GCELL

A SUB-CELLULAR LOCALIZATION TOOL

Rakesh Dhaval

Submitted to the faculty of the University Graduate School
In partial fulfillment of the requirements
For the degree
Master of Sciences
In the School of Informatics,
Indiana University
August, 2005

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics.

Dr. Mathew Palakal, Ph.D.

Dr. Snehasis Mukhopadhyay,Ph.D.

Dr. Douglas Perry,Ph.D.

Dr. Jake Yue Chen, Ph.D.

To my parents, Late Dr. R. M. Ravi and Mrs. Veena. They bore me, raised me, taught me, supported me, and loved me. To them I dedicate this thesis.

Acknowledgments

It is a pleasure to thank the people who made this thesis possible. I wish to express sincere appreciation to my supervisor Dr Mathew Palakal for many insightful conversations offering direction, resources and penetrating criticism during the thesis work. I would also like to thank the other members of my committee Dr. Snehasis Mukhopadhyay, Dr. Douglas Perry and Dr. Jake Chen for supporting my work, reading my thesis and offering valuable suggestions.

I wish to acknowledge my gratitude to the staff of the Department of Computer Science and the School of Informatics, and Jennifer Stewart, whose familiarity with the visualization interface was helpful during the early programming phase of this undertaking.

Many thanks to the grant from National Science Foundation which made this work possible. I am grateful to all the members of the *Sifter* group for their valuable input from time to time.

I thank for the constant support from my friend Manirupa Das who kept asking me all the year through: "Have you finished your thesis yet?" And most of all, my family and friends for their guidance, support, love and enthusiasm.

ABSTRACT

Rakesh Dhaval

GCELL - A SUB-CELLULAR LOCALIZATION TOOL

The aim of this thesis is to develop a biological database mining tool that incorporates mining of various publicly available heterogeneous databases and provides researchers with a reporting and visualization tool for sub-cellular localization of genes and proteins. Although there is little conservation of the primary structure, the general physiochemical properties are conserved to some extent among proteins that share sub-cellular location. Hence, the function of a protein is closely correlated with its sub-cellular location. Data in the field of genomics and proteomics are detailed, complex, and voluminous and distributed in heterogeneous databases. Most of the earlier work in information extraction from biological databases focused on database integration using wrapper techniques. However, little work has been done to mine specific data leading to the identification of pathway information and evolutionary relationship from heterogeneous biological databases. The need to develop an interactive information visualization tool leading to biological pathway detection for genes by using controlled vocabulary and various publicly available biological databases has led to the concept and implementation of GCell. This system provides a researcher to move from raw text data at a broader level to a much more detailed view of pathways representing complex biological interactions.

TABLE OF CONTENTS

I. INTRODUCTION

The explosion of data generated from research in life sciences during the past recent years has been on an ever-increasing note. The flood and heterogeneity of biological data from various ongoing genomic projects around the globe makes the issues of information representation, storage, structure, retrieval and interpretation critical and timely. The relevant data must provide analysis that can keep up, and which can decipher the inherent structure of information within the data. Data in this field are detailed, complex, and voluminous, representing the complete genetic blueprint for a living organism. Simultaneously, there is also a vast change in the user community accessing and benefiting from these data collections. Today as we enter the 21st century, a multitude of users ranging from ordinary students to researchers to even commercially interested parties make use of biological databanks on a daily basis to answer routine questions such as finding sequences similar to a newly sequenced gene, retrieving bibliographic references, or investigating fundamental problems of modern biology. Database systems today are facing the task of serving ever-increasing amounts of data of exponentially growing complexity. To further add to the gravity of the currently prevailing situation, the user community is also growing in terms of number and needs, nearly as fast as the data.

Information that brings further insights to the research conducted by scientists does reside in some database(s). However, it is a challenge for them to locate and select the data sources, integrate the heterogeneous data, resolve conflicts, and finally interpret the results. Mining biological data from disparate sources is a challenging problem due to the nature of biological data. Most data pertaining to genomics and molecular biology is characterized by being highly evolving and semi-structured. Moreover, genomics data available from public and private repositories is voluminous, highly heterogeneous, and not consistently represented. The heterogeneity results from (i) evolving understanding of complex biological systems, (ii) numerous disparities in modeling biological systems across organisms, across tissue, in different environments and over time and (iii) disparities across the scientific community in their understanding of these systems. A number of tools are available today for performing such operations, yet there is always the question of processing the multi-gigabytes of data in a feasible amount of time. The lack of standardization in the scientific domains and the dynamics of the data sources make the current data mining tools inadequate in biological domains.

A. SUB-CELLULAR LOCALIZATION

Eukaryotic proteins are found in membrane bound organelles. The major sites for localization in eukaryotic cells are the plasma-membrane, nucleus, mitochondria, peroxisome, endoplasmic reticulum, golgi apparatus, lysosome, endosome, and others such as chloroplasts, vacuoles and cell wall in plant cells. Proteins that share a target site also share some general characteristics in their peptide sequences, sequence length and charge distribution (Chou et al. 1999). Sub-cellular localization of proteins may be based on these characteristics. Although there is little conservation of primary structure, the general physiochemical properties are conserved to some extent among proteins that share sub-cellular location. Hence, the function of a protein is closely correlated with its sub-cellular location. Sub-cellular localization has become important in research today because of several reasons.

(i) Function is dependent on context – A gene may be present in different organisms and different tissues. They may share the function across species and across tissues. However, it is also possible that they may have different functions across organisms and across tissues. Thus function depends on the tissue where a particular gene product is active.

(ii) Localization is dynamic and changing - Localization of gene products in cells of different organisms and tissues may be the same or different. It may also be dynamic in nature and hence its products may appear to play a role at different locations within the cell.

Although sub-cellular localization is of utmost importance, specifying it is usually considered to be difficult. This is because proteins may have entirely different biological context according to the localization in the cell. The boundaries may be hard to define due to dynamic nature of some proteins. Sub-cellular localizations of gene products are assigned by wet lab experiments and machine learning / automated processes. Assignments may be done by direct assays- in-situ hybridization, high-throughput methods, prediction based on sequence (e.g. PSORT predicts proteins in mitochondria, nucleus, peroxisome, chloroplast, ER, vesicles), Bayesian methods and natural language processing (Marcotte et al 2000).

The sub-cellular localization of proteins specifies where they are, and determines their ability to interact with other proteins and small metabolites in their local environment. As sub-cellular localization of genes and proteins is a key functional characteristic, there has been much work to computationally predict the localization of proteins based on sequence and expression data (Drawid and Gerstein 2000). In addition to prediction methods, computational methods have been developed to classify sub-cellular localization based on natural language processing of existing abstracts and papers

(Stapley et al. 2002). Not only are computational techniques being developed to obtain more localization information, but new high–throughput experimental techniques are also being developed (Kumar, Agarwal et al. 2002).

B. BIOLOGICAL PATHWAYS

Modern biology has witnessed the complete sequencing of the genomes of hundreds of organisms. In recent years, this has transformed the focus and the practice of modern biology. In order to understand the biological systems, it has become important to understand the interplay of genes and their protein products. Biological pathways and networks arise from various interactions between genes and proteins. These causal pathways and networks are responsible for the development, maintenance, regulation, and responsiveness of all living systems. The complexity of biological systems comes into full view when we delve further into the analysis of the relationship between biological molecules.

Pathways may be considered to be a subset of networks comprising of a collection of interactions. In other words, a pathway can be defined as a biological network that relates to a known physiological process or phenotype. The partitioning of networks into pathways is somewhat arbitrary although it is based on the start – finish compounds. The start/finish points are chosen based on importance or easily understood compounds since it gives us the ability to conceptualize the mapping of genotype to phenotype. The interactions in a pathway can be mapped to *Enzyme–Ligand* interaction, *Protein-Protein* interaction or *Gene Products* Interaction.

Enzyme – Ligand Interaction:

E.g.- Metabolic Pathways. These include pathways involved in carbohydrate metabolism, lipid metabolism, energy metabolism, nucleotide metabolism and amino acid metabolism.

Protein – Protein Interaction

E.g.- Cell Signaling Pathways and complexes for cell processes. These include pathways involved in broad effects on biological processes with specific receptors, signaling within local tissues, and neuronal signaling.

Gene Regulatory Elements - Gene Products Interaction

E.g.- Genetic Networks. These include pathways involved in genetic information processing (transcription, translation, replication, repair, etc), environmental information processing (membrane transport, signal transduction, etc) and cellular processes such as cell motility, cell communication, cell growth and death)

All the pathways are interlinked. They represent metabolites involved, enzymes/transport proteins, order of reactions, general biological function, reaction rates, expression data, inhibitors, activators, alternate pathways and genetic regulatory information. Pathways involve multiple enzymes, which may have multiple subunits, alternate forms and alternate specificities. Each enzyme may be involved in multiple pathways. For example - malate dehydogenase appears in 6 different pathways.

A lot of experimental data elucidating various aspects of biological interactions is being generated at an ever-increasing rate. Such information contains patterns that reflect the dynamics of pathway, and hence can be used to deduce causal pathway structures. Hence, intelligent analysis and mining of high-throughput functional genomics data may lead to infer pathways and their regulation.

ORTHOLOGS AND PARALOGS

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one. Homologs are genes related to a second gene by descent from a common ancestral DNA sequence. The term, *homolog*, may apply to the relationship between genes separated by the event of speciation or to the relationship between genes separated by the event of genetic duplication. An example from NCBI explaining the concept of orthologs and paralogs is shown in Figure 1 below. Gene duplication of “early globin gene” results in alpha-chain gene and beta-chain gene. Both the duplicated genes (alpha-chain and beta-chain) in different species like frog, chick and mouse may or may not have similar functions; however, they derive from a common ancestor. Hence they may be considered as orthologs. Mouse-alpha and mouse-beta are paralogous genes within a single species that diverged by gene duplication.

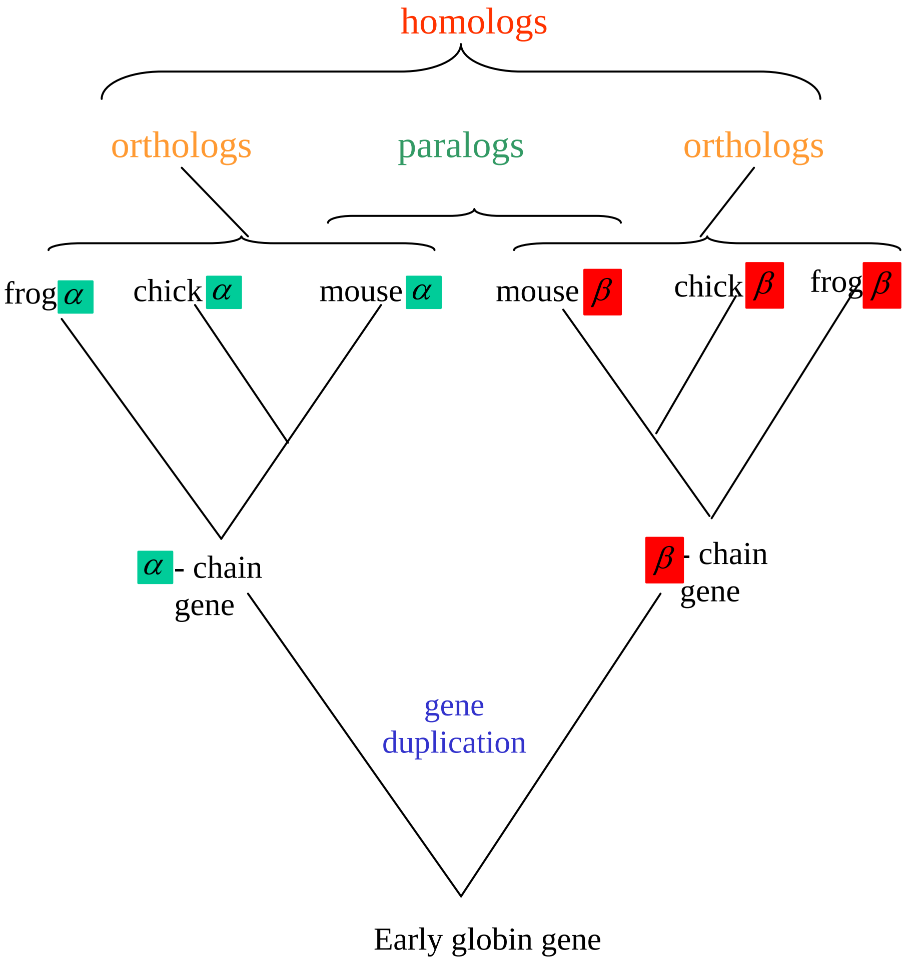


Figure1. Orthologs and Paralogs

C. NEED FOR ONTOLOGY

With an increasing amount of sub-cellular localization data, methods of standardizing data across different sources have become important. This need has been addressed by the Gene Ontology (GO) cellular component terms (http://www.geneontology.org). GO provides a hierarchy of cellular component concepts such as organelles, membranes and protein complexes. These concepts are used to annotate the location of gene products in a standard fashion. However, there remains a need to capture more aspects of the biological properties of cellular components in a standardized and computationally efficient way.

D. DATA HETEROGENEITY

Researchers invest a lot of time and effort in finding all the relevant information about genes and gene products of their interest. Data sources are spread across different communities and in various formats. The sources of heterogeneities originate from the storage of the data using different database systems (e.g. DBMS, semantic heterogeneity), operating systems (e.g. file systems, file types), and hardware (e.g. data formats and representation). This problem is further complicated due to the variations in the use of terminologies. The need to unify the data and provide reliable and accurate outputs to the user queries is urgent. Moreover, most of the current data sources only provide web links for navigational access along a network of biological relationships, making it impossible to perform automated analysis for large data sets across multiple dimensions. To facilitate a *one stop shopping* of heterogeneous data sources, a common approach is to perform semantic database integration.

E. MULTIPLE FORMAT DATABASES

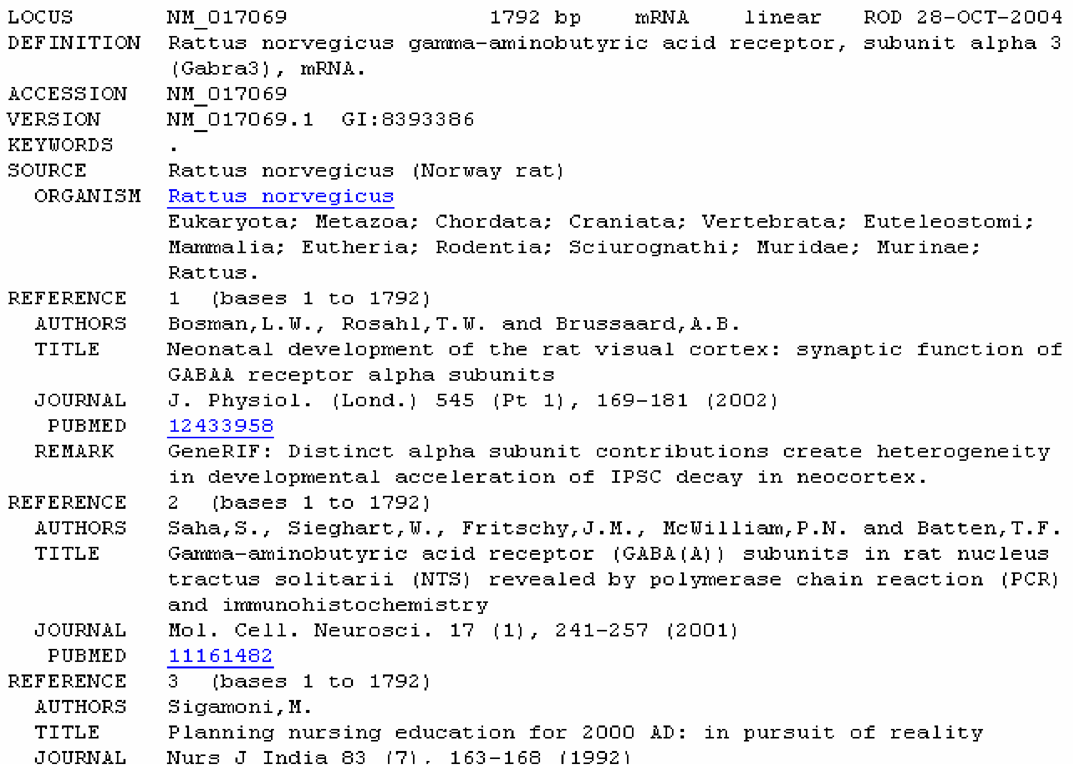
Biological data are usually organized in many different manners. These include (i) Flat text files databases, (ii) Relational databases, and (iii) Object oriented databases. In the following section, flat text files and relational databases are discussed briefly because these are the formats in which the data is organized in Gcell system.

Flat Text Files

Flat text file entries are stored in text. These text fields/attributes may be labeled with identifiers. They may use standard vocabulary for values of attributes (or not).

Search is performed by string matching of patterns using regular expressions. However, they can be indexed for faster search. Although they are easy to import/export and are not platform dependent, it becomes difficult to perform complicated queries with flat files.

Example: Below (Figure 2) is a part of a flat text file from Genbank.



```
LOCUS       NM_017069               1792 bp    mRNA    linear   ROD 28-OCT-2004
DEFINITION  Rattus norvegicus gamma-aminobutyric acid receptor, subunit alpha 3
            (Gabra3), mRNA.
ACCESSION   NM_017069
VERSION     NM_017069.1  GI:8393386
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae;
            Rattus.
REFERENCE   1  (bases 1 to 1792)
  AUTHORS   Bosman,L.W., Rosahl,T.W. and Brussaard,A.B.
  TITLE     Neonatal development of the rat visual cortex: synaptic function of
            GABAA receptor alpha subunits
  JOURNAL   J. Physiol. (Lond.) 545 (Pt 1), 169-181 (2002)
   PUBMED   12433958
  REMARK    GeneRIF: Distinct alpha subunit contributions create heterogeneity
            in developmental acceleration of IPSC decay in neocortex.
REFERENCE   2  (bases 1 to 1792)
  AUTHORS   Saha,S., Sieghart,W., Fritschy,J.M., McWilliam,P.N. and Batten,T.F.
  TITLE     Gamma-aminobutyric acid receptor (GABA(A)) subunits in rat nucleus
            tractus solitarii (NTS) revealed by polymerase chain reaction (PCR)
            and immunohistochemistry
  JOURNAL   Mol. Cell. Neurosci. 17 (1), 241-257 (2001)
   PUBMED   11161482
REFERENCE   3  (bases 1 to 1792)
  AUTHORS   Sigamoni,M.
  TITLE     Planning nursing education for 2000 AD: in pursuit of reality
  JOURNAL   Nurs J India 83 (7), 163-168 (1992)
```

Figure2. Flat text file from Genbank

Relational Database

A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. The standard user and application program interface to a relational database is the structured query language (SQL). SQL statements are used both for interactive queries for information from a relational database and for gathering data for reports.

A relational database has an important advantage of being easy to extend, other than being relatively easy to create and access. Even if the original database is already created, a new data category can be added without requiring that all existing applications be modified. A relational database is a set of tables containing data fitted into predefined categories. Each table (relation) contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns. For example, a typical database for gene information would include a table that describes a gene with columns for name, synonyms, symbols, function and so forth. Another table would describe the sequences: sequence, length, and so forth. A user of the database can obtain a view of the database that fits the user's needs. For example, a biological researcher might like a view or report on all genes that have a particular function. Another researcher might be interested in all the proteins and its sequences that a particular gene encodes for. Hence, a relational database can be considered to be very flexible and scalable.

Mining meaningful attributes from various biological databases is difficult due to a number of reasons. The vocabularies may not be shared leading to different terms for same concept or same term for different concepts, hence leading to different dependencies in the data. The queries in different forms of databases are very different from each other. Flat files may use text searches; RDBMS utilizes SQL whereas Object Oriented Databases makes use of the Object Oriented SQL (OOSQL). Hence, gathering of relevant data from heterogeneous sources with all incompatibilities *scrubbed out* becomes important.

F. INFORMATION VISUALIZATION

The exponential growth of the web is a dramatic demonstration of how information can be made more accessible by incorporating visualization techniques. Information visualization enables people to deal with all of this information by taking advantage of our innate visual perception capabilities. Presentation of information in graphical form makes it possible for the human brain to use more of its perceptual system in initially processing information, rather than immediately relying entirely on the cognitive system (Herman et al 2000). Information visualization applications rely on basic features that the human perceptual system inherently assimilates very quickly: color, size, shape, proximity, and motion. Pictorial display and representation of query results is necessary for large-scale genomic data. Visual tools now comprise as important components of any database systems and go far ahead than just replacing command-line queries through buttons and pull-down menus and displaying the retrieval results in a scrollable window. These allow the creation of complex views of large amounts of inter-related data, present various types of evidence in required context (e.g. genes together with regulatory elements), and increase the productivity of data mining.

Good visualizations enable us to not only perceive information more easily but also to perceive more information at one time. We can immediately see patterns in data that indicate trends and patterns. Information visualization applications enable us to better understand complex systems, make better decisions, and discover information that might otherwise remain unknown. It becomes difficult for researchers to visualize the location in cell where a particular gene and subsequently the associated gene products is active.

The use of consistent descriptions about biological objects and visualization of genes and gene products in the context of cell is of utmost importance to the researchers.

II. BACKGROUND

There are several databases that are available in public domain for searches on biological entities specially genes and gene products. Each has a different database format and schema. They have different interfaces to display the information. Some of the earlier works that have dealt in displaying information about genes and gene products have been described in the following paragraphs.

Biocarta (Galperin 2004) is a wrapper-based approach to integrate various heterogeneous databases. It is a web-based resource of information on gene function, proteomic pathways, and reagent exchange. It is a forum for information exchange and collaboration between researchers. Search on Biocarta may be formulated according to organism, area of research, or keyword search - using multiple online databases. Gene-specific information, including sequence data, publications and reviews, disease correlation, and interrelationship with other proteomic pathways can be located. The maps depict molecular relationships from areas of active research. It constantly integrates emerging proteomic information from the scientific community. It also catalogs and summarizes important resources providing information for over 120,000 genes from multiple species. Both classical pathways as well as current suggestions for new pathways may be found here. [http://www.biocarta.com]

Affymetrix® Analysis Data Model (AADM) (http://www.affymetrix.com) has a relational database schema that Affymetrix uses to store experiment results. It includes tables to support mapping and expression results. Affymetrix publishes AADM to support open access to experiment information generated and managed by Affymetrix software so the results may be filtered and mined with compatible analysis tools. The NetAffx™ Analysis Center enables researchers to correlate their GeneChip® array results with array design and annotation information. This resource provides with unprecedented access to array content information, including probe sequences and gene annotations. One can quickly search for genes and/or SNPs, compare and refine results, and export data into Excel friendly formats. [http://www.affymetrix.com/analysis/index.affx.

Universal Protein Resource (Uniprot) (Apweiler et al 2004) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. For convenient sequence searches, UniProt also provides several non-redundant sequence databases. The UniProt NREF (UniRef) databases provide representative subsets of the knowledgebase suitable for efficient searching. The comprehensive UniProt Archive (UniParc) is updated daily from many public source databases. The UniProt databases can be accessed online (http://www.uniprot.org) or downloaded in several formats (ftp://ftp.uniprot.org/pub). Uniprot provides high-quality annotation, manual annotation by curators based on literature and sequence analysis, standardized nomenclature and controlled vocabularies and integration with other databases with minimal redundancy. In short, UniProt is used to facilitate knowledge discovery by allowing researchers to integrate the enormous amount of data from the Human Genome Project and from structural and functional genomics and proteomics.[http://www.pir.uniprot.org/]

Biological And Chemical Information Integration System (BACIIS) (Ben Miled et al 2001) is a tightly coupled federated database system that uses wrappers to extract information from remote databases. The type of integration approach consists of mediator-wrapper architecture. The wrapper in this approach plays the role of extracting information from a given remote database and the mediator integrates the information retrieved from different remote databases. However, it currently relies on hand-constructed wrappers.

Investigators today, have to interpret many types of information from a variety of sources. These may include data generated from lab instruments, public databases, gene expression profiles, raw sequence traces, single nucleotide polymorphisms, chemical screening data, proteomics data, putative metabolic pathway models, and many others. In order to find new discoveries one needs a large set of genetic information to generate valid leads. In order to find valid leads, one needs to study gene function. It has become necessary to integrate all this information.

In order to get a better understanding of the molecular mechanisms for disease, metabolic and regulatory biochemical pathways must be inferred from this information. And finally, to stimulate the discovery of breakthrough healthcare products, therapeutics must be developed and tested in a pre-clinical, then clinical environment. Ongoing long-term clinical research must feed back into the discovery process as well. This would further generate more not-integrated data. Researchers in the same organization might be considering the problem, indeed the same gene, but in different domains and using different names for that gene, they would never make the connection.

A. LIFE SCIENCES DATABASES

Life science researchers today, make use of various publicly available databases like Pubmed, Uniprot, Gene, Kegg, Gene Ontology, Locuslink, BIND, and ENZYME. A brief description about the contents of each of these is described below.

KEGG

Kyoto Encyclopedia for Genes and Genomes (KEGG) (Kanehisa et al 2002) is an online genomic database that serves as a Bioinformatics resource for understanding higher order functional meanings and utilities of a living cell or the organism as a whole, from its genome information. It consists of a collection of databases and supporting software components, integrating our current knowledge on molecular interaction networks in biological processes (PATHWAY database), the information about the universe of genes and proteins (GENES/SSDB/KO databases), and the information about the universe of chemical compounds and reactions (COMPOUND/REACTION databases).

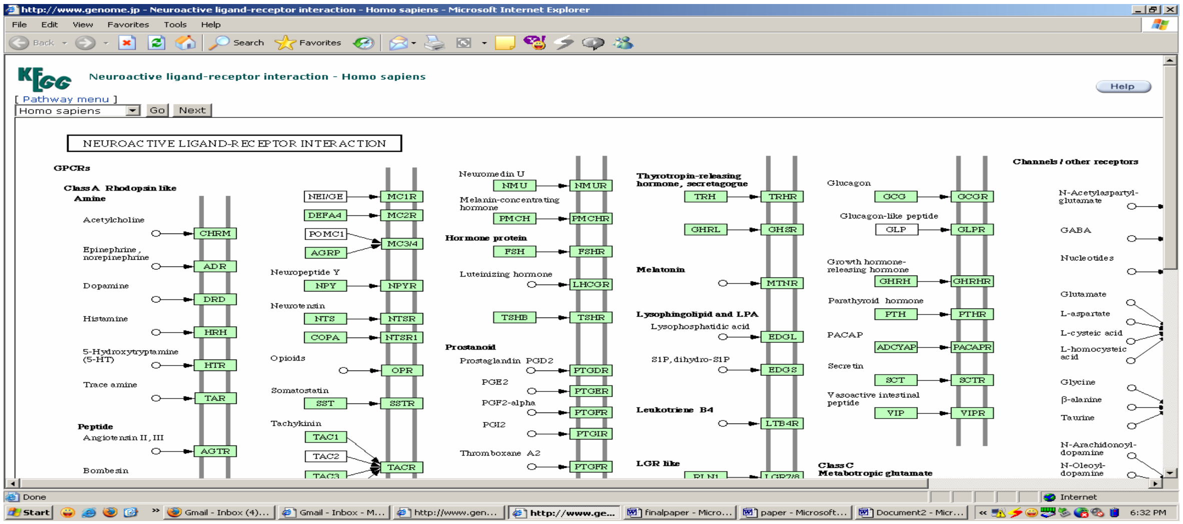
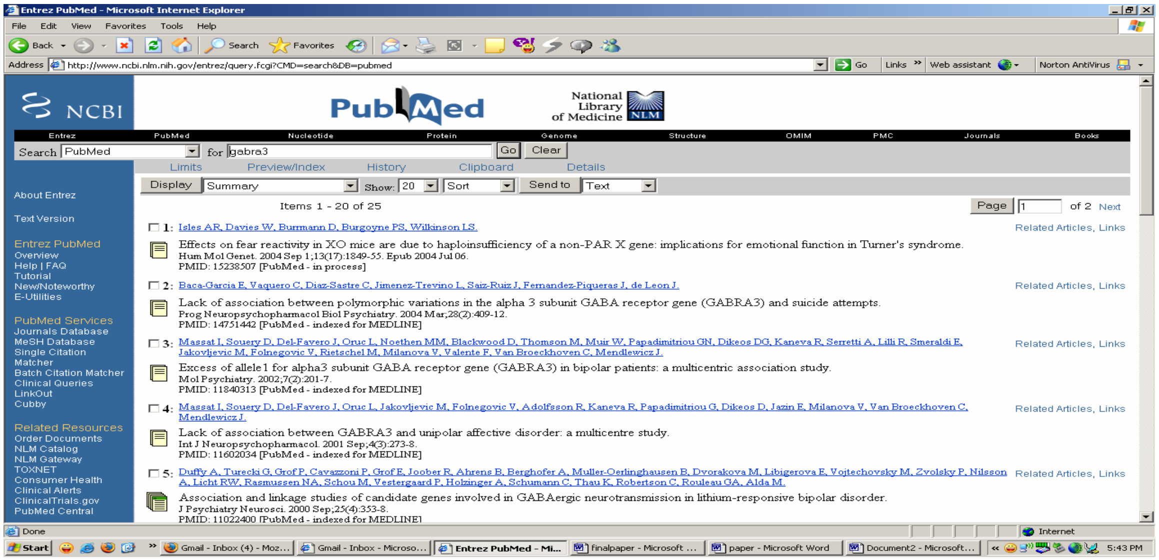


Figure 3. Snapshot of KEGG database

MEDLINE

MEDLINE (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, and the preclinical sciences. The articles are indexed and their citations are searchable. NLM's controlled vocabulary MeSH (Medical Subject Headings) is made use of in classifying various biological objects and hence in finding associations between them.


Entrez PubMed - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed
Go Links Web assistant Norton AntiVirus
NCBI
PubMed
National Library of Medicine NLM
Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search PubMed for gabra3 Go Clear
Limits Preview/Index History Clipboard Details
Display Summary Show: 20 Sort Send to Text
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
Page 1 of 2 Next
Done
Internet
Start
Gmail - Inbox (4) - Moz...
Gmail - Inbox - Microso...
Entrez PubMed - Mi...
finalpaper - Microsoft ...
paper - Microsoft Word
Document2 - Microsoft...
5:43 PM

Items 1 - 20 of 25

□ 1: Isles AR, Davies W, Burrmann D, Burgoyne PS, Wilkinson LS. Related Articles, Links

Effects on fear reactivity in XO mice are due to haploinsufficiency of a non-PAR X gene: implications for emotional function in Turner's syndrome.
Hum Mol Genet. 2004 Sep 1;13(17):1849-55. Epub 2004 Jul 06.
PMID: 15238507 [PubMed - in process]

□ 2: Baca-Garcia E, Vaquero C, Diaz-Sastre C, Jimenez-Trevino L, Saiz-Ruiz J, Fernandez-Piqueras J, de Leon J. Related Articles, Links

Lack of association between polymorphic variations in the alpha 3 subunit GABA receptor gene (GABRA3) and suicide attempts.
Prog Neuropsychopharmacol Biol Psychiatry. 2004 Mar;28(2):409-12.
PMID: 14751442 [PubMed - indexed for MEDLINE]

□ 3: Massat I, Souery D, Del-Favero J, Oruc L, Noethen MM, Blackwood D, Thomson M, Muir W, Papadimitriou GN, Dikeos DG, Kaneva R, Serretti A, Lilli R, Smeraldi E, Jakovljevic M, Folnegovic V, Rietschel M, Milanova V, Valente F, Van Broeckhoven C, Mendlewicz J. Related Articles, Links

Excess of allele1 for alpha3 subunit GABA receptor gene (GABRA3) in bipolar patients: a multicentric association study.
Mol Psychiatry. 2002;7(2):201-7.
PMID: 11840313 [PubMed - indexed for MEDLINE]

□ 4: Massat I, Souery D, Del-Favero J, Oruc L, Jakovljevic M, Folnegovic V, Adolfsson R, Kaneva R, Papadimitriou G, Dikeos D, Jazin E, Milanova V, Van Broeckhoven C, Mendlewicz J. Related Articles, Links

Lack of association between GABRA3 and unipolar affective disorder: a multicentre study.
Int J Neuropsychopharmacol. 2001 Sep;4(3):273-8.
PMID: 11602034 [PubMed - indexed for MEDLINE]

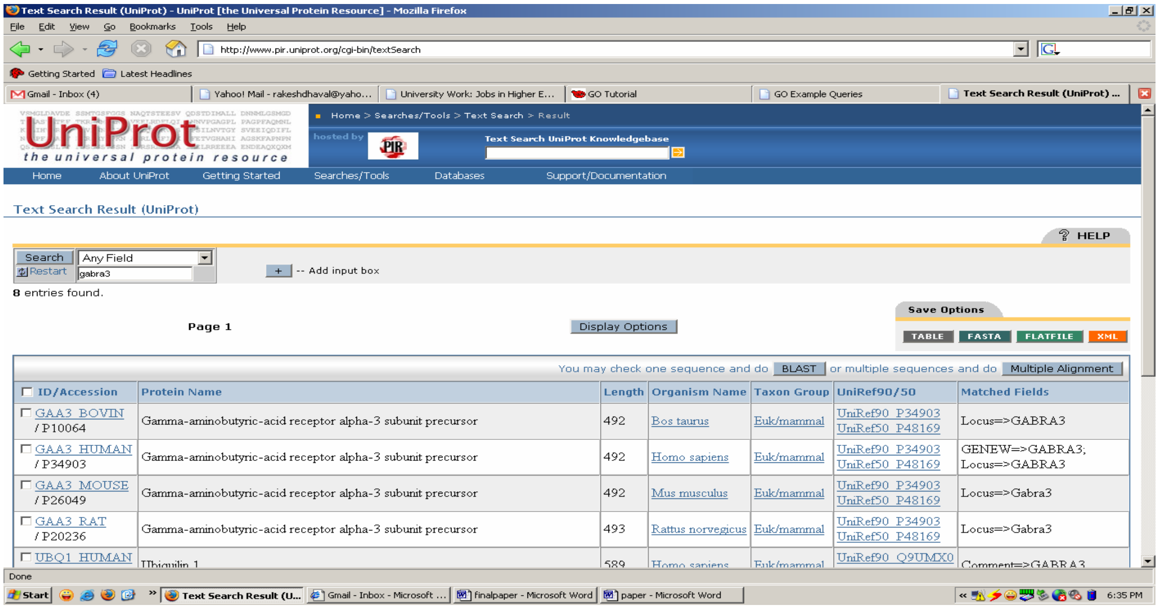
□ 5: Duffy A, Turecki G, Grof P, Cavazzoni P, Grof E, Joober R, Ahrens B, Berghofer A, Muller-Oerlinghausen B, Dvorakova M, Libigerova E, Vojtechovsky M, Zvolsky P, Nilsson A, Licht RW, Rasmussen NA, Schou M, Vestergaard P, Holzinger A, Schumann C, Thau K, Robertson C, Rouleau GA, Alda M. Related Articles, Links

Association and linkage studies of candidate genes involved in GABAergic neurotransmission in lithium-responsive bipolar disorder.
J Psychiatry Neurosci. 2000 Sep;25(4):353-8.
PMID: 11022400 [PubMed - indexed for MEDLINE]

Figure 4. Snapshot of PubMed database

UNIPROT

UniProt (Universal Protein Resource) (Apweiler et al 2004) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. UniProt Knowledgebase is the central access point for extensive curated protein information, including function, classification, and cross-reference.


Text Search Result (UniProt) - UniProt [the Universal Protein Resource] - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.pir.uniprot.org/cgi-bin/textSearch
Getting Started
Latest Headlines
Gmail - Inbox (4)
Yahoo! Mail - rakeshdhaval@yaho...
University Work: Jobs in Higher E...
GO Tutorial
GO Example Queries
Text Search Result (UniProt) ...
UniProt
the universal protein resource
Home > Searches/Tools > Text Search > Result
hosted by
PIR
Text Search UniProt Knowledgebase
Home
About UniProt
Getting Started
Searches/Tools
Databases
Support/Documentation
HELP
Search
Restart
Any Field
+
-- Add input box
Display Options
Save Options
TABLE
FASTA
FLATFILE
XML
BLAST
Multiple Alignment
Done
Start
Text Search Result (U...
Gmail - Inbox - Microsoft ...
finalpaper - Microsoft Word
paper - Microsoft Word
6:35 PM

Text Search Result (UniProt)

gabra3

8 entries found.

Page 1

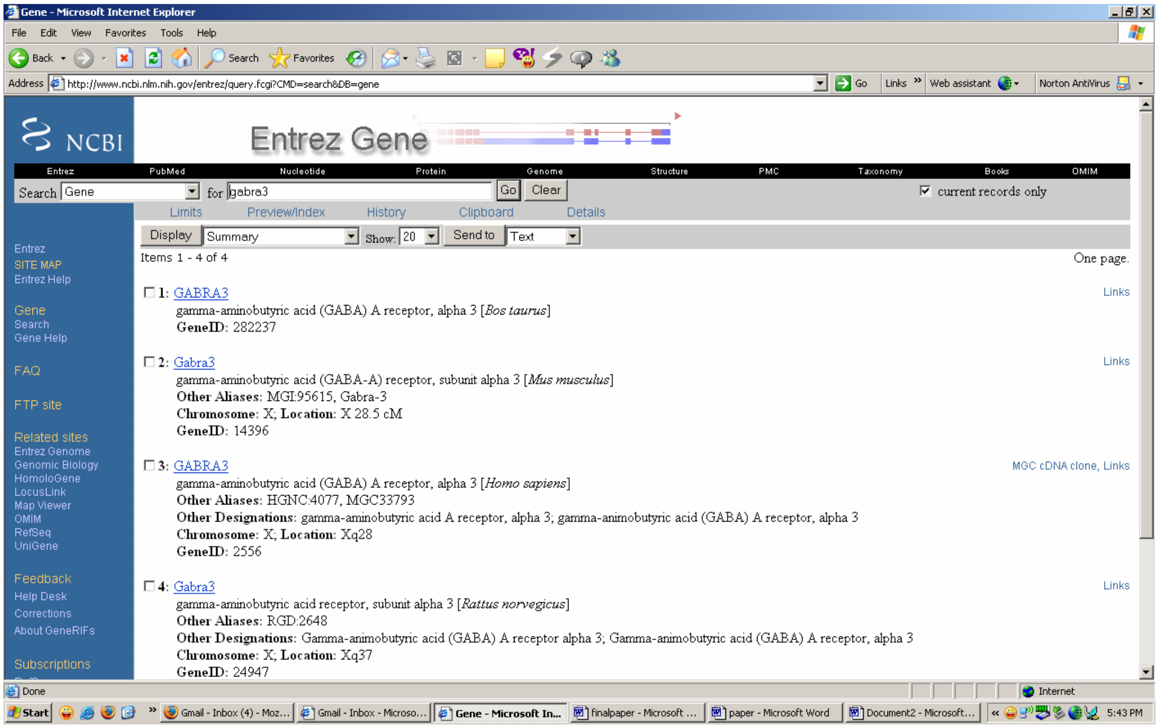
You may check one sequence and do BLAST or multiple sequences and do Multiple Alignment

□ ID/Accession	Protein Name	Length	Organism Name	Taxon Group	UniRef90/50	Matched Fields
□ GAA3_BOVIN / P10064	Gamma-aminobutyric-acid receptor alpha-3 subunit precursor	492	Bos taurus	Euk/mammal	UniRef90_P34903 UniRef50_P48169	Locus=>GABRA3
□ GAA3_HUMAN / P34903	Gamma-aminobutyric-acid receptor alpha-3 subunit precursor	492	Homo sapiens	Euk/mammal	UniRef90_P34903 UniRef50_P48169	GENEW=>GABRA3; Locus=>GABRA3
□ GAA3_MOUSE / P26049	Gamma-aminobutyric-acid receptor alpha-3 subunit precursor	492	Mus musculus	Euk/mammal	UniRef90_P34903 UniRef50_P48169	Locus=>Gabra3
□ GAA3_RAT / P20236	Gamma-aminobutyric-acid receptor alpha-3 subunit precursor	493	Rattus norvegicus	Euk/mammal	UniRef90_P34903 UniRef50_P48169	Locus=>Gabra3
□ UBQ1_HUMAN	Ubiquilin 1	589	Homo sapiens	Euk/mammal	UniRef90_Q9UMX0	Comment=>GABRA3

Figure 5. Snapshot of Uniprot database

GENE

Entrez Gene (http://www.ncbi.nlm.nih.gov) supplies key connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. Unique identifiers are assigned to any gene with defining sequence, genes with known map positions, and genes inferred from phenotypic information. These gene identifiers are tracked, and information is added when available. Entrez Gene is considered as the successor to LocusLink. The major difference between them is the greater scope of Gene database. Gene provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer. Gene supports query on names, symbols, accessions, publications, GO terms, chromosome numbers, E.C. numbers, and many other attributes associated with genes and the products they encode.



Gene - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene
Go Links Web assistant Norton AntiVirus
NCBI
Entrez Gene
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM
Search Gene for gabra3 Go Clear ☑ current records only
Limits Preview/Index History Clipboard Details
Display Summary Show: 20 Send to Text
Entrez
SITE MAP
Entrez Help
Gene
Search
Gene Help
FAQ
FTP site
Related sites
Entrez Genome
Genomic Biology
HomoloGene
LocusLink
Map Viewer
OMIM
RefSeq
UniGene
Feedback
Help Desk
Corrections
About GeneRIFs
Subscriptions
Done
Internet
Start
Gmail - Inbox (4) - Moz...
Gmail - Inbox - Microso...
Gene - Microsoft In...
finalpaper - Microsoft ...
paper - Microsoft Word
Document2 - Microsoft...
5:43 PM

Items 1 - 4 of 4 One page.

□ 1: GABRA3 Links
gamma-aminobutyric acid (GABA) A receptor, alpha 3 [*Bos taurus*]
GeneID: 282237

□ 2: Gabra3 Links
gamma-aminobutyric acid (GABA-A) receptor, subunit alpha 3 [*Mus musculus*]
Other Aliases: MGI:95615, Gabra-3
Chromosome: X; **Location**: X 28.5 cM
GeneID: 14396

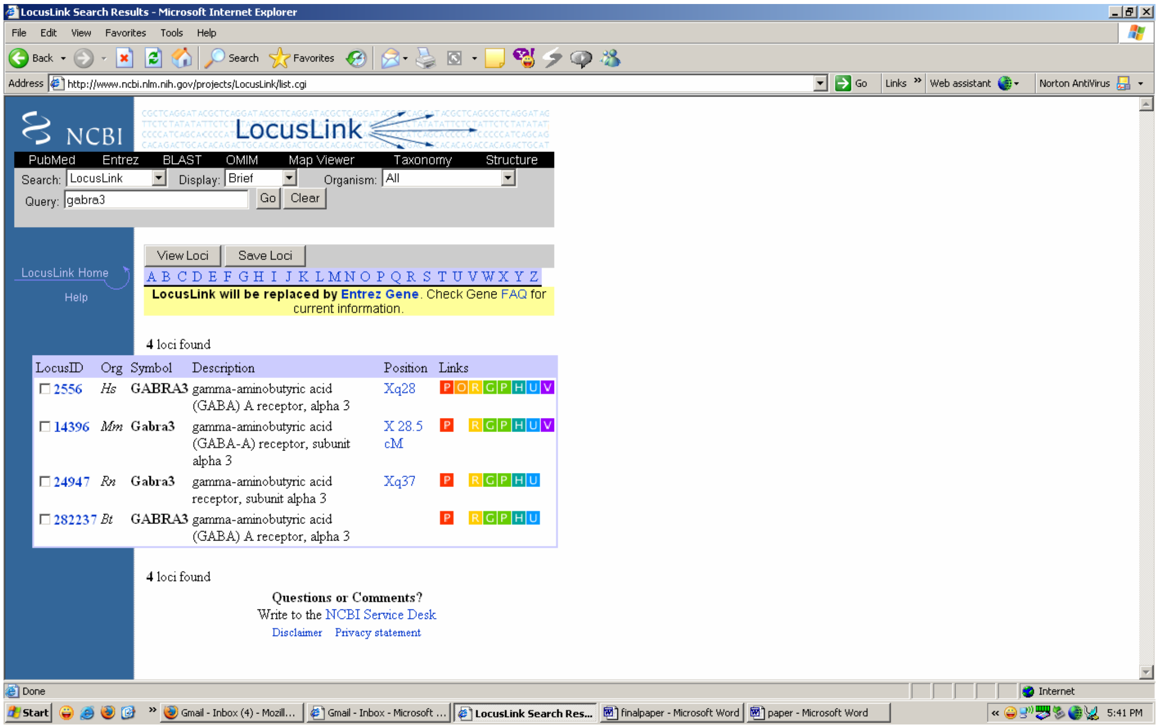
□ 3: GABRA3 MGC cDNA clone, Links
gamma-aminobutyric acid (GABA) A receptor, alpha 3 [*Homo sapiens*]
Other Aliases: HGNC:4077, MGC33793
Other Designations: gamma-aminobutyric acid A receptor, alpha 3; gamma-animobutyric acid (GABA) A receptor, alpha 3
Chromosome: X; **Location**: Xq28
GeneID: 2556

□ 4: Gabra3 Links
gamma-aminobutyric acid receptor, subunit alpha 3 [*Rattus norvegicus*]
Other Aliases: RGD:2648
Other Designations: Gamma-animobutyric acid (GABA) A receptor alpha 3; Gamma-animobutyric acid (GABA) A receptor, alpha 3
Chromosome: X; **Location**: Xq37
GeneID: 24947

Figure 6. Snapshot of Gene database

LOCUSLINK

LocusLink (http://www.ncbi.nlm.nih.gov) provides a query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites. This database is not being updated now, and has been replaced by Entrez Gene database.


LocusLink Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/projects/LocusLink/list.cgi
Go Links Web assistant Norton AntiVirus
NCBI LocusLink
PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure
Go Clear
LocusLink Home
Help
View Loci Save Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Done
Internet
Start Gmail - Inbox (4) - Mozill... Gmail - Inbox - Microsoft ... LocusLink Search Res... finalpaper - Microsoft Word paper - Microsoft Word
5:41 PM

Search: LocusLink Display: Brief Organism: All

Query: gabra3

LocusLink will be replaced by Entrez Gene. Check Gene FAQ for current information.

4 loci found

LocusID	Org	Symbol	Description	Position	Links
□ 2556	*Hs*	**GABRA3**	gamma-aminobutyric acid (GABA) A receptor, alpha 3	Xq28	P O R G P H U V
□ 14396	*Mm*	**Gabra3**	gamma-aminobutyric acid (GABA-A) receptor, subunit alpha 3	X 28.5 cM	P R G P H U V
□ 24947	*Rn*	**Gabra3**	gamma-aminobutyric acid receptor, subunit alpha 3	Xq37	P R G P H U
□ 282237	*Bt*	**GABRA3**	gamma-aminobutyric acid (GABA) A receptor, alpha 3		P R G P H U

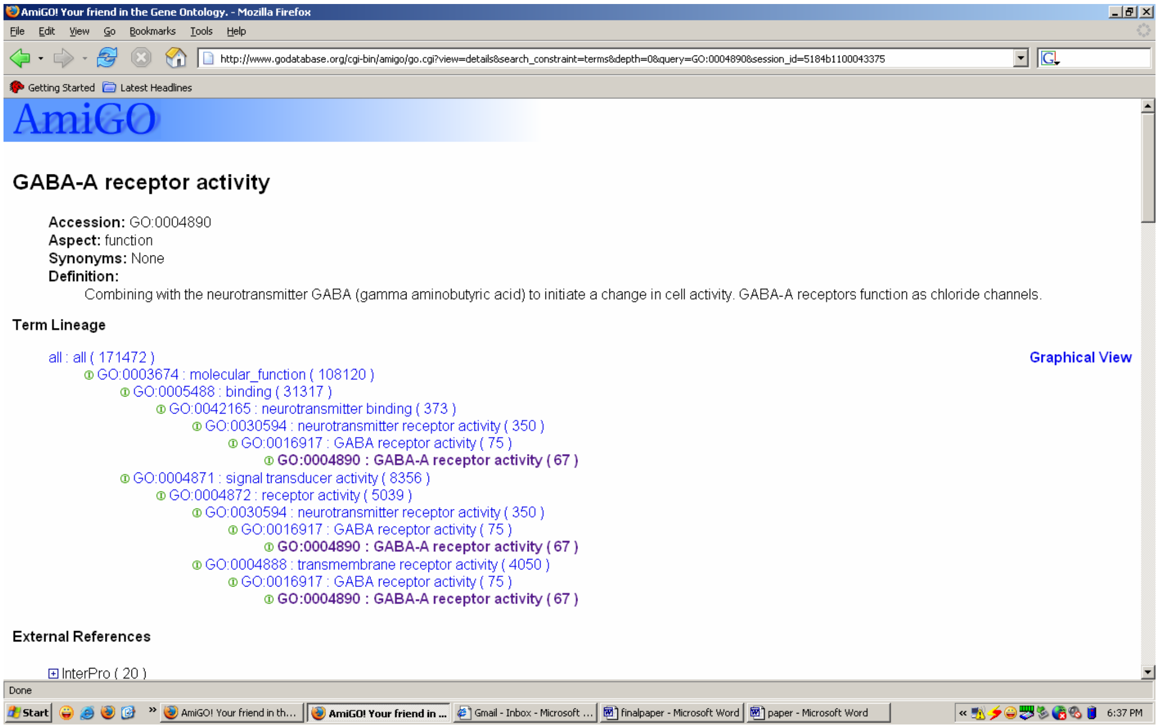
4 loci found

Questions or Comments?
Write to the NCBI Service Desk
Disclaimer Privacy statement

Figure 7. Snapshot of LocusLink database

GENE ONTOLOGY

The Gene Ontology (GO) (Ashburner et al 2000) project provides a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism. The GO includes three extensive ontologies to describe molecular function, biological process, and cellular component, and provides a community database resource that supports the use of these ontologies. The GO Consortium was initiated by scientists associated with three model organism databases: SGD, the Saccharomyces Genome database; FlyBase, the Drosophila genome database; and MGD/GXD, the Mouse Genome Informatics databases. Additional model organism database groups are joining the project. Each of these model organism information systems is annotating genes and gene products using GO vocabulary terms and incorporating these annotations into their respective model organism databases. Each database contributes its annotation files to a shared GO data resource accessible to the public at http://www.geneontology.org/. GO is used by the community to retrieve the GO vocabularies and to access the annotated gene product data sets from the model organism databases. The GO Consortium supports the development of the GO database resource and provides tools enabling curators and researchers to query and manipulate the vocabularies. This molecular annotation resource is intended to contribute to the unification of biological information.


AmiGO! Your friend in the Gene Ontology. - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.godatabase.org/cgi-bin/amigo/go.cgi?view=details&search_constraint=terms&depth=0&query=GO:0004890&session_id=5184b1100043375
Getting Started
Latest Headlines
AmiGO
Done
Start
AmiGO! Your friend in th...
AmiGO! Your friend in ...
Gmail - Inbox - Microsoft ...
finalpaper - Microsoft Word
paper - Microsoft Word
6:37 PM

GABA-A receptor activity

Accession: GO:0004890
Aspect: function
Synonyms: None
Definition:
Combining with the neurotransmitter GABA (gamma aminobutyric acid) to initiate a change in cell activity. GABA-A receptors function as chloride channels.

Term Lineage

Graphical View

- all : all (171472)
  - GO:0003674 : molecular_function (108120)
    - GO:0005488 : binding (31317)
      - GO:0042165 : neurotransmitter binding (373)
        - GO:0030594 : neurotransmitter receptor activity (350)
          - GO:0016917 : GABA receptor activity (75)
            - **GO:0004890 : GABA-A receptor activity (67)**
    - GO:0004871 : signal transducer activity (8356)
      - GO:0004872 : receptor activity (5039)
        - GO:0030594 : neurotransmitter receptor activity (350)
          - GO:0016917 : GABA receptor activity (75)
            - **GO:0004890 : GABA-A receptor activity (67)**
        - GO:0004888 : transmembrane receptor activity (4050)
          - GO:0016917 : GABA receptor activity (75)
            - **GO:0004890 : GABA-A receptor activity (67)**

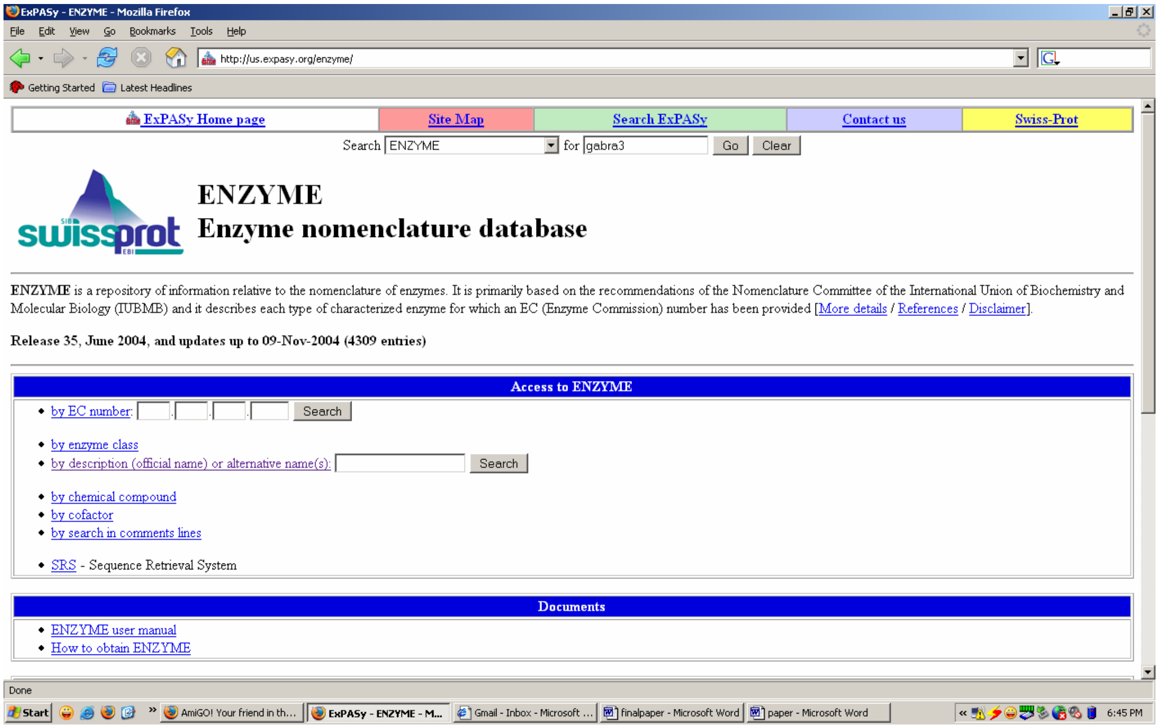
External References

⊞ InterPro (20)

Figure 8. Snapshot of Gene Ontology database

EC-ENZYME

The EC-ENZYME (http://us.expasy.org/enzyme/) data bank contains several data related to the enzyme for each type of characterized enzyme for which an EC number has been provided. It contains the following: EC number, Recommended name, Alternative names, Catalytic activity, Cofactors, Pointers to the SWISS-PROT entries that correspond to the enzyme, Pointers to disease(s) associated with a deficiency of the enzyme.


ExPASy - ENZYME - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://us.expasy.org/enzyme/
Getting Started Latest Headlines
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Go
Clear
swissprot
Done
Start
AmiGO! Your friend in th...
ExPASy - ENZYME - M...
Gmail - Inbox - Microsoft ...
finalpaper - Microsoft Word
paper - Microsoft Word
6:45 PM

Search ENZYME for gabra3

ENZYME
Enzyme nomenclature database

ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided [More details / References / Disclaimer].

Release 35, June 2004, and updates up to 09-Nov-2004 (4309 entries)

Access to ENZYME

- by EC number: Search
- by enzyme class
- by description (official name) or alternative name(s): Search
- by chemical compound
- by cofactor
- by search in comments lines
- SRS - Sequence Retrieval System

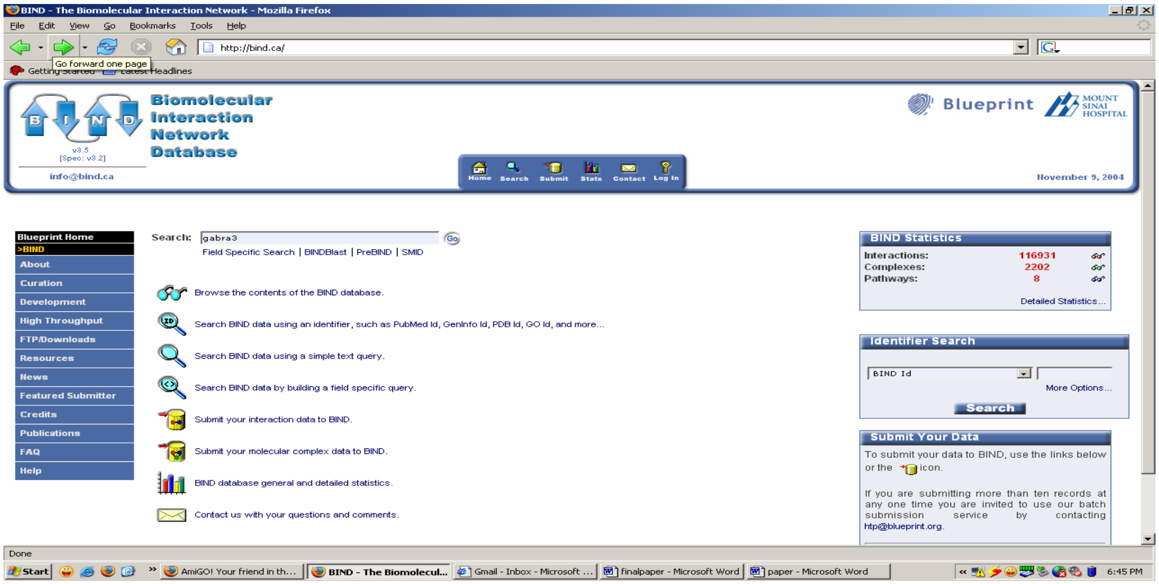
Documents

- ENZYME user manual
- How to obtain ENZYME

Figure 9. Snapshot of Enzyme database

BIND

Bio-molecular Interaction Network Database (BIND) contains information about bio-molecular interaction, molecular complex and pathway information using the internationally standard ASN.1 syntax and XML/DTD. In order to query, view and submit records, a web-based system is available (Bader et al 2003). The database grows with the addition of individual submissions as well as interaction data from the PDB and other complex mapping experiments. A graphical analysis tool provides users with a view of the domain composition of proteins in interaction. It helps to relate functional domains to protein interactions. BIND also has the ability to store detailed information about genetic interactions.



Search: gabra3

Browse the contents of the BIND database.

Search BIND data using an identifier, such as PubMed Id, GenInfo Id, PDB Id, GO Id, and more...

Search BIND data using a simple text query.

Search BIND data by building a field specific query.

Submit your interaction data to BIND.

Submit your molecular complex data to BIND.

BIND database general and detailed statistics.

Contact us with your questions and comments.

Interactions:	116931
Complexes:	2202
Pathways:	8

BIND Id

To submit your data to BIND, use the links below or the icon.

If you are submitting more than ten records at any one time you are invited to use our batch submission service by contacting htp@blueprint.org.

Figure 10. Snapshot of BIND database

B. THESIS STATEMENT

Most of the earlier work in information extraction from biological databases focused on database integration using wrapper technologies and provides only web links for navigational access along the network of biological relationships However; little work has been done to mine specific data leading to the identification of orthologs, paralogs and pathway information from heterogeneous biological databases. The need to develop an interactive information visualization tool leading to biological pathway detection by using controlled vocabulary and various other publicly available biological databases and providing a system to visualize the genes and gene products with respect to sub-cellular localization, has led to the concept and implementation of *GCell*. In this thesis we propose to develop a biological database-mining tool called Gcell that incorporates mining of various heterogeneous databases and provides researchers with a system to visualize and mine genomic data and offer the flexibility to view the results in multiple report formats. Considering the fact that all modern genomes have common ancestral genomes, and that, information gained from one can be applied to another has made studying of relationship(s) between different genomes inevitable. This becomes all the more relevant in finding the homologs, orthologs and paralogs of genes and proteins and biological pathways that are an important source of information, representing currently understood relationships (e.g., reactions and interactions) between biological factors.

Gcell system has been developed keeping in mind, the information need by the life science researchers. It is an integrated information and visualization system for genes and gene products. There are routes of providing input to the Gcell system: - 1. Gene list generated through microarray experiments. 2. Gene list generated from raw text (literature) by using biological entity recognition. 3. Gene list generated by wet lab experiments (Figure 11).

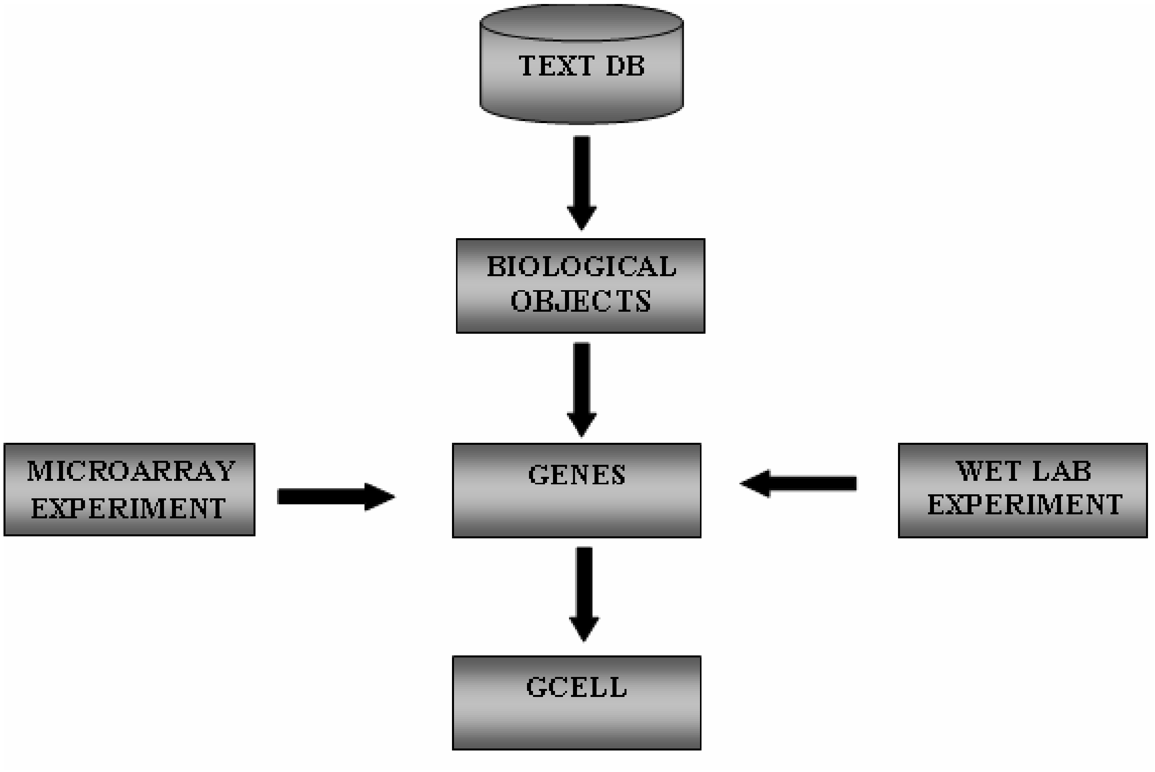


Figure11. Inputs to the Gcell System

The GCell system makes use of the raw text as the starting point from text databases. Scientific literature that is present in the form of abstracts appearing in different journals constitutes the text databases. From the text databases, biological entities are recognized and subsequently a gene list is generated. The gene list may be directly fed into the Gcell system for further information. However, deep mining on the genes may further generate a list of genes related to the initial set of genes. An iterative process of finding associations from scientific literature may achieve deep mining of genes. This newly generated list when used as an input to the Gcell system would incorporate all the information about the genes and proteins of interest to the user.

III. METHODS

A. DESIGN AND DEVELOPMENT OF GCELL

Text mining is about looking for patterns in natural language text. It is the process of analyzing text to extract information from it for particular purposes and focuses on extracting a small amount of information from text with high reliability (Witten et al 1999). Text Mining techniques were employed for cleaning and transforming the data obtained from heterogeneous databases. Thus, the data was exported from each of the sources into a data staging area. Here, all of the data is cleaned up, transformed as necessary, and linked with data from other sources. This staging process is a key to the success of this approach, and is fully automated. Then, when staging is complete, the data is placed in a unified central database, which is a composition of smaller databases. The one major drawback of this approach is the time it takes to extract, clean, transforms and loads the data into the warehouse. However, this problem can be addressed by scheduling smaller incremental updates. A broad schema for the database was developed in order to combat the heterogeneity of data.

Effective use of mined data requires a number of tools to be available to the users. GCell incorporates intuitive query interface, a variety of query tools to expand across multiple domains, and visualization tools to help users navigate through large volumes of mined data, finding patterns and trends that would otherwise go unnoticed. The system automatically and intelligently searches the GCell database to extract information leading to new insights.

The input to Gcell may come from various sources (Figure 12). Biosifter is an active personalized biological information system that presents the user with the documents of interest. The raw text documents come from the scientific literature database *Pubmed*. One of the key components of Biosifter system is the thesaurus creation module. This module creates a list of concepts that defines the domain of interest as specified by the user. These objects are then fed into the biological object identification module, which identifies the objects as genes, proteins, diseases, etc., and makes use of the BioMap knowledgebase. BioMap is a knowledge warehouse that stores the data for all biological objects extracted from the Pubmed database and the relationships between them. After the biological objects are identified using BioMap, genes and proteins can be fed into the TransMiner. TransMiner finds the transitive and direct associations between biological objects from the scientific literature. The associated objects may have similar functions or processes. These may lead to new hypothesis. In order to harness this knowledge and extend it to a much further level, Gcell system has been developed. Gcell also takes as input genes and proteins from microarray experiments and wet-lab experiments conducted by scientists. Any of the three methods may be used as input to the Gcell system.

B. BIOLOGICAL OBJECT IDENTIFICATION

Biological objects are of numerous types. One of the ways to identify biological objects is to use dictionaries. Dictionaries have been created for genes and proteins using publicly databases like Gene, LocusLink and Uniprot. A simple database string-matching search provides with the correct identification of the objects. There are 112874 unique proteins in the Protein dictionary. The number of unique genes extracted from Gene database by NCBI is 949588 unique gene symbols whereas 19365 unique gene names and 80024 unique gene symbols were extracted from Locuslink.

C. GENE SYNONYM RESOLUTION

Identifying all the genes and its symbols is an important aspect in the discovery of related information. Over a period of time, scientists have used different names and symbols for a gene. This may have happened because different communities of scientists working on different model organisms have not followed a global standard method of gene nomenclature. From the past few years, all the model organism researchers follow their own standard way of representing the gene by symbols. However, there is a need to further develop and extend a global gene nomenclature standard. This would allow a gene to have global reference. Hence, studies conducted for the action of genes in different organisms can provide deeper understanding of its functions.

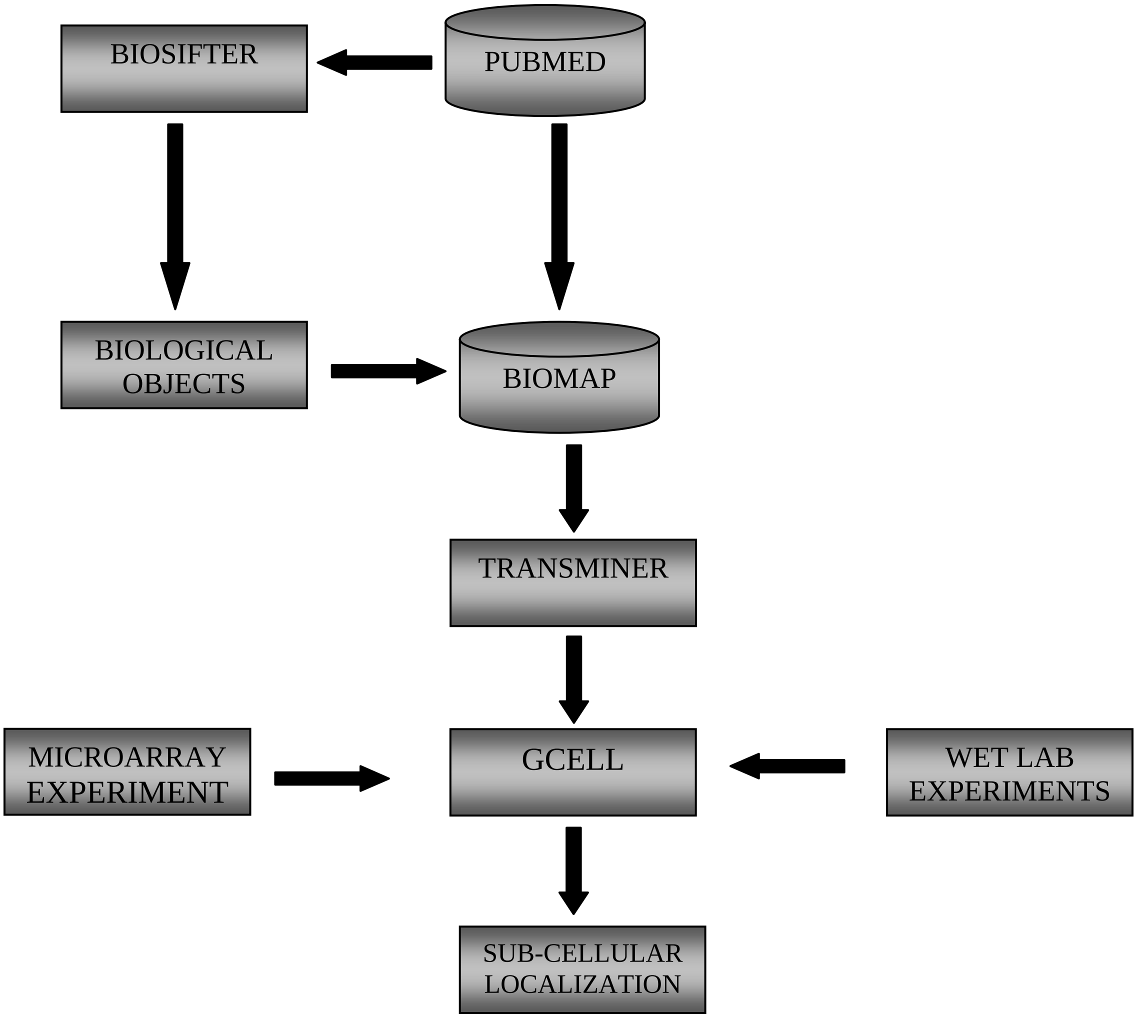


Figure12. Role of Gcell in gene and protein research

GCell takes as input a gene or a list of genes through the query interface and sends the request to the *GCell Query Engine.* The query engine then processes the request by further sending the requests to the GCell databases, fetches the results and integrates them to display to the user in a report form. The detailed report displays information regarding the gene and its synonyms with respect to the organism and tissue in which the gene products are formed or affected. It displays the ontology's for further references to biological processes, molecular function and cellular components. Some genes have the same name as its proteins. Hence a lot of information may be left out if a search among proteins was not made. Hence, GCell incorporates an explicit search on the Proteins database from Uniprot. A lot of information is also found by making use of the NCBI Gene database. GCell displays all the information present with the Entrez Gene database. Along with the display of different Geneid's, symbol, synonyms and chromosome map location, it also displays the external references to the other public databases. Transcripts and products associated to each gene product are displayed. Information from REFSEQ database and GENERIFS is also provided for further insight and deeper understanding of the genes and its products. A special feature of GCell is to provide the users with the references to the text/ journal paper in which that particular gene has been cited. Users can read the text online and may also download the abstracts of the text referencing them in text file format for further analysis. The ontology terms are hyper linked to the Gene Ontology browser for further analysis by the researchers. GCell provides an explicit interface to perform several queries not supported by other available online Gene Ontology browsers. Some of the queries supported include finding all the children, descendents, ancestors of a term, finding the shared parent of two nodes, sequences for a term, correlations between terms, transitive correlations of a term and finding External References. The genes in the query are integrated with the pathway database from KEGG to display the pathway in which the gene or its gene products belongs.

A unique feature of GCell is its visualization interface. The visualization interface is an interactive display of the cellular component where the gene products are active in a cell. This is displayed on a cell template using interactive Flash. It is in the form of a grid that spans across different species and tissues. This feature is intended for clear understanding about the orthologs and paralogs of the gene products. All the information available from GCell may be used for target identification for further research in the course of development of new drugs.

D. ARCHITECTURE OF GCELL

GCell is a classical 3-tier web application. We discuss the Three-Tier architecture of GCell in the following paragraphs. See Figure 13 below for an overview of the architecture.

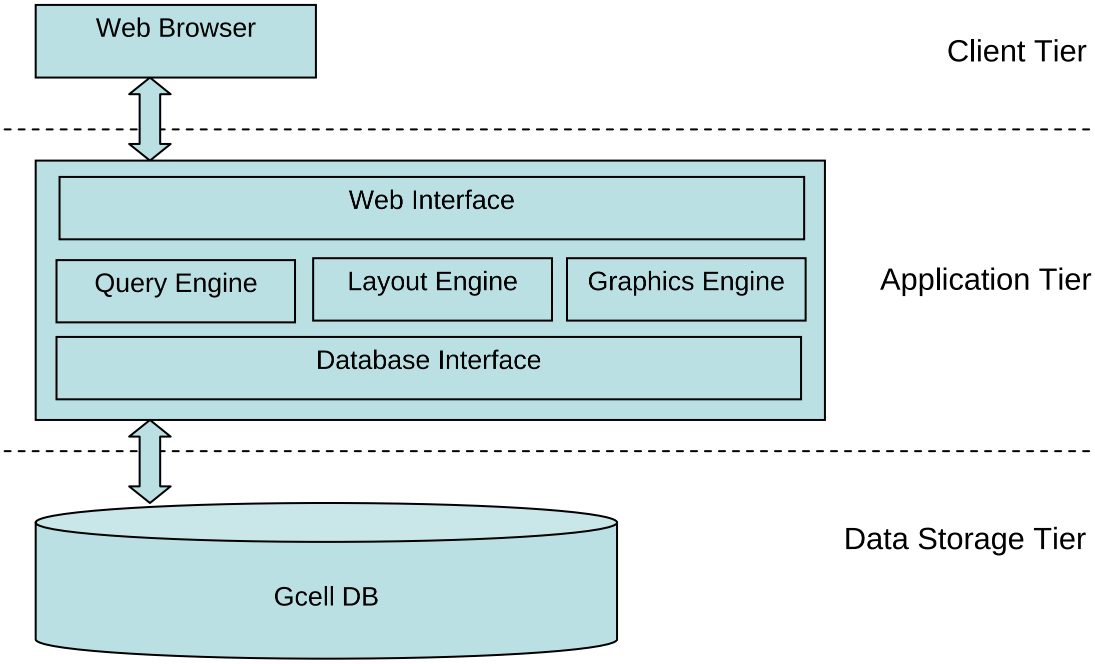


Figure13. Architecture of GCell

Client Tier

Users access the GCell service using a web browser like Netscape Navigator or Microsoft Internet Explorer. They enter their queries into HTML forms. The browser passes the query data to the GCell application server by sending a HTTP request. When the application server has finished query processing, the browser displays the returned query results. Clicking on internal links or image maps triggers another HTTP request.

Application Tier

The main part of GCell is the application server. It accepts queries, retrieves the corresponding data from the database, computes the result and delivers it to the client tier.

Data Storage Tier

The data of GCell is stored in a relational database management system (currently MySQL). The overall objective of the project is to enable web-based presentation of real time and archived data for the purposes of GCell system. Programs have been written to read download data from the ftp sites, parse the data, and store it in a database, all done in real time. Also, search and selection ability have been implemented, so that a user anywhere in the world can access and see information; search for particular data sets.

In pursuing this objective, several significant questions involving data storage, filtering, searching, optimization, real-time display, and system architecture were addressed initially. Questions like the ones specified below were taken into consideration.

(i) Should data be stored in a file system or a database?

(ii) What structure should the storage system have?

(iii) What data is relevant to store? How should data be filtered to eliminate noise?

(iv) How can the data be searched and analyzed efficiently in spite of the large quantity of data stored? Can other optimizations be performed?

(v) What system architecture, OS, web server, and database should be used?

(vi) How should data be displayed?

The design of the system incorporates the flow of data between different components of the Gcell system as described in the Architecture of GCell. A small example about the scripts written to create the database is shown below.

```
--
-- Table structure for table 'EX_COMPMAP'
--
CREATE TABLE EX_COMPMAP (
  id int(11) NOT NULL auto_increment,
  COMPONENT varchar(200) default NULL,
  SYMBOL varchar(20) default NULL
) TYPE=MyISAM;

--
-- Table structure for table 'EX_GOCOMPS'
--
CREATE TABLE EX_GOCOMPS (
  id int(11) NOT NULL auto_increment,
  GOID varchar(20) default NULL,
  CPF varchar(200) default NULL,
  CPFDETAIL varchar(200) default NULL
) TYPE=MyISAM;

--
-- Table structure for table 'GE_GENE2REFSEQ'
--
CREATE TABLE GE_GENE2REFSEQ (
  id int(11) NOT NULL auto_increment,
  TAX_ID varchar(20) default NULL,
  GENEID varchar(20) default NULL,
  STATUS varchar(20) default NULL,
  RNAACC varchar(20) default NULL,
  RNAGI varchar(20) default NULL,
  PROTACC varchar(20) default NULL,
  PROTGI varchar(20) default NULL,
  GENNUCACC varchar(20) default NULL,
  GENNUCGI varchar(20) default NULL,
  START varchar(20) default NULL,
  END varchar(20) default NULL,
  ORIENTATION varchar(20) default NULL,
  KEY GENEID_GE_GENE2REFSEQ (GENEID)
) TYPE=MyISAM;
```

```
--
-- Table structure for table 'GE_GENERIFS'
--

CREATE TABLE GE_GENERIFS (
  id int(11) NOT NULL auto_increment,
  TAXID varchar(20) default NULL,
  GENEID varchar(20) default NULL,
  PUBMEDID varchar(20) default NULL,
  LASTUPDATE varchar(20) default NULL,
  GENERIFTEXT varchar(255) default NULL,
  KEY GENEID_GE_GENERIFS (GENEID)
) TYPE=MyISAM;
```

E. COMPONENTS OF GCELL

The various components of GCell systems comprises of the following:

Application server

The heart of the GCell system is the application server. It consists of several components, some of which are implemented in Java. The application server has been setup on Linux operating system and is easily portable to any platform for which Java is available.

Web Interface

The web interface is responsible for the communication with the client tier. It receives the query in terms of a HTTP request with associated parameters. It parses the request and triggers the corresponding functionality of the query engine which processes the query and returns the result as a graph. The web interface uses the layout and graphics engine to transform the graph into a picture and delivers it to the client as a GIF or PNG image with a corresponding image map and a HTML page. The web interface is implemented in Java based on the Java Servlet Technology.

Query Engine

The query engine executes the user queries (Figure 15). Upon a request from the user, the query engine further breaks down the query for fetching relevant information from the Gcell database. Initially the synonyms are resolved which is done by resolving a given gene name against mined data from NCBI’s Gene and Locuslink databases and the Uniprot database. The original flat files of these databases comprises of synonyms for known genes. Relevant data from all three databases are correlated by common gene names or/ and symbols. Mining of this data yields a comprehensive list of gene names and symbols. The user query is matched against Gcell database for name and symbols. All the unique geneid’s are retrieved and a backward search for all the names and synonyms are retrieved for the list of geneid’s (Figure 14). The same procedure is followed for NCBI’s LocusLink and Uniprot database.

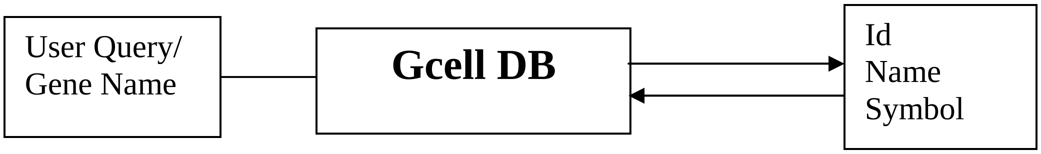


Figure 14. Gene synonym resolution

Along with synonyms resolution, query is further sub-divided into looking for detailed information from Gcell database comprising of NCBI’s Gene, LocusLink, PubMed, Refseq, GeneRif’s, Uniprot, Kegg, Enzyme, BIND and Gene Ontology databases. Eventually, it builds a visualization grid, represented as a grid of animal cell template, which spans across different organisms and tissues. The query engine is implemented in Java.

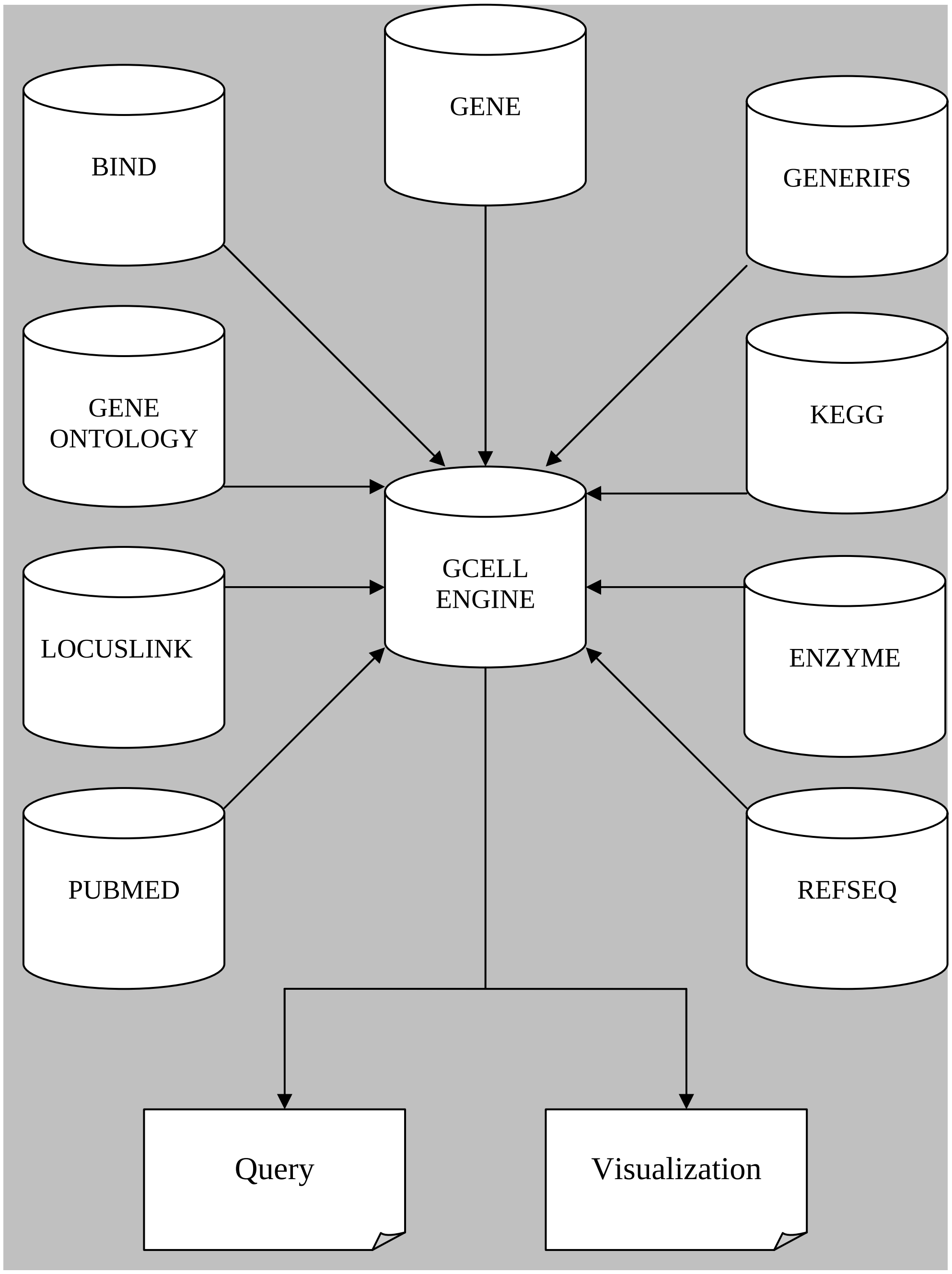
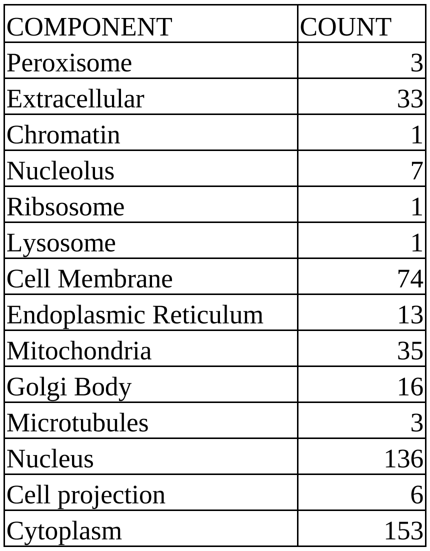


Figure 15 Database interface with Gcell query engine

Layout Engine

The data generated by the query engine is represented in an XML format and written to an XML file. The different components of the cell are mapped on the animal cell template and the layout engine decides on the components to be animated for a particular gene and tissue. The layout engine comprises of a cell component mapping that has been developed using 630 different types of cell components as found in Gene Ontology database. All the 630 different components were manually check and mapped to 14 different basic components of an animal cell (Figure 16).



COMPONENT	COUNT
Peroxisome	3
Extracellular	33
Chromatin	1
Nucleolus	7
Ribsosome	1
Lysosome	1
Cell Membrane	74
Endoplasmic Reticulum	13
Mitochondria	35
Golgi Body	16
Microtubules	3
Nucleus	136
Cell projection	6
Cytoplasm	153

Figure 16. Number of cell components in each category

However, 139 components were not clear in the ontology and scientists are not sure about it. With reference to plant cell, 2 components for chloroplasts and 7 for vacuoles have been mentioned. Hence in Gcell, chloroplasts, vacuoles and unknown components have not been taken into consideration.

Graphics Engine

The graphics engine generates images and image maps from the data computed by the layout engine. It is implemented in Java using Flash.

Database Interface

The communication between the query engine and the database is done via the database interface. This encapsulation simplifies adapting GCell to other data sources. The database interface utilized is JDBC.

F. IMPLEMENTATION

1. System Architecture

System architecture is the hardware and software platform that is used. The hardware platform is defined by specifying the processor, ram, hard drive, etc. The software platform is defined by specifying the operating system, type of database, web server, etc. For this project, we decided to use Fedora Linux. The backend database and web server is a MySQL Server and Tomcat 4.2. The benefits of this system architecture include inter-compatibility, compatibility with existing systems, cost, and performance.

Hardware Architecture

CORSAIR.CS.IUPUI.EDU (134.68.140.208) is a Dell Server with a P4-1.7 GHz processor, 3GB RAM, and a 36 GB hard drive on Fedora Server. Installed on Corsair.cs.iupui.edu are Fedora Server, MySQL, Apache and Tomcat 4.2, which act as the application server.

Software Architecture

Fedora is used as the operating system. Since Fedora is the operating system of choice in the Department of Computer Science, most of the development work is usually done on Red Hat workstations. It made the most sense to keep Fedora as the server operating system. Hence I had the opportunity to explore the open source software and check the compatibility with Fedora.

Application Server

For an application server, Jakarta-tomcat is used. It handles Hyper-Text Transfer Protocol (HTTP) requests in the form of JSP/Servlet. When one enters a URL into a web browser (e.g. Netscape or Internet Explorer), one performs an HTTP request for the web/ application server to retrieve and serve the desired web page. Tomcat allows interactions with a relational database and can generate web pages.

Relational Database

MySQL is used as a relational database, as it is designed for and runs on Linux Servers. To perform many of the desired searches requires a relational database, a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. The incoming information from various heterogeneous data sources is automatically parsed and inserted into a relational database with the help of scripts. The information is stored on a file system as well for backup and archival purposes. Once the information for each record in a database is inserted, a standard Structured Query Language (SQL) statement is used to select the appropriate information from the relational database.

2. Database / Data Model

One benefit of using a relational database (MySQL) is that the information is stored in an organized fashion. To organize this data efficiently, decisions regarding the structure of the database, what data should be stored, and how records should be indexed are crucial.

The data model considered comprises of several tables holding data from various heterogeneous sources, having primary and foreign keys to relate the unique identifiers throughout. Each table is indexed according to the query attribute. The definition of the tables used to store data is called a data model. A data model is a plan or map that defines the units of data and specifies how each unit is related to the others. It is contained in the database it describes and is available to any program that uses the database. The data model defines both the names of the data items and their respective data types. The data model created a database structure that allows for searches to be performed quickly and efficiently. Furthermore, the data model is easily modifiable and maintainable.

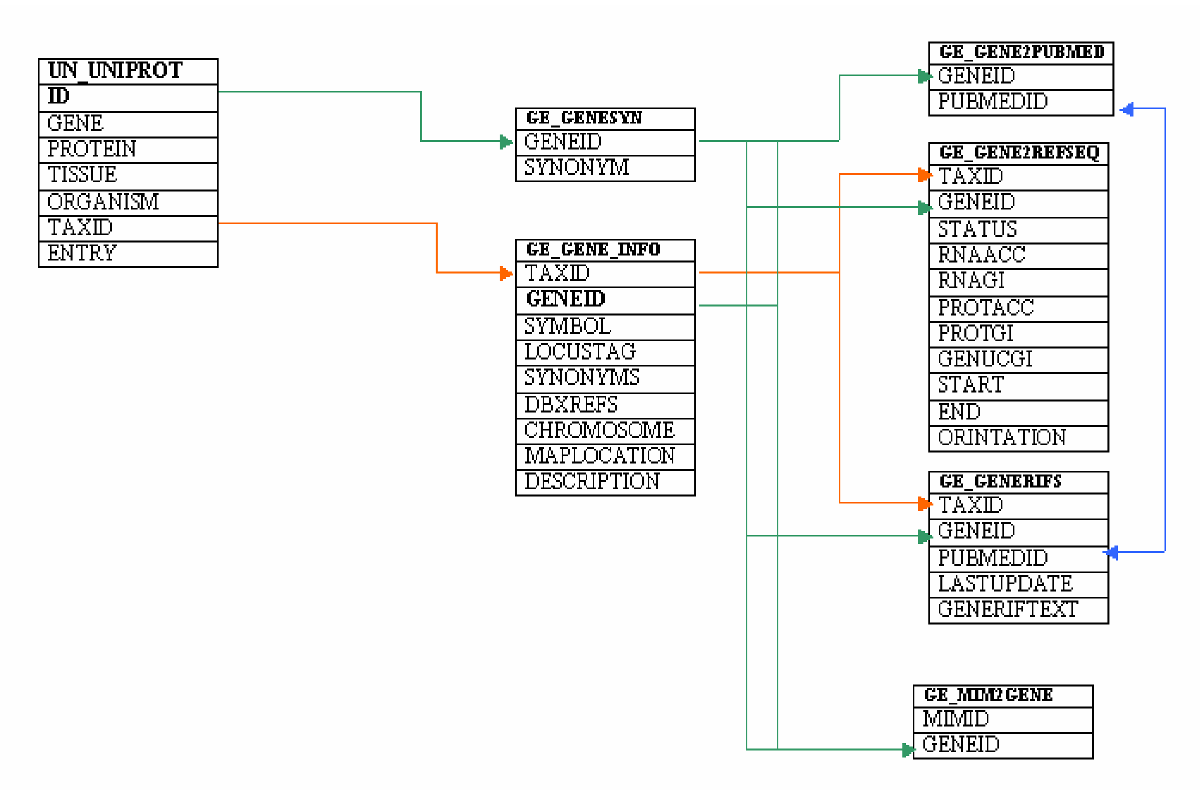


Figure17 Partial Schema of the Gcell database

The above figure (Figure17) displays a partial schema designed for Gcell database. In this schema, databases like Uniprot, Gene, RefSeq, generifs, OMIM and PubMed have been brought together. Selected fields from each of these databases contribute to the structure of the Gcell database. This is designed to return data quickly for searches that are likely to be performed. In this case, likely searches span across different databases. So, the data model is designed with that in mind, as genes are indexed by the unique "id". Indexing by *id* allows for those *id* searches to run faster because not all the rows in the table need to be examined. The search simply needs to go to the index and where values in a certain interval are located.

This data model is easy to maintain and modify. Since it is likely that other students will continue to work on this project and similar projects, easy maintenance and modifiability of the data model is important. Extra care has been taken to name the tables and columns in an intelligent, consistent, and descriptive fashion. For example, the column names are descriptive and are uniform in the database and software programs. This may seem obvious, but it is important since it simplifies maintenance. Moreover, if the initial data model accurately represents the data that will be kept in the database, modification of the data model adding new columns, deletion of columns, etc is minimized. This is an important point, and minimizing modifications to the data model reduces later complexity.

3. Data Processing

Data processing constitutes a major part of this project. In this system, a data collection solution is developed in the form of scripts, which run as *cron* jobs on the server. Filtering is done prior to insertion of data into the database. This reduces insignificant entries in the table, which subsequently reduces the size of the database. The various databases incorporated in the GCell database are updated at different intervals. Hence, the script keep track of the updates, and subsequently downloads the databases, parses them using the parser scripts and finally uploads the data in the database for filtering, analysis, and display with software.

Data Archival

Shell scripts are made use of to download the individual database updates via ftp as *cron* jobs. The downloaded data are then parsed in the required format to satisfy the database schema.

Data Upload

At the server where the Gcell system is installed, CORSAIR.CS.IUPUI.EDU, data is processed and simultaneously files are uploaded to the database by 'mysqlimport' statements. When the data update system is online and synchronized, from the ftp server is put into the server and subsequently after parsing, into database continuously. This is the phase when the system might be down for a short period of time for database updates. Following commands were used in order to upload the data files for the GCell system.

```
>>cat *.sql | mysql -h localhost.localdomain -u gcell --password=***** gcell_db
>>mysqlimport -h localhost.localdomain -u gcell --password=***** gcell_db *.tab
```

Batch insertion is a convenient way to insert multiple rows simultaneously into the database. Since there are approximately millions of rows inserted into the database periodically, performing an insert would cause unnecessary load on the database and it would be very time consuming and an inefficient way. Performing updates in batch significantly decreases load on the database by both decreasing the number of interactions with the database and increasing the amount of data inserted per interaction.

4. Database Interactions

Database Connectivity

We need to establish connectivity to the relational database in order to upload data and select data. Furthermore, the database connectivity needs to work with Java programming language of choice. We use both the Open Database Connectivity (ODBC) application programming interface (API) along with the Java Database Connectivity (JDBC) application programming interface specifications to connect our Java programs to our database. For our purposes the JDBC-ODBC bridge that comes with the Java SDK works well. The API allows database access statements written in SQL to be encoded and then passed to a program managing the database.

The client accesses the database through the middleware Servlet/Beans. They perform database queries and simply transmit results to the client. The standard JDBC (Java Database Connectivity) API is used allow our Java programs to connect to the database. A class for database access was created, called Database.java. In the class, the driver is registered with the driver manager one time, and then Servlets and other programs that need a database connection call a static method to return a connection to the database. This makes changing database drivers easier. Also, connection pooling can be easily implemented. We can use a simple line in Java in Database.java to create a bridge between JDBC and ODBC.

```
DriverManager.getConnection("jdbc:odbc:GCell_db", "GCell", "*****");
```

This line attempts to establish a connection to the database specified in the first argument. In the code sample, we are using JDBC to connect to an ODBC database GCell, which resides on the localhost CORSAIR.CS.IUPUI.EDU. The second and third arguments are the username and password required to access the database. Essentially, our Java programs use a JDBC-ODBC bridge to connect to a SQL Server 2000 database.

Searching

After data from the parsed files have been inserted into the database properly, SQL statements are used to search and retrieve desired data. Using SQL, we perform a variety of desired functions:

(i) Select gene related information from a variety of sources such as Uniprot, Gene and Locuslink databases.

(ii) Select pathway information from KEGG, BIND and ENZYME database

(iii) Select ontology information from Gene Ontology

(iv) Select bibliographic references from Pubmed database and other databases

Various SQL statements are used for insertion and selection of data from the database. To select data from the database, a number of different SQL statements are used depending on what the end user has requested. For example, if the end user wants to view transitive associations between ontology term "hexokinase activity" and all other terms, the following SQL statement is used:

```
SELECT term2.name, term2.acc, count(distinct asso2.gene_product_id) AS gpc
FROM
term term1, term term2, graph_path path1, graph_path path2, association asso1,
association asso2, evidence evid1, evidence evid2
WHERE
term1.id = path1.term1_id AND
term2.id = path2.term1_id AND
asso1.term_id = path1.term2_id AND
asso2.term_id = path2.term2_id AND
evid1.association_id = asso1.id AND
evid2.association_id = asso2.id AND
asso1.is_not = 0 AND
asso2.is_not = 0 AND
evid1.code != 'IEA' AND
evid2.code != 'IEA' AND
asso1.gene_product_id = asso2.gene_product_id AND
term1.acc = 'GO:0004396'  AND
asso1.term_id != asso2.term_id
GROUP BY term2.name, term2.acc
ORDER BY gpc DESC;
```

The result obtained comprises of 57 records.

SL.NO.	TERM	ACCESSION	COUNT
1	Gene_Ontology	GO:0003673	24
2	biological_process	GO:0008150	18
3	physiological process	GO:0007582	18
4	hexose metabolism	GO:0019318	17
5	alcohol metabolism	GO:0006066	17
6	metabolism	GO:0008152	17
7	carbohydrate metabolism	GO:0005975	17
8	monosaccharide metabolism	GO:0005996	17
9	alcohol catabolism	GO:0046164	13
10	carbohydrate catabolism	GO:0016052	13
11	catabolism	GO:0009056	13
12	monosaccharide catabolism	GO:0046365	13
13	cellular_component	GO:0005575	13
14	glucose catabolism	GO:0006007	13
15	glucose metabolism	GO:0006006	13
16	hexose catabolism	GO:0019320	13
17	adenyl nucleotide binding	GO:0030554	12
18	ATP binding	GO:0005524	12
19	Main pathways of carbohydrate metabolism	GO:0006092	12
20	binding	GO:0005488	12
21	molecular_function	GO:0003674	12
22	nucleotide binding	GO:0000166	12
23	purine nucleotide binding	GO:0017076	12
24	energy derivation by oxidation of organic compounds	GO:0015980	12
25	energy pathways	GO:0006091	12
26	glycolysis	GO:0006096	12

27	cell	GO:0005623	10
28	intracellular	GO:0005622	9
29	cytoplasm	GO:0005737	9
30	mitochondrion	GO:0005739	4
31	fructose metabolism	GO:0006000	4
32	plastid	GO:0009536	3
33	cellular_component unknown	GO:0008372	3
34	chloroplast	GO:0009507	3
35	cytosol	GO:0005829	3
36	membrane	GO:0016020	2
37	cellular process	GO:0009987	2
38	hexose mediated signaling	GO:0009757	1
39	intracellular signaling cascade	GO:0007242	1
40	mitochondrial membrane	GO:0005740	1
41	carbohydrate mediated signaling	GO:0009756	1
42	mitochondrial outer membrane	GO:0005741	1
43	cell communication	GO:0007154	1
44	cell cycle	GO:0007049	1
45	cell growth and/or maintenance	GO:0008151	1
46	nucleus	GO:0005634	1
47	cell proliferation	GO:0008283	1
48	outer membrane	GO:0019867	1
49	cellular physiological process	GO:0050875	1
50	regulation of cell cycle	GO:0000074	1
51	response to carbohydrate stimulus	GO:0009743	1
52	response to endogenous stimulus	GO:0009719	1
53	response to hexose stimulus	GO:0009746	1
54	response to stimulus	GO:0050896	1
55	signal transduction	GO:0007165	1
56	sugar mediated signaling	GO:0010182	1
57	hexokinase-dependent signaling	GO:0009747	1

Optimization

There is a large quantity of data to be stored into the database. For this reason, the structure of the database and search optimization is of utmost importance. The most direct way to speed up selection of data is to use an index. An index is essentially a structure of pointers that point to rows of data in a table. Indexes were created for all the required attributes over different tables in the database. Example: -

```
CREATE INDEX SYN_GE_GENESYN ON GE_GENESYN(GENEID,SYNONYM);
CREATE INDEX GENE_UN_GENE ON UN_GENE(GENE);
CREATE INDEX ID_UN_ENTRY ON UN_ENTRY(ID);
CREATE INDEX ID_UN_TAXID ON UN_TAXID(ID);
CREATE INDEX ID_UN_ORGANISM ON UN_ORGANISM(ID);
CREATE INDEX ID_UN_TISSUE ON UN_TISSUE(ID);
CREATE INDEX ID_UN_PROTEIN ON UN_PROTEIN(ID);
CREATE INDEX PROTEIN_UN_PROTEIN ON UN_PROTEIN(PROTEIN);
CREATE INDEX GENEID_GE_GENE_INFO ON GE_GENE_INFO(GENEID);
CREATE INDEX GENEID_GE_GENE2REFSEQ ON GE_GENE2REFSEQ(GENEID);
CREATE INDEX GENEID_GE_GENE2PUBMED ON GE_GENE2PUBMED(GENEID);
CREATE INDEX GENEID_GE_GENERIFS ON GE_GENERIFS(GENEID);
```

An index optimizes the performance of database queries by ordering rows to speed access. To understand what a database index does, a simple analogy is the index of a book. To find something in the book, you simply flip to the index of the book and look up the page number the desired subject is located on. For a database, the idea of an index is the same examining the index tells you where in the table the desired information is located.

5. User Interface

Finally, the system allows display of user-selected data through HTML/Flash and a website. These all have been designed with targeted end user researchers and students in mind. Factors such as usability, clarity, simplicity, speed, etc have been considered in the design of the website. The Appendix portion of this thesis concentrates on the user interface developed for the system.

IV. RESULTS

GCell system incorporates all possible information for any gene and gene products belonging to various species. The input to the system may be any gene or a list of genes that may be of interest to the user. However, looking at the way research progresses, in the beginning, the researcher may be interested in a specific disease or condition, but eventually after literature survey and the subsequent research, the user gets more and more involved in the function of genes and gene products at the molecular level. Hence, making use of the BioSifter, TransMiner (Figure A1) and GCell, a researcher would be able to pursue further research and generate hypothesis based on the knowledge from these systems. In order to demonstrate this, a “brca1” problem domain was created (Figure A2, A3) in BioSifter. BRCA1 gene participates in transcriptional regulation and causes breast cancer in cows, mouse and humans. The BioSifter system created a thesaurus of terms from an initial document set for the ‘brca1’ problem domain. The terms among others ((Figure A4)), included ovarian, prostate, p53, brct, genetic, tumours, mastectomy, ashkenazi, atm, brca, mammary, 185delag, mutations, cancers, loh, apos, women, methylation, tumour, breast, carcinoma, oophorectomy, cancer, prophylactic, sporadic, carcinomas, brca2 and bard1. Later on the system presented the user with the Pubmed articles for ‘brca1’. In total, there were 3788 (as of December 2nd, 2004) articles in PubMed, which referred to this gene. The system after processing (vector space document representation and classification) presented the user with the result. Each article was reviewed online and a relevance feedback was provided for the same (Figure A5). This was carried out for around 100 iterations and each iteration displayed 25 articles at a time. Over a period of time, the system learned (Figure A7) and displayed only the articles of interest. This was confirmed by the learning curve that the system generated. The system also has the capability to provide for automatic feedback to the remaining documents (Figure A6), which the user has not yet seen. Practically, it is almost impossible for any user to go through each of the articles for a particular query. Hence, the need for an automated process.

TransMiner (Figure A8) was used to find the direct relations and transitive associations between objects. A set of 56 gene symbols related to breast cancer was used to find the transitive associations. Some of them include *APC, APS, ATM, BCL1, BCL2, BRCA1, BRCA2, CCND1, CDKN2A, COL18A1, DCC, EGF, EGFR, EMS1, ERBB2, ERBB3, MSH2, MLH1, FGF3, FGF4, FGFR1, FGFR2, FGFR4, GH1, GRB7, HRAS, IGF1R, KIT, KRAS2, MYCL1, IGF2R, MCC, MDM2, MET, MYC, NF2, NRAS, PGR, PHB, PLAT, PLG, PRL, PTH, PTPN1, RB1, SSTR1, SSTR2, SSTR3, SSTR4, SSTR5, SRC, TGFA, TP53, TSG101, VIM, WNT10B.* A total of 4392 documents were retrieved from Pubmed database. 132 direct and 619 potential transitive associations were found. After the iterative process, using each pair with 'AND' operator, 304 transitive associations and 255 direct associations were found. The direct associations and transitive associations are visualized in the TransMiner applet [Figure A9, A10]. The transitive associations are depicted by the use of 'dashed' lines whereas straight lines depict direct associations with the thickness indicating the strength of the associations. The interface aids in the display and eventually understanding the different types of relationships between the biological objects. Strength of the association may be adjusted according to the user in order to view associations satisfying certain criteria. Different levels of neighbor may be selected in order to view the neighbors of each of these objects.

The same set of 56 gene symbols was input to the GCell system. The GCell query engine [Figure A11] upon receiving the request sends the query to different databases such as Gene, Uniprot, LocusLink, Refseq, Generifs, etc to generate a report [Figure A12,] and Visualization of the sub-cellular localization for the given gene or protein. The visualization shows a grid of the cell template across different organisms and different tissues. Biological researchers can easily see the cell components where a particular gene or protein is active. An overall view would provide an insight into finding the homologs, orthologs and paralogs of genes and proteins. From here, scientists may move on to view the pathway(s) affected by the genes and proteins from KEGG database. The information displayed on the report moves across the species and tissues in which the gene is found, the proteins it affects and the cellular component, molecular function and biological processes it affects [Figure A13]. It also displays hyperlinks to Uniprot and Gene Ontology database [Figure A14]. The same entry is used for a search among in the Protein databases too [Figure A15]. A lot of interesting results are found here. The various links also display information from the Generifs and Refseq databases [Figure A16, Figure A17]. Titles of documents from Pubmed database are displayed which are also linked to the online Pubmed database for immediate retrieval of abstracts [Figure A18]. On clicking the Kegg database link, all the pathways in which the particular gene plays a role by any of its actions, are displayed [Figure A19]. It gives a clear picture to the researcher about the places and pathways, which may be altered in the event of any gene mutations. As additional feature, The GCell system has the capability to find the ancestors [Figure A20], children and sequences of the Gene Ontology terms. It is also capable of finding transitive associations between the GO terms [Figure A21].

V. CONCLUSION

A system has been designed with an approach to the problem of sub-cellular localization of gene products leading to biological pathways. This involves moving from the raw text data at a broader level to a much more detailed view of pathways representing complex biological interactions. The study of pathways would help to understand the mechanism of organism and provide guidance for biologists to design experiments. Drug designing would become much easier and would save time and money for the biologists to do experiments.

A better understanding of the interaction networks in complex biological systems will enable numerous advances in biotechnology. The field of interaction bioinformatics includes many challenges and holds much promise. It would provide with an enhanced ability to target therapeutics appropriately in diseased cells. The Kegg database has up to 18735 pathways available for 231 species. However, Kegg does not provide pathway metadata or compartment information. Hence, Gcell can be used in combination with Kegg database for annotating metadata for pathways. The Gcell system is specifically designed and developed keeping the needs of biological researchers in mind. The fact that a gene may be present in different organism and different tissues and may share function(s) across species and across tissues has been taken into account.

The system provides a visualization interface for the cell components where a particular gene or protein may be active. Use of Gcell can be made in the field of evolutionary biology for identifying a possible relation between the genes and proteins, which may have common ancestors. Evolutionary functional counterparts in different species and lineage specific adaptations may be identified. This is made much easier by the report generated in Gcell. All the relevant information is displayed in an easy to understand manner and major data points are hyper-linked to the parent databases from where the data was extracted. To an evolutionary biologist and researchers, information regarding whether a particular gene or protein is conserved in different organism may be of interest in formulating new hypothesis.

An example of "GCK" gene was fed into the Gcell system, which generated a visualization grid for sub-cellular localization results in Figure A23 and the report as shown in Figure A22. The grid clearly shows the cellular components where the gene GCK plays a role (see Figure 18). The GCK gene matched four entries

(I) HXK4_RAT

(II) HXK4_MOUSE

(III) M4K2_HUMAN

(IV) HXK4_HUMAN

Out of these three, no pathway data was available for HXK4_RAT entry in Kegg database. The various pathways (Figure19) in which the other three entries played a role are listed in the Table1,2,3 below. Although the pathways involved are same for HXK4_MOUSE and HXK4_HUMAN, there is a distinct difference in the pathway affected by M4K2_HUMAN. Identification of such scientific knowledge may help the biologists to perform further research leading new hypothesis.

Table 1. HXK4_MOUSE

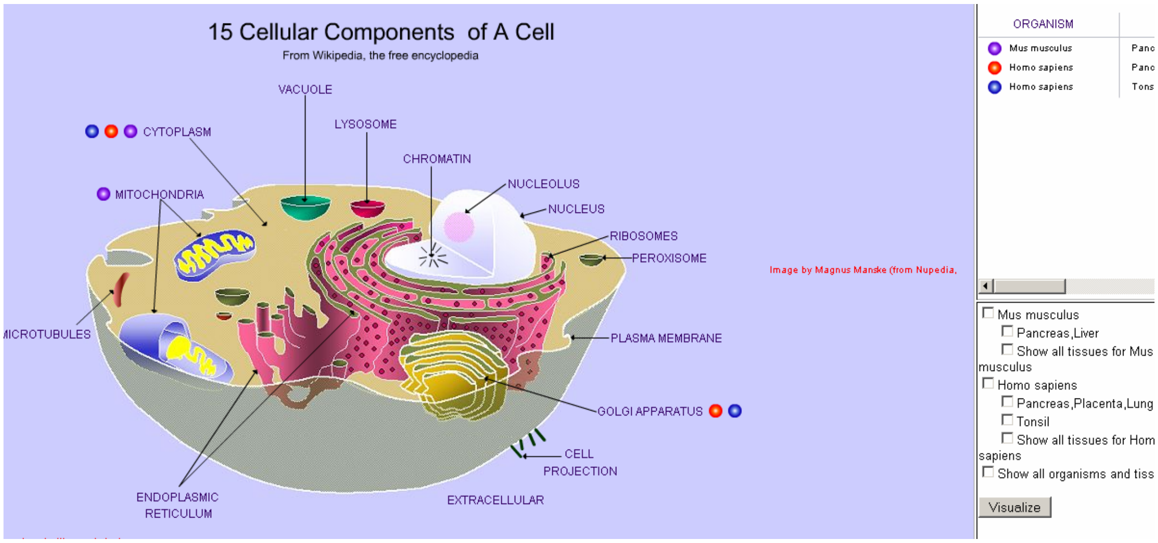
Type	Category	Pathway	Kegg ID
Metabolism	Carbohydrate Metabolism	Glycolysis / Gluconeogenesis	PATH:mmu00010
Metabolism	Carbohydrate Metabolism	Fructose and mannose metabolism	PATH:mmu00051
Metabolism	Carbohydrate Metabolism	Galactose metabolism	PATH:mmu00052
Metabolism	Carbohydrate Metabolism	Starch and sucrose metabolism	PATH:mmu00500
Metabolism	Biosynthesis of Secondary Metabolites	Streptomycin biosynthesis	PATH:mmu00521
Metabolism	Carbohydrate Metabolism	Aminosugars metabolism	PATH:mmu00530

Table 2. HXK4_HUMAN

Type	Category	Pathway	Kegg ID
Metabolism	Carbohydrate Metabolism	Glycolysis / Gluconeogenesis	PATH:mmu00010
Metabolism	Carbohydrate Metabolism	Fructose and mannose metabolism	PATH:mmu00051
Metabolism	Carbohydrate Metabolism	Galactose metabolism	PATH:mmu00052
Metabolism	Carbohydrate Metabolism	Starch and sucrose metabolism	PATH:mmu00500
Metabolism	Biosynthesis of Secondary Metabolites	Streptomycin biosynthesis	PATH:mmu00521
Metabolism	Carbohydrate Metabolism	Aminosugars metabolism	PATH:mmu00530

Table 3. M4K2_HUMAN

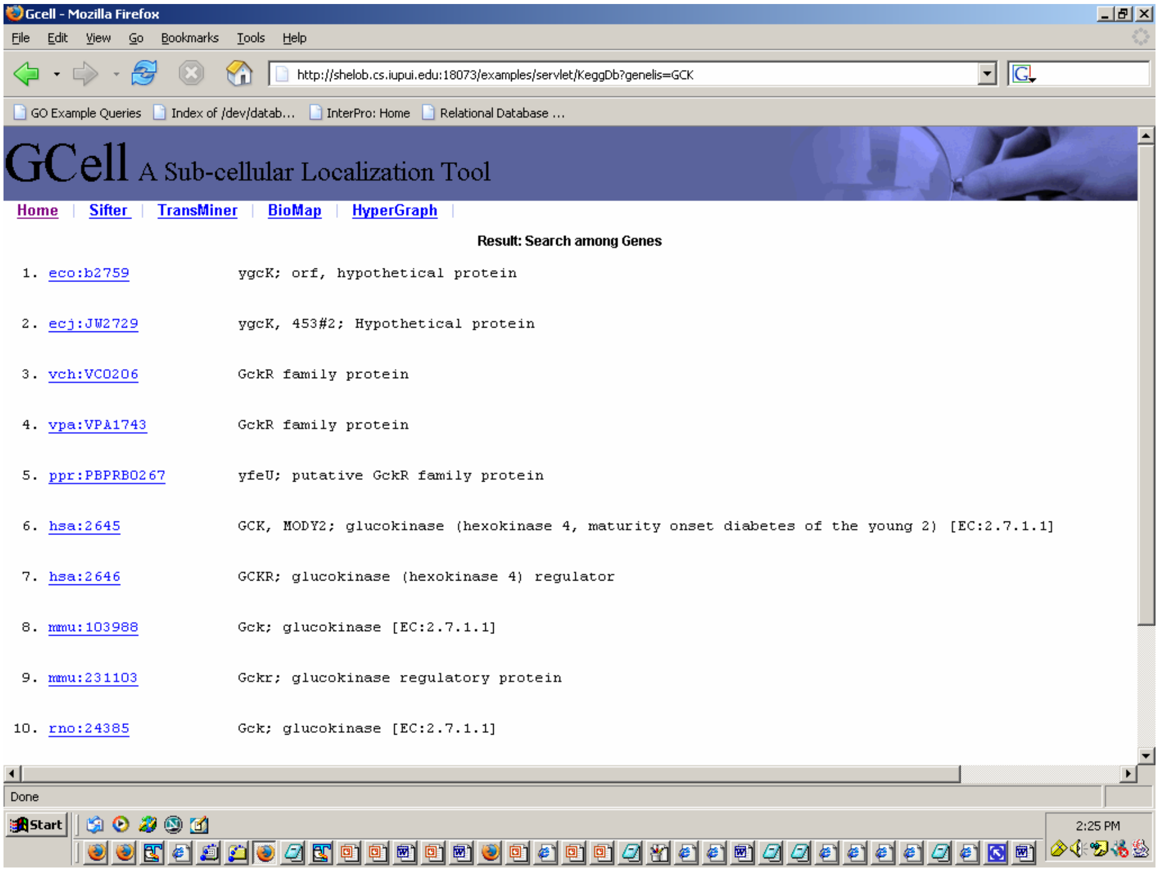
Type	Category	Pathway	Kegg ID
Environmental Information Processing	Signal Transduction	MAPK signaling pathway	PATH:hsa04010



ORGANISM	
● Mus musculus	Panc
● Homo sapiens	Panc
● Homo sapiens	Tons

□ Mus musculus
□ Pancreas,Liver
□ Show all tissues for Mus musculus
□ Homo sapiens
□ Pancreas,Placenta,Lung
□ Tonsil
□ Show all tissues for Hom sapiens
□ Show all organisms and tiss

Figure 18 Sub-cellular localization of GCK gene


Gcell - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:18073/examples/servlet/KeggDb?geneIs=GCK
GO Example Queries
Index of /dev/datab...
InterPro: Home
Relational Database ...
GCell A Sub-cellular Localization Tool
Home | Sifter | TransMiner | BioMap | HyperGraph |
Done
Start
2:25 PM

Result: Search among Genes

```
 1. eco:b2759          ygcK; orf, hypothetical protein

 2. ecj:JW2729         ygcK, 453#2; Hypothetical protein

 3. vch:VC0206         GckR family protein

 4. vpa:VPA1743        GckR family protein

 5. ppr:PBPRB0267      yfeU; putative GckR family protein

 6. hsa:2645           GCK, MODY2; glucokinase (hexokinase 4, maturity onset diabetes of the young 2) [EC:2.7.1.1]

 7. hsa:2646           GCKR; glucokinase (hexokinase 4) regulator

 8. mmu:103988         Gck; glucokinase [EC:2.7.1.1]

 9. mmu:231103         Gckr; glucokinase regulatory protein

10. rno:24385          Gck; glucokinase [EC:2.7.1.1]
```

Figure 19. Pathways for GCK gene available from Kegg Database

One thing worth mentioning here is that the biologists have a fair idea about what pathways goes on at certain locations. Hence, if they are presented with localization information, they would have a pretty good idea about the pathways in which a particular gene would play a role. From this display, one can clearly see that HXK4_HUMAN and M4K2_HUMAN entries are paralogs. They are genes that derive from a single gene that was duplicated within the genome. On the same report, one can also view the orthologs - HXK4_HUMAN and HXK4_MOUSE. These may be considered to be genes from two different species that derive from a single gene in the last common ancestor of the species. Looking at the molecular function one can figure out that other than "protein binding" activity, all the other functions remain the same. It is important to know here, that the molecular functions of orthologs may or may not be the same.

Gene expression microarray data shows expression level of thousands of genes at the same time. Hence, through the use of Gcell, we can find sequence of pathways based on microarray database in terms of time. In the area of drug design, this application can provide vital information regarding the tissues in different organisms. Drugs may be developed to target a tissue in one organism if a similar drug or chemical alters a pathway in another organism. Tissue similarity may be used to identify target drugs.

VI. DISCUSSION

Biological data in public repositories focus on deriving and providing one particular type of data, be it sequences (e.g., GenBank), protein information (e.g., UniProt), molecular interactions (e.g., BIND), literature database(e.g. Pubmed), diseases (e.g., OMIM).

Several attempts have been made to integrate data that would enable the researchers in discovering new associations between the data, or validate existing hypotheses (Shah Et. al.). Several recent studies have demonstrated the power of integrative bioinformatics. Illustrating the potential of querying integrated public data to reveal novel relationships, Mootha et al were able to identify one of the disease genes responsible for Leigh syndrome (Mootha et. al.). In another example, Stuart et al generated hypotheses about the functional roles of gene sets using publicly available data (Stuart et. al.).

Entrez System (Wheeler et. al.) from NCBI is a web-based system which is extremely extensive in the scope of data it provides, however, large queries may take a long time to return or may not be returned at all due to server resource restrictions. Also, the level of data integration is only at the services level, rather than at a field-based level which can provide much better resolution for queries. These are the disadvantages of the currently available integrated bioinformatics solutions.

The Gcell system may be considered to be a biological data warehouse that locally stores and integrates functional annotations of genes and proteins, biological ontologies, molecular interactions and biological sequences. This system is based on relational data model and the major advantage of GCell over Entrez is that it is installed as a local instance providing access to the data. It gives flexible access to the data by means of SQL queries. The data from the system can be easily queried on local instances of Entrez Gene, LocusLink, GenBank, RefSeq, UniProt, KEGG, BIND and Gene Ontology. The users get access to the system on a high-bandwidth internal network with lower latency. However, one of the limitations of Gcell system is that it shows the visualization grid of cell templates only for genes and proteins whose organism, tissue and cellular component have been reported in publicly available biological databases. This can be further improved upon to display the sub-cellular localization even if the organism and/ or tissue is unknown. The system can be made more extensive by using more URL links to the biological databases.

Like any other informatics approach, Gcell is not intended to bypass scientific experiments, but to serve as a tool to find new information and increase the chances of generating new hypothesis. Thus it helps in increasing the productivity and efficiency of researchers in the biological field. This thesis achieves intelligent mining of biological databases to arrive at pathways for a given gene from raw text.

VII. FUTURE WORK

Following are some of the future work that could be undertaken in order to extend the Gcell system. A general purpose multiple sequence alignment program for DNA or proteins could work in combination with GCell. This calculates the best match for the selected sequences, and displays the identities, similarities and differences between them. This would provide basic information for identification of conserved sequence regions that could be useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families. Gcell system with required additions can also be used for annotating metadata of pathway for specific tissues, environmental conditions and development stage of organism. Some more pathway databases like BIND and Database of Interacting Proteins (DIP) may be mined and integrated with the existing system to make the system more extensive.

VIII. REFERENCES

1. Kenneth Giffiths et al. "Approaches to Integrating Biological Data" ISMB 2000 Tutorial

2. Chou KC, Elrod DW., Protein subcellular location prediction.Protein Eng. 1999 Feb;12(2):107-18, 1999

3. Blake J. (2001). "Creating the Gene Ontology Resource: Design and Implementation" Genome Research Vol. 11, Issue 8, 1425-1433, August 2001.

4. Ng, S.K., and Wong M. (1999). Toward routine automatic pathway discovery from on-line scientific text abstracts. Genome Informatics, 10:104-112.

5. Galperin MY. The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D3-22.

6. Mostafa, J., Mukhopahyay, S., Lam, W., & Palakal, M. A Multi-level approach to intelligent information filtering: Model, system, and evaluation, Technical Report-96-01, Center for Social Informatics, Indiana University, Bloomington, 1996.

7. Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. (1999) Automatic extraction of biological information from scientific text: Protein-protein interactions. International Conference on Intelligent Systems for Molecular Biology. 60-67.

8. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A. From gene networks to gene function. Genome Res. 2003 Dec;13(12):2568-76

9. Koehler, J. “Semantic heterogeneity in biological databases”, Symposium Integrative Bioinformatics, August 4th - 5th, 2003, Bielefeld University, http://cweb.uni-bielefeld.de/agbi/home/index,id,142.html

10. Ben Miled, Z., Bukhres, O., Wang, Y., Li, N., Baumgartner, M., Sipes, B.Biological and Chemical Information Integration System. Network Tools and Applications in Biology, May, Genoa, Italy, 2001.

11. Ben Miled, Z., Liu, Y., Li, N., Bukhres, O., He, Yue., and Lynch, E., On the Integration of a Large Number of Life Science Web Databases.

12. Iwei Yeh et al. "An Ontology for Subcellular Localization". BioOntologies 2002

13. Kumar, A., S. Agarwal, et al. (2002). "Subcellular localization of the yeast proteome." Genes Dev 16(6): 707-719.

14. Dickerson, J.A., Berleant, D., Cox, Z., Qi, W., and Wurtele, E. (2001) Creating Metabolic Network Models using Text Mining and Expert Knowledge, Atlantic Symposium on Computational Biology and Genome Information Systems & Technology. 26-30.

15. Herman I, Melancon G, Marshall MS. Graph visualization and navigation in information visualization: a survey. IEEE Transactions on Visualization and Computer Graphics, 6(1): 24-43, 2000.

16. Edward M. Marcotte, Ioannis Xenarios, Alexander M. van der Bliek, and David Eisenberg. Localizing proteins in the cell from their phylogenetic profiles, Proc Natl Acad Sci U S A. 2000 October 24; 97(22): 12115–12120, 2000

17. Hu ZZ, Mani I, Hermoso V, Liu H, Wu CH. iProLINK: an integrated protein resource for literature mining. Comput Biol Chem. 2004 Dec; 28(5-6):409-16.

18. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D115-9.

19. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A.From gene networks to gene function.Genome Res. 2003 Dec;13(12):2568-76

20. Enright, A, J. and Ouzounis, C, A. (2001). BioLayout An automatic graph layout algorithm for similarity visualization. Bioinformatics.17 (9): 853-4.

21. Gary Geisler, "Making Information More Accessible: A Survey of Information Visualization Applications and Techniques", January 31, 1998.

22. Stapley, B. J., L. A. Kelley, et al. (2002). Predicting the Sub-Cellular Location of Proteins from Text Using Support Vector Machines. Pacific Symposium on Biocomputing.

23. Kanehisa M, Goto S, Kawashima S, Nakaya A., The KEGG databases at GenomeNet, Nucleic Acids Res. 2002 Jan 1; 30(1):42-6, 2002

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G,

Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nature Genetics 2000 May;25(1):25-9, 2000

25. Fu, Y., Bauer, T., Mostafa, J., Palakal, M., and Mukhopadhyay, S. (2002). Concept extraction and association from cancer literature. WIDM 2002: 100-103.

26. Witten I,Bray Z, Mahoui M, Teahan B, Text Mining: A New Frontier for Lossless Compression. ACM Proceedings of the Conference on Data Compression Page: 198, 1999

27. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.Nucleic Acids Res. 2003 Jan 1;31(1):219-23.

28. Drawid, A. and M. Gerstein (2000). "A Bayesian system integrating expression data with sequence patterns for localizaing proteins: comprehensive application to the yeast genome." J Mol Biol. 301(4): 1059-1075.

29. Mrowka, R., (2001) A Java applet for visualizing protein-protein interaction. Bioinformatics. 17(7): 669-71.

30. Mukhopadhyay, S., Mostafa, J., Palakal, M., Lam, W., Xue, L., and Hudli, A. (1996). An Adaptive Multi-level Information Filtering System. Proceedings of the Fifth International Conference on User Modeling. 21-28.

31. Palakal, M., Mukhopadhyay, S., Mostafa, J., Raje, R., N'Cho, M., and Mishra, S. (2002). An intelligent biological information management system. Bioinformatics 18(10): 1283-1288.

32. Swanson, D.R. and Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91: 183-203.

33. Karp, P. D., M. Riley, et al. (2002). "The EcoCyc Database." Nucleic Acids Research 30(1): 56.

34. Salton, G. (1989) Automatic text processing: The transformation, analysis, and retrieval of information by computer, Addison–Wesley, Reading, MA, USA.

35. Tan, A-H. (1999) Text mining: The state of the art and the challenges. In Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases. 65-70.

36. Bader GD, Betel D, Hogue C, BIND: the Biomolecular Interaction Network Database, Nucleic Acids Res. 2003 Jan 1; 31(1):248-50, 2003

37. Smalheiser, N.R. (2001). Predicting emerging technologies with the aid of text-based data mining: a micro approach. Technovation 21: 689-693.

38. Stephens, M. (2001) Extracting Biological Relationships from Text. Masters Thesis, Purdue University.

39. Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. and Mostafa, J. (2001) Detecting Gene Relations from Medline Abstracts. Pacific Symposium on Biocomputing, Honolulu, Hawaii. 483-95.

40. Martucci D, Masseroli M, Pinciroli F. Gene Ontology application to genomic functional annotation, statistical analysis and knowledge mining. Ontologies in Medicine, IOS Press, 2004

41. Shah SP, Huang Y, Xu T, Yuen MMS, Ling J, Ouellette BBF: Atlas: a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005, 6:34

42. Mootha V, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell G, Morin C, Mann M, Hudson T, Robinson B, Rioux J, Lander E: Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci U S A 2003, 100(2):605-610

43. Wheeler D, Church D, Edgar R, Federhen S, Helmberg W, Madden T, Pontius J, Schuler G, Schriml L, Sequeira E, Suzek T, Tatusova T, Wagner L: Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004, (32 Database):35-40.

44. Stuart J, Segal E, Koller D, Kim S: A gene-coexpression network for global discovery of conserved genetic modules. Science 2003, 302(5643):249-255

45. http://homepage.usask.ca/~ctl271/857/def_homolog.shtml

46. http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

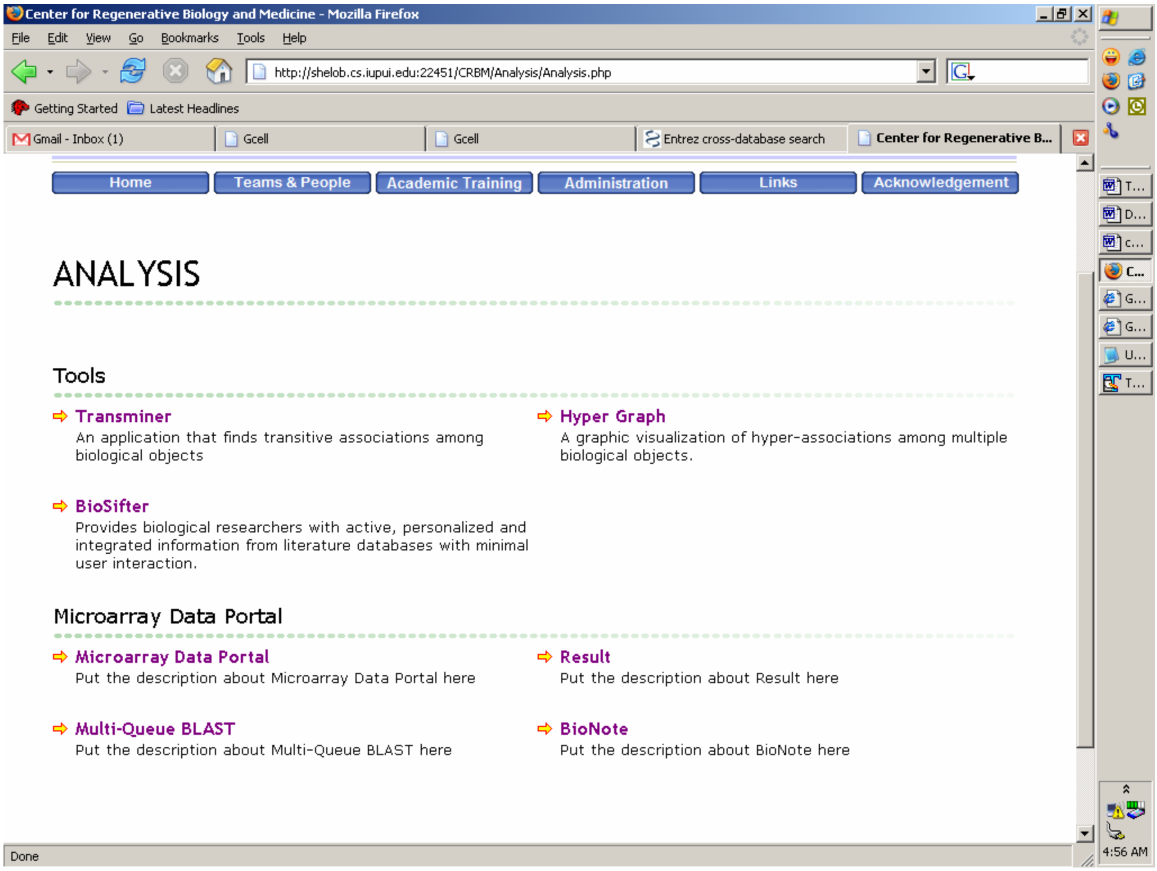
47. http://www.ncbi.nlm.nih.gov/

48. http://www.genome.jp/kegg/

49. http://bind.ca/

IX. Appendix

A. Snapshots of the Gcell System


Center for Regenerative Biology and Medicine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:22451/CRBM/Analysis/Analysis.php
Getting Started
Latest Headlines
Gmail - Inbox (1)
Gcell
Gcell
Entrez cross-database search
Center for Regenerative B...
Home
Teams & People
Academic Training
Administration
Links
Acknowledgement
Done
4:56 AM

ANALYSIS

Tools

⇨ **Transminer**
An application that finds transitive associations among biological objects

⇨ **Hyper Graph**
A graphic visualization of hyper-associations among multiple biological objects.

⇨ **BioSifter**
Provides biological researchers with active, personalized and integrated information from literature databases with minimal user interaction.

Microarray Data Portal

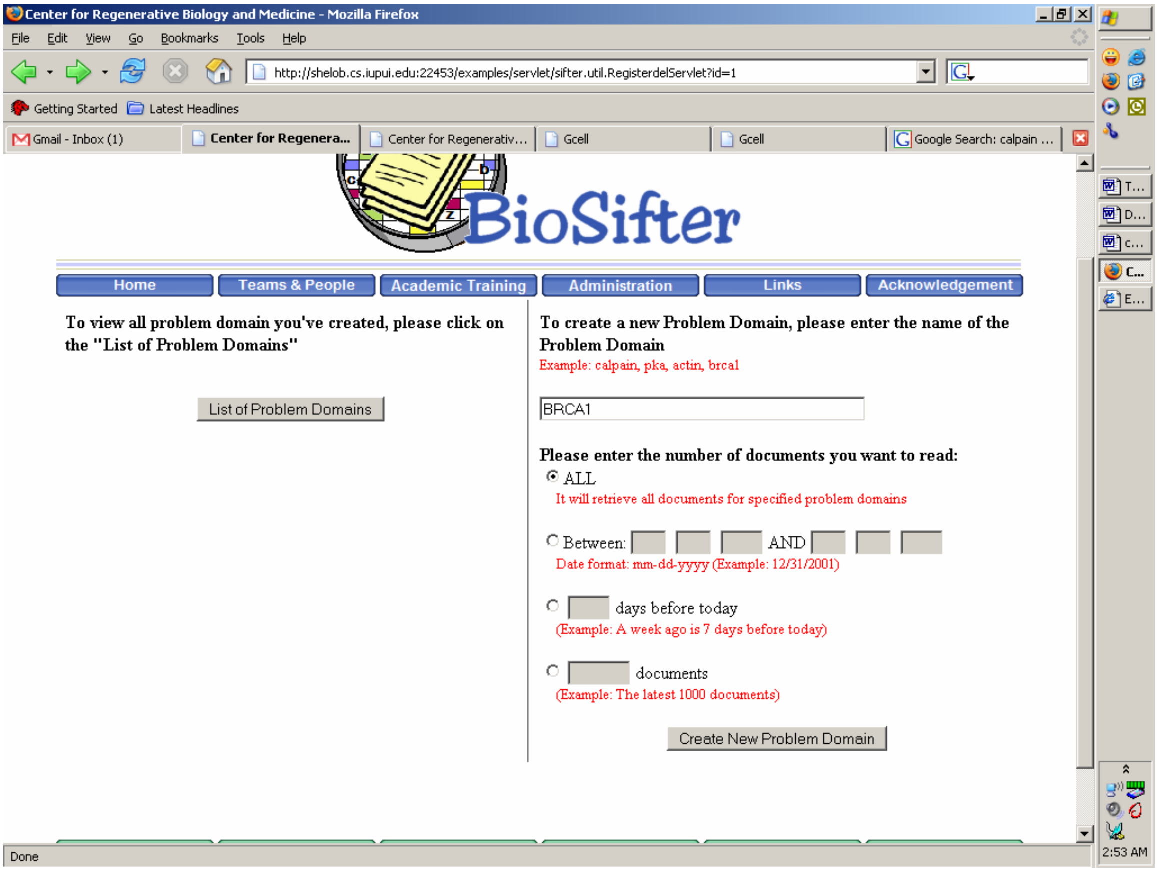
⇨ **Microarray Data Portal**
Put the description about Microarray Data Portal here

⇨ **Result**
Put the description about Result here

⇨ **Multi-Queue BLAST**
Put the description about Multi-Queue BLAST here

⇨ **BioNote**
Put the description about BioNote here

Figure A1. Home Page for BioSifter and TransMiner


Center for Regenerative Biology and Medicine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:22453/examples/servlet/sifter.util.RegisterdelServlet?id=1
Getting Started
Latest Headlines
Gmail - Inbox (1)
Center for Regenera...
Center for Regenerativ...
Gcell
Gcell
Google Search: calpain ...
BioSifter
Home
Teams & People
Academic Training
Administration
Links
Acknowledgement
Done
2:53 AM

To view all problem domain you've created, please click on the "List of Problem Domains"

List of Problem Domains

To create a new Problem Domain, please enter the name of the Problem Domain

Example: calpain, pka, actin, brca1

BRCA1

Please enter the number of documents you want to read:

● ALL

It will retrieve all documents for specified problem domains

○ Between: ___ ___ ___ AND ___ ___ ___

Date format: mm-dd-yyyy (Example: 12/31/2001)

○ ___ days before today

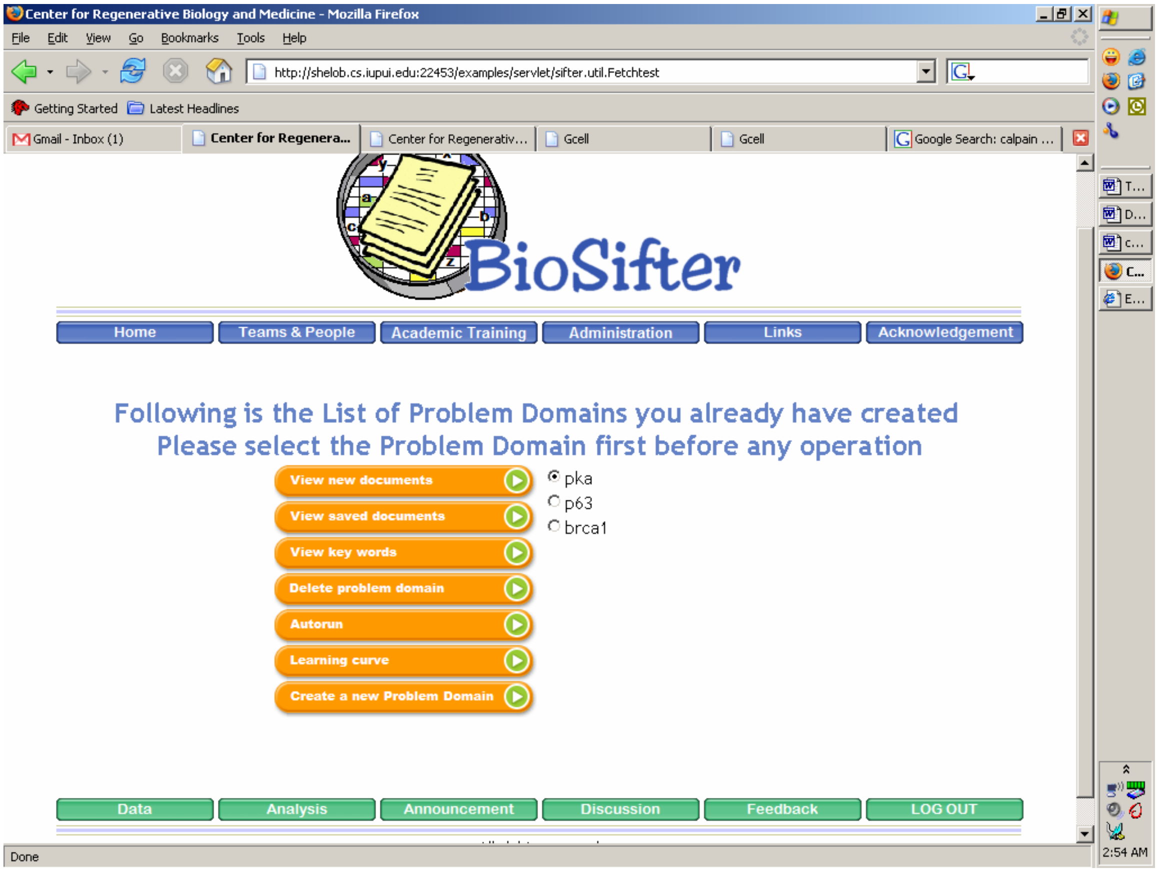
(Example: A week ago is 7 days before today)

○ ___ documents

(Example: The latest 1000 documents)

Create New Problem Domain

Figure A2. BioSifter: Creating a new problem domain and/or viewing the existing ones

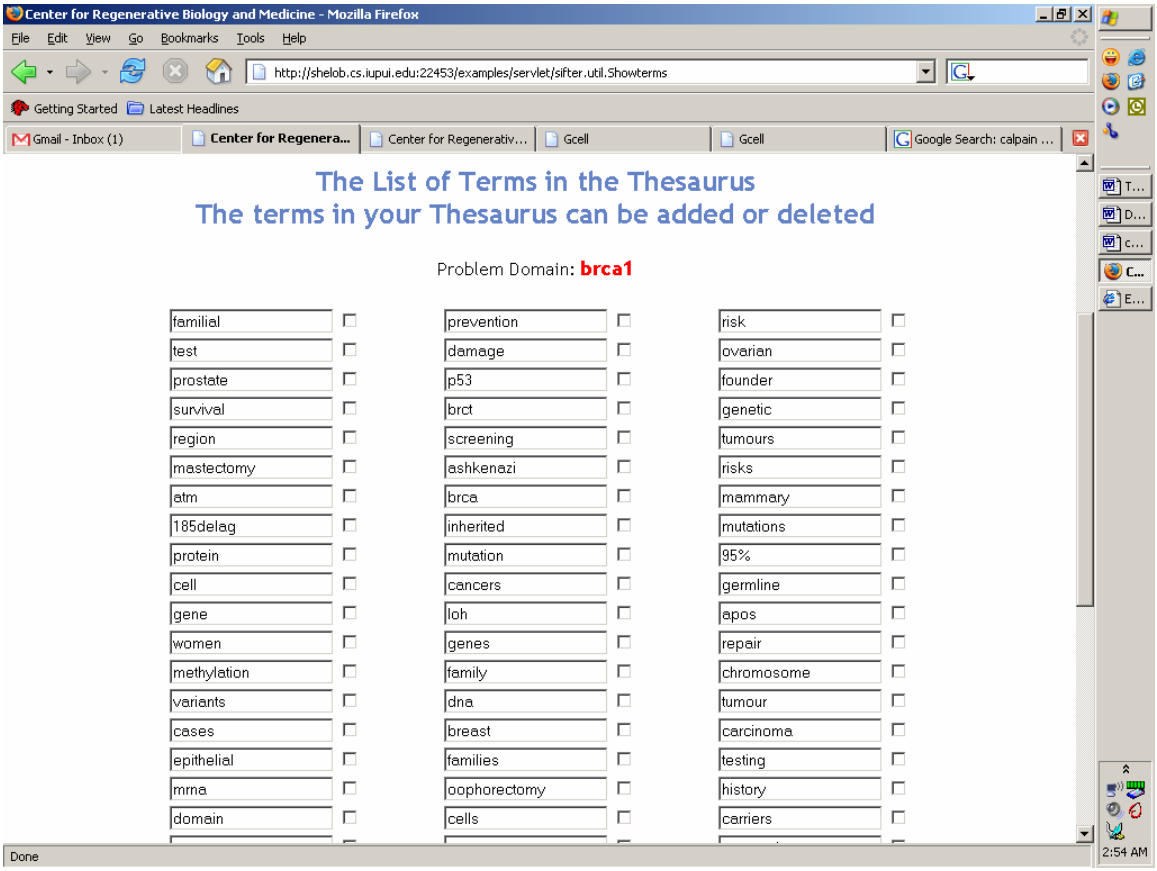


Center for Regenerative Biology and Medicine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:22453/examples/servlet/sifter.util.Fetchtest
Getting Started Latest Headlines
Gmail - Inbox (1)
Center for Regenera...
Center for Regenerativ...
Gcell
Gcell
Google Search: calpain ...
BioSifter
Home
Teams & People
Academic Training
Administration
Links
Acknowledgement
View new documents
View saved documents
View key words
Delete problem domain
Autorun
Learning curve
Create a new Problem Domain
Data
Analysis
Announcement
Discussion
Feedback
LOG OUT
Done
2:54 AM

Following is the List of Problem Domains you already have created
Please select the Problem Domain first before any operation

● pka
○ p63
○ brca1

Fugure A3. BioSifter: Various options for the problem domains created



Center for Regenerative Biology and Medicine - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://shelob.cs.iupui.edu:22453/examples/servlet/sifter.util.Showterms
Getting Started
Latest Headlines
Gmail - Inbox (1)
Center for Regenera...
Center for Regenerativ...
Gcell
Gcell
Google Search: calpain ...
T...
D...
c...
C...
E...
Done
2:54 AM

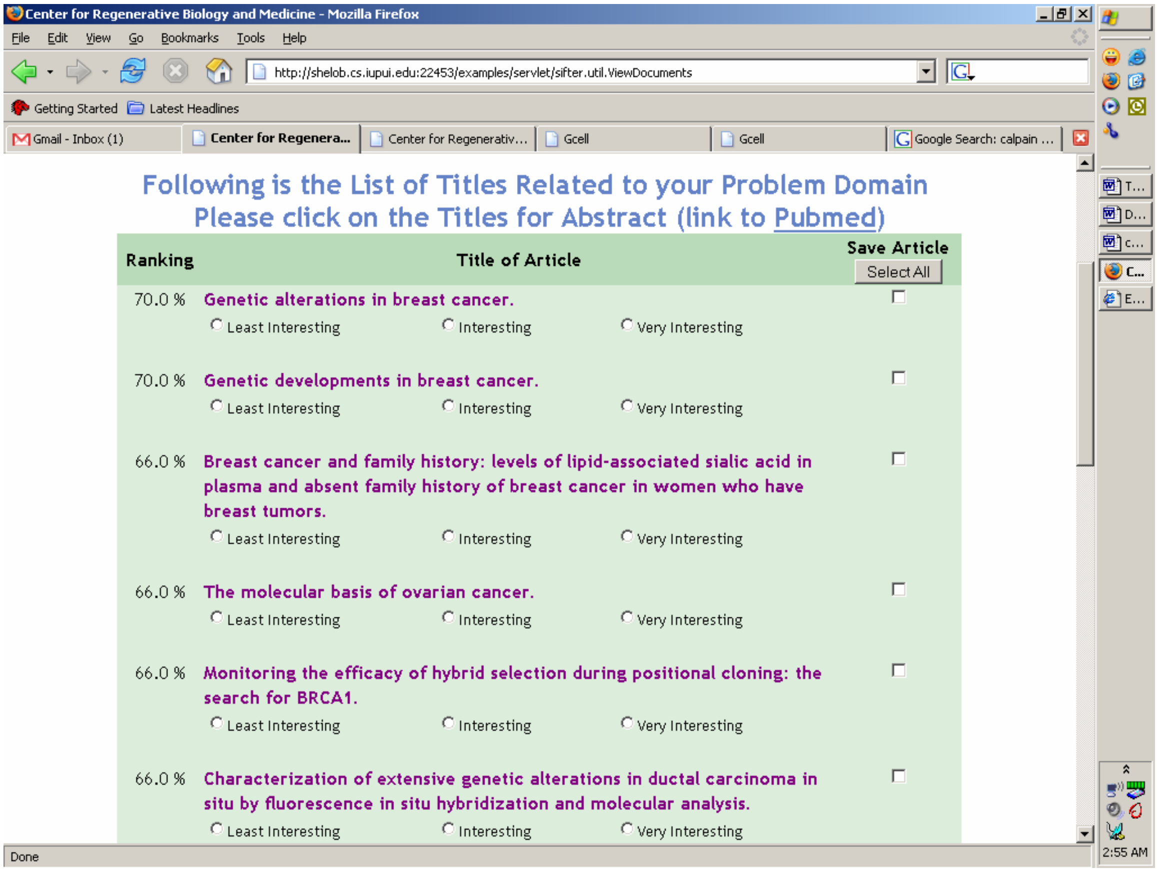
The List of Terms in the Thesaurus

The terms in your Thesaurus can be added or deleted

Problem Domain: **brca1**

familial	☐	prevention	☐	risk	☐
test	☐	damage	☐	ovarian	☐
prostate	☐	p53	☐	founder	☐
survival	☐	brct	☐	genetic	☐
region	☐	screening	☐	tumours	☐
mastectomy	☐	ashkenazi	☐	risks	☐
atm	☐	brca	☐	mammary	☐
185delag	☐	inherited	☐	mutations	☐
protein	☐	mutation	☐	95%	☐
cell	☐	cancers	☐	germline	☐
gene	☐	loh	☐	apos	☐
women	☐	genes	☐	repair	☐
methylation	☐	family	☐	chromosome	☐
variants	☐	dna	☐	tumour	☐
cases	☐	breast	☐	carcinoma	☐
epithelial	☐	families	☐	testing	☐
mrna	☐	oophorectomy	☐	history	☐
domain	☐	cells	☐	carriers	☐

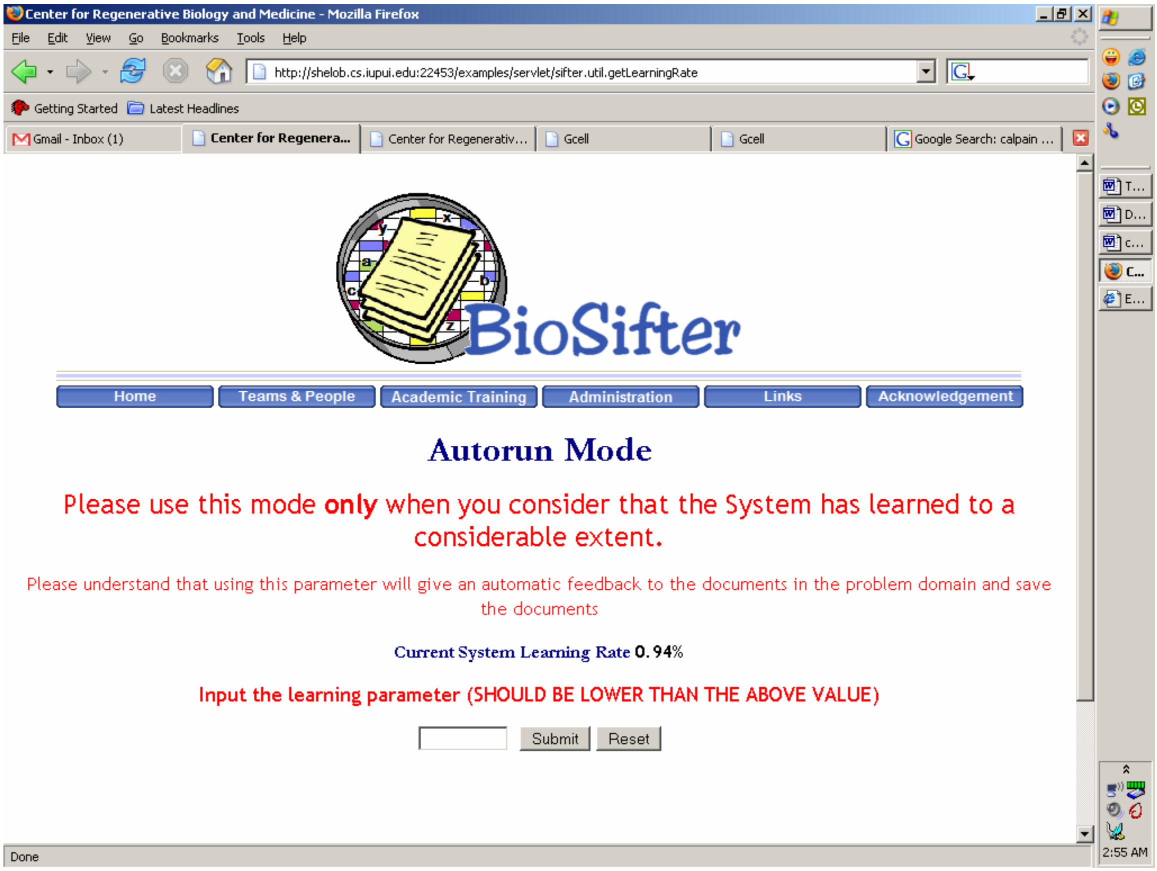
Figure A 4. BioSifter: Add/ Modify/ Delete thesaurus terms


Center for Regenerative Biology and Medicine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:22453/examples/servlet/sifter.util.ViewDocuments
Getting Started
Latest Headlines
Gmail - Inbox (1)
Center for Regenera...
Center for Regenerativ...
Gcell
Gcell
Google Search: calpain ...
Done
2:55 AM

Following is the List of Titles Related to your Problem Domain
Please click on the Titles for Abstract (link to Pubmed)

Ranking	Title of Article	Save Article Select All
70.0 %	Genetic alterations in breast cancer. ○ Least Interesting ○ Interesting ○ Very Interesting	☐
70.0 %	Genetic developments in breast cancer. ○ Least Interesting ○ Interesting ○ Very Interesting	☐
66.0 %	Breast cancer and family history: levels of lipid-associated sialic acid in plasma and absent family history of breast cancer in women who have breast tumors. ○ Least Interesting ○ Interesting ○ Very Interesting	☐
66.0 %	The molecular basis of ovarian cancer. ○ Least Interesting ○ Interesting ○ Very Interesting	☐
66.0 %	Monitoring the efficacy of hybrid selection during positional cloning: the search for BRCA1. ○ Least Interesting ○ Interesting ○ Very Interesting	☐
66.0 %	Characterization of extensive genetic alterations in ductal carcinoma in situ by fluorescence in situ hybridization and molecular analysis. ○ Least Interesting ○ Interesting ○ Very Interesting	☐

Figure A 5. BioSifter: View the Ranked documents



Center for Regenerative Biology and Medicine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:22453/examples/servlet/sifter.util.getLearningRate
Getting Started
Latest Headlines
Gmail - Inbox (1)
Center for Regenera...
Center for Regenerativ...
Gcell
Gcell
Google Search: calpain ...
BioSifter
Home
Teams & People
Academic Training
Administration
Links
Acknowledgement
Submit
Reset
Done
2:55 AM

Autorun Mode

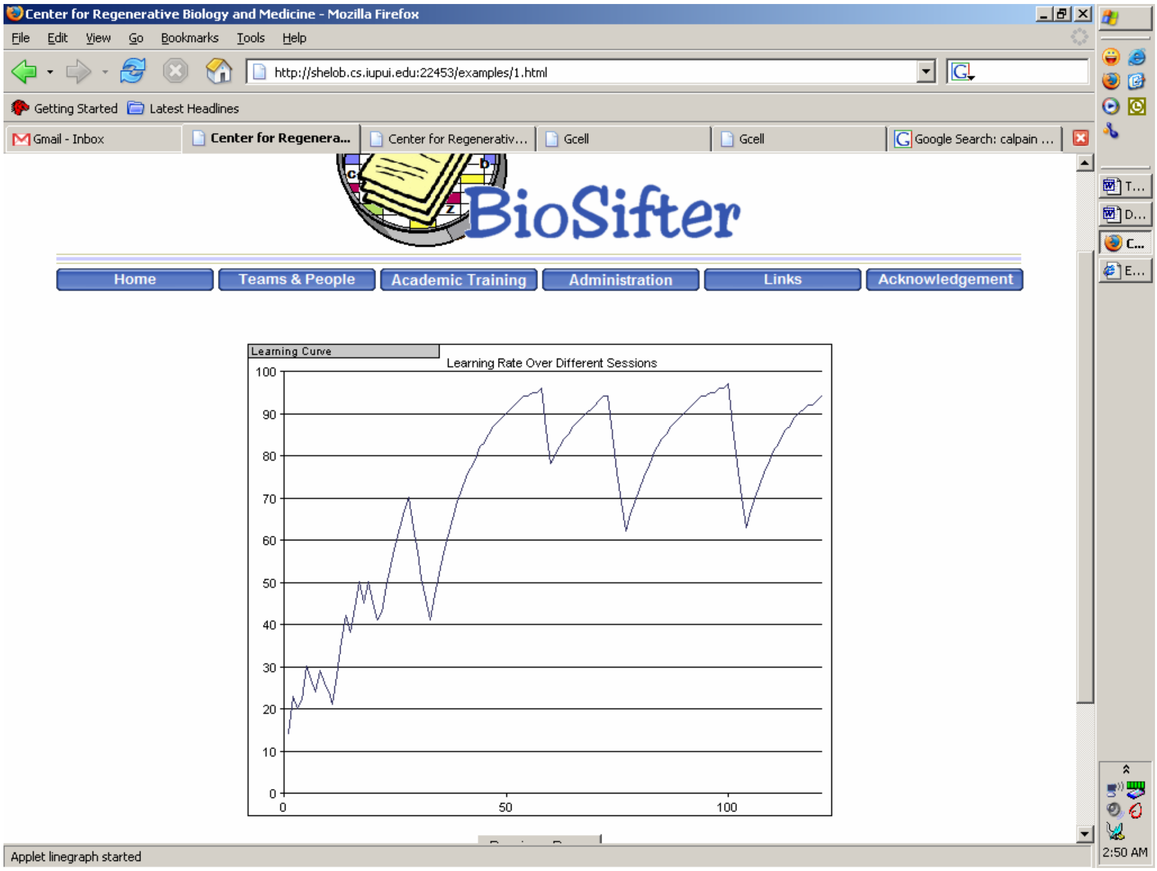
Please use this mode **only** when you consider that the System has learned to a considerable extent.

Please understand that using this parameter will give an automatic feedback to the documents in the problem domain and save the documents

Current System Learning Rate 0.94%

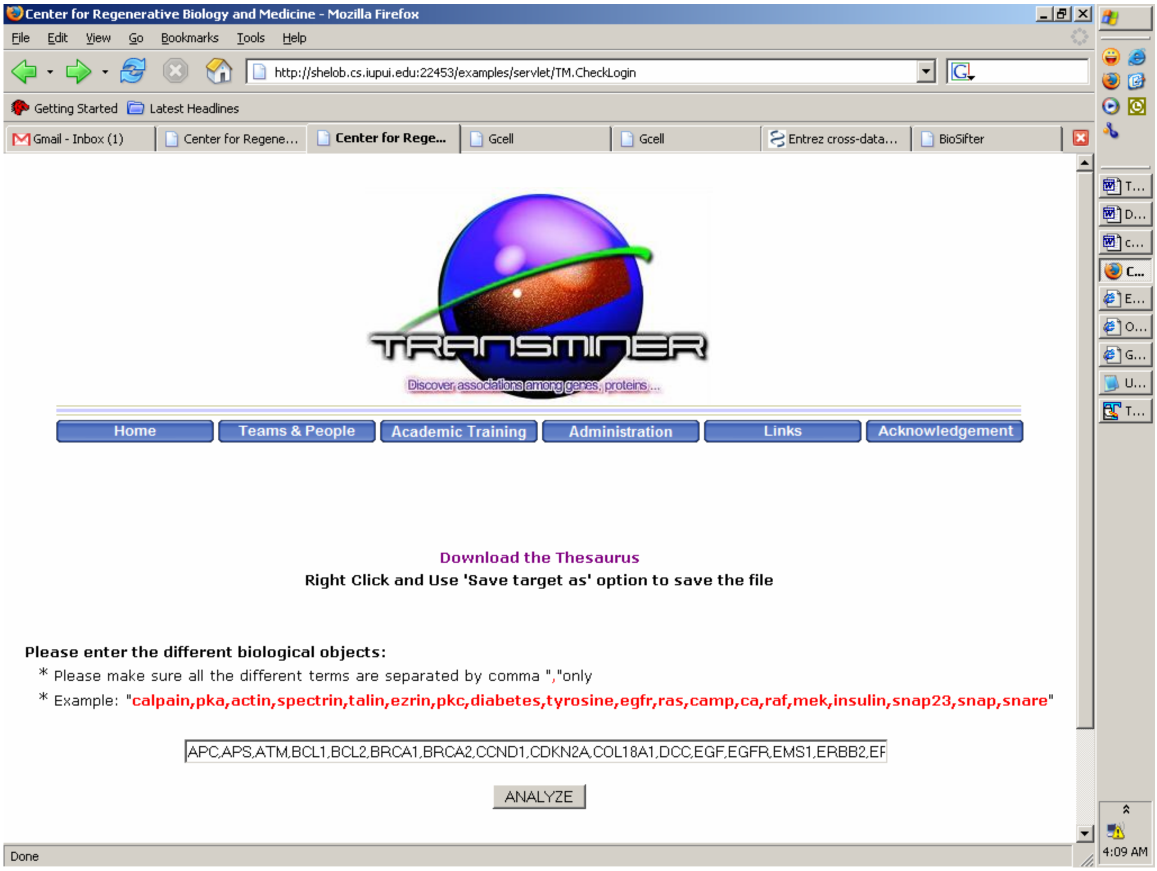
Input the learning parameter (SHOULD BE LOWER THAN THE ABOVE VALUE)

Figure A 6. BioSifter: Providing automatic feedback to the unseen documents


Center for Regenerative Biology and Medicine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:22453/examples/1.html
Getting Started
Latest Headlines
Gmail - Inbox
Center for Regenera...
Center for Regenerativ...
Gcell
Gcell
Google Search: calpain ...
BioSifter
Home
Teams & People
Academic Training
Administration
Links
Acknowledgement
Learning Curve
Learning Rate Over Different Sessions
100
90
80
70
60
50
40
30
20
10
0
0
50
100
2:50 AM

Applet linegraph started

Figure A 7. BioSifter: Learning Curve for the user profile



http://shelob.cs.iupui.edu:22453/examples/servlet/TM.CheckLogin

Download the Thesaurus

Right Click and Use 'Save target as' option to save the file

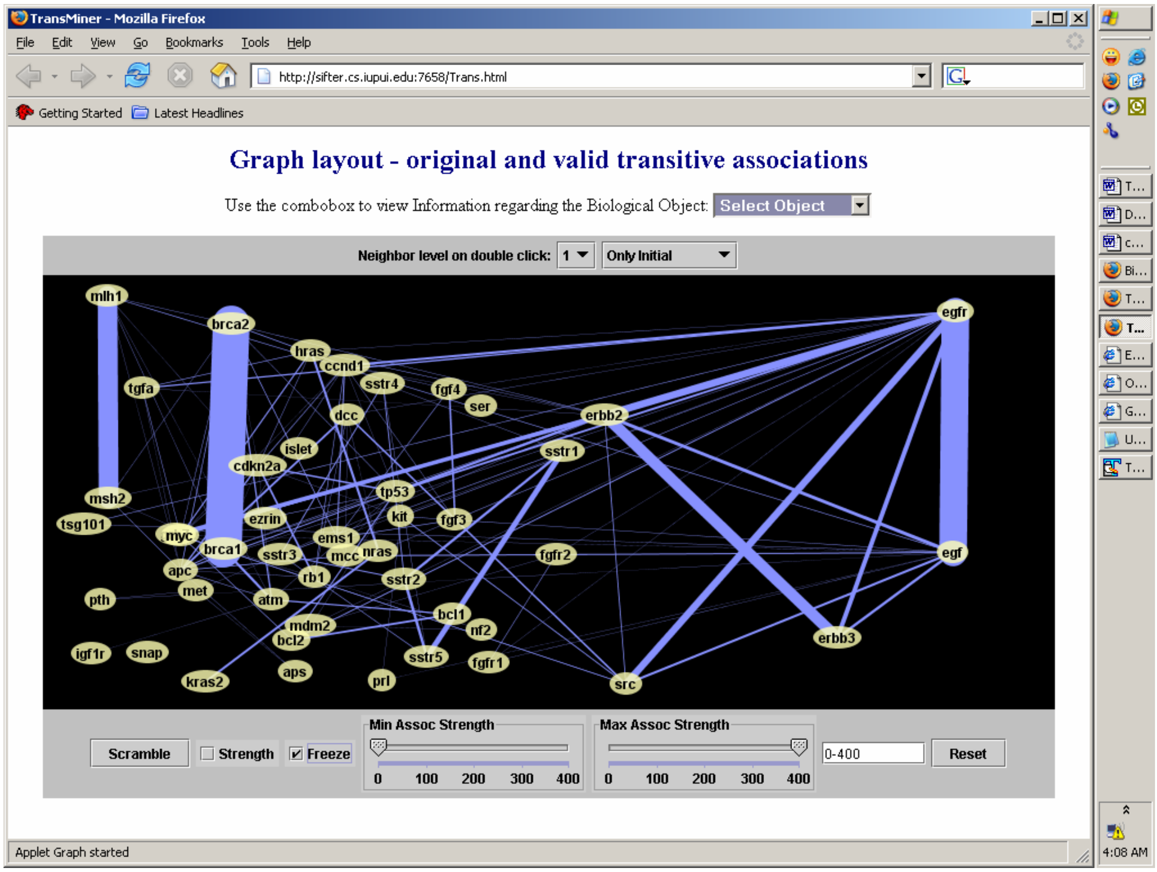
Please enter the different biological objects:

* Please make sure all the different terms are separated by comma ","only

* Example: "calpain,pka,actin,spectrin,talin,ezrin,pkc,diabetes,tyrosine,egfr,ras,camp,ca,raf,mek,insulin,snap23,snap,snare"

APC,APS,ATM,BCL1,BCL2,BRCA1,BRCA2,CCND1,CDKN2A,COL18A1,DCC,EGF,EGFR,EMS1,ERBB2,EF

Figure A 8. TransMiner: Main Page for entering the object names



Graph layout - original and valid transitive associations

Use the combobox to view Information regarding the Biological Object: Select Object

Neighbor level on double click: 1 Only Initial

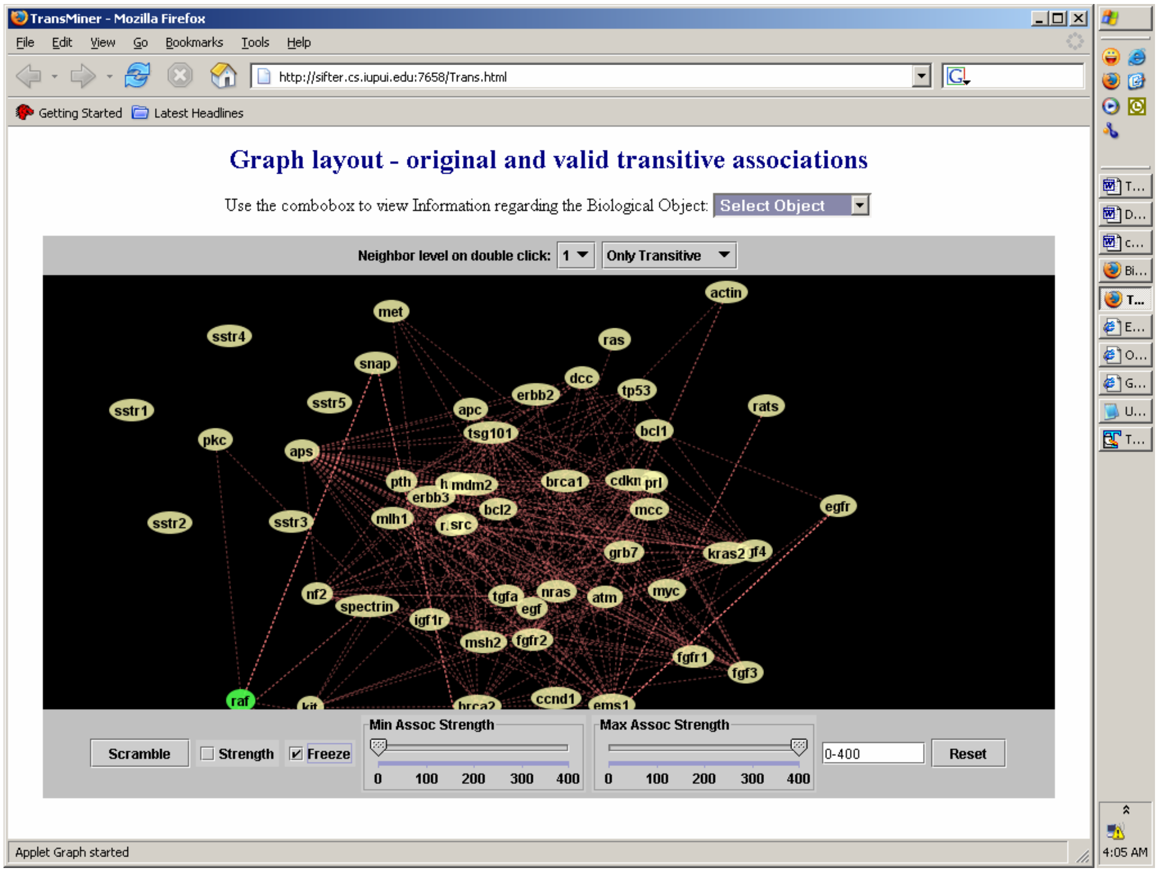
Min Assoc Strength

Max Assoc Strength

0-400

Applet Graph started

Figure A 9. TransMiner: Viewing the Direct Associations between entities



Graph layout - original and valid transitive associations

Use the combobox to view Information regarding the Biological Object: Select Object

Neighbor level on double click: 1 Only Transitive

□ Strength ☑ Freeze

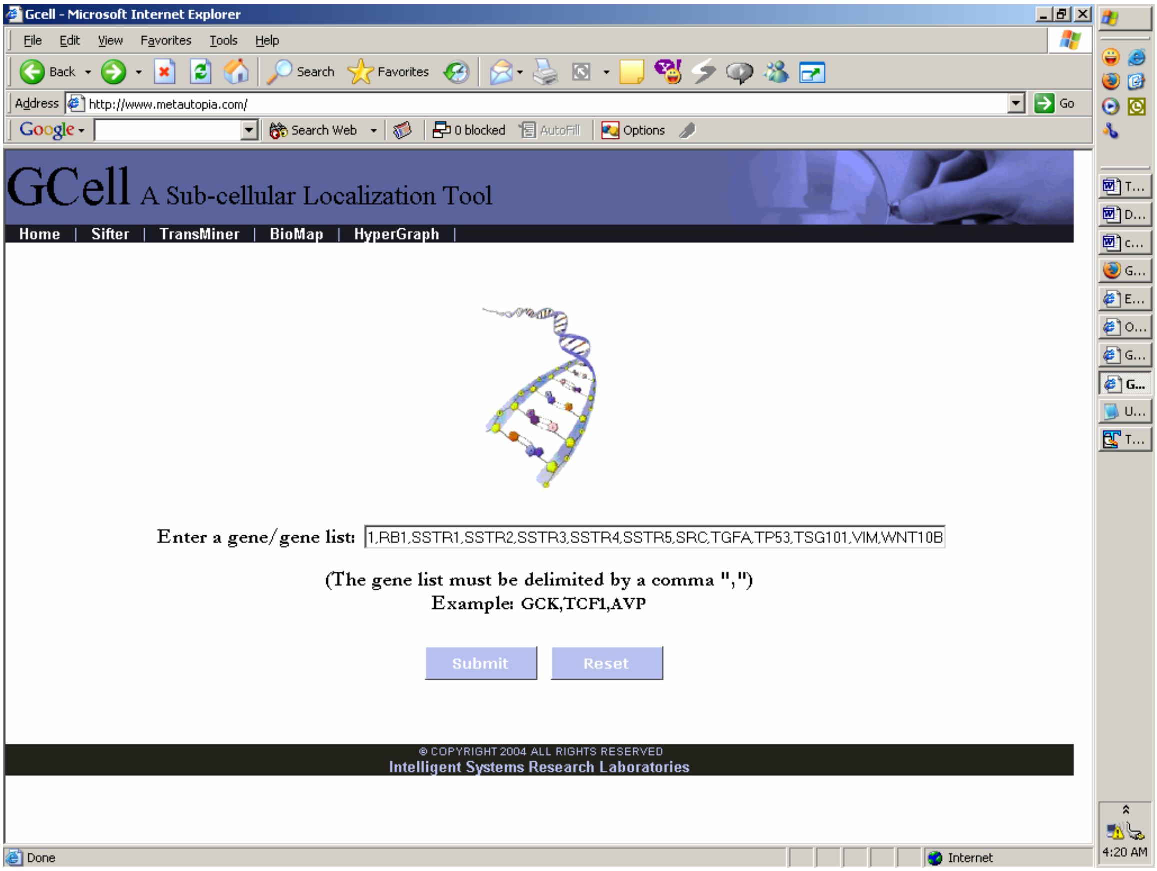
Min Assoc Strength

Max Assoc Strength

0-400

Applet Graph started

Figure A 10. TransMiner: Viewing the Transitive Associations



Address: http://www.metautopia.com/

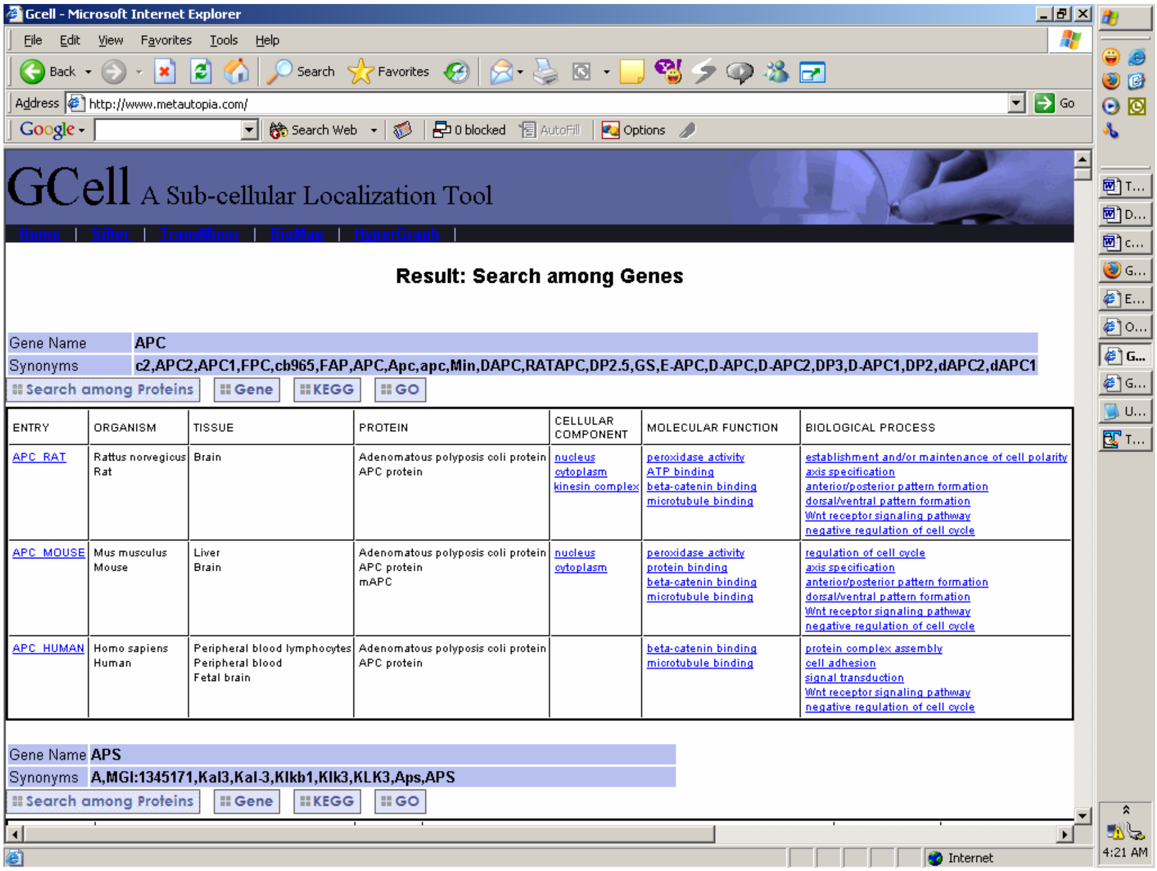
GCell A Sub-cellular Localization Tool

Enter a gene/gene list: 1,RB1,SSTR1,SSTR2,SSTR3,SSTR4,SSTR5,SRC,TGFA,TP53,TSG101,VIM,WNT10B

(The gene list must be delimited by a comma ",")
Example: GCK,TCF1,AVP

© COPYRIGHT 2004 ALL RIGHTS RESERVED
Intelligent Systems Research Laboratories

Figure A 11. GCell: Main Page for entering the Genes / Proteins


Gcell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.metautopia.com/ Go
Google Search Web 0 blocked AutoFill Options
GCell A Sub-cellular Localization Tool
Home | Sifter | TransMiner | BioMap | HyperGraph |
Internet
4:21 AM

Result: Search among Genes

Gene Name	APC
Synonyms	c2,APC2,APC1,FPC,cb965,FAP,APC,Apc,apc,Min,DAPC,RATAPC,DP2.5,GS,E-APC,D-APC,D-APC2,DP3,D-APC1,DP2,dAPC2,dAPC1

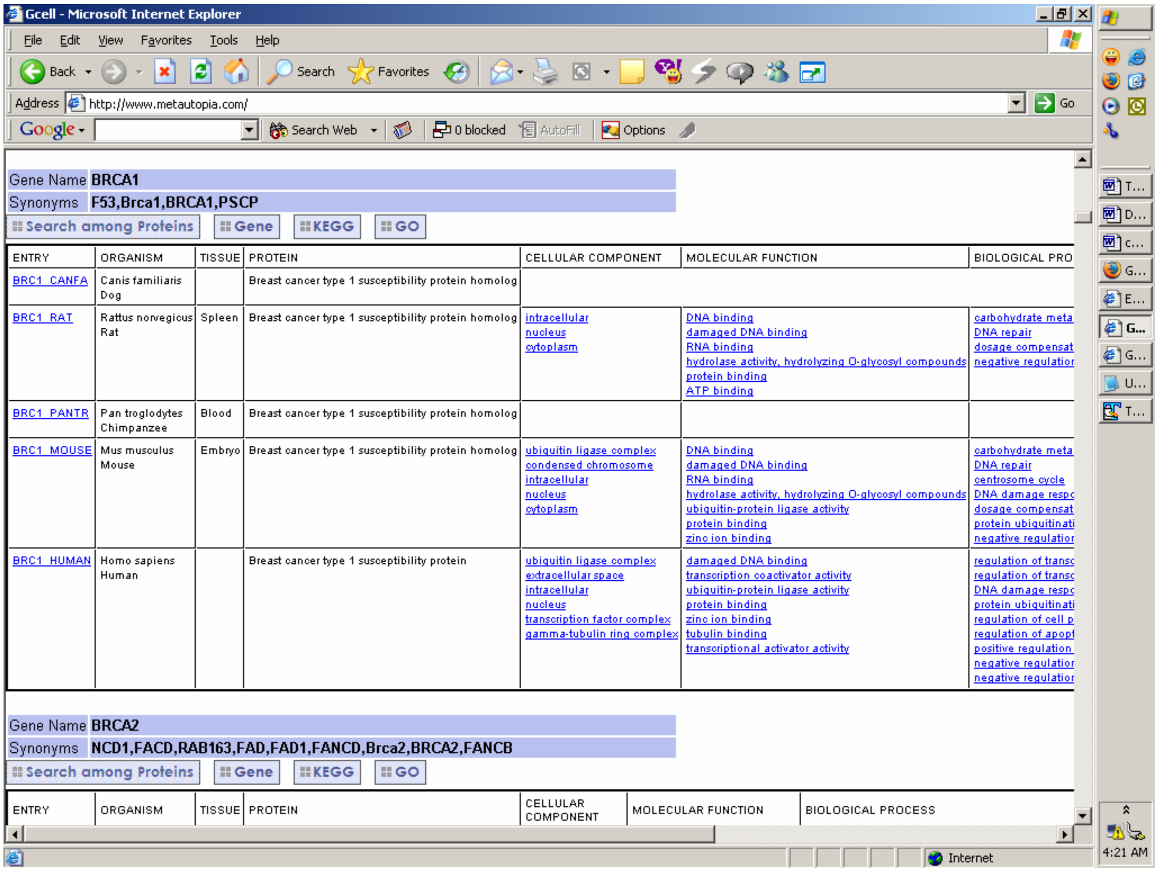
Search among Proteins | Gene | KEGG | GO

ENTRY	ORGANISM	TISSUE	PROTEIN	CELLULAR COMPONENT	MOLECULAR FUNCTION	BIOLOGICAL PROCESS
APC_RAT	Rattus norvegicus Rat	Brain	Adenomatous polyposis coli protein APC protein	nucleus cytoplasm kinesin complex	peroxidase activity ATP binding beta-catenin binding microtubule binding	establishment and/or maintenance of cell polarity axis specification anterior/posterior pattern formation dorsal/ventral pattern formation Wnt receptor signaling pathway negative regulation of cell cycle
APC_MOUSE	Mus musculus Mouse	Liver Brain	Adenomatous polyposis coli protein APC protein mAPC	nucleus cytoplasm	peroxidase activity protein binding beta-catenin binding microtubule binding	regulation of cell cycle axis specification anterior/posterior pattern formation dorsal/ventral pattern formation Wnt receptor signaling pathway negative regulation of cell cycle
APC_HUMAN	Homo sapiens Human	Peripheral blood lymphocytes Peripheral blood Fetal brain	Adenomatous polyposis coli protein APC protein		beta-catenin binding microtubule binding	protein complex assembly cell adhesion signal transduction Wnt receptor signaling pathway negative regulation of cell cycle

Gene Name	APS
Synonyms	A,MGI:1345171,Kal3,Kal-3,Klkb1,Klk3,KLK3,Aps,APS

Search among Proteins | Gene | KEGG | GO

Figure A 12. GCell: Displaying variety of Information regarding the genes


Gcell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.metautopia.com/ Go
Google Search Web 0 blocked AutoFill Options
Internet
4:21 AM

Gene Name BRCA1
Synonyms F53,Brca1,BRCA1,PSCP

Search among Proteins | Gene | KEGG | GO

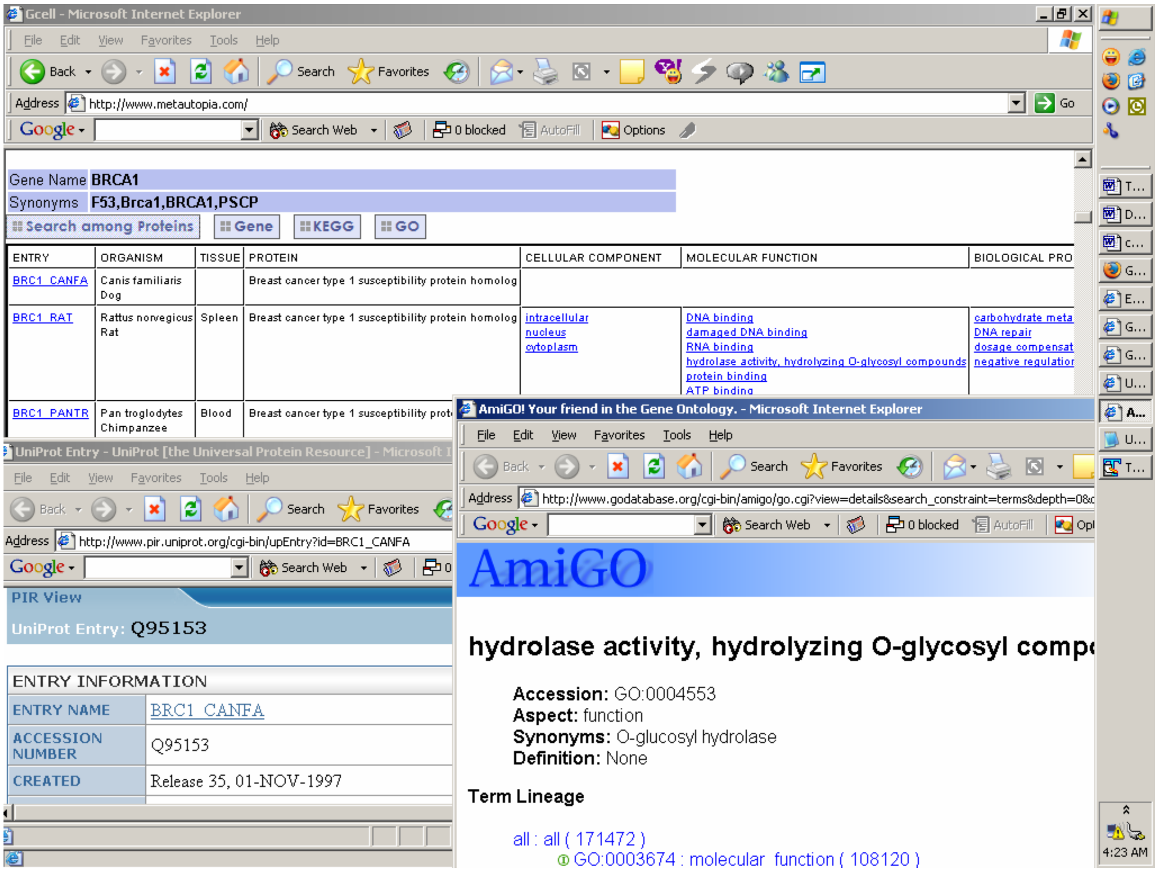
ENTRY	ORGANISM	TISSUE	PROTEIN	CELLULAR COMPONENT	MOLECULAR FUNCTION	BIOLOGICAL PRO
BRC1_CANFA	Canis familiaris Dog		Breast cancer type 1 susceptibility protein homolog			
BRC1_RAT	Rattus norvegicus Rat	Spleen	Breast cancer type 1 susceptibility protein homolog	intracellular nucleus cytoplasm	DNA binding damaged DNA binding RNA binding hydrolase activity, hydrolyzing O-glycosyl compounds protein binding ATP binding	carbohydrate meta DNA repair dosage compensat negative regulatior
BRC1_PANTR	Pan troglodytes Chimpanzee	Blood	Breast cancer type 1 susceptibility protein homolog			
BRC1_MOUSE	Mus musculus Mouse	Embryo	Breast cancer type 1 susceptibility protein homolog	ubiquitin ligase complex condensed chromosome intracellular nucleus cytoplasm	DNA binding damaged DNA binding RNA binding hydrolase activity, hydrolyzing O-glycosyl compounds ubiquitin-protein ligase activity protein binding zinc ion binding	carbohydrate meta DNA repair centrosome cycle DNA damage respo dosage compensat protein ubiquitinati negative regulatior
BRC1_HUMAN	Homo sapiens Human		Breast cancer type 1 susceptibility protein	ubiquitin ligase complex extracellular space intracellular nucleus transcription factor complex gamma-tubulin ring complex	damaged DNA binding transcription coactivator activity ubiquitin-protein ligase activity protein binding zinc ion binding tubulin binding transcriptional activator activity	regulation of transc regulation of transc DNA damage respo protein ubiquitinati regulation of cell p regulation of apopt positive regulation negative regulatior negative regulatior

Gene Name BRCA2
Synonyms NCD1,FACD,RAB163,FAD,FAD1,FANCD,Brca2,BRCA2,FANCB

Search among Proteins | Gene | KEGG | GO

ENTRY	ORGANISM	TISSUE	PROTEIN	CELLULAR COMPONENT	MOLECULAR FUNCTION	BIOLOGICAL PROCESS

Figure A 13. GCell: Displaying variety of Information regarding BRCA1 gene


Gcell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Google
Search Web
0 blocked
AutoFill
Options
Search among Proteins
Gene
KEGG
GO
UniProt Entry - UniProt [the Universal Protein Resource] - Microsoft I
AmiGO! Your friend in the Gene Ontology. - Microsoft Internet Explorer
PIR View
AmiGO
4:23 AM

Address http://www.metautopia.com/

Gene Name BRCA1

Synonyms F53,Brca1,BRCA1,PSCP

ENTRY	ORGANISM	TISSUE	PROTEIN	CELLULAR COMPONENT	MOLECULAR FUNCTION	BIOLOGICAL PRO
BRC1_CANFA	Canis familiaris Dog		Breast cancer type 1 susceptibility protein homolog			
BRC1_RAT	Rattus norvegicus Rat	Spleen	Breast cancer type 1 susceptibility protein homolog	intracellular nucleus cytoplasm	DNA binding damaged DNA binding RNA binding hydrolase activity, hydrolyzing O-glycosyl compounds protein binding ATP binding	carbohydrate meta DNA repair dosage compensat negative regulatior
BRC1_PANTR	Pan troglodytes Chimpanzee	Blood	Breast cancer type 1 susceptibility prot			

Address http://www.pir.uniprot.org/cgi-bin/upEntry?id=BRC1_CANFA

UniProt Entry: Q95153

ENTRY INFORMATION	
ENTRY NAME	BRC1_CANFA
ACCESSION NUMBER	Q95153
CREATED	Release 35, 01-NOV-1997

Address http://www.godatabase.org/cgi-bin/amigo/go.cgi?view=details&search_constraint=terms&depth=0&c

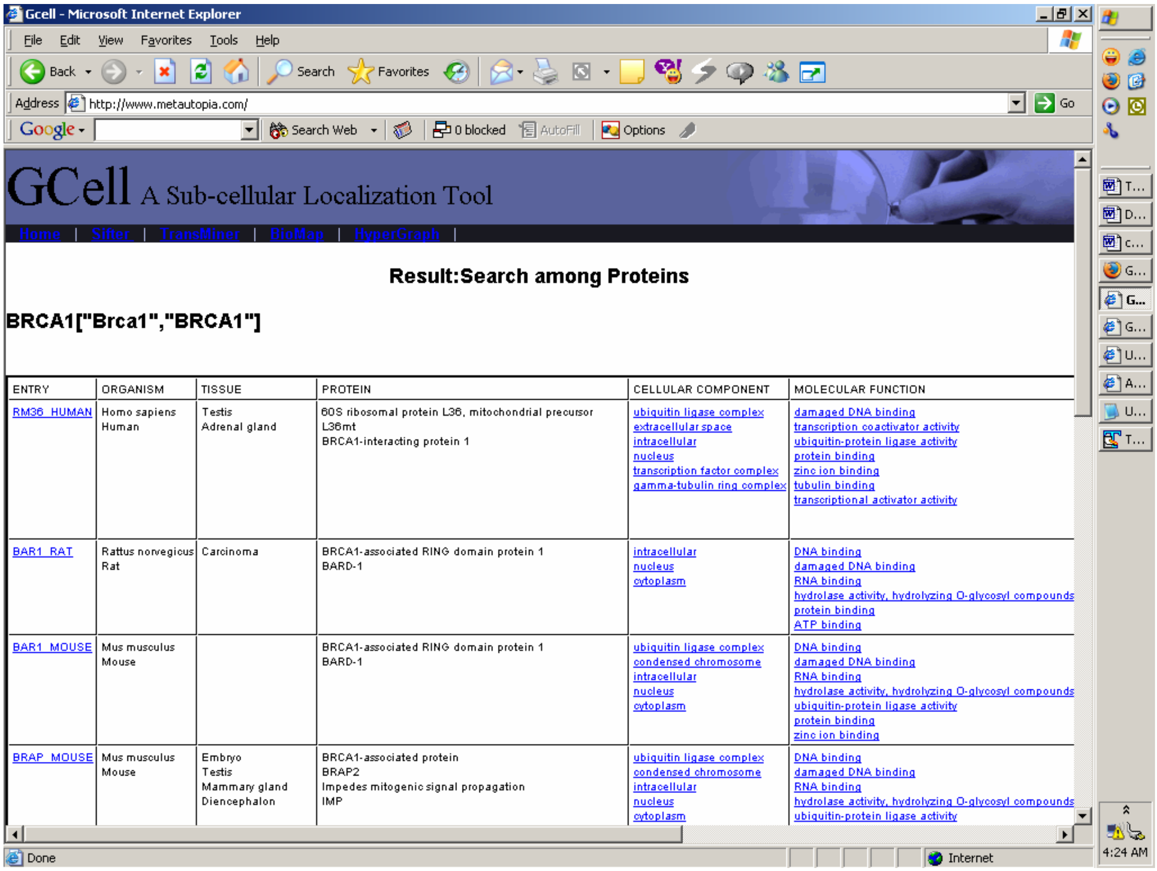
hydrolase activity, hydrolyzing O-glycosyl compo

Accession: GO:0004553
Aspect: function
Synonyms: O-glucosyl hydrolase
Definition: None

Term Lineage

all : all (171472)
GO:0003674 : molecular function (108120)

Figure A 14. GCell: Hyperlinks to Uniprot and Gene Ontology databases


Gcell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.metautopia.com/ Go
Google Search Web 0 blocked AutoFill Options
Home | Sifter | TransMiner | DioMap | HyperGraph |
Done
Internet
4:24 AM

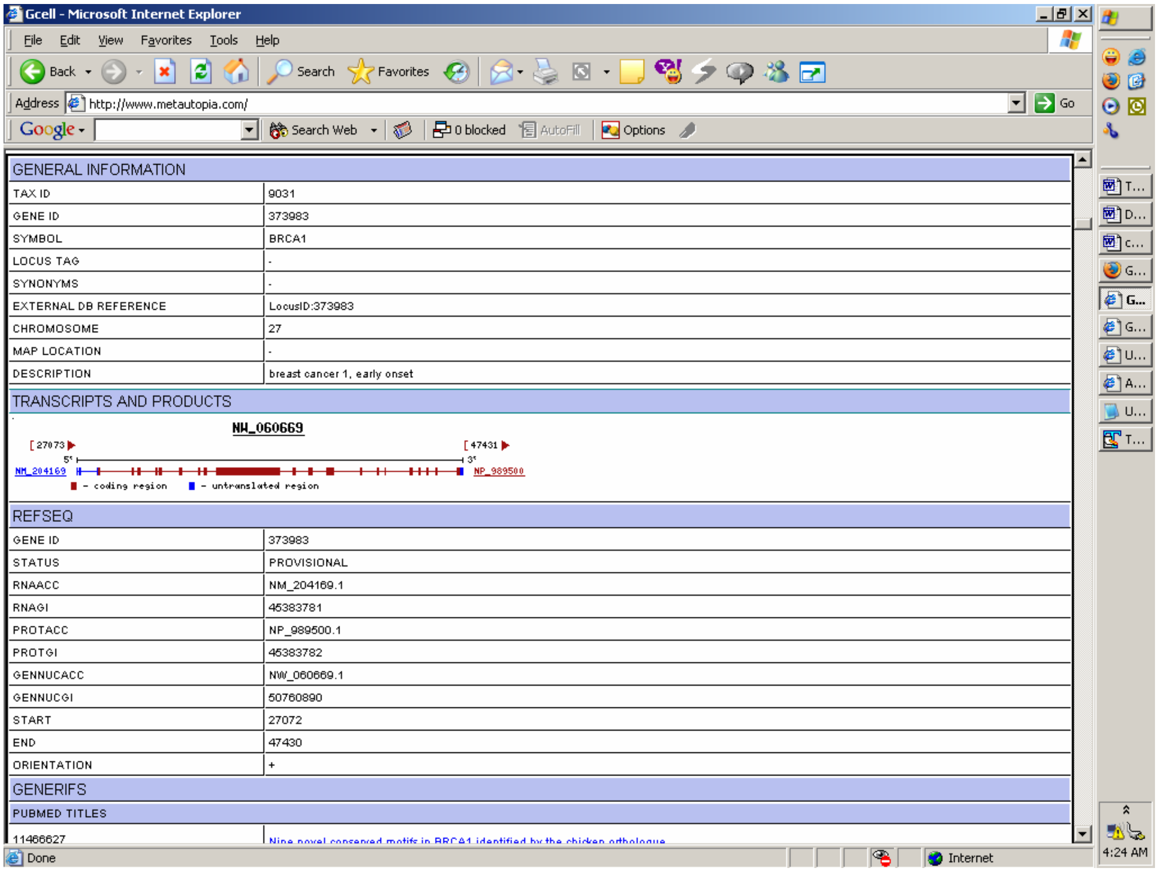
GCell A Sub-cellular Localization Tool

Result:Search among Proteins

BRCA1["Brca1","BRCA1"]

ENTRY	ORGANISM	TISSUE	PROTEIN	CELLULAR COMPONENT	MOLECULAR FUNCTION
RM36_HUMAN	Homo sapiens Human	Testis Adrenal gland	60S ribosomal protein L36, mitochondrial precursor L36mt BRCA1-interacting protein 1	ubiquitin ligase complex extracellular space intracellular nucleus transcription factor complex gamma-tubulin ring complex	damaged DNA binding transcription coactivator activity ubiquitin-protein ligase activity protein binding zinc ion binding tubulin binding transcriptional activator activity
BAR1_RAT	Rattus norvegicus Rat	Carcinoma	BRCA1-associated RING domain protein 1 BARD-1	intracellular nucleus cytoplasm	DNA binding damaged DNA binding RNA binding hydrolase activity, hydrolyzing O-glycosyl compounds protein binding ATP binding
BAR1_MOUSE	Mus musculus Mouse		BRCA1-associated RING domain protein 1 BARD-1	ubiquitin ligase complex condensed chromosome intracellular nucleus cytoplasm	DNA binding damaged DNA binding RNA binding hydrolase activity, hydrolyzing O-glycosyl compounds ubiquitin-protein ligase activity protein binding zinc ion binding
BRAP_MOUSE	Mus musculus Mouse	Embryo Testis Mammary gland Diencephalon	BRCA1-associated protein BRAP2 Impedes mitogenic signal propagation IMP	ubiquitin ligase complex condensed chromosome intracellular nucleus cytoplasm	DNA binding damaged DNA binding RNA binding hydrolase activity, hydrolyzing O-glycosyl compounds ubiquitin-protein ligase activity

Figure A 15. GCell: Result from Protein search for BRCA1



GENERAL INFORMATION

TAX ID	9031
GENE ID	373983
SYMBOL	BRCA1
LOCUS TAG	-
SYNONYMS	-
EXTERNAL DB REFERENCE	LocusID:373983
CHROMOSOME	27
MAP LOCATION	-
DESCRIPTION	breast cancer 1, early onset

TRANSCRIPTS AND PRODUCTS

REFSEQ

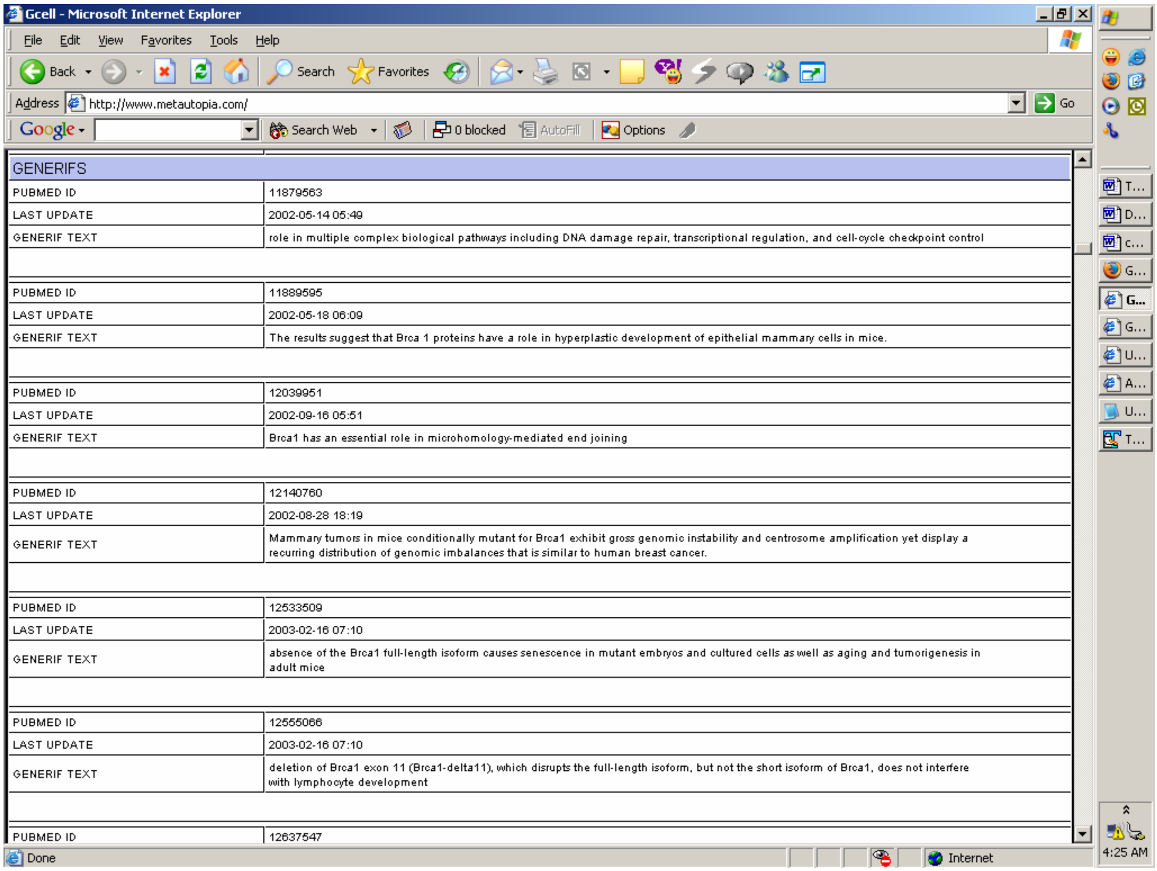
GENE ID	373983
STATUS	PROVISIONAL
RNAACC	NM_204169.1
RNAGI	45383781
PROTACC	NP_989500.1
PROTGI	45383782
GENNUCACC	NW_060669.1
GENNUCGI	50760890
START	27072
END	47430
ORIENTATION	+

GENERIFS

PUBMED TITLES

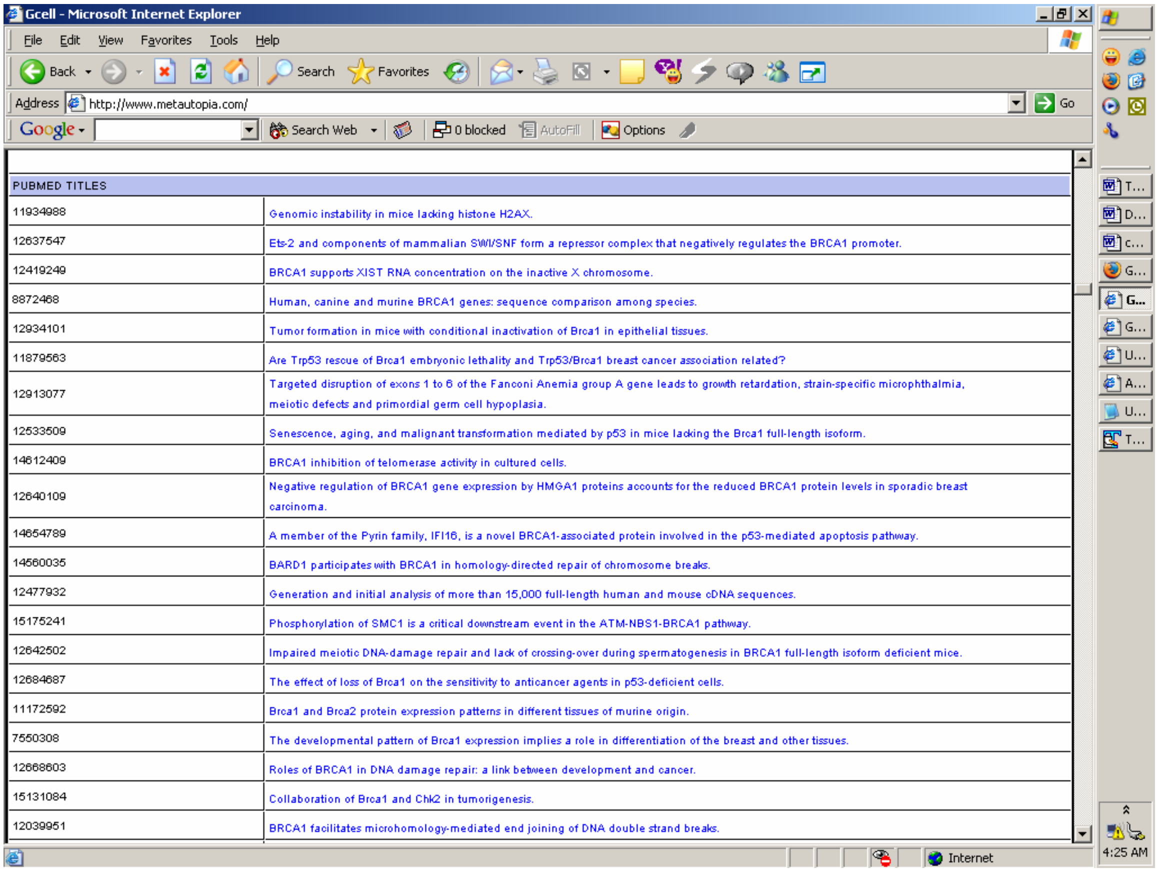
11466627	Nine novel conserved motifs in BRCA1 identified by the chicken orthologue

Figure A 16. GCell: Information from Gene and Refseq databases


Gcell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.metautopia.com/
Go
Google Search Web 0 blocked AutoFill Options
Done
Internet
4:25 AM

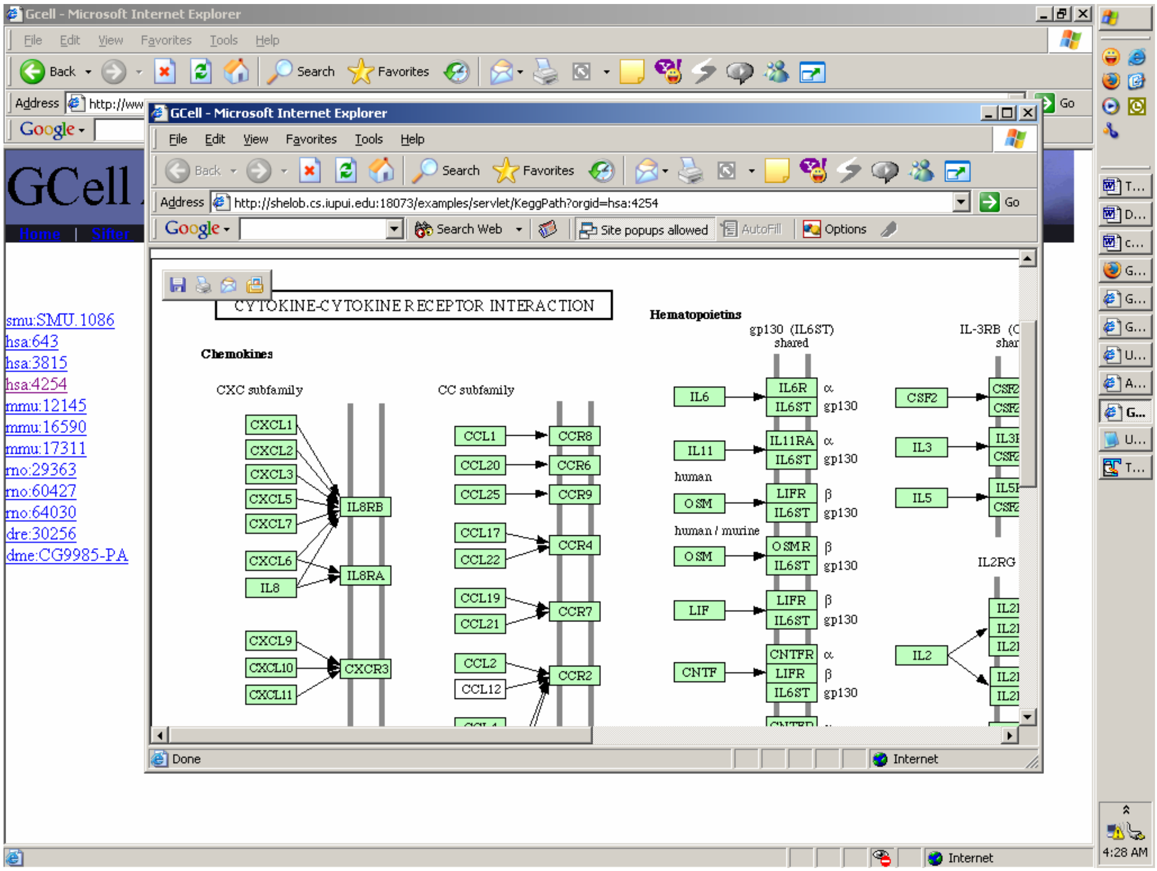
GENERIFS	
PUBMED ID	11879563
LAST UPDATE	2002-05-14 05:49
GENERIF TEXT	role in multiple complex biological pathways including DNA damage repair, transcriptional regulation, and cell-cycle checkpoint control
PUBMED ID	11889595
LAST UPDATE	2002-05-18 06:09
GENERIF TEXT	The results suggest that Brca 1 proteins have a role in hyperplastic development of epithelial mammary cells in mice.
PUBMED ID	12039951
LAST UPDATE	2002-09-16 05:51
GENERIF TEXT	Brca1 has an essential role in microhomology-mediated end joining
PUBMED ID	12140760
LAST UPDATE	2002-08-28 18:19
GENERIF TEXT	Mammary tumors in mice conditionally mutant for Brca1 exhibit gross genomic instability and centrosome amplification yet display a recurring distribution of genomic imbalances that is similar to human breast cancer.
PUBMED ID	12533509
LAST UPDATE	2003-02-16 07:10
GENERIF TEXT	absence of the Brca1 full-length isoform causes senescence in mutant embryos and cultured cells as well as aging and tumorigenesis in adult mice
PUBMED ID	12555066
LAST UPDATE	2003-02-16 07:10
GENERIF TEXT	deletion of Brca1 exon 11 (Brca1-delta11), which disrupts the full-length isoform, but not the short isoform of Brca1, does not interfere with lymphocyte development
PUBMED ID	12637547

Figure A 17. GCell: Information from Gene Reference Into Function(Generifs)


Gcell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.metautopia.com/
Go
Google
Search Web
0 blocked
AutoFill
Options
Internet
4:25 AM

PUBMED TITLES	
11934988	Genomic instability in mice lacking histone H2AX.
12637547	Ets-2 and components of mammalian SWI/SNF form a repressor complex that negatively regulates the BRCA1 promoter.
12419249	BRCA1 supports XIST RNA concentration on the inactive X chromosome.
8872468	Human, canine and murine BRCA1 genes: sequence comparison among species.
12934101	Tumor formation in mice with conditional inactivation of Brca1 in epithelial tissues.
11879563	Are Trp53 rescue of Brca1 embryonic lethality and Trp53/Brca1 breast cancer association related?
12913077	Targeted disruption of exons 1 to 6 of the Fanconi Anemia group A gene leads to growth retardation, strain-specific microphthalmia, meiotic defects and primordial germ cell hypoplasia.
12533509	Senescence, aging, and malignant transformation mediated by p53 in mice lacking the Brca1 full-length isoform.
14612409	BRCA1 inhibition of telomerase activity in cultured cells.
12640109	Negative regulation of BRCA1 gene expression by HMGA1 proteins accounts for the reduced BRCA1 protein levels in sporadic breast carcinoma.
14654789	A member of the Pyrin family, IFI16, is a novel BRCA1-associated protein involved in the p53-mediated apoptosis pathway.
14560035	BARD1 participates with BRCA1 in homology-directed repair of chromosome breaks.
12477932	Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.
15175241	Phosphorylation of SMC1 is a critical downstream event in the ATM-NBS1-BRCA1 pathway.
12642502	Impaired meiotic DNA-damage repair and lack of crossing-over during spermatogenesis in BRCA1 full-length isoform deficient mice.
12684687	The effect of loss of Brca1 on the sensitivity to anticancer agents in p53-deficient cells.
11172592	Brca1 and Brca2 protein expression patterns in different tissues of murine origin.
7550308	The developmental pattern of Brca1 expression implies a role in differentiation of the breast and other tissues.
12668603	Roles of BRCA1 in DNA damage repair: a link between development and cancer.
15131084	Collaboration of Brca1 and Chk2 in tumorigenesis.
12039951	BRCA1 facilitates microhomology-mediated end joining of DNA double strand breaks.

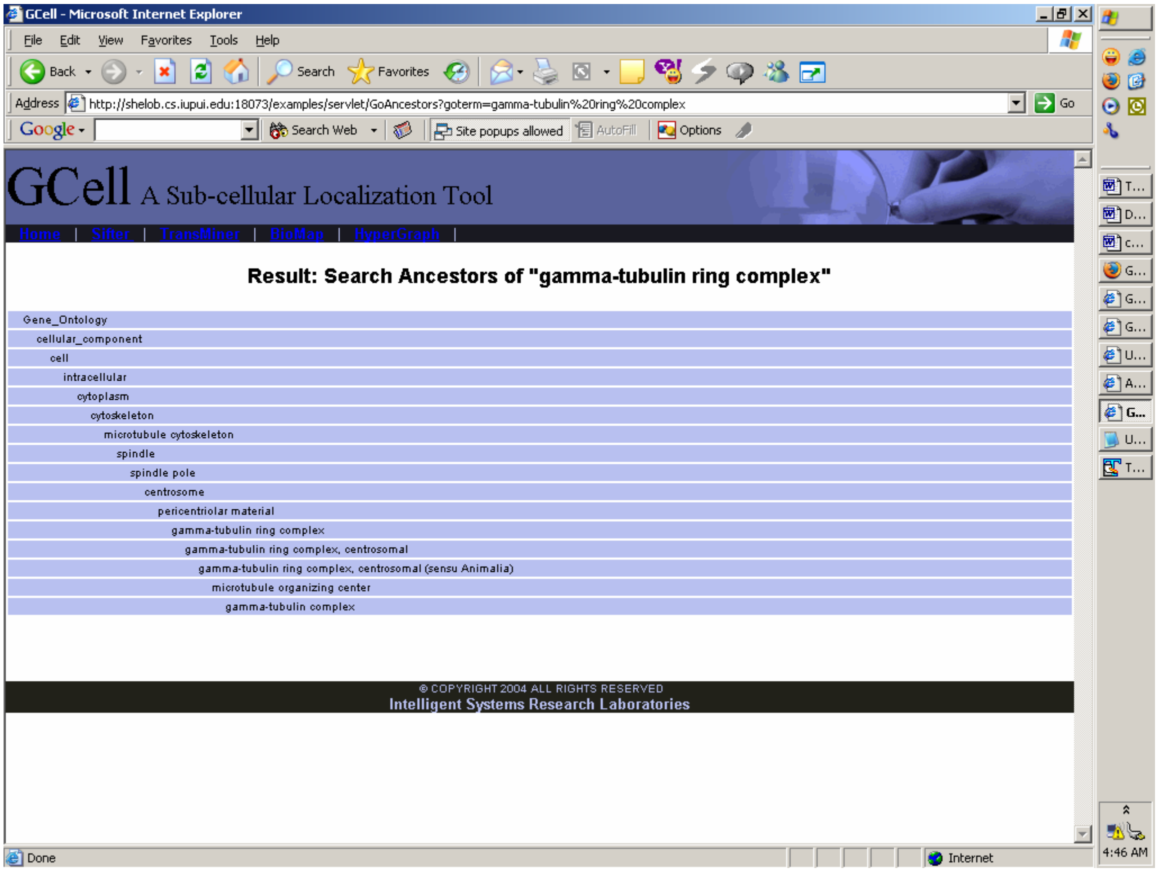
Figure A 18. GCell: Titles as hyperlinks to the Pubmed Database



Address http://shelob.cs.iupui.edu:18073/examples/servlet/KeggPath?orgid=hsa:4254

smu:SMU.1086
hsa:643
hsa:3815
hsa:4254
mmu:12145
mmu:16590
mmu:17311
rno:29363
rno:60427
rno:64030
dre:30256
dme:CG9985-PA

Figure A 19. GCell: Pathway Information from Kegg Database



http://shelob.cs.iupui.edu:18073/examples/servlet/GoAncestors?goterm=gamma-tubulin%20ring%20complex

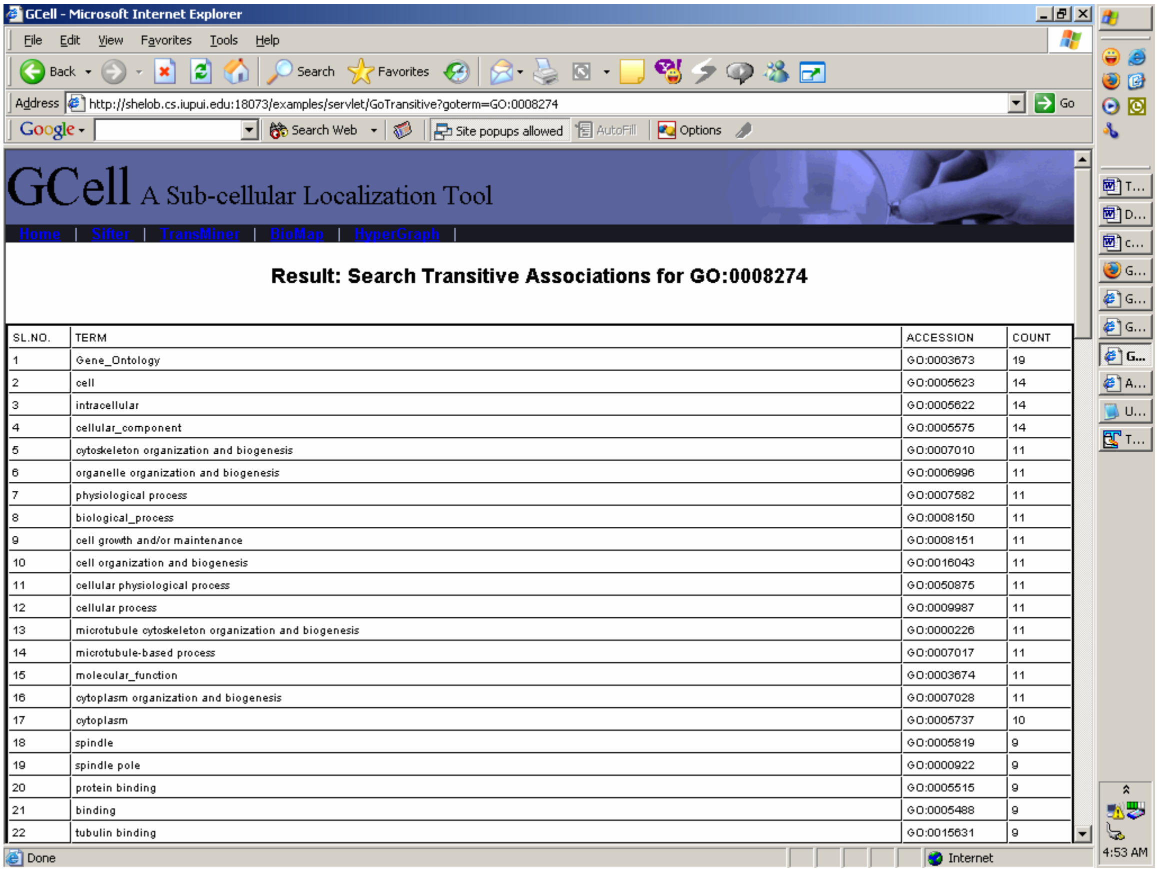
GCell A Sub-cellular Localization Tool

Result: Search Ancestors of "gamma-tubulin ring complex"

```
Gene_Ontology
  cellular_component
    cell
      intracellular
        cytoplasm
          cytoskeleton
            microtubule cytoskeleton
              spindle
                spindle pole
                  centrosome
                    pericentriolar material
                      gamma-tubulin ring complex
                        gamma-tubulin ring complex, centrosomal
                          gamma-tubulin ring complex, centrosomal (sensu Animalia)
                            microtubule organizing center
                              gamma-tubulin complex
```

© COPYRIGHT 2004 ALL RIGHTS RESERVED
Intelligent Systems Research Laboratories

Figure A 20. GCell: Ancestors to a particular term extracted from Gene Ontology database

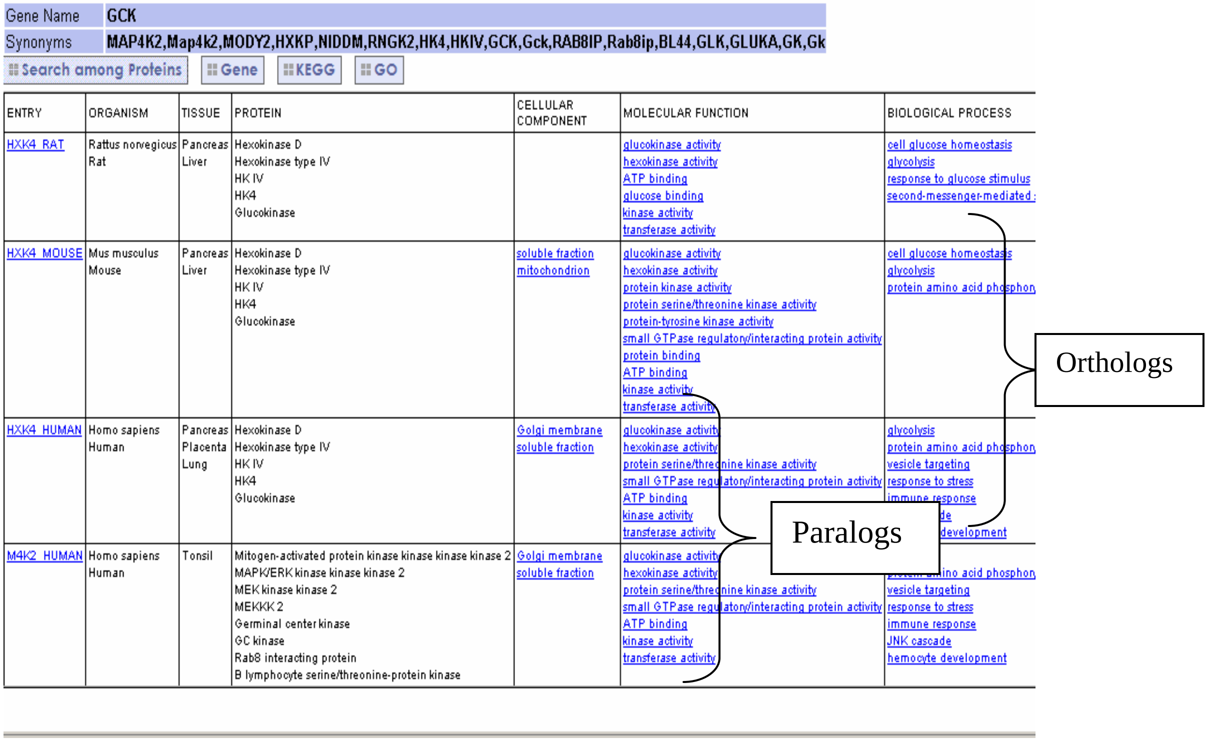


GCell - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://shelob.cs.iupui.edu:18073/examples/servlet/GoTransitive?goterm=GO:0008274
Go
Google Search Web Site popups allowed AutoFill Options
GCell A Sub-cellular Localization Tool
Home | Sifter | TransMiner | BioMap | HyperGraph |
Done
Internet
4:53 AM

Result: Search Transitive Associations for GO:0008274

SL.NO.	TERM	ACCESSION	COUNT
1	Gene_Ontology	GO:0003673	19
2	cell	GO:0005623	14
3	intracellular	GO:0005622	14
4	cellular_component	GO:0005575	14
5	cytoskeleton organization and biogenesis	GO:0007010	11
6	organelle organization and biogenesis	GO:0006996	11
7	physiological process	GO:0007582	11
8	biological_process	GO:0008150	11
9	cell growth and/or maintenance	GO:0008151	11
10	cell organization and biogenesis	GO:0016043	11
11	cellular physiological process	GO:0050875	11
12	cellular process	GO:0009987	11
13	microtubule cytoskeleton organization and biogenesis	GO:0000226	11
14	microtubule-based process	GO:0007017	11
15	molecular_function	GO:0003674	11
16	cytoplasm organization and biogenesis	GO:0007028	11
17	cytoplasm	GO:0005737	10
18	spindle	GO:0005819	9
19	spindle pole	GO:0000922	9
20	protein binding	GO:0005515	9
21	binding	GO:0005488	9
22	tubulin binding	GO:0015631	9

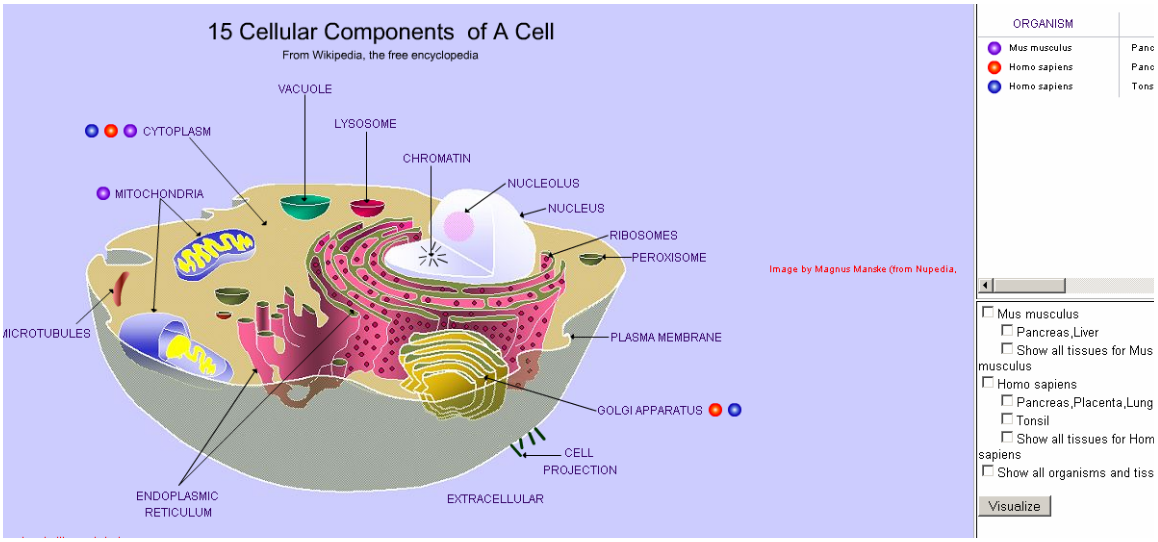
Figure A 21. GCell: Transitive Association for a gene product



Gene Name	GCK
Synonyms	MAP4K2,Map4k2,MODY2,HXKP,NIDDM,RNGK2,HK4,HKIV,GCK,Gck,RAB8IP,Rab8ip,BL44,GLK,GLUKA,GK,Gk

ENTRY	ORGANISM	TISSUE	PROTEIN	CELLULAR COMPONENT	MOLECULAR FUNCTION	BIOLOGICAL PROCESS
HXK4_RAT	Rattus norvegicus Rat	Pancreas Liver	Hexokinase D Hexokinase type IV HK IV HK4 Glucokinase		glucokinase activity hexokinase activity ATP binding glucose binding kinase activity transferase activity	cell glucose homeostasis glycolysis response to glucose stimulus second-messenger-mediated :
HXK4_MOUSE	Mus musculus Mouse	Pancreas Liver	Hexokinase D Hexokinase type IV HK IV HK4 Glucokinase	soluble fraction mitochondrion	glucokinase activity hexokinase activity protein kinase activity protein serine/threonine kinase activity protein-tyrosine kinase activity small GTPase regulatory/interacting protein activity protein binding ATP binding kinase activity transferase activity	cell glucose homeostasis glycolysis protein amino acid phosphon
HXK4_HUMAN	Homo sapiens Human	Pancreas Placenta Lung	Hexokinase D Hexokinase type IV HK IV HK4 Glucokinase	Golgi membrane soluble fraction	glucokinase activity hexokinase activity protein serine/threonine kinase activity small GTPase regulatory/interacting protein activity ATP binding kinase activity transferase activity	glycolysis protein amino acid phosphon vesicle targeting response to stress immune response de development
M4K2_HUMAN	Homo sapiens Human	Tonsil	Mitogen-activated protein kinase kinase kinase kinase 2 MAPK/ERK kinase kinase kinase 2 MEK kinase kinase 2 MEKKK 2 Germinal center kinase GC kinase Rab8 interacting protein B lymphocyte serine/threonine-protein kinase	Golgi membrane soluble fraction	glucokinase activity hexokinase activity protein serine/threonine kinase activity small GTPase regulatory/interacting protein activity ATP binding kinase activity transferase activity	ino acid phosphon vesicle targeting response to stress immune response JNK cascade hemocyte development

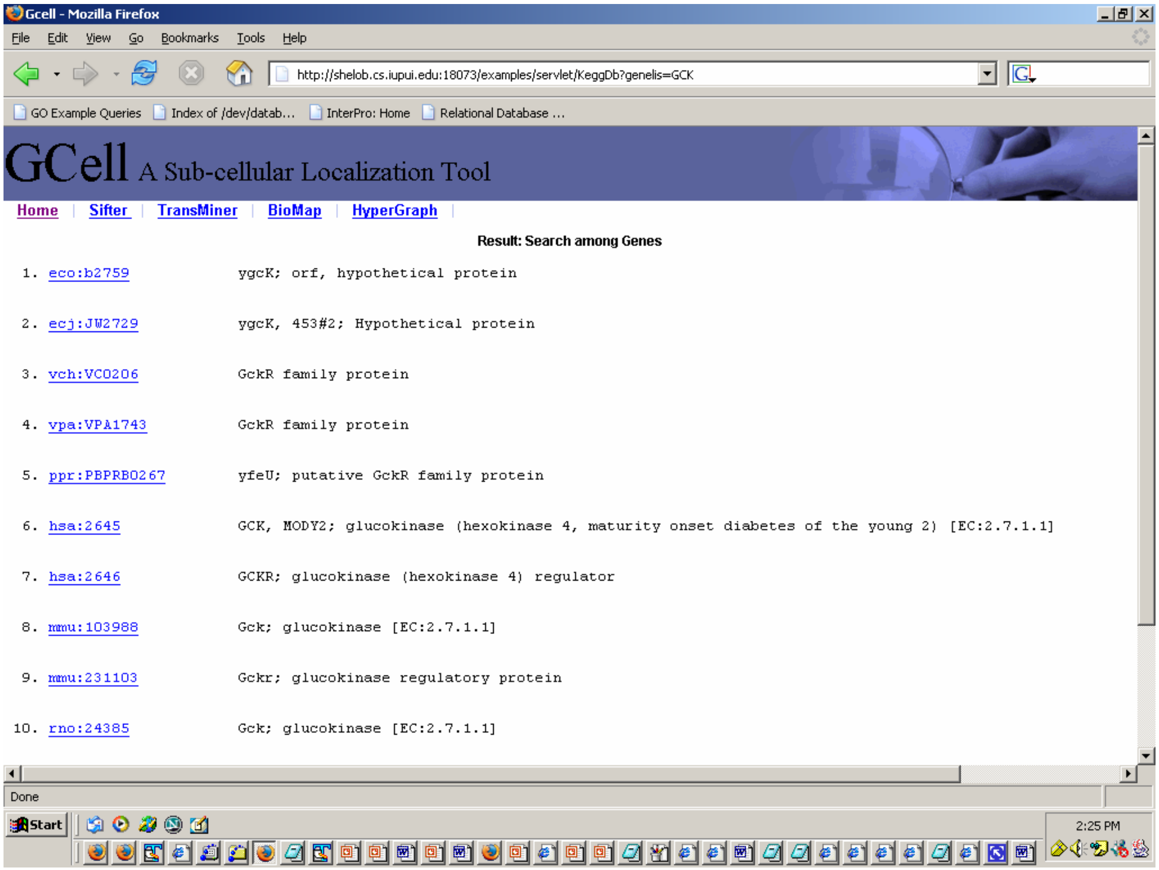
Figure A 22. GCell: Orthologs and Paralogs



ORGANISM	
● Mus musculus	Panc
● Homo sapiens	Panc
● Homo sapiens	Tons

☐ Mus musculus
☐ Pancreas,Liver
☐ Show all tissues for Mus musculus
☐ Homo sapiens
☐ Pancreas,Placenta,Lung
☐ Tonsil
☐ Show all tissues for Hom sapiens
☐ Show all organisms and tiss

Figure A23 Sub-cellular localization


Gcell - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://shelob.cs.iupui.edu:18073/examples/servlet/KeggDb?genelis=GCK
GO Example Queries
Index of /dev/datab...
InterPro: Home
Relational Database ...
GCell A Sub-cellular Localization Tool
Home | Sifter | TransMiner | BioMap | HyperGraph |
Done
Start
2:25 PM

Result: Search among Genes

```
 1. eco:b2759           ygcK; orf, hypothetical protein

 2. ecj:JW2729          ygcK, 453#2; Hypothetical protein

 3. vch:VC0206          GckR family protein

 4. vpa:VPA1743         GckR family protein

 5. ppr:PBPRB0267       yfeU; putative GckR family protein

 6. hsa:2645            GCK, MODY2; glucokinase (hexokinase 4, maturity onset diabetes of the young 2) [EC:2.7.1.1]

 7. hsa:2646            GCKR; glucokinase (hexokinase 4) regulator

 8. mmu:103988          Gck; glucokinase [EC:2.7.1.1]

 9. mmu:231103          Gckr; glucokinase regulatory protein

10. rno:24385           Gck; glucokinase [EC:2.7.1.1]
```

Figure A 24 All the available pathways linked to Kegg Database

B. Text Mining And Biological Information Management System

Scientists today are burdened with the growth of data in biological field and subsequent information overload. The rapidly growing number and size of the databases makes it very difficult for biological researchers to obtain the customized information in an effective manner. As most information (over 80%) is stored as text, text mining has a high commercial potential value. Text mining is an interdisciplinary field, which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT), text mining refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Application of text mining in bioinformatics yields automatic extraction of experimental results from a large corpus of text and then processed computationally. It aims to extract useful knowledge from unstructured or semi-structured text. Although, there are several tools, which are available today for accessing this high volume of information in the biology domain, there is a necessity for an existence of a single tool for accessing multiple data sources and formats. Locating highly relevant information is of utmost importance. Palakal *et al.* (2002) defined the overall problem of information management as a two level mapping. The first level mapping is from data to information and the second level mapping is from information to knowledge. The process of retrieving relevant documents from the data sources and presenting to the user according to the interest can be obtained by an active personalized information management system. The system makes use of knowledge regarding the documents and users requirement in a continuously changing document stream. The information management process of mapping biological text data into useful relevant information is obtained by the implementation of BioSifter (Palakal et al 2002). The user is presented with relevant information in a seamless manner.

Three important entities comprise the filtering environment – document source, filter and the user. The input to the system consists of the incoming documents from various data sources. The storage of the documents before filtering is done by the Document Acquisition and Management (DAM) component. The filter comprises of a sub-module that represents the documents (determines input space), classifier (maps the resulting vector representation to the classification space) and profile manager (maps the categories to the relevance values). The filtering system presents these in a sorted fashion to the user. In order to perform its sorting tasks efficiently, the filtering system makes use of several modules, and operates in a feedback configuration with the user. Presentation and Access Management (PAM) is present at the output end of the filtering model as a user interface. The documents are presented to the user by PAM, which also provides for the controls for relevance feedback for documents.

The biological information management system comprises of three modules viz. document representation module, classification module and user profile learning module. The document representation module converts incoming documents into structures that are representative of original documents in a form that can be easily parsed by the component in the next level. The vector-space model accepts as input a set of textual documents and produces as output a set of vectors (representing individual documents) whose location in the document space is fixed. The technique used for generating weights for the vectors is to generate frequencies for individual terms appearing in documents and use those frequencies as weights. Not all the words in the English language can be used for weight generation. Certain common terms are eliminated from the list by using a negation set (known as “stop word list”). After the negation set is applied to the documents, they are converted to weight vectors. For a document set, a table is generated that contains the total frequencies of all unique terms. Another table is generated that contains the frequencies of terms occurring in individual documents for each document in the set.

After the document vectors are formed, each document vector is assigned to a particular class by the classification module. This module consists of two processing stages: an unsupervised learning stage and a vector classification stage. During the learning stage, an initial cluster hypothesis based on Maximin Distance clustering algorithm (Tou & Gonzalez 1974) is generated for the given set of document vectors. Centroids of classes are generated over the document vector space. During the second stage of online classification, each incoming document vector is classified into a particular class whose centroid has the minimum distance to the document vector using the learned cluster centroids from the previous stage. The distance between two vectors is calculated by the cosine similarity measure (Salton 1989). The distance is computed as one minus the similarity.

The user-profile learning module consists of a learning agent that interacts directly with the user. The users preference for the different classes of information is determined and incoming documents are prioritized accordingly. The learning agent maintains and updates a simplified model of the user based on a reinforcement-learning algorithm (Narendra & Thathachar 1989). On the basis of relevance feedback, the learning agent updates so as to improve its performance. The user provides relevance feedback as discrete value (0: not relevant, 0.5: relevant, 1: very relevant) for documents. Reinforcement learning approach is suitable in the context of biological information filtering due to its real-time learn-while-perform aspect. The learning agent rank-orders the incoming documents according to the users interests. It does this by maintaining two vectors – relevance probability vector and action probability vector, of dimensions equal to the number of classes. Hence, after a long learning time, ranking of the documents is performed according to user relevance values for the corresponding classes.

The Graphical User Interface (GUI) gathers all recent documents and displays them in a ranked list based on the ranking provided by the user profile learning module. It allows users to select individual documents and read them and also collects user feedback and passes the feedback to the user-profile learning module. A set of 25 titles ranked in order is shown on the web page. These titles appear as links to the Pubmed database in order to read the complete abstracts of the documents. The users can create/delete a problem domain, view/modify/delete the thesaurus terms, view titles and also view learning profile graph. After the system has learned for sometime, the users may use the 'autorun' feature to automatically feed the relevance feedback for all the documents. The control flow of the entire system is coordinated by a filter daemon, which makes calls to the individual modules. The daemon is activated when a user submits a request and invokes the classification module. The classifier reads the weight vectors and outputs their classes into another file. The engine subsequently activates the profile-learning module that ranks the documents according to their classes and the current user profile. The control then passes to the GUI, which presents the ranked documents to the user as described earlier. The user relevance feedback is stored in the database. The engine then invokes the user-profile learning module again to update the user model based on the relevance feedbacks. This constitutes one complete cycle of the engine. The user can view a new set of documents and initiate a new iteration.

Deep Mining Of Biological Objects

Knowledge from large volume of scientific literature may be extracted in the form of associations among biological objects such as genes, proteins, processes, diseases and chemicals. The documents required for this are downloaded from MEDLINE. Association between different biological objects may be found by deep mining of biological objects. Associations that are not explicitly found in Medline can be found using a transitive association discovery method [NarayanSwamy et al 2004]. The current implementation of deep mining of associations between biological objects assumes that Medline has both the set of complementary literatures. It finds many ranked associations at a time by using transitive closure principle. The documents required are downloaded from Medline using an HTML wrapper tool, for a set of terms specified by a user. Documents are represented as vectors using the tf-idf model (Salton 1983). This vector space model attempts to compute the importance of terms on the basis of term frequencies within a document and within an entire document collection. Thus each document vector consists of tf.idf weight of the query terms.

After the vector representation of all documents is computed, the association between the two objects is computed as association value, which is used as a measure of the strength of the relationship between any two terms. An association matrix is calculated that represents the strength between different biological objects. For any pair of object terms co-occurring in even a single document, the association strength is non-zero and positive. The association matrix is a symmetric matrix. The non-zero and non-diagonal values from the matrix are used for creating the undirected association graph. The newly discovered potential transitive associations must be checked to see if those associations are direct (found in any Medline document) or transitive. If objects co-occur, documents will be retrieved, and hence it could be stated that the association must be direct because of the principle of co-occurrence. The remaining potential transitive associations with non-zero strength are of implicit or transitive nature. Hypothesis can be generated on the basis of these transitive associations.

The discovered associations need to be evaluated which is done by experts in biology. Visualization of these associations is presented in the form of an undirected graph. Each object is represented as a node and the relationship as an edge between objects. Strengths between the objects are displayed by the thickness of lines between the objects. Transitive associations are represented by *dashed* edges.

Not all associations are interesting to all the users. The 'interesting' associations are subjective. Different users may like to view the associations according to different criteria. Hence, in order to cater to this requirement, different levels of association strength can be selected by the users. Associations may be selected on the basis of upper and lower association strength values, depending on whether the user is interested high strength, moderate strength or less strength associations. Neighbor identification capability for any object was implemented using a tree scan algorithm as implemented by Mrowka. The newly discovered transitive association enables the user to investigate further. However, strength can be calculated for these associations as in the case of direct associations where co-occurrence can be measured. The transitive strength is the sum of weight of all objects B that co-occur with nodes A and C of a transitive association. This is based on the idea that if there is a strong link in the form of A-B-C then the possibility of the AC association becoming true is increased. Several experiments were done in order to demonstrate the use of TransMiner in hypothesis generation. However, we will be undertaking some new experiments in combination with the results from Biosifter and GCell system to demonstrate its use.

C. Vita

Rakesh Dhaval
Email: rdhaval@gmail.com

EDUCATION

2002-2005	Indiana University, Indianapolis	MS Bioinformatics
1998-2000	Birla Institute of Technology, India	MS Information Sc.
1994-1998	Birla Institute of Technology, India	BS Pharmacy

RELEVANT COURSES

Distributed Databases, Database Management, Data Mining, Statistical Computing, Artificial Intelligence, Networking, Information Processing, Management Information Systems, Genetics and Molecular Biology, Biotechnology, Biochemistry, Pharmacology, Microbiology, Pharmaceutical Chemistry, Medicinal Chemistry

SKILL SET

Languages & Scripts: Java, Python, C
Web Technologies: JSP/Servlets, Struts, CGI, XML (SAX/DOM), HTML, DHTML, Javascript
Web Platforms: Tomcat, Apache, J2EE
OS Platforms: Linux, Sun Solaris, MS Win NT/9X
Databases: Oracle 8, MySQL, MS Access
Project Planning/Modeling: MS Visio (UML)
Statistical and Other Packages: SAS, Matlab

Knowledge of HIPAA, GMP/GLP and biological databases like PubMed, LocusLink, Gene Ontology, Unigene, OMIM, GenBank, SwissProt, UMLS, KEGG, Uniprot, BIND, DIP, MESH

PRINCIPAL FIELDS OF INTEREST

Information Filtering and Management, Data Mining, Machine Learning, Database Management

WORK EXPERIENCE

Research Assistant @ Intelligent Systems Laboratory, Indiana University Purdue University at Indianapolis

-Worked on projects funded by Eli Lilly and Company, the Digital Library Initiative and NSF. Duties included design, development and validation of text mining applications in the field of bioinformatics research.

Associate Consultant @ Zensar Technologies Ltd, India

-Offshore client support for Fujitsu, which involved system analysis and design and development in the Wireless Technologies group

INTERNSHIPS

Exelixis Inc, South San Francisco

-Worked on the development teams for systems such as the Distributed Annotation System (DAS), and the Cloning and Protein Information Management System (CPIMS)

National Aerospace Laboratories, Bangalore, India

-Web based Information System and Client Server application for reservations, room service, house keeping and inventory management of Guest Houses

CDAC, Ranchi, India

-Online Coal Sale Billing System

Birla Institute of Technology, Ranchi, India

-Literature survey on Immuno-Modulatory Agents

PART TIME EMPLOYMENT

Tutor at Mathematics Assistance Center, University College, Indianapolis
Mentor for Undergraduate Computer Courses, University College, Indianapolis
Instructor at Institute for Personality Development and Communication Studies, India

COURSE PROJECTS/PAPERS

Paper: Database Issues in Bioinformatics

Projects: Analysis of Microarray gene expression data using Neural Networks and SVM
PCA & SVD of Microarray gene expression data
Clustering of Microarray gene expression data
Information Management System for E. coli genome

AWARDS & HONORS

Poster Presentation at First Annual Bioinformatics Conference of Indiana: GCell
Brainbench Certification Master Level in C
University College Scholarship Recipient
Paper presentation at IEEE National Paper Presentation Contest

PROFESSIONAL AFFILIATION

ISCB