

**CBS 584
Internship Report**

**Title:
Exploring and Exploiting the Biological Maze**

**Internship Advisor
Dr. Zoe Lacroix**

**An Internship Report Presented in Partial Fulfillment
Of the Requirements for the Degree
Master of Science
In
Computational Biosciences
August 2004**

**Submitted By
Vidyadhari Edupuganti
993-74-2420**

TABLE OF CONTENTS

TITLE	PAGE
ABSTRACT	3
INTRODUCTION	4
OBJECTIVE	7
METHODOLOGY	8
EXPERIMENT	9
IMPLEMENTATION	10
QUERYING NCBI DATABASES	10
RESULTS	12
ANALYSIS	18
CONCLUSION	24
EXPLOITING RESULTS	25
FUTURE WORK	25
FUTURE DIRECTIONS OF THE PROJECT	26
ACKNOWLEDGEMENTS	26
REFERENCES	27
APPENDIX	28
PAPER SUBMITTED AT CSB CONFERENCE	39

ABSTRACT

The field of life sciences research partially depends on the collection of information related to multiple scientific objects of interest (e.g., "Retrieve all genes involved in brain cancer", "Retrieve all citations related to diabetes").

Scientists are interested in exploring multiple data sources in order to explore relationships between scientific objects. Each data source provides specific capabilities that allow scientists to access, navigate, and analyze the data. The research work presented mainly focuses on the fact that resource selection (data source and capability) in the data collection process significantly affects the quality and completeness of the data. Preliminary research begins with the hypothesis that the data collection process depends on two orthogonal variables: the number of data sources involved in the process, and the selection of capabilities available at these resources. We tried to prove our hypothesis by collecting results for four commonly used biological resources: the NCBI Nucleotide, Protein, PubMed and OMIM databases. The results obtained during this preliminary research prove our initial hypothesis that data collection process indeed depends on the data source and the capability selected.

INTRODUCTION

One of the main challenges that a life sciences scientist faces today is collecting information about scientific entities. These entities could be genes, citations, structures or sequences. The challenge is due to the abundance of biological data sources. For example, if a scientist wants to look at a simple query like looking at

protein sequences he could either go to the Protein database of NCBI, Swiss-Prot or EMBL. To answer a simple a scientist has to choose from numerous sources. Each source is different whether it is in terms of its interface, or the types of links it provides, or its capability. The situation is confounded when a scientist has to gather information from two or more sources, which are inter-related. Scientists have to select the resources to exploit in the exploration process and then fully traverse links and paths through these sources given some start object.

The fundamental challenge to information collection is that paths between data sources potentially have different properties and yield different benefits. The situation becomes more complex as the sources involved in answering a query increase. Indeed in general, no single source has all the required information required and one is faced with the challenge of exploring interrelationships (paths) between sources.

Moreover, the results obtained might be different from source to source or even within the same source. Yet such results are very crucial in terms of quality, cost (time and space), and completeness (number of entries and characterization) for any scientific discovery. Also, some scientists may be interested in the quality of the data, others in speed of accessing the data, while others may use a certain source out of convenience.

There is no integrated system that guides a scientist to effectively choose between sources and their capabilities. The building of such a system is only possible when one can fully understand the sources, their capabilities and the different paths in which they are involved. Our motivation is to develop an approach suitable for exploitation of biological resources representing multiple properties, such as to optionally meet the needs of the specific query.

Consider a preliminary test based on two simple queries. The first query was to “Retrieve all sequences related to a disease condition Cancer”. Here we are looking at three different data sources to answer the query .The results are shown in Table 1. The second query was to “Retrieve all sequences related to a disease condition diabetes mellitus” using the same data source NCBI with the same capabilities but with different paths.

To answer the query one has to first get the data about the disease condition in terms of genes involved and then link that data to the citations. The data source used here is NCBI. The results are shown in Table 2.

1. **Query 1:** Retrieve all sequences related to “Cancer”

Data Source	NCBI	SWISSPROT	TrEMBL
Entries Retrieved	126,174	148,516	1,067,463

Table 1: Results for Query 1

2. **Query 2:** Retrieve all genes and citations related to “diabetes mellitus”

	NCBI Nucleotide → PubMed	OMIM → PubMed	Gene → PubMed
Genes	743	296	228
Citations	4277	6906	4147
Capability	PubMed Link	PubMed Link	PubMed Link

Table 2: Results for Query 2

The results from query one indicate that indeed different data sources produce different results for the same query. The results from the second query indicate that even though we query the same data source using the same capability but

different paths we get retrieve different results. These observations motivate exploring the different paths and capabilities offered by standard data sources.

This paper provides a sound basis for the comparison of properties of different paths through interlinked data sources, and thus a means for optimization recognizing the importance of both execution time, as well as quality of the results.

This paper is divided into different sections, which includes the main objective of this research, the methodology followed to accomplish the objective, the experimentation procedure, implementation of the experiment, results obtained and a preliminary analysis based on the results. Then this paper concludes with and directing towards future work.

OBJECTIVE

There are many paths to answer a query as shown above in the second query. Each possible evaluation path may have different properties. The main objective of this project is to demonstrate that the data collection process is affected by two orthogonal variables:

1. The number of data sources involved in the process
2. The selection of capabilities available at these resources.

METHODOLOGY

We developed a methodology where we can look at all possible ways to prove our objective.

In order to accomplish the objective data was collected and analyzed. Data was collected based on a common query. In order to answer this query we need to select data sources that can answer the query. Sources have to be chosen very carefully. We need to find a source that has good information regarding the query. Based on the source selected we have to select the capabilities. Different implementations of the same query may yield different results. We considered looking at other sources that are involved at the intermediate level. By introducing sources, we would be able to look at different paths and see if the length of the path affects data collection process. Following different paths to collect data might yield different results. The results might vary when the path traversed contain more than two sources. Analysis of the data collected will be done in terms of no of entries retrieved by each path and implementation, time involved in the data collection process and space occupied.

Analyzing the results would help us choose among a good set of paths and implementations to maximize benefit and minimize cost.

EXPERIMENT

Scientists are interested in exploring relationships between scientific objects, e.g., genes and bibliographic citations. Therefore, we considered the query “*Return all citations that are related to some disease or condition.*” The data was collected based on this query.

The most commonly used source for disease conditions is OMIM (Online Mendelian Inheritance in Man), which contains information related to human

genetic diseases. The most commonly used data source for citations is PubMed, which includes over 14 million citations for biomedical articles. These two sources are hosted at the National Center for Biotechnology Information (NCBI). In order to look at different paths we added two intermediary resources, Protein and Nucleotide that are available at NCBI. Three capabilities that offered by NCBI website were chosen to collect the data. Therefore, we looked at three paths, three capabilities, and nine implementations of the query.

The experimental procedure involved four commonly used biological resources: the NCBI Nucleotide, Protein, PubMed and OMIM. The basic steps involved in the experiment are to first start from OMIM source, which contains information related to human genetic diseases and then retrieve all PubMed citations related to those diseases. The study focuses on three medical conditions: *cancer*, *aging*, and *diabetes*. For each of these conditions domain experts provided a list of relevant keywords. The execution of this query explores all paths from the starting object OMIM, and end object PubMed.

However, intermediate resources such as the NCBI Nucleotide and Protein databases may also be involved. The starting point is always OMIM and the ending point is always PubMed. The choice of the capability of the data source also has an impact on the result. To show this we use three capabilities provided by NCBI to express the relationships between scientific objects. The number of citations retrieved at the end of each path was recorded along with the time taken and space occupied by the results.

IMPLEMENTATION

A wrapper, implemented in Java, was used to make successive calls to the E-Link interface of NCBI, to follow the links from OMIM records to each of the other 3 sources, as well as to traverse all three paths from OMIM to PubMed. Three separate programs were written in order to query each path. Separate code snippets were used to remove duplicates and to calculate overlaps.

QUERYING NCBI DATABASES

We consider three different evaluation paths (without loops) starting from OMIM and terminating in PubMed (see Fig.1).

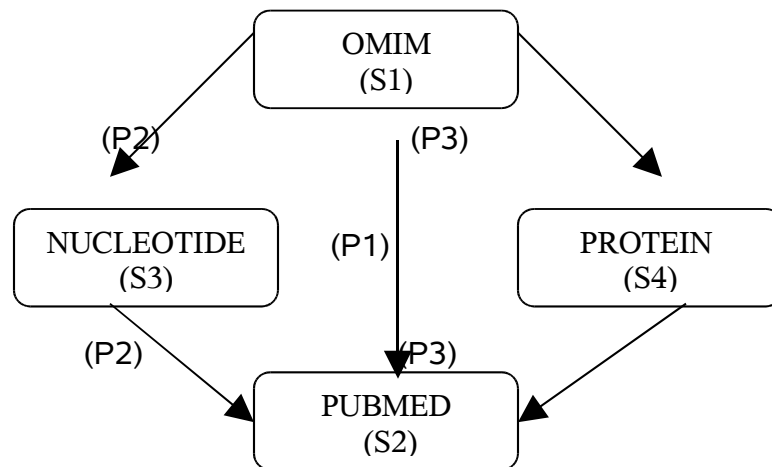


Figure 1: Three Evaluation Paths

The choice of paths has an impact on the result [1]. For example, traversing a path via the Nucleotide source might yield fewer and different citations, as compared to a path via the Protein source. The properties of these paths and their effects are of interest from a number of perspectives [1,2,3]: From a query evaluation viewpoint, one can estimate the cost and benefit of evaluating a query given some specific sources and paths. Query execution time is important to researchers, especially when sources are remote; they are equally interested in the quality of the result.

The study focuses on three medical conditions: cancer, aging, and diabetes. For each of these conditions domain experts provided a list of relevant keywords. Fourteen keywords characterize the disease condition diabetes. Seventy-one keywords characterize the disease condition aging and more than three hundred keywords characterize the disease condition cancer. The keywords for diabetes are shown in Table 3. The keywords for aging and cancer are attached as an appendix.

Type II diabetes	Pancreas Beta Cells
Neuropathy	MHC
Retinopathy	HLA
Kidney Failure	Autoimmunity Genetic Predisposition
Insulin	Diabetes candidate genes
Glucagon	NOD mouse
Animal Models	Type I Diabetes

Table 3: Keywords that characterize “diabetes”

These keywords were used to retrieve relevant genes from OMIM via the E-Search utility of NCBI. These relevant genes constitute the starting set of objects. The execution process then explores the three paths of Fig.1. Each path was explored

by using three capabilities and results for all the three disease conditions for all the three paths using all the three capabilities were recorded.

The three capabilities explored are

- *Link (C1)*: Entrez provides hyperlinks linking an OMIM entry to PubMed citations.
- *Parse (C2)*: Entrez provides parsing each OMIM entry to retrieve its related PubMed citations.
- *All (C3)*: Entrez provides an index from OMIM to PubMed with the PubMed Links in the display option which allows consideration of a set of OMIM entries at a time.

RESULTS

The data was collected from Dec04 - Jan04. The results for each disease condition for all the three paths are given below. The results shown here have total citations for all the paths with duplicates and without duplicates. There are two levels where duplicates were removed. The duplicates obtained with path P1 were removed only at the end (PubMed citations). The duplicates for paths P2 and P3 were removed at intermediate level (Protein/Nucleotide records), as well as at the end (PubMed citations).

- Diabetes: 3638 OMIM records without duplicates were retrieved for the disease condition *diabetes*. The results presented in Tables 4 and 5 and 6 are obtained for the three different paths and three different retrievals for the disease condition diabetes. Table 4 shows

the results of one experiment for the condition *diabetes*, where the records are extracted using the Link option, starting with 3,638 OMIM records. The *measured* values for the first three paths shown in Figure 1 using the Links option are shown in table 4.

PATH	P1	P2	P3
PUBMED Entries With duplicates	49,739	140,946	141,239
PUBMED Entries without duplicates	43,890	42,969	59,959
Total Execution Time	00:30:00	14:24:00	7:17:00
Size	724KB	6,843KB	4,690KB

Table 4: Results of Link option for the condition *diabetes*.

Table 5 shows the results of one experiment for the condition *diabetes* where the records are retrieved using the Parse option, starting with 3,638 OMIM records. It shows the *measured* values for the first three paths shown in Fig. 1 by using the Parse option.

PATH	P1	P2	P3
PUBMED Entries With duplicates	50,028	136,770	137,799
PUBMED Entries without duplicates	43,747	43,090	51,906
Total Execution Time	2:45:00	12:50:00	22:40:00
Size	731KB	6,849KB	4,696KB

Table 5: Results of Parse option for the condition *diabetes*.

Table 6 shows the results of one experiment for the condition *diabetes* where the records are extracted by submitting all the records at once, starting with 3,638

OMIM records. It shows the *measured* values for the first three paths shown in Fig 1.

PATH	P1	P2	P3
PUBMED Entries	44,037	43,581	49,719
Total Execution Time	00:2:57	00:39:02	00:48:40
Size	517KB	1,692KB	1,551KB

Table 6: Results of All option for the condition *diabetes*

- Aging: 4,997 OMIM records without duplicates were retrieved for the disease condition *aging*. The results presented in Tables 7 and 8 and 9 are obtained for the three different paths and three different retrievals for the disease condition *aging*.

Table 7 shows the results of one experiment for the condition *aging* where the records are extracted using the Link option, starting with 4,997 OMIM records. It shows the *measured* values for the first three paths shown in Figure 1 using the Links option.

PATH	P1	P2	P3
PUBMED Entries With duplicates	55,646	175,335	157,548
PUBMED Entries without duplicates	48,393	51,712	60,129
Total Execution Time	00:31:00	10:41:00	10:30:00
Size	888KB	9,509KB	5,508KB

Table 7: Results of Link option for the condition *aging*.

Table 8 shows the results of one experiment for the condition *aging* where the records are retrieved using the Parse option, starting with 4,997 OMIM records. It shows the *measured* values for the first three paths shown in Fig. 1 by using the Parse option.

PATH	P1	P2	P3
PUBMED Entries With duplicates	56,510	176,666	162,424
PUBMED Entries without duplicates	48,398	51,855	61,260
Total Execution Time	00:56:00	19:03:00	11:30:00
Size	896KB	9,528KB	5,560KB

Table 8: Results for Parse option for the condition *aging*.

Table 9 shows the results of one experiment for the condition *aging* where the records are extracted by submitting all the records at once, starting with 4,997 OMIM records. It shows the *measured* values for the first three paths shown in Fig 1.

PATH	P1	P2	P3
PUBMED Entries	48,393	51,474	60,938
Total Execution Time	00:8:36	01:12:00	00:57:11
Size	591KB	2,953KB	2,052KB

Table 9: Results of All option for the condition *aging*

- Cancer: 5,699 OMIM records without duplicates were retrieved for the disease condition *cancer*. The results presented in Tables 10 and 11 and 12 are obtained for the three different paths and three different retrievals for the disease condition cancer.

Table 10 shows the results of one experiment for the condition *cancer* where the records are extracted using the Link option, starting with 5,699 OMIM records. It shows the *measured* values for the first three paths shown in Figure 1 using the Links option.

PATH	P1	P2	P3
PUBMED Entries With duplicates	65,096	183,529	171,060
PUBMED Entries without duplicates	56,315	54,487	62,686
Total Execution Time	00:50:34	12:05:00	8:43:00
Size	1,133KB	9,817KB	5,820KB

Table 10: Results of Link option for the condition *cancer*.

Table 11 shows the results of one experiment for the condition *cancer* where the records are retrieved using the Parse option, starting with 5,699 OMIM records. It shows the *measured* values for the first three paths shown in Fig. 1 by using the Parse option.

PATH	P1	P2	P3
PUBMED Entries With duplicates	67,008	184,460	173,468
PUBMED Entries without duplicates	56,315	54,607	63,367
Total Execution Time	1:1:13	21:08:00	13:18:00
Size	1,150KB	9,832KB	5,845KB

Table 11: Results for Parse option for the condition *cancer*.

Table 12 shows the results of one experiment for the condition *cancer* where the records are extracted by submitting all the records at once, starting with 5,699 OMIM records. It shows the *measured* values for the first three paths shown in Fig 1.

PATH	P1	P2	P3
------	----	----	----

PUBMED Entries	56,532	52,488	60,033
Total Execution Time	00:10:10	1:31:00	1:12:00
Size	810KB	3,709KB	2,168KB

Table 12: Results of All option for the condition *cancer*

ANALYSIS

The results shown here have total citations for all paths without duplicates. The first step in our analysis was to remove duplicates. The second was to look at cost (time/space) and finally look at overlap.

Overlap was calculated in two ways

- Between Paths: Probability that citations retrieved from one path are also retrieved by the path compared
 -
- Between Implementations: Probability that citations retrieved from one implementation of Path1 are also retrieved by the other implementation of Path1

-

Overlap detection is mainly to look at equivalence of sources. If the overlap between two sources is high then a scientist can traverse those sources in either direction. If two sources do not overlap significantly than one has to look at both

sources. The three steps of analysis allow choice among a good set of paths/implementations in order to maximize benefit and minimize cost.

The results shown below are the overlap results for the three disease conditions “diabetes,” “aging”, and “cancer”, respectively. Overlaps were calculated by using a wrapper developed in Java. The process involved counting the citations from one path/implementation that are also retrieved by the compared path/implementation. Then the percentage was calculated according to the recorded values. Detailed tables for the overlap are attached as an appendix for all the disease conditions. The tables shown below only contain the percentage values.

- Diabetes: Overlap results between the three paths and the three types of implementations are shown in Tables 13 and 14.

		P1	P2	P3
C1	P1	100%	25.82%	21.95%
	P2	25.28%	100%	70.00%
	P3	29.98%	97.68%	100%
C2	P1	100%	23.93%	22.87%
	P2	29.18%	100%	81.20%
	P3	33.60%	97.81%	100%
C3	P1	100%	24.75%	24.29%
	P2	24.64%	100%	79.49%
	P3	27.42%	90.68%	100%

Table 13: Overlap Results for the three paths for the condition *diabetes*

		C1	C2	C3
P1	C1	100%	100%	99.64%
	C2	80.52%	100%	80.23%
	C3	99.97%	99.96%	100%
P2	C1	100%	99.71%	94.33%
	C2	99.99%	100%	94.46%
	C3	95.67%	95.53%	100%
P3	C1	100%	99.71%	95.24%
	C2	86.32%	100%	95.33%
	C3	78.97%	91.32%	100%

Table 14: Overlap Results for the three implementations for the condition *diabetes*

- Aging: Overlap results between the three paths and the three types of implementations are shown in Tables 15 and 16.

		P1	P2	P3
C1	P1	100%	26.69%	25.88%
	P2	28.52%	100%	82.39
	P3	32.15	95.80%	100%
C2	P1	100%	26.64%	25.82%
	P2	28.54%	100%	82.41%
	P3	32.68%	97.36%	100%
C3	P1	100%	26.75%	25.88%
	P2	28.45%	100%	82.05%
	P3	32.59%	97.14%	100%

Table 15: Overlap Results for the three paths for the condition *aging*

		C1	C2	C3
P1	C1	100%	99.96%	99.96%
	C2	99.97%	100%	100%
	C3	99.96%	99.98%	100%
P2	C1	100%	99.71%	99.80%
	C2	99.86%	100%	99.98%
	C3	99.34%	99.25%	100%
P3	C1	100%	98.13%	98.58%
	C2	99.97%	100%	99.97%
	C3	99.91%	99.45%	100%

Table 16: Overlap Results for the three implementations for the condition *aging*

- Cancer: Overlap results between the three paths and the three types of implementations are shown in Tables 17 and 18.

		P1	P2	P3
C1	P1	100%	27.97%	27.29%
	P2	27.06%	100%	82.69%
	P3	30.38%	95.14%	100%
C2	P1	100%	27.94%	27.22%
	P2	27.09%	100%	82.78%
	P3	30.63%	96.06%	100%
C3	P1	100%	28.38%	27.66%
	P2	26.35%	100%	82.19%
	P3	29.38%	94%	100%

Table 17: Overlap Results for the three paths for the condition *cancer*

		C1	C2	C3
P1	C1	100%	100%	99.52%
	C2	100%	100%	99.52%
	C3	99.91	99.91	100%
P2	C1	100%	99.73%	99.36%
	C2	99.95%	100%	99.41%
	C3	95.72%	95.55%	100%
P3	C1	100%	98.90%	98.84%
	C2	99.97%	100%	99.81%
	C3	94.66%	94.56%	100%

Table 18: Overlap Results for the three implementations for the condition *Cancer*

The results presented in the Tables 4 to 12 indicate that even though the query asked is similar, we get different results when we choose different paths and capabilities. This proves the initial hypothesis; the data collection process depends on the data source selected. Paths P2 and P3 retrieve more results then compared to Path P1. This proves that the longer the path, the more results we get (Recall that paths P2 and P3 have Nucleotide and Protein as intermediate resources). The three capabilities C1, C2 and C3 retrieve different results for same path. For example, for the disease condition “diabetes”, if we look at path P1 the three capabilities retrieved 43,890, 43,747 and 44,037 pubmed citations, respectively. This proves our second hypothesis: data collection is indeed affected by the capability chosen at the resource level.

The overlap results presented in Tables 13 to 18 indicate that there is a property of inclusion. This indicates that the citations retrieved from one path to some extent are included in the paths compared. For example if we look at the disease condition “aging” the citations retrieved by Path P1 overlap 26.69% with path P2 and 25.88% with path P3. Path P2 overlaps 28.52% with path P1 and 82.39% with path P3. Path P3 overlaps 32.15% with path P1 and 95.80% with path P2.

No matter what disease condition we choose, or what capability we choose, the results indicate that there is a greater overlap between paths P2 and P3, when compared to P1 overlapping with either P2 or P3.

If we look at the different capabilities, for example for the disease condition “cancer”, capability C1 overlaps 100% with capability C2, and 99.52% with capability C3. This indicates that there is a very good overlap between capability C1 and C2.

All the results presented prove our initial hypothesis that the data collection process is affected by two orthogonal variables: The number of data sources involved in the process and the selection of capabilities available at these resources. The length of the path involved in the data collection process also makes a difference. There is a significant overlap between the long paths.

Paths that are more complex may be exploited to find the optimum path with the help of these results. Such results may be exploited to optimize the evaluation of queries [2], or to guide the user in his selection of resources to query [3].

CONCLUSIONS

The results indicate that the selection of resources and available links among them may affect significantly the output, as well as the cost of the evaluation process.

Three different paths from OMIM to PubMed return significantly different number of distinct objects in PubMed

Looking at the results obtained, if were to order the distinct links between paths and capabilities it would be $P1 < P2 < P3$ most of the time and $All < Link < Parse$ most of the time. Most of times, the parse option has retrieved more citations and obviously occupied space and consumed a lot of time. When expressing a query, the scientist may express whether his priority is to retrieve as many entries as possible, as few entries as possible to retrieve them as fast as possible, etc. The results do not demonstrate there is a unique "best path" to collect scientific data, but rather a "best path with respect to ones needs".

In addition, only a system able to exploit a range of information about the resources it integrates may provide scientists the execution plans best suitable for ones queries. Despite many experiments on NCBI data sources, there is yet much data to explore. A result of this comparison will give hints on improvement of the automated linking mechanisms. The presented research is only a starting point of understanding Web life sciences sources and their relationships with one another.

Future work concentrates on both the extension and generalization of the set of properties and on the usage of the presented properties for different scenarios.

Additionally, we plan to extend our model by allowing other distributions of links, by including multiple sources for individual scientific entities, and by considering more complex link structures, including ordering of paths.

EXPLOITING THE RESULTS

The results obtained can be used to develop an integrated platform that guides the scientists to choose the resources they could exploit. The scientist would be able to select his own resources, or the system can automatically select the resources for the scientist based on his query. The results can also be used in data curation. For instance, the results of this comparison will give hints to NCBI so that they can improve their automated linking mechanisms.

FUTURE WORK

The extension of this project would be

- Looking at attributes for all the data collected so far and then performing a complete analysis.
- Collecting data using DB2 Discovery Link, which creates wrappers for different data sources.

- Looking at an interesting property of ordering paths. This gives us estimation whether ordering of paths makes any difference in the data collection process.
 - OMIM - Nucleotide - Protein - PubMed
 - OMIM - Protein - Nucleotide – PubMed
- Finally perform combined analysis for all the datasets collected

FUTURE DIRECTIONS OF THE PROJECT

The results obtained in this study can be used in other projects.

- Equivalence results from Overlap Detection could be used for metadata for the capability maps.
- The queries can be used to look at some pipelines in BQL
- Capabilities can be exploited in the algorithm that ranks the path

ACKNOWLEDGEMENTS

- This research is partially supported by NSF grant IIS 02230042 and NIH National Library of Medicine grant R03 LM008046-01.

- I wish to thank Dr. David Lipman of NCBI for sharing his expertise on NCBI data sources
- Damayanti Gupta for early data collection
- Dr. Marta Janer and Dr. Michael Jazwinski for identifying relevant keywords.

REFERENCES

1. Z. Lacroix, L. Raschid, and B. Eckman (2004) “Exploiting Biomolecular Source Capabilities for Query Optimization” To appear in the *Journal of Bioinformatics and Computational Biology*.
2. Z. Lacroix, L. Raschid and M-E. Vidal (2004) “Links and Paths through Life Sciences Data Sources” In *Proc. International Workshop on Data Integration in the Life Sciences, Leipzig, Germany (to appear in the Springer-Verlag Lecture Notes in Computer Science)*.
3. B. Eckman, K. Deutsch, M. Janer, Z. Lacroix and L. Raschid (2003) “A Query Language to Support Scientific Discovery” In *Proc. 2nd IEEE International Computer Society*

APPENDIX

Keywords for Aging

Parkinson disease	Yeast
Alzheimer disease	Wrn
Aging	Blm
Ageing	Xpd
Myocardial infarction	Recql
Mi	Klotho
Acute mi	P
Osteoporosis	P shc
Impotence	Ku
Arthritis	Glucocorticoids
Degenerative joint disease	Dhea
Osteoarthritis	Dheas
Prostate	Immunosenescence
Huntington's disease	Senescence
Hand grip strength	Centenarian
Renal blood flow	Nonagenarian
Maximum breathing capacity	Leptin
Maximum work rate	Stress
Maximum oxygen uptake	Stress resistance
Resting cardia index	Life extension
Nerve conductance	Successful aging
Dietary restriction	Healthy aging
Food restriction	Adl
Caloric restriction	Iadl
Calorie restriction	Mmse
Resting metabolic rate	Neuropeptide y
Total energy expenditure	Apoe
Blood glucose	Fas
Diabetes	Proiomelancortin
Glycation	Growth hormone
Rage	Igf
Ros	Insulin
Werner's syndrome	Dwarf mouse
Hutchinson gilford syndrome	Biodemography
Progeria	Cardiovascular
Progeroid syndrome	disease
Life span	Drosophila
Saccharomyces	C. Elegans

Keywords for Cancer

Tumor	Hepatocarcinoma
Cancer	Hepatocellular carcinoma
Oncogene	Hepatoma
Suppressor	Histoma
Acidophil adenoma	Hodgkin's disease
Acoustic neuroma	Homologous tumour
Actinomycind	Hydatidiform mole
Adenocarcinoma	Hyperfractionation
Adenolymphoma	Hypernephroma
Adenoma	Immunochemotherapy
Adenomatosis	Infiltrating cancer
Adjuvant chemotherapy	Infiltrating ductal cell carcinoma
Adjuvant therapy	Insulinoma
Adrenal gland tumour	Interstitial radiation therapy
Aflatoxin	Isolated limb perfusion
Agrobacterium tumifaciens	Juvenile angiofibroma
Alkylating agent	Kaposi sarcoma
Alopecia	Kaposi's sarcoma
Ameloblastoma	Keratoses
Anaplasia	Kidney tumour
Anaplastic	Large cell lymphoma
Anaplastic carcinoma of the thyroid	Lipoma
Angiofibroma	Liver cancer
Angioma	Liver metastases
Angiosarcoma	Localised
Anthracycline	Long term survival
Anticarcinogen	Lucke carcinoma
Antiemetic	Lymphadenoma
Antitumour	Lymphangioma
Apudoma	Lymphocytic leukaemia
Argentaffinoma	Lymphocytic lymphoma
Arrhenoblastoma	Lymphocytoma
Ascites tumour	Lymphoedema
Ascitic tumour	Lymphoma
Astroblastoma	Malignant
Astrocytoma	Malignant melanoma
Ataxia telangiectasia	Malignant mesothelioma
Atrial myxoma	Malignant teratoma
Bacteriophytoma	Mastocytoma
Basal cell carcinoma	Medulloblastoma
Bence jones protein	Melanoma
Benign	Meningioma
Benign tumour	Mesothelioma
Biological response modifier	Metaplasia
Biological therapy	Metastases
Bone cancer	Metastases to the liver
Bone marrow suppression	Metastasis
Bone marrow transplantation	Metastasize
Bone tumour	Metastatic spread
Brain metastases	

Total Overlap Results for Diabetes

(P1) OMIM → PUBMED	Capability1	Capability 2	Capability 3
No entries	43,890	43747	44,037
Size	724KB	731KB	517KB
Time	30Min.33sec	2hr: 50min	2m: 57sec
No of failing entries	0	58	0
(number and % of) C1 overlapping with ...	43,890 100%	35,344 100%	43,879 99.64%
(number and % of) C2 overlapping with ...	35,344 80.52%	35,344 100%	35,333 80.23%
(number and % of) C3 overlapping with ...	43,879 99.97%	35,333 99.96%	44,037 100%

(P2) OMIM→ NUCLEOTIDE→ PUBMED	Capability1	Capability 2	Capability 3
No entries	42,969	43,090	43,581
Size	6843KB	6849KB	1692KB
Time	14Hr.24Min	12hr:49 min	39m:02sec
No of failing entries	0	0	0
(Number and % of) C11 overlapping with ...	42,969 100%	42968 99.71%	41110 94.33%

(Number and % of) C2 overlapping with ...	42968 99.99%	43,090 100%	41167 94.46%
(Number and % of) C3 overlapping with ...	41110 95.67%	41167 95.53%	43,581 100%

(P1) OMIM →Protein→PUBMED	Capability1	Capability 2	Capability 3
No entries	59,959	51,906	49,719
Size	4690KB	4696KB	1551KB
Time	7Hr.52sec	22hr: 82min	48m: 40 sec
No of failing entries	0	0	0
(Number and % of) C1 overlapping with ...	59,959 100%	51759 99.71%	47354 95.24%
(Number and % of) C2 overlapping with ...	51759 86.32%	51,906 100%	47402 95.33%
(Number and % of) C3 overlapping with ...	47354 78.97%	47402 91.32%	49,719 100%

Implementation1	Path (P1)	Path (P2)	Path (P3)
No entries	43890	42969	59959
Size	724KB	6,843KB	4690KB
Time	30Min.33sec	14Hr.24 min	7Hr.50Min
No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	43890 100%	11097 25.82%	13162 21.95%
(Number and % of) P2 overlapping with ...	11097 25.28%	42969 100%	41974 70.00%
(Number and % of) P3 overlapping with ...	13162 29.98%	41974 97.68%	59959 100%

Implementation2	Path (P1)	Path (P2)	Path (P3)
No entries	43747	43090	51906
Size	731KB	6849KB	4696KB
Time	2hr: 50min	13hr: 49min	23hr: 12min
No of failing entries	58	0	0
(Number and % of) P1 overlapping with ...	35,344 100%	10314 23.93%	11876 22.87%
(Number and % of) P2 overlapping with ...	10314 29.18%	43090 100%	42149 81.20%
(Number and % of) P3 overlapping with ...	11876 33.60%	42149 97.81%	51906 100%

Capability3	Path (P1)	Path (P2)	Path (P3)
No entries	44,037	43,581	49,719
Size	517KB	1692KB	1551KB
Time	2m: 57sec	39m:02sec	48m:40 sec
No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	44,037 100%	10855 24.75%	12079 24.29%
(Number and % of) P2 overlapping with ...	10,855 24.64%	43,581 100%	39,523 79.49%
(Number and % of) P3 overlapping with ...	12,079 27.42%	39,523 90.68%	44,037 100%

Total Overlap Results for Aging

(P1) OMIM → PUBMED	Capability1	Capability 2	Capability 3
No entries	48393	48398	48393
Size	888KB	896KB	591KB
Time	31Min.64sec	2hr: 47min	8m: 36sec
No of failing entries	0	0	0
(number and % of) C1 overlapping with ...	48393 100%	48383 99.96%	48378 99.96%
(number and % of) C2 overlapping with ...	48383 99.97%	48398 100%	48393 100%
(number and % of) C3 overlapping with ...	48378 99.96%	48393 99.98%	48393 100%

(P2) OMIM → NUCLEOTIDE → PUBMED	Capability1	Capability 2	Capability 3
No entries	51712	51855	51474
Size	9509KB	9528KB	2953KB
Time	10Hr.41 min	19hr: 03min	1hr:12min
Nb of failing entries	0	0	0
(number and % of) C1 overlapping with ...	51712 100%	51706 99.71%	51375 99.80%
(number and % of) C2 overlapping with ...	51706 99.86%	51855 100%	51468 99.98%

(number and % of) C3 overlapping with ...	51375 99.34%	51468 99.25%	51474 100%
---	-----------------	-----------------	---------------

(P1) OMIM →Protein→ PUBMED	Capability1	Capability 2	Capability 3
No entries	60129	61260	60938
Size	5508KB	5560KB	2052KB
Time	10Hr.30Min	11hr: 30min	57m: 11 sec
No of failing entries	0	0	0
(Number and % of) C1 overlapping with ...	60129 100%	60115 98.13%	60077 98.58%
(Number and % of) C2 overlapping with ...	60115 99.97%	61260 100%	60924 99.97%
(Number and % of) C3 overlapping with ...	60077 99.91%	60924 99.45%	60938 100%
Capability1	Path (P1)	Path (P2)	Path (P3)
No entries	48393	51712	60129
Size	888KB	9509KB	5508KB
Time	31Min.64sec	10Hr.41 min	10Hr.30Min
No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	48393 100%	13803 26.69%	15563 25.88%
(Number and % of) P2 overlapping with ...	13803 28.52%	51712 100%	49543 82.39
(Number and % of) P3 overlapping with ...	15563 32.15	49543 95.80%	60129 100%

Capability2	Path (P1)	Path (P2)	Path (P3)
No entries	48398	51855	61260
Size	896KB	9528KB	5560KB
Time	2hr: 47min	19hr: 03min	11hr: 30min
No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	48398 100%	13816 26.64%	15820 25.82%
(Number and % of) P2 overlapping with ...	13816 28.54%	51855 100%	50488 82.41%
(Number and % of) P3 overlapping with ...	15820 32.68%	50488 97.36%	61260 100%

Capability3	Path (P1)	Path (P2)	Path (P3)
No entries	48393	51474	60938
Size	591KB	2953KB	2052KB
Time	8m: 36sec	1hr:12min	57m: 11 sec

No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	48393 100%	13772 26.75%	15775 25.88%
(Number and % of) P2 overlapping with ...	13772 28.45%	51474 100%	50002 82.05%
(Number and % of) P3 overlapping with ...	15775 32.59%	50002 97.14%	60938 100%

Total Overlap Results for Cancer

(P1) OMIM → PUBMED	Capability1	Capability 2	Capability 3
No entries	56315	56315	56532
Size	1133KB	1150KB	810KB
Time	50Min.34sec c	61min: 13sec	10m: 10sec
No of failing entries	0	0	0
(number and % of) C1 overlapping with ...	56315 100%	56315 100%	56266 99.52%
(number and % of) C2 overlapping with ...	56315 100%	56315 100%	56266 99.52%
(number and % of) C3 overlapping with ...	56266 99.91%	56266 99.91%	56532 100%

(P2) OMIM → NUCLEOTIDE → PUBMED	Capability1	Capability 2	Capability 3
No entries	54487	54607	52488
Size	9817KB	9832KB	3709KB
Time	12Hr.05 min	21hr: 08min	1hr:31min
Nb of failing entries	0	0	0
(number and % of) C1 overlapping with ...	54487 100%	54463 99.73%	52157 99.36%
(number and % of) C2 overlapping with ...	54463 99.95%	54607 100%	52179 99.41%
(number and % of) C3 overlapping with ...	52157 95.72%	52179 95.55%	52488 100%

(P1) OMIM →Protein→ PUBMED	Capability1	Capability 2	Capability 3
No entries	62686	63367	60033
Size	5820KB	5845KB	2168KB
Time	8Hr.43Min	13hr: 18min	1hr: 12 min
No of failing entries	0	0	0
(Number and % of) C1 overlapping with ...	62686 100%	62671 98.90%	59342 98.84%
(Number and % of) C2 overlapping with ...	62671 99.97%	63367 100%	59920 99.81%
(Number and % of) C3 overlapping with ...	59342 94.66%	59920 94.56%	60033 100%

Capability1	Path (P1)	Path (P2)	Path (P3)
No entries	56315	54487	62686
Size	1133KB	9817KB	5820KB
Time	50Min.34se c	12Hr.05 min	8Hr.43Min
No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	56315 100%	15243 27.97%	17113 27.29%
(Number and % of) P2 overlapping with ...	15243 27.06%	54487 100%	51839 82.69%
(Number and % of) P3 overlapping with ...	17113 30.38%	51839 95.14%	62686 100%

Capability2	Path (P1)	Path (P2)	Path (P3)
No entries	56315	54607	63367
Size	1150KB	9832KB	5845KB
Time	1hr: 1min	21hr: 08min	13hr: 18min
No of failing entries	0	0	0
(Number and % of) P1 overlapping with ...	56315 100%	15260 27.94%	17253 27.22%
(Number and % of) P2 overlapping with ...	15260 27.09%	54607 100%	52458 82.78%
(Number and % of) P3 overlapping with ...	17253 30.63%	52458 96.06%	63367 100%

Capability3	Path (P1)	Path (P2)	Path (P3)
No entries	56532	52488	60033
Size	810KB	3709KB	2168KB
Time	10m: 10sec	1hr: 31min	1hr: 12 min
No of failing entries	0	0	0

(Number and % of) P1 overlapping with ...	56532 100%	14899 28.38%	16610 27.66%
(Number and % of) P2 overlapping with ...	14899 26.35%	52488 100%	49342 82.19%
(Number and % of) P3 overlapping with ...	16610 29.38%	49342 94%	60033 100%

•