Searching Databases of Metabolic Pathways Using Inverted Term Lists

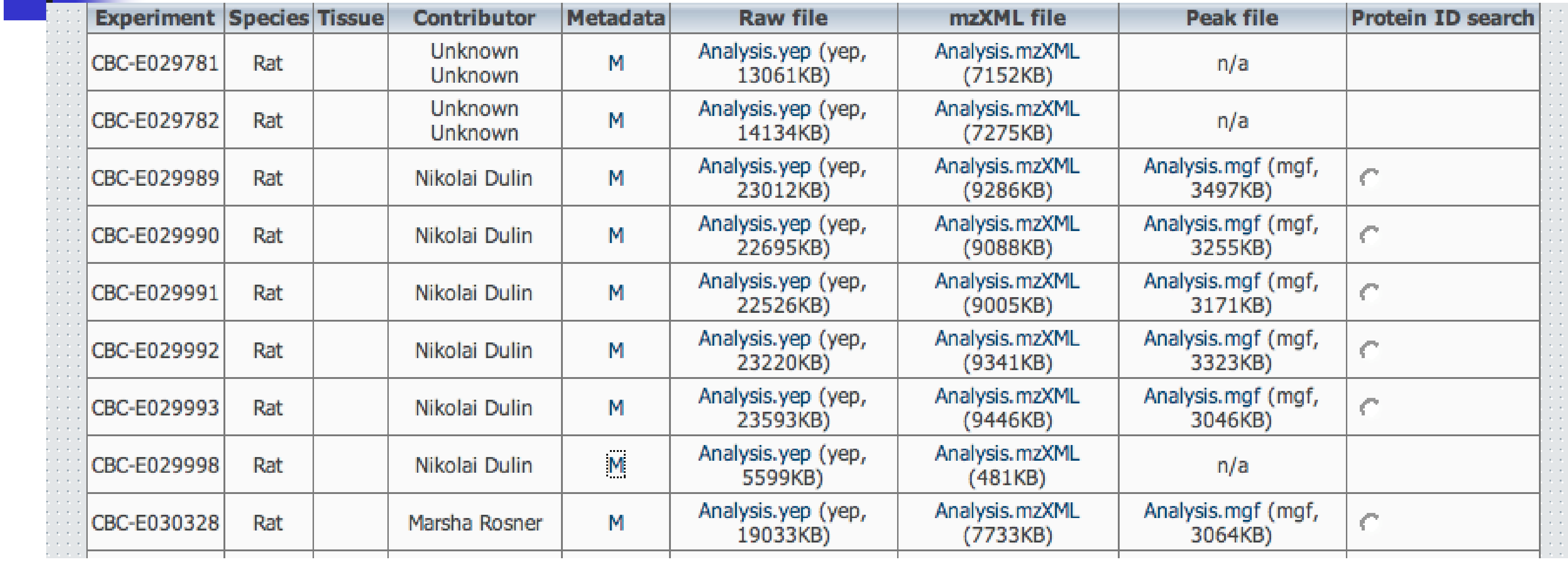
Greeshma Neglur, Robert Grossman, and Clement Yu

University of Illinois at Chicago

Natalia Maltsev

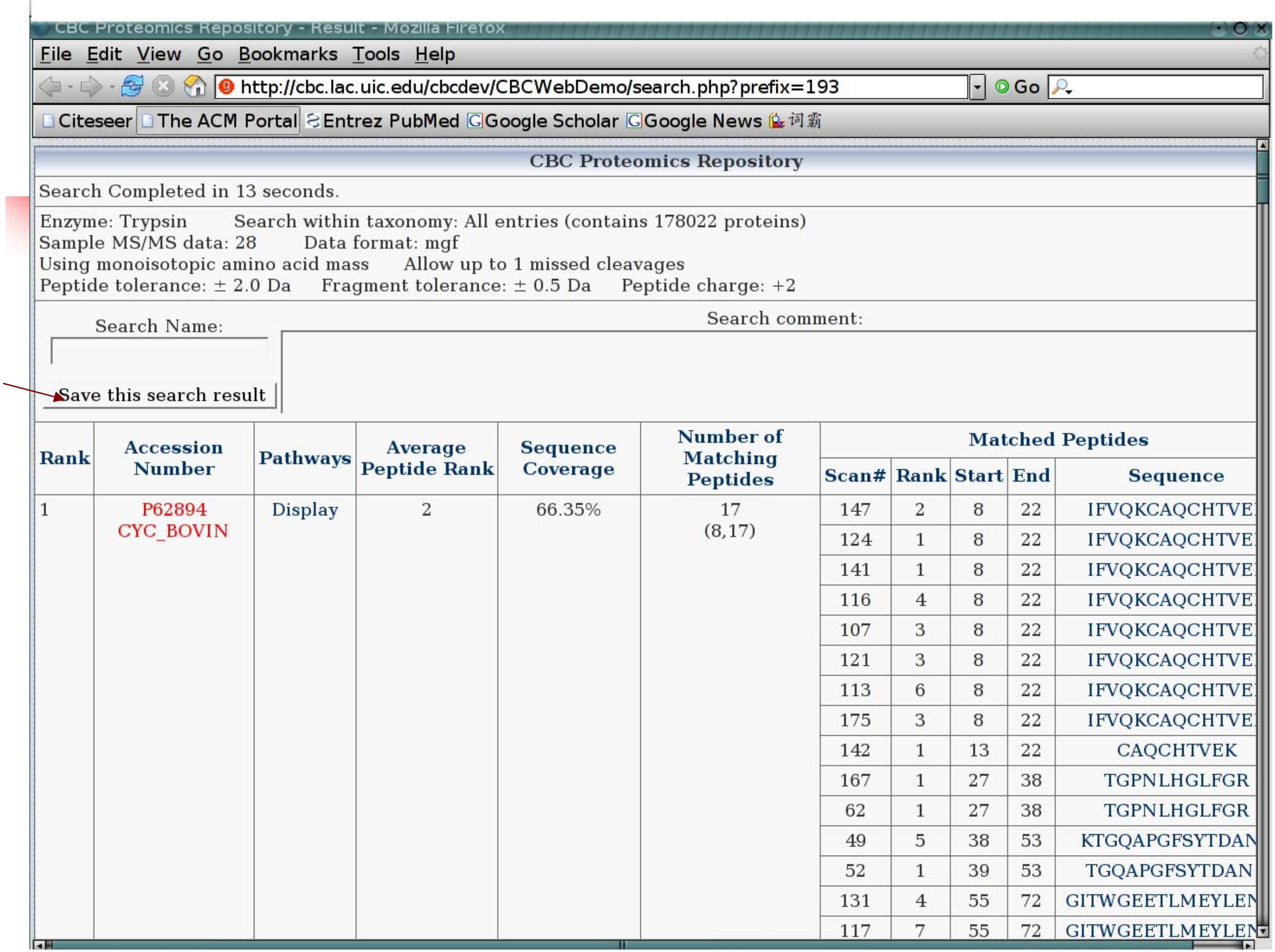
Argonne National Laboratory

Overall Goal - Add Pathway Search to CBC Proteomics Repository



Experiment	Species	Tissue	Contributor	Metadata	Raw file	mzXML file	Peak file	Protein ID search
CBC-E029781	Rat		Unknown Unknown	M	Analysis.yep (yep, 13061KB)	Analysis.mzXML (7152KB)	n/a	
CBC-E029782	Rat		Unknown Unknown	M	Analysis.yep (yep, 14134KB)	Analysis.mzXML (7275KB)	n/a	
CBC-E029989	Rat		Nikolai Dulin	M	Analysis.yep (yep, 23012KB)	Analysis.mzXML (9286KB)	Analysis.mgf (mgf, 3497KB)	○
CBC-E029990	Rat		Nikolai Dulin	M	Analysis.yep (yep, 22695KB)	Analysis.mzXML (9088KB)	Analysis.mgf (mgf, 3255KB)	○
CBC-E029991	Rat		Nikolai Dulin	M	Analysis.yep (yep, 22526KB)	Analysis.mzXML (9005KB)	Analysis.mgf (mgf, 3171KB)	○
CBC-E029992	Rat		Nikolai Dulin	M	Analysis.yep (yep, 23220KB)	Analysis.mzXML (9341KB)	Analysis.mgf (mgf, 3323KB)	○
CBC-E029993	Rat		Nikolai Dulin	M	Analysis.yep (yep, 23593KB)	Analysis.mzXML (9446KB)	Analysis.mgf (mgf, 3046KB)	○
CBC-E029998	Rat		Nikolai Dulin	M	Analysis.yep (yep, 5599KB)	Analysis.mzXML (481KB)	n/a	
CBC-E030328	Rat		Marsha Rosner	M	Analysis.yep (yep, 19033KB)	Analysis.mzXML (7733KB)	Analysis.mgf (mgf, 3064KB)	○

- Chicago Biomedical Consortium is a consortium of 3 major Chicago area universities
- This is a CBC Project to develop search engine for metabolic pathways for the CBC Proteomics Repository


CBC Proteomics Repository - Result - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cbc.lac.uic.edu/cbcdev/CBCWebDemo/search.php?prefix=193
Go
Citeseer
The ACM Portal
Entrez PubMed
Google Scholar
Google News
词霸

CBC Proteomics Repository

Search Completed in 13 seconds.

Enzyme: Trypsin Search within taxonomy: All entries (contains 178022 proteins)
Sample MS/MS data: 28 Data format: mgf
Using monoisotopic amino acid mass Allow up to 1 missed cleavages
Peptide tolerance: ± 2.0 Da Fragment tolerance: ± 0.5 Da Peptide charge: +2

Search Name:

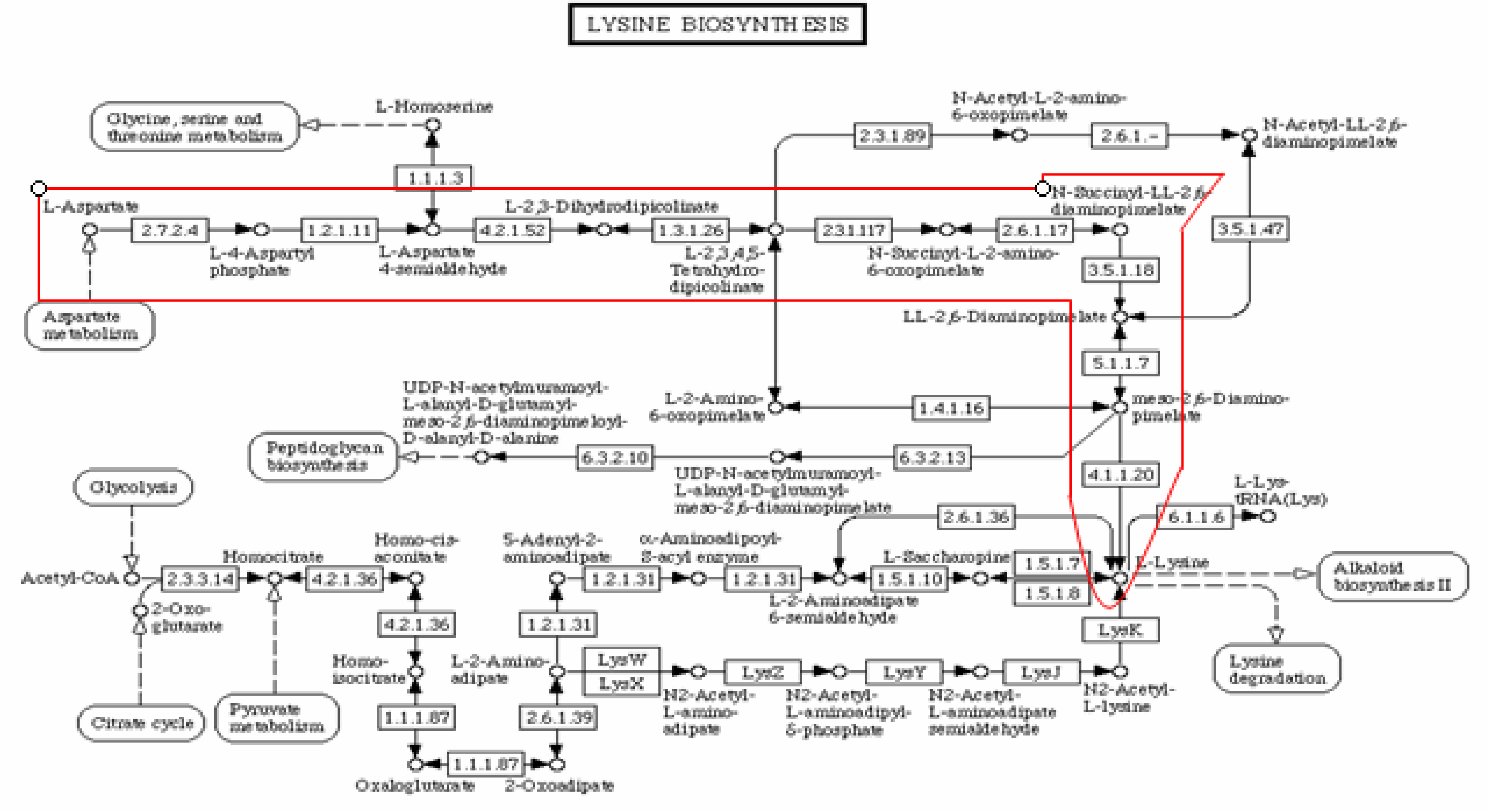
Search comment:

Save this search result

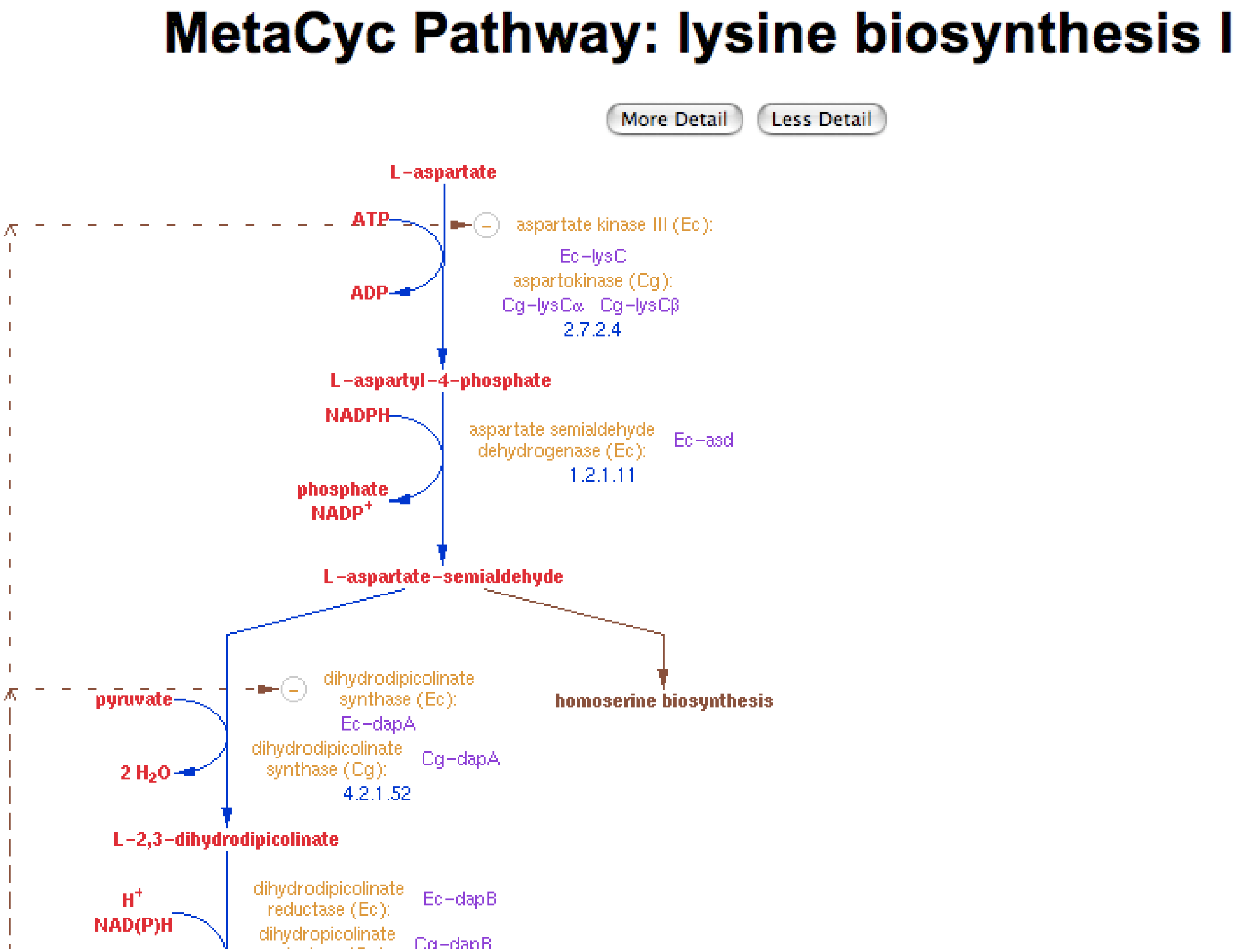
Rank	Accession Number	Pathways	Average Peptide Rank	Sequence Coverage	Number of Matching Peptides	Matched Peptides				
						Scan#	Rank	Start	End	Sequence
1	P62894 CYC_BOVIN	Display	2	66.35%	17 (8,17)	147	2	8	22	IFVQKCAQCHTVE
						124	1	8	22	IFVQKCAQCHTVE
						141	1	8	22	IFVQKCAQCHTVE
						116	4	8	22	IFVQKCAQCHTVE
						107	3	8	22	IFVQKCAQCHTVE
						121	3	8	22	IFVQKCAQCHTVE
						113	6	8	22	IFVQKCAQCHTVE
						175	3	8	22	IFVQKCAQCHTVE
						142	1	13	22	CAQCHTVEK
						167	1	27	38	TGPNLHGLFGR
						62	1	27	38	TGPNLHGLFGR
						49	5	38	53	KTGQAPGFSYTDAN
						52	1	39	53	TGQAPGFSYTDAN
						131	4	55	72	GITWGEETLMEYLEN
						117	7	55	72	GITWGEETLMEYLEN

Example: Similar Pathways Different Databases

KEGG database : Lysine biosynthesis



Example (cont’d)



MetaCyc Pathway: lysine biosynthesis I

Overview

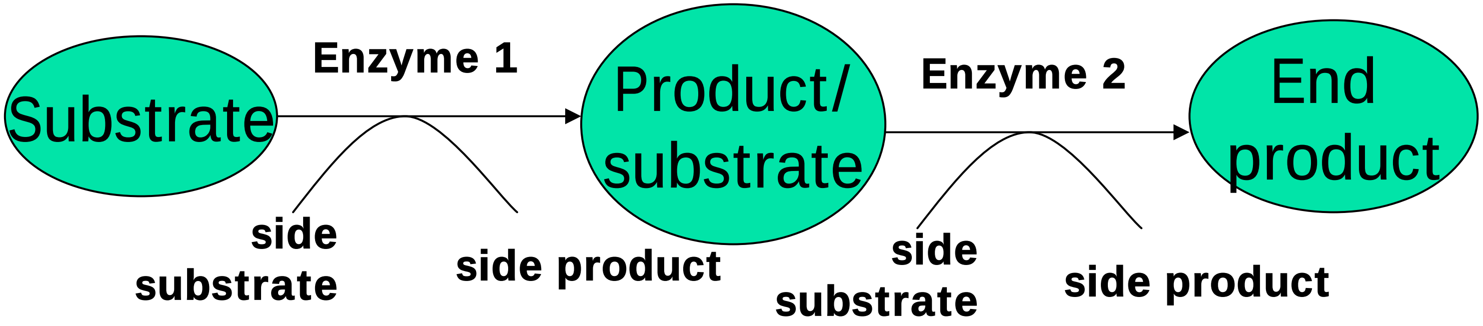
- We view metabolic pathways as labeled directed graphs where the nodes represent chemical compounds.

- We use Universal Chemical Keys or UCKs to attach unique labels to each node

- By maintaining an inverted file that indexes all pathways in a database on their edges, our algorithm finds and ranks all pathways similar to the user input query pathway in time, which is linear in the total number of occurrences of the edges in common with the query in the entire database.

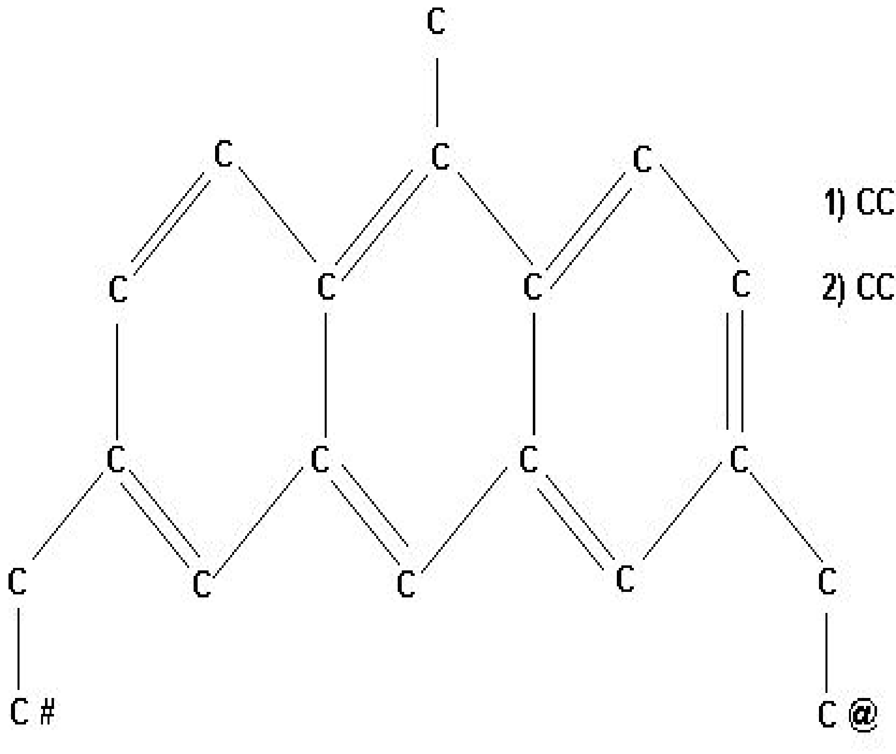
We Model Metabolic Pathways as Directed Graphs

- Definition :
  - A series of 2 or more interconnected enzyme-mediated chemical reactions that take place in a cell.
- Structure :



Chemical Compounds Mapped to Labeled Nodes

Name : 2,7-diethyl-10-methylanthracene



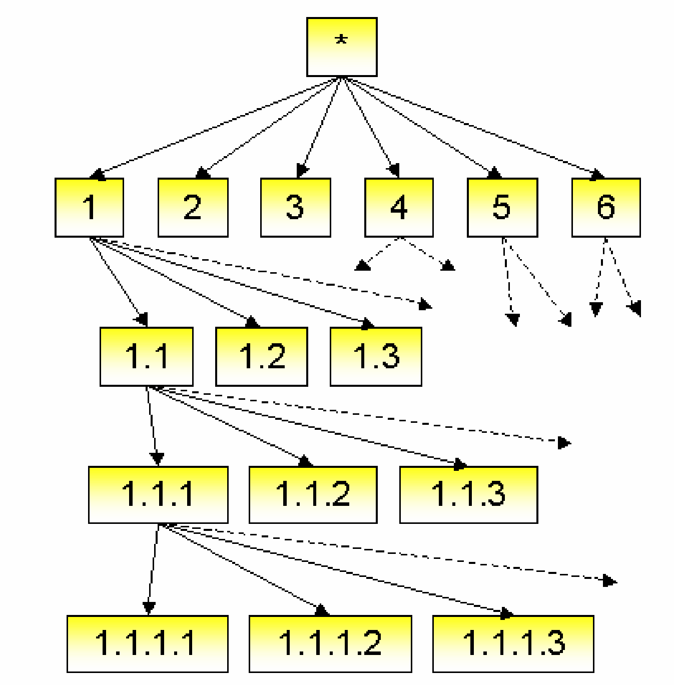
1) CCC1=CC2=CC3=CC(=CC=C3C(=C2C=C1)C)CC (starting at #)

2) CCC1=CC=C2C(=C3C=CC(=CC3=CC2=C1)CC)C (starting at @)

Enzymes Mapped to Labeled Edges

- Edges correspond to enzymes
- Each enzyme has an IUBMB EC number expressed as a string of 4 digits.

  eg : [1.2.3.4]



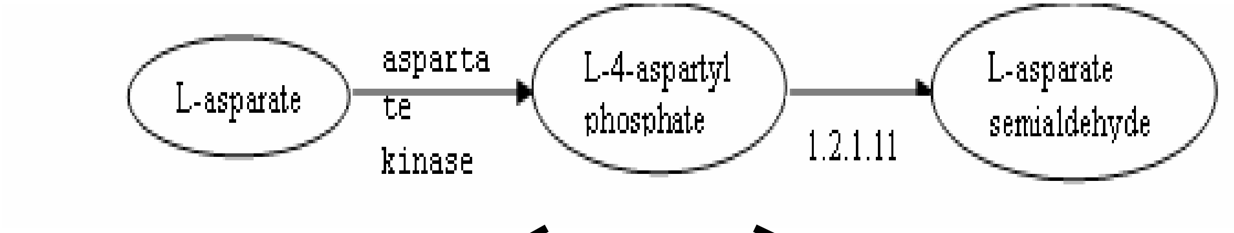
Related Work ...

- A popular XML indexing technique called HOPI provides support for path expression search with wildcards
- GraphGrep: index structure is a hash table consisting of hash values of the labeled paths and the corresponding pathways containing the labeled path
- Another approach outlined in GIndex by Han et al. uses frequent substructures as a basic indexing unit
- Different measures of node similarities include Sequence similarity, Structural similarity, Reaction/ EC similarity, Semantic similarity (comparison of gene ontology)

Idea 1: Create Uniquely Labeled Graph Associated with a Pathway

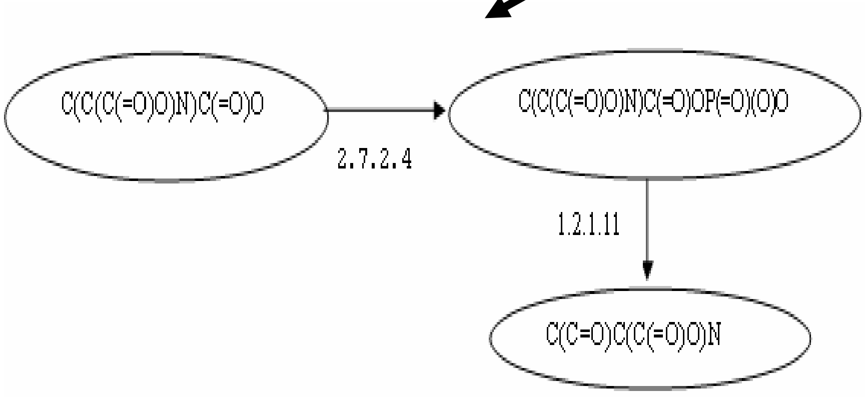
- Method 1
  - We label the nodes with Canonical SMILES string of the chemical compound associated with the node.
  - We identify all nodes whose labels are the same and associate a $G' = G / \sim$, where $\sim$ is the equivalence relation defined as follows: $u \sim v$ in case the nodes u and v in G have the same label. G’ is the uniquely labeled pathway graph
- Method 2
  - We label the nodes with the Unique Chemical Key or UCK associated with the chemical compound (DILS 05)
  - UCKs are unique but, the chemical structure cannot be recovered from them

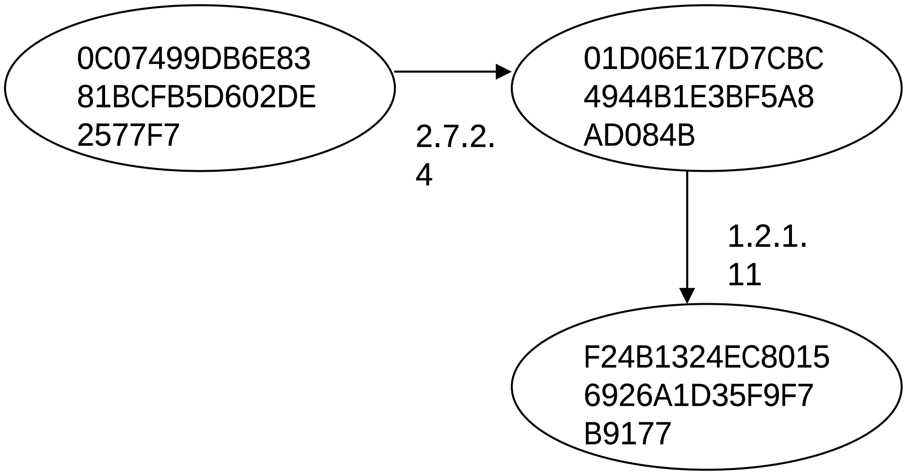
Example of uniquely labeled directed pathway graph



Using USMILES

Using UCK

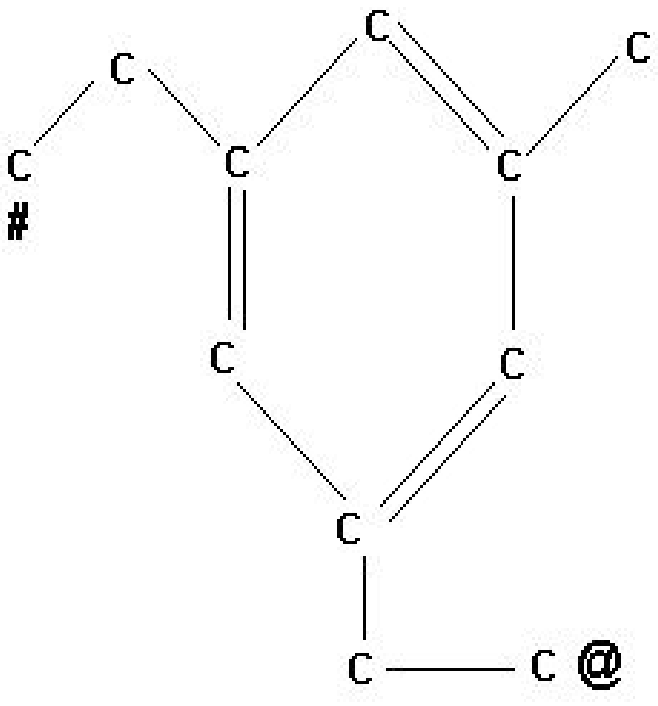



0C07499DB6E83 81BCFB5D602DE 2577F7
2.7.2. 4
01D06E17D7CBC 4944B1E3BF5A8 AD084B
1.2.1. 11
F24B1324EC8015 6926A1D35F9F7 B9177

May change the topology of the graph.

Universal Chemical Key (UCK) - Example 1

Name : 3,5-diethyl toluene



Two different Unique SMILES :

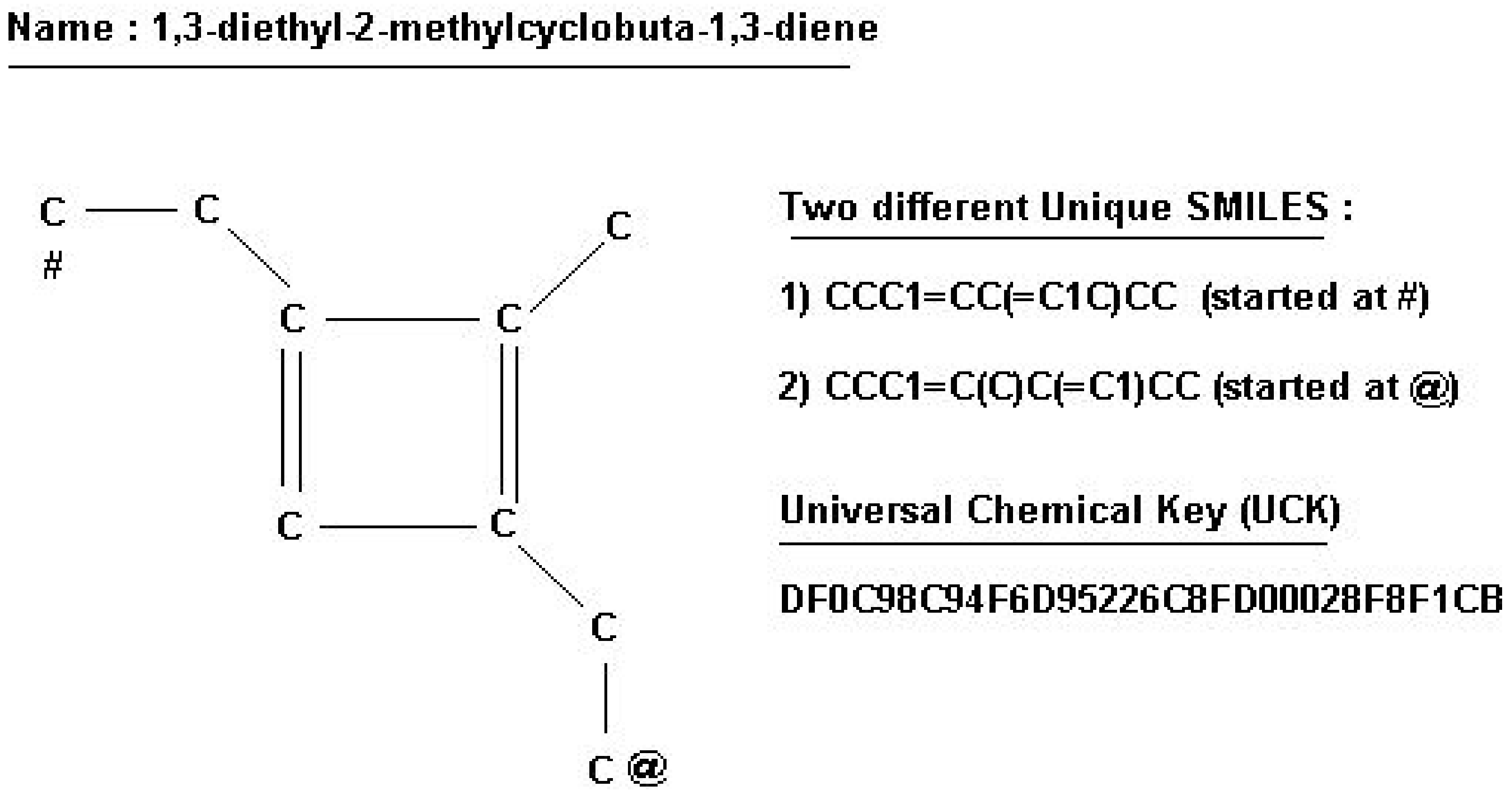
1) CCC1=CC(=CC(=C1)C)CC (started at #)

2) CCC1=CC(=CC(=C1)CC)C (started at @)

Universal Chemical Key (UCK)

85C7DC186897FD83D8ECB6B167D988BE

UCK - Example 2



Name : 1,3-diethyl-2-methylcyclobuta-1,3-diene

Two different Unique SMILES :

1) CCC1=CC(=C1C)CC (started at #)

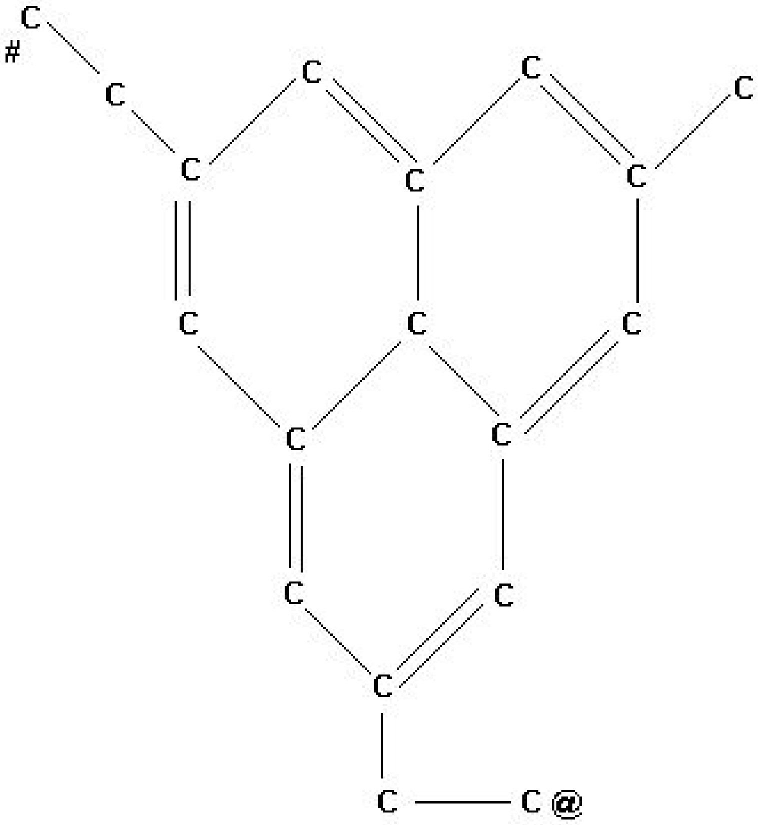
2) CCC1=C(C)C(=C1)CC (started at @)

Universal Chemical Key (UCK)

DF0C98C94F6D95226C8FD00028F8F1CB

UCK - Example 3

Name : 2,5-diethyl-8-methyl-3*H*-phenalene



Two different Unique SMILES for the graph :

1) CCC1=CC2=CC(=CC3=CC(=CC(=C1)C23)C)CC (starting at #)

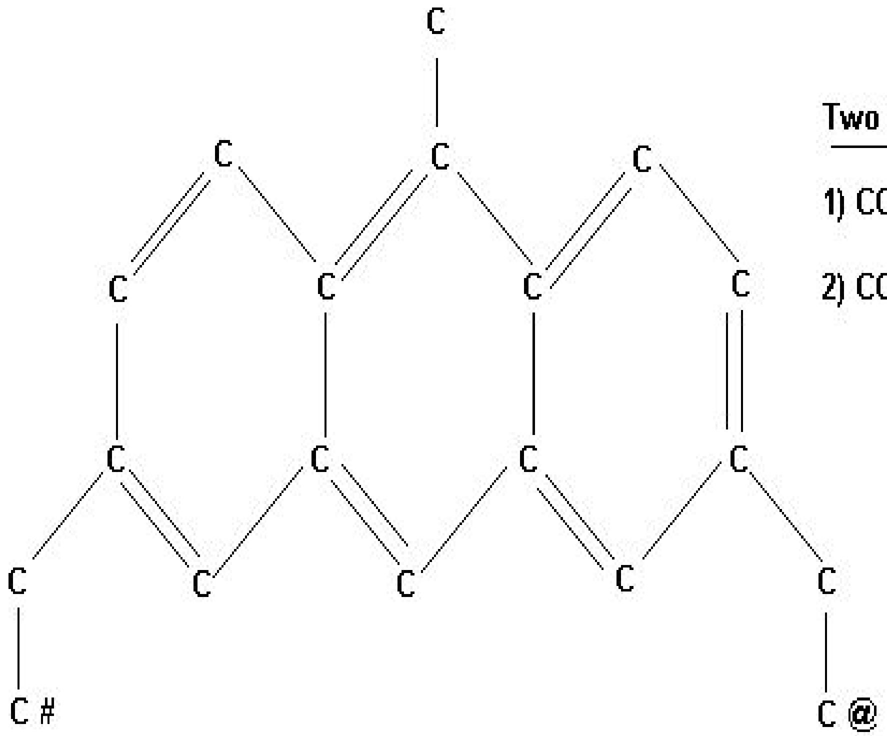
2) CCC1=CC2=CC(=CC3=CC(=CC(=C1)C23)CC)C (starting at @)

Universal Chemical Key (UCK)

EAE7F5CD89F839505ACAF3CFE040B7BF

UCK - Example 4

Name : 2,7-diethyl-10-methylanthracene



Two different Unique SMILES for the graph :

1) CCC1=CC2=CC3=CC(=CC=C3C(=C2C=C1)C)CC (starting at #)

2) CCC1=CC=C2C(=C3C=CC(=CC3=CC2=C1)CC)C (starting at @)

Universal Chemical Key (UCK)

807EC425B863D72C8897A9AC72809076

Analysis of NCI Database Using UCKs

Description	Number	Remark
Total number of chemical compounds	236,917	Some compounds have duplicate entries
Number of chem. comp. with single entry	202,384	All gave unique UCK
Number chem. comp. 2 or more entries	33,533	UCK gave same key to same compounds

Idea 2: Use Bag of Terms

	t1	t2	t3	t4	t5	t6	…
d1		1		2	1		
d2	1		3				
d3			1		1		
d4	2					2	
…							

- Basic approach - divide text into terms (e.g. words)
- Form document-term count matrix capturing frequencies of terms in data (i.e. view terms as basis for vector space)
- Normalize

Terms for Pathway Databases

- We view edges as terms; more precisely a term is an ordered-triplet consisting of a substrate, enzyme and product, which we denote as follows:

(coef) substrate : enzyme : product (term)

- represents an edge in the uniquely labeled graph of the pathway. Coefficient is the number of times edge occurs

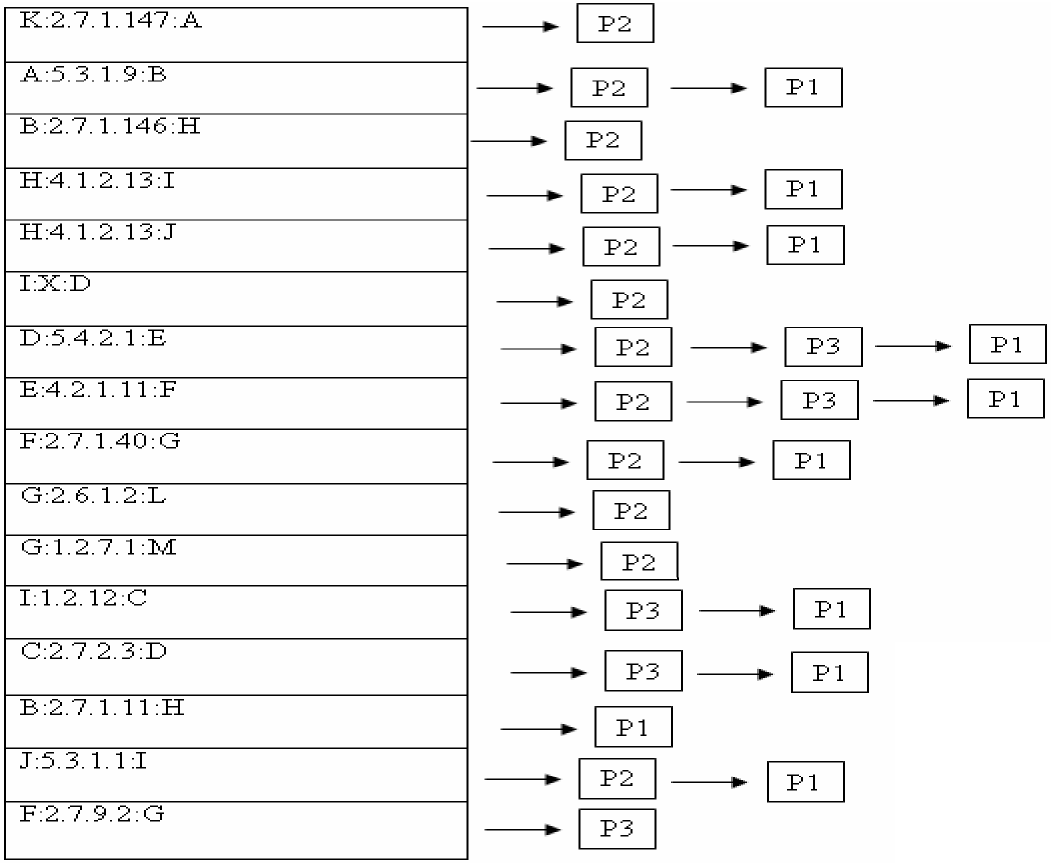
- Example

3 C(C(C(=O)O)N)C(=O)O **:** 2.7.2.4 **:** C(C(C(=O)O)N)C(=O)OP(=O)(O)O

Idea 3: Use an Inverted File to Index Pathways

- Use the following inverted file as the index structure for the pathway search system

A, B, C, … chemical compounds



K:2.7.1.147:A	P2		
A:5.3.1.9:B	P2	P1	
B:2.7.1.146:H	P2		
H:4.1.2.13:I	P2	P1	
H:4.1.2.13:J	P2	P1	
I:X:D	P2		
D:5.4.2.1:E	P2	P3	P1
E:4.2.1.11:F	P2	P3	P1
F:2.7.1.40:G	P2	P1	
G:2.6.1.2:L	P2		
G:1.2.7.1:M	P2		
I:1.2.12:C	P3	P1	
C:2.7.2.3:D	P3	P1	
B:2.7.1.11:H	P1		
J:5.3.1.1:I	P2	P1	
F:2.7.9.2:G	P3		

Similarity Functions

- Cosine Similarity: measure of number of edges in common [Salton and McGrill 1983]

$$F(Q,G) = \frac{\sum_{i}^{n} q_i G_i}{\sqrt{\sum q_i^2}\sqrt{\sum G_i^2}} \qquad (1)$$

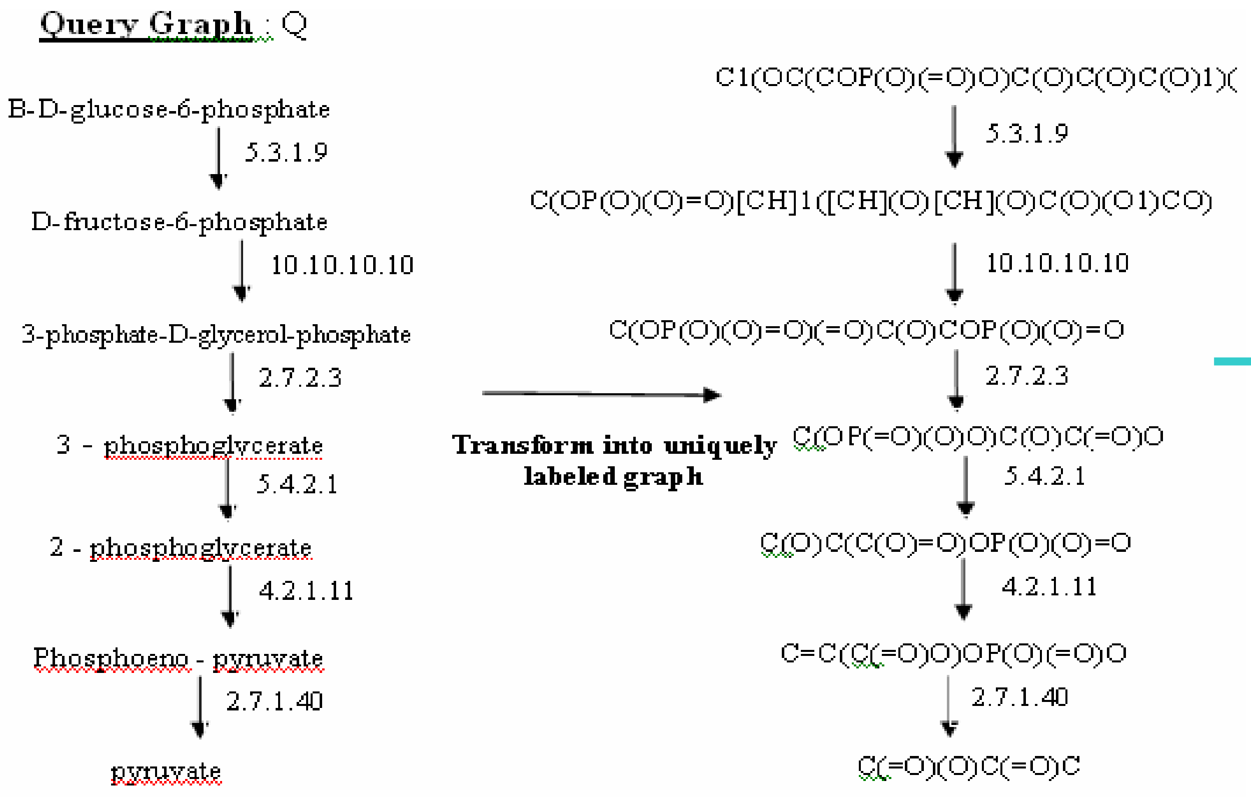
where $Q = (q_1, q_2, \ldots\ldots, q_n)$ and $G = (G_1, G_2, \ldots\ldots, G_n)$

- MCS based similarity: mcs(Q, G) is the Maximal Common Subgraph between Q and G and |G| is the size of the graph in terms of number of edges (E) in the graph.

$$Sim\,(Q, G) = \frac{|mcs(Q, G)|}{Max(|Q|\,,\,|G|)} \qquad (2)$$

Searching and computing similarity ...

- Convert the user query to uniquely labeled directed graph



Query Graph : Q

B-D-glucose-6-phosphate

5.3.1.9

D-fructose-6-phosphate

10.10.10.10

3-phosphate-D-glycerol-phosphate

2.7.2.3

3 - phosphoglycerate

5.4.2.1

2 - phosphoglycerate

4.2.1.11

Phosphoeno - pyruvate

2.7.1.40

pyruvate

Transform into uniquely labeled graph

C1(OC(COP(O)(=O)O)C(O)C(O)C(O)1)(

5.3.1.9

C(OP(O)(O)=O)[CH]1([CH](O)[CH](O)C(O)(O1)CO)

10.10.10.10

C(OP(O)(O)=O)(=O)C(O)COP(O)(O)=O

2.7.2.3

C(OP(=O)(O)O)C(O)C(=O)O

5.4.2.1

C(O)C(C(O)=O)OP(O)(O)=O

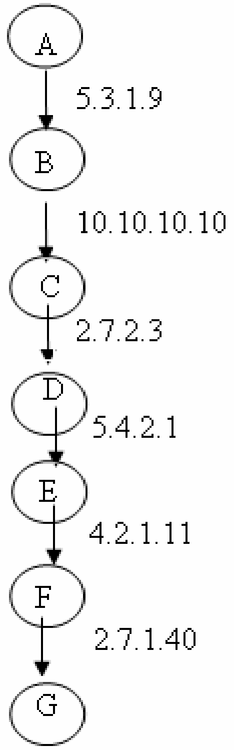
4.2.1.11

C=C(C(=O)O)OP(O)(=O)O

2.7.1.40

C(=O)(O)C(=O)C

For brevity the symbols are transformed



Searching and computing similarity …

- **Step 1** For each edge given in the query pathway; find all the database pathways that have the edge.
  - Time Complexity = O(sum over all edges in the query) n_i) = O(n)
  - For the i’th edge in the query graph, let n_i be the number of pathways that have the edge
- **Step 2** For each pathway obtained in Step 1; find all the common edges between the pathway and the query graph. Time = O(n)

P1 = { A:5.3.1.9:B, C:2.7.2.3:D, D:5.4.2.1:E, E:4.2.1.11:F, F:2.7.1.40:G} = 5 common edges

P2 = { A:5.3.1.9:B, D:5.4.2.1:E, E:4.2.1.11:F , F:2.7.1.40:G} = 4 common edges

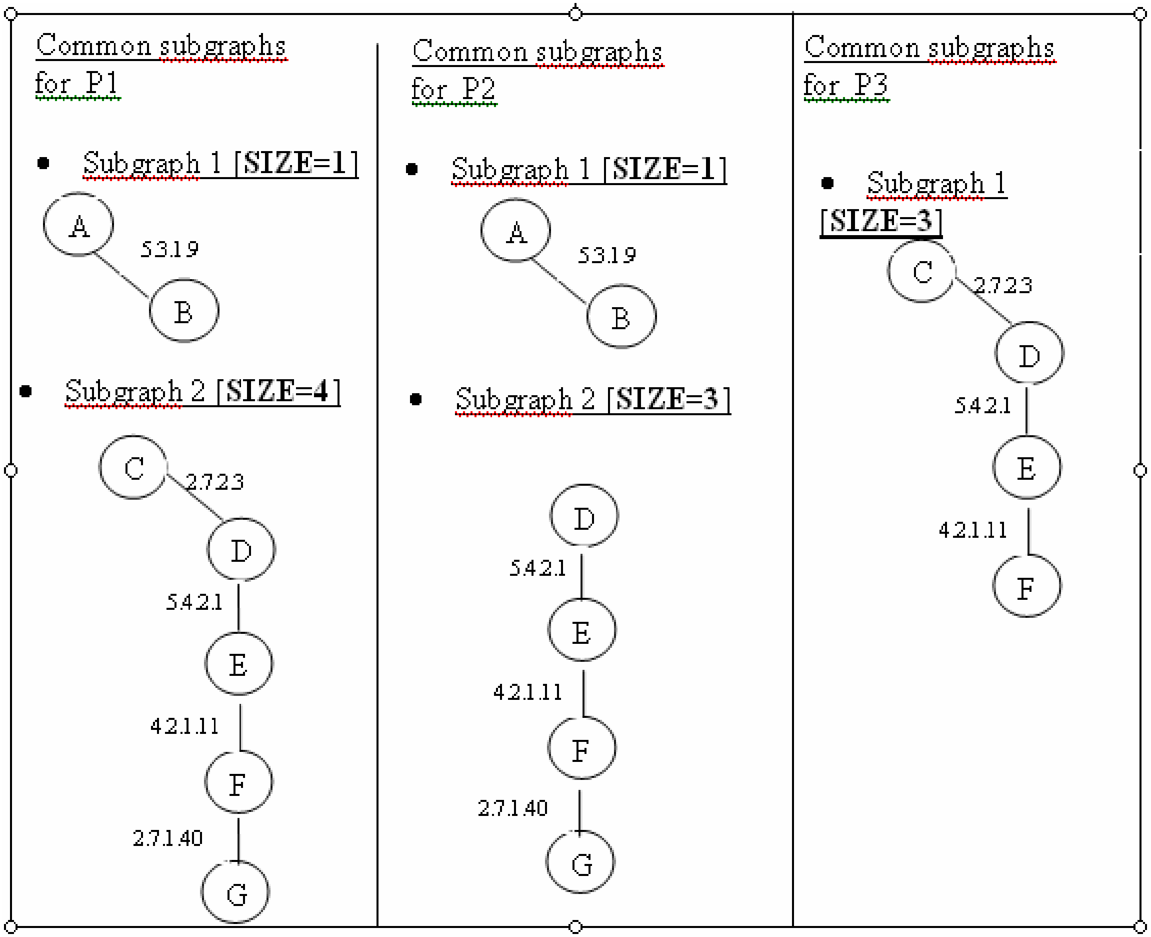
P3 = { C:2.7.2.3:D, D:5.4.2.1:E, E:4.2.1.11:F} = 3 common edges

Searching and computing similarity ...

- **Step 3.** For each pathway with common edges found above, perform a simple Depth First Traversal (DFT) on the undirected graph obtained in Step 3.

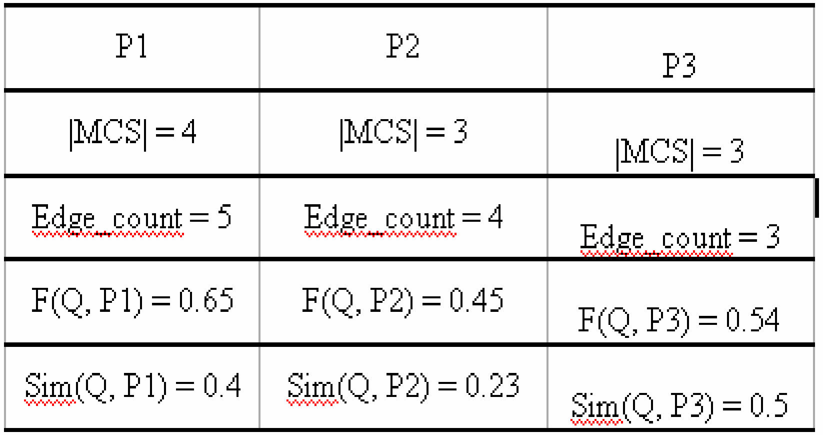
  Time = O(n)

- The connected components (trees) obtained in the Depth First Traversal forest will represent the common subgraphs between Q and the pathway.

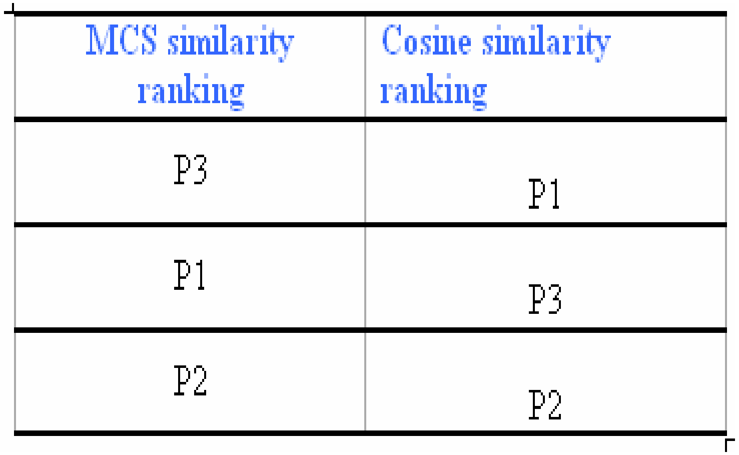


Searching and computing similarity …

- **Step 4.** Find a maximal subgraph and use it to compute the similarity measure based on Equation 1 and 2 . Merge and Rank the pathways in descending order of similarity based on the similarity measure chosen by the user. Time = O(n)



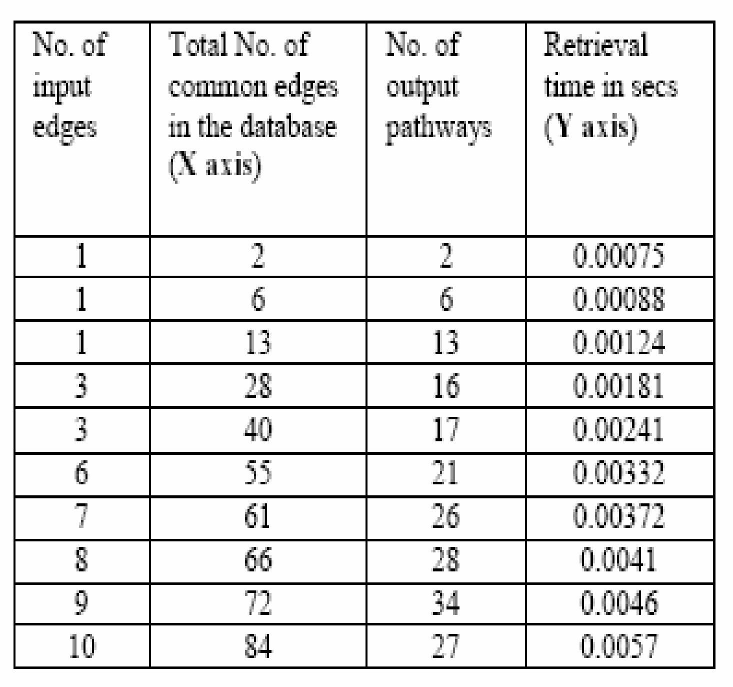
P1	P2	P3
|MCS| = 4	|MCS| = 3	|MCS| = 3
Edge_count = 5	Edge_count = 4	Edge_count = 3
F(Q, P1) = 0.65	F(Q, P2) = 0.45	F(Q, P3) = 0.54
Sim(Q, P1) = 0.4	Sim(Q, P2) = 0.23	Sim(Q, P3) = 0.5



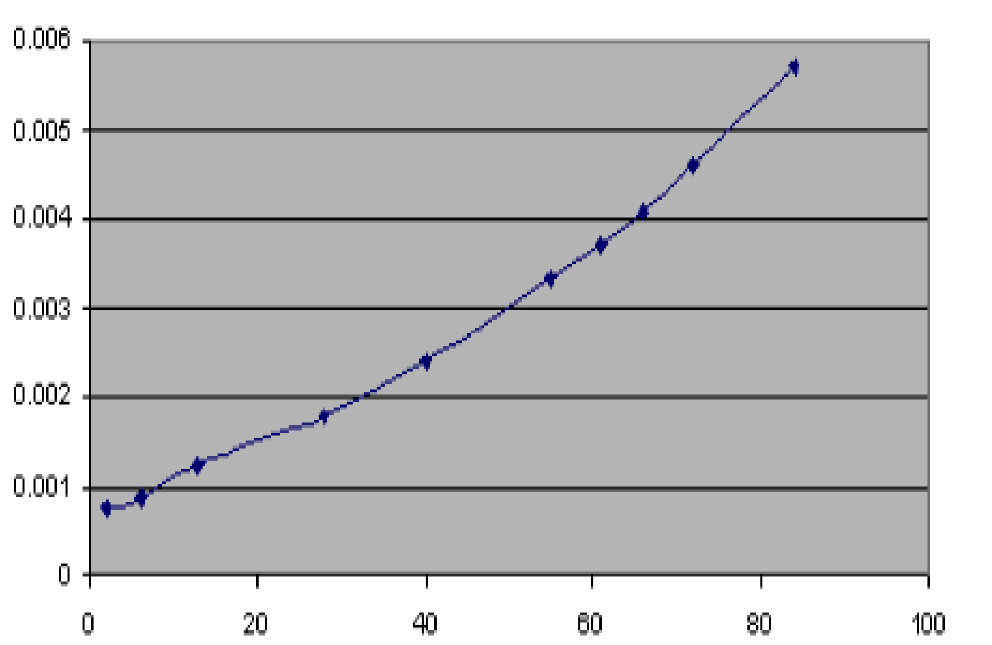
MCS similarity ranking	Cosine similarity ranking
P3	P1
P1	P3
P2	P2

- *The search time/retrieval time given a query pathway graph is linear in the total number of edges (n) in common with the query in the entire database.*

Experimental Studies ...



No. of input edges	Total No. of common edges in the database (X axis)	No. of output pathways	Retrieval time in secs (Y axis)
1	2	2	0.00075
1	6	6	0.00088
1	13	13	0.00124
3	28	16	0.00181
3	40	17	0.00241
6	55	21	0.00332
7	61	26	0.00372
8	66	28	0.0041
9	72	34	0.0046
10	84	27	0.0057



X-axis: total no. of edges in common with the query in the entire database, Y-axis: retrieval time in seconds.

Conclusion and Future Work

- We have described a search engine for the distributed searching of metabolic pathways
- We used Unique Chemical Keys (UCK) to create a uniquely labeled graph
- We then viewed edges as terms and used an inverted file list so that search is linear in the number of terms n that are shared by the query and the edges in the database of pathways
- This is one of the tools being developed for with the Chicago Biomedical Consortium (CBC) Proteomics Repository

Questions ?

For more information:

www.ncdm.uic.edu

For publications:

www.rgrossman.com

Thank You !