

# Exploring and Exploiting the Biological Maze

*Presented By*

Vidyadhari Edupuganti

*Advisor*

Dr. Zoe Lacroix

# Motivation

- An abundance of biological data sources contain data about scientific entities, such as genes and sequences
- Scientists are interested in exploring relationships between scientific objects
- Explore a multiplicity of inter-related sources
- E.g: For looking at protein sequences one could either go to the Protein database of NCBI ,Swiss-Port or EMBL.

# Continued ...

- Each data source has different capabilities
  - Data Access
  - Links
  - Navigation
  - Analysis
- Different data sources and different capabilities raise various challenges to scientists
  - Same Entries
  - Quality
  - Cost (time / space)

# Common Queries

- Query1:
  - Retrieve all sequences linked to a disease condition
    - NCBI Protein
    - SwissProt
    - TrEMBL
- Query2 :
  - Retrieve all genes and its citations linked to a disease condition
    - NCBI Nucleotide → PubMed
    - OMIM → PubMed
    - Gene → PubMed

# Different sources / Different answers

Query 1: Retrieve all sequences linked to “cancer”

- Data Source : NCBI Protein
  - Entries Retrieved : 126174
- Data Source : SwissProt
  - Entries Retrieved : 148516
- Data Source : TrEMBL
  - Entries Retrieved : 1067463

# Same sources/Same capability/Different answers

- Query2 : Retrieve all genes and its citations linked to “diabetes mellitus”
  - NCBI Nucleotide → PubMed
    - Genes: 743
    - Citations: 4277
    - Capability: PubMed Link
  - OMIM → PubMed
    - Genes: 296
    - Citations: 6906
    - Capability: PubMed Link

# Continued ...

- Query2 : Retrieve all genes and its citations linked to “diabetes mellitus”
  - Gene → PubMed
    - Genes: 228
    - Citations: 4147
    - Capability: PubMed Link

# Objective

- Demonstrate that the data collection process depends on two variables
  - Data Source Selected
  - Type of Capability chosen for that source
- Collect data using a script
- Deploy the same using Discovery Link

# Querying Data Sources

- Analysis done on four databases of NCBI
  - OMIM
  - PubMed
  - Nucleotide
  - Protein
- Choose OMIM source, which contains information related to human genetic diseases
- Retrieve all PubMed citations related to those diseases

# Process

- Start with a keyword at OMIM and retrieve all PubMed citations following three paths.
- Three different implementations
  - Link (ESearch utility of NCBI)
  - Parse (Parse each entry to retrieve citations)
  - All (Query database for all the records)
  - E.g.: [NCBI Homepage](#)
- Does choice of paths and the type of retrieval has an impact on the result ?

# Continued...

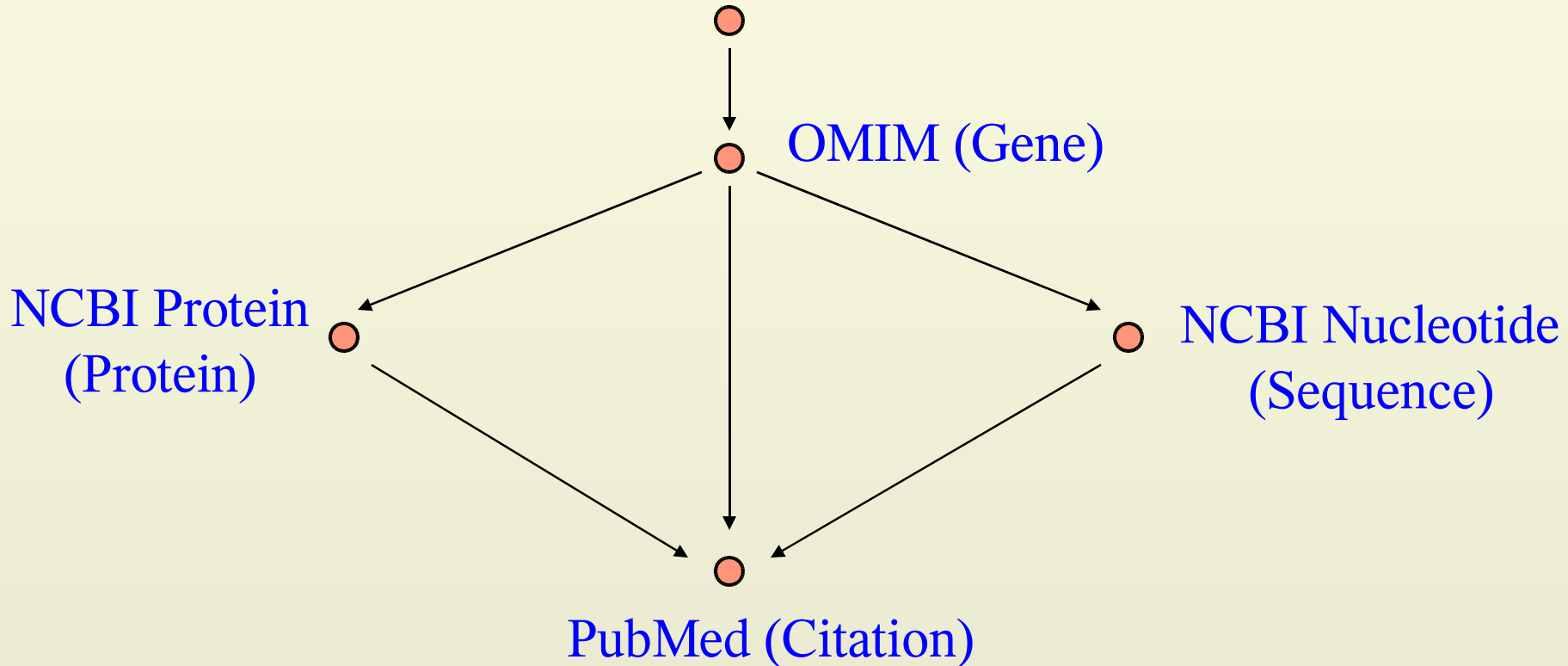
- Three paths were followed to retrieve PubMed citations
- Three disease conditions chosen
  - **Ageing** (71 Keywords)
  - **Diabetes** (11 Keywords)
  - **Cancer** (391 Keywords)
- Keywords were given by domain experts
- Query : *Return all citations of PubMed that are linked to an OMIM entry that is related to some disease or condition*

# Implementation

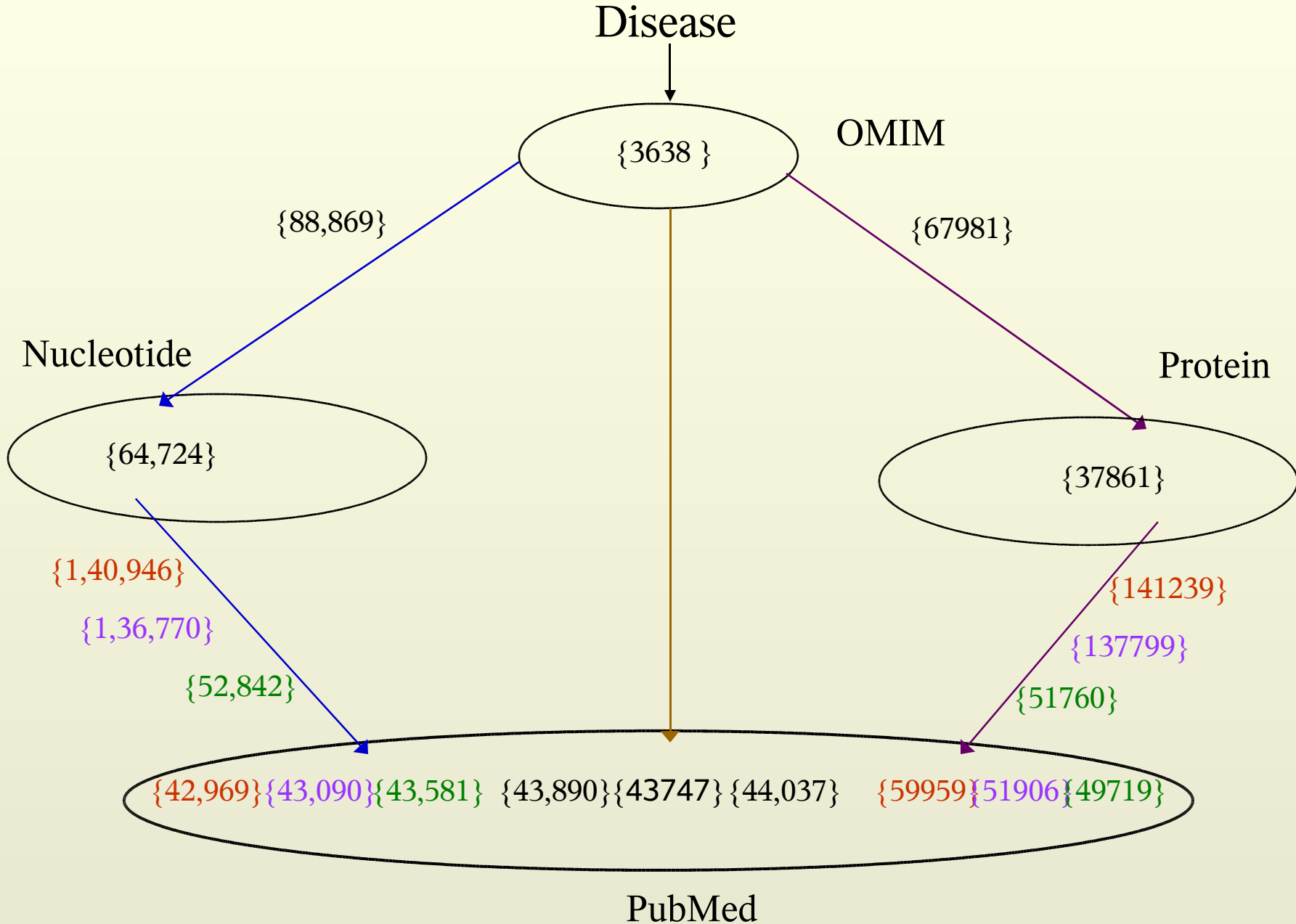
- The execution of this query explores all paths from the starting objects of OMIM, and ended in objects of the end point source PubMed
- A wrapper implemented in Java, made successive calls to the E-Link interface of NCBI

# Source graph for NCBI data sources

- Disease/Condition



# Results for Diabetes



# Results for Aging

PubMed Citations	Path1 (P1)	Path2 (P2)	Path3 (P3)
Implementation 1	48393	51712	60129
Implementation 2	48398	51855	61260
Implementation 3	48393	51474	60938

# Results for Cancer

PubMed Citations	Path1 (P1)	Path2 (P2)	Path3 (P3)
Implementation 1	56315	54487	62686
Implementation 2	56315	54607	63367
Implementation 3	56532	52488	60033

# Steps for Analysis

- Remove Duplicates
- Cost ( Time/Space)
- Calculate Overlap
  - Between Paths
  - Between Implementations
- Choosing among a good set of Paths/implementations to maximize benefit and minimize cost.

# Overlap detection

- Check for Equivalence
  - Two sources overlap
  - Sources don't overlap
- Choose Better Path
- Choose Better Implementation

# Detection Process

- Calculate overlap among different paths
  - Probability that citations retrieved from one path are also retrieved by the path compared
- Calculate overlap among different implementations
  - Probability that citations retrieved from one implementation of Path1 are also retrieved by the other implementation compared of Path1

# Overlap for Disease Condition “Diabetes”

		P1	P2	P3
IMP1	P1	100%	25.82%	21.95%
	P2	25.28%	100%	70.00%
	P3	29.98%	97.68%	100%
IMP2	P1	100%	23.93%	22.87%
	P2	29.18%	100%	81.20%
	P3	33.60%	97.81%	100%
IMP3	P1	100%	24.75%	24.29%
	P2	24.64%	100%	79.49%
	P3	27.42%	90.68%	100%

# Continued ...

		<b>IMP1</b>	<b>IMP2</b>	<b>IMP3</b>
<b>P1</b>	<b>IMP1</b>	<b>100%</b>	<b>100%</b>	<b>99.64%</b>
	<b>IMP2</b>	<b>80.52%</b>	<b>100%</b>	<b>80.23%</b>
	<b>IMP3</b>	<b>99.97%</b>	<b>99.96%</b>	<b>100%</b>
<b>P2</b>	<b>IMP1</b>	<b>100%</b>	<b>99.71%</b>	<b>94.33%</b>
	<b>IMP2</b>	<b>99.99%</b>	<b>100%</b>	<b>94.46%</b>
	<b>IMP3</b>	<b>95.67%</b>	<b>95.53%</b>	<b>100%</b>
<b>P3</b>	<b>IMP1</b>	<b>100%</b>	<b>99.71%</b>	<b>95.24%</b>
	<b>IMP2</b>	<b>86.32%</b>	<b>100%</b>	<b>95.33%</b>
	<b>IMP3</b>	<b>78.97%</b>	<b>91.32%</b>	<b>100%</b>

# Overlap for Disease Condition “Aging”

		P1	P2	P3
IMP1	P1	100%	26.69%	25.88%
	P2	28.52%	100%	82.39
	P3	32.15	95.80%	100%
IMP2	P1	100%	26.64%	25.82%
	P2	28.54%	100%	82.41%
	P3	32.68%	97.36%	100%
IMP3	P1	100%	26.75%	25.88%
	P2	28.45%	100%	82.05%
	P3	32.59%	97.14%	100%

# Continued ...

		IMP1	IMP2	IMP3
P1	IMP1	100%	99.96%	99.96%
	IMP2	99.97%	100%	100%
	IMP3	99.96%	99.98%	100%
P2	IMP1	100%	99.71%	99.80%
	IMP2	99.86%	100%	99.98%
	IMP3	99.34%	99.25%	100%
P3	IMP1	100%	98.13%	98.58%
	IMP2	99.97%	100%	99.97%
	IMP3	99.91%	99.45%	100%

# Overlap for Disease Condition “Cancer”

		P1	P2	P3
IMP1	P1	100%	27.97%	27.29%
	P2	27.06%	100%	82.69%
	P3	30.38%	95.14%	100%
IMP2	P1	100%	27.94%	27.22%
	P2	27.09%	100%	82.78%
	P3	30.63%	96.06%	100%
IMP3	P1	100%	28.38%	27.66%
	P2	26.35%	100%	82.19%
	P3	29.38%	94%	100%

# Continued ...

		IMP1	IMP2	IMP3
<b>P1</b>	IMP1	<b>100%</b>	<b>100%</b>	<b>99.52%</b>
	IMP2	<b>100%</b>	<b>100%</b>	<b>99.52%</b>
	IMP3	<b>99.91</b>	<b>99.91</b>	<b>100%</b>
<b>P2</b>	IMP1	<b>100%</b>	<b>99.73%</b>	<b>99.36%</b>
	IMP2	<b>99.95%</b>	<b>100%</b>	<b>99.41%</b>
	IMP3	<b>95.72%</b>	<b>95.55%</b>	<b>100%</b>
<b>P3</b>	IMP1	<b>100%</b>	<b>98.90%</b>	<b>98.84%</b>
	IMP2	<b>99.97%</b>	<b>100%</b>	<b>99.81%</b>
	IMP3	<b>94.66%</b>	<b>94.56%</b>	<b>100%</b>

# Comparisons

- Different Implementations give different results
- Three different paths from OMIM to PubMed return significantly different number of distinct objects in PubMed
- Ascending order of distinct links
  - Paths
  - Implementations

# Conclusions

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

# Exploiting Results

- Develop Integrated Platform
  - Guide the Scientists to choose the resources they could exploit
    - Based on their selection
    - Automatically select resources
      - Ex: Discovery Link
- Data Curation

# Future Work

- Look at attributes
- Collect data using DB2 Discovery Link
- Look at Ordering of paths
  - OMIM - Nucleotide - Protein - PubMed
  - OMIM - Protein - Nucleotide – PubMed
- Perform combined analysis for all the datasets collected

# Further Development

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

# Problems Encountered

- Initial Script from UMD
- Data previously collected
- Collection Process
- Discovery link

# Acknowledgements

- This research is partially supported by a NSF grant, and the NIH National Institute of Aging.
- We wish to thank David Lipman of NCBI for sharing his expertise on NCBI data sources
- Damayanti Gupta for early data collection, and Marta Janer
- Michael Jazwinski for identifying relevant keywords.

# References

- *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.*
- *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).*
- *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*
- *Z. Lacroix, L. Raschid and M-E. Vidal (2004) “Efficient Techniques to Explore Paths in Life Science Data Source” In Proc. International Workshop on Data Integration in the Life Sciences, Leipzig, Germany (to appear in the Springer-Verlag Lecture Notes in Computer Science).*