A graph theoretic method for comparing TCA cycles of some species

Losiana Nayak and Rajat K. De
{lomiana_t,rajat}@isical.ac.in
Machine Intelligence Unit
Indian Statistical Institute
203 B. T. Road
Kolkata-700108, India
Objective

• Comparison of metabolic pathways

Future use

• Evolution/ Development of metabolic pathways

• Its similarity to conventional evolution

• Patterns in development of metabolic pathways

• Stages of major change
We took TCA Cycle for a certain set of species from KEGG/Pathway database

http://www.genome.jp/kegg/pathway/map/map00020.html
ENZYME COMMISSION NUMBERS

EC numbers (Enzyme Commission numbers) are a numerical classification scheme for enzymes based on the chemical reactions they catalyze. According to IUPAC-IUBMB 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 number indicates the subclass.

The third one gives the sub-subclass.

The fourth one is the serial number of the enzyme in its sub-subclass.
ENZYME GRAPH?

An enzyme graph derived from a metabolic pathway can be represented as $G = (V, E)$

- $V$ is the set of nodes representing EC numbers
- $E \subseteq V \times V$ is the set of edges representing connections between two EC numbers of two successive reactions
P. aerophilum

CITRATE CYCLE (TCA cycle)
CITRATE CYCLE (TCA cycle)

P. aerophilum
Symbols For Metabolites

A… Pyruvate,
B… Phosphoenolpyruvate,
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… Fumarate,
V… (S)-Malate.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>EC Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>6.4.1.1</td>
</tr>
<tr>
<td>b</td>
<td>4.1.1.49</td>
</tr>
<tr>
<td>c</td>
<td>4.1.1.32</td>
</tr>
<tr>
<td>d</td>
<td>2.3.3.1</td>
</tr>
<tr>
<td>e</td>
<td>2.3.3.8</td>
</tr>
<tr>
<td>f</td>
<td>4.1.3.34</td>
</tr>
<tr>
<td>g</td>
<td>2.8.3.10</td>
</tr>
<tr>
<td>h</td>
<td>6.2.1.18</td>
</tr>
<tr>
<td>i</td>
<td>4.1.3.6</td>
</tr>
<tr>
<td>j</td>
<td>4.2.1.3</td>
</tr>
<tr>
<td>k</td>
<td>1.1.1.41</td>
</tr>
<tr>
<td>l</td>
<td>1.1.1.42</td>
</tr>
<tr>
<td>m</td>
<td>1.2.4.2</td>
</tr>
<tr>
<td>n</td>
<td>2.3.1.61</td>
</tr>
<tr>
<td>o</td>
<td>1.8.1.4</td>
</tr>
<tr>
<td>p</td>
<td>1.2.7.3</td>
</tr>
<tr>
<td>q</td>
<td>6.2.1.4</td>
</tr>
<tr>
<td>r</td>
<td>6.2.1.5</td>
</tr>
<tr>
<td>s</td>
<td>3.1.2.3</td>
</tr>
<tr>
<td>t</td>
<td>1.3.99.1</td>
</tr>
<tr>
<td>u</td>
<td>1.3.5.1</td>
</tr>
<tr>
<td>v</td>
<td>4.2.1.2</td>
</tr>
<tr>
<td>w</td>
<td>1.1.1.37</td>
</tr>
</tbody>
</table>
CITRATE CYCLE (TCA cycle)

Pyrobaculum aerophilum

Diagram of the Citrate Cycle with labeled nodes and arrows indicating the flow of the cycle.
So $A \leftrightarrow B + C$

By no means $B$ and $C$ are connected.
(Individual Member Reactions of TCA)

1. C → D(d)
2. C → F(d)
3. E → F(d)
4. F → J(j)
5. J → F(j)
6. F → K(j)
7. K → F(j)
8. J → K(j)
9. K → J(j)
10. K → L(l)
11. L → K(l)
12. L → M(l)
13. M → L(l)
14. M → Q(p)
15. R → S(o)
16. S → R(o)
17. Q → T(r)
18. T → Q(r)
19. T → U(t)
20. U → T(t)
21. U → V(v)
22. V → U(v)
TCA Cycle (*Pyrobaculum aerophilum*)

\[
\begin{align*}
&d 
\quad \rightarrow 
\quad j 
\quad \leftarrow 
\quad l 
\quad \rightarrow 
\quad p \\
&w 
\quad \leftarrow 
\quad v 
\quad \rightarrow 
\quad t 
\quad \leftarrow 
\quad r
\end{align*}
\]
TCA Cycle, *Pyrobaculum aerophilum*
CONVERSION OF GRAPHS TO ADJACENCY MATRICES

If graph $G$ has $n$ vertices, then the adjacency matrix is an $n \times n$ matrix $A$ defined as

$$A(x, y) = 1, \text{ if } x \rightarrow y \text{ in } G$$
$$= 0, \text{ otherwise}$$
Adjacency matrix representing TCA Cycle enzyme graph of *P. aerophilum*

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>d</th>
<th>j</th>
<th>l</th>
<th>o</th>
<th>p</th>
<th>r</th>
<th>t</th>
<th>v</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>j</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>t</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>w</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Here a to z alphabets are taken as abbreviation of nodes which are actually enzyme commission numbers.
Two adjacency matrices representing two metabolic pathways is our input data for graph comparison.

The algorithm we follow is a modification of that by Heymans and Singh (Bioinformatics, 2003, vol. 19, suppl. 2, pages i138- i146).

Instead of Bipartite graph matching we simply took the pair of nodes \((x,y)\) having maximum similarity, where \(x \in G_1\) and \(y \in G_2\).
Initialization:
\[ S_0(a, b) = \text{Sim}(a, b) \]

Iterative step:
\[ S_{(k+1)}(a, b) = \left( \frac{A_{k1}(a, b) + A_{k2}(a, b) + A_{k3}(a, b) + A_{k4}(a, b) - D_{k1}(a, b) + D_{k2}(a, b) + D_{k3}(a, b) + D_{k4}(a, b))}{4} \right) \times \text{Sim}(a, b) \]

Normalization:
\[ S \leftarrow \frac{S}{\|S\|_2} \]

Computing similarity between two graphs:
\[ S_{(G1,G2)} = \sum_{a \in G1, b \in G2, M(a,b)=1} S(a,b) \]
\[ \frac{1}{\sqrt{n1.n2}} \]
### Species we have considered

<table>
<thead>
<tr>
<th>Species</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y. pestis</td>
<td>A. sp. CC7120</td>
</tr>
<tr>
<td>X. campestris</td>
<td>M. janaschii</td>
</tr>
<tr>
<td>P. luminescens</td>
<td>A. fulgidus</td>
</tr>
<tr>
<td>M. leprae</td>
<td>T. acidophilum</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>C. jejuni</td>
</tr>
<tr>
<td>S. solfataricus</td>
<td>M. mazei</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>P. furiosus</td>
</tr>
<tr>
<td>E. coli K12MG</td>
<td>C. acetobutylicum</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>C. pneumonia</td>
</tr>
<tr>
<td>M. musculus</td>
<td>C. muridarium</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>W. brevipalpis</td>
</tr>
<tr>
<td>C. elegans</td>
<td>P. horikoshii</td>
</tr>
<tr>
<td>A. aeolicus</td>
<td>F. nucleatum</td>
</tr>
<tr>
<td>P. aerophilum</td>
<td></td>
</tr>
</tbody>
</table>
Based on the similarity matrix we constructed phylogenetic tree of the TCA cycle of the species using tools (Fitch and Drawgram) from Phylip 3.65 package.
TCA cycle tree is not in accordance with Phylogenetic tree of life.

From evolution point of view closest species may have different kind of metabolism.
Some Useful References


Thank you!