CCR/NCI Integrated and Collaborative Knowledge Environment

David J. Goldstein[1], Javed Khan[1], Snorri Thorgeirsson[1], Jean C. Zenklusen[1], Richard W. Zhang[2], Hilmar Ilgenfritz[3], Wenzel Kalus[3], Klaus Heumann[3], Patrick M. Blake[2], and Shoshana Segal[1].

1. CCR, NCI 2. Sophic Systems Alliance, Inc., Rockville, Maryland 3. Biomax Informatics AG, 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. Its configuration is based on the specific research conducted in each lab and supports different types of cancer disease, various sources of data, and diverse research strategies. The 2007 Pilot Phase is complete and resulted in the development of the CCR Integrated Knowledge Environment, which includes predefined queries that can easily be applied to establish relationships between biological terms (e.g., genes, proteins, diseases, etc.). The Common CCR Knowledge Environment was configured to provide access to over 35 public databases 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 are to include the integration of additional databases, tools, and software.of complex relationships. Future plans are to include the integration of additional databases, tools, and software.

CCR Knowledge Environment Implementation Goals

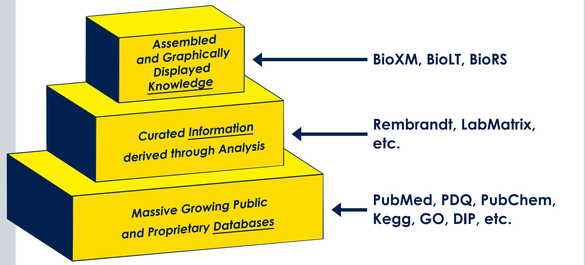
Scientists from Sophic Systems Alliance and Biomax are collaborating with CCR investigators to:

- 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.

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.

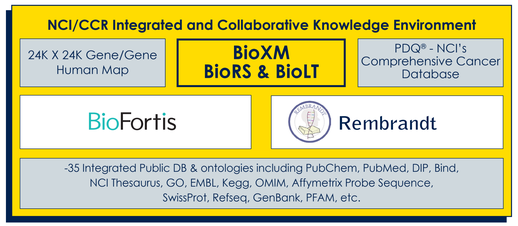
Examples of innovations implemented in the CCR Common Knowledge Environment include:

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.

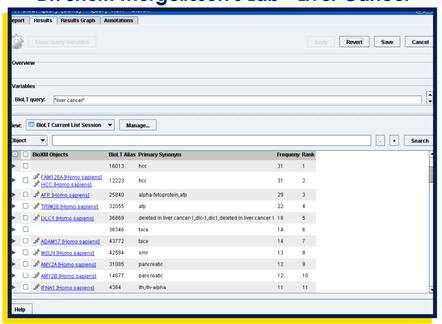
Publicly Available and Commercial Software Integration.



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

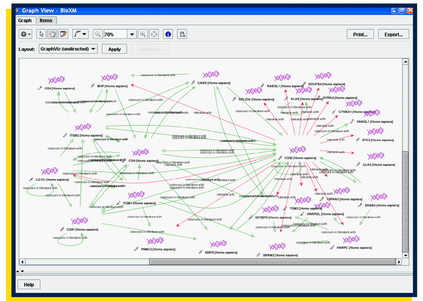
- **Biomax BioXM Knowledge Management System** – the software layer where semantic objects representing scientific elements are assembled into complex relationship networks.
- **Biomax BioLT Linguistic System** – accesses PubMed abstracts and may access any other text corpus.
- **Biomax BioRS Data Integration 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, HIPAA compliant, scientific application that serves as a central data repository merging clinical, genetic and molecular data.

Use Case #1
Dr. Snorri Thorgeirsson's Lab - Liver Cancer



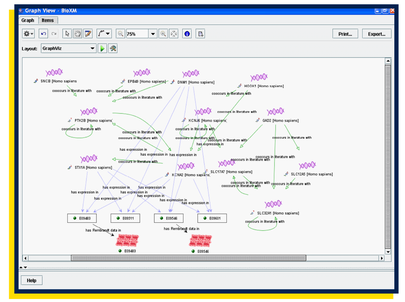
By using a cross-species comparative genomic approach, we identified a gene signature which can discriminate a clinically significant subset of human liver cancer. Based on the gene signature identified from our approach, we applied the BioLT text-mining tool to search for associations between the genes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

Use Case #2
CCR Office of Science and Technology Partnerships
Yeast 2-Hybrid Screening Program: Defining Functional Relationships



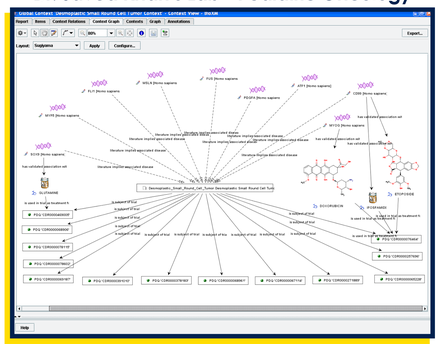
The OSTP manages a research partnership with Myriad Genetics to help CCR investigators identify novel protein-protein interactions and signaling pathways using Myriad's automated process based on the Yeast Two-Hybrid (Y2H) methodology. Since the programs inception, over 100 cancer- and HIV-related genes of interest to CCR investigators were analyzed, resulting in over 1800 novel protein interactions, many of which are now being validated and confirmed by CCR investigators. The stored queries in the CCR Common Knowledge Environment are now being used to assist with the analysis of this very large and complex data set of "bait" and "prey" relationships and to help identify potential relevant interactions in mammalian cells. In the example shown here, CD82 (also known as Kai1) is shown connected by red arrow with 17 preys from the Y2H screen. CD82 is also shown connected by 9 proteins with green arrows, which represent interactions shown previously in the literature. The BioXM framework is used to display the results from literature mining using the predefined BioLT human gene-gene co-occurrence algorithm. This algorithm searches all MedLine abstracts for occurrence of any combination of two genes from the list of known CD82 binding partners and CD82 preys identified in the Y2H screen. This analysis reveals that some of the preys of CD82, such as MVP, co-occur in the literature with a number of known CD82 binding partners (ITGB2, CD9 and CANX). This literature mining analysis using the BioLT algorithm and the BioXM visualization tool reveals a potential functional relationship between a novel and previously characterized CD82 binding partners. This approach is now being employed on several other Y2H bait/prey combinations to help identify interactions that are more likely to be functionally relevant and worth further investigation.

Use Case #3
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. Additionally, here we show how the BioXM integrates to both the LabMatrix (Electronic Lab management system) and Rembrandt (Repository of Molecular Brain Neoplasia Data) to obtain both information on the available resources at hand for the samples analyzed (LabMatrix) and the expression values of these genes in the aforementioned samples. The total time spent in creating such a "network" was less than 30 minutes.

Use Case #4
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.

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