model Driven Architecture (MDA) approach and employs an n-tier architecture. A data layer, supported by Oracle 9i, stores the UniProt Knowledgebase (UniProtKB). The object layer contains 48 domain objects with 51 attributes registered to caDSR, as required for caBIG™ semantic interoperability. A data access layer utilizing Hibernate provides the mapping between objects and the relational database. The gridPIR API exposed to caGrid was generated using caCORE SDK 1.0.3.1. Upon successful completion on August 1st, 2005, gridPIR became one of the first caBIG™ Silver compliant data services on caGrid 0.5. In compliance with caBIG™ compatibility guidelines, gridPIR will continue to improve its services to better serve as the central genomic and proteomic information resource for cancer research within caBIG™.

AUTHOR(S): Developer: Baris Suzek, M.S., Hongzhan Huang, Scott Chung, Hsing-Kuo Hua, Peter McGarvey, Cathy H. Wu; Adopter: Craig Street, Casey Overby, David Fenstermacher M.D., Ph.D.

AFFILIATION: Developer: Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, Washington, DC; Adopter: Biomedical Informatics Facility, University of Pennsylvania, Philadelphia, PA

33. Initial Vocabulary and Common Data Element Compatibility Review Process

Elements of a Compatibility Review process are described which support the reuse of well defined Common Data Elements (CDEs) that can be used to join data sets across institutions and systems. The Compatibility Review process aims to be consistent, transparent and scalable. The Silver Level Compatibility Review is performed to ensure consistency between UML models and CDEs stored in the NCI cancer Data Standards Repository (caDSR). The review focuses on well-specified data elements that can be used for joining across data sets. A number of items in the Compatibility Review can be automated via the Semantic Integration Workbench and UML Loader, which improves the efficiency of the review. The moderately difficult process of ensuring a properly constructed UML model remains (i.e., proper classes, attributes, associations, directionality, and multiplicity). The most challenging aspect of the review is ensuring reuse of appropriate CDEs and vocabularies already in the caDSR. There is a need for software tools to support proper CDE reuse. Initial reviews highlighted the need to develop a review reporting tool that improves the consistency, transparency, and scalability of the Compatibility Review process. A prototype tool that combines information from the CDEs and UML model is under development.

AUTHOR(S): Lewis Frey, Ph.D., and Robert Freimuth, Ph.D.

AFFILIATION: Vanderbilt University and Washington University in St. Louis

35. Cancer Clinical Central Participant Registry (C3PR) Adopter

C3PR is a component of NCI's Cancer Centralized Clinical Database System (C3DS). It is an open source Web-based application that provides a central repository for participant information across studies, sites, systems, and organizations. It also has a direct interface to the clinical trial management systems, such as C3D (Cancer Centralized Clinical Database). University of California, Irvine (UCI) and Georgetown University (GU), two caBIG™ participating cancer centers who have previously adopted C3D in caBIG™ Year One, are pioneering in adopting C3PR in Year Two. This poster will outline the C3PR adoption experiences and strategies of these two cancer centers, including software requirement gathering, system configuration and integration, deployment, data entry, and reporting.

AUTHOR(S): Jieping Li, M.B.A., Yi-Chen(Andrea) Hwang, M.S., M.B.A.

AFFILIATION: Georgetown University, University of California, Irvine Chao

36. Cancer Models Database 2.0

The NCI Center for Bioinformatics (NCICB) and NCI™ Mouse Models of Human Cancers Consortium (MMHCC) released a new version of the Cancer Models Database (caMOD) in December 2005. Cancer models that recapitulate many aspects of the genesis, progression, and clinical course of human cancers are valuable resources to cancer researchers engaged in a variety of basic, translational, clinical, and epidemiological investigations. caMOD is a Web-based resource that provides information about rodent models for human cancer to the research community. The redesigned version of caMOD is targeted for caBIG™ silver-level compliance according to the caBIG™ compatibility guidelines for creating and adopting software systems. The Cancer Models Database provides the following key capabilities to its users: Data Submission—Data in caMOD are extracted from scientific literature by curators and verified by the scientists who generated or worked with the models, or they are directly submitted by scientists themselves. Cancer Model Search—Users can retrieve information about the making of models, their genetic descriptions, histopathology, derived cell lines, associated images, carcinogenic interventions, microarray data, and therapeutic trials in which the models were used. caMOD provides links to PubMed for associated publications and other resources, such as mouse repositories, detailed information about altered genes, pathways affected, and information about human clinical trials that utilize the same compounds as the pre-clinical trials in animal models. System Function Administration—The Administrative function provides services for user registration, review of submitted models, and database management. The new features allow closer connections to the resources made available by other institutions such as the Jackson Laboratory and the Developmental Therapeutics Program at NCI. The application is available at http://cancermodels.nci.nih.gov

AUTHOR(S): Guruswami S2, Wagner U2, George D2, Pandya S2, Schroedl N2, Piparo M2, Steward D2, Kong D2, Luo Z2, Jill Hadfield, M.S., Marks C1, Tarnowski B1, Heiskanen M1

AFFILIATION: 1) National Cancer Institute Center for Bioinformatics 2) Science Applications International Corporation

37. caBIG™ Data (CDE) Standards Development

Over the past year, the caBIG™ Vocabulary and Common Data Element workspace has facilitated the review, adoption and deployment of data standards. caBIG™ success is dependent on use of these data standards in the development of caBIG™ models and applications. Reuse of standard common data elements (CDEs) by developers is the key to semantic and syntactic interoperability, and has the additional benefit of time savings for modelers and developers. Ten standards (representing over 100 data elements) have been reviewed and are either approved as caBIG™ standards or are in the process of being reviewed by the caBIG™ commu-

nity. The initial standards focused on the patient demographics domain, designed to standardize a set of commonly used patient attribute information (race and ethnicity, sex and gender). More recently, domain-specific standards (such as Gene Identifier and Unit of Measure) have been proposed by the caBIG™ domain workspaces. A current list of CDEs in the data standards process can be found by using the CDE Browser (http://cdebrowser.nci.nih.gov/CDEBrowser/). Data Standards are registered in the caBIG™ context under Classifications in the Browser tree and presented by registration status of Proposed, Candidate, and Standard. The caBIG™ Web site also contains information on the data standards (https://caBIG.nci.nih.gov/workspaces/VCDE/Data_Standards/index_html).

AUTHOR(S): Brian Davis, Ph.D., Michael Keller, Kathleen Gundry, Tommie Curtis, Mary Cooper, Brenda Maeske

AFFILIATION: Vocabulary and Common Data Element (VCDE) Workspace

38. VISDA: A caBIG™ Analytical Tool for Clustering and Beyond

VISDA (Visual Statistical Data Analyzer) is a caBIG™ analytical tool for cluster modeling, visualization, and discovery. Being statistically-principled and visually-insightful, VISDA exploits the human gift for pattern recognition and allows users to discover hidden clustered data structures within high dimensional and complex biomedical data sets. The unique features of VISDA include its hybrid algorithm, robust performance, and tree of phenotype. With global and local biomarker identification and prediction functionalities, VISDA allows users across the cancer research community to analyze their genomic/proteomic data to define new cancer subtypes based on the gene expression patterns, construct hierarchical trees of multiclass cancer phenotypic composites, or to discover the correlation between cancer statistics and risk factors. VISDA is being used in several ongoing projects at the Wistar Institute, for cancer diagnostic studies, and has been useful for sub-class identification and outlier predictions. Their experiences with VISDA have led to a number of suggestions for enhancing functionality. Implementing the ability to reselect genes at different node levels rather than using those genes selected on the whole data set is likely to be important for discovering new clusters that might be missed if gene selection is fixed at the first step. Another important function to be implemented is the ability to use different methods for the calculation of the optimal number of genes needed for class distinction, as this may be different at different levels of node dissection.

AUTHOR(S): Huai Li1, Jiajing Wang1, Yitan Zhu1, Jason Xuan1, Robert Clarke1, Yue Wang1, Malik Yousef2, Michael Nebozhyn2, Michael Showe2, Louise Showe, Ph.D., 2

AFFILIATION: 1) Computational Bioinformatics and Bioimaging Laboratory, Virginia Tech, and Lombardi Cancer Center, Georgetown University; 2) The Wistar Institute

39. RProteomics: An Application for Analyzing Mass Spectrometry Proteomics Data

In this poster, we describe the RProteomics application, which comprises a set of low-level and high-level statistical methods for analyzing MALDI and SELDI MS and LC-MS mass spectrometry proteomics data. Currently available routines include background removal, denoising, normalization, alignment, and calibration. The R analytics are exposed to Java using our in-house Open Statistical Systems package, which marshals data objects between Java and R using the StatML XML encoding. The Java layer handles large proteomics experiment files using Mako, an XML databases layered on MySQL, and binary files. Experiment metadata is encoded in ScanFeatures XML and stored in Mako, while the actual mass spectrometry scan data is stored as arrays in binary files. The analytical and data routines are exposed via silver level Web services using the Axis toolkit. The analytical routines are also exposed as grid services using the caGrid 0.5 toolkit. The entire system can be accessed directly by experts using the R console or by any end-user via a graphical user interface plug-in to caWorkbench. The GUI allows users to load data into the system, search for existing data, run analytical routines on single or multiple mass spectrometry scans, and plot data in 2D stacked plots.

AUTHOR(S): Patrick McConnell, Richard Haney, Salvatore Mungal, Mark Peedin, David Fenstermacher, M.D., Ph.D., Kimberly Sellers, Craig Street, Shannon McWeeney, Veena Rajaraman, Ted Laderas, Simon Lin, Christoph Borchers

AFFILIATION: Duke University, University of Pennsylvania, Oregon Health and Science University, Northwestern University, University of North Carolina

40. cPath: Pathway Database Software for Systems Biology

cPath is open-source pathway database software that eases data integration from multiple sources. It currently supports the BioPAX pathway exchange language and the PSI-MI protein interaction exchange format. It can be locally installed and can connect with Cytoscape for network visualization and analysis.

AUTHOR(S): Ethan Cerami, Gary Bader, Benjamin Gross, Chris Sander

AFFILIATION: Memorial Sloan-Kettering Cancer Center

42. Implementing a Learning Management System (LMS) to Maximize caBIG™ Training Benefits

To effectively manage training programs offered as part of the caBIG™ initiative, NCICB is implementing Training Partner, a Learning Management System (LMS). This system will enable the caBIG™ Training Workspace to set up and maintain a central portal for publicizing training programs, registering and communicating with trainees, recording progress, and evaluating course/program success. Working in close collaboration with the caBIG™ Training Workspace, the LMS implementation team has been setting up the Training Partner system using the caCORE curriculum as a pilot test program. This has involved defining appropriate terminology and use cases, evaluating alternatives, setting policies and procedures, and planning for long-term growth. Tracking program development, training, and acceptance enables the caBIG™ Training Workspace to provide quality customer service and content to the community's trainees as well as important feedback to its application developers.

AUTHOR(S): Kim Diercksen, Calla Pearce

AFFILIATION: caBIG™ Training Workspace

43. CTMS Best Practices SIG

The CTMS Best Practices SIG was founded to focus on establishing and fostering relevant clinical trial management best practices for clinical trials within the caBIG™ community. The identification and possible adoption of Best Practices are presented to the group from member experiences or developed as a combination of proven techniques or processes. The BP SIG is the parent group for the SOP Working Group, which continues the work started by the C3DS User Group. The SOP WG developed 18 technology neutral SOPs around the CTMS applications that would meet regulatory requirements for caBIG™ adopters. These active SOPs and its CBT module are available on the caBIG™ Web site. Another set of SOPs will be completed and delivered for training in May 2006. The BP SIG is identifying other domains of interest, such as CRF Management, mentoring caBIG™ adopters, Participant Registry, Informed Consent Workflow, Assuring Compliance with Protocols, and Defining Metrics for Clinical Research Trial Management. As the SIG identifies the need within these new areas of interest, other Working Groups will be established under this SIG. Participation and harmonization among the other SIGs and workspaces is essential to ensuring proper coverage and accuracy of the Best Practices developed by this SIG.

AUTHOR(S): Brenda Duggan, Andrea Hwang, M.S., M.B.A., Jieping Li, M.B.A., David Geismar, Hadley White, Michele Pontinen, Patrice Clark

AFFILIATION: National Cancer Institute Center for Bioinformatics, Georgetown University, University of California, Irvine Chao, Booz Allen Hamilton

44. Introducing caBIG™ Documentation and Training—Requirements, Processes, and Resources

This poster will introduce the guidelines, templates, and support resources that the Training Workspace has developed related to documentation and training. It will provide an overview of the documentation and training requirements and templates, and where they can be found; it will introduce the documentation/ training mentoring program; and

it will outline the steps involved with registering a project, requesting a documentation/training mentor, and submitting documentation and training modules for review.

AUTHOR(S): Jennifer Tucker, Ph.D., M.S., PMP, Leslie Derr, Ph.D.

AFFILIATION: Booz Allen Hamilton and National Cancer Institute

46. caElmer—The MMHCC Pre-Clinical Trials Laboratory Information Management System

The NCI Mouse Models of Human Cancers Consortium (MMHCC) has been given the task of studying and validating mouse models of human cancers. In preparation for pre-clinical trials, a Pre-clinical Trials Working Group was appointed by the MMHCC II Informatics Committee and tasked with developing a Pre-Clinical Trials Data Assessment Survey, which showed that the majority of the laboratories did not store their data in a retrievable digital format. The Pre-Clinical Trials Data Assessment Survey clearly showed that the majority of the MMHCC II-related laboratories do not have the necessary infrastructure to adequately store, retrieve and display the trials data in an electronic format. To remedy these infrastructure flaws, we are creating the MMHCC Preclinical laboratory information management system. The system will: a)Track individual mice through linkage with existing colony management systems b)Track individual studies, and experiments, experimental data c)Accurately model the different types of data sets collected in the individual studies. 2)Facilitate the upload of completed data sets to node repositories for sharing within the caBIG™ community.

AUTHOR(S): Jose Galvez, M.D.

AFFILIATION: University of California, Davis

52. A Validation Framework for Biological Pathways

Biological Pathway Exchange (BioPAX) is an ontology developed to describe biological pathway data. It uses OWL (Web ontology language), which is part of the W3C effort with regard to the semantic Web. We have developed a Java based tool, the BioPAX validator, that validates any pathway data expressed in BioPAX. The BioPAX ontology, which is expressed in OWL, also captures comments for each BioPAX object as defined in BioPAX. Those comments are best practices principles relating to how best to map each BioPAX object over any given pathway dataset. Our Java based validator will follow these principles to validate any given BioPAX pathway data. The current release supports Level 1 and Level 2 of the BioPAX data standard and is available for download at: http://hpc.ohsu.edu:88/biopaxvalidator/

AUTHOR(S): Vincent Yau (1), Ranjani Ramakrishan, M.S., (1), Shannon McWeeney (1,2,3)

AFFILIATION: (1)Informatics Shared Resource, Oregon Health and Science University Cancer Institute, (2) Biostatistics Shared Resource, Oregon Health and Science University Cancer Institute, (3) Division of Biostatistics, Department of Public Health and Preventive Medicine

53. A Diagnostic Test Suite for the RProteomics Statistical Routines

We present a statistical testing suite developed as part of our test strategy as caBIG™ adopters of the RProteomics statistical routines developed by Duke Comprehensive Cancer Center. RProteomics was developed in the open source statistical programming environment R, and consists of multiple processing routines for each step of the low-level analysis of proteomics data. Our test framework enables users to select the optimum processing routine for a single step of analysis by two software components developed in-house. The first component enables batch submission of a single dataset through multiple routines within multiple steps of analysis. Routines and associated parameters can be edited and selected through exchangeable pipeline files. The second component consists of metrics (developed both in-house and culled from current literature) that allow us to assess the impact of specific RProteomics low-level analysis routines on any dataset (e.g., evaluation of Analysis of Variance (ANOVA) tables before and after normalization). The diagnostic suite was also developed in the R environment, allowing it to be easily adapted to assess the impact of other routines in addition to RProteomics. We believe that our framework is an excellent example of added value that can arise

from successful Adopter/Developer collaboration within the caBIG™ framework.

AUTHOR(S): Ted Laderas, M.S., (1), Solange Mongoue-Tchokote (2), Veena Rajaraman (1), Kimberly Sellers (3), Shannon McWeeney (1,2,4)

AFFILIATION: 1.Informatics Shared Resource, Oregon Health and Science University Cancer Institute, Oregon Health & Science University, 2. Biostatistics Shared Resource, Oregon Health and Science University Cancer Institute, Oregon Health & Science University, (additional below)

55. The CAP Cancer Checklist Model and Common Data Elements—An Emerging Standard in Tissue Banking and Pathology Tool

Background: An important aspect of caBIG™ pilot is the development of consistent practices for data standards development, using a multi-tier approach that facilitates semantic interoperability of systems. The Tissue Banking and Pathology Tools (TBPT) Workspace has identified the College of American Pathologists (CAP) cancer checklists as one important standard set of standardized data elements in our domain. To our knowledge, there have been no previous attempts to create CAP protocols as ISO/IEC 11179 compliant CDEs. Goal: The purpose of this project is to develop the entire set of CAP Checklist CDEs in such a way that they may serve multiple purposes as (1) data elements for manual tissue annotation, (2) the target for information extraction from free-text of surgical pathology report by cancer Text Information Extraction System (caTIES), and (3) a conduit from laboratory information systems into caBIG™ data structures with the help of Clinical Annotation Engine (CAE). Results: We are developing all protocols as one UML model, targeting Breast, Prostate, and Melanoma guidelines. Concurrent vocabulary development elements occur within NCI Thesaurus. If these data elements prove to be useful within TBPT systems then they could form the basis for a future caBIG™ standard. Conclusion: The CAP Cancer Checklists can be used as the basis for an electronic data standard in pathology using the caBIG™ semantic modeling methodology. Furthermore, ISO/IEC 11179 gives concrete guidance on the formulation and maintenance of discrete data element description and semantic content (metadata) that shall be used to formulate data elements in a consistent and standard manner.

AUTHOR(S): Sambit K. Mohanty, Rebecca Crowley

AFFILIATION: Division of Pathology Informatics, Department of Pathology, University of Pittsburgh Medical Center

56. Translational Informatics Model for SPECS and SPORES

The SPECS in Lung Cancer involves a consortium of institutions, including SPORES from UCLA, Colorado, UTSW/MDA, Pittsburgh, Dana Farber, Vanderbilt, and non SPORES sites such as USC, Duke, Dartmouth, UC Davis, Michigan, as well as the SWOG, and Spanish Lung Cancer Cooperative groups. Many with their own substantive Bioinformatics operations. The challenge here is to frame a data delivery infrastructure that would not only provide reliable access, but also enable data integration specific for Translational Medicine. Locally generated SPECS and SPORES data from each collaborative site must be processed and stored such that relationships to external data sources can be expressed. Consistency and comparability across data sets requires annotation with controlled vocabularies and, further, metadata standards for data representation. Programmatic access to the processed data should be supported to ensure the maximum possible value is extracted. To confront these challenges for the SPECS in Lung Cancer, we have the need to develop, adopt, and implement a robust infrastructure for data management and integration that supports advanced biomedical applications utilizing existing NCICB and caBIG™ open source software and services within the consortium of participating institutions. This multi-tiered model approach will be outlined in support of Vanderbilt's SPECS and multiple SPORES programs, and in support of the NCI™ caBIG™ Pilot.

AUTHOR(S): Sorena Nadaf, David Carbone, Dan Masys

AFFILIATION: Vanderbilt-Ingram Cancer Center

57. mzHDF for High-Throughput Proteomics Data

We have been developing the open source mzHDF format and associated utilities for storing, accessing, exchanging and computing on very large proteomics data sets using the established Hierarchical Data Format (HDF) developed by the NCSA. The current implementations of mzXML have serious limitations for typical proteomic or metabolomic datasets, which are often larger than a terabyte. As recognized by the Proteomics SIG of the caBIG™, as scientists seek to obtain and analyze proteomics data across institutional boundaries, or even internally within institutions, simply exchanging very large proteomics data sets and establishing efficient data repositories will become bottlenecks for caGrid. We reviewed the current technology limitations of mzXML, mzData, and related MIAPE formats, and evaluated mzHDF with six use cases representing current trends in terabyte cancer proteomics data collection. The mzHDF format is backward compatible with mzXML. It has the exciting potential of handling datasets of unlimited size in a specified dimension. The basic concepts can be extended to unify microarray data and any other high throughput-omics data. mzHDF opens up the possibility that all -omics data across institutions can share the same data repository infrastructure and analytic strategies using services deployed on caGrid.

AUTHOR(S): Simon Lin, Rhett Sutphin, John Osborne, Andrew Winter, Lihua (Julie) Zhu, Moses Hohman, Ph.D, and Warren Kibbe

AFFILIATION: Robert H. Lurie Comprehensive Cancer Center, Northwestern University

58. BreastCancerTrials.org: Feasibility of a caMATCH Tool to Match Breast Cancer Patients to Clinical Trials

BreastCancerTrials.org (BCT) is a patient-centered clinical trial matching tool, piloted by caMATCH. Launched June 2006 in Northern California, efforts are underway to study acceptance, usability, and data quality. Patients registering with BCT self-report a detailed Personal Health Record (PHR) that is matched to trial eligibility criteria. Matches are displayed in a password-contolled Message Center with a trial summary and contact information. Patients can contact research staff directly or use the Message Center to send their PHR to selected sites. As of February 2006, 407 patients registered and consented to participate with BCT. Among these registrants 260 completed their Personal Health Record with 250 matching to at least one trial and 39 using the Message Center to contact a research site. Eleven research consortia were invited and are participating: 9 have obtained IRB approval and have entered trials. Sites include academic and community cancer centers, HMOs, private practices, and CCOPs. Patient satisfaction is being evaluated by online surveys; data quality by comparing a patient's BCT profile to his/her clinic chart. This pilot demonstrates that patients and research sites will use services such as BCT. Further research will evaluate the impact on trial accrual and acceptance by diverse patient populations.

AUTHOR(S): Elly Cohen, Ph.D, Joan Schreiner, Joanne Tyler, Brenda Duggan, Lakshmi Grama, M.S., Mary Jo Deering, Ph.D., Michael Hogarth, M.D., Morton Lieberman, Ph.D., John Park, M.D., and Laura Esserman M.D., M.B.A.

AFFILIATION: University of California San Francisco, Center of Excellence for Breast Cancer Care, and National Cancer Institute Center for Bioinformatics, National Cancer Institute Office of Communications

59. BRIDG: A Model of the Shared Semantics of Clinical Trials Research

In this poster, we will provide an update to the BRIDG project, a formal representation of the shared semantics of clinical trials research. We will describe some of the organizational changes made to the model in response to feedback from the caBIG™, HL7, and CDISC communities, and highlight some of the harmonization efforts to include CTOM, adverse event reporting, and the SDTM standard for the submission of data to the FDA. The BRIDG project is intended to support a Model Driven Architecture (MDA) and provides the shared semantics across various stakeholders.

AUTHOR(S): Douglas B. Fridsma, M.D., Ph.D.,Smita Hastak, M.S., Becky Kush, John Speakman

AFFILIATION: University of Pittsburgh, Clinical Data Interchange Standards Consortium, Health Level Seven

60. Sharing Cancer Microarray Data is Associated with Increased Citation Rate

Background Sharing research data on the caBIG™ grid will provide clear benefit to patients and to the researchers who use that data. For the researcher who makes his or her data available, the benefits are less obvious. In this study we used citation analysis to study the impact of sharing cancer microarray data on citation rates for authors. We collected the citation history and data availability of clinical trials published between 1999 and April 2003, which correlated gene expression patterns with cancer outcomes. Linear regression on the log number of citations received by each trial in 2004-2005 was used to quantify the contribution of data availability. The 41 of 85 trials (48%) which made their data publicly available received 5334 of the 6239 total citations (85%).

AUTHOR(S): Heather Piwowar, Roger Day, Douglas Fridsma, M.D., Ph.D.

AFFILIATION: University of Pittsburgh, Center for Biomedical Informatics

61. Integration Strategies and Methods for Legacy Systems to be caBIG™ Silver Compatible

For many cancer centers, the ultimate goal of participating in caBIG™ is to achieve caGRID availability by sharing their data resources onto the grid. The first step is to enable the legacy systems to be silver compatible. However, tremendous resources available at many cancer centers are still embedded in legacy databases. This poster will document the strategies Kimmel Cancer Center applied to fill the gap of moving our breast cancer tissue system to be caBIG™ silver compatible in caTISSUE CORE context. We will discuss how legacy relational databases can be transformed to UML models in XMI format. After annotating this raw XMI models with EA (Enterprise Architect), we will illustrate how SIW (Semantic Integration Workbench) can help to review, merge, and map legacy models with existing reference/standard models. A local database schema to track down metadata-level information is demonstrated. We intend to share our experience with other cancer centers having similar integration/migration concerns.

AUTHOR(S): Scot Li, Jianguo Yang, Jack London

AFFILIATION: Kimmel Cancer Center, Thomas Jefferson University

63. NCI-60 Dataset Evaluation for the caBIG™ Initiative

A panel of 60 human tumor cell lines (NCI-60) have been used extensively as a standard set of biological reagents for compound screening. The Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, NCI, has assembled a set of high throughput data generated on the NCI-60 set, and subsequently provided it to the caBIG™ initiative. This high throughput data consists of (a) cDNA microarray data, (b) Affymetrix Human Genome U95 set of high density oligonucleotide array data, (c) array comparative genomic hybridization (aCGH) data, (d) spectral karyotyping (SKY) data, and (e) pharmacologic screening data. In addition, detailed descriptions of the derivation of the cell lines have been curated and provided to caBIG™. We report on the detailed statistical and bioinformatical evaluation of these data, as well as their suitability for integration.

AUTHOR(S): Maureen Higgins, Nicholas Socci, Alex Steingart, Alex Lash

AFFILIATION: Bioinformatics Core, Computational Biology Center, Memorial Sloan-Kettering Cancer Center

64. The Impact of caBIG™ Products on the Cancer Patient's Experience

This poster will highlight the impact of caBIG™ products on the cancer patients experience. We will illustrate an individuals experience through diagnosis and treatment and show how caBIG™ products will influence their care. The caBIG™ initiative is developing an interconnected biomedical informatics grid that will facilitate the successful sharing of data, which will lead to more rapid translation of basic research to improve cancer patient care.

AUTHOR(S): The caBIG™ Patient Advocates

AFFILIATION: caBIG™ Patient Advocates

65. A Grid-Enabled Array Annotator Implementation

This poster will describe the implementation of a Grid-enabled application, called ArrayAnnotator. The application supports the data management and analyses processes for annotation of each clone of microarray chips to design custom arrays for DMH experiments in a P50 project studying epigenetic mechanisms in ovarian cancer. This project seeks 1) to identify target genes and promoter regions in a gene that are the binding sites of TGF-B/SMAD from ChIP-chip experiments, 2) to determine genes and regions of promoters that are hypermethylated, and 3) to examine the relationship of DNA methylation profile to clinicopathological features of the ovarian cancer. These research activities require the design of ChIP-chip and DMG experiments, analysis of microarray results, and correlation of results with clinical data. We have developed the ArrayAnnotator application using the caGrid infrastructure to enable synthesis of information from multiple heterogeneous data sources and analysis of the information using different analysis methods. In this implementation, caGrid services have been developed to expose microarray clone information provided by chip manufacturers, enzyme data, and gene sequence data. In addition, several data analysis methods, including microarray landscaping and clone annotation, have been wrapped as caGrid Analytical services.

AUTHOR(S): Junghee Han, Shannon L. Hastings, Stephen A. Langella, Scott W. Oster, Tahsin Kurc, Joel Saltz, M.D., Ph.D., Pearlly Yan, Tim Huang

AFFILIATION: The Ohio State University

66. Model Driven Design Approach to Extending Clinical Annotation Collection at the University of Pittsburgh

caTISSUE Clinical Annotation Engine (CAE) was developed using the caCore SDK Framework to provide a tool for integrating data from existing clinical and research systems. It provides XML-based batch import, a user interface for entering annotations manually and a query interface for finding cases with specific clinical characteristics. Both the batch and interactive modes of adding data to the system enforce the use of caBIG™-standard values as specified in Common Data Elements (CDEs) in the caDSR. The user interface for query is driven off the same CDEs. The labels and values displayed are generated directly from the CDEs as represented in the caDSR. CAE 1.0 is focused on anatomic pathology data. Specifically, the available annotations are based on the College of American Pathologists (CAP) protocol models and the vocabulary tokens for those models are associated with concepts in the EVS thesaurus. Clinical Annotation Engine version 2.0 will build on this foundation to expand on the model-driven approach along several dimensions. Capabilities currently under consideration for CAE 2.0 include: Expanded core model to support the annotation of treatment and outcome data; Model and metadata management capabilities to enable the addition of new annotation sets to model objects; Business rule mechanism for modifying the default persistence, display and query behavior for annotations; Deploy-time, model-driven code generation to enable the implementation of local model extensions. The initial Adopters for CAE include the Norris Cotton Cancer Center (Dartmouth), the Robert H. Lurie Cancer Center (Northwestern) and Siteman Cancer Center (Washington University); Case Comprehensive Cancer Center is expected to adopt the system very early in caBIG™, year 3.

AUTHOR(S): Andy Pople, Linda Schmandt, John Milnes, Michael J. Becich

AFFILIATION: University of Pittsburgh Medical Center

67. Model Driven Development of the caBIG™ CTMS Adverse Events Reporting System (caAERS) Prototype

The objective of the caAERS project is to capture adverse events and report as required to individuals and organizations responsible for trial conduct and for patient safety. CaAERS is envisioned as a flexible set of 12 integrated modules that will be configurable to meet the needs of cancer centers with little or no data management systems, as well as those with robust systems supplemented by selected modules. Through close collaboration with the AE SIG and Working Group members, City of Hope has incorporated a comprehensive set

of requirements into a model driven approach for software development. This poster provides a short course in the development of a prototype utilizing the caBIG™ framework and NCICB developed tools. Use case, activity, domain analysis, message design, and database design models comprise caAERS purpose-driven modeling approach. Specific model transforms drive CDE curation via the SIW and Semantic Loader, and enable code generation via the SDK. They also inform BRIDG of caAERS static information structures and dynamic behaviors. Code generation artifacts including the caAERS data base schema, Java Beans, Hybernate layer and associated web services will be described as well as current vetting activities with the HL7 RCRIM Technical Committee, BRIDG team and VCDE workspace.

AUTHOR(S): Joyce Niland, Ph.D., Doug Stahl, Ph.D., Caroline Song, RN, Susan Pannoni, Cindy Stahl, RN, Jennifer Neat, Hyeoneui Kim, Ph.D., Hemant Shah, M.D., Dave Ko, Edward Lee

AFFILIATION: City of Hope National Medical Center

68. Clinical Trials Object Model (CTOM)

The Clinical Trials Object Model (CTOM) is an initiative within NCICB to model the Clinical Trials arena and offer patient de-identified clinical trials data for research purposes. The UML classes designed in CTOM will be implemented into Java objects using the caCORE tool set. CTOM is a CDE driven UML class diagram and the CDE mappings have been captured in the model. It is intended to be a reference implementation of the BRIDG model.

AUTHOR(S): ScenPro, NCICB

AFFILIATION: ScenPro, Incorporated, National Cancer Institute Center for Bioinformatics

70. Clinical Trials Laboratory Data: A Real-World Application of the BRIDG Model

The BRIDG model is a formal representation of the shared semantics of clinical trials research. The majority of data collected in clinical trials originates in the laboratory; clinical laboratory systems are thus key source systems, and there is a pressing need for reusable interfaces between clinical laboratory systems and clinical trials systems. Based on a presentation at the January 2005 caBIG™ Clinical Trials Management Systems face-to-face meeting, this poster will demonstrate a real-world application of BRIDG in action. The poster will illustrate the step-by-step process of building an analysis model of the domain of the use of clinical laboratory data in clinical trials. The Lab Interface SIG of the CTMS workspace is currently engaged in this process, which builds upon our adoption of the CDISC Lab Model to represent the static information and our development of storyboards to collect some of the dynamic portions of the model.

AUTHOR(S): John Speakman

AFFILIATION: Memorial Sloan-Kettering Cancer Center

71. The Disease Ontology and Browser

We have been developing the open source Disease Ontology (DO) (http://diseaseontology.sourceforge.net) for several years, with the current release, version 3, based on UMLS concepts from unrestricted vocabularies that have been organized into a directed acyclic graph (DAG). The graph has been organized according to OBO principles of being open, machine computable, and the concepts are organized so that the path to the top is always true. DO provides external references to UMLS concepts, including ICD-9CM and SNOMED to allow DO consumers to link into existing sources of structured data such as medical records or to use UMLS-based parsers such as MMTx. We have developed the DO Browser to organize, display, and search a large genotype/phenotype/biospecimen repository for the NUgene project (http://www.nugene.org) by disease. We anticipate using the same concepts to prescreen incoming patients against eligibility criteria for Cancer Center trials. These concepts can be applied to nearly any kind of trial or repository, providing a uniquely generalized tool for collating electronic medical record data to provide evidence based reasoning for determining eligibility across a large cohort, such as all patient contacts at a healthcare organization.

AUTHOR(S): Warren A. Kibbe, John Osborne, Wendy A. Wolf, Maureen E. Smith, Lihua (Julie) Zhu, Simon Lin, and Rex L. Chisholm

AFFILIATION: The Robert H. Lurie Comprehensive Cancer Center, Northwestern University

72. Scalable, Federated, caBIG™-Compatible Clinical Trials—Technology in Use

A technology model delivering a unique infrastructure that has been developed and operationalized across multiple institutions in cancer clinical research is presented and discussed. This model provides collaborative clinical research across ten major research institutions and delivers caBIG™ compatible research services to the network, called the PCC (Prostate Cancer Consortium) Network. The model, built on the Velos system, allows centralized and decentralized management of trials, with institutions running independent instances of software on a members-only grid run by the Prostate Cancer Consortium. Member-run systems work seamlessly to facilitate simultaneous participation in in-house and PCC trials. Different institutions can act as lead centers in different studies, with the trial data then being aggregated to the appropriate lead center across this federated system. As each instance of this networked set of databases is caBIG™-compatible, it provides a unique instance of "NETWORK COMPATIBILITY to caBIG™" across major institutions. Velos is working to move from Bronze to Silver compatibility. The model will then offer Silver Level network-ready access to caBIG™ vocabularies and terminologies to the member institutions, and deliver data in caBIG™ grid-ready formats. The system and some aspects of its architecture are presented and discussed.

AUTHOR(S): Amar Chahal

AFFILIATION: Velos Incorporated

73. Cancer Experimental Protocol Assessment Tool: PADRE

In genomics, proteomics, and metabolomics, molecular-based experimental protocols usually introduce noisy variations in the high throughput data. Thus, manufactures have incorporated some naive statistics methodologies (e.g., GCRMA) to normalize the data. However, it is not enough to eliminate experimental-protocol-specific noisy variations, and the downstream analysis algorithms may produce misleading results for discovery. We have researched and developed caPADRE, a cancer Protocol Assessment and Data Recovery Engine, to assess all genomics and proteomics experimental protocols in NCI-sponsored Georgia Tech-Emory Center for Cancer Nanotechology Excellence (CCNE). caPADRE can identify noise and possible noise sources in experimental protocol. It also provides quality metrics and visualization as feedback for users to make intelligent decisions. For example, CCNE researchers have used caPADRE to process datasets in caARRAY public genomic data repository. They have seen different types of noisy variations linking to different problems in experimental protocols. They have seen different recommendation options provided by caPADRE quality metric. They can decide which portion of the data to use. caPADRE is Web-based and caBIG™ compatible. It provides standards for communications between data providers and data consumers. Thus, caPADRE can improve genomics and proteomics data usability in wider cancer research community supported by caBIG™.

AUTHOR(S): Todd Stokes, Richard Moffitt, Chang F Quo, John Phan, May D. Wang

AFFILIATION: The Wallace H. Coulter Department of Biomedical Engineering

83. The Past and Future of DWD in caBIG™

Distance Weighted Discrimination (DWD) has become a workhorse method for combining micro-array data sets. An old question as to why its empirical performance has proven to be better than PAM and other mean based methods, is resolved through the study of varying subpopulation sizes. Further examples illustrate some situations where conventional DWD may fail. While these situations are flagged by the diagnostic graphics, some expertise is required in their use. More failsafe methods, such as the Adjustment Quality Index and Back Classification are proposed for future development in caBIG™.

AUTHOR(S): Xingliang Zhou, Xuxin Liu, Saianand Balu, J. S. Marron

AFFILIATION: Linberger Comprehensive Cancer Center, University of North Carolina

85. Clinical Research Information Exchange (CRIX) Initiative

In early 2003, the National Cancer Institute (NCI) and the Food and Drug Administration (FDA) formed the Inter-Agency Operational Task Force (IOTF) with the goals to: Speed new research discoveries to the public; and Implement a common, standards based electronic infrastructure for regulatory data and document submission, review and analysis. Recognizing that a common solution would best serve all stakeholders, the Clinical Research Information Exchange (CRIX) initiative was launched in late 2004. CRIX is a collaborative effort among government, the bio-pharmaceutical industry, and academia to implement a common, secure standards-based electronic infrastructure to support the sharing of clinical research data for faster, more efficient development of new drugs. Ultimately CRIX will provide a portal for electronic submission of all regulatory filings, including continuous marketing applications, and e-clinical data collection. The CRIX Steering Committee, formed in 2005, is charged with guiding the CRIX effort through its initial start-up and to prepare the permanent operational, management, and governance structures for CRIX. The committee is responsible for identifying partners, establishing priorities, and providing guidance in the overall development and deployment of CRIX capabilities. To date, the Committee has developed the CRIX Business Plan and business strategy for the first CRIX module, Firebird. The next step is to align and confirm industry partnership in the formation of CRIX as a nonprofit business entity and transition Firebird to CRIX for government and commercial use.

AUTHOR(S): Booz Allen Hamilton: Kevin Vigilante, Patrice Clark, Hilary Jones, Greg Ferrante; SIG: Guy Tallent, Angelo Merola

AFFILIATION: Includes participation from Government, Biopharm Companies, Academia, Cancer Centers and Cooperative Groups, Standards Organizations, and Patient Advocates

86. Regulatory Requirements for Records and Signatures

Regulatory requirements for records and signatures are central to assuring data veracity, authenticity, and non-repudiation of records and/or information submitted to FDA for regulatory review. Title 21 Regulations & ICH Guidelines address the record and signature requirements for pre-clinical, pharmacogenomic, clinical research trials, as well as pharmcosurveillance. These regulatory statutes are categorized and detailed in their content and processes. This poster session will identify the what or predicate rules addressed by the regulations and highlight the importance of why compliance with the regulations is important to the caBIG™ community. We have mapped record and signature regulations to business processes for: Good Laboratory Procedures (cGLPs); Pharmacogenomic Guidance; Good Clinical Practices (cGLPs and ICH Guidelines); Pharmaco-surveillance & Data Mining Guidelines; and appropriately managing the creation, the processing and the retention, and the destruction of the record throughout its lifecycle means meeting the requirements of the regulations.

AUTHOR(S): Hadley White

AFFILIATION: Booz Allen Hamilton

87. InforSense LLC

Inforsense Ltd., working in collaboration with the Windber Research Institute, is building a next generation medical informatics solution to bring the power of translational medicine to the fight against breast cancer. At the foundation of the solution is a patient centric data warehouse that integrates a wide array of traditionally siloed data sources. From genotypes to pathology, PET scans to proteomics, data from the clinic and the lab are made available to researchers and care givers alike. Along with disparate data sources, are an equally diverse community of users. The creation and web-deployment of complex workflows enable informatics specialists to support the specfic needs of various user communites, while a common 'command and control center' fosters collaboration across these groups. By starting with the technology and techniques common in the domain of traditional business intel-

ligence, and adapting it to provide the flexibility demanded by a rapidly evolving research field, our solution enables a powerful top-down approach to research intelligence.

AUTHOR(S): Emma Cutting

88. Nutrition Ontology

A nutritional ontology was developed that integrates into the Thesaurus of the National Cancer Institute (NCI) to facilitate collaboration and data sharing as part of the caBIG™ project. Several agencies provide useful information about dietary components, including the US Department of Agriculture. The International Network of Food Data Systems (InFoods) of the UNFAO provides an administrative framework for the development of standards and guidelines for collection, compilation, and reporting of food component data. Other resources include the NCI Office of Dietary Supplements and Nutritional Science Research Group, the International Union of Pure and Applied Chemistry, and general nutrition books. However, no unified set of definitions for dietary components existed, with links to these sources. Our project developed a vocabulary that includes a list of commonly analyzed nutrients and other dietary components, with definitions, units and interrelationships between items. The final vocabulary will serve as a resource where researchers can unambiguously identify the dietary components and can search for dietary components that are included in a specific class. Additionally, new research studies will be to use the NCI concept codes to declare the definitions in use and thus help make the resulting study data more interoperable.

AUTHOR(S): Lynne R. Wilkens, Leo W. K. Cheung, Suzanne P. Murphy, Donna Lyn M. T. Au, Sherri De Coronado

AFFILIATION: Cancer Research Center of Hawaii, Honolulu, HI, National Cancer Institute, Washington, DC