Pathway Hunter Tool (PHT) – A Platform for Metabolic Network Analysis and Potential Drug Targeting

Inaugural – Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Syed Asad Rahman
aus Bhagalpur (India)

Köln, 2006

Berichterstatter	**Prof. Dr. Rainer Schrader** **Prof. Dr. Dietmer Schomburg**
Tag der mündlichen Prüfung	**14th February 2007**

ACKNOWLEDGEMENT

I would like to thank my supervisor, Prof. Dr. Dietmar Schomburg (CUBIC & Dept of Biochemistry, University of Cologne, Cologne, Germany) and guide, Prof. Dr. Rainer Schrader (CUBIC & ZAIK, University of Cologne, Cologne, Germany) for their invaluable guidance, support and endless patience throughout the course of my research. I am also grateful to my co-workers at CUBIC and to the students involved in the project for their hard work and dedication. I would like to thank my mentors Prof. Dr. Seyed Ehtesham Hasnain, Dr. H. A. Nagarajaram, Dr. Shekhar Mande, Dr. Akash Ranjan, Dr. Niyaz Ahmed and Dr. Noman Siddiqui (CDFD, Hyderabad, India) for their invaluable support and guidance – both professional and moral.

A special thanks to Pavan Advani, Saradhi Jonalagadda, Jyothi Padiadpu, Gill Egerton, Tsigereda Biru, Sebastian Buchinger, Kai Hartmann, Ralph Schunk & Michael Eiden, for their scientific support. I would also like to thank Michael Zimmerman, Beate Marx and Mark Lohmann who facilitated my fruitful stay at CUBIC. A special thanks to Srikanth Nagella Naidu an old colleague and friend who was always there to help me out in programming.

On a more personal note, I would like to extend my heartfelt gratitude to my parents and brother in India. I would not have been where I am if it were not for you. I could not have given my best had it not been for my loving and patient wife Filzah. Thank you for everything.

Last and certainly not the least, I would like to bow my head in gratitude and thank the Almighty for having bestowed upon me the opportunity to extend my mind and my horizons...

Abstract / Kurzzusammenfassung

Metabolic network analysis will play a major role in “Systems Biology” in the future as they represent the backbone of molecular activity within the cell. Recent studies have taken a comparative approach toward interpreting these networks, contrasting networks of different species and molecular types, and under varying conditions. We have developed a robust algorithm to calculate shortest path in the metabolic network using metabolite chemical structure information. A divide and conquer technique using Maximal Common Subgraph (MCS) approach and binary fingerprint was used to map each substrate onto its corresponding product. Then for the calculation of the shortest paths (using modified Breadth First Search algorithm) the two biochemical criteria “local” and “global” structural similarity were used, where “local similarity” is defined as the similarity between two intermediate molecules and “global similarity” is defined as the amount of conserved structure found between the source metabolite and the destination metabolites after a series of reaction steps. The pathway alignment was introduced to find enzyme(s) preference in the pathway of various organisms (a local and global outlook to metabolic networks). This was also used to predict potentially missing enzymes in the pathway. A novel concept called “load points” and “choke points” identifies hot spots in the network. This was used to find important enzymes in the pathogens metabolic network for potential drug targets.

Die Analyse von metabolischen Netzwerken wird eine große Rolle in der „Systembiologie“ der Zukunft spielen, weil sie das Rückgrat der molekularen Aktivitäten in der Zelle repräsentieren. Kürzlich erfolgte Studien verwenden vergleichende Ansätze zur Interpretation dieser Netzwerke, um zwischen verschiedenen Spezies und verschiedenen Bedingungen zu unterscheiden. Wir haben eine robusten Algorithmus zur Berechnung von kürzesten Pfaden in metabolischen Netzen unger Verwendung von chemischer Strukturinformation entwickelt. Eine Divide-and-Conquer-Technik unter Verwendung von Maximal Common Subgraph (MCS) und Fingerprint Algorithmen bildet jedes Substrat auf das korrespondierende Produkt ab. Für die Berechnung der kürzesten Pfade (mit einem modifizierten Breitensuche-Algorithmus) werden die beiden Kriterien “lokale” und “globale” strukturelle Ähnlichkeit verwendet, wobei die “lokale Ähnlichkeit” als Ähnlichkeit zwischen zwei Molekülen einer Reaktion und die “globale Ähnlichkeit” als Anteil der konservierten Struktur zwischen Ausgangs- und Endpunkt nach einer Reihe von Reaktionsschritten definiert ist. Das Alignment von Pfaden wurde eingeführt, um die Enzyme eines Pfades zwischen verschiedenen Organismen zu vergleichen (für eine lokale und globale Betrachtung von metabolischen Netzen). Das Alignment wurde ebenfalls benutzt, um potentiell fehlende Enzyme im Pfad vorherzusagen. Ein neues Konzept, das “load points” und “choke points” genannt wird, identifiziert Hotspots im Netzwerk. Es wurde verwendet, um wichtige Enzyme in metabolischen Netzen von pathogenen Organismen als potentielle Ziele für Medikamente zu finden.

1. Network

The advances in technology and communication have motivated the development of numerous random network models (Banavar, Maritan et al. 1999; Newman, Watts et al. 2002) contending to describe graphs of biological, technological, and sociological origin (Wasserman and Faust 1994; Newman 2001). The success of a graph model has been evaluated by how well it reproduces a few key features of the real-world data, such as degree distributions, mean geodesic lengths, and clustering coefficients (Newman 2001; Strogatz 2001; Barabasi 2002). The network can be described using graph theory, as a set of nodes and edges depicting a relationship between them (Newman, Strogatz et al. 2001; Watts 2003). For example, a graph model of **actors** (nodes) and their **movies** (edges) highlights the connection (often called as **ties**) between them. Both actors and ties can be defined in different ways depending on the questions of interest. An actor might be a single person, a team, or a company (Albert and Barabasi 2000). A tie might be a friendship between two people, collaboration or common member between two teams, or a business relationship between companies. A group of actors are related if they share common edges between them. Further these networks can be assigned as directed or undirected depending upon the relationship they depict. Consider the directed graph whose nodes correspond to static pages on the web, and whose edges correspond to hyperlinks between these pages. Some interesting properties of a network's topology include its diameter, degree distributions, connected components, and macroscopic structure (Newman 2001; Newman 2001).

Many of the existing networks such as biological networks, food webs, social acquaintance networks, the Internet or highway transportation networks appear to share common architectonic principles (Strogatz 2001; Barabasi 2002; Newman 2003). All these networks are sparse yet possess clusters. Many of them also feature a highly efficient pathway structure, in which only a small number of intermediate connections have to be passed in order to get from one node to another. In relation to social networks this has long been known as the "small-world" phenomenon where persons are linked with only six degrees (that means, six intermediate persons) of separation (Milgram 1967).

More recently, the growing availability of network data—from topological data about the Internet to metabolic and regulatory cellular networks—has attracted researchers to undertake a joint heading of network analysis (Krapivsky, Rodgers et al. 2001; Liljeros, Edling et al. 2001; Barabasi and Oltvai 2004; Lieberman, Hauert et al. 2005; Ma'ayan, Lipshtat et al. 2006). Moreover, the field was recently called *network science* to emphasize the interdisciplinary nature of the research.

1.1. Types of Graphs

The informal definition of a graph is a set of nodes with edges connecting some of them. However, there can be several variations on this theme, each of which has a special name. The graph can be directed (when edges point from one node to another) or undirected (edges point both ways). A graph can be a self-graph or a bipartite graph depending on whether the set of nodes pointed to by edges is the same as the set of nodes pointed from, or not. In addition, edges can have different weights, or not, which differentiates weighted graphs from unweighted graphs. All of these terms are formally defined below (Figure 1, Figure 2, Figure 3) show examples of these.

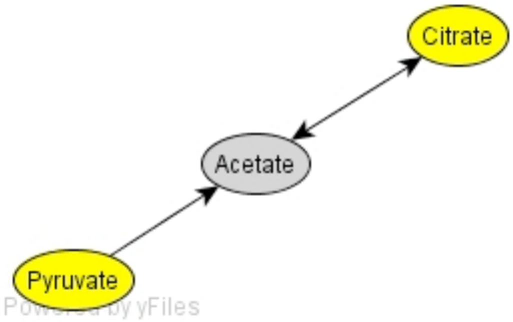

Figure 1. Edges in a directed graph point one way, such as a graph of metabolic network were edges implies relationship from one metabolite (substrate) to another (product)

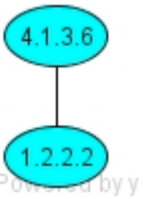

Figure 2. Undirected graphs have edges pointing both ways, such as a enzyme-enzyme graph. These represent self-graphs.

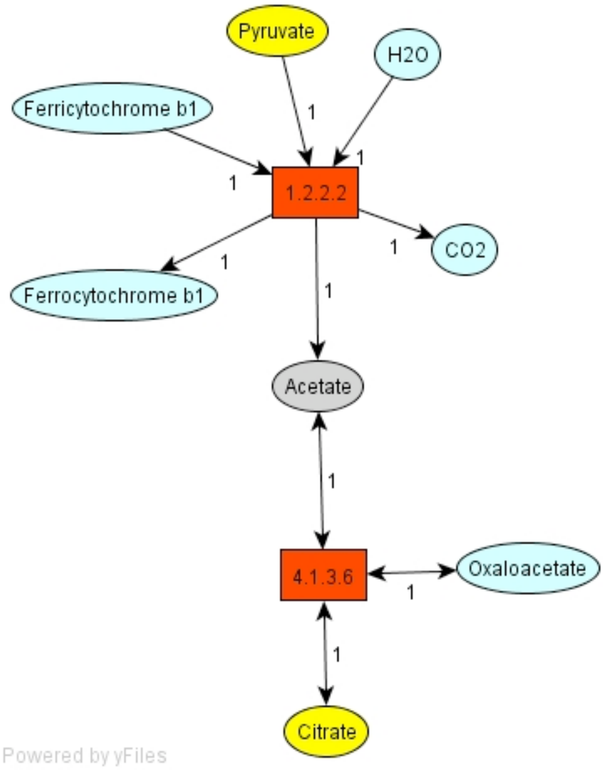

Figure 3. A bipartite graph has edges connecting two different sets of nodes, such as reactions and the metabolites coding it. A weighted graph has weights for each edge, such as the stoichiometric weight of the reaction.

Definition 1 (A simple Graph or Self-graph is defined as)

A graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ *is a set* $\mathcal{V}$ *of N nodes, and a set* $\mathcal{E}$ *of E edges between them. The number of nodes is denoted by* $N = \|\mathcal{V}\|$*, and the number of edges by* $E = \|\mathcal{E}\|$.

Definition 2 (Bipartite graph is defined as)

In a bipartite graph, *the set of nodes* $\mathcal{V}$ *consists of two disjoint sets of nodes* $\mathcal{V}_1$ *and* $\mathcal{V}_2$*:* $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$. *Any edge connects a node in* $\mathcal{V}_1$ *to a node in* $\mathcal{V}_2$*, that is,* $(i, j) \in$

$\mathcal{E} \Rightarrow i \in \mathcal{V}_1$ and $j \in \mathcal{V}_2$. The number of nodes in the two node set are $N_1 = \|\mathcal{V}_1\|$ and $N_2 = \|\mathcal{V}_2\|$. The number of edges is still $E = \|\mathcal{E}\|$.

Definition 3 (Directed and undirected graph is defined as)

In a directed graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, each edge $(i, j) \in \mathcal{E}$ points from node i to node j. An undirected graph is a directed graph where edges point both ways, that is, $(i, j) \in \mathcal{E} \Rightarrow (j, i) \in \mathcal{E}$.

Definition 4 (Weighted and unweighted graphs is defined as)

A weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, has a set of nodes $\mathcal{V}$, and a set of edges $\mathcal{E}$; and $\mathcal{W}$ represents the corresponding weights *of those edges.*

Weighted graphs can be both self-graphs or bipartite graphs. Unweighted graphs are a special case of weighted graphs, with all weights set to 1.

Definition 5 (Adjacency matrix of a self-graph)

The adjacency matrix **A** *of a weighted self-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ is an $\boldsymbol{N} \times \boldsymbol{N}$ matrix such that*

$$A_{i,j} = \begin{cases} w_{i,j} & if\,(i,j) \in \varepsilon \\ 0 & otherwise \end{cases} \quad for\ i \in 1.....N \tag{1.1}$$

The adjacency for an unweighted self-graph merely replaces $\mathbf{w_{i,j}}$ by 1.

Definition 6 (Adjacency matrix of a bipartite-graph)

The adjacency matrix A *of a weighted bipartite-graph $\mathcal{G} = (\mathcal{V},\mathcal{E},\mathcal{W})$ with $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ is an $\boldsymbol{N_1} \times \boldsymbol{N_2}$ matrix such that*

$$A_{i,j} = \begin{cases} w_{i,j} & if\,(i,j) \in \varepsilon \\ 0 & otherwise \end{cases} \quad for\ \begin{matrix} i \in 1.....N_1 \\ j \in 1.....N_2 \end{matrix} \tag{1.2}$$

The adjacency for an unweighted bipartite-graph merely replaces $\mathbf{w_{i,j}}$ by 1.

In addition, graph nodes and edges can have attached labels (i.e., categorical values); such graphs are called *labeled graphs*. However, our work has focused on unlabeled and unweighted graphs, both self- and bipartite, and both directed and undirected lists (Table 1) these symbols.

Table 1. Basic Graph Theory Notations

Symbols	Symbol Description
G	The graph. $\mathcal{G}$ $\mathcal{V}$ $\mathcal{E}$ $\mathcal{W}$ $\mathcal{G}$ $\mathcal{V}$ $\mathcal{E}$ unweighted graph
$\mathcal{V}$	The set of nodes in **G**. $\mathcal{V}$ $\mathcal{V}_1$ U $\mathcal{V}_2$ for a bipartite graph
$\mathcal{E}$	The set of edges
$\mathcal{W}$	Edge weights : zero when no edge exists, and positive when it does. Edges weights are 1 for an unweighted graph.
$\mathbf{A}$	The adjacency matrix of the graph
N	Number of nodes in **G**
N_1, N_2	Number of nodes in the partitions $\mathcal{V}_1$ and $\mathcal{V}_2$ of a bipartite graph **G**
E	Number of edges i n **G**

1.2. Properties of Networks

Network science offers a quantifiable description of the networks that characterize various systems. Here I will introduce and define the most basic network measures that allow us to compare and characterize different complex networks.

1.2.1 Degree

The most elementary characteristic of a node is its degree (or connectivity), k, which denotes the number of links the node has with other nodes (Barabasi and Oltvai 2004). For example, in the undirected network shown in Figure 4, node A has degree k = 5. Highly connected nodes in the network are often termed as "hubs" in the network, for example node A.

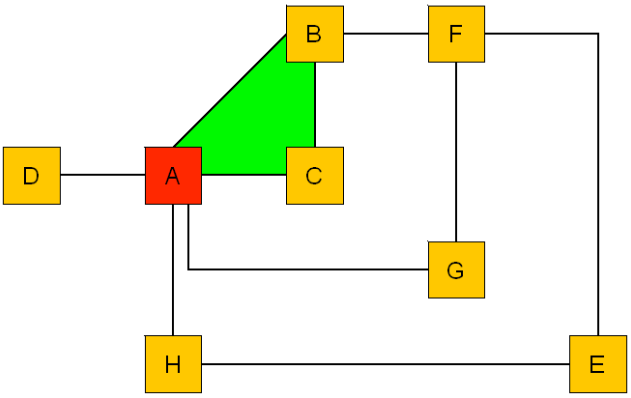

Figure 4. An example of an undirected graph. The shortest distance between node B to A is one step. The incoming/outgoing degree of node A is 5.

In a directed network (each link has a defined direction; see Figure 5), there is an incoming degree, k_{in}, which denotes the number of links that point to a node, and an outgoing degree, k_{out}, which denotes the number of links that start from it. For example, node A in Figure 5 has $k_{\text{in}} = 4$ and $k_{\text{out}} = 1$. An undirected network with N nodes and L links is characterized by an average degree $<k> = 2L/N$ (where <> denotes the average).

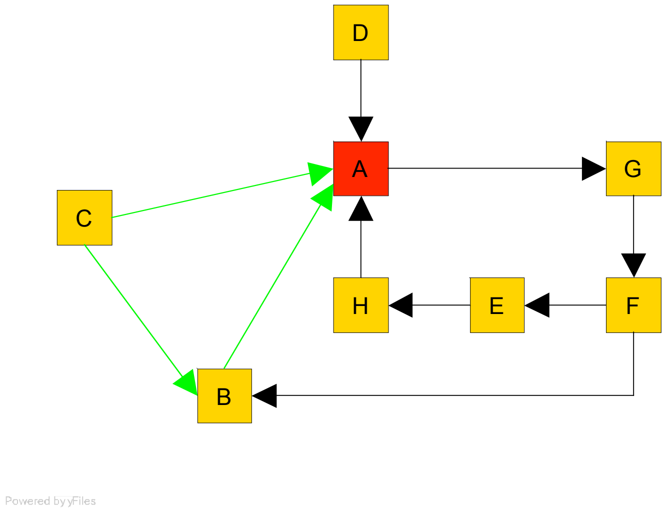

Figure 5. An example of a directed graph. The shortest path between node B to A is one step and between A to B is three steps. The incoming degree of node A is 4, whereas outgoing degree is 1.

1.2.2 Power Law and the Degree Distribution

Previous studies indicate that most of the biological networks are scale-free (Barabasi and Oltvai 2004), meaning their degree distribution approximates a power law, $P(k) \sim k^{-\gamma}$, where γ is the degree exponent and ~ indicates 'proportional to'. The degree distribution, $P(k)$, gives the probability that a selected node has exactly k links. It is obtained by counting the number of nodes $N(k)$ with $k = 1, 2\dots$ links and dividing it by the total number of nodes N (Barabasi and Oltvai 2004). The degree distribution allows us to distinguish between different classes of networks.

The value of γ determines many properties of the system i.e. the smaller the value of γ, the more important the role of the hubs is in the network. While for $\gamma>3$ the hubs are not relevant, for $2>\gamma>3$ there is a hierarchy of hubs, with the most connected hub being in contact with a small fraction of all nodes, and for $\gamma = 2$ a hub-and-spoke network emerges, with the largest hub being in contact with a large fraction of all nodes. In general, the unusual properties of scale-free networks are valid only for $\gamma<3$, when the dispersion of the $P(k)$ distribution, which is defined as $\sigma^2 =< k^2 > - < k >^2$, increases with the number of nodes (that is, σ diverges), resulting in a series of unexpected features such as a high degree of robustness against accidental node failures (Albert, Jeong et al. 2000). For $\gamma>3$ however, most unusual features are absent and in many respects the scale-free network behaves like a random one (Barabasi and Oltvai 2004) .

1.3 Shortest Path Analysis

Distance in networks is measured with the path length, which tells us how many links we need to pass through to travel between two nodes. As there are many alternative paths between two nodes, the shortest path — the path with the smallest number of links between the selected nodes — has a special role (Barabasi and Oltvai 2004). In directed networks, the shortest distance l_{AB} from node A to node B is often different from the distance l_{BA} from B to A. For example, in Figure 5 $l_{BA} = 1$, whereas $l_{AB} = 3$. Often there is no direct path between two nodes. As shown in Figure 5, although there is a path from C to A, there is no path from A to C.

1.3.1 Common Algorithms for Calculating Shortest Path in the Network

- Dijkstra's algorithm — solves single source problem if all edge weights are greater than or equal to zero. Without worsening the run time, this algorithm can in fact compute the shortest paths from a given start point s to all other nodes (Cormen, Leiserson et al. 2001; Jungnickel 2002). Time complexity O (|V||E| log |V|) or O$|V^3|$.
- Bellman-Ford algorithm — solves single source problem if edge weights may be negative (Cormen, Leiserson et al. 2001; Jungnickel 2002). Time complexity O ($|V|^2|E|$)
- A* search algorithm — solves for single source shortest paths using heuristics to try to speed up the search (http://en.wikipedia.org/wiki/A-star_search_algorithm).
- Floyd-Warshall algorithm — solves all pairs shortest paths (Jungnickel 2002). Time complexity varies from O $|V^3|$ to O |E|.

In the present work we will deal with un-weighted graphs, hence the complexity would be O (|E|).

1.3.2 Path Finding with Breadth First Search Algorithm (BFS) and its Improvised Version

The standard algorithm of BFS can be defined as *naïve* method (non-heuristic) that aims to expand and examine all nodes of a graph logically in search of a solution. In other words, it exhaustively searches the entire graph without considering the goal until it finds it. From the standpoint of the algorithm, all child nodes obtained by expanding a node are added to a FIFO (First In First Out) queue. In typical implementations, nodes that have not yet been examined for their neighbors are placed in some container (such as a queue or linked list) called "open" and then once examined are placed in the container "closed" (Jungnickel 2002). The standard runtime complexity for a BFS to find shortest path is O (|V||E|).

A modified form of standard Breadth First Search (BFS) algorithm was proposed by Newman (Newman 2001) to calculate shortest path in an un-weighted network. The algorithm allows to calculate the shortest paths from *all* vertices to the target *j* in a single run, and not just from the single vertex *i as per* original algorithm. Thus we can calculate *n* shortest paths in time *O*(*m*), where *n* is the number of vertices in the graph.

1.4 Mean Path Length

The Average Shortest Path (ASP) (1.3), <*l*> or, represents the average over the shortest paths between all pairs of nodes and offers a measure of a network's overall navigability (Watts 1999). Another term used for such an analysis is Average Path Length (AL), which is defined as the shortest path length averaged for every pair of metabolites in the whole network (Jeong, Tombor et al. 2000).

$$ASP = \frac{1}{N(N-1)} \sum_{i,j} d(i,j) \quad with \quad i \neq j \tag{1.3}$$

1.5 Clustering Coefficient

In many networks, if node A is connected to B, and B is connected to C, then it is highly probable that A also has a direct link to C (Figure 4, Figure 5). This phenomenon can be quantified using the clustering coefficient (Watts and Strogatz 1998) $C_I = 2n_I/k \times (k\text{-}1)$, where n_I is the number of links connecting the k_I neighbors of node I to each other.

The average clustering coefficient <*C*>, characterizes the overall tendency of nodes to form clusters or groups. An important measure of the network's structure is the function *C*(*k*),which is defined as the average clustering coefficient of all nodes with *k* links. For many real networks *C*(*k*) ~ *k* – 1, which is an indication of a network's hierarchical character (Ravasz, Somera et al. 2002).

Note: The average degree <*k*>, average path length <*l*> and average clustering coefficient <*C*> depend on the number of nodes and links (N and L) in the network. By contrast, the *P* (*k*) and *C* (*k*) functions are independent of the network's size and they therefore capture a network's generic feature, which allows them to be used to classify various networks.

1.6 Types of Networks

Network models are crucial for shaping our understanding of complex networks and help to explain the origin of observed network characteristics. One can also plot the **relative frequency of each node** as a function of **the number of connections** of the nodes. Both axes in the plot use logarithmic scales. Using data from real networks or from simulations of different types of randomized networks, one obtains log-log plots that have characteristic shapes depending on the properties of the networks and how they were generated. Examples are shown in the Figure 6 cited by (Barabasi and Oltvai 2004).

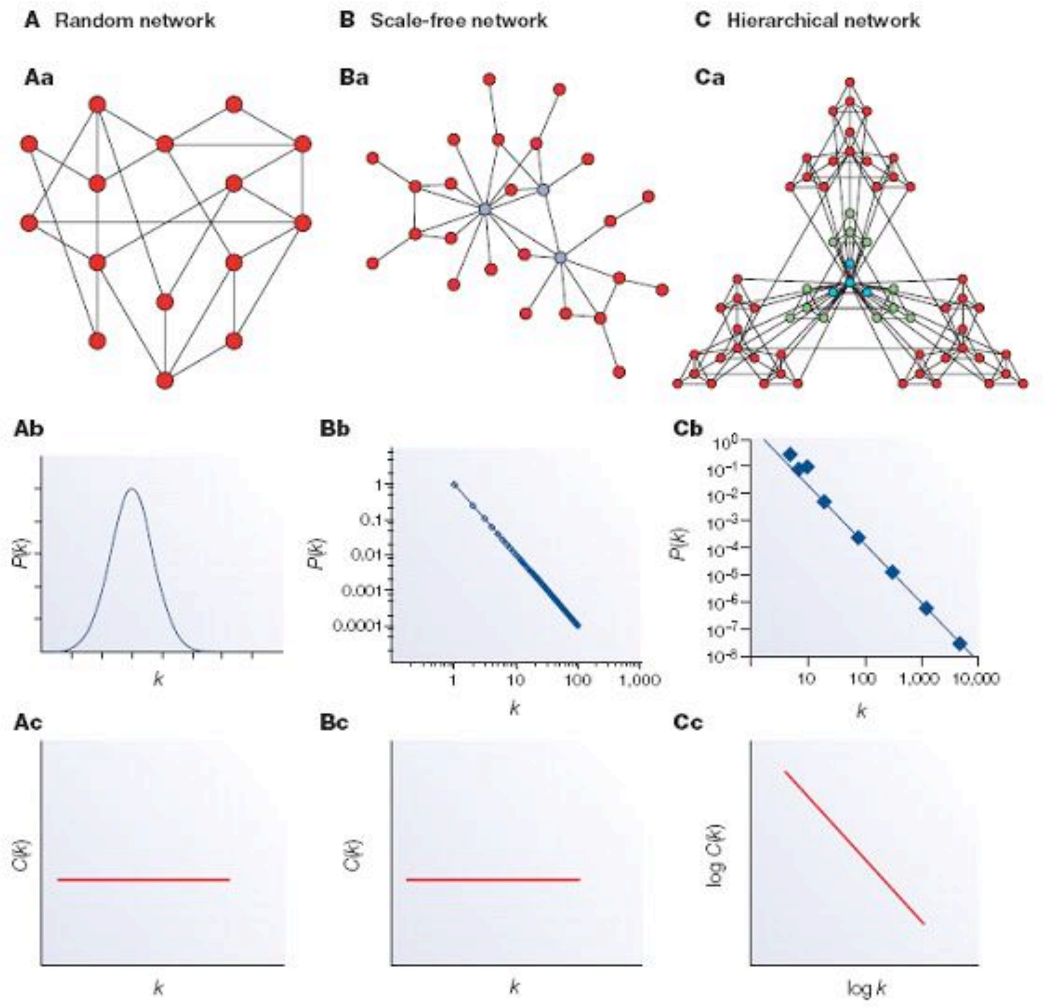

Figure 6. Three models had a direct impact on our understanding of biological networks. They were Random, Scale-free and Hierarchical networks

The **classical random network theory by Erdös and Renyi** (Erdös and Rényi 1960) states that given a set of nodes, the connections are made randomly between the nodes. This results in a network where most nodes have the same number of connections. The log-log frequency plot for this kind of network is shown above. (Figure 6).

Recent research has shown that this model does not fit the structure found in several important networks, such as the World-Wide Web, power grids, or social networks. Instead, a so-called scale-free, hierarchical model better describes these complex networks where most nodes have only a few connections, but a few nodes (called hubs) have a very large number of connections. Jeong, Barabasi *et al* (Jeong, Tombor et al. 2000), **indicate that metabolic networks are examples of such scale-free networks**. For example, Pyruvate would be a hub in metabolic networks as it is a metabolite that takes part in many chemical reactions. This result is important, and will probably lead to new insights into the function of metabolic networks, and into the evolutionary history of the networks.

The **characteristic shape of a scale-free, hierarchical network** can be represented in a log-log plot (Figure 6). Scale-free networks can be generated by growth and preferential attachment (Barabasi and Albert 1999). Starting with some connected nodes, at each step a new node is added to the network (growth). The new node establishes connections with the already existing nodes. However, nodes with many connections are preferred. This results in an imbalanced degree distribution, in which highly connected nodes of the existing network get even more new connections ("rich get richer") (Barabasi 2002). The network that is created by the Barabási–Albert model does not have an inherent modularity, so *C*(*k*) is independent of *k* (Figure 6). Scale-free networks with degree exponents $2<\gamma<3$, a range that is observed in most biological and non-biological networks, are ultra-small (Chung and Lu 2002; Cohen and Havlin 2003), with the average path length following $l \sim log\, log\, N$, which is significantly shorter than log N that characterizes random small-world networks.

1.7 Biological Networks

In the present “omics” era, it becomes increasingly more obvious that network analysis is essential for the analysis of genetic, proteomics and metabolomics data (Hartwell, Hopfield et al. 1999; Grigorov 2005). Large-scale, graph-based mathematical models have been developed to demonstrate the intrinsic hierarchical modularity of metabolic networks (Ravasz, Somera et al. 2002) and their robustness based on the shortest path analysis of the metabolic networks (Arita 2004; Barabasi and Oltvai 2004; Papin, Reed et al. 2004).

A typical metabolic network consists of reactions, metabolites and enzymes, which can be modelled using graph theory (Jeong, Tombor et al. 2000; Schuster, Fell et al. 2000; Girvan and Newman 2002; Oltvai and Barabasi 2002; Steinbeck, Han et al. 2003). These representations lead from a simple graph consisting of edges (reactions) and nodes (metabolites) or vice versa to a complex bipartite graph where two nodes (metabolites) share a common node (reaction/enzymes) (Rahman, Advani et al. 2005). Joining enzymes that share a common metabolite in a path can create enzyme-centric networks. The enzyme-centric view (Horne, Hodgman et al. 2004; Rahman, Advani et al. 2005) simplifies the representation of the metabolic network(Figure 7) by removing loose ends in the network (metabolites at the periphery of the network) and forming clusters of interacting enzymes. The gene-centric view has been successfully used in determining co-regulated genes in the metabolic and regulatory networks (Levchenko 2003; Covert, Knight et al. 2004; Luscombe, Babu et al. 2004; Ozbudak, Thattai et al. 2004; Barrett, Herring et al. 2005).

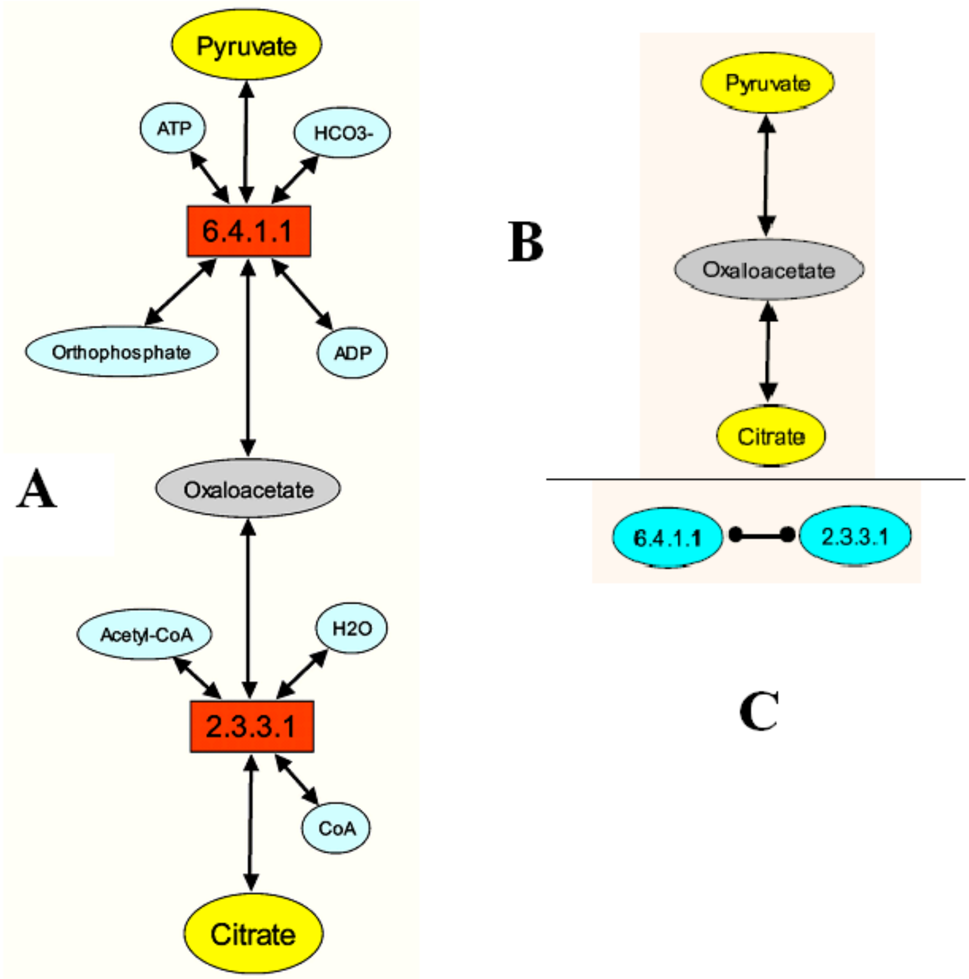

Figure 7. Bipartite view of one of the shortest paths between pyruvate and citrate in Bacillus subtilis 168 (A). Metabolic-centric view (B-top-right) and enzyme-centric view (C-bottom-right) of the above mentioned path.