GXplain

AN INTELLIGENT SYSTEM THAT OFFERS STUDY ADVICE ABOUT GENES

Ki Kim, Rashmi Raj, Ravi Tiruvury, Sine Zambach
Course Project for CS270/BMI210
Fall 2005, Stanford University
{kykim, rashmisu, raviteja, sinez}@stanford.edu

ABSTRACT:

Gene data is humongous and scattered over several data repositories. In addition, it is not comprehensive as it is recorded in varied representations. There are also very few tools that represent this data in a user friendly format. Due to these reasons, it becomes difficult for biologists to visualize the data and interpret its meaning and infer intelligent reasoning about genes. The National Center for Biotechnology Information – Gene Expression Omnibus (NCBI – GEO) provides the most comprehensive repository of microarray gene expression data. The Gene Ontology (GO) represents a controlled vocabulary that describes genes and gene products. Currently, there exist no systems that leverage the GEO and GO data collectively, to reason about genes. We hence developed GXplain, which an intelligent reasoning system that offers study advice about genes. GXplain takes a Gene or GO term as input, analyses the summary of microarray experiment results and gene product information – to come up with answers to questions such as: "Is a gene interesting study?" GXplain offers this kind of "study advice" by *heuristic classification* of gene characteristics using if-then-else rules. Furthermore, GXplain's graphical presentation of the values of microarray experiments efficiently summarizes the level of gene expression in different microarray experiments. Given the huge data, short time-span and the complex biological relationships amongst genes, we restricted our Knowledge Base to genes related to Insulin domain. GXplain is however extensible – it can certainly widen its functional scope given more computational resources. Nevertheless, the complex nature of gene poses several challenges and there is immense scope for the further development of GXplain.

1. INTRODUCTION AND PROBLEM STATEMENT

Biologists have traditionally found it very difficult to study gene and gene product information because of the high volume and complexity of gene data. This problem is compounded by factors like inconsistency of gene data and the lack of standards for representing gene data. There is a compelling need to obtain a more comprehensive representation of data – pertaining to both microarray experiment data and GO term information from the Gene Ontology.

1.1. Background Information and Motivation

Why study microarray data? According to NCBI, GenBank is a comprehensive database that contains publicly available DNA sequences for more than 165,000 named organisms and the INSDC (International Nucleotide Sequence Database Collaboration) announced the DNA sequence database has exceeded 100 gigabases (Benson et al., 2005). Nevertheless, the biological functions of most of these genes are in mystery. One possible way to solve the mystery of these genes is through repeated measurements of their RNA transcripts. For example, if we know that a certain gene is only expressed in a specific organ under a particular condition, then we might be able to infer the biological function of that gene. People from

functional genomics study gene function using microarray experiments. Microarrays are artificially constructed grids of DNA, such that each element of the grid probes for a specific RNA sequence. The expression value is reflecting how many copies of each gene are present in a certain sample. These expression values are standardized or normalized for further analysis.

The exponential increase of the number of microarray experiments necessitates the need to represent microarray experiment results in a unified way, to be able to gather and represent useful information. Often, researchers have to go back and forth between NCBI GEO and Gene Ontology (GO) datasets to study characteristics of genes. Though the Gene Expression Omnibus(GEO) database has information about microarrays experiments, there is no tool to extract and visualize the data in a comprehensive and a user-friendly format. More importantly, there is no application to intelligently reason over microarray gene data and its association with GO database. GXplain attempts to provide solutions to these problems.

1.2. Data

GXplain uses selected data from the GEO and the Gene Ontology. The GEO database contains data for genes, gene expressions and microarray experiments (Edgar et al, 2002). Each gene has unique ID called Locuslink, and each spot in the RNA sequence of a gene has a corresponding GSM number. Each GSM has a unique microarray experiment measurement and this value mirrors the expression level of a certain gene in a specific biological sample. The GO database contains controlled vocabularies for gene products, and its three organizing principles of GO are biological process, cellular component, and molecular function. The Uniprot ID uniquely recognizes each protein in the Gene Ontology. GXplain connects GEO and GO databases by mapping Locuslink to Uniprot ID.

1.3. Questions to Be Answered

GXplain answers two main questions: “Is This Gene Interesting One to Study?” and “Which Gene(s) is Interesting to Study a Specific GO term?” GXplain perform preliminary reasoning over user input data—Locuslink or Gene Name or GO term. It then heuristically classifies data to answer questions. These answers to these questions are not only useful to experienced biologists but also to a person with little background in gene data.

2. MATERIALS AND METHODS

An ontology is a specification of a conceptualization. The conceptualization is the way we think about a domain. Problem Solving Methods are abstract algorithms for achieving solutions to stereotypical tasks.

2.1 Ontology Description

We conceptualize the GXplain model as a system that uses a combination of the Gene Ontology and the GEO (Gene Expression Omnibus) concepts. The Gene Ontology describes three organizing principles, viz. Cellular Component, Biological Process and Molecular Function, to describe a gene product (The Gene Ontology Consortium, 2000). The GEO is a repository of microarray expression data.

The aim of GXplain is to leverage the benefit of the Gene Ontology by wrapping it around the concepts of the GEO. Figure 1 illustrates the concepts and relationships of the GXplain model. The concepts used are: GO terms (Biological Process, Molecular Function, Cellular Component), Gene, Locuslink, Protein, GSM, GSM Title and Rank.

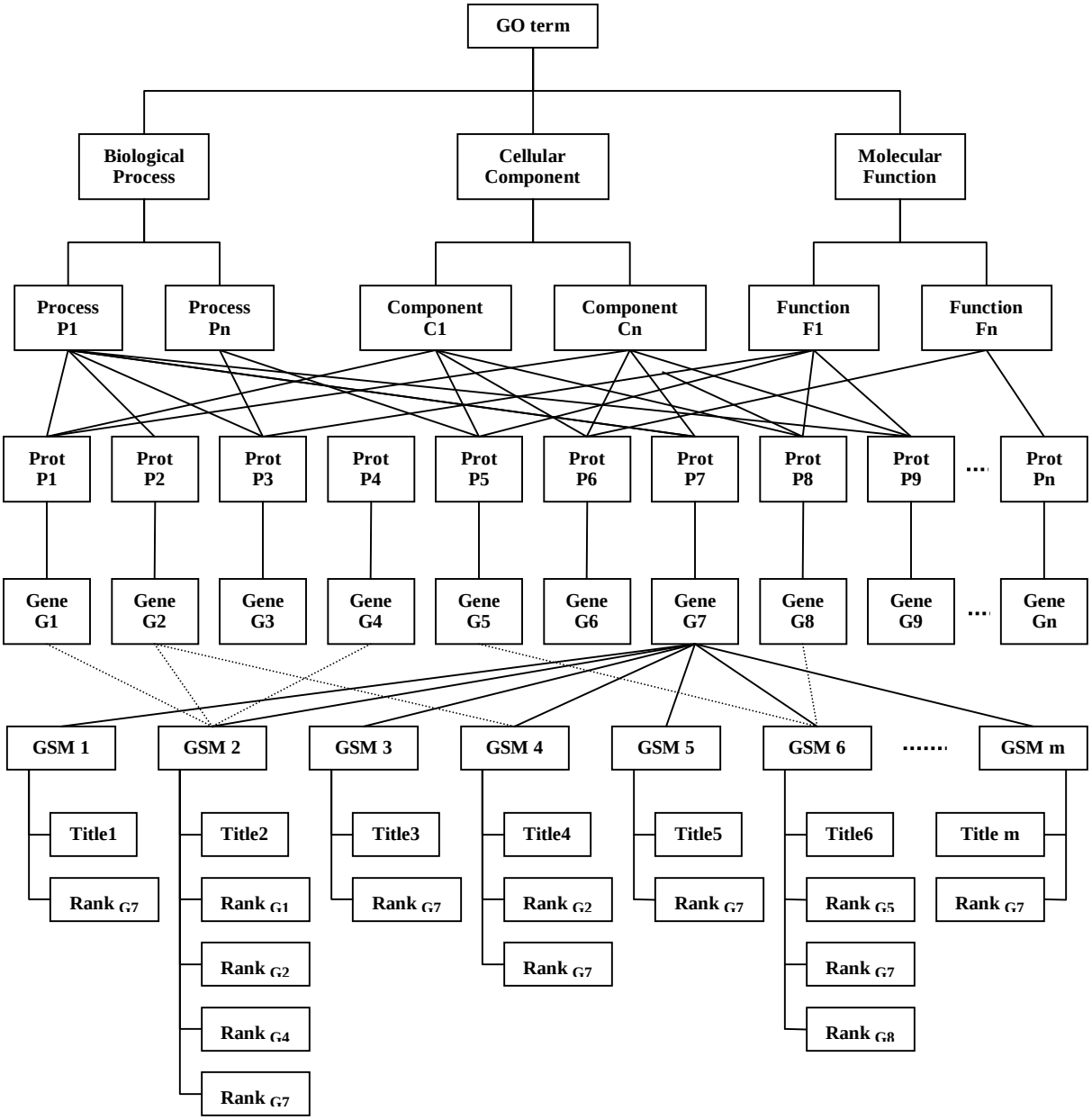


Fig. 1: GXplain Ontology

Concept Descriptions

Molecular Function: Describes activities at the molecular level.

Biological Process: A series of events accomplished by one or more ordered assemblies of molecular functions.

Cellular Component: It is just a component of a cell, with the proviso that it is part of some larger object, which may be an anatomical structure, or a gene product group.

Gene: A fundamental physical and functional unit of heredity.

Locuslink: A unique gene-identifier in the GEO Database.

Protein: A large complex molecule made up of one or more chains of amino acids.

GSM: A Sample record describes the conditions under which an individual sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it.

GSM Title: The name of the tissue/organ/part that the sample is associated with.

Rank: Rank normalized value which describes the level of expression of a gene in a particular sample.

Relationships and Constraints

- One gene may be associated with more than one GO terms (BP, MF or CC).
- One gene may be associated with multiple GSMs.
- One GSM may be associated with multiple genes.
- One GO term may be associated with several genes.
- One GSM may be associated with only one GSM Title.
- One GSM may be associated with different Rank Values, depending upon the gene involved.
- One gene may be associated only with one protein.
- Each gene is uniquely identified by a Locuslink.
- For a particular GSM, one gene is associated with only one rank value.

The GXplain ontology follows the "bottom-up" approach. We have observed the data and identified the relationships and constraints that exist within our system. After defining our questions, we were able to extract relevant concepts and represent them in the GXplain Ontology (see Fig. 1). It is relatively "light", as we have only described the concepts we used in our decision making and inference processes.

Since GXplain extensively depends on gene-specific and GO term specific information, we modeled a "domain-specific" ontology that is closely tied with the tasks that GXplain performs. It may not be applicable for usage in non-biological domains, but can certainly be extended to include more concepts and make more inferences about the gene behavior.

After finalizing our concepts, the next challenge was to come up with a representational structure for the GXplain Ontology. We first identified the crucial features required:

- The representation should be able to represent data of the order of gigabytes of gene data.
- The representation should allow computationally tractable relations.
- It should be in a format that is easily "query-able".
- It should allow cross-linking of data and concepts.
- It should aid in generating a fast and optimized outcome.

After a careful analysis of various representations, we narrowed down to representing data in Tabular format in a Relational Data Base Management System.

2.2 Problem Solving Method Development

In the GXplain project, we have tried to simulate the "Hypothetico-deductive approach". Based on the available data, we made some assumptions, and deduced intelligent reasoning. Thus, the results are consistent with the assumptions, if not with the real world. They have been listed below:

- The number of microarray experiments a gene is involved indicates how well it has been studied.
- The number of GO terms a gene is associated with is indicative of how "functional" the gene is.
- A Locuslink is assigned to a gene in the order in which it is discovered.
- If the Rank Value of a Gene in a sample lies between 0.8 to 1.0, then it is "Very Highly Expressed".
- If the Rank Value of a Gene in a sample lies between 0.6 to 0.8, then it is "Highly Expressed".
- If the Rank Value of a Gene in a sample lies between 0.4 to 0.6, then it is "Medium Expressed".
- If the Rank Value of a Gene in a sample lies between 0.2 to 0.4, then it is "Lowly Expressed".
- If the Rank Value of a Gene in a sample lies between 0.0 to 0.2, then it is "Very Lowly Expressed".

In developing our Problem Solving Method, we have evaluated several methods including Propose-and-Revise and Heuristic Classification. After an extensive analysis, we decided that Heuristic Classification would be the most appropriate PSM for GXplain. Our initial set of rules performs "Feature Abstraction" based on the above assumptions. A more complex rule set performs a "heuristic match" upon the abstracted features to answer the question: "Whether a gene is interesting to study". Once we get an answer to this question, we further refine our solution to include gene expression data and answer questions like: "In which samples is the gene highly expressed" and "Which samples is the gene very specific to".

We describe our Heuristic Classification as follows:

Input Data: PSM takes in three types of inputs – Locuslink, Gene Name or GO term.

Output Data: Gene Features such as "Well-Studied" or "Not Well-Studied", "More Functionality" or "Less Functionality", and "Old" or "New" gene. These features become input data to more rules, which GXplain uses to infer if a gene is interesting to study or not.

Constraints that establish relationships between Input and Output Data:

- Total Number of GSMs with respect to a Locuslink should be available
- Each Locuslink must have at least one GO term associated with it.
- Rank Normalized Value of each Gene with respect to each GSM must be available.
- Locuslink must be a valid positive number.

In Fig. 2, it can be observed that Rules 1, 2, 3 and 4 are used to abstract "features" of a gene – its level of expressiveness, whether or not it is well-studied and whether or not it has more functionality. Once these basic features are abstracted, Rules 5, 6, 7, 8 and 9 further refine them, and deduce if the gene is interesting to study or not. These rules do a "heuristic match" between the outcome of the Feature Abstraction Rules (1, 2, 3, & 4) and arrive at possible solutions. They don't necessarily guarantee a perfect outcome, but definitely suggest a list of possible outcomes. Rules 10 and 11 further refine the results produced by the previous rules, to make more complex decisions such as which gene is more specific to which GSM sample.

In a nutshell, Rules 1, 2, 3, 4 can be categorized as Feature Abstraction Rules. Rules 5 through 9 perform Heuristic Matching and Solution Refinement. Rules 10 and 11 further refine the results of the previous rules (Please refer to the Appendix to view the Rule Set). This is the process that works best with our system – and this is what the Heuristic Classification PSM advocates.

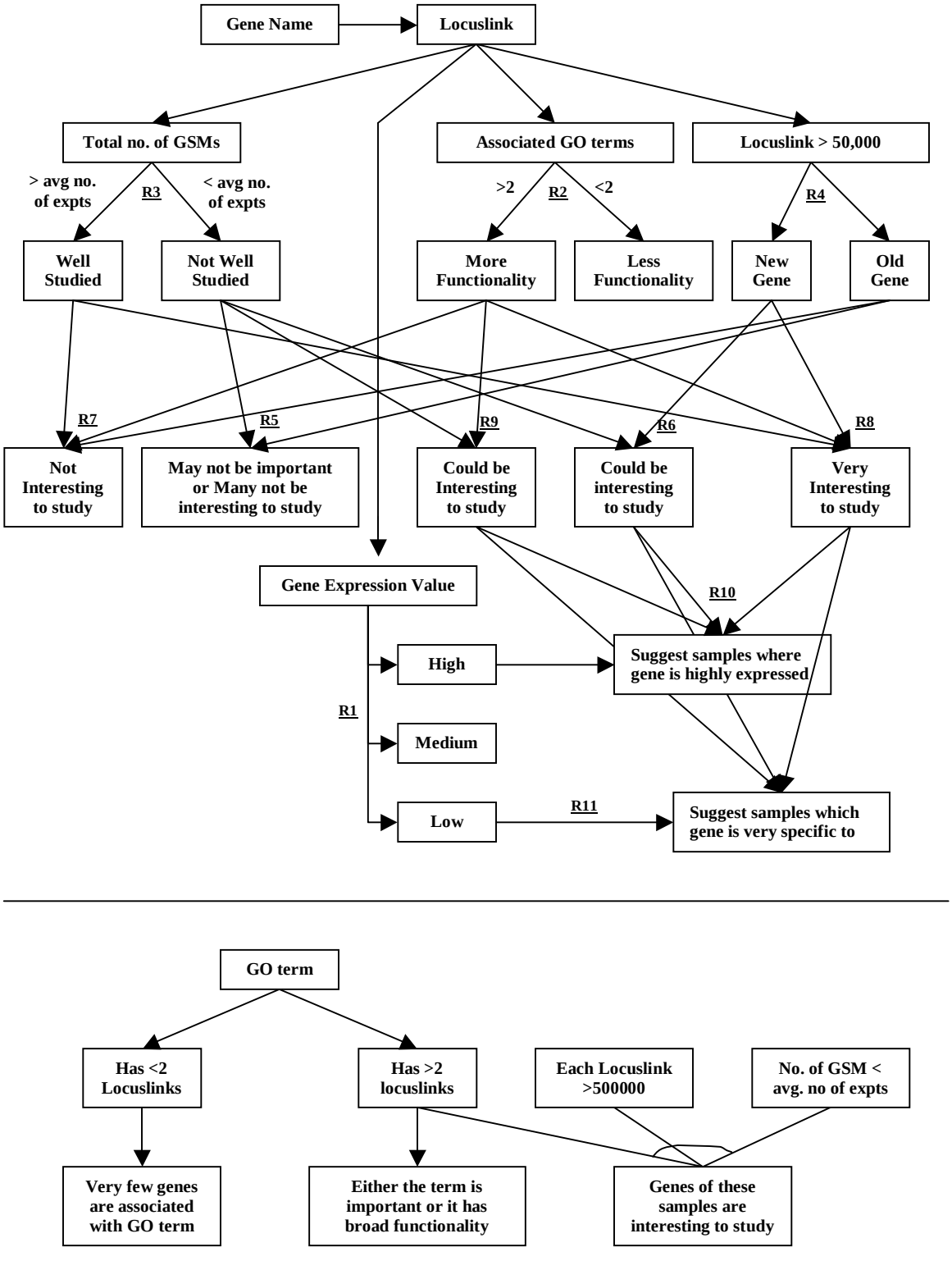


Fig 2: Heuristic Classification PSM of GXplain

2.3 Implementation Details

In the actual implementation, we have linked all of the concepts used through tables in a Relational Database Management System. We used "procedural knowledge" in GXplain – i.e. knowledge is stored as program code. This is usable within only a specialized problem-solving context and is highly useful in that domain context. Due to the unimaginable enormity of GO terms and the GEO Datasets, and to build a feasible, working application given the short duration, we have instantiated the ontology through the GO term "Insulin Secretion" and all the terms associated with it.

The Knowledge Base contains propositions about the domain concepts. In GXplain, we have used data from the Gene Ontology and the GEO Relational Database Tables as our domain knowledge. In our actual implementation, we selectively joined several tables in the GEO database to get our own customized table with fields Locuslink, GSM, GSM Title, and Normalized Rank Value. Another table is related to Gene Ontology – it has fields Locuslink, Gene Name, GO term, GO term Annotation and Number of Experiments. The two tables are linked through the Locuslink, and this is how we obtain the necessary relationship between GEO and GO.

As per Rudolph's recommendations (Rudolph, 2000), we chose to use a Rule-based system as there is significant decision making involved. The decisions made by the system are not very intuitive. The rules are static, but more rules could be added to infer more about genes and their effects. These are the reasons why we chose a rule-based system. An Inference Engine contains procedures that operate on the propositions made in the Knowledge Base. We have implemented our Inference engine as a Rule-based system (with a set of If-Then and If-Then-Else rules) that infers intelligent reasoning if the domain knowledge satisfies some constraints.

At the Client side, we built a simple-to-use User Interface using HTML and JavaScript, which accepts user input (i.e. Locuslink or Gene Name or GO term). We used PHP Scripting language to implement the Inference Rules and make intelligent reasoning upon the data. The entire application is deployed as a Web Service through an Apache Web Server.

3. RESULTS/APPLICATION BUILT

GXplain uses the Knowledge Base (as described in section 2.1) and Heuristic Classification Problem Solving Method for inferring useful information about a gene or a GO term. It provides a comprehensive overview of genes and GO terms and gives research advice for studying them. One important feature of GXplain is that it not only gives study advice but it also gives the reasons behind suggesting particular advice. The User Interfaces is simple to use and easy to understand – this helps people with little knowledge of genes to understand why a gene is interesting to study.

3.1. Description of GXplain application

When the GXplain application is invoked, the user can choose between three choices to enter: Locuslink, Gene Name or GO term.

Output for a Locuslink/Gene Name

1. Gene Overview: It presents a summarized display of the Locuslink, the Gene Name and the associated GO terms.
2. Graphical Representation of the Gene in Microarray experiments: This plots a bar graph that displays a categorized view of gene expression values of that gene in different experiments. The

classification into five categories, i.e. “Very Highly Expressed”, “Highly Expressed”, “Medium Expressed”, “Lowly Expressed” and “Very Lowly Expressed”, is done using the normalized rank values of the gene in different samples.

3. Inferences section: It answers the primary question whether a gene is “interesting” to study or not. It also lists out answers to several other smaller questions which are involved in answering the above big question.

Output for a GO term

1. Lists all genes (and their Locuslinks) associated with the GO term. User can click a gene for further study of a gene using GXplain.
2. GXplain uses Heuristic Classification and predicts which genes are useful for the study of this GO term.

3.2. Knowledge inferred by GXplain

The primary aim of GXplain is to give study advice for a gene and a GO term. For a given Locuslink/Gene Name it provides following knowledge along with study advice:

- Overview of a gene and GO terms associated with it.
- Degree of presence in the body: Depending upon gene’s data in different categories (High, medium, low), it infers about its existence in the body.

  For instance: *Rule1* IF a gene is highly expressed in most of the experiments THEN it is present almost everywhere in the body.
- Functionality of a gene: By counting the number of GO terms associated with a gene, GXplain can predict the functionality of a gene.

  For instance: *Rule2* IF a gene has more than two associated GO terms with it THEN this gene is more generic and performs several functions.
- Age (discovery) of a gene: The Locuslink is given to a gene depending upon when it is discovered. A higher Locuslink number than the mean Locuslink number indicates that this gene is relatively new.

  For instance: *Rule4* IF Locuslink > mean Locuslink THEN gene is recently discovered.
- Interest level: Interest level for further study or for performing experiments with a gene can be inferred depending upon the gene’s age and data.

  For instance, *Rule 8* IF a gene is newly discovered AND it is not well-studied THEN it is useful to further study the gene. Similarly, there are several other complex rules to predict interest level of a gene.
- Samples where gene is highly expressed or specific to: GXplain also predicts the most relevant sample title for the given gene be analyzing gsm title and its rank value.

For a GO term GXplain infers following knowledge:

- Nature of genes associated with a GO term: GXplain lists all genes associated with a GO term and provides the link for studying the gene individually using GXplain knowledgebase for the gene.
- List of genes which are most interesting for the study of a given GO term: GXplain lists relevant genes for a GO term by analyzing over GO term, associated genes, degree of expressiveness, amount of data available and its age.

4. DISCUSSION

As discussed in earlier sections, GXplain successfully provides research advice for a gene and a GO term along with reasons behind it. In addition, GXplain is also successful in summarizing and representing microarray experiments data in a comprehensive format through graphical representation. This is unique since, most of the formal ontological analysis of gene expression data and Gene Ontology has focused on getting a GO term for the genes that are highly expressed (Khatri and Dadhichi, 2005)

4.1 Benefits and Limitations

Benefits of our implementation: GXplain adheres to all the three properties of reasoning systems:

1. Soundness: Knowing that if a query is asked of the system, the result will be logically consistent with all other propositions in the KB (you will get right answer)

2. Completeness: Knowing that if it is possible to conclude a given proposition from a set of axioms, the system indeed will be able to make that conclusion. (you will get all the answers)

3. Decidability: Knowing that if a query is asked of the system, the system will indeed return a result in a reasonable period of time (i.e. you will get an answer)

Limitations of our implementation: The Gene domain is very complex and there are no standard answers to many questions. GXplain makes some assumptions for deducing inferences about a gene or a GO term. Some of these assumptions may be imprecise in the biological perspective. The other limitation is that GXplain's functionality is restricted to the Insulin domain. However this can be easily be extended to other domains.

4.2 Ontology/Data structures

For representing ontology of GXplain we evaluated many available options including First Order Logic, Frames, Semantic Network and Database Schema. The main aim was to optimize data representation and data extraction as GXplain's data is of the order of gigabytes. We also needed a more tractable system than a more expressive system. First Order Logic (FOL) is very expressive but not tractable, so that option was ruled out. Concepts in GXplain are not hierarchical and there is no need of inheritance of data, so Frames also were not a suitable representation for our system.

Relational Database schema offers an efficient way of storing and extracting data. In addition, RDBMS provides an easy way of merging different databases together, which was a key requirement in our system for mapping GEO and GO. Moreover, structure of RDBMS (MySQL-client-server architecture) is more suited for a web application like GXplain. A possible disadvantage of this representation is its lack of expressiveness. However, in GXplain, expressiveness is not a key requirement.

4.3 Algorithms/PSMs

GXplain uses Heuristic Classification for inferring knowledge about a gene or a GO term. The Underlying rules are divided into Feature Abstraction rules and Solution Refinement rules. We chose rules and heuristic classification for GXplain because it needs to answer several smaller questions before answering the main question. It also makes our system more extensible by allowing addition of new rules. However, unlike MYCIN (Clancey, 1997), we chose forward chaining because we did not have a goal at start but we had rich data.

We also considered propose- and-revise and cover-and-differentiate PSMs for GXplain, but it did not suit our specific requirement as GXplain does not do repeated prototyping or iterative matching. We could also use a Bayesian Belief Network for associating weights to output(s) of GXplain. However, given that we have a large amount of data, it may make our system computationally infeasible. . Similarly, open world assumption of DL makes DL unsuitable for GXplain.

REFERENCES

Davis, Buchanan, Shortliffe. Production rules as a representation for a knowledge based consultation Program. *Artificial Intelligence* 8, 15-45, 1977.

Rudolph : "Some Guidelines For Deciding Whether To Use A Rules Engine". JESS mail archives at http://www.mail-archive.com/jess-users@sandia.gov/msg01887.html, 2000.

Clancey. “Mycin’s Map”, In: *Situated Cognition, On Human Knowledge and Computer Representations.* Cambridge University Press, pp. 29-45, 1997.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25 (1):25-29, 2000.

Edgar et al. Gene Expression Omnibus: NCBI gene expression and hybridization array data respitory. *Nucleic Acid Research*, 2002, vol. 30 (1): 207-210, 2002.

Khatri and Draghici: Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics review,* vol 21 no 18, pp 3587-3595, 2005.

Benson et el, Genbank, *Nucleic Acids Research* January 13;33(Database Issue):D34-D36, 2005. http://www.ncbi.nlm.nih.gov/Genbank/index.html

DIVISION OF LABOR

Idea about the project	: Ki Kim, Rashmi Raj, Ravi Tiruvury, Sine Zambach
Exploring domain knowledge	: Ki Kim, Rashmi Raj, Ravi Tiruvury, Sine Zambach
Deciding ontology and Problem Solving Method	: Ki Kim, Rashmi Raj, Ravi Tiruvury, Sine Zambach
Implementation and coding	: Rashmi Raj, Ravi Tiruvury
Study of Insulin Domain	: Ki Kim, Sine Zambach
Preparing report and presentation	: Ki Kim, Rashmi Raj, Ravi Tiruvury, Sine Zambach

GLOSSARY - Concept and cross-references are in *italic*

Body Tissue: Organ or cell-type.

Expression (value)*:* Value that correlates with how many copies of each gene that is present in a certain tissue-mixture.

DNA: building blocks for genes. DNA are translated to *RNA*. Sometimes refered to as the genetic code since it is responsible for the genetic propagation of most inherited traits.

Expressiveness: Being able to say everything you would like about the world being modeled.

Gene: Used in the definition of the DNA-sequence, and is not necessarily equal to the *gene product*.

GeneID: number unique for each gene. Also known as *Locuslink*, developed by NCBI (American).

Gene Ontology: Ontology for classifying genes (The Gene Ontology Consortium, 2000).

Gene Product*:* Protein or functional RNA.

GEO: Gene Expression Omnibus – a database collecting almost all microarray-data in the world (Edgar et al, 2002).

GSM: identity of every microarray-value

GXplain: A System that Supports Biological Research using Information from *GEO* Database and the Gene Ontology.

Locuslink: Number unique for each gene. Also known as *GeneID*.

Microarray: Biochemical experimental methods that tells about each genes expression in different body tissues.

Normalization: statistical approach that makes different kinds of microarray data comparable.

Pathway: (or biological pathways). Used as an overall title of a large or small biological functional network.

Rank(value): Normalization method used by Dr Atul Butte to make the microarray-data comparable.

RDBMS: Relational Database Management System, where knowledge is stored as program code

RNA(mRNA)*:* Translated from *DNA*. Translate *gene* into *gene product*.

Tractability: Knowing that the inference is decidable and computationally efficient.

Uniprot, Unique ID for a protein, developed by EMBL (European).

CODE SUBMISSION: Attached

PROGRAM EXECUTABLES: Attached

INSTRUCTIONS ON RUNNING GXPLAIN: Currently, GXplain is not available publicly. But if required, we can switch our server on, which can be accessed at "http://rashmi.stanford.edu". An alternate approach for running the code is to create a virtual host and put the files in public_html/cgi-bin folder of the host. The Server should load PHP module and should support GD library and JP graph.

SAMPLE DATA:

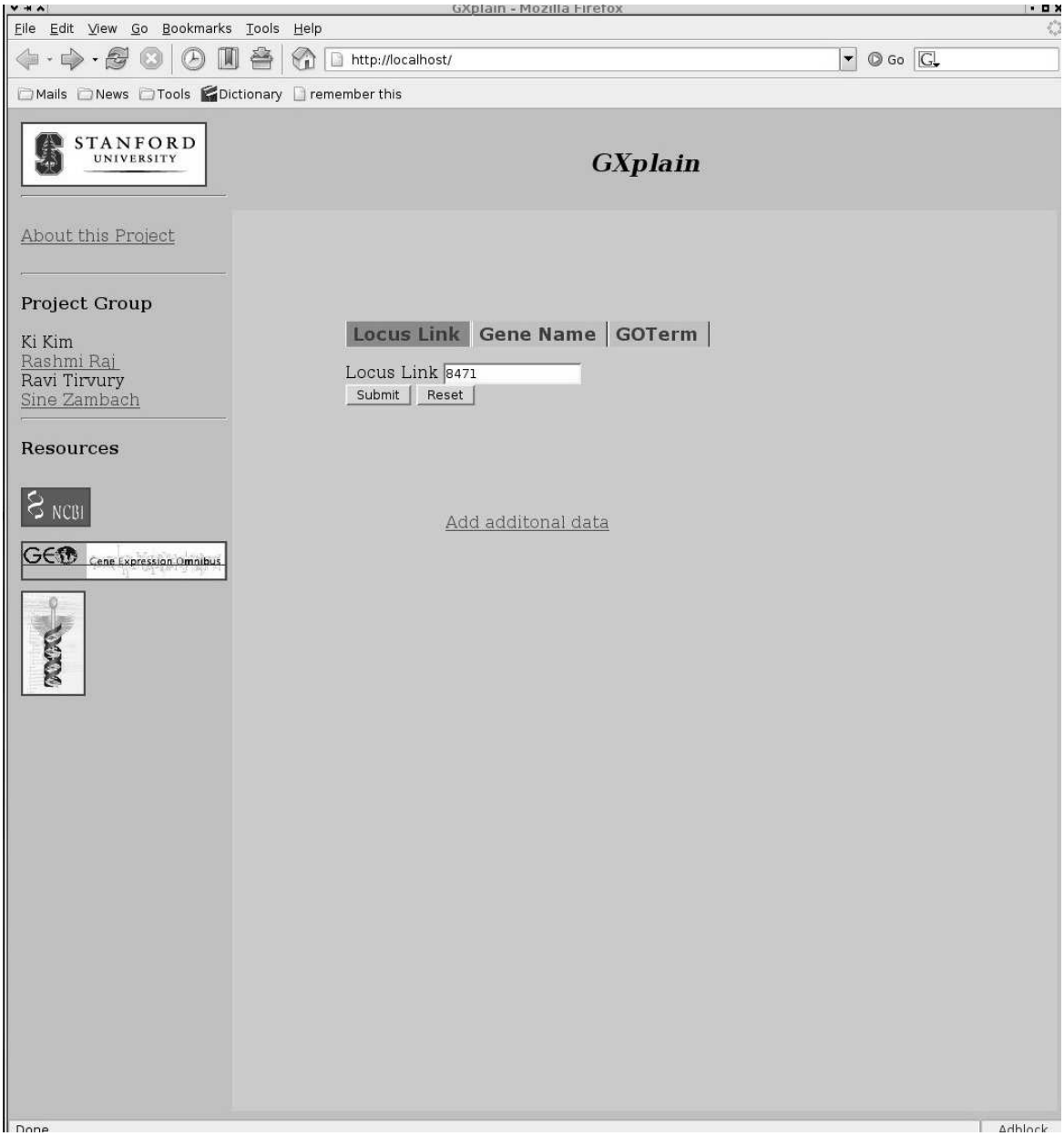
Gene Name	Locuslink	Process	Function
DOK3	1795	Insulin Receptor Binding	MF
GRB14	2888	Insulin Receptor Binding	MF
IRS4	8471	Insulin Receptor Binding	MF
SRBS1	10580	Insulin Receptor Binding	MF
FRS3	10817	Insulin Receptor Binding	MF
LIP5	51534	Insulin Receptor Binding	MF
PHIP	55023	Insulin Receptor Binding	MF
LIP4	57060	Insulin Receptor Binding	MF
LIP6	85406	Insulin Receptor Binding	MF
AHSG	197	insulin receptor signaling pathway	BP
AP3S1	1176	insulin receptor signaling pathway	BP
GAB1	2549	insulin receptor signaling pathway	BP
GRB10	2887	insulin receptor signaling pathway	BP
IGF1R	3480	insulin receptor signaling pathway	BP
IGF2	3481	insulin receptor signaling pathway	BP
PDPK1	5170	insulin receptor signaling pathway	BP
PIK3R2	5296	insulin receptor signaling pathway	BP
PIK3R3	8503	insulin receptor signaling pathway	BP
SORBS1	10580	insulin receptor signaling pathway	BP
SIK2	23235	insulin receptor signaling pathway	BP
CAMK2G	818	Insulin Secretion	BP
PPARD	5467	Insulin Secretion	BP
SNAP25	6616	Insulin Secretion	BP
STX1A	6804	Insulin Secretion	BP
GAL	51083	Insulin Secretion	BP
FAM3B	54097	Insulin Secretion	BP
FAM3D	131177	Insulin Secretion	BP
CTGF	1490	insulin-like growth factor binding	MF
IGFALS	3483	insulin-like growth factor binding	MF
IGFBP1	3484	insulin-like growth factor binding	MF
IGFBP3	3486	insulin-like growth factor binding	MF
INS	3630	insulin-like growth factor binding	MF
CG11910	3483	insulin-like growth factor binding protein complex	CC
Igfbp3	3486	insulin-like growth factor binding protein complex	CC
Igfbp5	3488	insulin-like growth factor binding protein complex	CC
CG32055	39188	insulin-like growth factor binding protein complex	CC
IGF1R	3480	insulin-like growth factor receptor activity	MF
MPRI_HUMAN	3482	insulin-like growth factor receptor activity	MF
IGF1	3479	Insulin-like growth factor receptor binding	MF
IGF2	3481	Insulin-like growth factor receptor binding	MF
YWHAG	7532	Insulin-like growth factor receptor binding	MF
YWHAH	7533	Insulin-like growth factor receptor binding	MF
SOCS1	8651	Insulin-like growth factor receptor binding	MF
FAM3D	131177	Negative Regulation of Insulin Secretion	BP
AHSG	197	regulation of insulin receptor signaling pathway	BP
SIK2	23235	regulation of insulin receptor signaling pathway	BP
PPARD	5467	Regulation of Insulin Secretion	BP
SNAP25	6616	Regulation of Insulin Secretion	BP
STX1A	6804	Regulation of Insulin Secretion	BP
FAM3D	131177	Regulation of Insulin Secretion	BP

HOW TO ADD ADDITIONAL DATA: GXplain has a link on the Home Page for adding data. After clicking the link, user is directed to a new page where he can enter data in text boxes and can submit it.

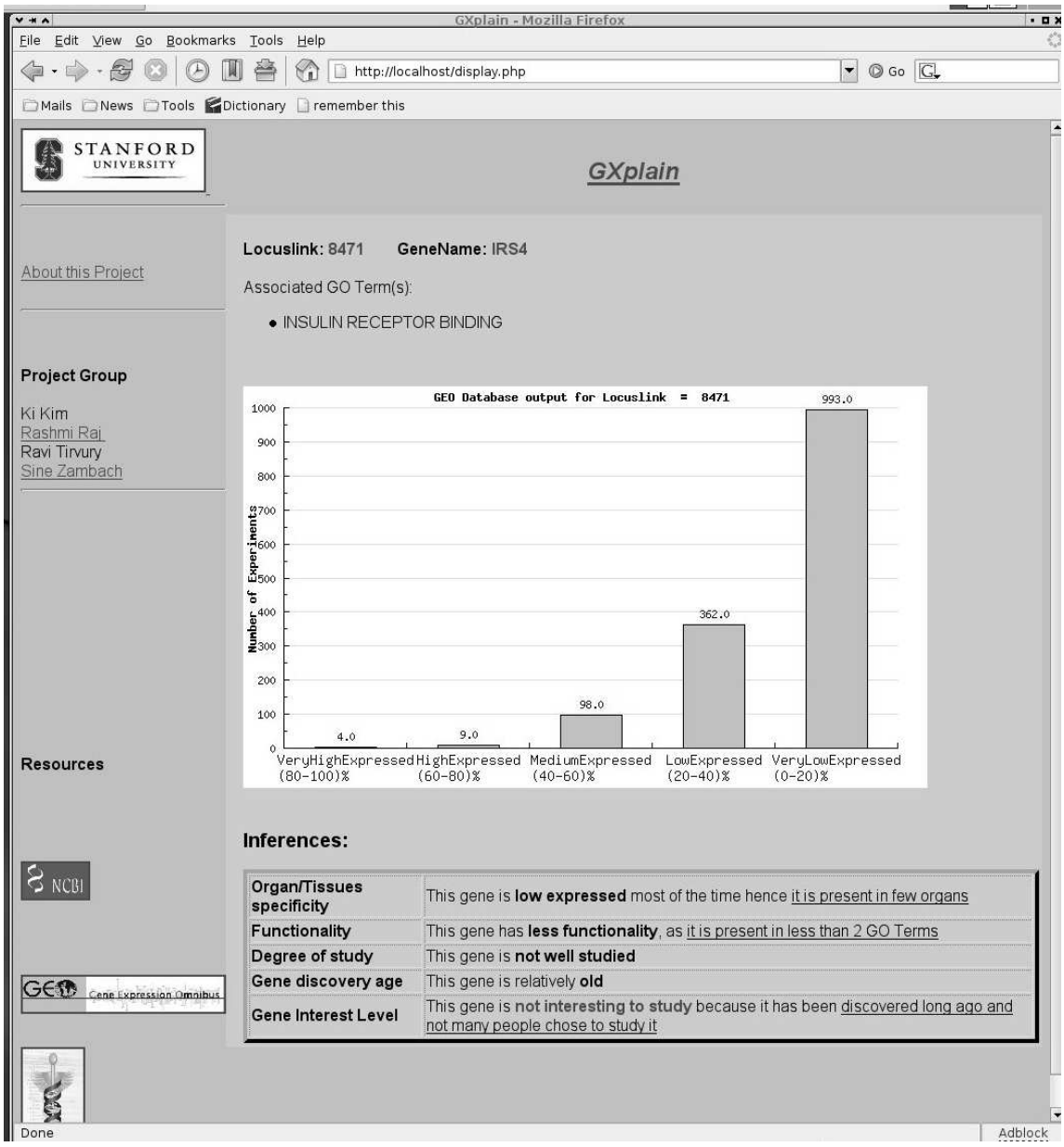
Rules:

Rule 1	a. IF a gene is Low Expressed THEN the gene is present in specific tissues. b. IF a gene is Medium Expressed THEN the gene is moderately present everywhere. c. IF a gene is High Expressed THEN the gene is present everwhere.
Rule 2	a. IF a gene has more than two GO terms associated with it THEN the gene is more functional. b. IF a gene has less than two GO terms associated with it THEN the gene is less functional.
Rule 3	a. IF a gene is present in more than average number of experiments THEN gene is well studied. b. IF a gene is present in less than average number of experiments THEN gene is not well studied.
Rule 4	a. IF a locuslink number is greater than the mean locus link number THEN the gene is recently discovered. b. IF a locuslink number is less than the mean locus link number THEN the gene is discovered long time back.
Rule 5	IF a gene is (not well studied) AND (old), THEN the gene may not be interesting to study.
Rule 6	IF a gene is (not well studied) AND (new) THEN the gene could be interesting to study.
Rule 7	IF a gene is (new) AND (more functional) THEN the gene could be interesting to study.
Rule 8	If a gene is (old) AND (well studied) THEN the gene is already studied and may not be interesting for further study.
Rule 9	IF a gene is (well) studies AND (not functional) THEN the gene is not very interesting to study.
Rule 10	IF a gene is (interesting to study) AND (highly expressed) THEN suggest tissues where the gene is highly expressed.
Rule 11	IF a gene is (interesting to study) AND (low expressed) THEN this gene is specific to an organ and this gene is good to study that organ. Suggest that organ where it is specific to.

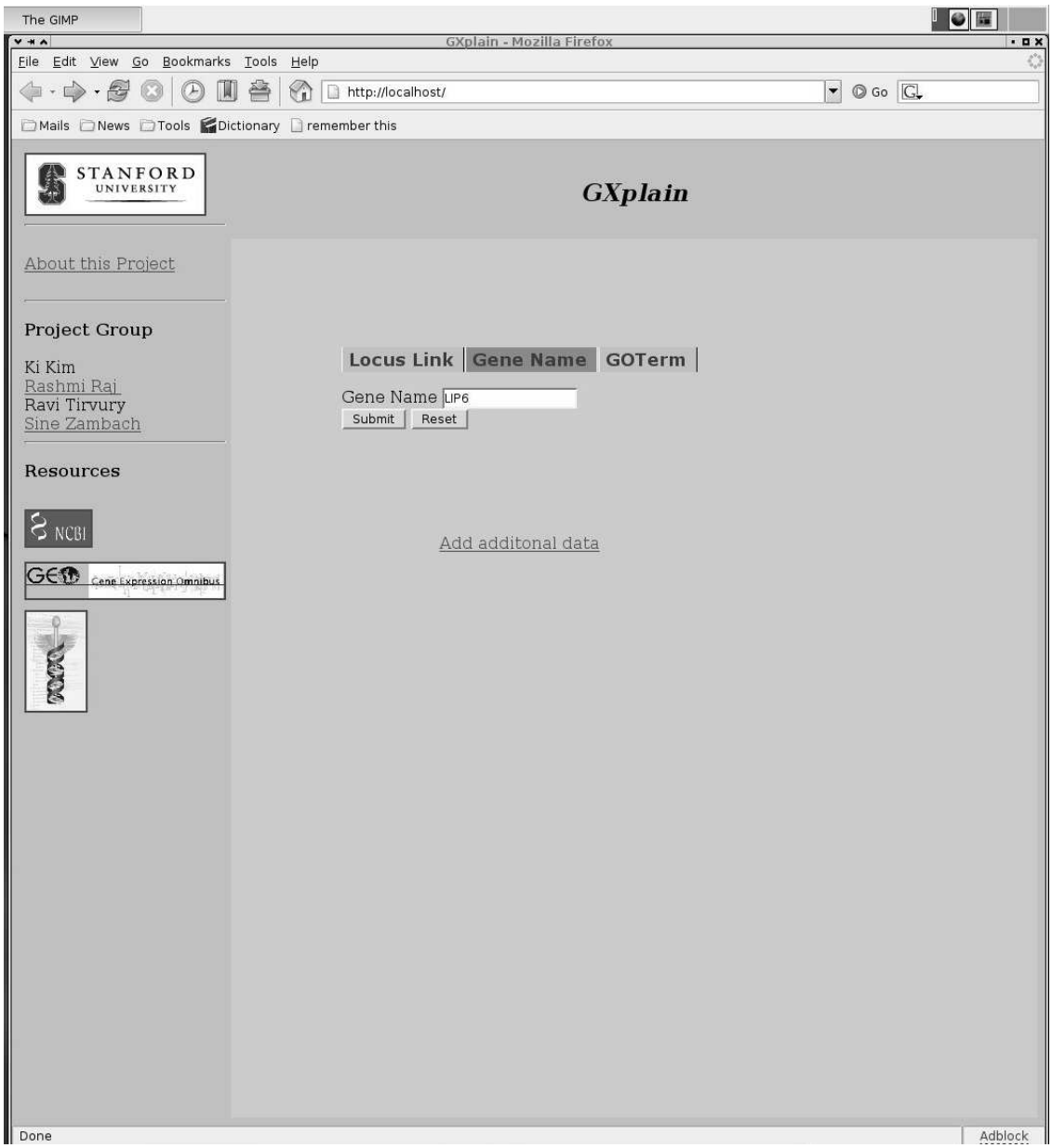
SCREEN SHOTS:



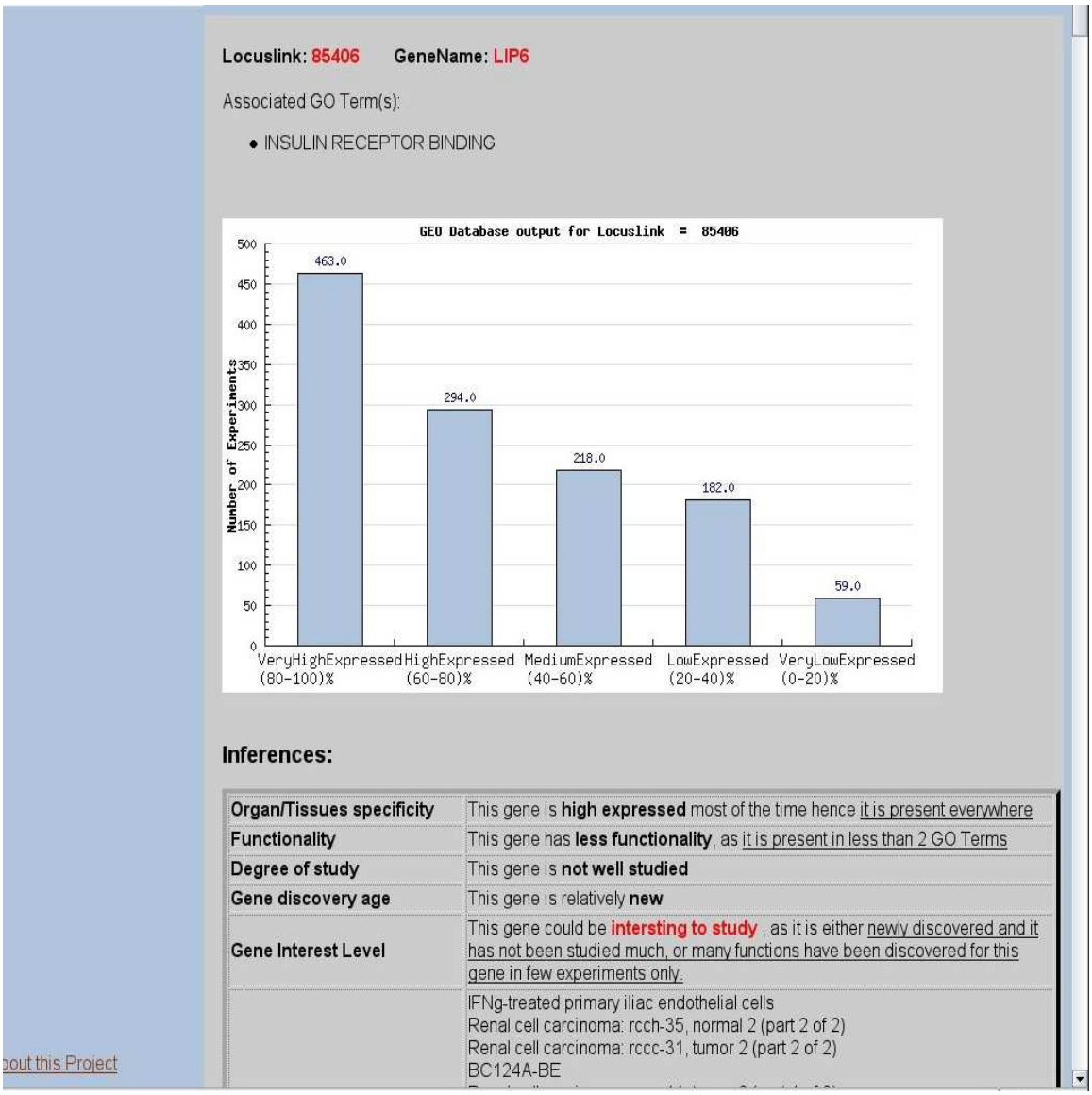
User enters Locus Link 8471



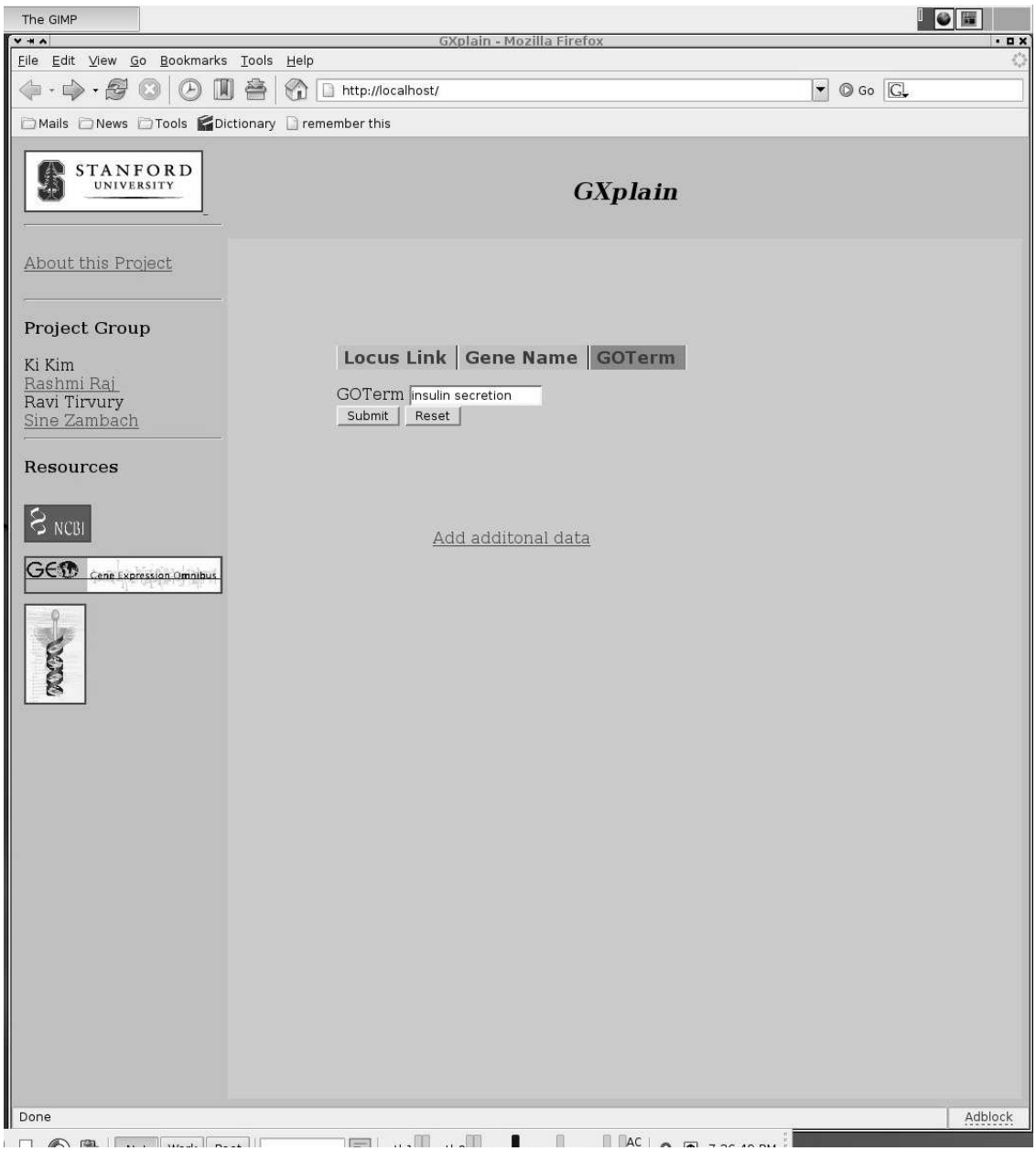
Result for Locus Link 8471



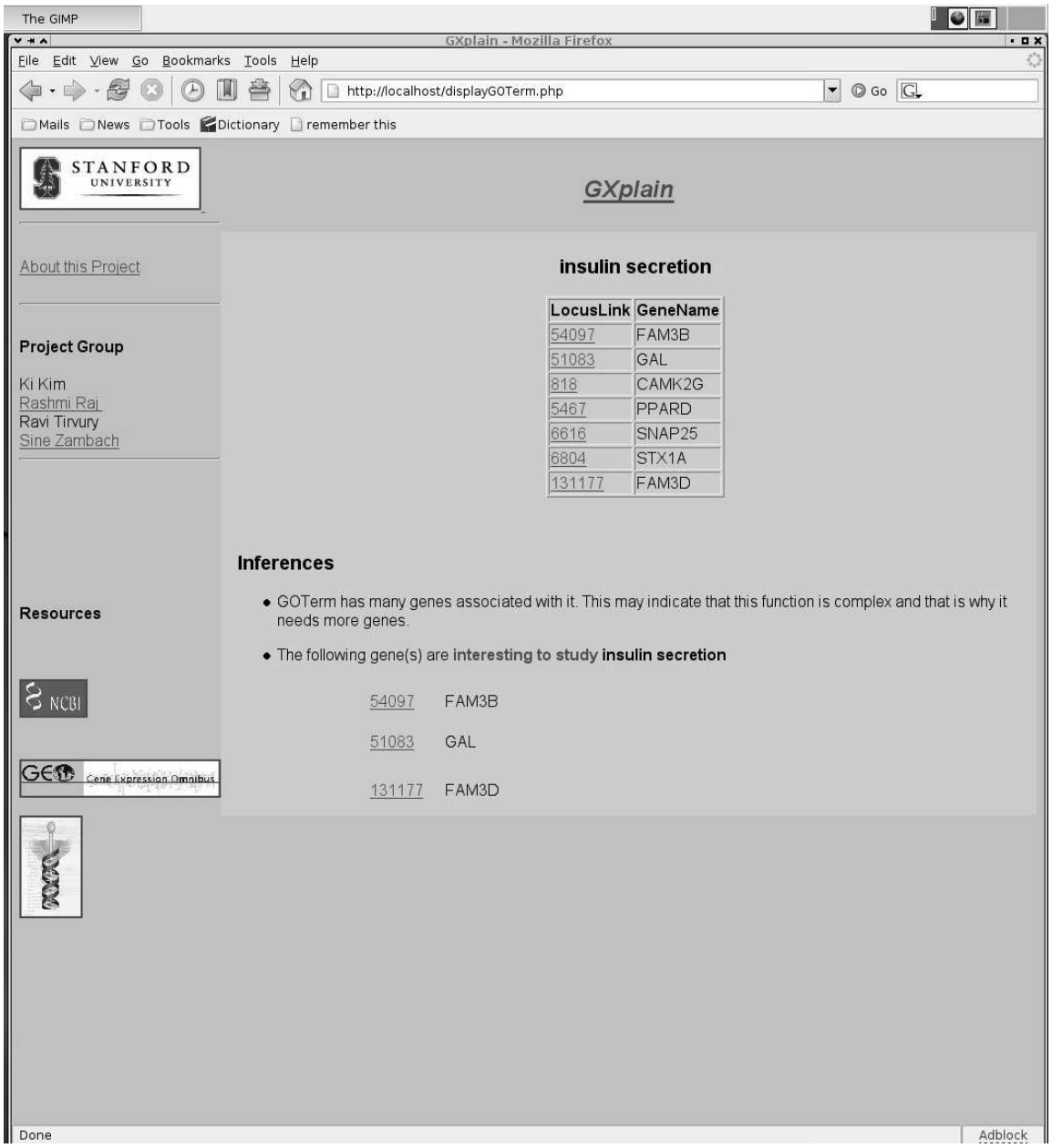
User enters Gene Name LIP6



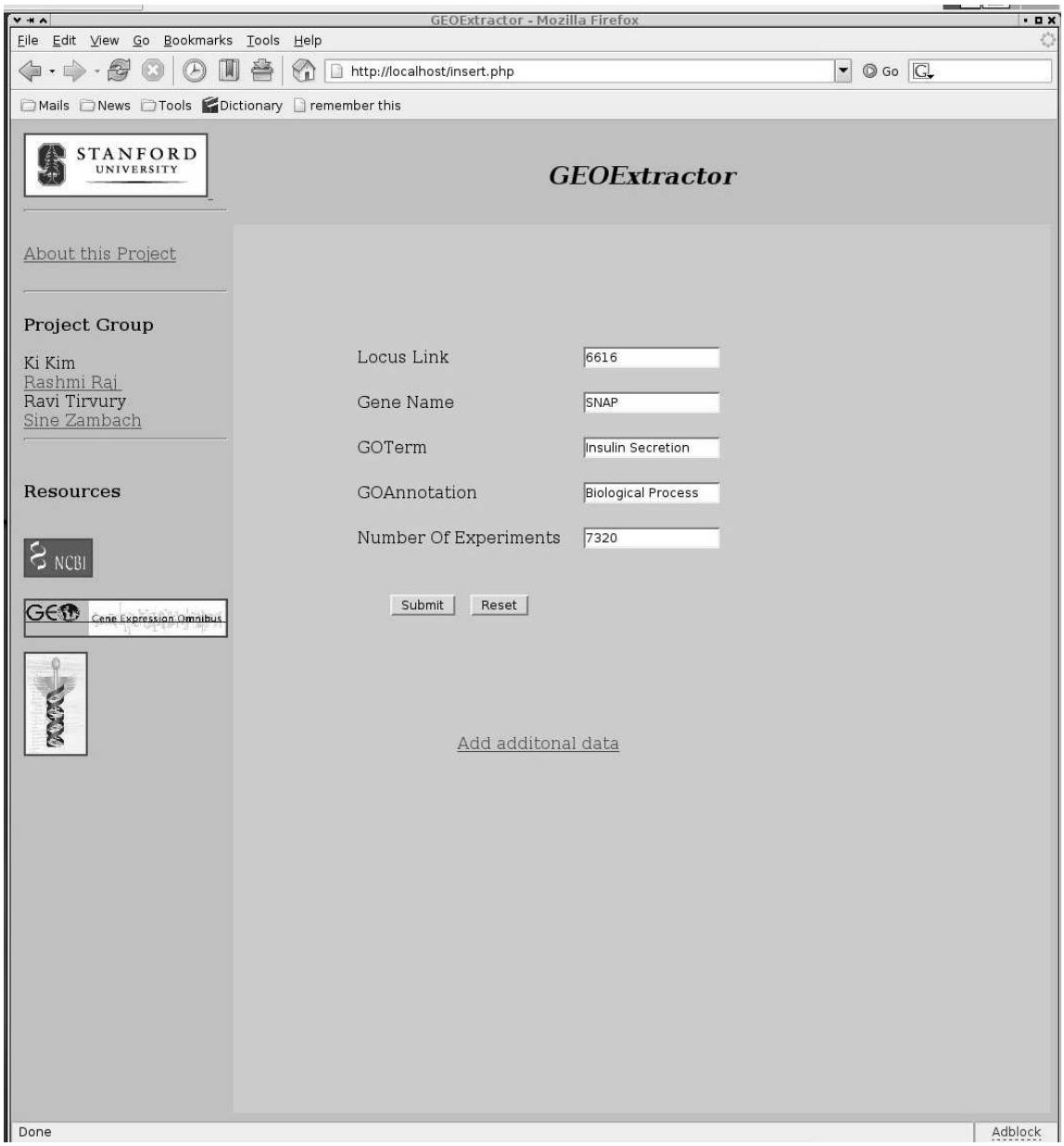
Result for Gene Name LIP6



User enters GO Term Insulin Secretion



Results for GO Term Insulin Secretion



Entering additional data into the system