National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

caBIG™ Tools Fact Sheet

The cancer Biomedical Informatics Grid (caBIG™) develops and freely provides software tools and datasets using common infrastructure and vocabularies. Cancer research progress and discovery are accelerated by facilitating data exchange across many disciplines, including pathology, molecular biology, clinical trials, and imaging. By uniting the cancer research community, caBIG™ projects are expected to lead to enhancements in health outcomes for cancer patients.

For further information, please visit https://caBIG.nci.nih.gov.

caBIG™ Tools

Following are brief descriptions of caBIG™ tools, grouped into three categories: Clinical Software, Data Analysis Software, and Infrastructure. Tools in these groups are further subcategorized by general and specific functions.

Visit https://caBIG.nci.nih.gov/inventory for more detailed information and access to caBIG™ resources.

For More Information:

Mary Jo Deering, Ph.D.
Director for Informatics Dissemination
Center for Biomedical Informatics and Information Technology
National Cancer Institute
301-496-3458
deeringm@mail.nih.gov

US DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

July 2007

Clinical Software

GENERAL Function	SPECIFIC Function	Name	Description
Clinical Trials Management	Clinical trial data collection	Cancer Central Clinical Database (C3D)	C3D provides clinical trial managers a secure tool for collecting, tracking, auditing, and electronically submitting trial data across multiple studies and sites. C3D collects clinical trial data using standard case report forms (CRFs) based on common data elements (CDEs). C3D utilizes security procedures to protect patient confidentiality and maintain an audit trail as required by FDA regulations. This web-based application can be hosted at NCICB or locally at an individual institution.
		Cancer Central Clinical Participant Registry (C3PR)	C3PR is a web-based application that helps organize and manage participant registration data collected in clinical trials.
	Clinical trial data submission	Clinical Data System (CDS)	The CDS is a web-based system for the submission of patient data in NCI-sponsored clinical trials. The CDS provides a centralized portal for viewing and generating reports on submitted data.
	Patient calendar management	Patient Study Calendar (PSC)	PSC enables clinical trial managers to schedule and manage treatment and care events for each participant in a clinical trial.
Biospecimen Banking	Biospecimen tracking and annotation	caTissue Core	caTissue Core is a biobank management tool to collect, manage, process, annotate and distribute biospecimens and associated information to selected users. caTissue Core 1.2 manages tissue, fluid, cell, and molecular biospecimen information. The next release, caTissue Suite 1.0, will integrate the annotation functionalities of caTissue clinical annotation engine and caTIES.
	Biospecimen annotation	cancer Text Information Extraction System (caTIES)	caTIES automates the extraction of coded information from surgical pathology reports and presents it in a standardized format using common data elements. In its current release, caTIES 2.3 can also de-identify the extracted information using a third-party de-identification tool. Users can take advantage of the standardized representation of information to effectively query, browse, and acquire annotated biospecimens.
Image Analysis	Cancer image archive and retrieval	National Cancer Imaging Archive (NCIA)	NCIA is a searchable repository of *in vivo* cancer images, such as CT, MRI, and Digital X-rays. NCIA also contains annotation files (PDF, image markup) and annotation data provided by a curator. Cancer images are integrated with clinical and genomic data.

Data Analysis Software

GENERAL Function	SPECIFIC Function	Name	Description
Data Integration	Molecular biology data analysis	geWorkbench	geWorkbench consists of more than forty different modules that provide integrated analysis and visualization of a variety of biomedical data, including microarrays, sequences, pathways, ontologies, and transcription factors. geWorkbench provides access from any repository with a MicroArray and Gene Expression Object Model Application Programming Interface (MAGE-OM API), such as caArray.
	Molecular biology data analysis	GenePattern	GenePattern is a powerful tool for the analysis of gene expression, proteomics, and genotyping. GenePattern can also read data directly from caArray, a caBIG™ microarray repository.
	Molecular biology and clinical data analysis	caIntegrator	caIntegrator allows researchers to integrate and analyze a variety of data types from multiple sources, including microarray, genomic, immunohistochemistry, imaging, and clinical data, through a single application.
Genome Analysis	Genome annotation	SEED	SEED is a framework that supports peer-to-peer annotation of genomes, with investigators having the ability to work independently and synchronize their work or update code versions. Analyses such as psi-blast can be performed on sequences that match a specified annotation search.
	Genome annotation to find disease-causing mutations	Transcript Annotation Prioritization and Screening System (TrAPSS)	TrAPSS permits the prediction of the likelihood of gene sub-sequences to contain disease-causing mutations; it utilizes annotation to prioritize focused regions of a gene during mutation screening or when searching for linkage between mutations and disease phenotype.
	Microarray probe annotation	Function Express (caFE)	caFE provides a system for automatically updating microarray probe annotations and a literature search to allow users to search for relationships between genes with similar expression profiles.
	Determination of biological functions using Gene Ontology	GOMiner	GOMiner classifies genes from microarray experiments into biologically coherent categories using the Gene Ontology (GO), a controlled vocabulary that describes gene and gene product attributes in organisms. GOMiner aids investigators in the interpretation of biological functions of genes found in an individual microarray experiment.

Statistical Analysis	Multivariate cluster-modeling	VIsual Statistical Data Analyzer (VISDA)	VISDA is a statistical analysis tool used for multivariate cluster-modeling, discovery, and visualization of high-dimensional data sets. VISDA includes functionalities for global and local biomarker identification and prediction.
	Statistical corrections	Distance Weighted Discrimination (DWD)	DWD performs statistical corrections to reduce systematic biases resulting from different laboratories, sources of RNA, batches of microarrays, and microarray platforms.
Data Services	Protein data collection and analysis	Protein Information Resource (PIR)	PIR is a data resource comprising an integrated and annotated protein database, containing more than 283,000 sequences covering the entire taxonomic range of organisms. The data set is part of UniProt, the central international resource of protein sequence and function that unifies the PIR, Swiss-Prot and TrEMBL databases.
	Microarray data collection and analysis	caArray	caArray is a MIAME 1.1 compliant microarray data repository. Data is accessible through a MicroArray and Gene Expression Object Model Application Programming Interface (MAGE-OM API) as well as a graphical user interface.
	Collection of data on animal models of human cancer	Cancer MODels Database (caMOD)	caMOD permits retrieval of information on animal models of human cancer ,including genetic descriptions, histopathology, images, and microarray data. Data are directly submitted by scientists or extracted from the public scientific literature by curators.
Protein Analysis	Mass spectrometry data analysis	RProteomics	RProteomics performs low-level analysis (denoising and peak alignment) of proteomics data from surface-enhanced laser desorption ionization/time of flight (SELDI-TOF) and of matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF).
		Q5	Q5 is an algorithm that supports probabilistic disease classification of expression dependent on proteomic data from mass spectrometry of human serum. Q5 has been integrated into the RProteomics suite of tools.
	Management of 2D gel lab processing	Proteomics Laboratory Information Management System (protLIMS)	protLIMS tracks the laboratory processes relevant to two-dimensional gel electrophoresis, with a schema to support the addition of emerging new data types.
	Collection and analysis of protein data relevant to cancer	Cancer Molecular Pages (CMP)	CMP is a catalog of automatically annotated cancer-related proteins, with integrated data from Genbank, computer generated annotations (such as predictions of 3-D structure and protein sequence similarity comparisons), and user's annotations. This web-based resource incorporates a range of homology tools. CMP links entries to relevant caBIG™ datasets and other similar online databases.

Pathway Analysis	Analysis of microarray and pathway data relevant to cancer	Quantitative Pathway Analysis in Cancer (QPACA)	QPACA provides quantitative analysis of microarray data in the context of pathway structure.
	Human pathway analysis	Reactome	Reactome Genome Knowledge Base (GKB) is a curated database that includes biological pathways and transformations in humans. The information in the database is cross-referenced with the sequence databases Ensemble and SwissProt..

Infrastructure

GENERAL Function	SPECIFIC Function	Name	Description
Core Infrastructure	Data sharing network	caGrid	caGrid is the underlying network architecture that provides the basis for connectivity between all of the cancer community institutions, allowing research groups to tap into the rich collection of emerging cancer research data while supporting their individual investigations. caGrid manages and securely shares information and analytic resources using locally managed access control policies and using strongly typed data objects in XML format.
	Standardization of clinical data exchange	Biomedical Research Integrated Domain Group (BRIDG) Model	The BRIDG model provides the basis for harmonization among standards within the clinical research domain and between clinical research and healthcare.
	Common data management and application development framework	cancer Common Ontologic Representation Environment (caCORE)	caCORE provides systems for implementing controlled terminologies and metadata standards in order to help caBIG™ software programs seamlessly work with each other.
	Development of controlled vocabularies	NCI Enterprise Vocabulary Services (EVS)	EVS develops standard, controlled vocabularies as part of caCORE. This service produces the NCI Thesaurus and the NCI Metathesaurus, which is based on NLM's Unified Medical Language System Metathesaurus and supplemented with additional cancer-centric vocabulary.
	Standardization of metadata in the form of common data descriptors	cancer Data Standards Repository (caDSR)	caDSR is a metadata registry in caCORE that stores and manages Common Data Elements (CDEs) which are developed by caBIG participants and various NCI-sponsored organizations.

	Creation of a “caCORE-like” software system	caCORE Software Developers Kit (caCORE SDK)	caCORE SDK provides a toolkit to create a caBIG™ Silver Compatible data system.
Vocabularies	Vocabulary installation and publication	LexBIG	LexBIG provides a hosting solution for terminology distribution.
	Standard vocabulary development for human and mouse anatomy	Mouse Human Anatomy Mapping (MHAP) Ontology	The MHAP ontology provides a mapping and harmonization of human and mouse anatomical descriptors as they are currently used by Mouse Genome Informatics and the NCI Thesaurus.
	Standard vocabulary development for cancer nutrition	Cancer Nutrition Ontology	The Cancer Nutrition ontology provides standard vocabularies for cancer nutrition research. The project is driven from studies that search for nutritional factors that alter the risk of getting cancer.