

Derivation of Graphs from Metabolic Pathways: Similarity Computation

Ms. Losiana Nayak, OUAT

M. Sc Bioinformatics

Lusiana_nayak@yahoo.co.in

PARTICULARS



Citric Acid Cycle

KEGG: Kyoto Encyclopedia of Genes and Genomes

The Pathway Database

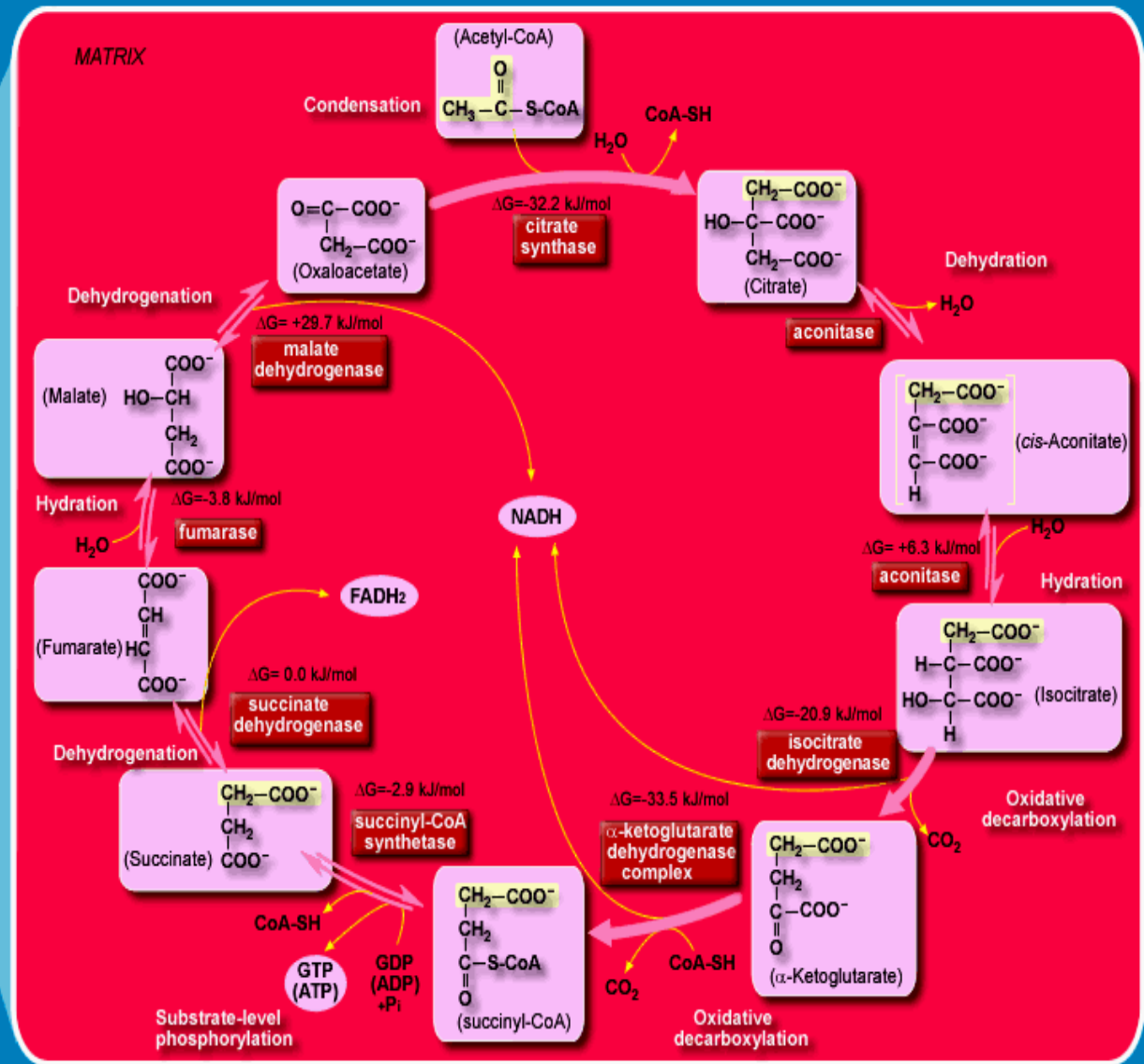
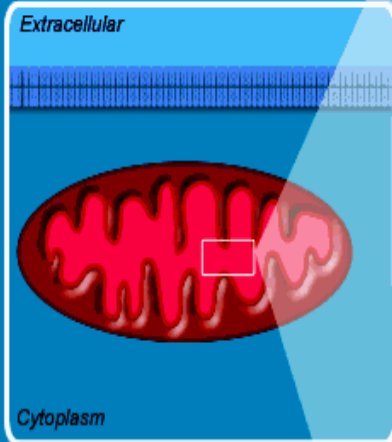
Enzyme Commission Numbers

The Data Generation Step

Conclusion

Bibliography

The TCA Cycle



KEGG:



Kyoto Encyclopedia of Genes and Genomes


"KEGG is an effort to computerize current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes and to provide links from the gene catalogs produced by genome sequencing projects."

<http://www.genome.ad.jp/kegg/kegg.html>

PATHWAY Database (Generalized protein interaction network)

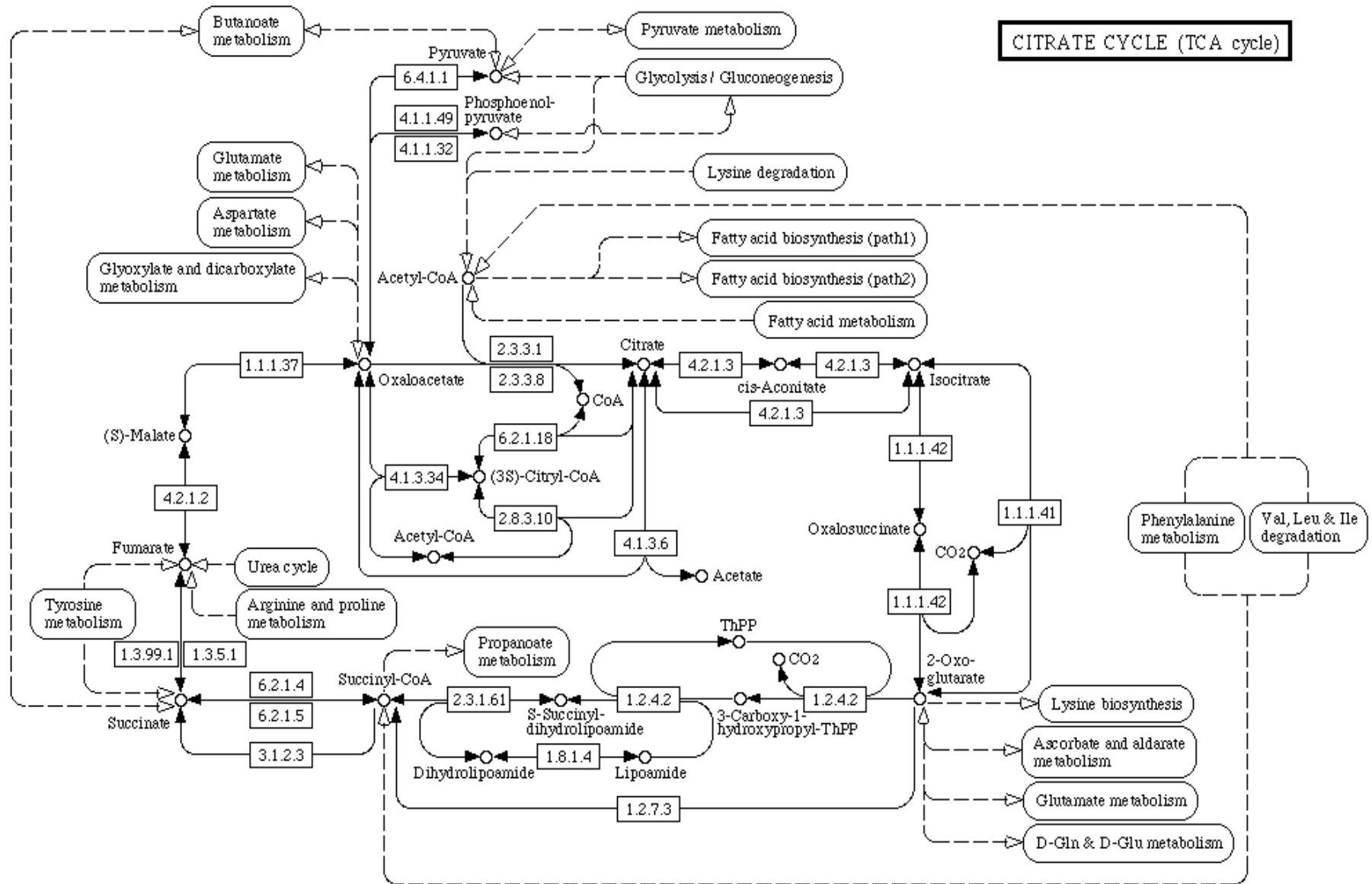


PATHWAY database is a collection of pathway maps, representing wiring diagrams of proteins and other gene products responsible for various cellular functions. Reflecting the map resolution and functional modules at different levels, these pathway maps are hierarchically classified.

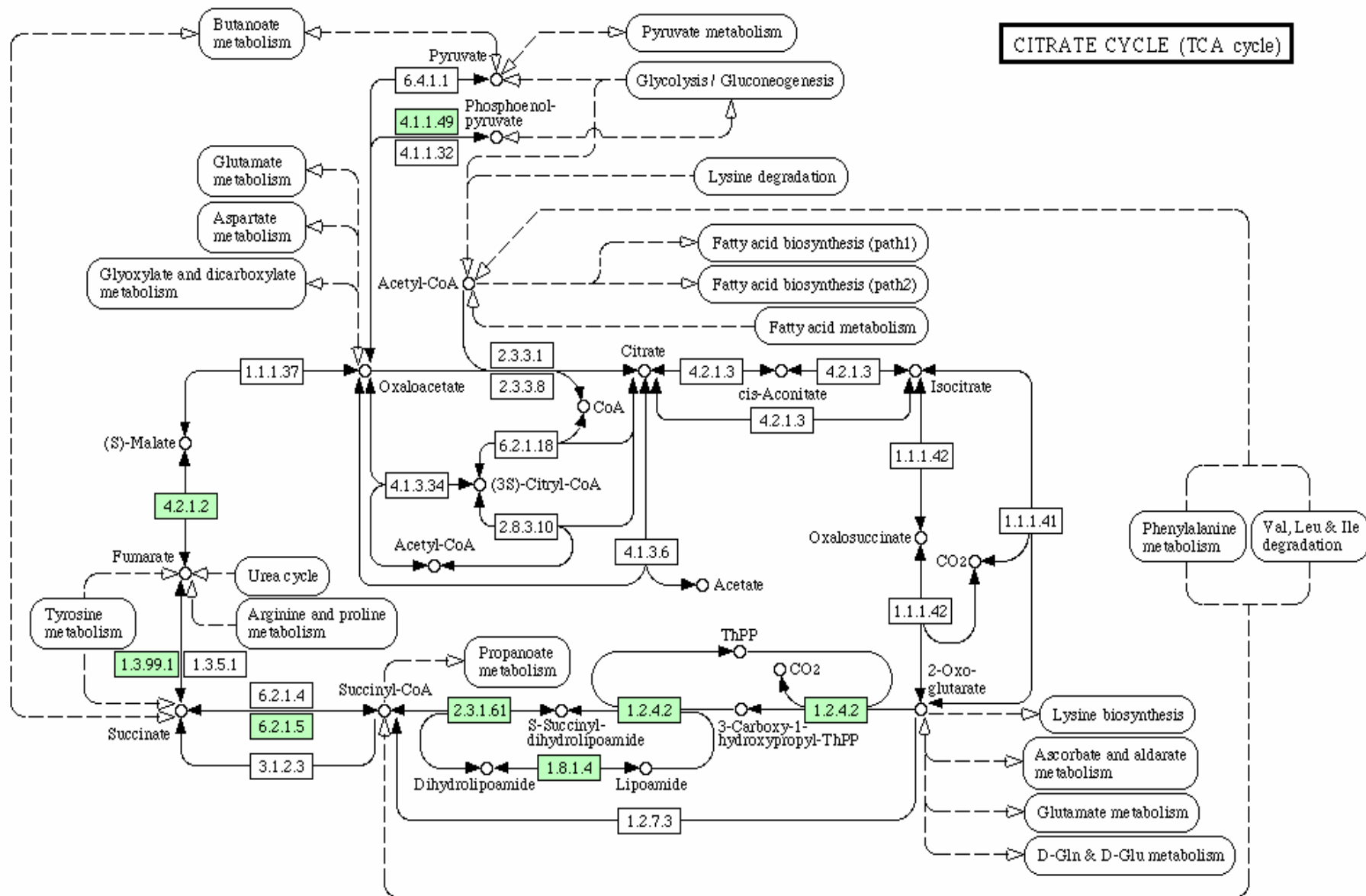


The PATHWAY database is a collection of manually drawn diagrams called the KEGG reference pathway diagrams (maps). From the manually drawn reference pathways, many organism specific pathways are automatically generated. The organism specific pathways are demarcated from each other by superimposing (coloring) genes specific to every organism. As of 25th September 2003, the database contains 13,457 entries including 235 reference pathway diagrams.

KEGG/PATHWAY: Generalized TCA Cycle




TCA Cycle of *Wigglesworthia brevipalpis*



ENZYME COMMISSION NUMBERS



EC numbers (Enzyme Commission numbers) are a numerical classification scheme for enzymes based on the chemical reactions they catalyze. As a system of enzyme nomenclature, every EC number is associated with a recommended name for the respective enzyme. Every enzyme code consists of the letters "EC" followed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzyme.



The first number shows to which of the six main divisions (classes) the enzyme belongs,

The second figure indicates the subclass,

The third figure gives the sub-subclass,

The fourth figure is the serial number of the enzyme in its sub-subclass.



Class 1. Oxidoreductases


Class 2. Transferases

Class 3. Hydrolases

Class 4. Lyases

Class 5. Isomerases

Class 6. Ligases.



Detailed level wise information can be
obtained by visiting the URL


[http://www.genome.jp/htbin/get_htext?
ECtable.](http://www.genome.jp/htbin/get_htext?ECtable)

The data generation step



IV. ASSUMPTIONS FOR SUBSTRATES:

- A... Pyruvate,
- B... Phosphoenol pyruvate,
- C... Acetyl Co A,
- D... Co A,
- E... Oxaloacetate,
- F... Citrate,
- G... (3S)-Citryl Co A,
- I... Acetate,
- J... Cisaconitate,
- K... Isocitrate,
- L... Oxaloacetate,
- M... 2-Oxo-glutarate,
- N... 3-carboxy-1-hydroxypropyl-Thpp



O... S-succinyl Co A
P... ThPP
Q... Succinyl Co A,
R... Dihydrolipoamide,
S... Lipoamide,
T... Succinate,
U... Fumerate,
V... (S)-Malate.

ASSUMPTIONS FOR EC NUMBERS:



a... 6.4.1.1

b... 4.1.1.49

c... 4.1.1.32

d... 2.3.3.1

e... 2.3.3.8

f... 4.1.3.34

g... 2.8.3.10


h... 6.2.1.18

i... 4.1.3.6

j... 4.2.1.3

k... 1.1.1.41

l... 1.1.1.42



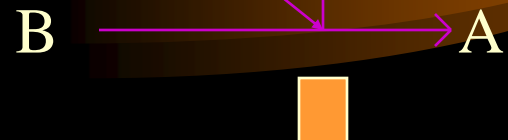
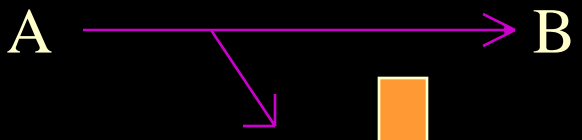
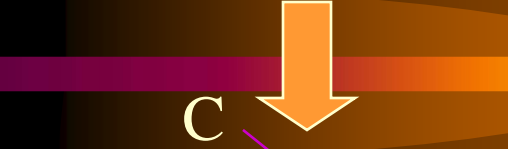
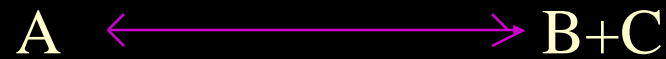
m... 1.2.4.2
n... 2.3.1.61
o... 1.8.1.4
p... 1.2.7.3
q... 6.2.1.4
r... 6.2.1.5
s... 3.1.2.3
t... 1.3.99.1
u... 1.3.5.1
v... 4.2.1.2
w... 1.1.1.37

RULES FOR CONVERSION OF MAPS INTO GRAPHS:

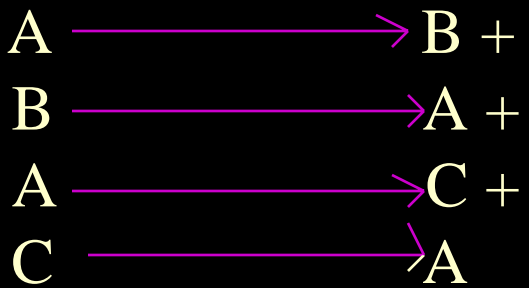


- 1. Avoid release of CO₂ and H₂O**

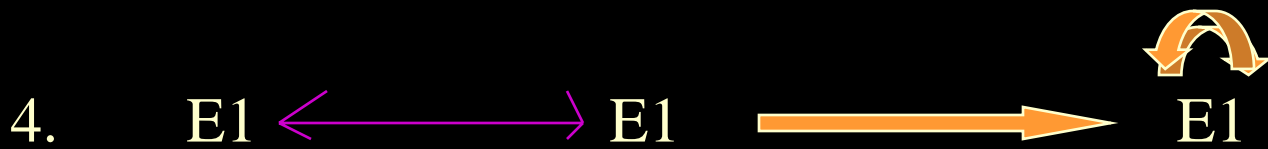
2.



So $A \longleftrightarrow B + C$



By no means B and C are connected.

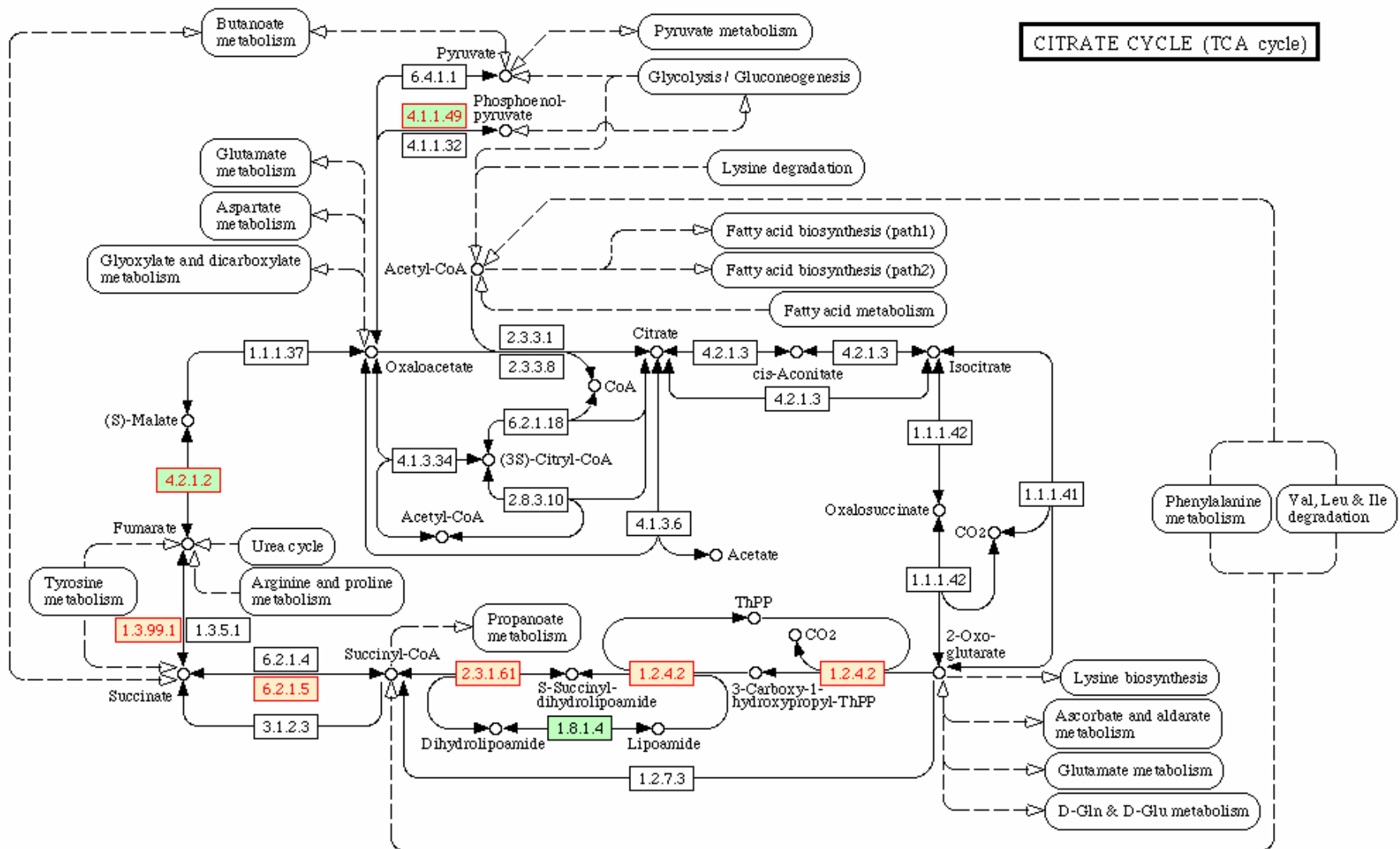


HOW TO CONSTRUCT THE ENZYME GRAPHS?

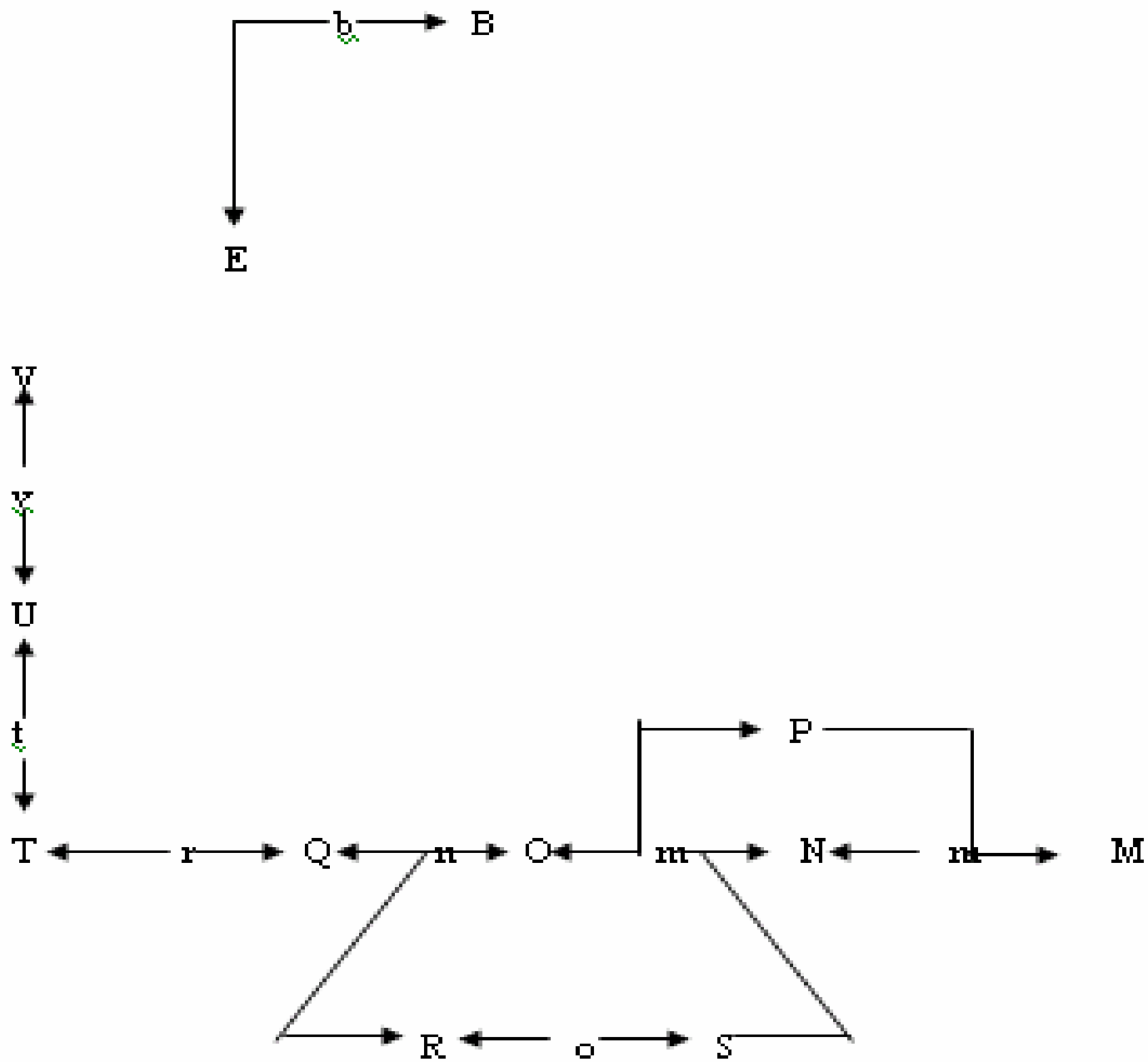
Each created enzyme graph can be represented as $G = (V, E)$ where V is the set of vertices (nodes) and $E \subseteq V \times V$ is the set of edges. If $V = \phi$, then G is empty graph. If for every edge $(a, b) \in E$ there exists the edge $(b, a) \in E$, then the graph is said to be undirected.

From a species specific metabolic map, only the colored part i.e. active reactions are taken separately. For example from the TCA Cycle map of Wigglesworthia brevipalpis the following active reactions are taken.

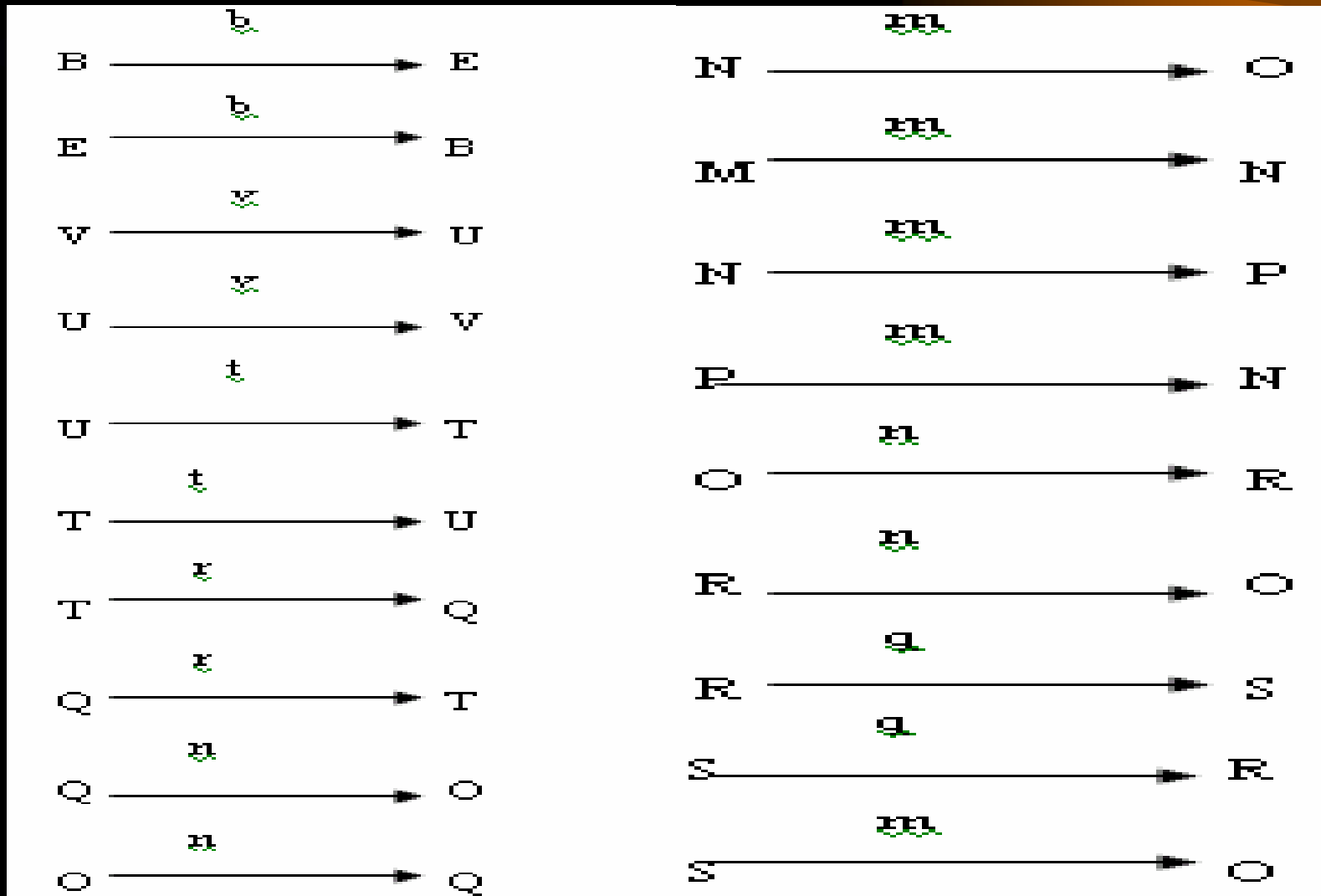
Citrate cycle (TCA cycle) - *Wigglesworthia brevipalpis*



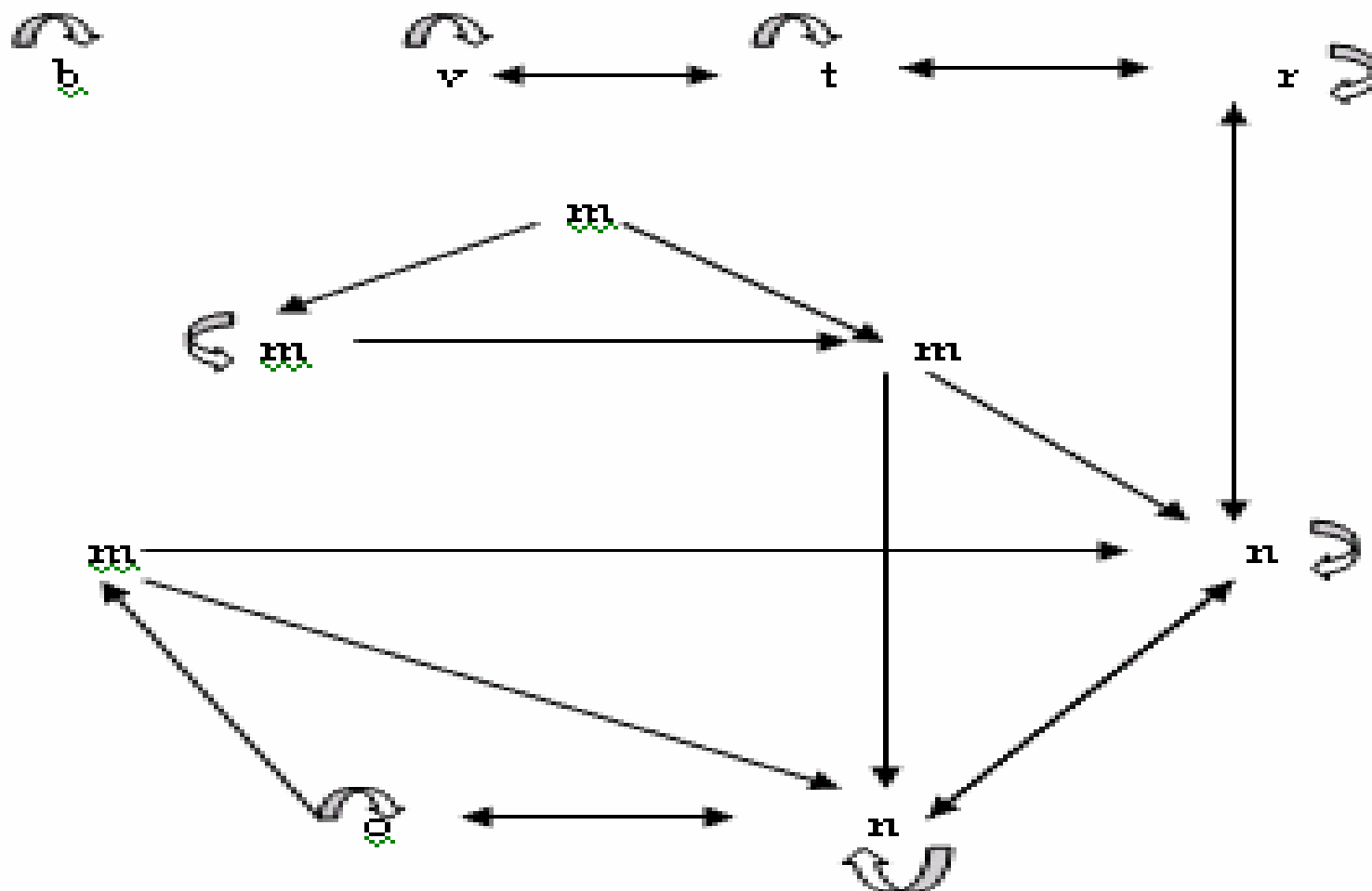
Wigglesworthioa brevipalpis: The Enzyme Graph



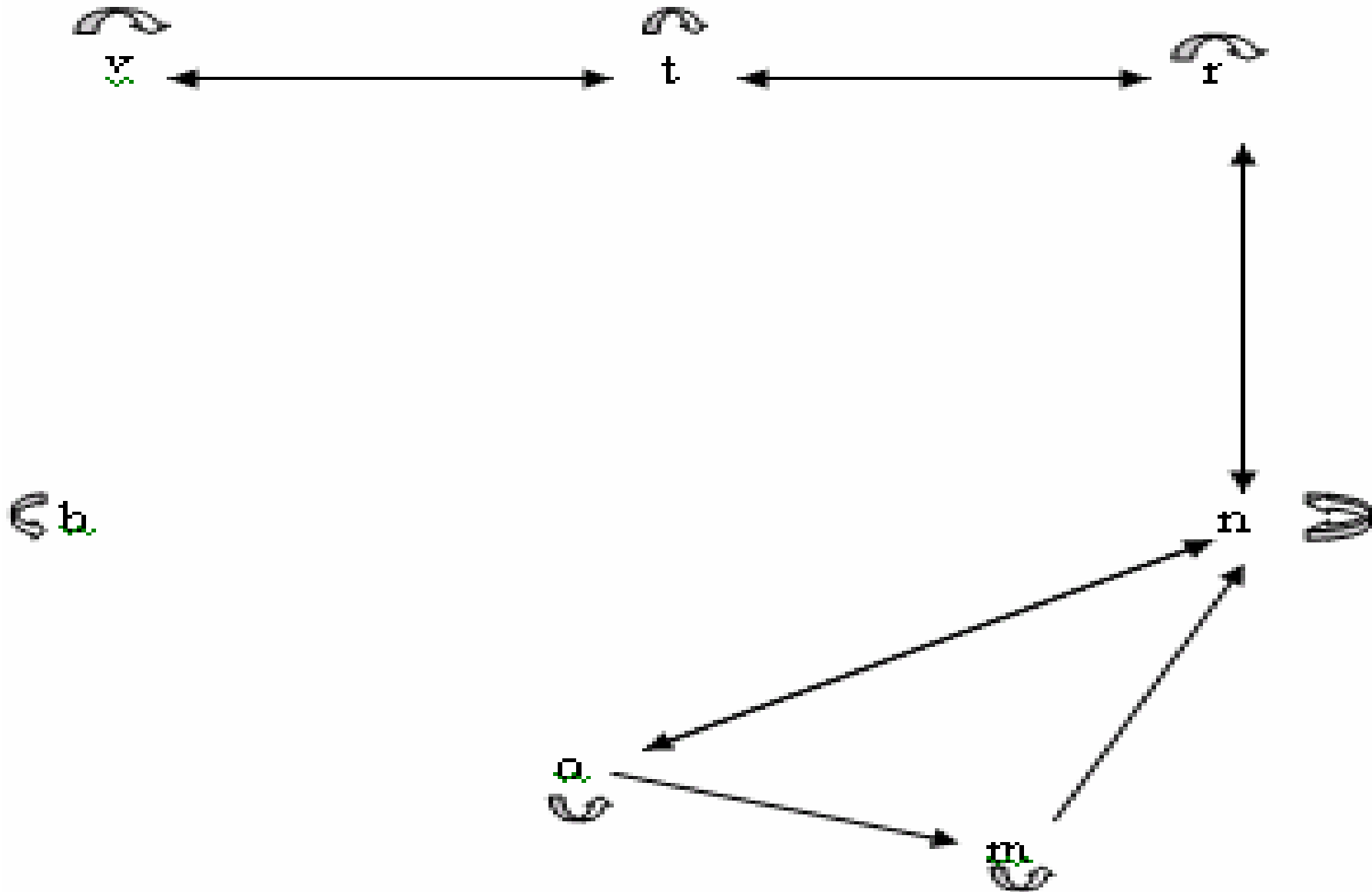
In second step, reactions are separated according to their reversibility or irreversibility



In next step a graph is constructed by taking the EC numbers as nodes and applying the rules described before.



Fourth and final step involves simplification of the graph obtained in 3rd step.



CONVERSION OF GRAPHS TO ADJACENCY MATRICES

If graph G has \underline{n} vertices, then the adjacency matrix is a ($n \times n$) matrix 'A' defined by

$$A(x, y) = \begin{cases} 1 & \text{if } x \longrightarrow y \text{ in } G, \\ 0 & \text{else} \end{cases}$$

x- indicates rows,

y- indicate columns,

Here a to z alphabets are taken as abbreviation of nodes which are actually enzyme commission numbers.

Adjacency matrix representing TCA Cycle enzyme graph of *W. brevipalpis*

	y	b	m	n	o	r	t	v
x								
b		1	0	0	0	0	0	0
m		0	1	1	0	0	0	0
n		0	0	2	1	1	0	0
o		0	1	1	1	0	0	0
r		0	0	1	0	1	1	0
t		0	0	0	0	1	1	1
v		0	0	0	0	0	1	1

The algorithm followed



is originally implemented by authors Maureen Heymans and Ambuj K. Singh. It is described in the paper “**Deriving phylogenetic trees from the similarity analysis of metabolic pathways.**” Department of Computer Science, University of California, Santa Barbara, USA, *Bioinformatics*, 2003, vol. 19, suppl. 2, pages i138- i146.

Initialization:

$$S_0(a, b) = \text{Sim}(a, b) \quad (1)$$

Iterative step:

$$S_{(k+1)}(a, b) = ((A_{k1}(a, b) + A_{k2}(a, b) + A_{k3}(a, b) + A_{k4}(a, b) - \\ D_{k1}(a, b) + D_{k2}(a, b) + D_{k3}(a, b) + D_{k4}(a, b)) / 4) \\ \times \text{Sim}(a, b) \quad (2)$$

Normalization:

$$S \leftarrow \frac{S}{\|S\|_2} \quad (3)$$

Graph matching

Computation of similarity scores between matched nodes

Computing similarity between two graphs

$$S_{(G1,G2)} = \frac{\sum_{a \in G1, b \in G2, M(a,b)=1} S(a,b)}{\sqrt{n1.n2}} \quad (4)$$

IMPLEMENTATION OF ALGORITHM IN C CODE

Input- Two adjacency matrices corresponding to TCA Cycle metabolic maps of Campylobacter jejuni and Escherichia coli taken from KEGG.

Final Output:

average sum of difference values of 1 iteration: 0.006764
average sum of difference values of 2 iteration: 0.002121
average sum of difference values of 3 iteration: 0.001216
average sum of difference values of 4 iteration: 0.000712
average sum of difference values of 5 iteration: 0.000424
average sum of difference values of 6 iteration: 0.000254
average sum of difference values of 7 iteration: 0.000159
average sum of difference values of 8 iteration: 0.000102
average sum of difference values of 9 iteration: 0.000067

Converging threshold Criteria:



Since it will be meaningless for the code to run for unlimited iterations, we have to fix the threshold criteria for the output. Here limitation is exercised on `asdv` value. If for n th iteration, the value stored in `asdv`, that gives **average of sum of difference values** between two successive iterations, is less than **0.0009**, then the output will be that of $(n-1)$ th iteration and program will be terminated.

result after 9 iterations:

0.000000,0.000000,0.000000,0.000000,0.000000,-0.000000,-0.000000,-0.000000,-0.000000,-
0.000000,-0.034276,-0.000000,-0.000000,0.000000,
0.228591,0.000000,0.101796,0.000000,0.094335,-0.004456,-0.000000,-0.000000,-0.000000,-
0.000000,-0.000000,-0.000000,-0.000924,0.000000,
0.000000,0.248196,0.000000,0.029754,0.000000,0.000000,0.000000,-0.000000,-
0.002744,0.000000,-0.000000,-0.000000,-0.000000,-0.000000,
-0.018374,0.000000,-0.010499,0.000000,0.014563,0.203080,0.000000,0.000000,
0.000000,0.000000,-0.000000,-0.000000,-0.204913,-0.000000,
-0.000000,0.000000,-0.000000,0.000000,0.000000,0.000000,0.257821,0.038899,
0.000000,0.021194,-0.000000,-0.080020,-0.000000,-0.115701,
-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,0.000000,0.029671,0.065330,
0.000000,0.000321,0.000000,0.011068,-0.000000,-0.036273,
-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,
0.000000,-0.000000,0.354955,0.000000,0.000000,-0.000000,
-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,-0.000000,-0.044440,-0.026595,
0.000000,-0.016802,0.000000,0.470998,0.000000,0.001294,
-0.001316,-0.000000,-0.004624,-0.000000,-0.024824,-0.155504,-0.000000,-0.000000,-
0.000000,-0.000000,0.000000,0.000000,0.479654,0.000000,
0.000000,0.000000,0.000000,-0.000000,0.000000,-0.000000,-0.115070,-0.016873,-
0.000000,-0.003882,-0.000000,-0.002806,0.000000,0.263947,

TECHNICAL DETAILS



OPERATING SYSTEM: WINDOWS 9X/2000, LINUX

LANGUAGE : C

SOFTWARE : EXCEED

DATABASE : KEGG

MODIFICATIONS:



Among the four rules involved in creation of adjacency matrices, two are newly implemented, one is modified and the rest one is taken from the original paper (Heymans and Singh, 2003). Also the graphs created by Heymans and Singh are not up to dated according to the updation of KEGG database. So the enzyme graphs obtained are not expected to bear total resemblance with enzyme graphs created by the authors.

SUMMARY OF WORK:



Here we tried to generate computer readable data from metabolic pathways which can be applied further to calculate similarity between metabolic pathways belonging to two different species. The similarity scores can be later converted to distance scores and used in evolutionary study by creating phylogenetic trees.

Bibliography

1. Maureen, S., K., Aambuj, “Deriving phylogenetic trees from the similarity analysis of metabolic pathways”, Department of Computer Science, University of California, USA, *Bioinformatics*, 2003, vol. 19, suppl. 2, pages i138- i146.
2. K., Minoru, G., Susumu, “KEGG: Kyoto encyclopedia of Genes and Genomes”, Institute of Chemical Research, Kyoto University, Japan, *Nucleic Acid Research*, 2000, Vol. 28, No. 1.

http://www.fact-index.com/m/me/metabolic_pathway.html

<http://www.genome.ad.jp/kegg/>

<http://www.chem.qmw.ac.uk/iubmb/enzyme/>

GUIDES:

Ms. K. Karunasree

Assistant Professor

Dept. of Bioinformatics

CPGS, OUAT

Bhubaneswar- 3

kk_sree@yahoo.com

GUIDES:

Dr. Rajat K. De

Assistant Professor

Machine Intelligence Unit

Indian Statistical Institute

Kolkata- 108

rajat@isical.ac.in



THANK YOU

Worth 1000.com