CCR/NCI Integrated and Collaborative Knowledge Environment

David J. Goldstein[1], Javed Khan[1], Chand Khanna[1], Snorri Thorgeirsson[1], Jean C. Zenklusen[1], Richard W. Zhang[2], Hilmar Ilgenfritz[3], Ioannis Kontodinas[3], Oliver Heinrich[3], Klaus Heumann[3], Patrick M. Blake[2], and Shoshana Segal[1]

1. Center for Cancer Research, National Cancer Institute • 37 Convent Drive • Bethesda, MD 20892 • USA **2. Sophic Systems Alliance, Inc.** • One Research Court, #450 • Rockville, MD 20850 • USA **3. Biomax Informatics AG** • Lochhamer Str. 9 • D-82152 Martinsried • Germany

Abstract: The CCR Office of Science and Technology Partnerships (OSTP) is responsible for exploring emerging technologies and making them available to CCR scientists through partnerships, collaborations, contracts, and other technology agreements. An ongoing OSTP project is the Knowledge Integration and Management System provided by Sophic and Biomax. This System enables the visualization of complex relationships between biological and biomedical data and information. Six laboratories within CCR/NCI were chosen to participate in a pilot study to evaluate the System and determine its benefits to cancer research at CCR. The areas of research include ovarian cancer, metastasis, liver carcinogenesis, neuroblastoma, radiation oncology, and neuro-oncology. The System is designed to institutionalize knowledge by enabling the annotation of information stored in a central repository. The system's configuration requirements depend on the research conducted in each lab. The Knowledge Environment was configured to support research on different types of cancer, sources of data, and research strategies. Over 35 public databases were integrated into a single user interface, allowing scientists to query a wide range of data sources. Information derived through genomics, proteomics, pathway analysis, and clinical studies are combined into graphical representations of complex relationships. Future plans for the CCR Knowledge Environment include the integration of additional databases, tools, and software.

CCR Knowledge Environment Implementation Project Trial Goals

Scientists from Sophic Systems Alliance and Biomax are collaborating with CCR investigators to:

- Integrate information from multiple systems inside labs with information from various public domain sources into a single system with a simple, easy-to-use interface.
- Support diverse areas of cancer research, different discovery strategies and evolving hypotheses and research processes.
- Provide a lab-centric research environment and enable sharing of information across the labs with a collaborative layer.
- Integrate caBIG and other public domain software with proven commercial software systems.

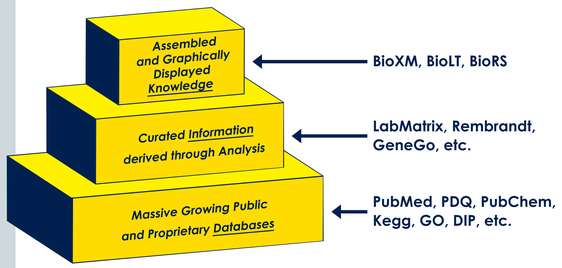
Dr. David J. Goldstein
Dr. Shoshana Segal
OSTP

Dr. Michael Birrer Ovarian Cancer	**Dr. Kevin Camphausen** Radiation Oncology	**Dr. Javed Khan** Pediatric Oncology
Dr. Howard Fine Neuro-Oncology	**Dr. Chand Khanna** Tumor Metastases	**Dr. Snorri Thorgeirsson** Experimental Carcinogenesis

Knowledge Architecture

Scientific information is typically found in public and proprietary databases or in free text such as PubMed abstracts. As scientists collect answers to questions derived from various sources, assembling and finding relationships between the disparate pieces of information is a significant challenge. The Knowledge Environment developed by Biomax Informatics AG includes three modules that access both text and databases and then "assemble" the information. The Knowledge layer sits on top of the text mining and database query systems, integrating information objects into relationship-based networks. The complex networks of semantic and experiment-based relationships are graphically represented and provide insight into the mechanisms of cancer.

The three layer architecture includes:



Integration and Implementation of the CCR Common Knowledge Environment

The laboratories featured herein provide examples of how their research and disease-specific focus can be supported while creating the CCR Common Knowledge Environment to be shared across CCR labs. The CCR Common Knowledge Environment integrated and implemented by Sophic, will provide a single interface for researchers. Scientists will have access to multiple sources of information inside NCI and throughout the public domain, while integrating a range of other commercial software systems.

Examples of innovations implemented in the CCR Common Knowledge Environment include:
24K X 24K Gene/Gene Human Map – The BioLT Linguistics System was used to identify all co-occurring gene/gene relationships in 80 Million PubMed sentences. The individual cells in the 24K X 24K Gene/Gene Human Map contain the "address" of individual sentences in PubMed that connect a gene with another gene. This network of complex relationships is graphically represented in BioXM and allows the researcher to explore relationships throughout the map.

Integration with the NCI PDQ (Physicians Data Query) Clinical Trials Database – Researchers focused on early discovery are now able to query into the clinical trials database to find information that would impact their research direction and strategy. The combination of genomic, proteomic, therapeutic and clinical information in single graphical representation allows the researchers to see complex networks that previously would have been difficult to find.

CCR Common Database Repository – The implementation included integrating over 35 public domain databases such as PubMed, PFAM, EMBL, OMIM, DIP, GO, etc. into a single source for researchers to access with a single semantic query. Answers to these queries are integrated into the CCR Common Knowledge Environment where this new information is mapped into graphically represented relationships.

Publicly Available and Commercial Software Integration.

NCI/CCR Integrated and Collaborative Knowledge Environment

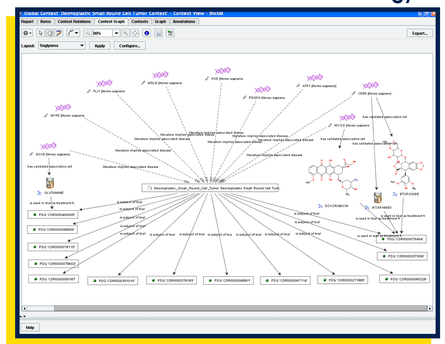
24K X 24K Gene/Gene Human Map	**BioXM BioRS & BioLT**	PDQ® - NCI's Comprehensive Cancer Database
gene	BioFortis	

-35 Integrated Public DB & ontologies including PubChem, PubMed, DIP, Bind, NCI Thesaurus, GO, EMBL, Kegg, OMIM, Affymetrix Probe Sequence, SwissProt, Refseq, GenBank, PFAM, etc.

The project entailed building interfaces with commonly used public domain and commercial software such as:

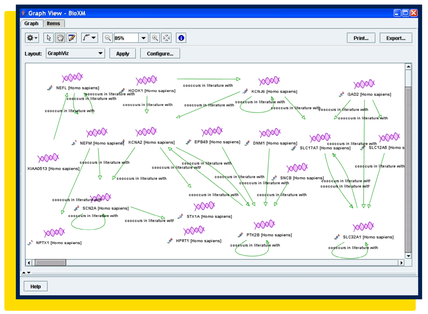
- **Biomax BioXM Knowledge Management System** – the software layer where semantic objects representing scientific elements are assembled into complex relationship networks.
- **Biomax BioLT Linguistic System** – accesses PubMed abstracts and may access any other text corpus.
- **Biomax BioRS Data Integration and Retrieval System** – middle-ware that accesses multiple flat files and relational databases.
- **caBIG Rembrandt** – robust knowledge-based framework that hosts and integrates clinical and functional genomics data from clinical trials involving patients suffering from gliomas.
- **BioFortis LabMatrix** – internet-based, HIPPA compliant, scientific application that serves as a central data repository merging clinical, genetic and molecular data.
- **GeneGo MetaCore** – data-mining analysis software system focused on ligand receptor interaction, cell signaling regulation and metabolic pathways.

Use Case #1
Dr. Javed Khan's Lab - Pediatric Oncology



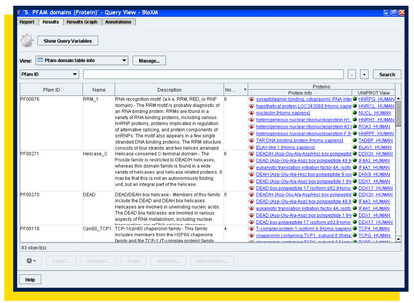
Despite aggressive therapy the survival rate for patients with metastatic cancers remains <30%, at diagnosis and <5% after the first relapse. Using BioXM we attempted to identify new drugs that could potentially treat a patient presenting with multiply relapsed Desmoplastic Round Cell Tumor (DRCT). We first identified all genes associated with the disease DRCT and compounds that targets these genes. We next identified which of these drugs are currently in clinical trials from the NCI PDQ database integrated in BioXM. This search identified 4 compounds that are currently in clinical trials. These results were obtained within a single session. The eventual goal of these types of queries is to identify, using evidence based queries, new treatments for patients with incurable cancers.

Use Case #2
Dr. Howard Fine's Lab – Dr. Jean C. Zenklusen Brain Cancer



Primary brain tumors are the fourth leading cause of cancer mortality in adults under the age of 54 and the leading cause of cancer mortality in children in the United States. Therapy for the most common type of primary brain tumors, gliomas, remains suboptimal. The development of new and more effective treatments will likely require a better understanding of the biology and molecular classification of these tumors. We have developed a molecular-based classification of gliomas (both high and low grades) that creates six distinct entities correlating with various clinical parameters. However, the classifiers' lists are, in general, a group of disparate genes, not related by any obvious pathway. This screen shot shows the results obtained with the BioXM system, employing its gene-gene correlation algorithm, we were able to establish direct or indirect (through an interpolated gene) relationships between the classifiers that allow for a better understanding of the biological relevance of the groups in question. The total time spent in creating such a "network" was less than 30 minutes.

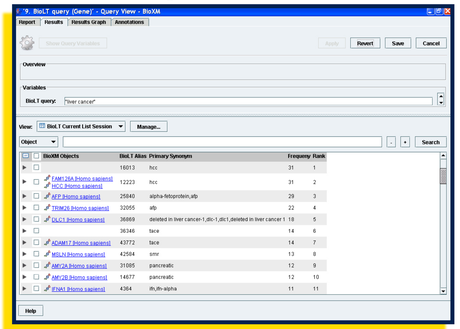
Use Case #3
Dr. Chand Khanna's Lab - Tumor Metastases



The focus of our laboratory's efforts thus far has been the study of the cytoskeleton-linker protein, ezrin, a potential metastasis-associated portein. Ezrin sits at the cell membrane establishing both physical and functional connections between the cell membrane and the actin cytoskeleton. This connection is the result of the N-terminus of ezrin binding to integral membrane or scaffolding proteins and the C-terminus binding to the actin cytoskeleton. These ezrin - protein interactions are globally referred to as the ezrin interactome, and consist of both direct and indirect binding partners. The interactome defines what signaling pathways and partners are available to proteins, may direct a proteins subcellular localization, and may modulate a proteins stability following translation. Using a non-candidate biochemical approach we have begun to shape the ezrin interactome and generated new hypotheses on ezrin's functions in metastasis.

Our initial proteomic assessment of the ezrin interactome has generated 77 proteins of strong interest. A difficulty in dealing with this proteomic data has been defining common features amongst binding proteins. The use of BioXM platform has begun to provide a single site, where protein functions, binding proteins, structural domains, and ontologies for these proteins can be reviewed and visualized. Identification of common features and connections between these proteins will allow ezrin:protein interactions to be considered as druggable targets for new ezrin-based cancer therapies.

Use Case #4
Dr. Snorri Thorgeirsson's Lab - Liver Cancer



By using a cross-species comparative genomic approach, we identified a gene signature which can discriminate a clinically significant subset of human liver cancer. Based on the gene signature identified from our approach, we applied the BioLT text-mining tool to search for associations between the genes of interest and "Liver Cancer" in the BioXM framework.

National Cancer Institute
Center for Cancer Research
Office of Science and Technology Partnerships
37 Convent Drive
Bethesda, MD 20892
+1 301 496 4347
www.nci.gov

Sophic
Integrated Life Science Solutions
Sophic Systems Alliance, Inc.
One Research Court, #450
Rockville, MD 20850
+1 508 495 3801
www.sophicalliance.com

BioMax
Biomax Informatics AG
Lochhamer Str. 9
B-82152 Martinsried Germany
+49 89 895574-0
www.biomax.com