Sophic
Integrated Life Science Software and Services

White Paper
A Cancer Gene Index Use Case Overview of The NCI CCR Common Knowledge Environment

The NCI CCR Common Knowledge Environment

The range of questions cancer researchers ask is endless. The potential answers to these questions are, unfortunately, stored in countless text files and public and proprietary databases. Non-technical scientists need a simple portal, as the single point of entry where they can begin the question, answer, and exploration process.

The NCI CCR Common Knowledge Environment consists of a configured and implemented BioXM™ Knowledge Management System[1] using the Cancer Gene Index (CGI) integrated with extensive public- and lab-centric databases. The GUI user interface provides easy access to an integrated bank of selected databases that allow scientists to begin or continue the question and answer process efficiently and effectively. Information in the Common Knowledge Environment can be confidential and used exclusively by a scientist, lab, or shared with external collaborators.

Many scientists begin their search of these complex topics with common, frequently asked questions about their list of genes, protein, compounds, pathways, etc. In this environment, each of the common questions a scientist may ask is pre-configured as a ***smart folder*** that allows scientists to insert their variables – gene lists for example - and query a bank of selected databases that are likely to have answers to their questions. These ***smart folders*** are only the beginning – a gateway for ongoing research and exploration that is virtually limitless within the Common Knowledge Environment.

Pre-configured Smart Folders Overview:

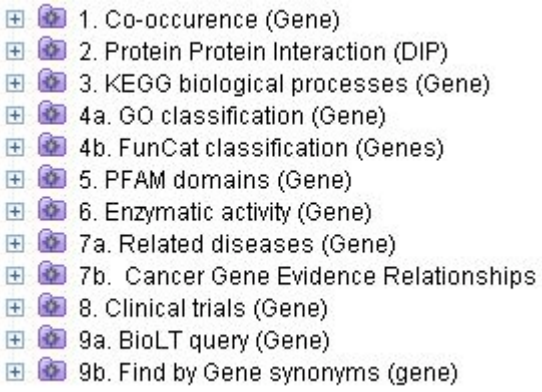
Some of the common questions asked by scientists that can be addressed using this knowledge environment already exist in the pre-configured smart folders:

1. How is my gene related to other genes and what evidence supports these relationships?
2. I have a list of genes. How are they related to one another?

[1] BioXM™ Knowledge Management System is a trademark of Biomax Informatics AG, Munich, Germany

2275 Research Blvd., #500, Rockville, MD 20850
Office (301) 296-4554 Fax (301) 330-0740
www.sophicalliance.com

3. What functional domains are represented in my list of genes?
4. Does my gene list have connections with any other known diseases?
5. What are the documented clinical trials that are underway for the list of genes I have?
6. Are there any papers in the literature linking some of my terms and the genes I have?
7. Are there any pathways mapping to the genes I have in mind?



The following figure shows a list of such *smart folders* that describe the specific questions and concepts arising out of a list of genes, and how the results can be used to generate further relationships.

Co-occurrence of Human Genes Smart Folder

Research scientists often ask whether a lists of their genes are related to one another, and if so, what evidence points to this. Alternatively, they may be focused on one gene, and wish to map and network all the other genes that have been documented to interact with this one gene.

The *Co-occurrence* Smart Folder addresses this particular need. It is preconfigured to take in a list of genes and fire them against a database of about 24,000 genes. All of the evidence for this entire list is stored in a huge matrix of each gene against the other. The result is a comprehensive table listing the source, target and evidence in the literature pointing to the relationship. The search function allows easy location and selection of specified records. If the output is large, the user can then select a list from the table and generate a visual schematic showing the relationships between all of these entities. A list of control buttons on top of the graphical output allows the user to maneuver the elements (drag and resize the objects). Further, there is a freedom to simultaneously display the results in diagrammatic and tabular formats.

A series of options using the mouse-pointer are available to the user that link a gene name to its products, all other possible relations evidenced from other databases, gene-probe relationships, gene compound relationships, changes in expression levels, etc.

Protein-Protein Interaction (DIP) Smart Folder

This smart folder generates a list of interacting proteins from the *file list* containing the source genes. The evidence gleaned is from the ***DIP*** (Database of Interacting Proteins hosted on servers at UCLA) is, a comprehensive catalog of experimentally determined interactions between proteins.

Some of the common questions that can be answered using this smart folder are:

1. How are proteins produced from the genes in my *file list* related to each other? Are they known to interact with each other and, if so, what literature evidence points to this?
2. What are the functions known about the proteins whose gene list I have uploaded?
3. What other families of proteins are the products from my genes related to?
4. What are the known biological functions assigned to my gene list (e.g. angiogenesis, chemotaxis, etc.)?

The user gets to upload a list of genes whose proteins products are being studied. The list is fired against the DIP database and a tabular output is generated showing the *Object*, *Source protein*, *Target protein*, *Medline source*. As mentioned earlier a search function allows the location and selection of specific records. Following this search, a graph can be generated, as mentioned above. The *right click* options for the individual elements in the graph are appropriately configured. For example, a *right click* on any protein allows the user to explore functional relationships of the proteins evidenced through different databases (e.g., *FunCat*, *GO*, *Pfam*, or even evidences of further *protein interaction)*.

KEGG Biological processes (Gene) Smart Folder

Sometimes, researchers want to know all about the pathways relating to a list of genes that they study. Typical questions in this category include:

1. Is a particular pathway assigned to the gene I am investigating?
2. What are the pathway components assigned to the list of genes I have?
3. How are the displayed pathways related to each other, and what other information can I gather from them?
4. What are the enzymatic activities related to the genes I am studying?

Answering the above questions means connecting to a database that maps all the metabolic pathways, their enzymatic activities and related pathway components. To address this, the queries are linked to the ***KEGG*** database to gather the related information. KEGG stands for **K**yoto **E**ncyclopedia of **G**enes and **G**enomes and was initiated by the Japanese human genome program. This database is a resource of online databases dealing with *genomes*, *enzymatic pathways* and related *biological chemicals*. Consequently, this smart folder queries a *file list* of gene names against the KEGG database and displays the results in tabular and graphical format.

The tabular format displays the mapped genes against the known pathways, the number of *pathway components*, and all related information regarding the *mapped genes* and their *protein* components. Clicking on any particular record displays all information relating to the genes mapped to the pathway mentioned in the record. There is a *Graph* button among the tabbed buttons on top of the record view. Clicking this button displays all *relations* in the selected pathway. The graphical output allows the user to select a few pathways and click on the *Graph* option to get a diagrammatic representation of the link between different pathways. The initial picture shows the components unconnected to each other. However, on selecting the icons, the user can use the *right click* option to explore all kinds of relationships between the selected pathways. The simultaneous view, also affords a parallel view of the results in both formats with the ability to search across the table.

GO Classification (Gene) Smart Folder

Suppose there is a list of genes and the user wants to use this list to explore relationships between scientific elements in different databases and analyze the data from these disparate sources. Exploring ech of these sources and then annotating the results can be a cumbersome and painstaking task. The ***GO*** database addresses this issue by giving the user the ability to connect his gene of interest to different assigned biological activities. The *Smart folder*, in turn, uses an uploaded list of genes and develops relationships between them to create a highly specific and focused graphical output.

The ***GO*** Classification stands for the ***G***ene ***O***ntology classification. It represents a collaborative effort to address the need for consistent descriptions of gene products in different databases. This *smart folder*, through its link to the GO database, allows the user to retrieve and analyze data from disparate sources[2] related to the pathway under analysis.

[2] Webpage of the GO database - http://www.geneontology.org/GO.indices.shtml . Clicking on the FAQ gives detailed information on the rationale behind the database design

The tabular format displays the mapped genes through the *GO IDs*, providing a brief *description* of the biological activity involved (e.g.: *angiogenesis*, *chemotaxis*, *response to stress*, *signaling)*. Clicking on any particular record gives complete access to all relations pointing to this concept. For example, if a record in tabular output said *response to drug*, then clicking on that record shows all the genes from the *file list* that are mapped here, and the number of components linked to this. The graphical output displays components separately which can then be used to study and understand inter-relationships between the mapped components.

FunCat Classification (Gene) Smart Folder

The *FunCat* is an annotation scheme for the functional description of proteins from a *file list* of human genes. This scheme takes into account the broad and highly diverse spectrum of known protein functions and is hence divided into 28 main functional categories that cover general fields, such as cellular transport, metabolism and cellular communication/signal transduction[3]. Consequently, this smart folder takes a *file list* of human genes and maps them using this database. The tabular format shows all the mapped genes from the uploaded *file list* linked to their appropriate FunCat IDs. This is followed by a brief *description* of the associated process.

Clicking on each record provides access to the *derived concepts*, *relations*, *contexts*, *graphs* and *annotations*. The graphical output from the selected records provides a graphical visualization from which other related information can be networked. The *Right click* options lead the user to explore and understand other *relations* and *ontology relationships* in the selected records. For example, a *right click* on a particular element in the graph, (e.g.: "06.07.03 modification by phosphorylation, depshosphorylation"), will lead the user to explore further *FunCat*, *GO database* relations OR ontologies between selected elements in the graph.

PFAM domains (Gene) Smart Folders

Often, research scientists wish to know if the sequence of the protein they are studying has any other protein sequences that are similarly aligned and if there is any particular database that contains such related information. Such information would allow the researcher to study hitherto unknown relationships to other proteins with other functions.

[3] Homepage of the ***Functional Catalogue:*** - http://mips.gsf.de/projects/funcat. The opening lines describe the scope and coverage of this database.

The ***Pfam database*** is a large collection of *multiple sequence alignments* and *hidden markov models* covering a variety of protein domains and families. Hence the use of this *smart folder* results in a complete analysis of the proteins coded by the genes uploaded from the *file list*. This is a huge *timesaving way* to look into multiple sequence alignments of a given protein against numerous other related proteins. The tabular format of the results display the mapped genes to the domains in their respectively encoded proteins, followed by a brief description of the encoded protein and a list of the genes mapped there. Clicking on each record provides information on all the *relations* and *contexts* for the selected domain. The graphical output from selected records can be visualized to develop a schematic and explore further relations. *Right click* options lead the user to look at further *annotations* and *ontology relationships*.

Enzymatic activity (Gene) Smart Folder

The *Enzymatic-activity smart folder* maps all the related enzymes to the uploaded *file list* of human genes. This *smart folder* is useful when one wants to discover all the possible enzymes linked to the genes in the list. Some typical questions that can be answered using this *smart folder* include:

1. Is a particular enzyme, in any way, related to the gene I am studying?
2. What lists of IUPAC nomenclature enzymes are associated with the list of genes I have?
3. Which co-factors associated with the enzymatic activities are associated with my gene list?

The tabular format of the displayed results maps the genes to their respective enzymes and to further annotations by the *EC Classification number*, its respective *nomenclature*, and the number of genes mapped here. Unlike the earlier Smart Folders, clicking on each record displays a ***BioLT analysis report*** that is further classified by medical terms in the literature. The graphical output of selected records from the tabular output allow the user to study information pertaining to *enzymatic activities*, *catalysis*, *functional mapping to the GO database*, etc.

Related Diseases (Gene) Smart Folder

Many times a researcher would like to start from a disease term and then delve all the way down to proteins, their functions, and the related pathways. Such an approach is often used when a researcher starts with a diffuse question in his mind about a disease and the gene he is studying. Typical questions, when starting in this way include:

1. Is this particular gene in anyway related to the study of other diseases? If so, what are the treatments associated with this gene?

2. Conversely, what other gene-related treatment is this particular compound associated with?

3. Can I generate a detailed report about the gene-disease relationships with a list of genes I have?

This *smart folder* maps all the genes from the uploaded *file list* to all the known diseases. This is a very ***comprehensive report generator*** that displays links starting from a disease up to the encoded proteins, their functional domains, related compounds/drugs, the cell tissues they are associated with, etc. The tabular output displays the mapped genes to all the diseases and the link to the appropriate NCI thesaurus. As with the *smart folder* described above, *double clicking* on each record displays a ***BioLT analysis report***. Clicking on the *Graph* tab in the *element view*, the user is shown all possible relations existing to this disease. These relations range from the associated genes to the related compounds and drugs. *Right click* options provide further degrees of exploration to link with tissue types, gene-product relations, and functional domain annotations. A graphical output from multiple records in the tabular output can be created to graphically view the inter-relations between the selected records

Cancer Gene Evidence Relationships Smart Folder

Carrying the above concept further, this smart folder allows the user to completely explore the full relationship between a gene and the association with different cancers. Hence, this *smart folder* is a minor variation of the above one in that it selectively explores all the information related to cancer for a list of uploaded genes from the *file list*. The results panel in the tabular format displays the mapped *genes* to all the different kinds of *tumors/cancers*, the respective *NCI concept* and the *evidence count* (as seen in the *Co-occurrence Smart Folder*). Clicking on each report gives a complete description of the associated tumor/cancer and evidences in literature to this effect. Clicking on the *Graph* tab displays the relationship between the gene and its associated tumor. *Right click* options on the gene allows the user to explore up to the structural level of the related protein. Similarly, a *right click* option for the associated tumor allows the user to explore the disease phenomenon including *expression level changes*, *loss of heterozygosity*, *hypermethylation*, *polymorphisms* etc. Multiple records in the tabular format can be selected to get a graphical output and investigate relations between the mapped evidences.

Clinical trials (Gene) Smart Folder

A physician often starts by looking at clinical trials relating to a disease or gene. From there, the physician may further want to explore a greater level of informational granularity regarding protein-related drugs, proteins, etc. The ***PDQ*** (**P**hysician's **Q**uery **D**atabase) is a valuable resource to address such questions and the smart folder under this name links the *file list* containing gene names to this database. This is the ***ultimate resource*** for a practicing physician associated with *gene-disease research*. The ensuing results presented in the tabular format and graphical outputs allow the Physician to look at a gene under clinical trials down to any level of desired granularity. The results displayed in the tabular format map the genes to the *PDQ database*, the *disease* involved, followed by *description* and brief *remarks*. Clicking on each record provides an exhaustive description of the relation. The *Graph* tab allows the user to explore the relationship in much finer details. The graphical output allows viewing of relationships between multiple records and delving into the details of the combined relationships.

BioLT Query (Gene) Smart Folder

Performing a literature search is a chore done by every researcher. The process often involves having to combine a variety of related terms, using the Booleans (AND, OR etc.) and then poring through the huge list of retrieved documents. The process gets even complex when the user has a list of related genes and wants to classify the results of the search by popular MESH terms. Alternatively, suppose the user may wish to look for certain terms related to the genes of interest.

The *BioLT Query Smart folder* links the uploaded *file list* of genes to the literature-mining tool – ***BioLT***. The process starts by uploading the *file list into the smart folder*. Then the user feeds in the terms as 'arguments' into the 'BioLT Query' box. The results are then displayed starting with BioXM 'objects' that can be assigned to the query definition. Clicking on the arrow preceding the record displays the instances in literature that have been mined. The user can select.

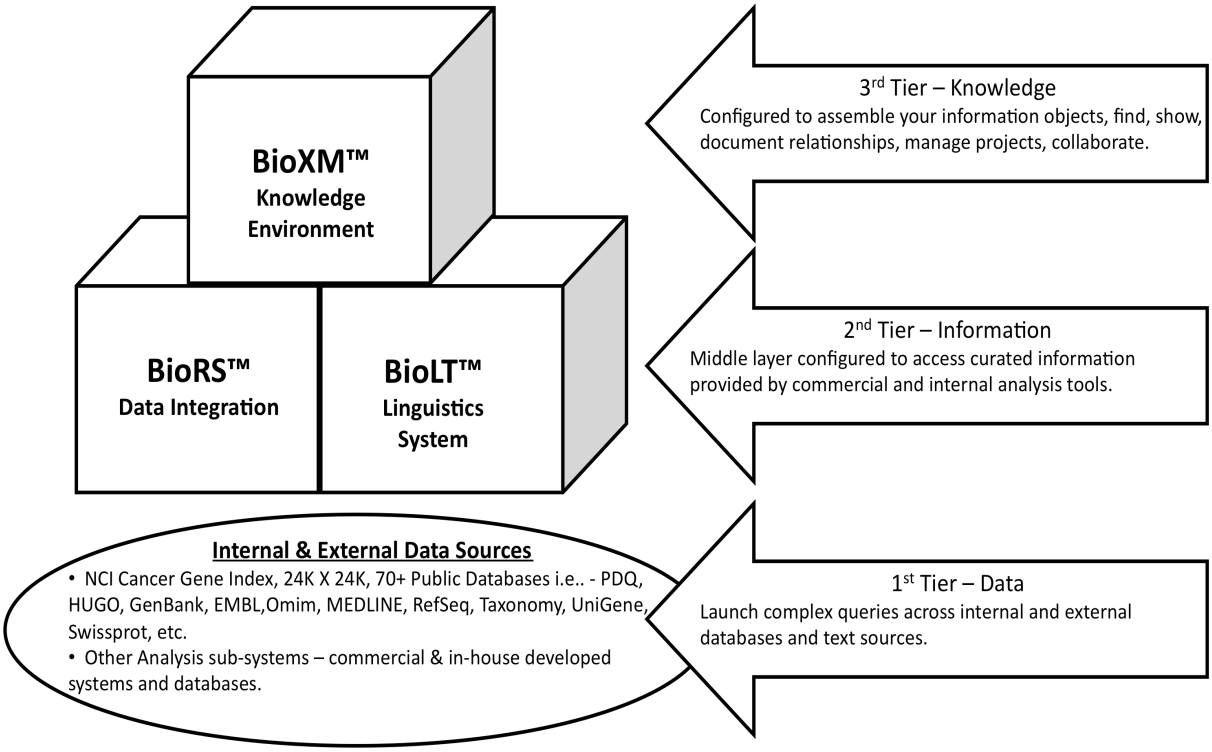
different records to obtain a graphical output by clicking on the *Graph* function present within the query results, NOT in the overall tab view. This click brings up a visual representation of the relationships between the records and allows the user to build on a gene-disease relationship going down to a very detailed level of information.

Find by Gene synonyms (Gene) Smart Folder

This *smart folder* is more of an aid to assist the user in locating the appropriate HUGO gene

names before starting to explore the CCR Common Knowledge Environment. This *smart folder* allows the user to locate the appropriate HUGO gene names using any other known gene synonyms associated with the gene. This functionality is very useful in generating a gene list on the fly and conducting a quick search for gene-disease relationships.

Overview of Software Modules



Scientific information is typically stored in either databases or free text. The foundation of the Knowledge Environment includes two modules that provide access to any database (BioRS) or text corpus (BioLT). They are the "repositories" that are integrated in BioXM, the Knowledge Integration and Visualization System.

BioLT™ Linguistic Tool

BioLT is a proven, powerful linguistics tool. Beginning in 2003, BioLT was used in a NCI-funded project to produce the first comprehensive Cancer Gene Index. BioLT has been used to text mine all Medline abstracts to identify all cancer genes, the gene-disease relationships, and gene-compound relationships with each entry manually annotated using NCI role codes and evidence codes. Teams of 15+ Ph.D. scientists used BioLT to mine over 18M Medline Abstracts leveraging the systems capabilities to analyze and parse each word in over 94M sentences.

BioLT uses the NCI Thesaurus as the foundation of the exhaustive and extensive structured dictionaries.

The BioLT™ Literature Mining Tool is a customizable application designed for intuitive and structured text mining. BioLT™ version 3.0 combines biological and medical term dictionaries with powerful free-text querying capabilities. The tool provides comprehensive and structured answers to complex questions. Search results can be used for iterative refinement and extension of queries.

BioRS™ Data Integration System

The BioRS™ Integration and Retrieval System quickly and efficiently retrieves biological data from public and proprietary databases. Multiple databases can be searched simultaneously using convenient Web interfaces. Flat-file and relational databases (Oracle, MySQL and DB2) are easily integrated using Web or command-line interfaces and standardized data formats based on XML. The BioRS application has been used as the data integration tool in a biological data management project for NCI, and in pharmaceutical and plant science companies.

BioXM™ Knowledge Management Environment

BioXM is a project-centric, distributed, client server system that provides a virtual research communication and collaboration environment. The UNIX-based system is scalable and serves as a central repository of scientific and biomedical information.

BioXM is designed to be configured to support all types of scientific and biomedical knowledge management including basic research in compounds, proteomics, genomics, genetics, biomedical translational medicine, biospecimens, pathology, oncology and epidemiology. BioXM provides a graphical single user interface that allows scientists to visualize the scientific objects to produce the pair-wise semantic or experimental relationships automatically mapped in the system.