VISUALIZATION AND ANALYSIS OF GENE EXPRESSION IN BIO-MOLECULAR NETWORKS

A thesis submitted in fulfilment of the requirement for
the degree of Doctor of Philosophy in
the School of Information Technologies at
The University of Sydney

David Cho Yau Fung
April 2010

Dedicated to
My long-suffering parents,
Alex Hon Ping Fung
and
Hoi Man Fong
Who worked hard to provide me a first-class education;

Also to my uncle and auntie,
Dick Ap
and
Jan Hing Ap
Who brought me to the Christian faith;

And to my Guardian and Creator,
Lord Jesus Christ
Who raised my understanding in
Biological Complex Systems
Through His Words.

Abstract

The research described in this thesis concerns with the visual exploration of gene expression in bio-molecular networks. Our focus is on investigating the merits of different visualizations as visual analysis methods with the eventual objective of hypotheses deduction. When designing each visualization, we focused on *concept model* visualization, i.e. each visualization is designed to capture a perspective of the biologist's understanding in molecular biology.

Our contributions span four areas: visual analysis, visualization methods, user evaluation, and analysis of hepatocellular carcinoma biology. In visual analysis, we contributed a visual analysis framework which captures the biologist's practice of incremental investigation (see Chapter 1). We demonstrate that the approach of "*filter first, zoom and details, overview if necessary*" can support hypotheses deduction in biology. As such, our visual analysis framework can provide an engineering framework for bioinformatics software designers in future.

In visualization methods, we designed the clustered circular layout for capturing the '*network within network*' organization of a GO_Process-defined Protein Interaction Network (see Chapter 4). We also designed the three-parallel plane layout as a novel method for visualizing the two-overlapping network (see Chapter 5). The uniqueness of our design is that, apart from the two heterogeneous bio-molecular networks G_1 and G_2, the overlap layer G_3 is explicitly visualized in the middle plane. The node set V_3 of G_3 is commonly shared by the node sets of G_1 and G_2, i.e. $V_1 \cap V_2$. Finally, we designed the circular plane layout as a novel method for visualizing the three-overlapping network (see Chapter 6). The uniqueness of our design is that the mappings between the three heterogeneous bio-molecular networks G_1, G_2, and G_3 are being explicitly visualized as inter-plane edges.

In user evaluation methods, we brought to the bioinformatics community the first set of benchmark tasks for evaluating usability of visualizations that display *gene_cluster*-GO relationships (see Chapter 3). These tasks define usability in terms readability and effectiveness in assisting analytical reasoning. They can be modified for evaluating data object-to-ontology relationships in any clustering pattern visualizations.

In cancer biology, we proposed a tentative explanation on how the protein-based gene regulatory interactions may co-operate with RNA-based gene silencing interactions and the *TGFB1* (transforming growth factor beta)-signaling interactions in promoting cancer growth (see Chapter 6). Our hypothesis can provide a direction to the cancer research community for future laboratory-based investigations.

In conclusion, we hope that this thesis will provide interested experts from the fields of bioinformatics, information visualization, and visual analytics, with a starting point for investigating visualization-related problems in molecular biology.

Acknowledgement

I would like to like to thank Dr Seok-Hee Hong (University of Sydney, Australia) for her supervision, Dr Kai Xu (ICT Centre, CSIRO, Australia) for his technical advice on coding the plugins for GEOMI, and Dr David Hart (Axogenic Proprietary Limited, Australia) for contributing half of the funding for this project as part of the Australian Research Council Linkage Grant LP0455334. I would also like to thank Professor Marc Wilkins (University of New South Wales, Australia) for his expert advice on the biological interpretation of protein interaction network visualization. I am also thankful to have a fruitful collaboration with Dr Falk Schreiber (Martin-Luther University, Germany) and Dr Dirk Koschützki (Furtwangen University of Applied Sciences, Germany) on investigating the problem of visualizing multiple heterogeneous networks. I am also indebted to Professor Ronald Trent (University of Sydney, Australia) and Associate Professor Graham Mann (University of Sydney, Australia) for their personal encouragement. Finally, I would like to thank the Australian Research Council for granting me the Australian Postgraduate Award Industry scholarship.

Last but not lest, I thank all the participants from the Ramaciotti Centre for Gene Function Analysis, Victor Chang Cardiac Research Institute, Viral Research Unit in the Prince of Wales Hospital, Westmead Millennium Institute, Drug Health Services (Research) in the Royal Prince Alfred Hospital, Institute for Biomedical Research, and the School of Biochemistry in the University of Sydney for taking part in this project. Without their efforts, this thesis would not be realized.

The research described in this thesis is supported by the Australian Research Council Linkage Grant number LP0455334.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

"A Picture means A Thousand Words" – OLD CHINESE SAYING

1.1. Motivation

One great advancement brought forward by the Human Genome Project in the last decade has been the increasingly routine application of high throughput technologies, e.g. DNA microarrays [97], protein-protein affinity microarray [115], CHiP-Chip array [13] and high-throughput mass spectrometry [165], in biological research. Their extensive application turns molecular cell biology from a data-poor to a data-intensive science within a decade. The copious amount of data generated continues to challenge the biologist's cognitive capacity to gain a holistic understanding on its biological meaning. The problem is two-folded.

The first has to do with the general conceptual view assumed by most biologists. Most of them have been trained in the reductionist approach towards biology. The term 'biologist' used in this thesis is defined as experts who research and study biology at the molecular and cellular level. They are often known as molecular biologists and cell biologists respectively. This view assumes that we can understand the biological meaning of a complete single-cell network by having it dissected down to its individual components. In other words, we can understand the whole if we know how many molecules there are and more importantly, the supposed biological function of each molecule. Such an assumption leads to the emphasis on the choice of data mining methods available to biologists. The rationale has been that the biological meaning of a large-scale gene expression dataset can largely be explained by a handful of differentially expressed genes. However, recent discoveries that many low-copy expressed genes are functionally important to cancer progression show the fallacy of this rationale [94].

The second reason has to do with the multi-disciplinary approach required. Each of the disciplines, i.e. statistics, data mining, information visualization, and molecular biology, contributes to parts of the solution in the process of biological research (see FIGURE 1.1). Yet very few professionals in any of these disciplines are practitioners of adjacent disciplines. For example, researchers in bioinformatics visualization often focus on designing new visualizations of data generated by statistical or data mining algorithms, rather than capturing one or more concept models in biology. To achieve the latter will require a background in biology.

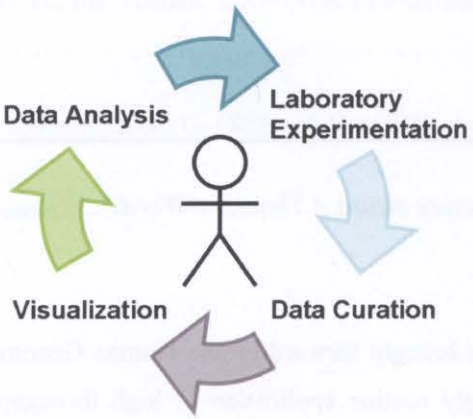


FIGURE 1.1. Key steps in a typical molecular biology research project in the post Human Genome Project era. A biologist represented by the ball-and-stick model needs to participate in a research process of four broad steps, i.e. (1) Laboratory experimentation, (2) Data curation, (3) Data and bio-informatics visualization, and (4) Data analysis. Computing supports every step in this process, and visualization serves as the medium for the biologist to proceed from data curation to data analysis.

Some visualization systems address the above problem partially by allowing the biologists to overlay gene expression values or expression correlation scores onto the nodes of a molecular network as colour hues. The protein interaction network (PIN) and metabolic-network (MN) are the two most commonly used. The assumption behind has been that such an integration should provide a glimpse on the systems-level interaction dynamics within a single cell. The molecular network is then generated as a node-edge network visualization in a variety of layouts, e.g. force-directed layout, circular layout, or grid layout. The same visualization also allows the biologist to integrate biological ontologies with the network in order to give it a functional context [104]. Without which, biologists will have difficulty deducing the meaning of the data, let alone hypothesis formulation. The latter is central to knowledge discovery [117].

While it is true that network visualization has indeed facilitated the conceptual shift from *reductionist* to *systems* biology, the current network visualizations have two limitations. The first is that most visualization systems provide a whole-cell molecular network as the first step in network exploration. This idea is deeply rooted in the information visualization mantra of "*overview, zoom and filter, details on demand*" [141] rather than the biologist's work practice of incremental investigation. Furthermore, the network visualizations are generated using generic layout algorithms commonly used for addressing the issue of scalability. Yet the layout algorithm does not account for any biological context such as biological processes, molecular function, intracellular distribution, pathways, and disease association.

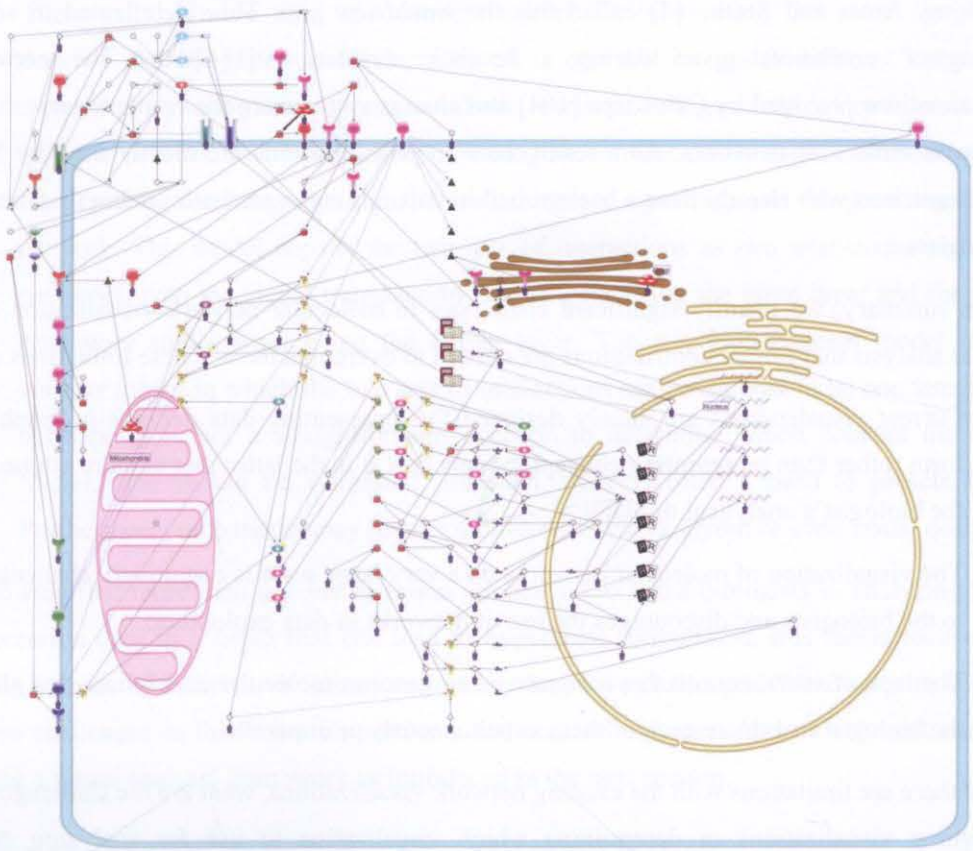


FIGURE 1.2. A computer-generated visualization of a human protein interaction network (PIN) superimposed on a cellular organization plan. Reproduced from Kojima et al. 2007 [87].

The result is often an unreadable network visualization that contains substantial edge crossings and node overlaps. The second is that most network visualizations can handle only one type of molecular interaction. The expanding variety of interaction types raises many practical and theoretical problems related to how these datasets should be integrated into the network representations (or models), how best to visualize the integrated network and how the network should be explored [148].

Using an integrated network of gene regulatory and protein-protein interactions as an example, the fact that the latter occur between proteins which are at least transiently co-localized in a particular cellular component may favour an approach that seeks to generate a visualization that mimics sub-cellular components (see FIGURE 1.2). However the same approach may be unsuitable for gene regulatory interactions where the interacting proteins may be involved in multiple distinct and distally located biological processes [148]. One can argue that this can be resolved by simply changing layouts. Yet there is very little understanding on how different network layouts influence biological reasoning.

The limitations discussed above impose a steep learning curve on bench biologists. They also expose the misalignment between the designer's intention and the biologist's analytical

thinking. Amar and Stasko [4] called this the *worldview gap*. This is reflected in some biologists' comments given during a heuristic evaluation [134] that the network visualizations provided by Cytoscape [104] and similar systems are the computer scientist's view on molecular networks. As a result, network visualizations are mostly used by bio-informaticians who already have a background in data mining or statistics, rather than bench biologists.

In summary, we identify significant challenges in molecular network visualization and visual analysis that further contributions are needed to overcome them. These limitations are:

1. Current visualizations are mainly designed for representing data patterns in graphical form rather than representing concept models. Yet it is the latter that is more attuned to the biologist's analytical thinking.
2. The visualization of molecular networks on a very large scale is cognitively challenging to the biologists and discourages the use of networks in data exploration.
3. The lack of visualizations that integrate heterogeneous molecular interactions and allow the biologist to explore each of them simultaneously or in parallel.

If there are limitations with the existing network visualizations, what are the challenges in designing visualizations or determining which visualization to use for analyzing gene expression? There are at least four challenges to answer.

1. The first challenge is the "curse of scale". While molecular interactions can be mapped to network representations, network visualizations worked well for small networks. As the molecular network approaches a few thousand nodes, node overlaps, node label overlaps, and edge crossings make the network visualization confusing and unreadable [61]. Protein interaction network (PIN) is the case to the point. Though it can be represented by a simple node-edge graph, the scale of even a bacterial PIN ($|V| \geq 5000, |E| \geq 6000$) is beyond the human cognitive capacity to comprehend and therefore presents a steep challenge to layout design and interactivity design.
2. Filtering is often used as a solution for reducing scale. Network reduction inevitably leads to the loss of information but will also reduce visual complexity. The challenge is to find a trade off between information loss and visual complexity, that is acceptable to biologists with diverse motivations.
3. Biologists like to cluster data according to their ontological classification, e.g. Gene Ontology (GO). However, explicit visualization of the original network as inter-connected clusters of molecules demands a change in network layout, which may affect the biologist's analytical reasoning.

4. There can be more than one concept model for interpreting the same *integrated network*. An integrated molecular network is one that integrates different interaction types, e.g. signaling interactions, metabolic reactions, and gene regulatory interactions, into one network. For example, bench biologists conventionally use the *cascade* model for depicting the integrated signal transduction (or signaling) and the gene regulatory network. This model depicts the two sets of interactions as two inter-connected sub-networks, with the signal transduction sub-network being the input layer and the gene regulatory sub-network being the output layer. The emerging concept model is the *systems* model in which the two sets of interactions are considered to be one integrated network. It is not a straightforward decision to determine which concept model is '*better*' and should be visualized. As such, there is often a need to present both. Furthermore, each model may require a different network layout or even visual design.

Knowing that more than one visualization are needed to assist biologists in analyzing gene expression data to a depth that can lead to hypothesis formulation, this thesis focuses on using a series of different visualizations to meet the above challenges. We attempt to answer these challenges in four research problems and conduct our research in a coherent manner using a visual analysis framework as introduced in the next section.

1.2. A Visual Analysis Framework for Molecular Biology

Nowadays, biologists perform microarray experiments for measuring gene expression on the genome-wide scale. He/she will then perform the two pre-processing steps, i.e. normalization and filtering to extract a set of quality data amendable to data mining or statistical analysis [145]. Following that, the biologist may want to extract a set of co-expressed genes. Then he/she will apply pairwise correlation coefficients such as Pearson or Spearman correlations to the dataset. Our framework is designed to mimic a series of visual analysis tasks with the assumption that a bench biologist would like to investigate gene co-expression as the first step (see FIGURE 1.3).

The objective of our proposed framework is to assist biologists in deducing biological hypotheses incrementally using visual analysis. The motivation behind is to provide a framework that is based on modeling the biologist's work practice, i.e. *filter first, zoom and details*, and *overview if necessary*. The visual analysis framework consists of three steps, with each step focusing on a different type of network visualization, i.e.

1. visual analysis of co-expressed gene clusters;
2. visual analysis of protein interaction networks;
3. visual analysis of integrated network.

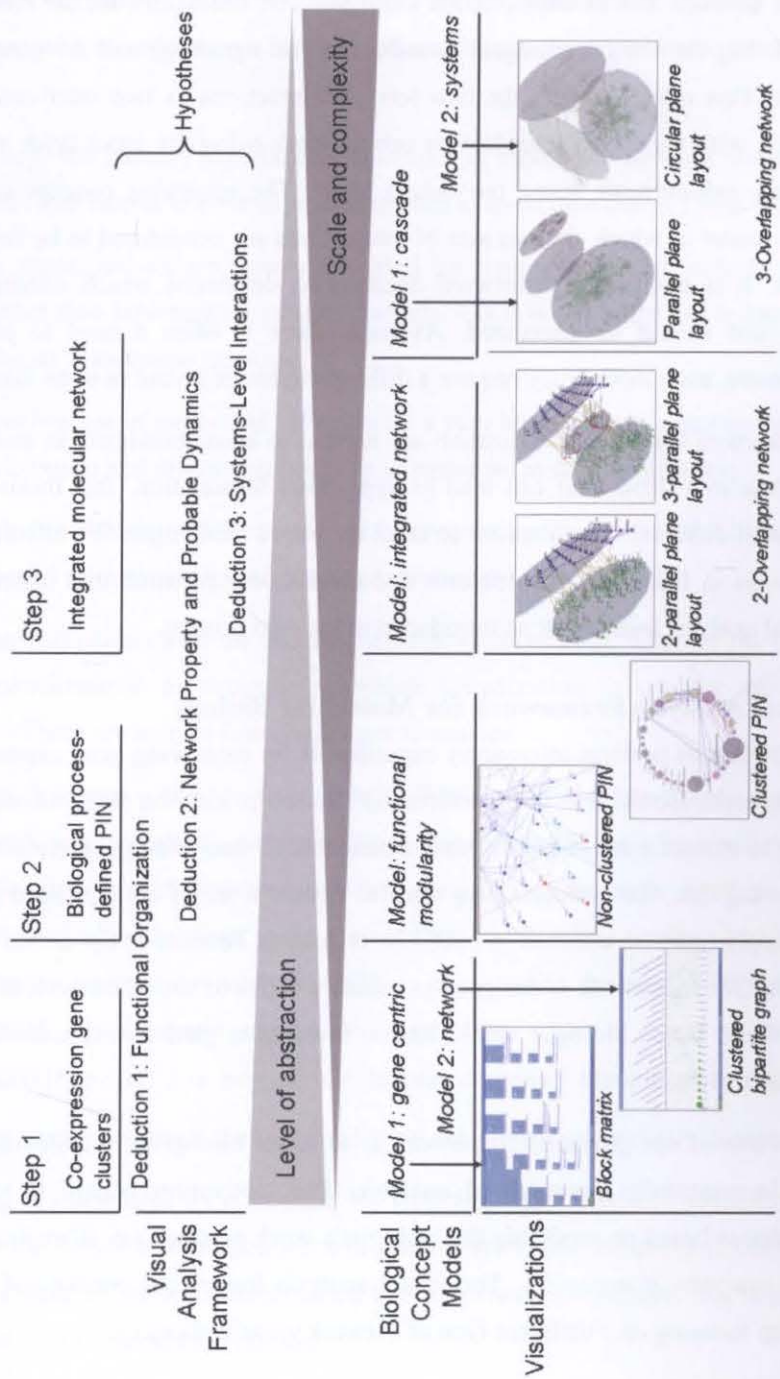


FIGURE 1.3. A schematic representation of our proposed three-step visual analysis framework in relation to the biological concept models and the visualizations studied in each step. The visualizations used in each visual analysis step decrease in their level of abstraction but increase in scale and visual complexity.

Each step is a visual analysis task that employs one or more computer-generated visualizations. The purpose of each task is to support biological deductions through visual analyses. The level of abstraction represented by the visualization decreases with each step but increases in scale and visual complexity (see FIGURE 1.3). Therefore, the biologist uses increasingly informative but complex network visualization to extract biological meaning from the dataset on hand. In this way, any deductions made in one visual analysis step should assist the visual analysis in next step and so forth. This should reduce the cognitive challenge that the biologist has to face when analyzing a large gene expression dataset, while at the same time, allow them to perform progressively in-depth biological analysis. In this way, hypothesis formulation is achieved by the collaborative use of multiple visualizations. In the following sub-sections, the objective that each visual analysis step can achieve for the biologist and its role in the entire framework will be explained. During our elaboration, the various research problems will also be brought out followed with our proposed solutions.

1.2.1. Research Objectives and Rationale

1.2.1.1. Visualization and Analysis of Gene Ontology-annotated Co-expressed Gene Clusters

Co-expression analysis is often performed on gene expression data as the first step, simply because it implies synchronized expression of genes. Furthermore, the selection of co-expressed genes using correlation coefficients has the side benefit of filtering down a large dataset by 90%. When clustered in groups using shared GO Process category as the criterion, the resulting set of clusters informs the biologist on the set of co-regulated biological processes. This provides an abstraction on the functional organization of a cell at a level higher than any bio-molecular network visualizations. In other words, GO-annotated co-expressed gene clusters is to give the biologist a glimpse of the functional organization within a single cell without having to examine the underlying molecular network. The significance of this step is that it assists the biologist to prioritize the biological processes for further investigation in the next visual analysis step because the clustering pattern when visualized will inform the biologist on the relative activity between different biological processes. Therefore, any visualization applied in this step must effectively display the *gene*-to-*gene* co-expression and the *gene_cluster-biological_process* relationships.

The issue of contention, however, is that the same set of gene clusters can be viewed in two different presumed biological concept models. The first is the *gene-centric* model. Simply speaking, this model assumes that knowing the biological process(es) a particular cluster of genes belongs to is adequate for making biological deductions. The second is the *network* model which assumes that the biological processes are inter-connected and are held together by the co-expressed gene clusters. These diverse views call for two different representations to be visualized. We introduce a non-graph matrix-like representation for

capturing the first concept model and a clustered bipartite graph for capturing the second model. As a consequence, a few questions arise.

How do different representations affect biological reasoning? Which representation when visualized is better in assisting biological deduction? What analytical tasks does each visualization analysis support? The last two questions define the meaning of usability in the context of biological analysis. Answering these questions is crucial as the outcome will either validate the currently popular approach of color matrix visualization or the alternative approach of network visualization. As such, the outcome will either support or challenge the use of network visualizations in the last two steps of our framework (see FIGURE 1.3). In the light of this, we attempt to answer the posed questions using a case study and a task-oriented user evaluation. In the case study, visual analysis will be conducted using a published set of co-expressed genes which came in two sample sets, hepatocellular carcinoma (HCC) and normal hepatocytes [58]. In the user evaluation, we evaluate the usability of each representation by measuring user performance in three different variables, task completion time, accuracy and user satisfaction score. The user evaluation studies published to date are limited to visualizations in microarray analytical software [134, 135]. This is the first time that a comparative evaluation on concept-based visualizations has been conducted.

1.2.1.2. Visualization and Analysis of GO-defined Protein Interaction Networks

Since biological processes are largely driven by *protein-protein* interactions, the next step is to examine the protein interaction network (PIN) that drives each of the biological processes identified in the previous step. The rationale behind visualizing gene co-expression in the context of a protein interaction network (PIN) is to provide a more detailed view on the protein-protein interactions required for the functioning of co-regulated biological processes. More importantly, the biologist wants to identify proteins that are coded by co-expressed genes. Co-expressed genes often imply comparable molecular abundance of their respective proteins. If a pair of co-expressed genes is shown to interact with each other in the PIN, that implies their protein-protein interaction is active and the proteins are co-regulated.

Another important piece of information is the intracellular distribution of the proteins, i.e. the spatial information for a biological process-defined PIN. For many biological processes to occur, their protein-protein interactions have to be co-localized in certain cell components, e.g. nucleus, cytoplasm, mitochondrion, and etc. When clustered in groups using the GO Cell Component ontology, the resulting clusters inform the biologists on the spatial organization of a particular biological process. Therefore, visualizations used in this step must display not only the *protein-protein* interactions and the gene-gene co-expression relationship but also the *gene_cluster-cell_component* relationship.

The significance of this step is to provide the biologist with the first glimpse on the network dynamics of each biological process. If the biologist examines a set of related process-specific PINs, he/she will be able to deduce the intersecting subset of protein-protein interactions that connects the set of biological processes of interest. Any co-expression between interacting proteins will suggest that they are co-regulated such that the processes can be co-activated. The intracellular distribution of co-expressed genes in multiple cell components often raises questions about protein transportation which is often used for regulating the activity of biological processes. The limitation of this approach is that the biologist is viewing a canonical PIN with a gene expression correlation score overlay rather than a real time PIN because the latter cannot be detected by microarray technology.

There are currently two challenges in PIN visualizations. The first challenge is the network scale that needs to be visualized. Current solutions can be categorized to two approaches. One approach employs layouts that took biological knowledge into account [71, 148]. Another approach uses data filtering [178]. The second challenge is to integrate the GO annotations with the PIN visualization, in especially for capturing the functional modularization of the PIN. Proteins belonging to a functional module can be defined by their membership in a biological process. Such a functional module can itself be subdivided into smaller modules (or sub-networks) if its proteins co-exist in multiple cellular components. This type of modularization is known as '*network within network*' or '*module within module*'. The use of colour coding in correspondence to GO category membership has been the most common solution. If a protein is annotated with terms from multiple GO categories, the node will be coloured as a pie chart [178]. However, this approach faces the challenges of readability and colourability. Another solution is to partition the PIN according to GO category membership [7]. However, this approach can only present partitioning according to a single GO category and is suitable only for visualizing a PIN with only a few hundred nodes.

To answer the first challenge, we will filter the human PIN using the GO Biological Process ontology. We called the filtered PIN a GO_Process-defined PIN. We then visualized the same GO_Process-defined PIN using two different methods, i.e. non-clustered PIN and clustered PIN visualizations. The first method is to visualize the GO_Process-defined PIN in a force-directed layout. This is the layout generally used for PIN visualizations. As the second method for visualizing the GO_Process-defined PIN, we design and implement the *clustered circular layout* as a new layout. On the one hand, the non-clustered PIN visualization uses a generic layout that does not take into account any further partitioning by biological knowledge. On the other hand, the clustered PIN visualization uses a clustered layout that represents the distribution of proteins in a variety of cellular components. Each

cellular component is represented by a cluster node label derived from the GO Cell Component ontology. By this method, we can present the physical localization of the protein-protein interactions for every known GO_Process in a human cell. The result will be a visualization that models closely to the '*network within network*' property of PIN. This will be a crucial feature for expanding the biologist's cognitive ability.

The availability of two different visualizations for the same GO_Process-defined PIN brought to our attention a few questions. How do different visualizations affect biological reasoning? Which visualization is better in assisting biological deduction? What analytical tasks can each visualization support? What biological insights can be generated out of PIN visualizations with gene expression overlay? To investigate these questions, we conduct a case study on HCC again and also a domain expert evaluation on the non-clustered PIN and the clustered PIN visualizations. In the case study, the co-expressed genes of the HCC sample set are overlaid onto the full-scale human PIN. We then select seven GO Process-defined PINs and perform a visual analysis on each using both the non-clustered and the clustered PIN. These GO Processes are biological processes underlying the characteristics defined by the current cancer model [68].

1.2.1.3. Visualization and Analysis of Two-Overlapping Integrated Networks

PIN visualization allows the biologist to make deductions based on physical interactions between proteins. However, it is not a complete view of the cellular molecular network for two reasons. The first reason is that protein-protein interactions have specialized molecular functions. Some are signaling proteins which relay an activation/de-activation signal from protein to protein. Others are metabolic enzymes which catalyze the conversion of metabolites. The second reason is that there are other interaction types besides protein-protein interactions. For example gene regulation requires protein-DNA interactions. With the discovery of *non-coding* genes (also known as RNA genes), RNA-RNA interactions become the latest addition to the full-scale molecular network.

To provide the biologist with a more complete view of the molecular network, a visualization which integrates the variety of interaction types becomes necessary. We called such a network an *integrated network*. The significance of this step is to provide the biologist with an overview of the molecular network so that he/she can add further details to the deductions made in the previous steps. This should assist the biologist in finalizing hypotheses formulation.

When trying to visualize an integrated network, the issues of scale and scalability will arise. If all the interaction types are visualized in a single large network, the biologist will be faced with a myriad of edges in a variety of visual encoding. The other problem is how to

visualize proteins that have multiple molecular functions and are expected to participate in multiple interaction types. For example, a certain bacterial protein can be both a metabolic enzyme and a gene regulator. In this case, the protein has to be visually represented as separate nodes, each in a distinct colour hue or in a unique shape. In addition, an undirected edge, also distinctly coloured, has to be included to show that the two nodes are in fact representations of the same protein. Even more problematic is trying to identify such proteins in a large and complex network visualization.

In view of these problems, we introduce the 2.5-dimensional two-overlapping network visualization in which each of the heterogeneous networks is drawn on a separate plane. Each heterogeneous network is of a distinct interaction type. The nodes and edges of each network are distinctly coloured. The planes are stacked in parallel in the 2.5-dimension. The two-overlapping network can also come in the two-plane and the three-plane representations. The difference between them lies in the additional overlap layer in the latter representation. The two-plane representation requires the *two-parallel plane* layout. The three-plane representation requires the *three-parallel plane* layout. Both are new layouts designed and implemented for visualizing the two-overlapping network.

Here again, we ask the questions: which of the two visualizations is better in assisting biological deduction? Which of them are more readable at different network sizes? What analytical tasks can each visualization support? In a more general scope, we also ask whether the two-overlapping network visualization, no matter which representation, is able to capture the current biological knowledge, i.e. as a visual knowledge representation. Furthermore, we ask whether it is effective for formulating new hypotheses, i.e. as a knowledge discovery method.

To answer the above questions, we conduct two case studies to find anecdotal evidence for supporting our choice of using the two-overlapping network as a method for visualizing an integrated network. The first case study employs network data from *E. coli*. Since it has been very well studied, *E. coli* is an excellent case for testing the effectiveness of any visualization as a visual concept model. The second case study is an extension of the HCC study conducted with the GO Process-defined PINs. The intention is to make new deductions that should be complementary with the latter. In each case study, we also compare the two visualization methods for their usability.

1.2.1.4. Visualization and Analysis of Three-Overlapping Integrated Networks

The rationale behind visualizing the three-overlapping network is that the two-overlapping network visualization is still not a complete view of an integrated network. Like its two-overlapping counterpart, each sub-network in the three-overlapping network is of a distinct

interaction type drawn on separate planes. However, the three-overlapping network can come in two different representations, a *parallel plane* representation and a *circular plane* representation. Each captures a different biological concept model. The parallel plane representation captures the *cascade* model. This model assumes that the three sub-networks have a linear functional order. In other words, the interactions within network G_1 will influence interactions within network G_2 which in turn will influence interactions within network G_3 and vice versa. The circular plane representation captures the *systems* model which assumes that the three sub-networks function co-operatively. That means the interactions within each network will influence those in the other two networks. Each representation requires a different layout when visualized. The parallel plane representation requires the *parallel plane* layout much like the one seen in the two-overlapping network visualization. The circular plane representation requires the *circular plane* layout which has the three planes arranged in a triangular formation.

With two different visualization methods capturing two different biological concept models, a number of questions arise. With more inter-connected networks to analyse, what combination of networks can help the biologist to initiate the analytical process? Which visualization is better in assisting biological deduction? What analytical tasks can each visualization support? Again, we investigate these questions using *E. coli* and human networks as case studies. It is noteworthy to state that our work on the two- and three-overlapping networks is highly experimental. Although multi-plane (or level) biological networks have been investigated before [14], the use of heterogeneous networks as multi-layers has yet to be experimented.

1.3. Research Methodology

The research methodology followed the three-step process of design, implementation and evaluation.

1.3.1. Design

We first design a visual analysis framework that captures the biologist's practice of incremental investigation. We then design the visualization methods for each step of the visual analysis framework. The choice of visualization design is based on the concept models used by biologists in their analytical reasoning. Each design is intended to capture either a certain biological concept model or to capture two different perspectives of the same concept model (see FIGURE 1.3). In Chapter 3, we design the block matrix to capture the gene-centric view of the cell's functional organization, and the clustered bipartite graph to capture the network view of the same functional organization. In Chapter 4, we design the clustered circular layout for visualizing the GO_Process-defined PIN. The design is intended

to capture the nested modular organization of the PIN. In Chapter 5, we design two methods for visualizing the two-overlapping network, i.e. two-parallel and three-parallel layouts. The two-parallel layout captures the direct mapping between the two heterogeneous bio-molecular networks using inter-plane edges. The three-parallel layout includes an overlap layer that explicitly represents the nodes commonly represented by the two heterogeneous bio-molecular networks. In Chapter 6, we designed two methods for visualizing the three-overlapping network, i.e. three-parallel plane and circular plane layouts. The three-parallel plane is designed to capture three bio-molecular networks as a path whereas the circular plane layout is designed to capture the same networks as a cycle.

1.3.2. Implementation

The drawing algorithms are implemented as plug-ins to the network visualization tool GEOMI [2]. All the visualizations are implemented as prototypes. For the first step in our visual analysis framework (see FIGURE 1.3), the prototypic implementations of GO-annotated gene clusters generate static visualizations. This is intentionally done to satisfy the design requirements of the user evaluation. However, the implementations in the subsequent visual analysis steps become more sophisticated. For the final step in the framework, navigation by rotation and zooming are provided with the overlapping network visualizations. Case studies are performed on the implemented visualizations to evaluate their effectiveness as visual analysis methods. Each study involves experimenting and analyzing a visualization with one or more publicly available biological datasets as the input. Each visual analysis relates the biological deductions to the design of the visualization used. The overall objective here is to evaluate the merits (strength and limitations) of each design as a visual analysis method.

1.3.3. Evaluation

Of the three visual analysis steps in our framework, we conduct user evaluations on the prototypic implementations for the first two steps. Since clustering gene expression data using GO Process as a criterion has been widely practiced, we have accumulated enough understanding on the biologist's analytical objectives in general. This allows us to design analytical tasks that closely mimic those in the real life scenario, and in addition able to recruit a pool of biologists for the evaluation[a]. All these allow us to collect user performance data and opinions useful for deducing design guidelines. We make an additional survey on the same group of biologists and find that none had experience in reading PIN visualizations. For this reason, we conduct an expert evaluation in which the biologist will be asked to evaluate a GO_Process-defined PIN visualization according to a set of evaluation criteria.

[a] The Human Research Ethics Committee approval number 9418.

They are modified from the published heuristics originally designed for evaluating pathway visualizations [135]. Because the two- and the three-overlapping network visualizations are relatively new, there are no biologists experienced in applying them to visual analysis. Furthermore, they are intended for in-depth analysis. Any analytical task will require long hours to complete. Hence, it is not practical for us to conduct a user evaluation. Instead, we rely on the case studies to provide anecdotal evidence on the usability of the two-and the three-overlapping network visualization.

1.4. Contributions

The contributions of this thesis fall into four areas: visual analysis, visualization methods, user evaluation, and analysis of hepatocellular carcinoma biology.

1.4.1. Visual Analysis

We made two contributions to visual analysis as follows.

- The visual analysis framework (see FIGURE 1.3) which models after the biologist's practice of incremental investigation. We changed the conventional information visualization mantra of "*overview, zoom and filter, details on demand*" [141] to "*filter first, zoom and details, overview if necessary*". Using a series of visualizations with decreasing levels of abstraction, our framework has the advantage of guiding the biologist step-by-step into network exploration. Hypotheses deduced this way are more likely to be based on one or more biological concept models than on statistical scoring or data mining output alone.
- The experimentation of different visualizations in each step of the visual analysis framework also provides an understanding on how '*design influences reasoning*' in biological analysis.

1.4.2. Visualization Methods

In regards to visualization, we contribute new methods for visualizing bio-molecular networks as follows:

- We design the clustered circular layout for capturing the '*network within network*' organization of a GO_Process-defined PIN (see Chapter 4). Our algorithm automatically generates sub-networks enclosed within cluster nodes with each cluster node representing a sub-cellular component.
- We design the three-parallel plane layout as a novel method for visualizing the two-overlapping network (see Chapter 5). The uniqueness of our design is that, apart from the two heterogeneous bio-molecular networks G_1 and G_2, the overlap layer G_3 is

explicitly visualized in the middle plane. The node set V_3 of G_3 is commonly shared by the node sets of G_1 and G_2, i.e. $V_1 \cap V_2$.

- We design the circular plane layout as a novel method for visualizing the three-overlapping network (see Chapter 6). The uniqueness of our design is that the mapping between the three heterogeneous bio-molecular networks G_1, G_2, and G_3 is being explicitly visualized as inter-plane edges. Two algorithms are designed for handling two possible cases. The first case is *fixed-free-fixed* where G_1 and G_3 are drawn in a given layout whereas is drawn using the force-directed layout [44]. The second case is *free-fixed-free* where only G_2 is drawn in a given layout whereas G_1 and G_3 drawn in the force-directed layout [44].

Our visualization methods on overlapping networks resolve the issue of scale and visual complexity that arise from integrating multiple interaction types and node types in a single network visualization. This contribution is not only applicable to molecular networks but also in other domains.

1.4.3. User Evaluation

We also make two contributions to user evaluation.

- We bring to the bioinformatics community the first set of benchmark tasks for evaluating usability of visualizations that display *gene_cluster*-GO relationships. These tasks define usability in terms readability and effectiveness in assisting analytical reasoning. They are also designed with an understanding on how biologists interpret gene_cluster-GO relationships (see Chapter 3).
- Our experience in conducting user evaluation may inform other practitioners what measurements are useful for assessing user performance when the tasks involve analytical reasoning in a knowledge-intensive domain like biology.

1.4.4. Analysis of Hepatocellular Carcinoma Biology

Our contribution to hepatocellular carcinoma (HCC) biology is the hypothesis generated from our visual analyses (see Chapters 4, 5 and 6). Our hypothesis provides a tentative explanation on how the protein-based gene regulatory interactions co-operate with RNA-based gene silencing interactions and the *TGFB1* (transforming growth factor beta)-signaling interactions in promoting cancer growth. This hypothesis should provide a direction to the cancer research community for future laboratory-based investigations.

1.5. Thesis Organization

This thesis is organized in seven chapters. In Chapter 2, the general background knowledge on information visual analysis, molecular network visualization, user evaluation studies, and systems biology is given.

In Chapter 3, the two visual representations of GO-annotated co-expressed gene clusters are described. The design criteria and the drawing algorithm of each representation are introduced and also the case study on HCC is presented. We also present the design and results of the task-oriented user evaluation. These results have been presented in a visualization conference [54].

In Chapter 4, the non-clustered and the clustered visualizations of the GO Process–defined PIN are described. The drawing algorithm for the layout of each visualization is introduced. The concept model of cancer biology [68] is introduced in the case study which served as a guide for our visual analysis. The HCC dataset used in Chapter 3 is re-applied to the case study. This is followed with an elaboration on the results of a domain expert evaluation.

Chapters 5 and 6 are where we introduce the visualization problem of *overlapping network*. In Chapter 5, the two representations of the two-overlapping network are elaborated. The drawing algorithm for visualizing each representation is introduced. Two case studies are presented. The first concerns with *E. coli* networks in two combinations and the second concerns with human cancer-related networks in two combinations. The result of the visual analysis conducted in each case study is also presented. This work has been published in a bioinformatics journal [55].

In Chapter 6, the two representations of the three-overlapping networks, the drawing algorithms and the two use cases are elaborated. The first use case concerns with the integration of three *E. coli* networks. The second concerns with the integration of three human cancer-related networks. This work has been accepted for oral presentation in a visualization conference paper [56].

Finally, general conclusions on the research results in Chapters 3 to 6 are presented in Chapter 7 and so are the directions for future work.

{End of Chapter 1}

CHAPTER 2

Background

"Graphics reveal Data"—EDWARD R. TUFTE

2.1. Information Visualization

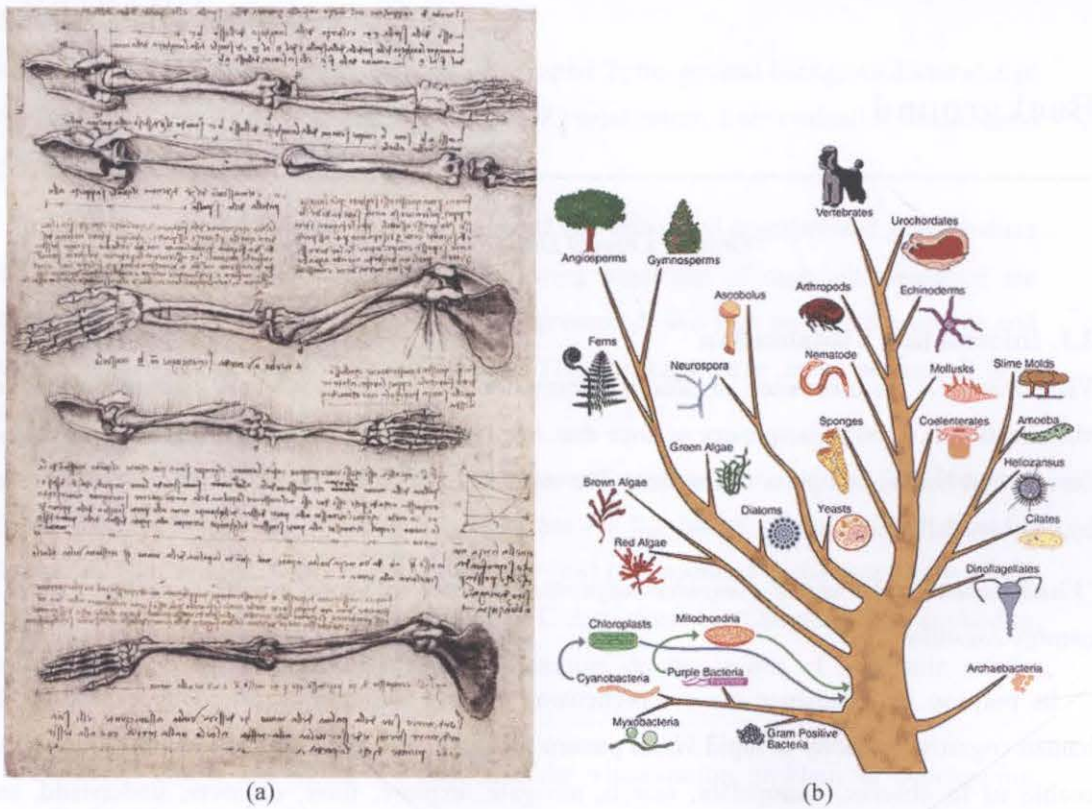
Visualization is the translation of data or information into graphics whereas computer-generated visualization is a multidisciplinary science that involves computer science, psychology, graphics design, and human-computer interaction. The most concise definition of *visualization* was given by Card *et al.* [22].

'*Visualization is the use of computer-supported, interactive, visual representations of data to amplify cognition.*'

Its purpose is to enhance the comprehension of data and/or information by exploiting the human cognitive capacity in rapid visual pattern recognition [159]. Effective visualization should enable us to observe, manipulate, search, navigate, explore, filter, discover, understand, and interact with large volumes of data more effectively to discover hidden patterns [61]. The mandate of visualization research is to search for new methods for encoding data in graphical forms so that the human user can comprehend, navigate and manipulate the data.

Visualization is historically divided into two categories, i.e. *scientific visualization* and *information visualization* [22]. Their categorization is primarily based on three criteria:

1. *Domain*: Is the domain scientific or non-scientific?
2. *Data*: Is the data physically based? The primacy of scientific visualization is to impart a visual representation to data that are measurements of physical objects, e.g. wind tunnel vector data. In contrast, the primacy of information visualization is to impart a visual representation to data that are an abstraction of information, e.g. document collections.
3. *Spatialization*: Is the spatialization given or chosen? In scientific visualization, the data is inherently spatial because of its physical nature. Therefore, the visual representation has a given spatialization. In information visualization, the data involved is not inherently spatial. For this reason, there is the need to design a spatial topology for the synthetic visual representation. In this case, the objective of spatialization is often to leverage the user's cognitive ability to unpack information out of the visual representation. In other words, the spatialization is chosen rather than given [127].



(a) (b)

FIGURE 2.1. The earliest forms of visualization in biology. (a) Anatomy of the human arm drawn by Leonardo Da Vinci in the 16th century (*circa* 1510). (b) The phylogenetic tree of the animal and plant kingdoms.

In summary, scientific visualization focuses on visualizing physical data that are inherently spatial. On the other hand, information visualization focuses on information which is often abstract and in many cases does not automatically map to the physical world.

2.2. Bioinformatics Visualization

Visualization has been applied to biology and medicine for a long time. There are examples in scientific visualization and also examples in information visualization. The manual drawing of the human anatomy was the earliest form of scientific visualization (see FIGURE 2.1(a)) because it visually represents the physical and spatial structure of the human body. The drawing of the taxonomy in the animal and plant kingdoms is the earliest example of a combined scientific and information visualization (see FIGURE 2.1(b)). It is a scientific visualization because the plants and animals are physical objects. It is also information visualization because the evolutionary relationships between the objects are a human projection.

In modern medicine, the most prominent examples of pure scientific visualization have been medical imaging. Technologies such as computerized tomographic (CT) scan, magnetic resonance imaging (MRI) and ultrasound scan render a visual representation of the tomographic volume data generated by electromagnetic emitting devices.

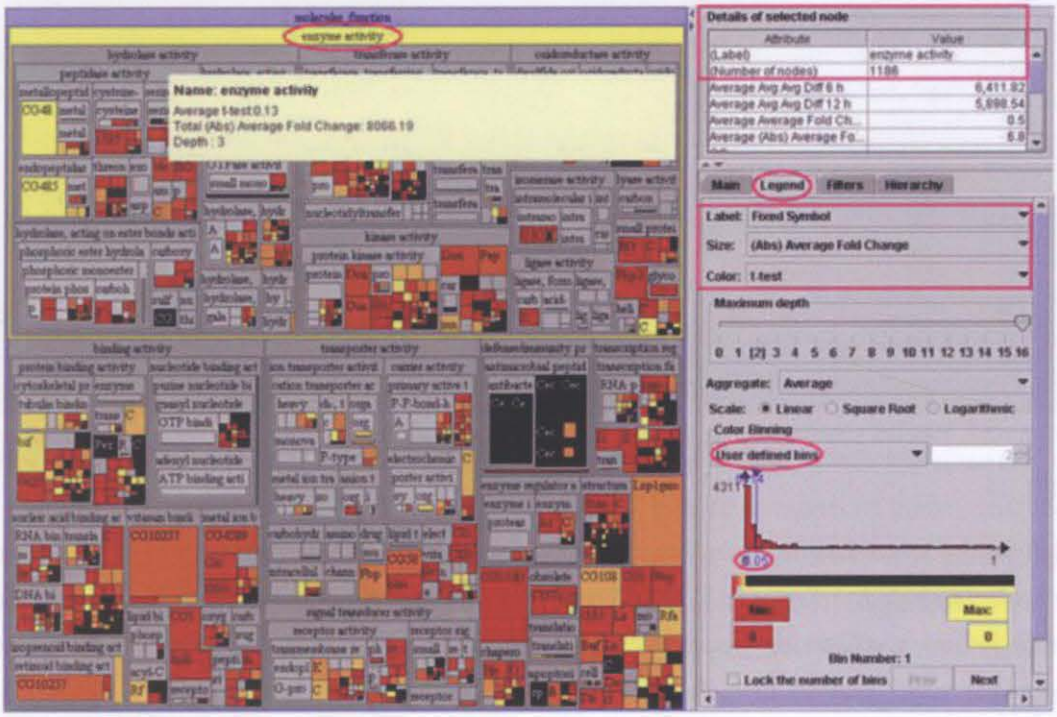


FIGURE 2.2. Visualization of microarray data in the context of the Gene Ontology hierarchy using Treemap. Reproduced from Baehrecke 2004 [5].

In molecular biology, the best example is the visualization of protein molecules constructed from X-ray diffraction data. In these examples, spatial information is crucial to the biologist's understanding of biology. There are a few if uncommon examples of pure information visualization in medicine and biology e.g. Gene Ontology visualized in Treemap [5] (see FIGURE 2.2), and medical literature visualized in three-dimensional contour map [117] (see FIGURE 2.3).

Rhyne [127] commented that information visualization is no less important than scientific visualization when it comes to genomics simply because of the qualitative and transient nature of the knowledge. Indeed, scientific visualization alone cannot enhance the user's understanding of the data generated by high-throughput technologies such as DNA microarray, protein array, and DNA sequencing because the spatial information inherent with the dataset has no biological meaning and the numerical or nominal values within the dataset are uninformative without some form of abstract (or contextual) data attached to it. That is because the abstract data assist the biologists in explaining the scientific data. In the view of this, we can give *bioinformatics visualization* a concise definition.

'*Bioinformatics visualization is scientific visualization that involves biological information visualization.*'

Because bioinformatics visualization contains both scientific and abstract data, it will serve as a tool for communicating a biological concept to the biologist and for exploring the data to the point where hypotheses can be formulated. Usually, the hypotheses are based on novel biological-

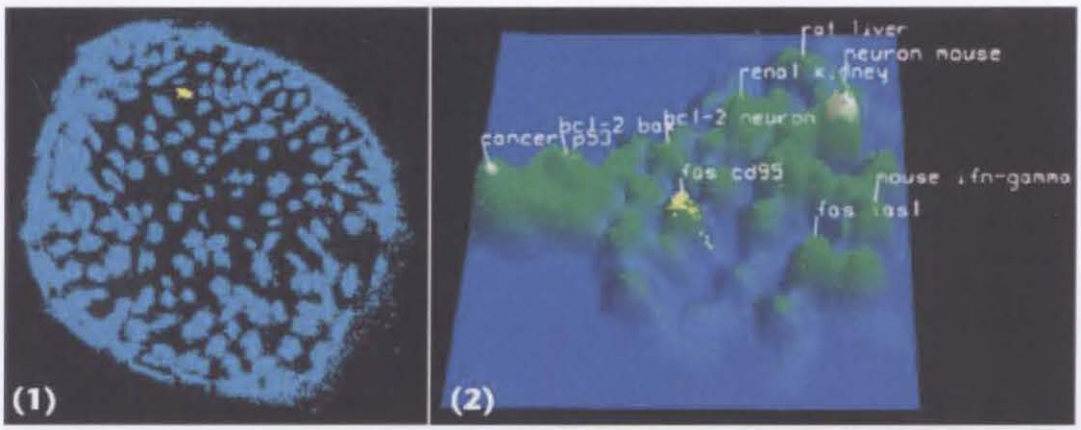


FIGURE 2.3. Parallel visualization of the therapeutic chemical clusters in (1) OmniViz Galaxy in relation to (2) pharmaceutical citation volume in the three-dimensional contour map. Reproduced from Saffer 2004 [117].

relationships uncovered from the visualization. This is by far the most important pre-requisite to *biological knowledge discovery* [117]. Therefore the mandate in bioinformatics visualization is to search for new methods for capturing the biologist's concept model(s) in graphical forms so as to enable the visual analysis of biological data [106]. The biological data here means the scientific data generated from experimental data collected during biological research.

2.3. Visual Information Analysis

Proponents of visualization argued that by exploiting human visual perception, computer generated images can assist users to explore and comprehend data. Indeed, human vision is very adapted to pattern recognition in especially pattern anomaly. The emerging area of *visual data mining* aims at leveraging this perceptual capability by mapping data to different visual representations. The hope is that unexpected properties of the data set will get augmented by a particular visual representation thereby not only to enhance the user's understanding of the experimental data, but also to encourage one to explore the data to the point where hypotheses can be formulated. This exploratory aspect of visualization is often known as *visual analysis* [143]. Therefore, visualization is more than just a tool for communicating information about data to the users. It is also a tool for generating new knowledge from data.

Put simply, *visual information analysis* is the application of information visualization on investigative or exploratory data analysis [143]. *Visual analytics* is the science of analytical reasoning supported by interactive visualizations as interfaces. The purpose is to highlight hidden patterns of the dataset so that the user can generate insights that would otherwise be impossible. In this sense, visual analytics can be seen as a data mining method because *data mining* is commonly defined as the extraction of patterns or concept models from experimental data [173]. Concept model visualization is the process of using visualization methods to make the discovered knowledge understandable by the human user. Because the cognitive process of extracting high-

level potentially useful knowledge from low-level data is crucial to knowledge discovery, visual data exploration plays an important role in this. Furthermore, visual data exploration is especially useful when little is known about the data and/or the analytical objective is vague. Hence, it is of value as a hypothesis generation method [174]. Visual data exploration in the form of concept model visualization is more intuitive than statistics-based data mining methods. For this reason, it suits domain experts whose lack background in statistics and mathematics.

Several researchers have outlined *visual analytical frameworks* that describe how users apply information visualization to data analysis. These frameworks share the common characteristic of modeling a user's participation in the visual analytical process as an iterative sequence of steps. Each step has a different focus and a different the level of abstraction.

The model proposed by Card *et al.* [22] put forward a high-level model of human analytical activity. They called it the *knowledge crystallization cycle* where the objective is to gain insights from data relative to some tasks. The analytical steps involved in this model range from 'foraging for data' to 'deciding or acting on the findings'. Spence [144] extended Card *et al.*'s model by specifically investigating the 'foraging for data' step in terms of visual navigation. He related visual navigation to cognitive activities such as concept model formation and information interpretation, and argued that the way users navigate, explore, and visualize a dataset will influence how they think about the dataset.

The cognitive process of visual analysis has also been investigated from a task-centric perspective. Shneiderman [141] proposed a two-step visual analysis framework, i.e. "*overview then detail*". He further suggested that information visualization systems need to support seven tasks in order to facilitate the problem-solving process. These tasks are overview, zoom, filter, details on demand, relate, history, and extract. They are obviously an extension of his well known information visualization mantra of "*overview, zoom, filter*, and *details-on-demand*".

A more recently proposed framework [4] considered data analytical tasks as high-level knowledge-based analytical activities. Therefore, the Amar and Stasko's framework emphasized heavily on supporting decision making and domain learning by identifying useful relationships from data. It presented analytical steps that users of a visualization system would typically perform, e.g. complex decision making, domain learning, identifying, explaining, and predicting trends. However, visualization systems that employ Amar and Stasko's framework are yet to be seen.

2.4. Graph Drawing and Network Visualization

Network visualization has emerged in recent years as an actively research area in information visualization. In chemistry and biology, network visualization has been applied to evolutionary trees, phylogenetic trees, molecular interaction networks, genetic maps, and biochemical

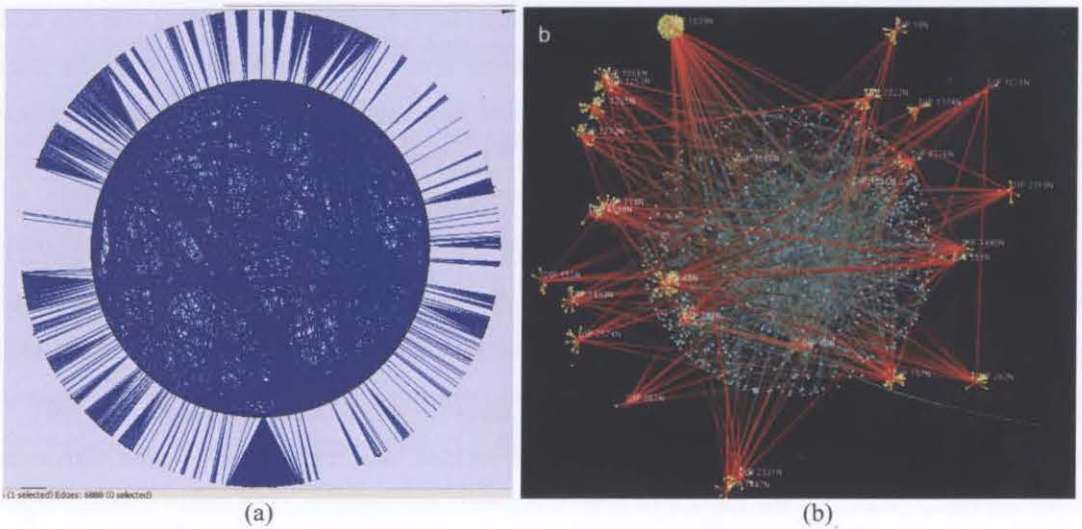
pathways [175]. It is especially important to molecular biology since molecules that interact or regulate one another are readily visualized as a network (or graph) [32]. Therefore, one can argue that network visualization concerns mainly with *graph drawing*.

At the simplest level, biological molecules within a single cell can be displayed as nodes (or vertices) and interactions can be represented as edges. The edges can be undirected or directed, i.e. specifying a source and a target. The biological molecules come in different types, i.e. genes (DNA sequences), gene products (proteins and RNA), and metabolites (glucose, pentose phosphates, and lipids). Thus the interactions among genes, gene products, and metabolites can be visualized as networks with directed or undirected edges. Directed edges are suitable for visualizing the flow of metabolites in a metabolic reaction or the flow of information from a gene regulator to the target gene which expression it regulates. Undirected edges are suitable for visualizing physical interactions between molecules such as protein-protein interactions.

The visual encoding and topology of the graph can be presented in a variety of ways. Nodes can be represented by spheres, boxes, circles, squares, and a combination of these or none of the above but implicitly by their name labels. Edges can be displayed as straight line, Bezier curves, and etc. Additional information on a node or an edge can be represented by using the properties of visual entities such as colour and size or using text labels positioned next to the node or the edge. A network can be drawn on a two-dimensional plane or in a three-dimensional space. A combination of these visual design elements forms the *visual representation* of a network [49].

After deciding on the visual representation needed, the next question to be resolved is how to automatically position the nodes and edges in a readable form (layout). The *layout* of a graph is the geometrical mapping of the nodes and edges onto a two-dimensional plane or three-dimensional space. The choice of layout often determines how comprehensible the graph is to human cognition. Different applications may require different layouts and therefore different criteria for determining whether the layout is good or bad. These criteria are appropriately called *aesthetics criteria*. There have been a few general aesthetics criteria applicable to a wide range of graph drawing [176]. They are listed as follows:

1. Crossing minimization
2. Bend minimization
3. Area minimization (2D layout)
4. Volume minimization (3D layout)
5. Good angular resolution
6. Total edge length minimization
7. Symmetries maximization



(a) (b)

FIGURE 2.4. Visualizations of the yeast protein interaction network using (a) two-level circular layout, and (b) hub-satellite spherical layout. Reproduced from Lee and Megeney 2005 [90].

Not all of these criteria are important to the layout of every biological network. For example, bend minimization will be more important than area minimization to the graph layout of metabolic network because the emphasis is on the flow of metabolites through a series of reactions. It is easier for the human eye to follow an edge with only a few or no bends than one with frequent zig-zags.

In large scale graphs containing a few thousand nodes and edges, it is difficult to satisfy more than one criterion simultaneously because the criteria may contradict each other. More evaluation on the aesthetics criteria of graph drawing in the field of bio-information visualization is required. To date, most bioinformatics tools automatically draw static layouts of networks that roughly fall into three categories: circular layout, force-directed layout, and hierarchical layout. The fourth category known as the multi-plane 2.5D layout has not yet been widely used and is in an experimental stage for bioinformatics applications.

2.4.1. Circular Layout

In its simplest form, each node is placed on the circumference of a circle and edges are drawn as straight lines. The drawing algorithm consists of three basic steps:

1. Crossing reduction: This step computes the ordering of the nodes such that the number of edge crossings is minimized [10].
2. Node positioning: This step assigns the x,y-coordinates to every node such the nodes are arranged in a circle.
3. Edge drawing: This step draws the edges, usually straight lines, between nodes.

A more complex variation is the *concentric* layout [140]. Each concentric level contains nodes with a higher degree than those in the adjacent outer level. As a result, nodes with the lowest degree are being arranged in the innermost level and the ones with the highest degree are being placed on the outermost level (see FIGURE 2.4(a)). The purpose is to expose the highly connected nodes within the network while revealing their connection with the rest of the network. This approach is particularly suitable for representing scale-free networks, e.g. canonical proteome and metabolic networks. Its limitation is that substantial edge crossings may occur in large-scale networks and will make identifying *k*-neighbours per node difficult.

The three-dimensional variation of the circular layout is the concentric hemispheres [90]. The nodes are being ranked by their node degrees and divided into three groups. Nodes within each group are being placed on the surface of three concentric hemispheres. The nodes with the highest range of node degrees are being placed on the outermost layer whereas those with the lowest range of node degrees are being placed on the innermost layer (see FIGURE 2.4(b)). This layout also has the advantage of separating the high degree nodes from the low degree ones while reducing edge crossings. Its limitation is the occlusion formed by the aggregation of high degree hubs.

2.4.2. Force-directed Layout

This method models the nodes in an undirected graph as repelling charged balls. This is the *repelling force*. In order to limit the distance between the repelling nodes, the edges in the graph are modeled as springs. This is the *attraction force* or the *spring force* [177]. The original force-directed algorithm now known as *spring embedder* was proposed by Eades [44]. Given an undirected graph $G = (V, E)$, let $p = (p_v)_{v \in V}$ be the vector of node co-ordinates $p_v = (x_v, y_v)$. Eades' algorithm [44] defines the repelling force between every pair of non-adjacent nodes $u, v \in V$ as:

$$f_{rep}(p_u, p_v) = \frac{c_\partial}{\|p_v - p_u\|^2} \cdot \overrightarrow{p_u p_v}$$

where c_∂ is the repulsion constant, $\|p_v - p_u\|$ is the length of the difference vector $p_v - p_u$ which is the Euclidean distance between positions p_v and p_u, and $\overrightarrow{p_u p_v}$ is the unit length vector $\frac{p_v - p_u}{\|p_v - p_u\|}$ pointing from p_u to p_v.

The edge connecting the pair of non-adjacent nodes $u, v \in V$ is modeled as the spring force which is defined as:

$$f_{spring}(p_u, p_v) = c_\sigma \cdot \log \frac{\|p_u - p_v\|}{l} \cdot \overrightarrow{p_v p_u}$$

where c_σ is the constant controlling the strength of the spring, l is the natural length of the spring. The placement of the nodes is computed iteratively until force equilibrium is attained when the spring force and the repelling force almost cancel out each other.

Force-directed layouts are ubiquitous in visualization tools because of their ease of implementation. One limitation of force-directed algorithms is that they require $O(n^2)$ time to attain equilibrium where n is the number of nodes. For this reason, some implementation allows users to terminate the algorithm at will. Yet its biggest limitation is poor predictability. Repeated running of the algorithm does not necessarily generate the same layouts, therefore, demanding the reconstruction of a new mental model on the user's part. At the large scale, the force-directed layout algorithm generated the 'hair ball' effect typically seen in many PIN visualizations [148].

2.4.3. Hierarchical Layout

The hierarchical layout is used for directed graphs. A hierarchical graph is a graph that exhibits a hierarchy of parent-child relationships. Because the nodes within the hierarchical graph are organized into a hierarchy, they can be drawn on k-levels such that the hierarchy is displayed as a series of parallel and horizontal levels [149]. Each inter-level edge represents the parent-child relationship between two nodes. The drawing algorithm usually consists of four main steps:

1. Level assignment: This step assigns each node with a y-coordinate. The node set within each level has a distinct y-coordinate to ensure the clear separation of levels.
2. Crossing minimization: This step computes the ordering of the nodes in each level such that the number of inter-level edge crossings is minimized or otherwise reduced. This is usually done by examining adjacent levels and the inter-level edges (see below).
3. Node positioning: This step converts the node ordering of each level into x-coordinates.
4. Edge drawing: This step draws the edges, usually straight lines, connecting nodes within different levels.

For step 2, one commonly used method is the barycenter heuristic which is also called *averaging* [149]. Because it is easy to implement, linear time complexity, and generally gives good results, the barycenter heuristic is a very popular crossing minimization method. Given a bipartite graph $G = (V_1, V_2, E)$ in which $u \in V_1$ and $v \in V_2$, the position of the node u is defined as the average of the x-coordinates of its neighbours $N(u)$ where $N(u) := \{v : \{u,v\} \in E\}$. The barycenter score of node u can be computed as the follows:

$$barycenter\ (u) = \frac{1}{\deg(u)} \sum_{v \in N(u)} pos(v)$$

where $pos(v)$ is the relative ordering node v. The nodeset V_1 is then sorted by the barycenter scores. For every node u, its barycenter can be computed in $O(N(u))$ time. Hence, the barycenters of all nodes can be computed in linear time.

The greatest strength of the hierarchical layout is the clear layout of the parent-child relationship structure between objects. However, visual complexity due to edge crossings increases with the number of nodes in each level.

2.4.4. Multi-Plane (or Level) 2.5D Layout

All the above layout methods discussed so far are used for 2D graph drawings. These methods tend to have one or more of the following limitations:

1. Lack of scalability: 2D layouts can at best accommodate a few thousand nodes without running into two constraints. The first is computational efficiency. The runtime of the 2D layouts becomes increasingly prolonged with the increase in the number of nodes and edges by an order of magnitude. The second is visual complexity. The inclusion of over ten thousand nodes in a graph will be cluttered due to the high levels of edge crossing and node overlaps. This will certainly reduce readability making it difficult to recognize patterns and inhibiting good insight on the data set.
2. Restricted capacity for domain complexity: There are certain network properties pertaining to a particular knowledge domain that requires multiple visual encoding and multiple layout conventions. Molecular biology is a knowledge intensive domain. It is difficult to produce a readable biological network that has multiple glyphs for representing different types of molecules and multi-colour edges for representing different types of interactions between molecules (also see section 2.6.5).

The 2.5D multi-plane layout resolves the limitations of 2D layout by using a divide-and-conquer approach [183]. A graph (or network) is first divided into a series of sub-graphs (or sub-network), and then each sub-graph is drawn on a separate plane using one or more of the 2D layout methods listed in the previous sections. In general, the drawing algorithm consists of four steps:

1. Graph partitioning: This step partitions a graph G into a set of sub-graphs $\{G_i : 1 \le i \le m\}$.
2. Sub-graph drawing: This step draws each sub-graph G_i, for each i in the range $1 \le i \le m$, on a plane P_i using a certain 2D drawing algorithm.
3. Plane arrangement: This step arranges each plane P_i in a 3D or 2.5D formation by satisfying some chosen criteria.

4. Inter-plane connection: This step draws all the edges in between planes.

This algorithm is very flexible since arbitrary choices can be made according to some domain knowledge or computational optimization criterion. For step 1, the appropriate partitioning can be determined by the domain application. In molecular biology, this partitioning can be determined by the functionality of different molecular interaction networks. Otherwise, graph partitioning can be a classic optimization problem, e.g. finding triangular motifs, finding minimum cuts, or a balanced partitioning. In most cases, such problems are NP-hard. However, linear time heuristics are available [184, 185]. For step 2, one can select a preferred 2D graph drawing algorithm based on the application domain [176, 186]. For step 3, some criteria are involved. For example, the number of planes should be kept to the minimum. Otherwise, the visualization will lose its 2.5D attitude. The other criterion is to avoid intersection between planes. There is also the need to minimize inter-plane edge crossing, i.e. where at least one crossing edge has endpoints in two different planes. In the same theme, one can also consider other criteria such as minimizing the total inter-plane edge length in the drawing.

The time complexity of the multi-plane layout algorithm depends on the time complexity of the method chosen for each step. At present, multi-plane layout is not generally available in visualization tools but has been experimented successfully on metabolic networks [14, 187].

2.5. Visualization of Gene Expression Pattern

The existing methods for visualizing gene expression data generated from microarray technologies come in four main approaches.

Visualization of gene expression patterns. This is the visualization of the data pattern as an output of a certain clustering algorithm. The resulting visual pattern is entirely determined by the grouping of gene expression values. The aim is to assist the biologist in the task of identifying groups of co-expressed genes and groups of differentially expressed genes throughout a series of experimental conditions or time points.

Contextual visualization of gene expression. This is the visualization of gene expression data being mapped to the restricted controlled vocabulary schema curated by the Gene Ontology Consortium [60].

Visualization of gene co-expression network. This is the visualization of gene expression data that have been filtered according to a statistical correlation score, e.g. Pearson correlation.

Visualization of gene expression in molecular networks. This is the visualization of a biological network being overlaid with gene expression data.

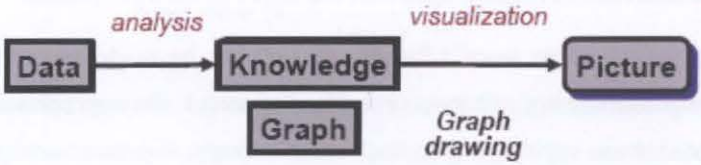


FIGURE 2.5. Visualization pipeline for mapping data to a network representation and then visualizing the network as a picture.

Of the above approaches, the last three employs variations of the visualization pipeline shown in FIGURE 2.5. It is a process for mapping abstract data to a visual representation and broadly involves two essential steps. The first is the *analysis* step which extracts an information model from the abstract data. This information model can be a type of ontology schema or a type of graph-theoretic representation. The second is the *visualization* step which generates either a non-graph or graph visualization. In the following sub-sections, the methods used in each approach and the strength and weaknesses of each will be elaborated.

2.5.1. Visualization of Gene Expression Patterns

By far, the earliest visualization method and the most widely used is the '*dendrogram + colour matrix*' (see FIGURE 2.6). In order to give the colour matrix a visible pattern, the dataset is analysed with a data mining algorithm and gene expression profiles are organized into a 'dendrogram + colour matrix'. A *gene expression profile* is the vector of gene expression values represented by each column on the matrix. The colour matrix is being used to visualize clusters of similar expression profiles whereas the dendrogram indicates the degree of similarity (or distance) between clusters.

The underlying data pattern for the 'dendrogram + colour matrix' visualization is usually an output of the hierarchical clustering algorithm [145]. Other data mining algorithms, e.g. Spearman correlation, Pearson correlation, and Self-Organising Map have also been used [145]. This approach basically employs the visualization pipeline shown in FIGURE 2.5 and is predominant among microarray analysis applications. It is particularly strong in presenting an overview on the hierarchical and modular structure of the gene expression pattern, and makes full use of the available screen space.

Many colour matrices provide interactions to facilitate exploration. For example, a brush over on a colour spot in maxdView [67] can show the corresponding expression value and gene symbol and a click on the right mouse button on the same spot can trigger a menu that provides hyperlinks to public databases.

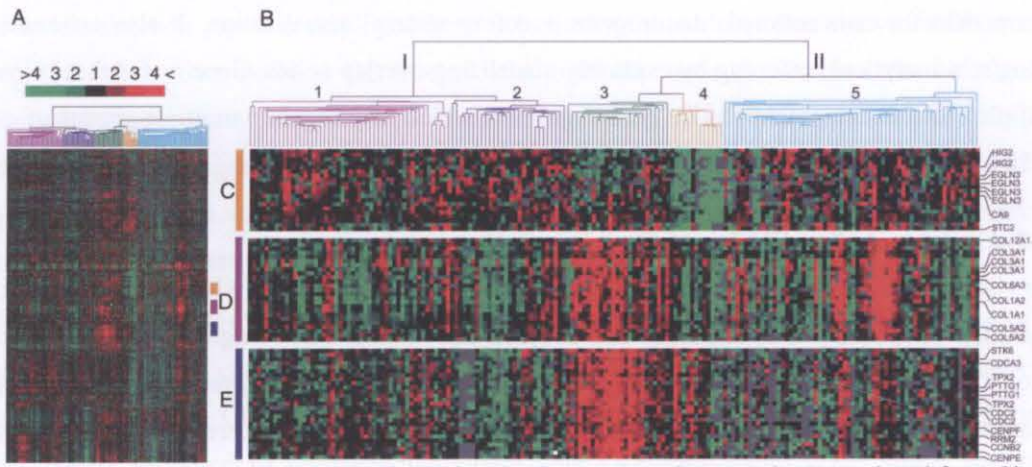


FIGURE 2.6. Visualization of microarray data using dendrogram + colour matrix. Reproduced from Chen et al. 2002 [29].

The 'dendrogram + colour matrix' visualization has four limitations.

1. The visualization is stronger in revealing positive correlations than is otherwise because the immediate neighbour of each expression profile is always the one with the highest similarity scoring and the same trend applies on the cluster level. In other words, it is easier for users to identify clusters of positively co-expressed genes rather than the opposite.
2. Given the same dataset and the same hierarchical clustering algorithm, the resulting dendrogram can be drawn in 2^{n-1} ways. Thus the order of the clusters within the dendrogram can vary from one instance to another.
3. The drawing of the dendrogram next to the colour matrix leads to the misinterpretation that all clusters are non-intersecting [133].
4. Colour hue is not the best visual encoding for numerical data because hue variation at the extreme ends of the data range is often too subtle to be detected by human vision. Rather, size and length coding is more effective.

The latest visualization on the clustered gene expression pattern is the *bicluster visualization* in the force-directed layout [133]. It resolved one of the limitations seen with the 'dendrogram + colour matrix' visualization, i.e. the lack of explicit visualization of multiple intersecting clusters. The visualization resembles a *zone graph* [88, 109] with each semi-transparent cluster node containing a node set of genes and a node set of experimental conditions (see FIGURE 2.7). They are visually represented in different glyphs. If the same node belongs to multiple clusters, it is visually represented by *overlap* nodes (also called *hub* nodes) and is positioned in the intersection areas of the cluster nodes (see FIGURE 2.7). Instead of using the convention of colouring gene nodes in green for representing down-regulated genes and colouring in red for representing up-regulated genes, the *minus* and *plus* signs are used as node labels. The bicluster visualization therefore provides a more readable and aesthetic visual representation of the gene expression

pattern than the conventional 'dendrogram + colour matrix' visualization. It also enhances the biologist's analytical reasoning by explicitly visualizing overlap nodes. Gene nodes that belong to multiple clusters can potentially be gene regulators which may regulate multiple groups of genes [133].

2.5.2. Contextual Visualization of Gene Expression

The visualization of Gene Ontology (GO) is truly a type of information visualization since it is a controlled vocabulary that describes the accumulated human knowledge on every discovered gene. Visualizing microarray dataset in GO-annotated clusters is therefore valuable to biologists because it provides the means for identifying sets of genes that share the same biological process(es). The overlaying of gene expression correlation values on GO-annotated clusters further assists the identification of functioning biological processes. Such an analytical task is based on the biologist's presumption that genes involved in the same biological process are more likely to be highly correlated in their expression level. Biologists called this type of gene expression dynamics as *gene co-expression* [147].

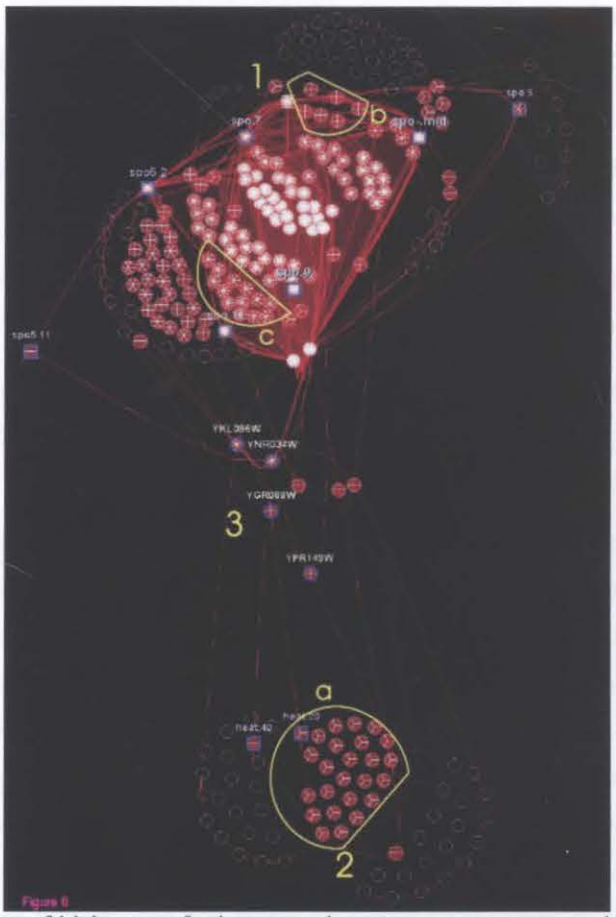


FIGURE 2.7. Visualization of biclusters of microarray data on yeast gene expression profiles. Two groups can be seen at the (1) top and (2) bottom, and (3) a small number of genes connecting both groups. Subsets of nodes are clustered together in (a) three, (b) four, or (c) six biclusters. Circular nodes represent genes and square-shaped nodes represent experimental conditions. Reproduced from Santamaria et al. [133].

The visualization pipeline employed here is basically the same as that shown in FIGURE 2.5 except the information model is GO instead of a molecular network. To date, the visualization generated by microarray analysis tools such as GenMapp [41, 131] and GOTree [171] are predominantly dendrograms and the *n*:*m* gene-GO relationships are often obscured (see FIGURE 2.8). Furthermore, as a one-dimensional solution, dendrograms can not present intersecting clusters where the same set of genes is associated with more than one ontology terms [82].

HTP-GOMiner™ [169] and Exploratory Visual Analysis [125] provide visualizations of the gene-GO and gene_cluster-GO relationships respectively while hiding the parent-child relationships between GO terms. The cluster map presented in HTP-GOMiner™ is a form of colour matrix designed to represent individual gene-GO relationship. The cluster pattern is formed by the aggregation of red coloured squares with no clear cluster boundaries (see FIGURE 2.9(a)). Exploratory Visual Analysis (EVA) is another form of colour matrix (see FIGURE 2.9(b)). The global cluster pattern is formed by a series of clearly bounded GO-annotated clusters arranged in a grid layout. Within each cluster is a matrix of nodes representing the subset of genes. Co-expressed genes are the nodes that share the same colour hue. In EVA, the *m*:*n* gene-GO relationship is being visualized as a 1:1 relationship leading to the same gene being drawn into multiple clusters. Therefore, unlike GOMiner™, the visual semantics of EVA is GO-centric rather than gene-centric.

The GO hierarchy has also been visualized as Treemaps (see FIGURE 2.2). Treemap is a space-filling visual representation designed for visualizing hierarchies. Each GO-annotated gene cluster is visually encoded as a rectangle with an area proportional to the cluster size, i.e. the number of member genes [5]. Each cluster is a leaf node in the GO hierarchy. Biologists can also gain a global view on the differential expression of biological processes by overlaying correlation scores on to the Treemap. The limitation of using Treemap is that the GO schema is a directed acyclic graph. Yet Treemap is designed for visualizing hierarchical data. Therefore an extra data processing step that maps the GO schema to a hierarchical structure is required. As a result, any *n*:*m* gene-ontology relationships are lost in the mapping process. Nevertheless, Treemaps are better than dendrograms because it preserves the biologist's modular view of molecular biology. More importantly, the treemap captures the '*module within module*' structure of a molecular biological system. Using treemaps, biologists can distinguish between up-regulated and down-regulated biological processes easier than reading dendrograms.

An alternative to Treemaps is the Venn diagram [82]. Similar to the Treemap, each GO cluster is visually encoded as a distinctly coloured polygon with an area directly proportional to the cluster size. Both Venn diagram and Treemap do not explicitly display the gene nodes in each cluster but that is where their similarity ends. The Venn diagram presents genes shared by multiple GO clusters as intersections between polygons (see FIGURE 2.10).

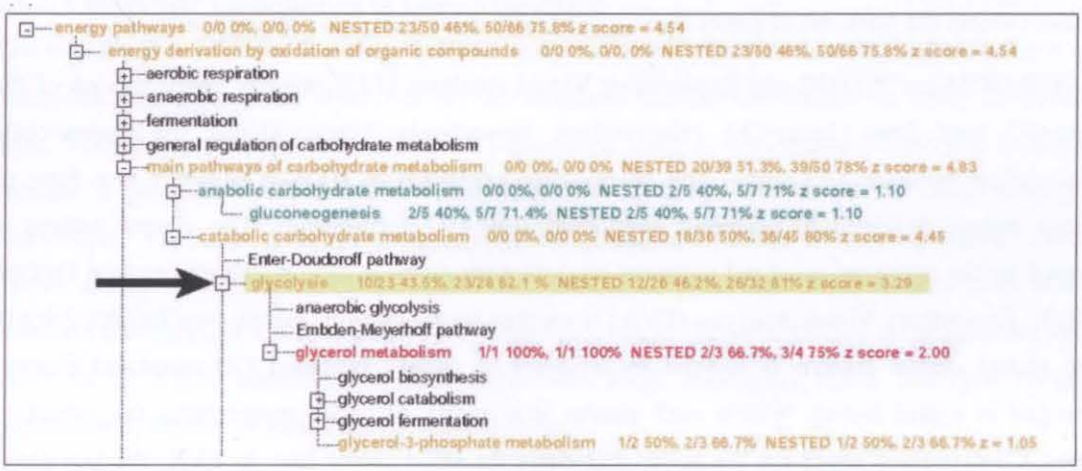
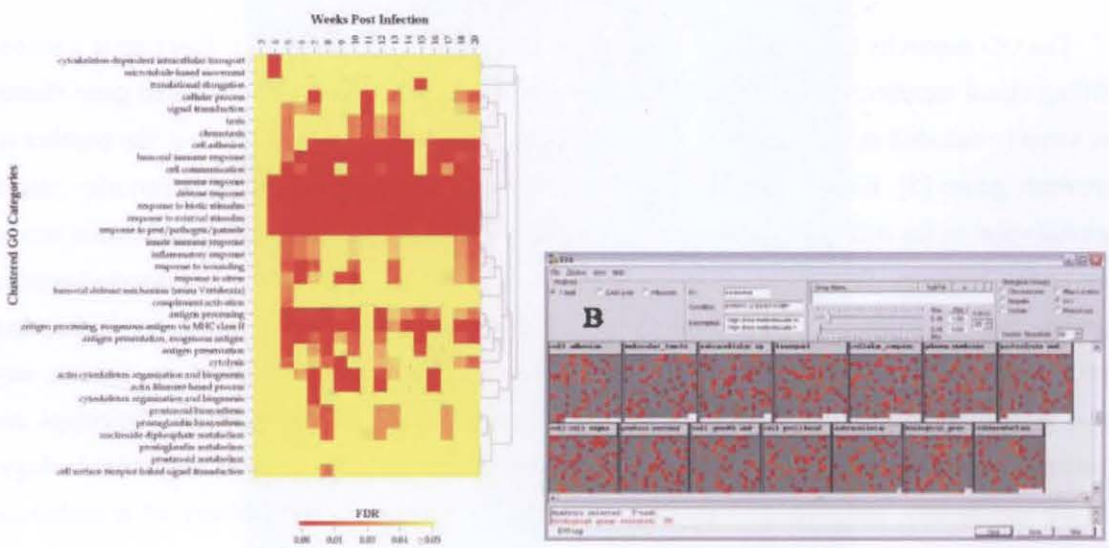


FIGURE 2.8. Visualization of the Gene Ontology hierarchy in a dendrogram. Reproduced from Doniger 2003 [41].



(a) (b)

FIGURE 2.9. Visualization of the gene-GO relationship in (a) GOMiner™. Reproduced from Zeeberg 2006 [169], and (b) Exploratory Visual Analysis. Reproduced from Reif 2005 [125].

The intersecting area is again directly proportional to the number of overlap gene nodes. As such, the Venn diagram is better than the Treemap in revealing the complex inter-dependency between GO clusters, especially if each cluster represents a biological process. As an example, FIGURE 2.10(a) shows the biological processes that are up-regulated in pancreatic ductal carcinoma when compared to normal pancreatic duct cells. The intersection of the ontology term "*regulation of cell cycle*" with the two other terms "*response to wounding*" and "*GTPase activity*" suggested that a malfunctioning signal transduction process that has a functional role in wound healing could be the cause.

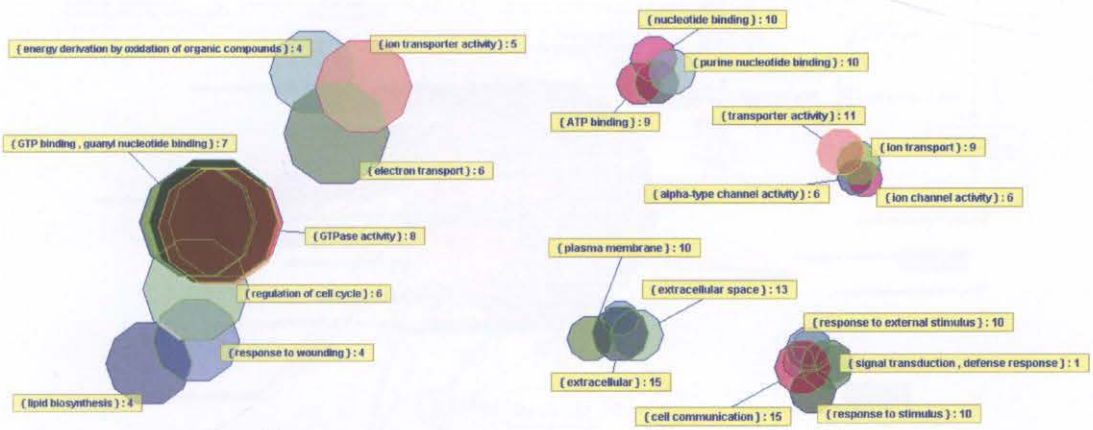


FIGURE 2.10. Visualization of differentially expressed gene clusters in Venn diagram. (a) GO Process-annotated gene clusters that are up-regulated in pancreatic ductal carcinoma compared to normal pancreatic cells, and (b) gene clusters that are down-regulated.

The Venn diagram also has its limitation. It collapses the parent-child hierarchy of a particular class of biological processes to a set of overlapping clusters. FIGURE 2.10(b) shows the biological processes that are down-regulated in pancreatic ductal carcinoma. The sets in each cluster belong to a single class of biological processes and there are no intersections between any of the clusters. For example, the cluster on the top left-hand corner contains four intersecting sets of genes all belonging to ontology hierarchy of the nucleotide binding biological process. The term "*nucleotide binding*" is the parent of "*purine nucleotide binding*" which in turn is the parent of "*ATP binding*". The intersection here simply represents genes that have been fully annotated in all levels of the "nucleotide binding" hierarchy. Since many biological processes such as DNA transcription, ion channel activity, protein degradation, glucose metabolism and etc, require "*ATP binding*" to function, the significance of down-regulated "*ATP binding*" could not be understood unless it intersects with some other biological processes.

2.5.3. Visualization of Gene Co-expression Network

In distinction to a dendrogram generated by hierarchical clustering in which a given gene can have at most one neighbour, any given gene in a graph-theoretic representation can have multiple

neighbours [100]. It is also easier to integrate qualitative information, e.g. gene ontology and gene names, with graphs.

One visualization of the gene expression pattern as a network is the *co-expression network* [130]. Each gene is visually represented as a node. Each edge representing co-expression is encoded as a coloured line. Positive co-expression is represented by red colour whereas negative co-expression is represented by green colour. To impart biological meaning to the network, the gene nodes can be clustered using GO terms as a criterion. The result is a network visualization of GO-annotated co-expressed gene clusters.

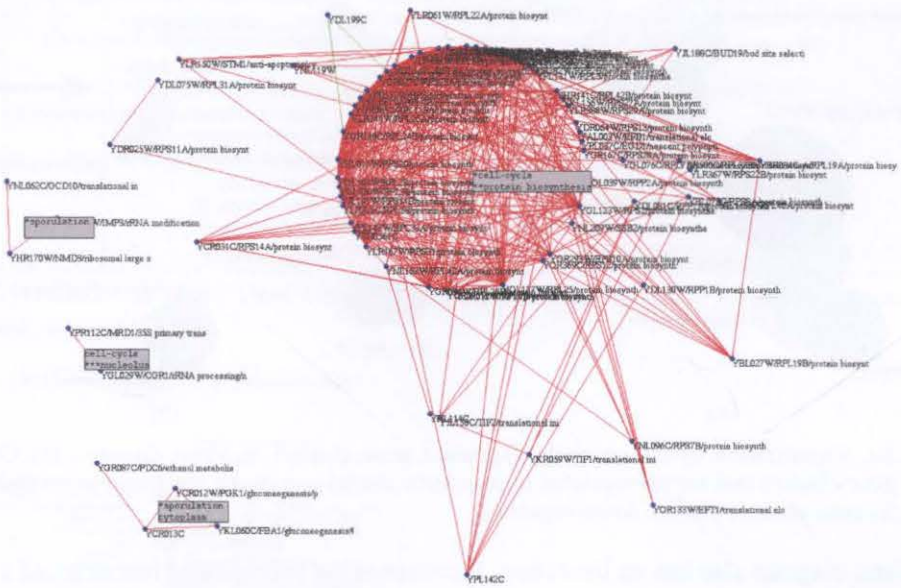
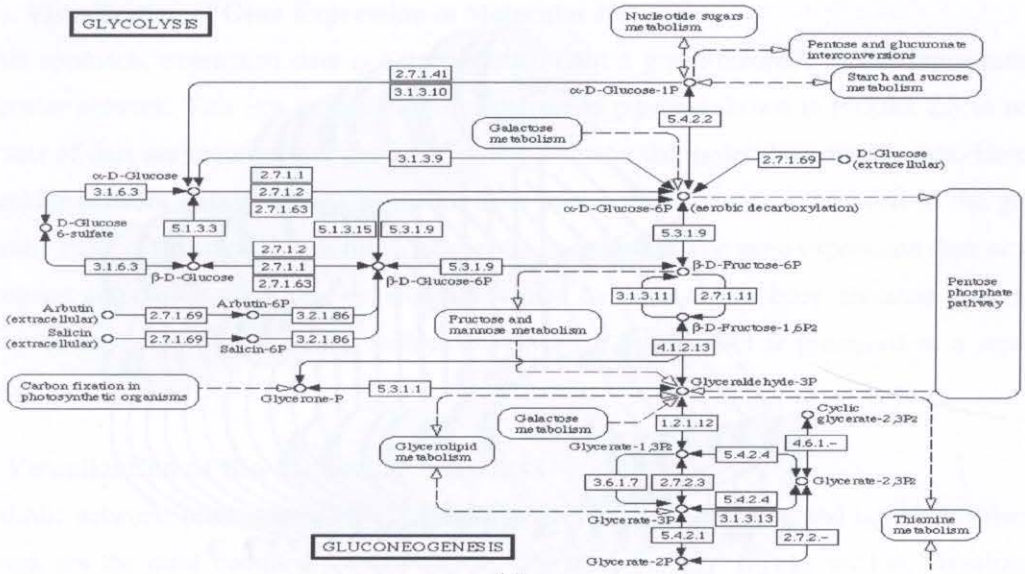
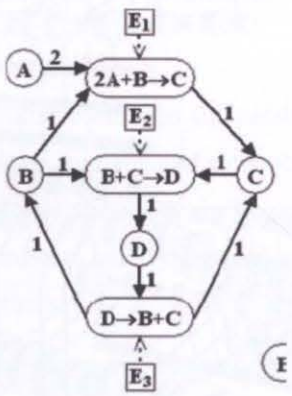


FIGURE 2.11. Visualization of the gene co-expression network in the circular layout. Reproduced from Rougemont 2003 [130].

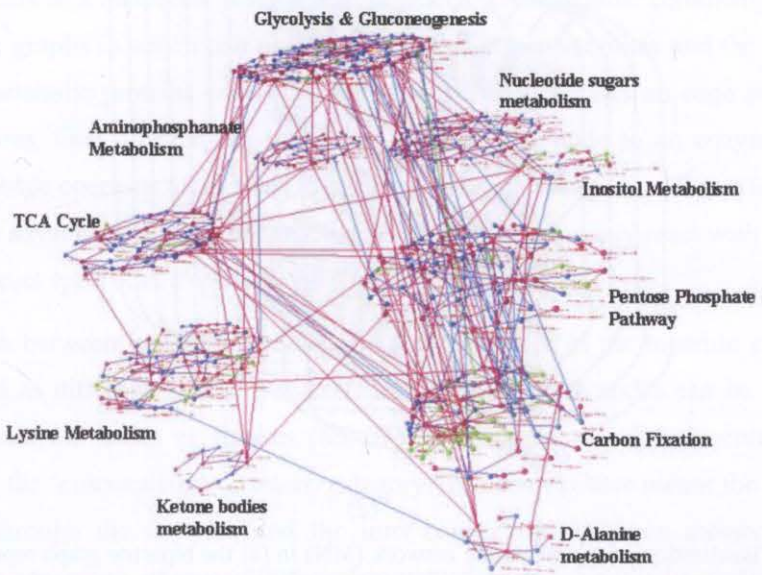
In terms of layout, the gene co-expression network has been drawn using the circular layout [130]. Each cluster of gene nodes is being arranged in a circular layout (see FIGURE 2.11). In terms of interactivity, users can isolate a particular cluster in a separate window by simultaneously pointing and clicking each node. Zooming mechanism is provided so that the biologist can inspect the network at different level of details. Although over-represented GO terms are being displayed within each cluster, it is difficult for the biologists to tell which subsets of genes are related to which particular GO term because the labels are being displayed at the centre of each cluster. When more than five GO labels are being displayed, label overlaps are frequently observed. Hence the clustered gene co-expression network visualization is good for displaying intra- or inter-cluster gene co-expression, but very limited in presenting the gene-GO relationship.


GLYCOLYSIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
α-D-Glucose-1P
Galactose metabolism
D-Glucose (extracellular)
α-D-Glucose
D-Glucose 6-sulfate
β-D-Glucose
α-D-Glucose-6P
(aerobic decarboxylation)
β-D-Glucose-6P
β-D-Fructose-6P
Pentose phosphate pathway
Arbutin (extracellular)
Arbutin-6P
Salicin (extracellular)
Salicin-6P
Fructose and mannose metabolism
β-D-Fructose-1,6P2
Carbon fixation in photosynthetic organisms
Glycerone-P
Glyceraldehyde-3P
Glycerolipid metabolism
Glycerate-1,3P2
Cyclic glycerate-2,3P2
Glycerate-2,3P2
Glycerate-3P
Thiamine metabolism
GLUCONEOGENESIS
Glycerate-2P

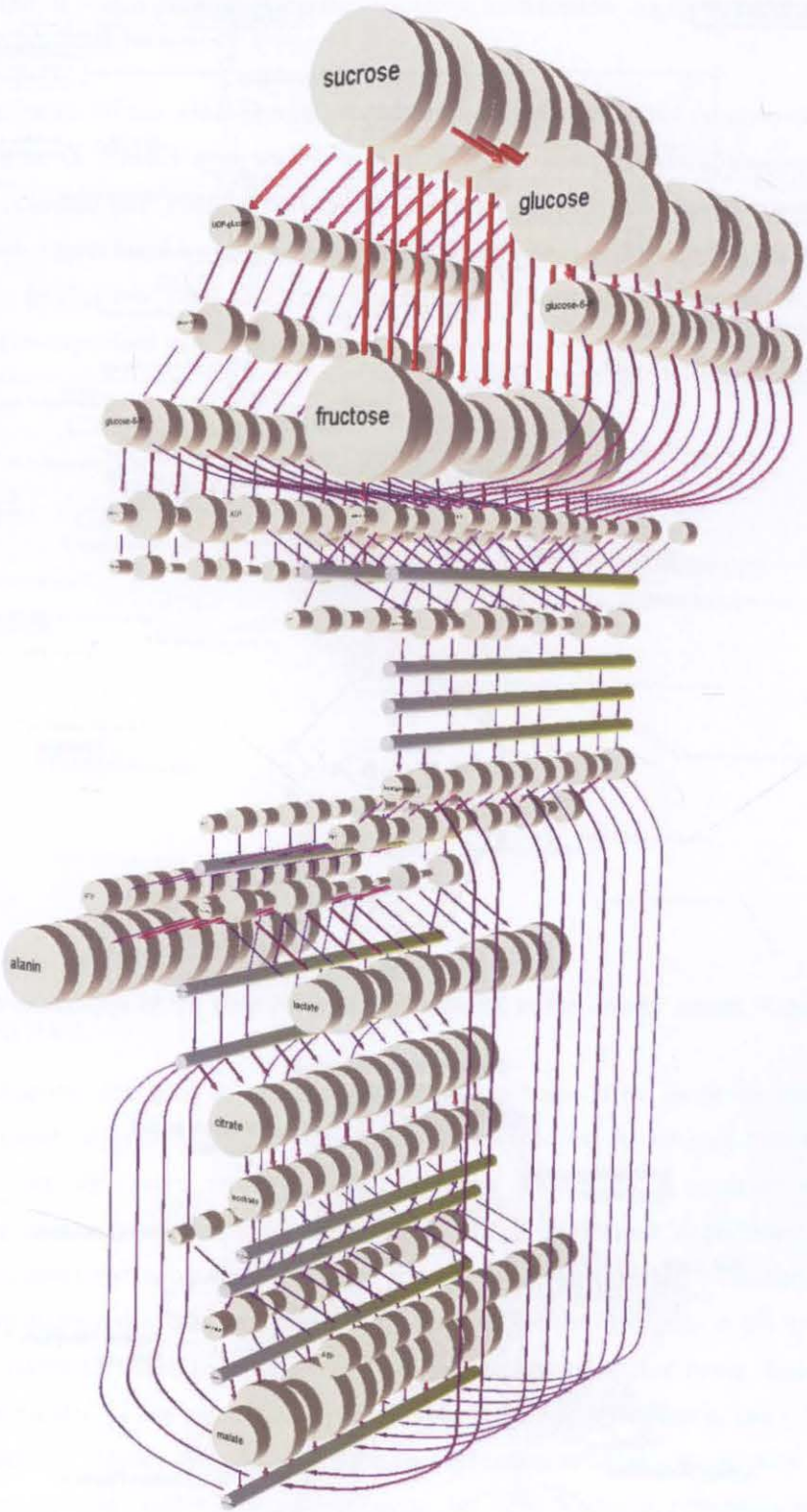
(a)


E1
A
2
2A+B→C
1
E2
B
B+C→D
C
D
D→B+C
E3

(b)


Glycolysis & Gluconeogenesis
Nucleotide sugars metabolism
Aminophosphanate Metabolism
Inositol Metabolism
TCA Cycle
Pentose Phosphate Pathway
Lysine Metabolism
Carbon Fixation
Ketone bodies metabolism
D-Alanine metabolism

(c)



(d)

FIGURE 2.12. Visualization of the metabolic network (MN) in (a) the bipartite graph representation with an orthogonal layout. Reproduced from KEGG [79], (b) the tripartite graph representation. Reproduced from Christensen 2007 [32], (c) the hierarchical clustered graph. Reproduced from Ho 2005 [187], and (d) the temporal dynamics visualization of glycolysis in parallel planes. Reproduced from Brandes 2004 [14].

2.5.4. Visualization of Gene Expression in Molecular Networks

In this approach, expression data is being overlaid onto a graph-theoretic model representing a molecular network. This is a variant of the visualization pipeline shown in FIGURE 2.5 in which two sets of data are required, the gene expression data and the molecular network data. Here the molecular network data, which is available from public data sources, is mapped to the graph-theoretic model. This model is in turn visualized as a network. The gene expression data or gene expression correlation scores are usually represented by node colour hues. Because there are a variety of methods for visualizing molecular networks, this subject is presented in a separate section.

2.6. Visualization of Bio-Molecular Networks

Metabolic network, protein interaction network, gene regulatory network, and signal transduction network are the most commonly visualized in molecular biology. In this section, visualization methods applied to these molecular networks are introduced.

2.6.1. Metabolic Network

A metabolic network (MN) is the entire collection of metabolic reactions within a single cell. Metabolic reactions are a combination of anabolic and catabolic reactions. In an anabolic reaction, new chemical compounds are created and energy is being consumed in the process. In a catabolic reaction, chemical compounds are being degraded and energy is being released in some processes while being consumed in others.

From the biological point of view, the representation primacy of a metabolic network is the flow of metabolites. Biologists often call a particular network path which metabolizes a certain class of compounds as a *metabolic pathway*. In practice, MNs are most commonly visualized as directed bipartite graphs in which one node set represents the metabolites and the other node set represents the metabolic proteins called *enzymes* [14]. It also contains an edge set representing metabolic reactions. One directed edge connects a metabolite node to an enzyme node and a second directed edge connects an enzyme node to another metabolite (see FIGURE 2.12(a)). This approach has the advantage of separating the metabolites that commonly react with most enzymes from those that react specifically with a small set of enzymes.

To distinguish between metabolite nodes and enzyme nodes in the bipartite graph, they are often represented as different glyphs. For example, the metabolite nodes can be represented as circles and the enzyme nodes as squares [63, 73]. In terms of visual representation, bipartite graphs belong to the '*connectivity + context*' category. The context here means the temporal flow of metabolites through the network and the inter-connections between metabolic pathways. However, a bipartite graph of over 1000 nodes and edges is cognitively challenging for the biologist to comprehend.

Another visual representation for MNs is the weighted tripartite graph [32] but is seldom used. It contains three node sets representing the metabolites, the metabolic reactions, and enzymes respectively. It also contains two directed edge sets representing metabolite flow and enzyme catalysis. For the metabolite flow, each directed edge connects the metabolite node to the metabolic reaction node. For the enzyme catalysis, each directed edge connects the enzyme node to the metabolic reaction node (see FIGURE 2.12(b)).

The visual encoding required for the tripartite graph visualization is more complex than its bipartite graph counterpart. Three types of glyphs are needed for representing the three node types and two edge formats for representing the two edge types. For example, the metabolite nodes can be represented as circles, the enzyme nodes as squares and the metabolic reaction nodes as ovals. The metabolic flow can be represented by solid lines whereas the enzyme catalysis can be represented by broken lines (see FIGURE 2.12(b)).

The greatest strength of the tripartite graph over its bipartite graph counterpart is the explicit visualization of the metabolic reaction equation as a node label (see '**2A + B → C**' in FIGURE 2.12(b)). This greatly reduces the biologist's cognitive loading since there is no need for him/her to cognitively extract a mental picture of a certain metabolic reaction from the visualization. The same task is necessary with the use of the bipartite graph visualization.

The limitation of MN visualizations, regardless of the visual representation, is the lack of explicit display of the enzyme-enzyme interactions. These are the specialized protein-protein interactions that mediate all metabolic reactions [83]. In terms of layout, the KEGG layout is commonly used which resembles the orthogonal layout [79] (see FIGURE 2.12(a)). Force-directed layout has also been used but is mainly for exposing the high degree hubs in the MN [6].

By far, the most novel layout is the multi-plane layout in 2.5D. One design is the hierarchical clustered graph on multiple planes. This method has been experimented on the metabolic network to visualize its modularity and the relationship between modules [187]. The visualization shows the metabolic pathways involved in glucose metabolism from the KEGG database (see FIGURE 2.12(c)). The top plane contains the glycolysis and gluconeogenesis network. The networks on the next plane are those that are directly connected to the glycolysis and gluconeogenesis network. The networks on the furthest plane at the fourth level are the most distant from the glycolysis and gluconeogenesis network. At this level, all the networks are involved in amino acid metabolism (see FIGURE 2.12(c)). Another design is have a series of networks stacked in parallel planes (see FIGURE 2.12(d)). Each plane represents a time point and the network drawn on each plane represents the glycolytic metabolic network at that time point. The resulting supergraph shows the changing topology of the same metabolic network as a representation of the temporal dynamics of glycolysis [14].

2.6.2. Protein Interaction Network

A protein interaction network (PIN) is the collection of all physical interactions between proteins. The most basic graph representation of a protein interaction network is an undirected graph and most commonly drawn using the force-directed layout [52]. This method produces readable network visualization up to a few hundred nodes. For a full-scale PIN of an organism that typically has several thousand nodes, the force-directed layout gives the 'hair ball' appearance which makes the visualization unreadable [148].

To resolve the limitation of the force-directed layout, an alternative layout for visualizing medium to large PINs called the *large graph layout algorithm* has been proposed [1]. This algorithm uses a tree as a guide to determine the order in which nodes are included in the spring force calculation. Nodes from a single connected network are laid out iteratively starting with a root node and incorporating additional nodes as guided by a minimum spanning tree of the network. The minimum spanning tree is defined as the minimum edge set necessary to keep the network connected and the sum of all the weights of the edges is minimized, where each edge is weighted by its associated statistical significance score. The statistical significance score is the calculated by the sequence comparison program BLAST [91].

The resulting visualization positions the largest connected component that has the highest edge density at the centre of the network drawing. An example of the large graph layout visualization is the protein homology network constructed by comparing 40 different bacterial genomes [1]. The visualization contains 111,604 protein nodes and 1,912,684 protein-protein interaction edges organized in 11,516 connected components (see FIGURE 2.13(a)). More importantly, the largest connected component at the core of the PIN visualization consists of clusters of essential proteins. Each cluster represents a biological process that is important to cell viability (see FIGURE 2.13(b) and (c)). Hence the large graph layout represents the self-organization property of the PIN that is natural to bacterial species.

PIN has also been visualized in the two-level circular layout [140] in which the low-degree nodes are in the inner level and the high-degree nodes are being arranged on the outer level (see FIGURE 2.4(a)). This approach has the advantage of separating the protein hubs from the sparsely connected proteins while the connectivity between the two remains clearly displayed. Node overlapping frequently seen in the force-directed layout is also resolved in the circular layout since all nodes are arranged side by side along a circular circumference. However, the substantial edge crossings in the centre space can make identifying k-neighbours per node difficult (see FIGURE 2.4). This limitation can be compensated by adding interactivity, e.g. highlighting neighbours by brushing the pointer over a certain node. The three-dimensional version of the two-level circular layout is the Satellite-Hub layout [90] (see FIGURE 2.4(b)).

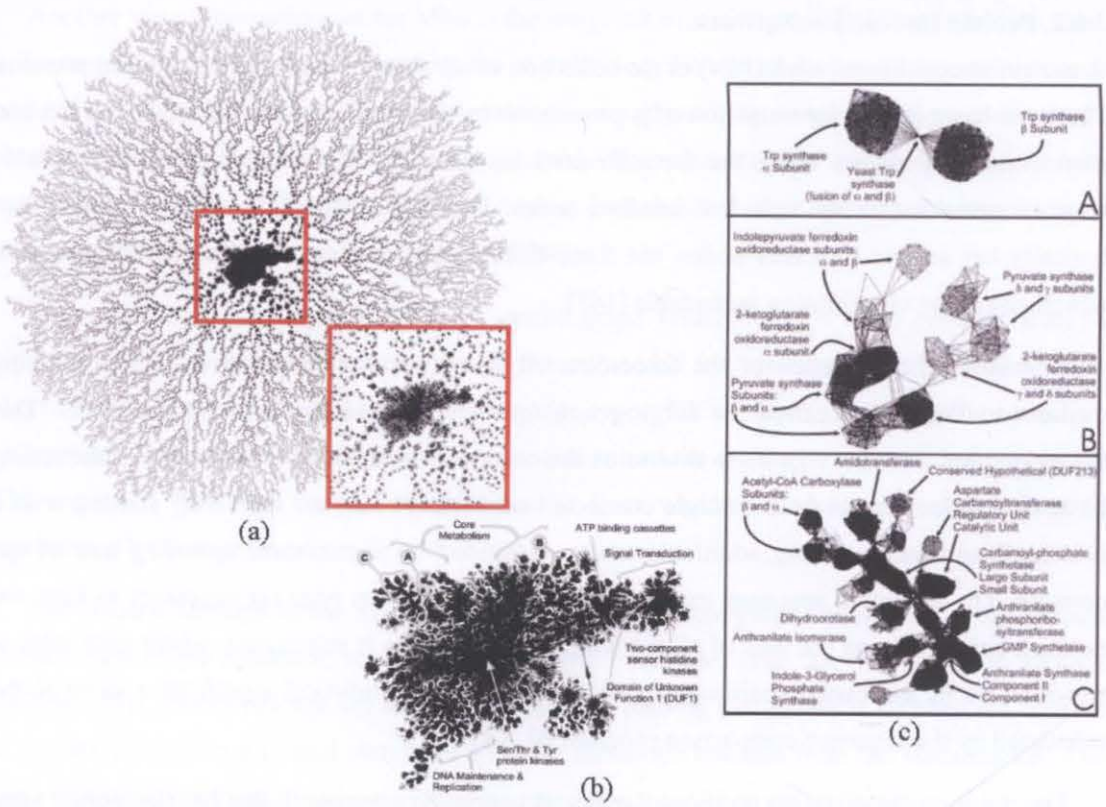


FIGURE 2.13. Visualization of full-scale protein interaction network common to 40 bacterial species in the large graph layout. (a) Overview of the PIN. The largest connected component is bound by the red box. (b) A zoom in view of the largest connected component representing 30,727 proteins. (c) Specific protein complexes in the largest connected component are shown. Reproduced from Adai 2004 [1].

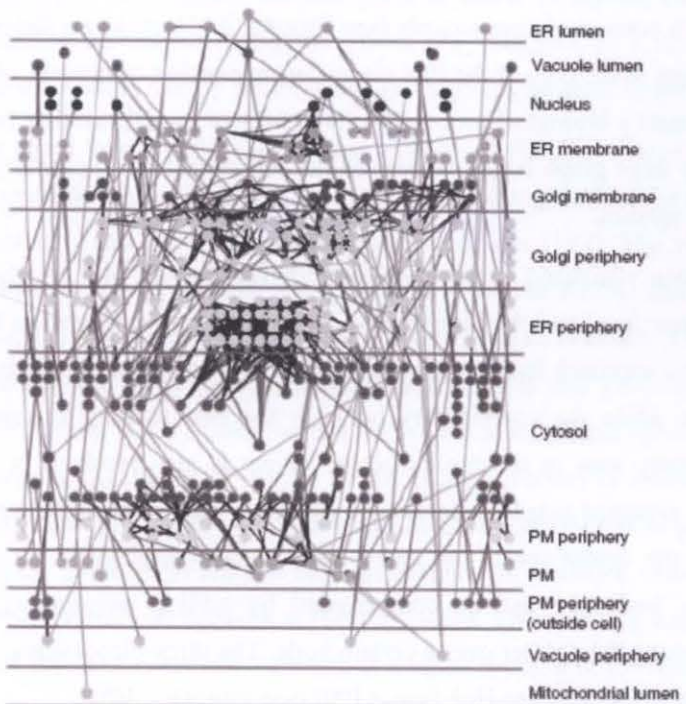


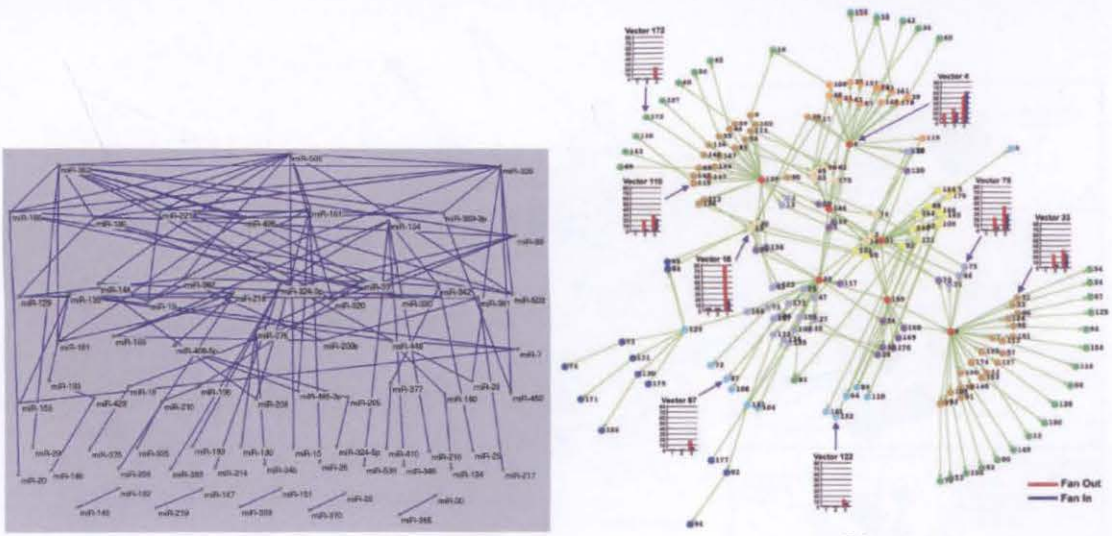
FIGURE 2.14. Visualization of the protein interaction network generated by the Cytoscape plug-in Cerebral. Each parallel partition represents a cell component. Reproduced from Suderman 2007 [148].

FIGURE 2.15. Visualization of the mouse protein interaction network using the betweenness fast layout (BFL). The size of the blue nodes corresponds to the magnitude of the node betweenness centrality. Reproduced from Hashimoto 2009 [71].

The greatest strength of this layout lies in being able to expose the protein interactions between high-degree nodes. These nodes are thought to represent proteins that coordinate the activities of various biological processes and are crucial to the viability of the living cell.

Recently, a number of layouts for PIN visualizations have been published. They share the common design criterion of accounting for biological knowledge in the layout. One approach takes biological ontology into account, e.g. the visualization of a PIN in a *parallel level layout* [7]. Each level defines a partition that represents a cellular component or a biological process (see FIGURE 2.14). With each partition, member nodes and edges presenting protein-protein interactions are drawn. This approach provides with the biologist a clear visual separation of proteins by their ontological classification, but is only suitable for visualizing PINs with a few hundred nodes. The other approach uses the biological significance of node or edge betweenness centrality as a layout optimization criterion.

In PIN, a high node betweenness centrality is often associated with bottleneck proteins, and systems biologists have been suggesting that the expression of these proteins fluctuate more frequently than others [168]. On the other hand, a high edge betweenness centrality is often associated with hubs that are central to other hubs. Systems biologists often associate such hubs as *essential* proteins [75]. Given that betweenness centrality is useful in identifying functionally important proteins, the *betweenness fast layout algorithm* (BFL algorithm) optimizes the positioning of high-betweenness nodes (see FIGURE 2.15) [71]. For a PIN with n nodes, BFL algorithm can achieve $O(n^2)$ runtime when only edge crossings minimization is considered, and $O(n \log n)$ runtime when only edge length and edge density minimization are considered.



(a) (b)

FIGURE 2.16. Visualization of the gene regulatory network (GRN) in two different layouts. (a) The human microRNA GRN in the hierarchal layout. Reproduced from Shalgi 2007 [138]. (b) The *E. coli* GRN in the Kamada-Kawai layout. Reproduced from Wong 2006 [164].

The limitation of this algorithm is that, at present, only biologists who are studying networks will understand the meaning of the node and edge betweenness. In this regard, the current biological knowledge is mostly projected from the yeast PIN. It is not clear whether the same understanding on the biological meaning of node and edge betweenness is projectable to human PIN or to other networks such as the MN. These are biological problems yet to be investigated.

2.6.3. Gene Regulatory Network

A gene regulatory network (GRN) is a collection of gene-gene interactions. It is often visualized as a multi-level hierarchal graph in the parallel level layout [149]. The nodes at the top level represent the master gene regulators which connect to their target genes at the lower level with a directed edge. Directed edges are particularly important to the visualization of regulatory networks because they present the control flow through the regulatory hierarchy (see FIGURE 2.16(a)).

GRN has also been visualized in a force-directed layout [78] but this approach does not explicitly separate the master regulators from their target genes by level assignment (see FIGURE 2.16(b)). Instead, high degree nodes tend to localize near the centre of the network. To identify the master regulators, the biologist has to exercise his/her reasoning that high degree nodes are more likely to be the master gene regulators. This path of reasoning is applicable only to bacteria because their GRNs contain few master regulators. For example, half of the *E. coli* genes are directly regulated by only seven master regulators [101]. The same is not true in multi-cellular organisms such as humans. In the human GRN, high degree nodes are not necessarily master gene regulators but could be co-factors, i.e. co-activators or co-repressors.

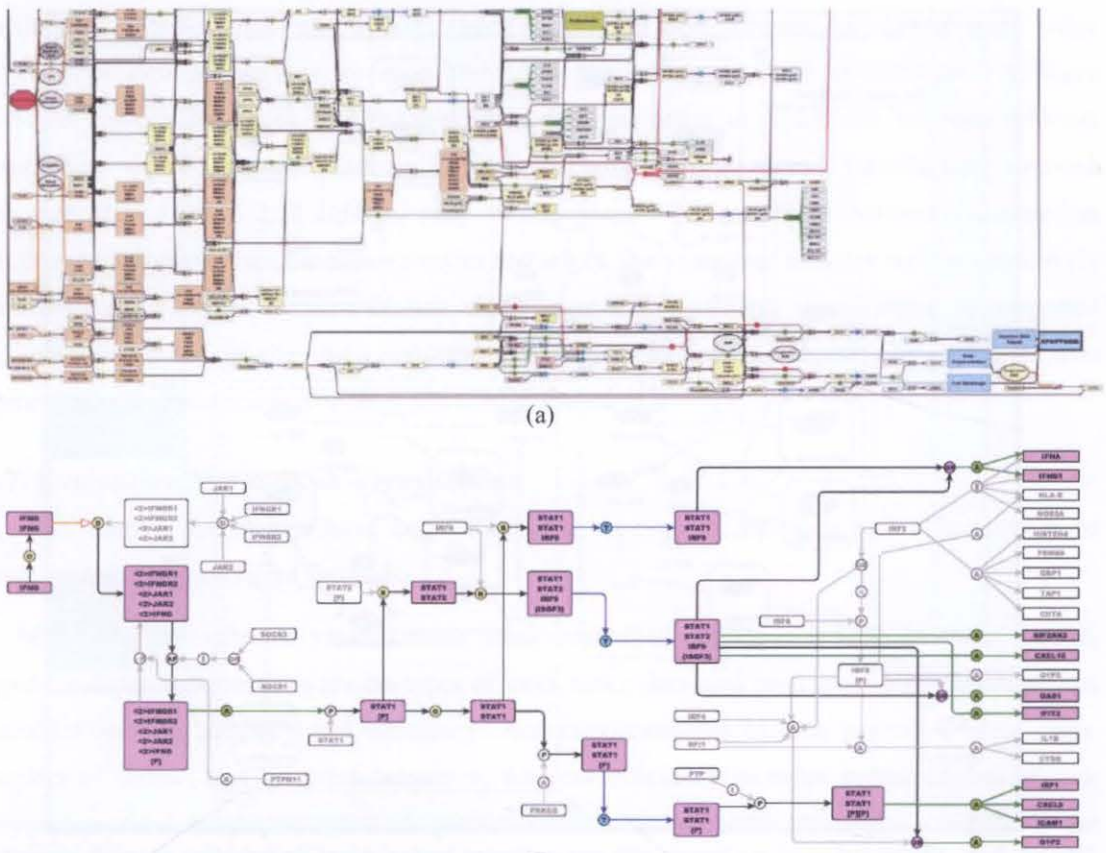


FIGURE 2.17. Visualization of the human signal transduction network (STN) in two different layouts. (a) The human STN in the hierarchal layout with orthogonal edge routing and (b) in the orthogonal layout. Reproduced from Raza 2008 [123].

They are proteins which interact with the gene regulators in order to facilitate the gene regulatory process [83]. Therefore the force-directed layout is not always the most effective visualization for biological analysis.

2.6.4. Signal Transduction Network

A signal transduction network (STN) is a collection of specialized protein-protein interactions which serve the purpose of signal relay and propagation. As with the GRN, STN is a directed network in order to account for the flow of signals. Therefore the visualization methods applicable to GRN can also be applied to the STN. For example, human STN has recently been visualized in the hierarchal layout [123] (see FIGURE 2.17(a)). Another approach is the orthogonal layout [123] (see FIGURE 2.17(b)). Both approaches provide a readable view on the directional flow of an STN. STN has also been visualized in the force-directed layout [34]. This approach is effective in exposing signaling hubs, i.e. nodes that have relatively high node degrees in a network of several hundred nodes [34]. Biologists can identify which hubs are convergent points and which are divergent points based on the combination of incoming and outgoing edges to and from the hub respectively.

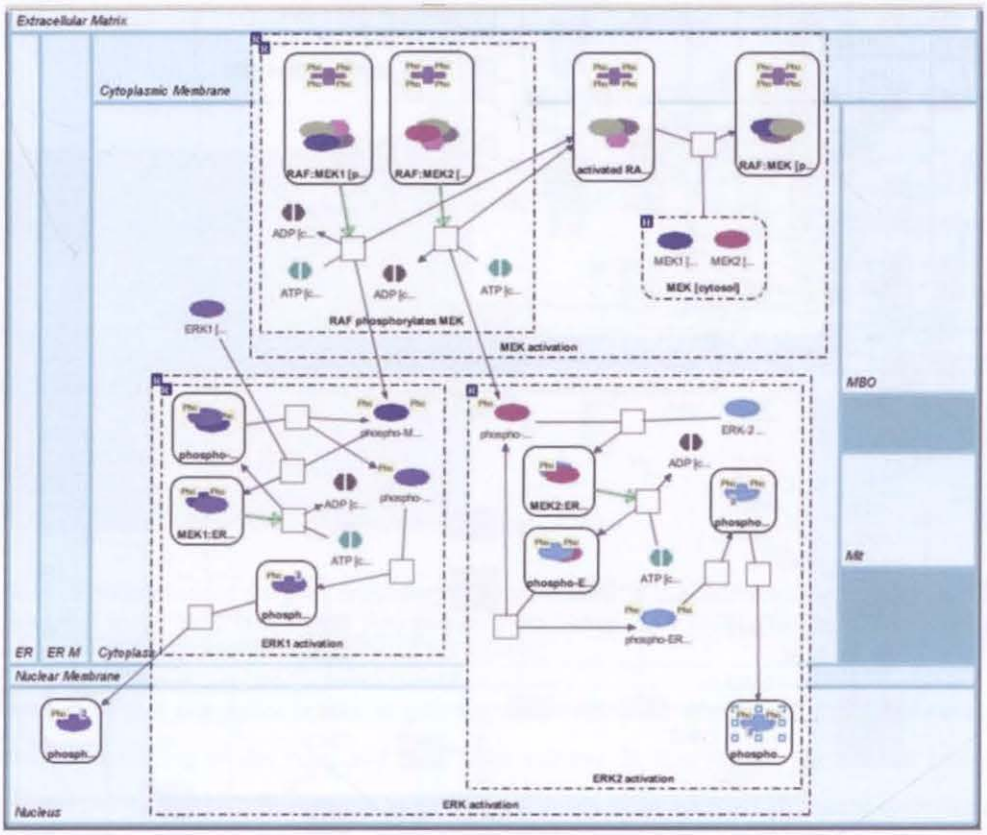


FIGURE 2.18. Visualization of the human mitogen-activated kinase signal transduction network generated by PATIKA [42]. The visualization contains different types of nodes and edges. Oval nodes represent proteins. Hexagonal nodes represent metabolites. Square-shaped nodes represent common molecular functions. Rectangular nodes with round corners represent cluster nodes containing a protein complex. Undirected edges represent physical interactions (blue colour). Directed edges with solid arrows represent signaling interactions (blue colour). Directed edges with hollow arrows represent metabolic reactions (green colour). Rectangular nodes bound by dotted edges represent biological processes. Rectangular nodes bound by solid light blue edges represent cell components.

However, edge crossings in the force-directed layout often disrupt the directional flow displayed in the visualization and reduce the effectiveness of this layout as a visual analysis method.

2.6.5. Integrated Network

An integrated network is one that combines different types of bio-molecular interactions. There have been very few examples of integrated networks which display multiple interaction types in one network visualization. The motivation is often based on the rationale that just visualizing a single interaction type fails to capture the biological reality. The reality is that biological processes usually involve more than one interaction type. Therefore, an integrated network which captures multiple interaction types should enhance biological analysis.

To date, only a few studies have been done on the automatic visualization of integrated networks, such as the Patika system [42] (see FIGURE 2.18). The challenges in visualizing integrated networks are the issues of scalability and complexity. Because the different interaction types are visualized in the same network drawing, complex visual encoding is required to

distinguish between molecular species (node types) and also between interaction types (edge types). For example, it requires six different glyphs in FIGURE 2.18 to distinguish between different molecular species, and requires three different edges to distinguish between different interaction types. Yet the human mitogen-activated kinase signal transduction network represented in FIGURE 2.18 contains only 44 nodes and 40 edges. If the network visualized in FIGURE 2.18 contains two times more nodes and edges, the visualized network will be cognitively challenging to interpret. Because of their visual complexity, readable visualization of integrated networks is limited in size. As a result, integrated networks are used mostly by biologists who come from computer science or other physical sciences.

2.7. Evaluation Methods on Visualization

A large number of studies have been conducted to evaluate usability and effectiveness of visualization using different methods.

Many studies evaluated visualizations using controlled experiments [27]. In these studies, typical independent variables are the types of tools, tasks, data, and participant classes. Dependent variables include accuracy and efficiency. Accuracy measures include precision, error rates, number of correct and incorrect responses, whereas efficiency includes measurements of task completion time where the tasks are pre-defined benchmark tasks. A classic example is the comparison among three different visualization systems on different tasks in terms of task completion time and accuracy published by Kobsa [86].

Usability tests usually evaluate visualizations to identify and solve user interface problems. The method involves observing the participants as they perform designated tasks using a 'think aloud' protocol, noting the usability incidents that may suggest incorrect use of interface, and comparing results against a pre-defined usability specification [69]. An example of a usability test is published by Rao and Mingay [122].

Different from empirical evaluations are inspections of user interfaces by experts, e.g. heuristics. Heuristic evaluation is a well known discount evaluation method used for finding usability problems at different developmental stages of a user interface. The procedure involves a small number of participants visually inspecting a user interface or visualization system according to a set of heuristics or guidelines. The heuristics used are relevant to the selected user interface and they exist as a shared knowledge on its design. To date, there are three known sets of heuristics proposed for information visualization, i.e. (1) selection of perceptual and cognitive heuristics [172], (2) visual information-seeking mantra [141], and (3) knowledge and task-based framework [4]. An example on evaluating a visualization system using these three heuristics is presented by Zuk et al. [172].

Longitudinal study examines the user's long-term exploratory learning of a user interface. The process involves noting usability incidents and recording the user's learning process. This type of evaluation usually takes a few days to accomplish rather than a few hours. An example of longitudinal study is presented by Rieman [128].

There were a few evaluations conducted with biologists to understand either their analytical tasks in biological research [134, 146] or the effectiveness of a visualization system in supporting analytical reasoning and hypotheses formulation [134, 135]. The earliest work was a survey of bioinformatics tasks commonly performed by biologists [146]. By using a combination of interviews and heuristic evaluations, a list of user requirements for an effective pathway visualization system was elucidated in another study [134].

2.8. Introduction to Molecular Systems Biology

From the computer science viewpoint, a single cell molecular system can be viewed as a state machine made of molecules that form a large and complex network. Within this network, cellular components and processes emerge from complex interactions among biological molecules. A *biological process* is a recognized series of events accomplished by one or more ordered assemblies of molecular functions [60]. The molecules in a single cell network can belong to one of the four classes: carbohydrates, lipids, proteins and nuclei acids [83]. It has been widely recognized by biologists that a majority of biological processes rely on the functioning of proteins and the information for encoding the chemical (amino acid) sequence of each protein is stored in the deoxyribonucleic acid (DNA). For this reasoning, it is worth elaborating of the biological roles of these two classes of molecules.

DNA is a form of digital encoding much like the magnetic storage tape that stores data in a linear sequence. Its primary function is to encode the instruction sets for synthesizing proteins. Each instruction set is encoded by a DNA sequence known as a *gene*. The two biological processes by which the information in a gene becomes a protein are known as transcription and translation. Biologists often called the decoding of the DNA instruction set which leads to the eventual synthesis of proteins as *gene expression*. The first step is transcription, i.e. the DNA is transcribed into mRNA. In essence, the mRNA is a template for synthesizing a particular protein. The production plant is embodied by a combination of proteins and tRNA. Together they form a protein complex known as the *ribosome*. The second step is translation, i.e. the ribosome translates the mRNA to proteins. If we compare gene expression to code compilation in software engineering, we will see similarities between the two (see FIGURE 2.19).

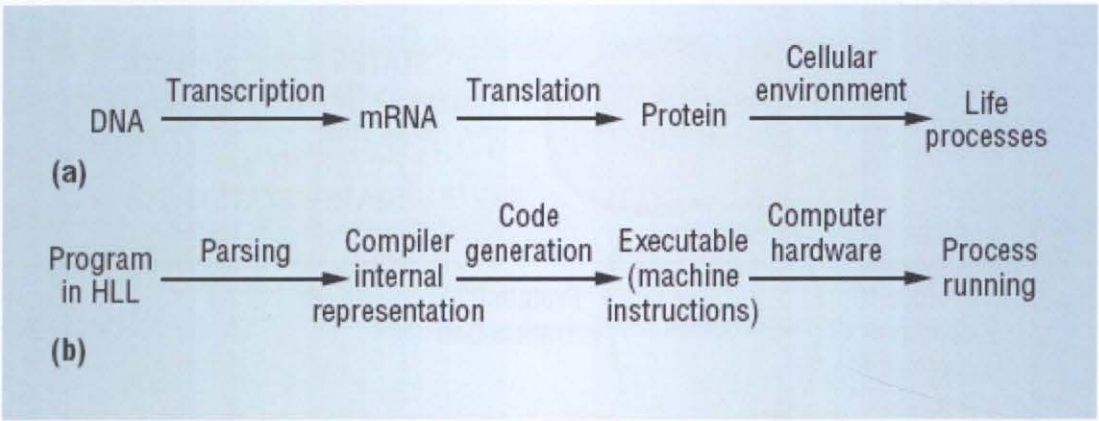


FIGURE 2.19. Similarities between (a) biological process function and (b) computer program execution. Both linear sequences progress from information to function. The acronym 'HLL' stands for *high level language*. Reproduced from Feitelson 2002 [48].

Because the initiation of transcription requires the direct physical interactions between proteins and DNA, the latter has to encode more than just the instruction sets for synthesizing proteins. It has to contain also a DNA sequence which serves as a docking site for proteins known as *gene regulators*. Such a special sequence is known as the *promoter*. The promoter is a control point that couples the DNA with its environment. Therefore the intracellular environment influences which proteins to synthesize within what extracellular environment. A living cell generally does not need to synthesize all the proteins encoded in its DNA. Only a small subset is synthesized. Which subset is synthesized will depend on which subset of gene regulators is present in the cell compartment called *nucleus*.

In computer science terms, a gene is therefore similar to the transition rule of a state machine [48]. The composition of the cell's interior, i.e. the protein set in all the cell compartments, determines the current state of the cell. The initiation of transcription by some members of the protein set will lead to the synthesis of new proteins thereby altering the state of the cell (see FIGURE 2.20). Hence, we can model a single cell molecular system as a state machine. This *state* is a snapshot of the system at a given time point that contains enough information to predict the behaviour of the system for future times [83].

A genome is therefore a library of transition rules [48]. From any given state, a cell can transit to any other states according to the currently enabled transition rules, i.e. the set of promoters being utilized. We can further simplify the state of a cellular molecular system as a string of bits indicating for each gene whether it is expressed (state = '1') or not expressed (state = '0'). The length of this string will be equal to the number of genes in a particular genome, i.e. the size of the genome. The two sub-domains in molecular cell biology, functional genomics and proteomics, are devoted to deciphering the cellular state machine. In functional genomics, microarrays are being used for measuring the cell's state by surveying the expression state of the genome.

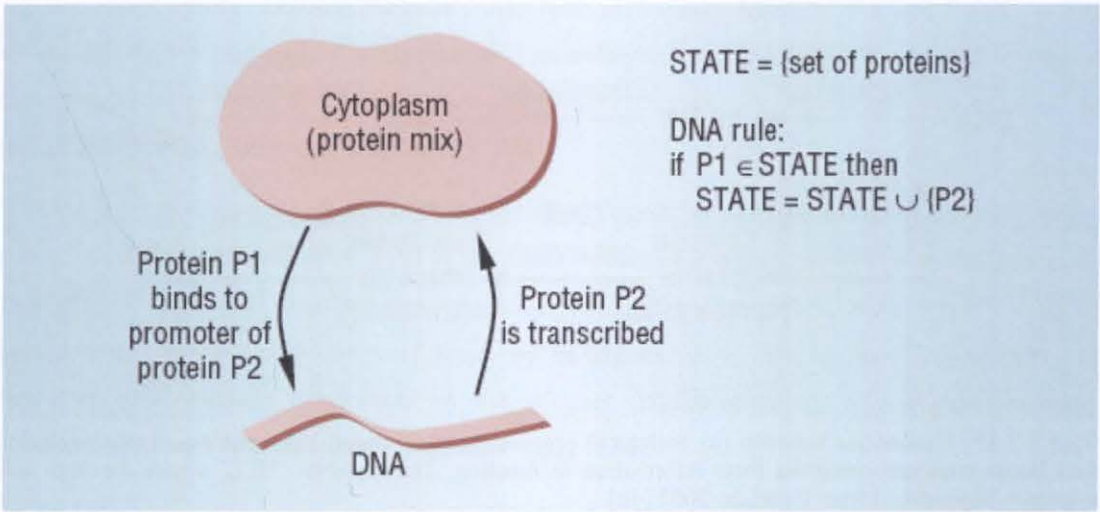


FIGURE 2.20. A simplistic concept model of a cell being a state machine. The expressed proteins are the state, and their interaction with the DNA causes new proteins to be expressed, thus changing the cell's state. Reproduced from Feitelson 2002 [48].

An event in which genes adopt similar expression states are known as a *co-expression*. Biologists often search for co-expression because co-expressed genes are more likely to function within the same biological processes (see Chapter 3).

In proteomics, high throughput mass spectrometry is being used for measuring the output of the genome. This way the expression state of the genome can be confirmed by examining the molecular abundance of each detectable protein. Different laboratory techniques, e.g. yeast two-hybrid, have been used to detect the protein-protein interactions in a cell. If two proteins are expressed and are known to interact with each other, the protein-protein interaction is in operation (see Chapter 4). Because a cell's state is associated with the various normal and disease conditions, the sub-domain of molecular medicine serves the objective of identifying abnormal cell states that can become our diagnostic marker.

The description of a single cell as a state machine so far ignores many biological complexities known to biologists. For example, proteins form protein complexes whose formation depends on the delicate quantitative balances that integrate inputs from many signalling paths, might regulate transcription. Nevertheless, the most basic concept model of a single cell molecular system is one which functioning requires the interaction between two fundamental types of information. The transition rules are largely encoded in the genome and the cellular state is largely encoded in the proteome. Most bio-molecular systems models are based on this principle.

{End of Chapter 2}

CHAPTER 3

Visualization of Gene Ontology-Annotated Co-Expressed Gene Clusters

"There is no Life without Organization"

3.1. Introduction

Gene Ontology (GO) has been used as a proxy for biological function and has increasingly been used for annotation and as a data mining dimension [121]. Because GO provides a higher level of abstraction than the network models of bio-molecular interactions, biologists often used a set of GO Process terms as a cryptic description on the functional organization of a cell. For this reason, the visualization of GO-annotated co-expressed gene clusters often serves as an entry point in microarray analytics. As a start, the biologist determines the set of co-expressed genes by computing pairwise correlation coefficients using Pearson or Spearman correlations [145]. This is then followed by clustering the co-expressed genes according to their commonly shared GO Process labels.

Co-expression between a pair of genes means that they have similar expression dynamics. This implies that they have a comparable functional context. If GO Process ontology is being used as an abstraction for such a functional context, then co-expressed genes are more likely to be involved in the same set of biological processes. The level of correlation as measured by the correlation coefficient is therefore an indication of their functional relatedness.

Recent studies suggest that genes which are exclusively clustered with a unique set of biological processes tend to co-express, giving rise to a functional module that exists in a steady state [147, 168]. A functional module is defined as a set of genes or their products (proteins or RNAs) which are related by one or more molecular interactions, e.g. co-regulation, co-expression, or membership of a protein complex, of a metabolic pathway, of a signal transduction pathway, or of a cellular component [70]. Many metabolic enzymes, e.g. glycolytic enzymes, are co-expressed because metabolic processes rely on the linear processing of metabolites. Many signaling proteins also exhibit similar expression dynamics under certain conditions because of the need to propagate an on/off signal through a specific path. In both cases, the co-expressed genes are induced or repressed by a specific set of gene regulators. Hence, the biological implication of co-expression is genetic co-regulation.

Most visualization provided as part of a software package is simply designed for capturing data structure not biological perspectives [53]. For this reason, we designed two prototype representations, i.e. the *block matrix* and the *clustered bipartite graph*, to capture one of the two current perspectives in molecular biology. The first perspective has its roots in reductionism and *genetic determinism*, i.e. '*DNA is the blueprint of life*' [12]. This is the conventional perspective among biologists. They consider genes as the basic building blocks of biology and so the functional organization of a cell is dictated by the information content of its genome. It is based on the presumption that one can understand the biology of a cell if one knows the function of every gene or different groups of genes within the genome. Biologists found that genes with similar functions are likely to be involved in the same biological process(es). This *gene-centric* perspective is being captured by the block matrix representation.

The second perspective, captured by the clustered bipartite graph, is the *network* view which stemmed from the emerging systems thinking. This view states that any attempts to reduce the whole system to smaller parts will destroy the properties emerged from the original scale of the system. Furthermore, this view regards molecular interactions as the building blocks of life, i.e. the network is the biology. Therefore all biological processes are operated by molecular interactions. The biological processes are themselves inter-connected because the entire molecular network is comprised of a highly inter-connected group of nested networks [3]. Each of these networks is involved in a particular biological process. As such, this view recognizes that it is the self-organizing property of the molecular network that gives rise to the cellular functional organization [6]. This network consumes information stored in the genome which in turn imposes a system constraint on the interaction types and the upper and lower limits of the network size.

Despite the frequent need to visualize co-expressed gene clusters in microarray analysis, the question on how different representations affect biological reasoning and usability is yet to be investigated. To answer the first part, a case study involving visual experimentation and biological analysis was performed on both representations. The domain application is hepatocellular carcinoma (HCC). The objective was to compare their effect on biological reasoning. Any hypotheses deduced in the case study also conveniently fulfill the auxiliary objective of understanding the biology of HCC based on its different functional organization from normal hepatocytes. To answer the second part, a usability evaluation was conducted with a group of bench biologists. The results obtained can fulfill two objectives. The first objective was to find out which representation is more suitable for microarray analytics. The second was to derive a set of design guidelines for visualizing GO-annotated gene clusters.

Our user evaluation is the first comparative evaluation on a specific pair of concept-based visualizations with the aim of testing their effectiveness in representing alternative biological concept models. It differs very much from the evaluations published by Saraiya *et al.* [134, 135] which aims at testing the effectiveness of the visualization choices provided by the various microarray analysis tools in hypothesis deduction. As such, our evaluations will answer the fundamental question as to whether a bioinformatics visualization should support the gene-centric or the network view, whereas Saraiya *et al.* answered the pragmatic question of which microarray software supports visual analytics better.

The rest of this chapter is divided into five sections. The representations are defined in section 3.2. The design criteria and the drawing algorithm of each representation are listed in section 3.3. The biological analysis on HCC is introduced in section 3.4. The design of the usability evaluation, the participants' background, and the results of the evaluation are elaborated in section 3.5. Of note, the analytical tasks listed in section 3.5.2 are the first benchmark tasks designed specifically for evaluating GO-annotated gene clusters and can be modified for evaluating any visualization of ontology-based clusters. Finally, section 3.6 serves as a concluding remark for this chapter.

3.2. Representation of Co-expressed Gene Clusters

3.2.1. Block Matrix

Given a set of co-expressed genes which are clustered by their common set of GO Process term(s), the block matrix is simply a set of non-intersecting clusters, i.e. each cluster has a unique set of co-expressed genes. Each gene is represented by a *gene node*. Each cluster is represented by a *cluster node* and has at least one GO Process as its node label. If the same GO Process term is common to more than one cluster, it will be redundantly displayed in multiple clusters. Hence, the *m*:*n* gene cluster-to-GO relationship is decomposed to a 1:*n* relationship. An identifier for each cluster `CLUSTER_ID` has been manually assigned to each cluster to facilitate their identification.

3.2.2. Clustered Bipartite Graph

From the same set of clusters mentioned in section 3.2.1, a unique set of GO Process terms have been derived and are represented explicitly by a set of nodes. The gene_cluster-GO relationships are then represented by edges connecting the cluster nodes and the GO nodes. The graph theoretic model of the clustered bipartite graph is defined as the following:

Definition 3.1. *A clustered bipartite graph is a graph* $G = (V, C, E)$ *in which* V and C *are the two finite and disjoint node sets. V denotes the set of GO nodes and C denotes the set of clusters. Each cluster* $c \in C$ *contains a unique set of gene nodes. It is possible that some*

clusters may share the same set of genes. Suppose c_i and c_j are any two clusters of gene nodes within the set of the clusters C, the intersection between c_i and c_j is non-empty, i.e. $c_i \cap c_j \neq 0$. E denotes the set of gene cluster-GO relationships.

When visualized, there is no redundant display of the GO Process terms as node labels. This is the biggest difference between the clustered bipartite graph representation and its block matrix counterpart.

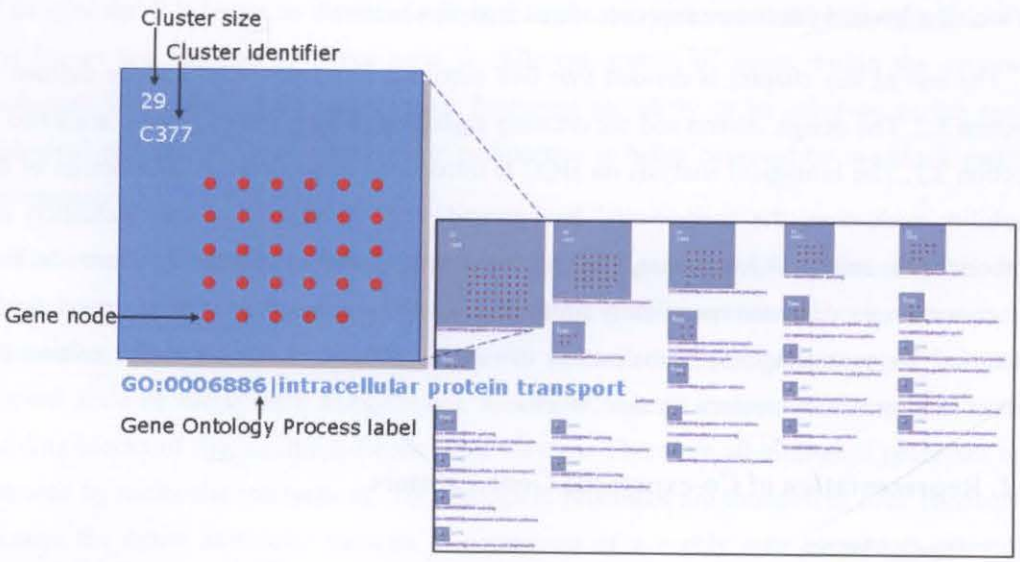


FIGURE 3.1. Visual representation of the block matrix.

3.3. Visualization of Co-expressed Gene Clusters

3.3.1. Block Matrix

When visualized, the block matrix is a set of matrices drawn within a grid layout (see FIGURE 3.1). Many grid layout algorithms were designed for maximizing space usage [17]. They generally gave a highly compartmentalized grid that is challenging to read. Our design emphasized on readability rather than optimum space usage.

The design criteria are:

1. The GO Process labels must be of a readable font size (12 points);
2. A maximum string size of 30 characters per GO Process label is allowed;
3. Each cluster must be homogeneously coloured and is in sharp contrast to the node colour;
4. Each cluster must be clearly bounded, and
5. No overlapping between clusters.

To highlight the cluster pattern, the cluster nodes and the gene nodes are assigned different shapes and colour. Each cluster node is a blue coloured square. Each gene node within a cluster is a red coloured circle meaning that the expression level of the represented gene has been detected. The size of each cluster is displayed on the upper left-hand corner of the corresponding square. If the area of the cluster is too small, the cluster size value will be drawn on the right-hand side of the square. This design requirement was recommended by two biologists in the preliminary user evaluation. While it lowers the information-to-ink ratio, the present design does fit the biologist's mental model of a gene cluster. They identify a red coloured node with an actively expressed gene. A group of red nodes within a square means that the genes in the cluster are also positively correlated in their expression levels. This design is also supported by the 'common field' principle proposed by Chmielewski *et al.* [31]. It stated that the user tends to see a set of objects as a group if they are being drawn within an explicitly bounded, homogeneously coloured or textured region.

The drawing algorithm involves four steps:

Algorithm 3.1. Block matrix algorithm

1. Sort the clusters by their sizes in the descending order using mergeSort.
2. Compute the grid formation (rows × columns). This is done by rounding the square root of the number of clusters to the nearest integer. If the initial number of rectangular partitions computed is smaller than the total number of clusters, an additional column is added.
3. For the largest cluster, compute the maximum width of the square area that can fit into the rectangular partition on the upper left-hand corner.
4. For the other clusters, compute the width of the square area in proportion to its relative size to the largest cluster. The origin of each cluster is the (x, y) Cartesian coordinates of the upper left-hand corner of each rectangular partition.
5. Reduce the vertical distance between clusters in each column so that the width of the resulting drawing becomes more compact.
6. Draw the clusters as blue-colored squares and in the descending order of their size across the grid from the left to the right.
7. For each cluster, compute the grid formation (rows × columns) required for positioning the gene nodes within. This is done by rounding the square root of the number of genes to the nearest integer. If the initial number of square partitions computed is smaller than the cluster size, an additional column is added.

8. Within each cluster, draw the gene nodes on the grid formation. The origin of each gene is the (x, y) Cartesian coordinates of the upper left-hand corner of each square partition.
9. Draw the GO Process labels under each rectangular cluster node.

3.3.2. Clustered Bipartite Graph

The clustered bipartite graph is drawn in a two parallel level layout (see FIGURE 3.2). The GO nodes and the gene cluster nodes are assigned to the upper and the lower levels respectively. The upper level is displayed as a black coloured line. The purpose is to impart a visual separation between the two node types. Each cluster of co-expressed genes is being enclosed within a circular node in green colour. The size of each cluster is written in parentheses at the bottom of the corresponding circle in a vertical orientation. Co-expressed genes within a cluster are being represented as circular nodes in red colour and GO nodes in blue colour. Edges between the GO nodes and the cluster nodes are in green colour. Its design criteria include those for the block matrix representation and the two additional criteria:

1. The edges between the GO nodes and the cluster nodes must be clearly displayed.
2. Edge crossing should be kept to the minimum with the cluster nodes being fixed in the descending order according to their size.

The drawing algorithm involved five steps.

Algorithm 3.2. Clustered bipartite graph algorithm

1. Arrange the gene clusters in the descending order of their size from left to right. For clusters of the same size, they will be laid out by random ordering.
2. Compute the area of each cluster node which is directly proportional to the size of the gene cluster.
3. Pack the gene nodes of each cluster into the appropriate cluster node using a phyllotactic layout [23] which has the advantages of spatial compactness and algorithmic simplicity. For the k-th gene node, it has the polar coordinates $(r, \Delta\theta)$.

$$r = q\sqrt{k}$$

$$\Delta\theta = k \cdot 0.753 \text{ radians}$$

where q is the *packing factor*. The Cartesian coordinates of the k-th node can then be computed as:

$$x_k = x_0 + r \cdot \cos(\Delta\theta)$$

$$y_k = y_0 + r \cdot \sin(\Delta\theta)$$

where (x_0, y_0) is the Cartesian coordinates of the centre of the cluster node, and r is the distance of the k-th node from the centre of the cluster node.

4. Minimize edge crossings by applying the barycenter algorithm [149]. The barycenter score $b(v)$ of every GO node $v \in V$ is defined as the average of the relative positions of its neighbouring cluster nodes. Thus,

$$b(v) = \frac{1}{|N(v)|} \sum_{i=1}^{n} pos(w_i)$$

where $N(v)$ is the number of neighbours to the GO node v and $pos(w_i)$ is the relative ordering of the cluster node w_i. The GO nodes are then sorted according to the ascending order of their barycenter scores and are drawn from left to right in regular spacing.

5. Draw the inter-level edges between the GO nodes and the cluster nodes.

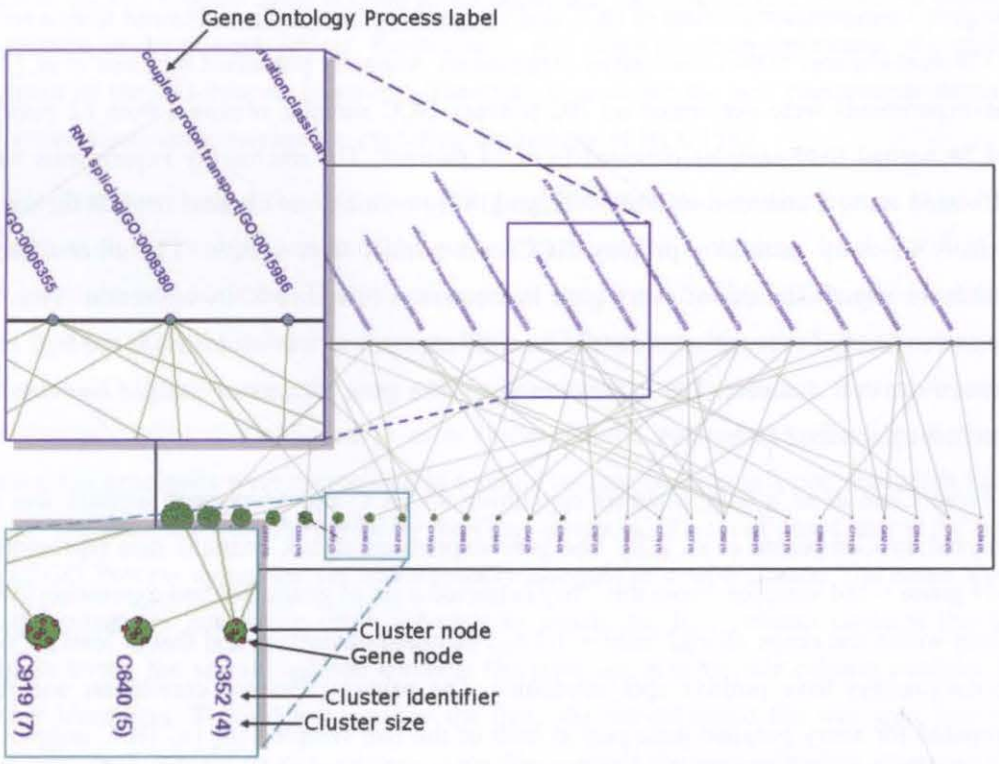


FIGURE 3.2. Visual representation of the clustered bipartite graph.

3.3.3. Implementation

The drawing algorithms for both representations were implemented using the *Processing* IDE version 1.15 [124]. Data for constructing each representation can be loaded into the *Processing* application as a tab-delimited file containing column values of the three

attributes: `CLUSTER_ID`, `GENE_ID`, and `GO_PROCESS`. Each attribute eventually became the labels for cluster, gene, and GO nodes respectively.

3.4. Case Study: Functional Organization of Hepatocellular Carcinoma

To compare the effect of the two representations on biological reasoning, the co-expression profiles of HCC and normal hepatocytes were applied to each. Through visual experimentation, the overall functional organization of each cellular phenotype was deduced. Because the visualizations were prototypes, interactivity for identifying the gene symbol[a] of each gene node was not yet available. However, the cluster patterns and the GO Process labels in each representation should provide a glimpse into the possible roles of the various biological processes in HCC and how the difference in functional organization as compared to normal hepatocytes could play a role in cancer development.

3.4.1. DataSet

3.4.1.1. Human liver gene expression data

The gene co-expression profiles of HCC and normal hepatocytes were obtained from a series of 176 dual-channel cDNA microarray experiments originally published by Chen *et al.* [29]. The experiments were performed on 102 primary HCC samples obtained from 82 patients and 74 normal liver samples obtained from 74 patients. The microarray experiments were performed using a common reference design [145] in which one channel records the signal intensity of every gene in a primary HCC or a normal liver sample. The other channel records the signal intensity of every gene in a common reference RNA collection. Thus the expression level of every gene in each microarray experiment was measured as the log2 ratio between the two channels. The expression level of a gene in a tissue sample was deemed significant if the log2 ratio ≥ 1.5.

From Chen *et al.*'s data, a subset containing 95 HCC and 66 normal samples was re-analyzed by Gamberoni *et al.* [58]. The gene expression matrix contains data representing 7449 genes × 161 samples. From this, they extracted a set of genes that had expression levels falling within the range of log2 ratio = 1.5 ± 3 standard deviations and that at least 75% of the microarrays have positive spot intensities. The pairwise Pearson correlation was then computed for every possible gene pair in each of the two sample sets, i.e. HCC samples as one and normal liver samples as another. Gamberoni *et al.* then extracted a set of co-expressed genes by filtering out gene pairs which Pearson correlation coefficients (PCC) are statistically insignificant, i.e. $p < 10^{-9}$. At a significance of $p \geq 10^{-9}$, the PCC ≥ 0.57617 for the HCC sample set and the PCC ≥ 0.66657 for the normal liver sample set. Finally, they

[a] The gene symbol is the standard nomenclature assigned to every human gene by the Human Genome Organization.

constructed a gene list from each sample set in which every pair of co-expressed genes share a common GO Process term. The gene lists are available as comma-delimited files in which the first column contains the GO identifiers, the second and the third columns contain the CloneIDs of the co-expressed gene pairs.

For the HCC sample set, the final gene list contains 163 genes annotated with 87 GO Process terms in 827 pairwise correlations. For the normal liver sample set, the final gene list contains 205 genes annotated with 88 GO Process terms in 419 pairwise correlations. For both gene lists, the GO Process terms were derived from levels 3-11 of the GO hierarchy in the Process category. Thus in each gene list, the same gene pair can be annotated with more than one GO term.

3.4.1.2. Data extraction

For the visual representations mentioned in this chapter, only gene clusters annotated with terms from levels 6 and 8 of the GO Process hierarchy were being visualized. This was because the GO Process terms at these levels accounted for 74% (1281/1732) of the pairwise correlation in the current dataset. Furthermore, this range of depth represented the middle segment of the GO Process hierarchy. Therefore it provided the best compromise between informativeness and coverage for exploring the biology of HCC [58].

We first downloaded the GO annotated gene lists provided by Gamberoni *et al.* [58] mentioned in the previous section. The gene pairs annotated with either level 6 or 8 of the GO hierarchy was extracted. The filtered gene list was first sorted by GO Process identifiers. This exposed the gene pairs that share multiple GO Processes. If a set of gene pairs shared the same GO Process term, the gene pairs were manually assigned to a cluster. A unique alphanumeric label was assigned to each cluster for the ease of identification. For each cluster, the gene pairs were normalized to a single column containing a non-redundant list of genes. The gene list was then sorted by the Gene Symbols. If a set of genes shares the same set of GO Process terms, the set was manually assigned to a new cluster. The result was a tab-delimited file containing three columns in which the first column contains the GO Process terms, the second column contains the gene list, and the last column contains the cluster identifiers. To further normalize the data, the tab-delimited file was split into two files with each containing two columns. One file contains the non-redundant cluster-to-GO Process mapping and the other contains the non-redundant cluster-to-gene mapping. These were the input files used for generating the visualizations described in the next section.

3.4.2. Visualization and Analysis

This section is divided into three subsections. In section 3.4.2.1, we described the visual effect of the block matrix and the clustered bipartite graph representations after the HCC

data was applied to each representation as an input. In section 3.4.2.2, we first described the rationale behind the visual analysis conducted and then the results of the analysis in relation to overview of the cluster pattern. In section 3.4.2.3, we described the rationale behind the visual analysis followed with the results of the analysis in relation to specific clusters. Finally, our conclusion to section 3.4.2 was given in section 3.4.2.4. In the following sections, the HCC dataset was referred to as the 'disease' sample whereas the normal hepatocyte dataset is referred to as the 'normal' sample. Where available, the Gene Ontology [60] identifier was given in parentheses for every biological process mentioned. Similarly, the Entrez Gene [99] identifier was given in parentheses for every human gene mentioned.

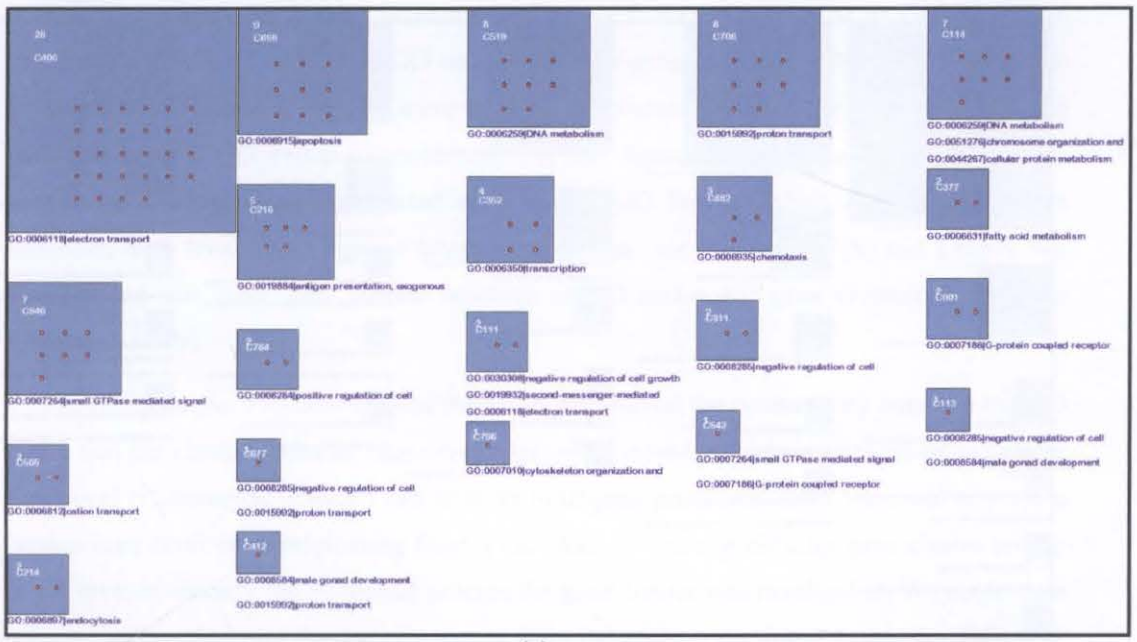
3.4.2.1. Visual effect of different representations

I. Block matrix

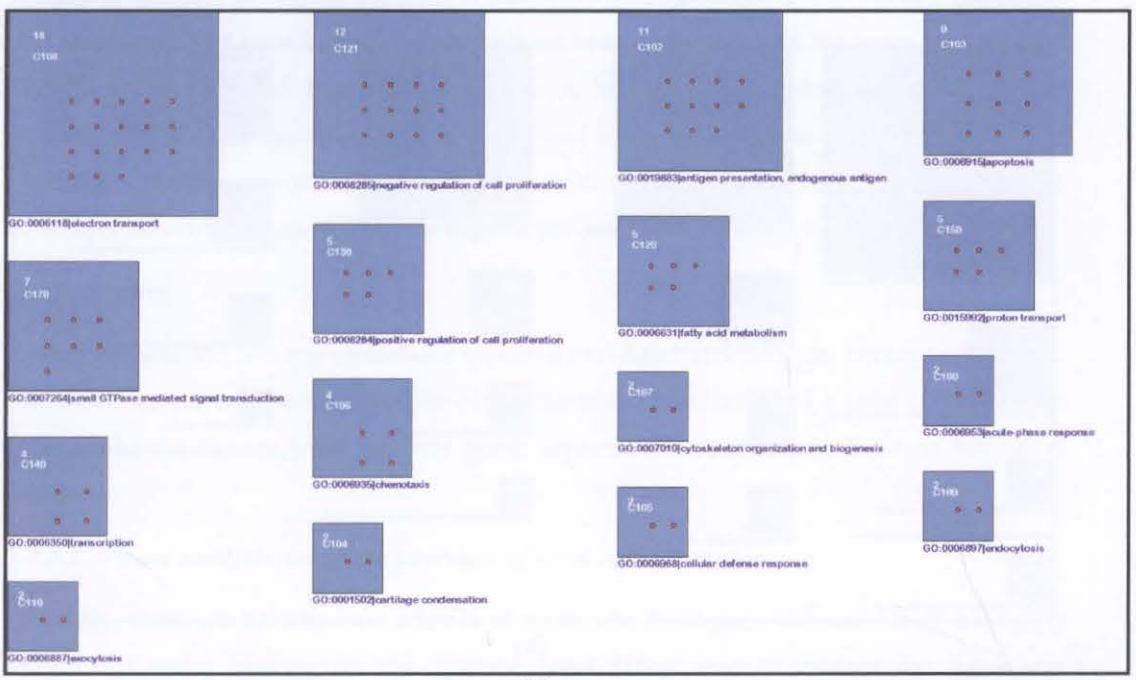
FIGURE 3.3 showed the block matrix representations of the two sample sets annotated with GO labels at level 6 of the GO Process hierarchy. FIGURE 3.3(a) showed the cluster pattern of the normal sample and FIGURE 3.3(b) showed the cluster pattern of the disease sample. FIGURE 3.4 showed the block matrix representations of the same sample sets but were annotated with GO labels at level 8. For the normal sample, the cluster pattern annotated with level 6 GO Process labels was visually less complex than its counterpart annotated with level 8 GO Process labels (see FIGUREs 3.3(a) and 3.4(a)). This was suggesting that as the set of GO labels became more informative, more functional relationships between gene clusters were being revealed. For the disease sample, the visual complexity between the cluster pattern annotated with level 6 GO Process labels and its counterpart annotated with level 8 GO Process labels were comparable (see FIGUREs 3.3(b) and 3.4(b)). For each sample, FIGUREs 3.3 and 3.4 clearly displayed the functional clustering of the co-expressed gene set. To identify the biological process(es) that the co-expressed genes in each cluster was involved in, we simply inspected the GO Process labels underneath the cluster. Therefore the block matrix representation effectively captured the functional modularity of gene co-expression. The block matrix representation was also visually simpler than the clustered bipartite graph representation (see FIGUREs 3.5 and 3.6) and therefore more readable.

II. Clustered bipartite graph

FIGURE 3.5 showed the clustered bipartite graph representations of the two sample sets annotated with GO labels at level 6 of the GO Process hierarchy. FIGURE 3.5(a) showed the connectivity between the gene clusters and the GO nodes of the normal sample and FIGURE 3.5(b) showed the case of the disease sample.

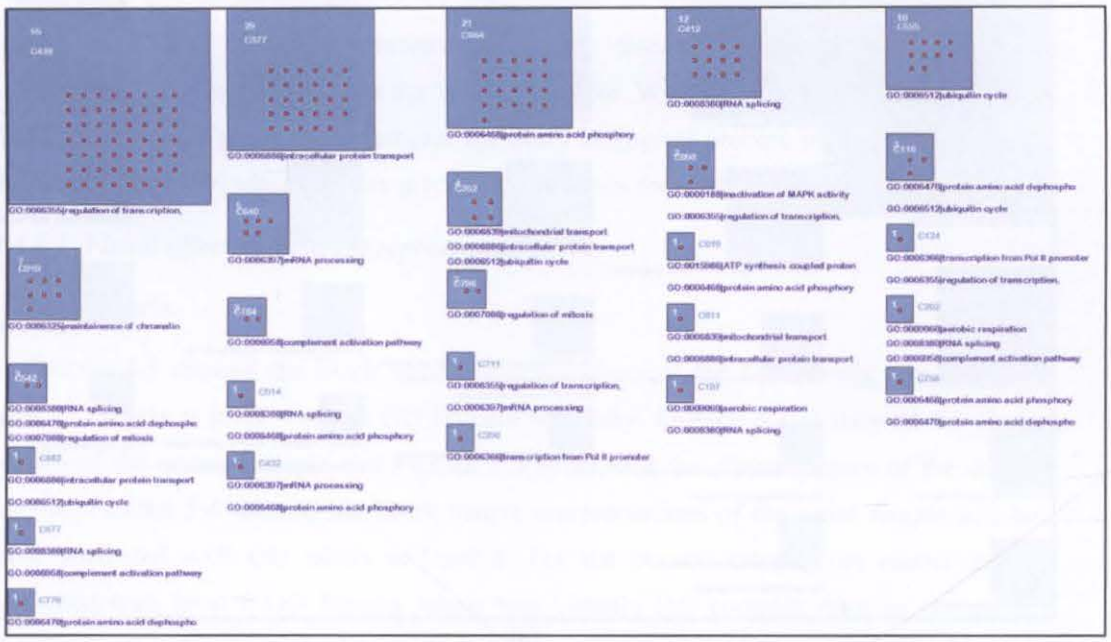


(a)

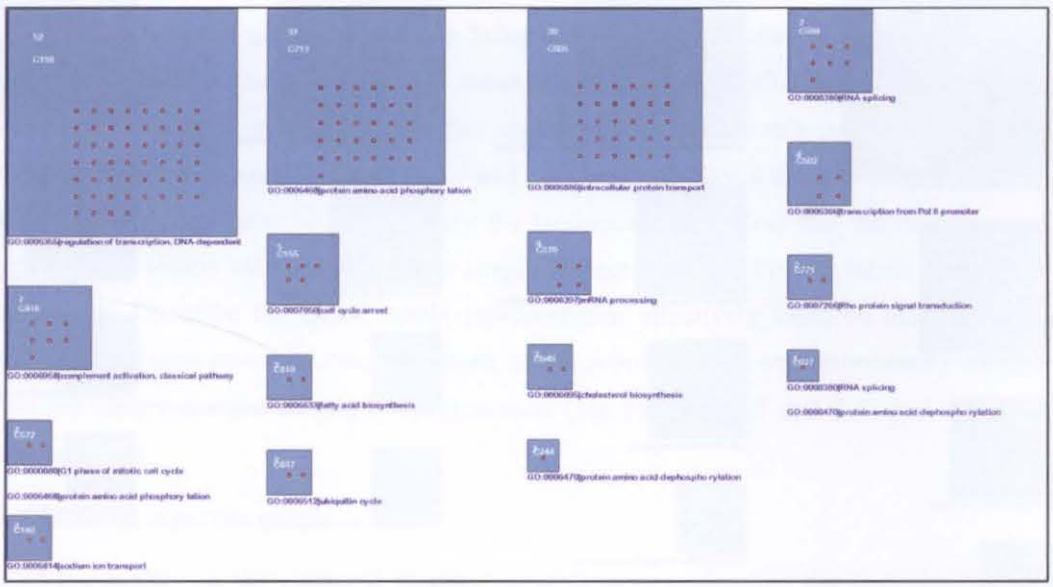


(b)

FIGURE 3.3. Visualization of co-expressed gene clusters in the block matrix representation with Level 6 GO Process annotation. (a) Normal hepatocyte. (b) Hepatocellular carcinoma.



(a)



(b)

FIGURE 3.4. Visualization of co-expressed gene clusters in the block matrix representation with Level 8 GO Process annotation. (a) Normal hepatocyte. (b) Hepatocellular carcinoma.

FIGURE 3.6 showed the clustered bipartite graph representations of the same sample sets but was annotated with GO labels at level 8. For the normal sample, the clustered bipartite graph representation annotated with level 6 GO Process labels was visually less complex than its counterpart annotated with level 8 GO Process labels (see FIGURES 3.5(a) and 3.6(a)). There were fewer edge crossings in FIGURE 3.5(a) as compared to FIGURE 3.6(a). We noticed that there were more GO nodes but fewer gene clusters in FIGURE 3.5(a) than in FIGURE 3.6(a). That was because some of the co-expressed genes in the normal sample did not have level 6 GO Process annotations. For the disease sample, the visual complexity between the representation annotated with level 6 GO Process labels and its counterpart annotated with level 8 GO Process labels were similar (see FIGURES 3.5(b) and 3.6(b)). We also noticed that there were similar numbers of GO nodes and gene clusters in FIGURES 3.5(b) and 3.6(b).

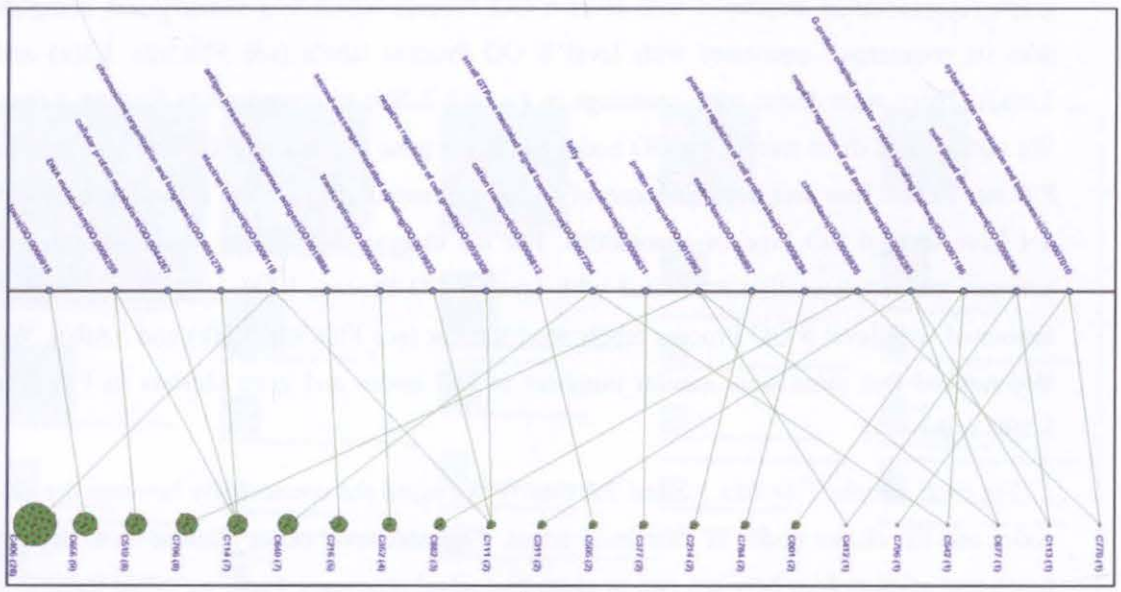
For each sample, FIGURES 3.5 and 3.6 clearly displayed the connectivity between the GO nodes and the cluster nodes as inter-level edges. The inter-level edges allowed us to deduce functional relationships between two or more biological processes. First, we could traverse a certain inter-level edge originating from a GO node to its neighbouring gene cluster on the lower level to identify the biological process the gene cluster was involved in. We could then traverse another inter-level edge originating from the same gene cluster to another GO node to identify the next biological process. In this way, we could deduce that the two biological processes identified must be functionally related because they shared the same gene cluster. Therefore the clustered bipartite graph was a better representation for capturing the connectivity between biological processes without losing the modular organization of gene expression. For this reason, it could provide us with an initial high-level abstract view of the molecular network in a normal human hepatocyte or an HCC cell.

III. Summary

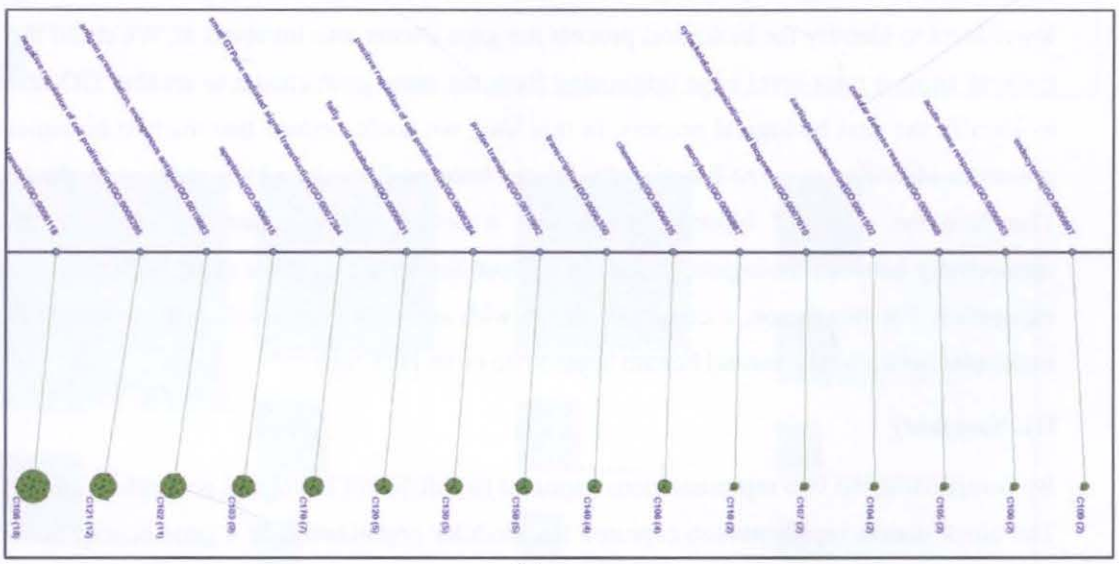
By comparison, the two representations captured two different biological conceptual models. The block matrix representation captured the modular organization of a gene co-expression dataset whereas the clustered bipartite graph captured the connectivity between biological processes.

3.4.2.2. Visual analysis using the overview of each representation

Currently, there are broadly two schools of molecular biologists. One school of biologists studies molecular biology on the systems level. They tend to deduce the functional organization of a single living cell as a system of biological processes held together by a network of co-regulated genes. The other school studies molecular biology in the reductionist approach.

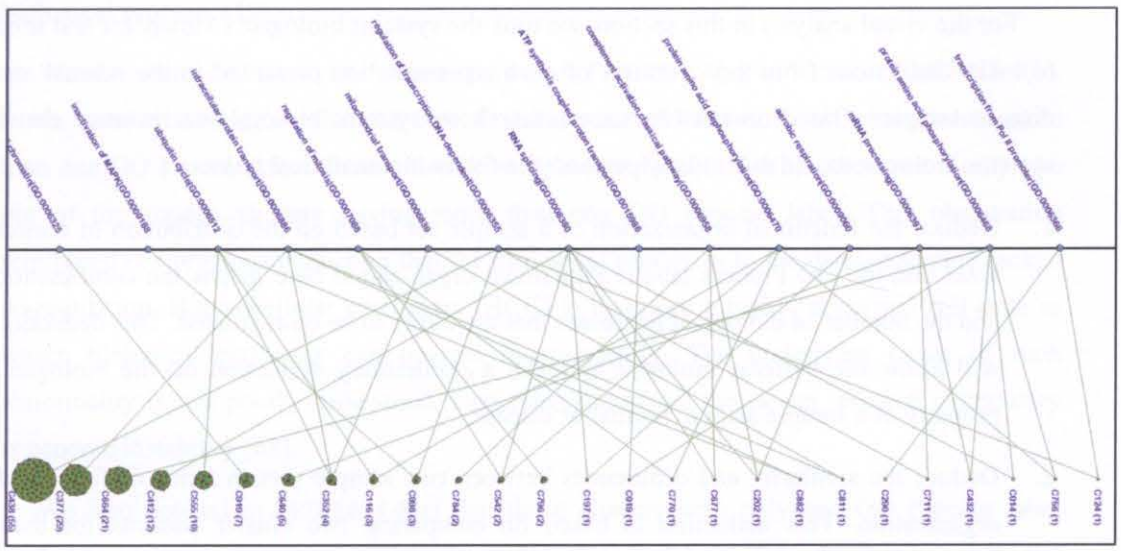


(a)

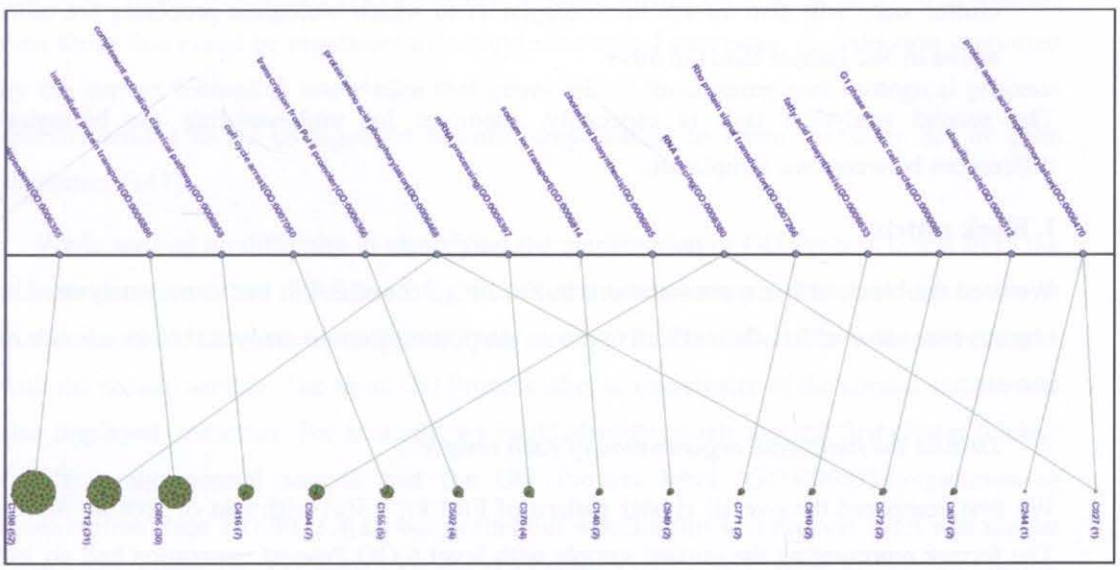


(b)

FIGURE 3.5. Visualization of co-expressed gene clusters in the clustered bipartite graph representation with Level 6 GO Process annotation. (a) Normal hepatocyte. (b) Hepatocellular carcinoma.



(a)



(b)

FIGURE 3.6. Visualization of co-expressed gene clusters in the clustered bipartite graph representation with Level 8 GO Process annotation. (a) Normal hepatocyte. (b) Hepatocellular carcinoma.

They will focus on those biological processes that they have special knowledge about and then identify gene clusters that have the GO Process labels of interest.

For the visual analysis in this section, we took the systems biologist's viewpoint and tried to make deductions from the overview of each representation presented in the normal and disease samples. Based on our literature research on systems biology, we assumed that a systems biologist would most likely perform the following analytical tasks.

1. Deduce the functional organization of a sample set based on the distribution of cluster sizes and the GO Process labels. Functional organization here means the combination and the number of biological processes that are likely to be co-regulated. This deduction will allow the systems biologist to make a preliminary deduction on the biological property of a sample set, i.e. normal or disease.
2. Deduce the similarity and differences between two sample sets in terms of functional organization. This deduction is based on comparing two cluster patterns for their similarities and differences in the combination of GO Process labels. Furthermore, the cluster sizes will also inform the biologist as to which biological processes are more active in one sample than the other.

The second analytical task is especially important for understanding the biological differences between two sample sets.

I. Block matrix

We used the block matrix representations in FIGUREs 3.3 and 3.4 in our visual analyses. The objective was to evaluate their effectiveness in supporting the two analytical tasks mentioned above.

1. Deduce the functional organization of each sample

We first compared the overall cluster pattern of FIGURE 3.3(a) with that of FIGURE 3.3(b). The former representing the normal sample with level 6 GO Process annotation had six out of the twenty one clusters labeled with more than one GO Process label. If a cluster had only one GO Process label, it informed us that its member genes were specialized in functioning within one particular biological process. If a cluster had multiple GO Process labels, its member genes were very likely to express proteins that were gene regulators of multiple biological processes.

We found that deducing the number of co-regulated biological processes in each sample is very easy. In the disease sample, none of the clusters had multiple GO Process labels (see FIGURE 3.3(b)) suggesting that there were no gene regulators being expressed. Therefore none of the biological processes in the disease sample was co-regulated. This deduction lent

support to the recent proposition that malfunctioning genes in human diseases were not necessarily biological process-specific but could be genes that function within multiple biological processes [98].

We also compared the overall cluster pattern of FIGURE 3.4(a) with that of FIGURE 3.4(b). In the normal sample (see FIGURE 3.4(a)), fifteen out of the twenty six clusters had more than one GO Process label. In the disease sample (see FIGURE 3.4(b)), there were only two out of the sixteen clusters having more than one GO Process label. This observation reinforced our previous deduction that the biological processes in the disease sample lacked co-regulation. Hepatocellular carcinoma (HCC) is therefore a highly abnormal cell state in which biological processes can operate independently. The underlying cause of such abnormality is still poorly understood. One suggestion from the cancer research community is genome instability [68].

We also noticed in FIGURE 3.4(a) that those clusters with only one GO Process label tended to be larger than those with multiple GO Process labels. This observation suggested that genes which functioning was biological process-specific were more likely to co-express than those that could be regulators of multiple biological processes. A deduction supported by the current biological knowledge that genes which functioning was biological process-specific tended to be co-regulated because they tended to share the same set of gene regulators [147].

While we had no difficulty in identifying the combination of GO Process labels from the disease sample, we had difficulty doing so with the normal sample. This was because most of the clusters in the disease sample had only one GO Process label. The same was not true with the normal sample. The same GO Process label in one cluster of the normal sample was also displayed in another. For example, we could identify clearly that the first cluster (cluster C438) in the normal sample had the GO Process label '*GO:0006355|regulation of transcription*' (see FIGURE 3.4(a)) but to find out whether this GO Process label was unique to cluster C438, we had to search for the '*GO:0006355|regulation of transcription*' label in all other clusters. This became tedious when we needed to perform the same action with every GO Process label. Therefore, the block matrix was ineffective for identifying the combination of GO Process labels in complex cluster patterns where the majority of clusters had more than one GO Process labels.

2. Compare the similarity and difference between samples

We tried to identify every GO Process labels that were either commonly shared or different in between the normal and disease sample sets using the block matrix representations in FIGURE 3.4, but found the analytical process tedious. It involved comparing each cluster in

the normal sample (see FIGURE 3.4(a)) with every cluster in the disease sample (see FIGURE 3.4(b)) for commonly shared as well as outstanding GO Process labels. After comparing the each of the first five clusters with every cluster in the disease sample, we decided to abandon the process and used the clustered bipartite graph representation instead (see the next section).

Thus far, we found that the block matrix representation did give us a glimpse on the functional organization of the normal and the HCC cells, but it did not support the comparative analysis between samples.

II. Clustered bipartite graph

After using the block matrix representations, we performed the same analytical tasks using the clustered bipartite graph representations in FIGUREs 3.5 and 3.6.

1. Deduce the functional organization of each sample

We compared the clustered bipartite graph representation in FIGURE 3.5(a) with that in FIGURE 3.5(b). FIGURE 3.5(a) showed that nine out of the twenty GO nodes were connected to more than one gene cluster with inter-level edges. By traversing inter-level edges originating from different GO nodes in FIGURE 3.5(a), we found that each of the nine GO nodes shared a co-expressed gene cluster with at least two other GO nodes. This finding suggested that some of the biological processes in the normal sample were co-regulated. On the other hand, by examining the node degree of the various gene clusters in the same figure, we found that seven out of the twenty two gene clusters were connected to more than one GO node. This finding suggests that those seven gene clusters were functionally related.

FIGURE 3.5(b) showed that each GO node in the disease sample was connected to a single gene cluster by an inter-level edge. Therefore, we deduced that the gene clusters in the disease sample were not functionally related and none of the biological processes were co-regulated. We could also identify the combination of GO Process labels readily for each sample because the GO nodes were clearly laid out on the first level and they were not redundantly represented.

We performed the same analyses using FIGUREs 3.6(a) and 3.6(b). We found that it was more difficult to identify GO nodes that shared the same co-expressed gene clusters with one another in the normal sample (see FIGURE 3.6(a)) because of the increased number of inter-level edges and edge crossings. We did not have any difficulties using the disease sample (see FIGURE 3.6(b)) to perform the same task owing to their visual simplicity. Therefore the increase in visual complexity due to the edge crossings reduced the usability of the clustered

bipartite graph representation somewhat even though it could still be used for identifying GO nodes that shared the same co-expressed gene clusters.

2. Compare the similarity and difference between samples

To compare the similarity and difference in functional organization between samples, we used FIGURE 3.6 because the normal sample gave a more complex clustered bipartite graph representation at level 8 GO annotation (see FIGURE 3.6(a)). We first identified in the normal sample those GO nodes that had a node degree ≥ 4. The rationale was that the higher was the node degree of a GO node, the more important was the represented biological process to the integrity of the cell's functional organization. We identified six such nodes. They were labeled '*GO:0006512|ubiquitin cycle*', '*GO:0006886|intracellular protein transport*', '*GO:0006355|regulation of transcription, DNA-dependent*', '*GO:0008380|RNA splicing*', '*GO:0006470|protein amino acid dephosphorylation*', and '*GO:0006468|protein amino acid phosphorylation*'. We then tried to identify GO nodes that had the same GO Process labels and found all six of them in the disease sample. We therefore deduced that the above six biological processes were essential to cell survival whether in normal cells or HCC cells.

When we compared the same six biological processes between samples for their differences in node degrees, we immediately recognized that in the disease sample, three of them had node degrees of two and the other three had node degrees of one (see FIGURE 3.6(b)). In other words, each of the six GO nodes connected to at most two gene clusters. This was markedly different from what we observed with the normal sample.

The deduction that could be drawn from this between-sample comparison was that these six biological processes were more actively co-regulated in the normal cell than in the HCC cell. This conclusion supported the notion that cancer is a state of systems failure [161]. We could also identify GO Process nodes that were unique in the normal sample but not in the disease sample using the same clustered bipartite graph representations (see FIGURE 3.7). To achieve this, we compared the GO node labels on the first level between the two sample sets. We found that there were four GO Process labels unique to the normal sample (see FIGURE 3.7(a)). They were '*GO:0007088|regulation of mitosis*', '*GO:0006325|maintainence of chromatin*', '*GO:0009060|aerobic respiration*', and '*GO:0015986|ATP synthesis coupled proton transport*'. The first GO Process label represented the biological process for regulating cell division. The second GO Process label represented the biological process for maintaining genome integrity. The last two represented biological processes for generating the energy molecules ATP (adenosine triphosphate). On the other hand, we found that there were four GO Process labels unique to the disease sample. They were '*GO:0000080|G1 phase of mitotic cycle*', '*GO:0007050|cell cycle arrest*', '*GO:0006633|fatty acid synthetic-*

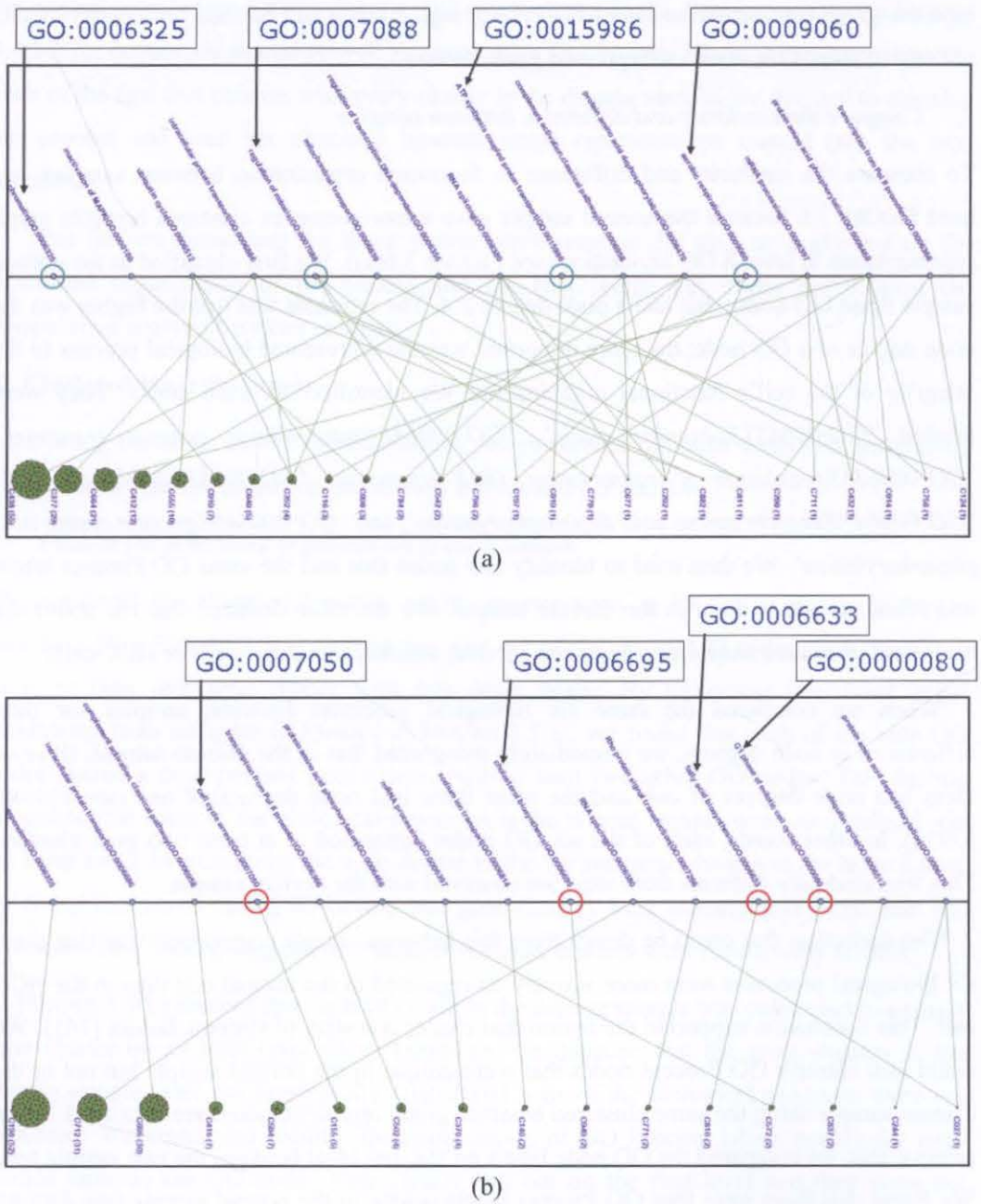


FIGURE 3.7. Visualization of the GO nodes unique to each sample seen in FIGURE 3.6. (a) The GO nodes unique to the 'normal' sample are circled in green and (b) those unique to the 'disease' sample are circled in red.

process' and '*GO:0006695|cholesterol biosynthetic process*' (see FIGURE 3.7(b)).

The first GO Process label represented the biological process for initiating cell division since G1 phase was the first phase of the cell cycle [83]. The second GO Process label represented the biological process for arresting cell division. The last two represented biological processes for synthesizing lipid molecules which were the main structural ingredients for the cell membrane and organelles [83].

Together, the different sets of sample-specific GO Processes suggested to us the different biological properties between normal cells and HCC cells. There could be a loss in genome stability in HCC cells because there were no co-expressed genes involved in chromatin maintenance (GO:0006325). Furthermore, there could be a loss in co-ordination between cell division initiation (GO:0000080) and its arrest (GO:0007050) in HCC cells. Cancer biologists had long suspected that uncontrolled cell division could lead to genome instability [43] and new evidence had been forthcoming [24]. They had also suggested that uncontrolled cell division could also lead to the increase in lipid biosynthesis (GO:0006633) because new cell membrane and organelles needed to be synthesized frequently [40]. We therefore deduced that HCC cells required active synthesis of new cellular components in order to prolong cell survival.

In general, the above analytical tasks were better supported by the clustered bipartite graph representation than by its block matrix counterpart. We conducted the same analyses with the block matrix representation, and had to tediously compare all possible pairs of clusters for their similarity in GO Process labels. This demonstrated the advantage of the clustered bipartite graph representation in preserving the original *m:n* gene_cluster-GO relationship as opposed to decomposing it to a 1:*n* relationship in the block matrix representation. That said, we relied heavily on inter-level edge traversal to confirm the gene_cluster-GO relationships seen in the clustered bipartite graph representations. Therefore the inter-level edges were the visual entities that enhanced our analytical reasoning during the visual analyses.

III. Summary

In summary, we found that the overview of the clustered bipartite graph representation allowed us to make the following tasks easier than using its block matrix counterpart.

1. Deduce the functional organization of each sample.
2. Deduce the biological differences between the samples.

In terms of HCC biology, the most important deduction here was that HCC is a highly abnormal cell state in which the co-regulation of biological processes is being lost.

3.4.2.3. Visual analysis using clusters pairs

As mentioned before, many biologists preferred examining those biological processes that they are familiar with and then identify gene clusters that have the GO Process labels of interest. They do not use overviews as their initial step of analysis. They tend to use a gene cluster set for two analytical tasks.

1. Identify co-expressed genes with similar biological functions. This is given by the GO Process labels of each cluster alone, but the relevance of these gene_cluster-GO relationships grow if two clusters share one or two identical GO Process labels. The relevance is that the clusters of concern are likely to be functionally related. Each cluster of genes could be responsible for a sub-process in a larger biological process.
2. Identify pairwise relationships between two samples. This involves examining a pair of gene clusters that are related in different samples, e.g. normal and disease. These clusters may have different GO Process labels but also share some identical GO Process labels.

For the visual analysis in this section, we took the reductionist biologist's viewpoint and tried to make deductions from selected clusters in each representation presented in the normal and disease samples.

I. Block matrix

We used the block matrix representations in FIGUREs 3.3 and 3.4 in our visual analyses. The objective was to evaluate their effectiveness in supporting the two analytical tasks mentioned above.

1. Identify co-expressed genes with similar biological functions

Since uncontrolled cell division is one hallmark of HCC [68], cancer biologists are often interested in identifying gene clusters that are involved in the regulation of mitosis (GO:0007088). Using the block matrix representations annotated with level 8 GO Process, we identified in the normal sample (see FIGURE 3.4(a)) that cluster C542 had three GO Process labels: '*GO:0006470|protein amino acid dephosphorylation*', '*GO:0007088| regulation of mitosis*', and '*GO:0008380|RNA splicing*' (see also FIGURE 3.8(a)). The proximity of the GO Process labels allowed us to identify the biological processes of which the co-expressed genes in C542 were involved in.

In the disease sample (see FIGURE 3.4(b)), we identified that cluster C572 had two GO Process labels: '*GO:0006468|protein amino acid phosphorylation*' and '*GO:0000080|G1 phase of mitotic cell cycle*' (also see FIGURE 3.9(a)). Because the two biological process -- the regulation of mitosis (GO:0007088) biological process and G1 phase of mitotic cell cycle (GO:0000080) biological process, are sub-processes of the larger biological process known as mitosis (GO:0007067) [60], we deduced that cluster C542 in the normal sample was functionally related to cluster 572 in the disease sample. We found that the block matrix representation was also useful for comparing juxtaposing clusters for their functional relatedness.

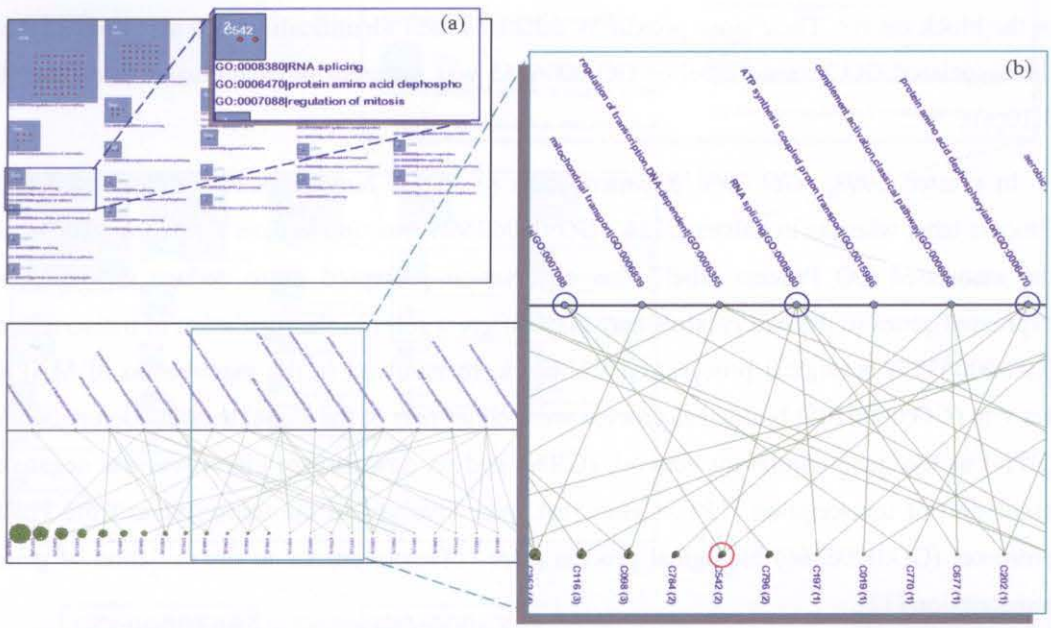


FIGURE 3.8. Visualization of the gene cluster C542 in different representations. (a) The zoom-in views of cluster C542 in FIGURE 3.4(a) (*blue box*) and in FIGURE 3.6(a) (*green box*) are shown in the insets. (b) The C542 is circled in red and its neighbouring GO nodes are circled in blue.

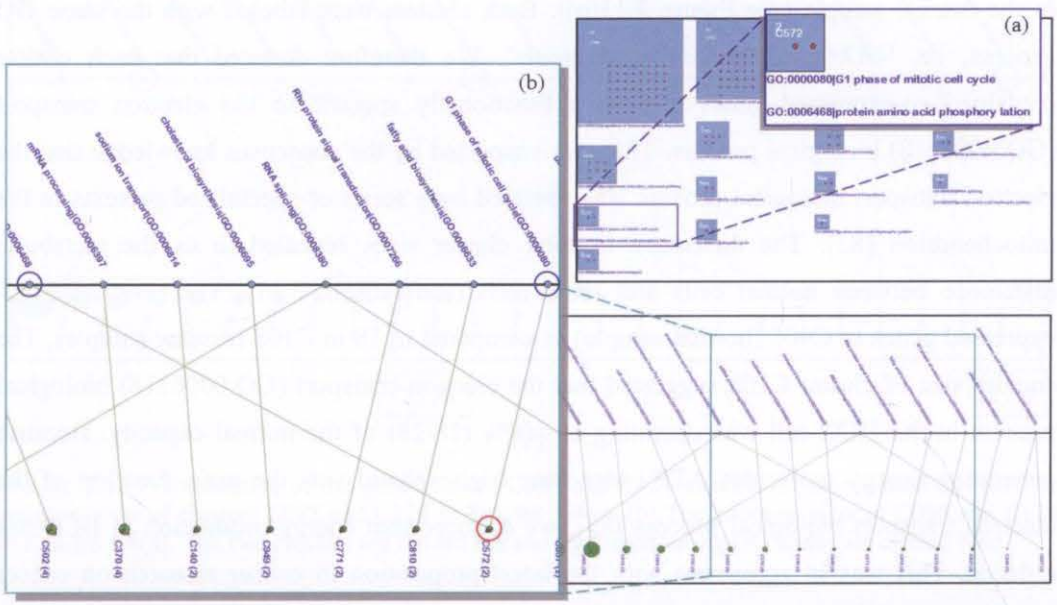


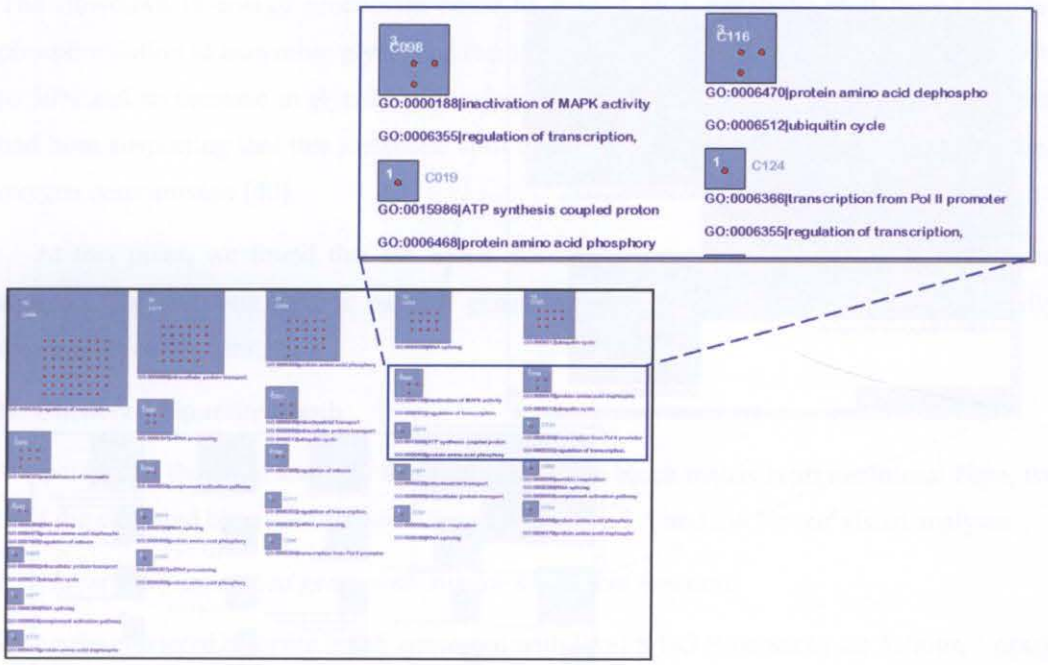
FIGURE 3.9. Visualization of the gene cluster C572 in different representations. (a) The zoom-in view of cluster C572 in FIGURE 3.4(b). (b) The zoom-in view of C572 in FIGURE 3.6(b). In the bipartite graph representation, Cluster C572 is circled in red and its neighbouring GO nodes are circled in blue.

This feature was especially applicable to the more complex cluster pattern of the normal sample (see FIGURE 3.10(a)). For example, we identified that '*GO:0006355|regulation of transcription*' was the GO Process label commonly shared between clusters C098 and C124 in the block matrix. Their close proximity enhances their identification. We also noticed that the associated GO Process label of GO:0006355 was different in each cluster (see FIGURE 3.10(a)).

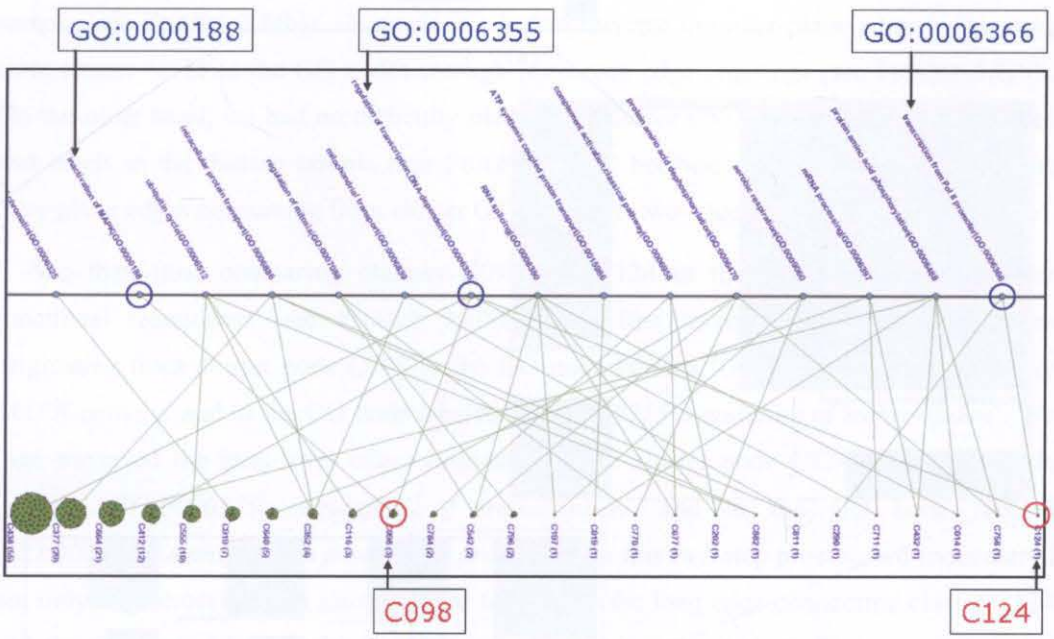
In cluster C098, '*GO:0000188|inactivation of MAPK activity*' was the associated GO Process label whereas in cluster C124, '*GO:0006366|transcription from PolII promoter*' was the associated GO Process label. This observation prompted us to deduce that the co-expressed genes in the two clusters may play different roles in the regulation of transcription (GO:0006355) biological process. Genes that were involved in the inactivation of MAPK activity (GO:0000188) biological process were either part of the signal transduction network (STN) or the gene regulatory network (GRN) and its downstream effect was the negative regulation of transcription [158]. Genes that were involved in the transcription from PolII promoter (GO:0006366) biological process were cofactors crucial to the initiation of gene transcription [12].

2. *Identify pairwise relationships between different samples*

Using the block matrix representations annotated with level 6 GO Process, we compared the largest cluster C406 in the normal sample (see FIGURE 3.11(a)) with the largest cluster C108 in the disease sample (see Figure 3.11(b)). Both clusters were labeled with the same GO Process, i.e. '*GO:0006118|electron transport*'. We therefore deduced that each cluster contained co-expressed genes that were functionally specific to the electron transport (GO:0006118) biological process. This was supported by the consensus knowledge that the electron transport biological process was operated by a series of specialized proteins in the mitochondrion [83]. The difference in their cluster sizes revealed to us the metabolic difference between normal cells and HCC cells (see FIGURE 3.11). There were 28 co-expressed genes in C406 (normal sample) as compared to 18 in C108 (disease sample). The smaller size of cluster C108 suggested that the electron transport (GO:0006118) biological process in the HCC cell was operating at ≈60% (18/28) of the normal capacity. Because generating energy molecules ATP (adenosine triphosphate) was the main function of the electron transport biological process [83], we deduced that energy production in HCC has reduced. This was in agreement with the latest proposition in cancer research on cancer metabolism.



(a)



(b)

FIGURE 3.10. Visualization of the gene clusters C098 and C124 in different representations. (a) The zoom-in view of clusters C098 and C124 in FIGURE 3.4(a). (b) The zoom-in view of C098 and C124 in FIGURE 3.6(a). The two clusters are circled red and its neighbouring GO nodes are circled blue.

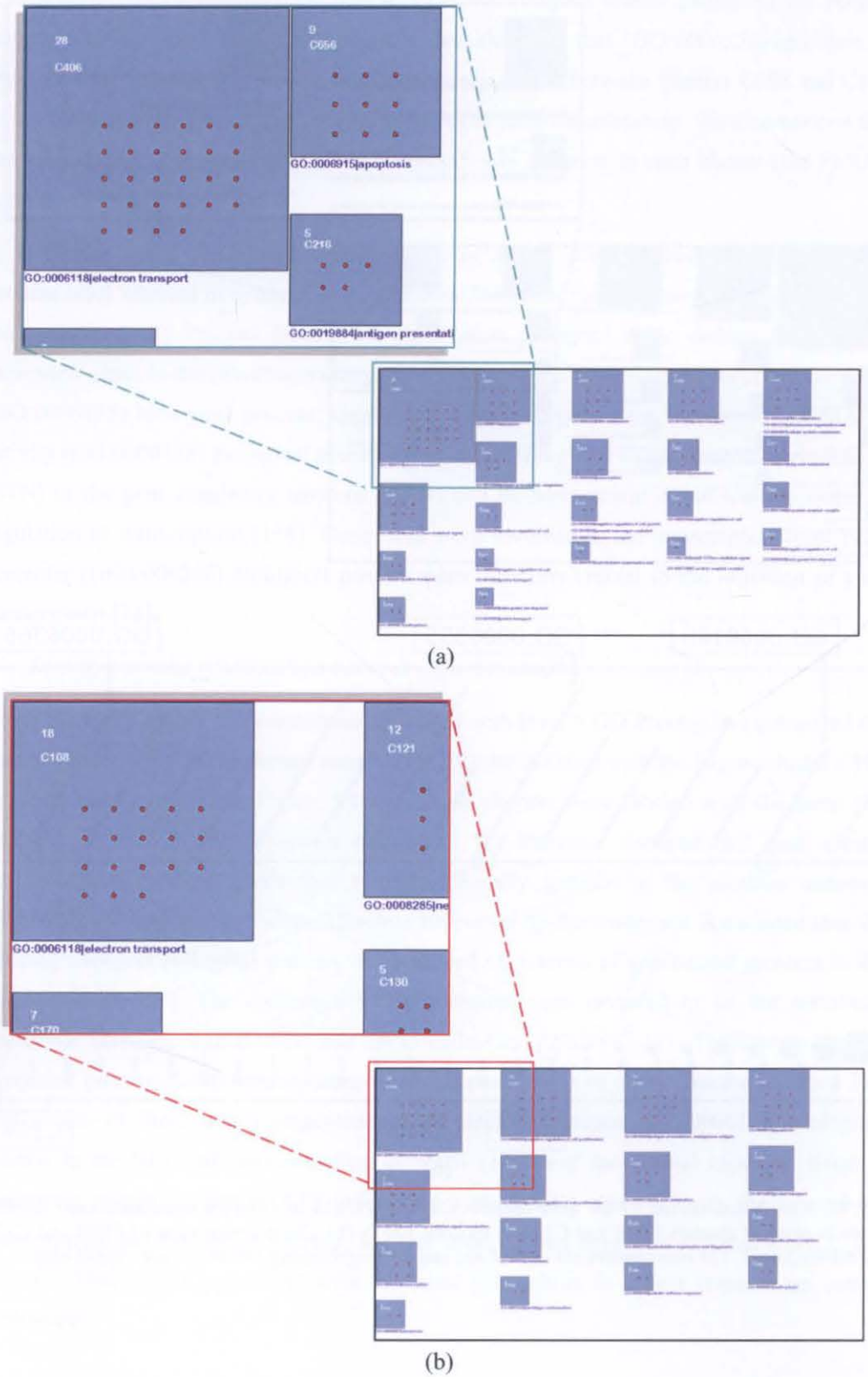


FIGURE 3.11. Visualization of the electron transport-specific gene clusters in the block matrix representations of FIGURE 3.3. (a) A zoom-in view of cluster C406 for normal hepatocytes (*green box*) and (b) cluster C108 for hepatocellular carcinoma (*red box*) are shown in the insets.

The slowdown in energy production could be caused by a metabolic shift from oxidative phosphorylation to anaerobic glycolysis resulting in a reduced production of ATP from 90% to 50% and an increase in glycolysis-based production from 10% to 50%. Cancer biologists had been suspecting that this metabolic shift could promote cancer cell survival by reducing oxygen consumption [40].

At this point, we found that the block matrix representation could effectively support analyses that involved specific pairs of gene clusters and allowed us to make biologically meaningful deductions.

II. Clustered bipartite graph

We performed the same analyses that we did with the block matrix representations. Here, we used the clustered bipartite representations in FIGUREs 3.5 and 3.6 for our visual analyses.

1. Identify co-expressed genes with similar biological functions

Using the clustered bipartite graph annotated with level 8 GO Processes (see FIGURE 3.6(a)), we could also identify cluster C542 and its GO Process nodes and labels in the normal sample (see FIGURE 3.8(b)). However, we had to traverse the inter-plane edges originating from cluster C542 to the GO nodes through numerous edge crossings (see FIGURE 3.8(b)). On the other hand, we had no difficulty identifying cluster C572 and its GO Process nodes and labels in the disease sample (see FIGURE 3.9(b)) because we only had to traverse the inter-plane edges originating from cluster C572 through two edge crossings.

We then tried comparing clusters C098 and C124 in the normal sample for their functional relatedness (see FIGURE 3.10(b)). We first traversed the inter-level edges originating from cluster node C098 to the GO node labeled '*GO:0000188| inactivation of MAPK activity*' and to the GO node labeled '*GO:0006355| regulation of transcription*'. We then traversed the inter-level edges originating from cluster node C124 to the GO node labeled '*GO:0006355| regulation of transcription*' and to the GO node labeled '*GO:0006366| transcription from Pol II promotor*'. In this two-step process, we encountered not only edge crossings but also the need to traverse the long edge connecting cluster C124 and the GO node labeled '*GO:0006355| regulation of transcription*'. The use of clustered bipartite graph representation did not alter our previous deduction made with the block matrix representation, i.e. the co-expressed genes in clusters C098 and C124 might play different roles in the regulation of transcription (GO:0006355) biological process. However, through this analysis, we recognized that the absence of edges in the block matrix representation made it a better choice for pairwise cluster comparison.

2. Identify pairwise relationships between different samples

Using the clustered bipartite graph annotated with level 6 GO Processes (see FIGURE 3.5), we could identify cluster C406 in the normal sample (see FIGURE 3.5(a)) and cluster C108 in the disease sample readily (see FIGURE 3.5(b)). In each sample, the respective cluster was the leftmost cluster and was connected to a GO node labeled '*GO:0006118| electron transport*'. In the normal sample, we traversed the inter-level edge originating from cluster C406 through only five edge crossings. In the disease sample, there were no edge crossings between C108 and its neighbouring GO node. Therefore, the visual simplicity in both representations allowed the ready identification of the selected gene_cluster-GO_Process relationship. Therefore, the clustered bipartite graph was as usable as its block matrix counterpart, only when there were few edge crossings interfering with inter-level edge traversal.

III. Summary

In summary, we found that the block matrix representation supported the following tasks better than using its clustered bipartite graph counterpart. This could due to the absence of edges in the block matrix representation.

1. Identify pairwise functional relationship between clusters within the same sample.
2. Identify pairwise relationships between clusters of different samples.

In terms of HCC biology, the most important deduction here was that HCC has a reduced energy production compared to normal cells.

3.4.2.4. Conclusion

In conclusion, both representations had their strengths and limitations. On the one hand, the block matrix seemed to be more suitable for examining the biological processes of an individual cluster and pairwise inter-cluster comparison for functional relationships. On the other hand, the clustered bipartite graph seemed to be more suitable for comparing between sample sets thereby facilitating the deduction of biological differences between different samples. Since this was the deduction fundamental to hypothesis formulation, the clustered bipartite graph would be better suited to microarray analysis than the block matrix representation.

In the next section, a usability evaluation conducted with a group of biologists was being presented. The purpose was to examine empirically their experience in using each representation in biological analysis. The results should inform us as to whether they identified the same strength and limitations in each visual representation.

3.5. Usability Evaluation

A task-oriented evaluation was performed to systemically quantify the biologist's performance and experience with each representation. The evaluation was designed to identify the strength and limitation of each representation. We anticipated that the participants evaluating the block matrix representation should perform better than its clustered bipartite graph counterpart. The reason behind was that the block-matrix representation captures the gene-centric model that biologists were most familiar with. We further hypothesized that edge crossing in the clustered bipartite graph representation will impede the participant's performance in the competency tasks and the conceptual tasks.

3.5.1. Experimental Design

This experiment was setup to examine three independent variables and three between-subject dependent variables. The three independent variables were: microarray datasets (normal hepatocytes and HCC [58]), representations (block matrix and clustered bipartite graph), and task types (competency and conceptual tasks). The three dependent variables were: task completion time, accuracy, and user confidence score. Accuracy was defined as the percentage of the total number of tasks being correctly answered. Each representation was presented as a static visualization without any interactivity to ensure that the participant's performance was influenced only by the cluster pattern of each representation.

Based on our understanding of microarray analysis, the tasks were designed with an emphasis on finding and interpreting the visual features of GO-annotated gene clusters that a typical microarray user will perform. There were ten analytical tasks. The first five (tasks A-E) were *competency tasks* designed to test the readability of each representation. The last five (tasks F-J) were *conceptual tasks* designed to test the usability of each representation in analytical reasoning. The analytical tasks and the use case scenario of each task were presented in the following section.

3.5.2. Analytical Tasks

3.5.2.1. Competency tasks

The five competency tasks are listed as in the following:

A. Find the gene cluster that is linked to the largest number of GO IDs (Questions 1 for normal sample; Question 2 for disease sample).

Use case scenario. A biologist may want to identify co-expressed genes that are involved in multiple biological processes. A cluster linked to multiple biological processes is an indication that its member genes could be control points for coupling or decoupling some of the biological processes associated with a cellular phenotype.

B. Find the gene cluster that is linked to the smallest number of GO IDs from each sample set (Question 3 for normal sample; Question 4 for disease sample).

Use case scenario. A biologist may want to identify co-expressed genes that are functioning in a particular biological pathway. A cluster linked to only one or two biological processes is an indication that its member genes are functionally specific.

C. Find the GO ID(s) that has the largest number of co-expressed genes from each sample set (Question 5 for normal sample; Question 6 for disease sample).

Use case scenario. A biologist may want to identify the biological processes that are the most active in the normal or the disease sample. A biological process with a larger number of co-expressed genes than others often indicates that it is relatively more active.

D. Find the GO ID(s) that has the smallest number of co-expressed genes from each sample set (Question 7 for normal sample; Question 8 for disease sample).

Use case scenario. A biologist may want to identify the biological processes that are the least active in the normal or the disease sample. The rationale is the opposite of task C.

E. Find the GO IDs that are active only in the NORMAL or DISEASE sample set (Question 9 for normal sample; Question 10 for disease sample).

Use case scenario. A biologist may want to identify the biological processes that are specific to a particular phenotype. This task is frequently performed not only in bio-technology and medicine but also in agriculture where biologists want to compare the functional differences between plant or livestock species.

3.5.2.2. Conceptual tasks

The five conceptual tasks are listed as in the following:

F. Deduce which biological process is the most highly regulated (Question 11 for normal sample; Question 12 for disease sample).

Use case scenario. A biologist may want to identify biological processes that are highly regulated relative to others. Biologists interpret co-expression as synchronized activity among a group of genes and therefore must be co-regulated. The number of co-expressed genes linked to a biological process is an indicator of how highly regulated it may be.

G. Deduce which biological process is likely to be the least regulated (Question 13 for normal sample; Question 14 for disease sample).

Use case scenario. A biologist may want to identify which biological process(es) is the least regulated relative to others. The rationale is the opposite of task F.

H. Deduce which biological processes are likely to be co-regulated with the ubiquitin cycle (GO:0006512) and has the largest number of co-expressed genes (Question 15 for normal sample; Question 16 for disease sample).

Use case scenario. A biologist may want to use a particular biological process as a focus for investigating its connection with the other biological processes.

I. Deduce which human tissue the diagrams could most likely represent (Question 17 for both samples).

Use case scenario. While this task is not a reflection of the real-world scenario, it was designed to test the usefulness of GO-annotated gene clustering in biological deduction. In particular, it is a test on whether the GO terms in the visualization are representative of liver physiology.

J. Deduce which disease could the DISEASE TISSUE most likely represent (Question 18 for both samples).

Use case scenario. The rationale is similar to task I. In this case, it is a test on whether the GO terms in the visualization are representative of liver cancer.

3.5.3. Participants

Since the representations had been designed for use in biological research, results obtained from the evaluation were informative only if the participants were expert biologists with different research interests. Our choice of participants emphasized on their quality as domain experts in biology. Fourteen participants were recruited from four medical research institutes and two university biology departments. They had research interests in various fields of biology, e.g. biochemistry, cardiology, immunology, oncology, pharmacology, and virology. The group consisted of two group leaders who were also holding lecturer positions, four postdoctoral fellows, two research assistants, and six doctorate degree students. Among them, three were also practicing clinicians from two teaching hospitals. All of them were practicing bench biologists with no formal qualifications in computer science or information technology.

3.5.4. Procedure

Each session started with the evaluator explaining to the participant the design of the representation, the nature of each task and how to fill out the questionnaire. The participant

was given a trial session to familiarize oneself with the procedure using a synthetic dataset. In the proper session, each participant performed tasks A to H twice, once on the normal sample set and once on the disease sample set. The exceptions are tasks I and J. They were performed once because the tasks demand the participants to compare both samples using the same representation. This gave rise to 18 questions on the questionnaire. For each task, the participant was timed, observations were gathered, and the answers to the questionnaire were collected. At the end of each session, the participant was asked to provide a subjective rating on a five-point scale (0 to 4) to indicate one's confidence in performing each task. The higher the score, the higher is the participant's confidence. The participant was also free to express one's opinion about the evaluated representation in writing.

In the following sections, the group of participants in the block matrix evaluation was referred to as the 'block matrix group' and those in the clustered bipartite graph evaluation were referred to as the 'bipartite graph group'.

3.5.5. Results

3.5.5.1. Competency tasks

A. Find the gene cluster that is linked to the largest number of GO IDs (Questions 1 and 2).

Evaluation result. The median time spent by both groups on questions 1 (normal sample set) and 2 (disease sample set) was comparable (see FIGURE 3.12(a)). Both groups gave the same number of correct responses to questions 1 and 2 (see FIGURE 3.12(b)). When asked to give a user satisfaction score on a 5-point scale, both groups gave a median score of 3 to question 1 and a median score of 4 to question 2 (see FIGURE 3.12(c)). Thus participants within each group were quite confident in performing task A with their confidence in answering question 2 slightly higher than question 1. In general, user performance on task A by both groups was comparable. Therefore neither representation has an advantage over the other when being used to perform task A.

B. Find the gene cluster that is linked to the smallest number of GO IDs from each sample set (Questions 3 and 4).

Evaluation result. The median time spent by both groups on questions 3 and 4 was comparable (see FIGURE 3.12(a)). For question 3 (normal sample set), the number of correct responses given by the bipartite graph group was higher than its block matrix counterpart (6 *cf.* 4). With respect to user confidence scores, the bipartite graph and block matrix groups gave median scores of 2 and 3 respectively. Thus the block matrix group was more confident in finding a solution to question 3 than its bipartite graph counterpart even though the latter group gave a higher number of correct responses.

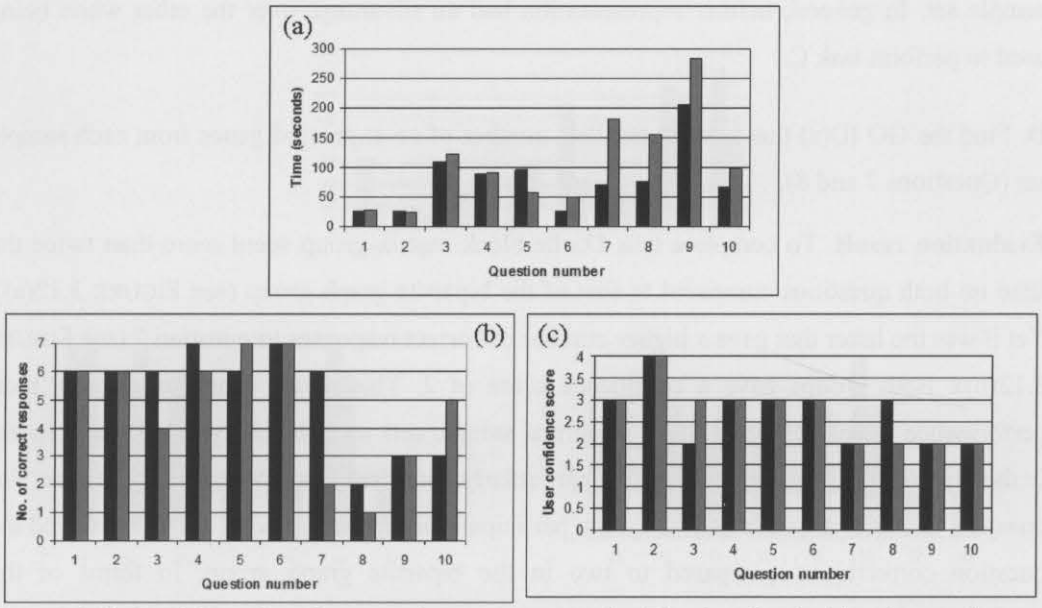


FIGURE 3.12. Participants' performance in competency tasks. The data for the bipartite graph group and its block matrix counterpart are shown in *dark* and *grey* bars respectively. (a) Median time spent per question. (b) Number of correct responses per question. (c) Participant's confidence score per question.

There was little performance discrepancy between the two groups for question 4 (disease sample set). The bipartite graph group gave a slightly higher number of correct responses than its block matrix counterpart (FIGURE 3.12(b)). Both groups gave the same median confidence score of 3 indicating that they were equally confident in answering the question (FIGURE 3.12(c)). In general, the bipartite graph helped improving the number of correct responses but not in task completion time and user confidence when used to perform task B on the normal sample set.

C. Find the GO ID(s) that has the largest number of co-expressed genes from each sample set (Questions 5 and 6).

Evaluation result. To complete task C, the bipartite graph group spent nearly twice the time on question 5 but spent only half the time on question 6 compared to that of the block matrix group (see FIGURE 3.12(a)). For question 5 (normal sample set), the block matrix group gave a slightly higher number of correct responses than its bipartite group counterpart (7 *cf.* 6). For question 6 (disease sample set), both groups gave the same number of correct responses (see FIGURE 3.12(b)). For both questions, both groups gave the same median confidence score of 3 indicating that they were equally confident in performing task C (see FIGURE 3.12(c)). In general, both groups were comparable in terms of the number of correct responses and user confidence score. Results on the task completion time suggested that the bipartite graph group encountered greater difficulty in performing the task on the normal

sample set. In general, neither representation had an advantage over the other when being used to perform task C.

D. Find the GO ID(s) that has the smallest number of co-expressed genes from each sample set (Questions 7 and 8).

Evaluation result. To complete task D, the block matrix group spent more than twice the time on both questions compared to that of the bipartite graph group (see FIGURE 3.12(a)). Yet it was the latter that gave a higher number of correct responses to question 7 (see FIGURE 3.12(b)). Both groups gave a confidence score of 2. This means that even though their performance in answering question 7 (normal sample set) was obviously better, participants in the bipartite graph group did not feel particularly confident in answering this question. For question 8 (disease sample set), only one participant in the block matrix group answered the question correctly as compared to two in the bipartite graph group. In terms of the confidence score, the bipartite graph and block matrix groups gave a median score of 3 and 2 respectively (see FIGURE 3.12(c)). Thus, participants in the bipartite graph group were more confident than their block matrix counterpart in answering this question. In general, the bipartite graph helped improving the participant's overall performance in task D.

E. Find the GO IDs that are active only in the NORMAL or the DISEASE sample set (Questions 9 and 10).

Evaluation result. To complete task E, the block matrix group spent 38% and 47% longer on questions 9 (normal sample set) and 10 (disease sample set) than the bipartite graph group respectively (see FIGURE 3.12(a)). For question 9, both groups gave the same number of correct responses. For question 10, the block matrix group gave a higher number of correct responses (see FIGURE 3.12(b)). For both questions, both groups gave the same median confidence score of 2 indicating that they are neutral in their level of confidence on performing task E (see FIGURE 3.12(c)).

To answer question 9, the participant was required to compare the normal tissue with the disease tissue and vice versa for question 10. The results indicated that the clustered bipartite graph was more suitable for answering question 9 by reducing the task completion time, whereas the block matrix was more suitable for answering question 10 by improving the number of correct responses. Therefore, the block matrix does not have an absolute advantage over the bipartite graph for deducing the biological differences between phenotypes.

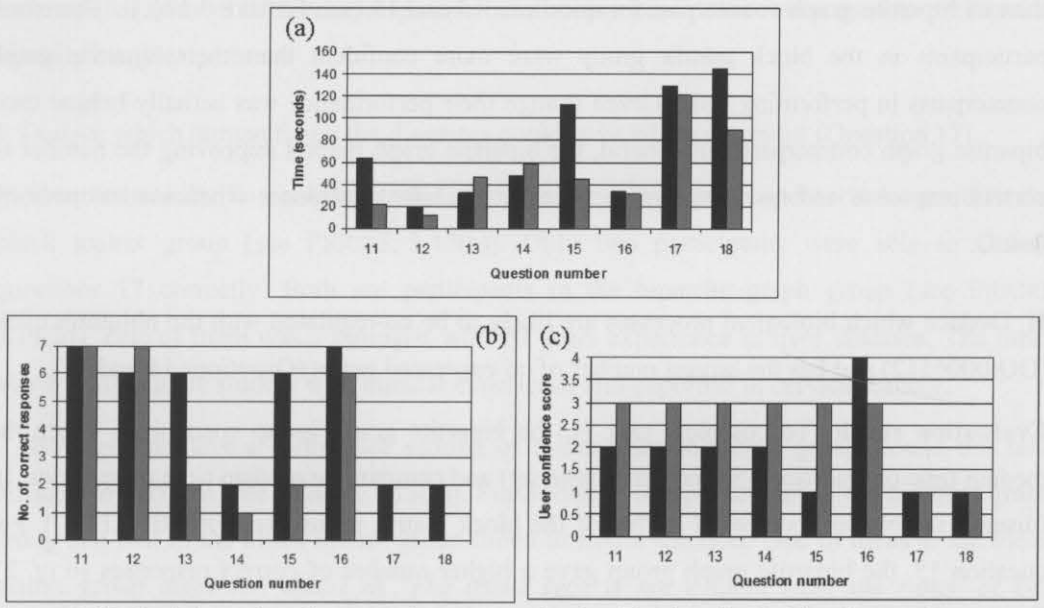


FIGURE 3.13. Participants' performance in conceptual tasks. The data for the bipartite graph group and its block matrix counterpart are shown in *dark* and *grey* bars respectively. (a) Median time spent per question. (b) Number of correct responses per question. (c) Participant's confidence score per question.

3.5.5.2. Conceptual tasks

F. Deduce which biological process is the most highly regulated (Questions 11 and 12).

Evaluation result. To complete task F, the bipartite graph group spent three times the median time on question 11 (normal sample set) and spent twice the time on question 12 (disease sample set) compared to that of the block matrix group (see FIGURE 3.13(a)). However, the two groups were comparable in the number of correct responses (see FIGURE 3.13(b)). In this respect, the participants' performance on accuracy was comparable to task C (questions 5 and 6). For both questions, the block matrix group gave a higher confidence score than its bipartite graph counterpart (see FIGURE 3.13(c)). Therefore participants in the block matrix group were more confident than their bipartite graph counterparts in performing task F. In general, the results were obviously indicating that the block matrix helped improving the task completion time and user confidence when used to perform task F.

G. Deduce which biological process is likely to be the least regulated (Questions 13 and 14).

Evaluation result. To complete task G, the block matrix group spent 32% longer on question 13 (normal sample set) and spent 17% longer on question 14 (disease sample set) than the bipartite graph group (see FIGURE 3.13(a)). Yet it was the latter that gave a higher number of correct responses to question 13 (see FIGURE 3.13(b)). For question 14, only one in the block matrix group answered the question correctly as compared to two in the bipartite graph group. However, it was the block matrix group that gave a higher confidence score

than its bipartite graph counterpart for questions 13 and 14 (see FIGURE 3.13(c)). Therefore, participants in the block matrix group were more confident than their bipartite graph counterparts in performing task G even though their performance was actually behind their bipartite graph counterparts. In general, the bipartite graph helped improving the number of correct responses and task completion time but not user confidence when used to perform task G.

H. Deduce which biological processes are likely to be co-regulated with the ubiquitin cycle (GO:0006512) and has the largest number of co-expressed genes (Questions 15 and 16).

Evaluation result. To complete task H, the bipartite graph group spent three times the median time on question 15 (normal sample set) and comparable median time on question 16 (disease sample set) compared to that of the block matrix group (see FIGURE 3.13(a)). For question 15, the bipartite graph group gave a higher number of correct responses (6 *cf.* 2). For question 16, the bipartite graph group gave a slightly higher number of correct responses (see FIGURE 3.13(b)).

For question 15, the block matrix group gave a higher confidence score than its bipartite graph counterpart (3 *cf.* 1). For question 16, the bipartite graph group gave a higher confidence score than its block matrix counterpart (see FIGURE 3.13(c)). Thus participants in the block matrix group are equally confident in performing task H on both sample sets, whereas the participants in the bipartite graph group are more confident in performing the same task only on the disease sample set.

With the block matrix, task H involved identifying all the gene clusters that had GO:0006512 as a listed GO Process label and then comparing between clusters for the GO Process labels associated with GO:0006512 in each sample set. The shorter median task completion time given by the block matrix group did support the view that the block matrix facilitated pairwise inter-cluster comparison for functional relationships. However, it might not improve the number of correct responses if the cluster pattern is visually complex as seen in the normal sample set.

With the bipartite graph representation, task H involved traversing edges that link GO:0006512 with other GO nodes via the gene clusters. The longer median task completion time by the bipartite graph group could be due to the numerous edge crossings present in the normal sample set, but that did not result in a lower number of correct responses than the block matrix group. With the lower visual complexity of the disease sample set, both groups became more comparable in their median task completion time and the number of correct responses for question 16. Thus the bipartite graph might help improving the task completion

time and the number of correct responses but not user confidence when used to perform on visually complex patterns.

I. Deduce which human tissue the diagrams could most likely represent (Question 17).

Evaluation result. To complete task I, the bipartite graph group spent 9% longer than the block matrix group (see FIGURE 3.13(a)). Only two participants were able to answer questions 17 correctly. Both are participants in the bipartite graph group (see FIGURE 3.13(b)). One of them was a biologist with 20 years experience in liver diseases. The other was a postgraduate student with clinical experience and expertise in cervical cancer.

Both groups gave a confidence scoring of 1 indicating that both groups found this task difficult to perform (see FIGURE 3.13(c)). Furthermore, one participant in the bipartite graph group and two in the block matrix group failed to find a solution. One of them in the block matrix group answered task I as '*The tissue type is not evident from the range of GO classifications represented as they are too general to draw an assumption.*' The answers deduced by all participants and their scientific interests are listed in TABLE 1.

J. Deduce which disease could the DISEASE TISSUE most likely represent (Question 18).

Evaluation result. To complete task J, the bipartite group spent 1.6 times the median time compared to that of the block matrix group (see FIGURE 3.13(a)). Only two participants were able to answer question 18 correctly. Both were participants in the bipartite graph group (see FIGURE 3.13(b)). One of them was an expert in liver disease who also answered question 17 correctly. The other was a postgraduate student who was a practicing clinician with expertise in Marfan's syndrome. One participant in the bipartite graph group and one in the block matrix group failed to find a solution. Both groups gave a confidence scoring of 1 indicating that both groups found this task difficult to perform. The answers deduced by all participants and their scientific interests are listed in TABLE 2.

3.5.6. Participants' Post-Task Comments

In the post-task briefing, participants in each group were asked to give their opinion about their respective representation and whether the representation was relevant to their research interest in biology. Four participants from the bipartite graph group commented that the edge crossings in the normal sample set were interfering with their graph reading. However, five of the six participants thought that with some improvements such as colour-coded edges, colour-coded GO Process labels or more readable font size, the bipartite graph would be still be useful for their microarray analysis.

TABLE 1. Participants' deductions for question 17 (Task I)

No. of participants	Participants' deduction	Participants' rationale	Participants' domain expertise
2	Blood	Presence of complement pathway (involved in immune cell function) and Rho protein expression.	Virology Cardiology
1	Kidney	There is a lot of fatty acid biosynthesis and cholesterol and etc.	Immunology
1	Liver	Not stated.	Hepatology Pharmacology
1	Lung	Less cell growth in the disease.	Clinical genetics
1	Muscle	There are ion transport genes, as well as aerobic respiration, and cell cycle, protein biosynthesis gene clusters. Significant co-expression of protein metabolic processes involved.	Cardiology Population genetics
1	Thyroid	Not stated.	Oncology Microbiology Virology
2	Heart	Diseased tissue has cholesterol biosynthesis. Presence of mitochondria components, need of aerobic respiration and maintenance of chromatin to maintain contractility.	Virology Immunology
1	Immune cells	Because of the presence of GO:0006958|complement activation pathway.	RNA molecular biology Lipid biochemistry
1	Vascular	Energy production is strong in normal tissue, whereas fat accumulation seems to be predominant in disease tissue.	Virology Oncology
3	Don't know	The tissue type is not evident from the range of GO classifications represented as they are too general to draw assumptions.	Oncology Clinical genetics

TABLE 2. Participants' deductions for question 18 (Task J)

No. of participants	Participants' deduction	Participants' rationale	Participants' domain expertise
3	Atherosclerosis	Due to the increased fatty acid and cholesterol biosynthesis and the increased expression of complement activation genes signifies inflammation. These processes are all characteristics of the development of atherosclerotic plaques. Diseased tissue has fatty acid and cholesterol biosynthesis, and sodium ion transport, whereas the normal tissue doesn't. The disease tissue may be a connective tissue and fat accumulation seems to be predominant in disease tissue. Macrophages turning into foam cells.	Immunology Lipid biochemistry Oncology Virology
1	Endocrine disease	Because genes in cholesterol and fatty acid synthesis are involved. I thought there may have been problems in the production of hormones which require these substances.	Immunology Cardiology
1	Hepatocellular carcinoma	Not stated.	Hepatology Pharmacology
1	Pulmonary carcinoma	Impaired cell growth and division and repair.	Microbiology Immunology
1	Herpes viral infection	Herpes virus infection as this virus shuts down cellular transcription and there are far few genes expressed in the disease tissue than the normal tissue.	Virology
2	Leukemia	Due to the increased number of complement activation pathway genes. Regulation of transcription has become 'detached' from other GO terms related to cell transformation.	RNA molecular biology Oncology
1	Muscle-related mitochondrial disease	Genes that are co expressed are involved in protein transport, phosphorylation processes.	Cardiology Population genetics
1	Nephritis	Due to the up-regulation of complement activation and cell cycle arrest genes.	Immunology Microbiology
1	Thyroid disease	This is due to regulation of immune function via the classical complement activation pathway, and cholesterol biosynthesis.	Oncology Microbiology Virology
2	Don't know	Not stated	Oncology Clinical genetics

Only one participant in this group expressed the negative view that the bipartite graph presented was not usable at all. Of note, one of the participants pointed out the limitation of using GO Process labels for visualizing biomedical microarray data (participant no. 3; see TABLE 3). In his opinion, GO did not describe the relationship between the various biological processes and human diseases and thus had very little use in biomedical research.

Two participants from the block matrix group commented that the redundant GO Process labels confused the functional relationships between the co-expressed gene clusters. One participant in this group found that the block matrix was difficult to use for comparing between sample sets (participant no. 10; see TABLE 3). However, the same participants still thought that the representation in its current form was useful to their microarray analysis work. Another three participants commented that the representation was easy to read and to understand. Collectively, participants' opinion indicated that the block matrix in its current form was acceptable in terms of usability but the clustered bipartite graph required further work.

3.5.7. Discussion

The evaluation results are summarized in FIGURE 3.14. With the competency tasks, the median accuracy (see FIGURE 3.14(c)) and the median user confidence score (see FIGURE 3.14(b)) of the bipartite graph group was comparable with its block matrix counterpart. A comparison between the worst cases showed that participants in the bipartite graph group gave a 50% higher accuracy than those in the block matrix group (see FIGURE 3.14(c)). Also, it took the block matrix group 40% longer to complete the competency tasks than the bipartite graph group (see FIGURE 3.14(a)). Thus the advantage of the clustered bipartite graph representation over its block matrix counterpart lay in faster task completion and, for some participants, reading accuracy.

A task-to-task analysis showed that the previously mentioned advantages of the clustered bipartite graph over the block matrix were evident mainly in tasks D and E (Questions 7 to 10; see FIGURE 3.12(a)). Indeed, it had been observed that an immediate stalling happened when the participants in the block matrix group were looking for a solution to question 7 of task D. This could be due to the redundant representation of the GO Process labels in the normal sample set which made the block matrix confusing to read, a view expressed by a few participants (see TABLE 3). On the other hand, four participants in the bipartite graph group were finger tracing the edges to confirm any perceived relationships between a gene cluster and its neighbouring GO nodes or vice versa. The same behaviour was absent with the block matrix group. The presence of edges seemed to give the participants a visual mean to confirm the presence of any perceived relationships.

TABLE 3. Participants' post-task comments

P	Comments P=particpants
Group: Clustered bipartite graph	
1	Visually easy to look at but the line crossings may be a bit confusing if the number of connections increases. It is a good visualization tool for array data.
2	The fonts of the GO terms are difficult to read. Recommend colour-coded lines and text for each GO term. I would like to have a 3D spherical arrangement of the correlated gene clusters to complement the 2D visualization so that I can interpret the data for tasks I and J.
3	The visualization is helpful because it does give you some idea of gene co-expression in the context of GO terms. However, when the relationship between various GO terms becomes more complex, the appeal of the visualization diminishes. The visualization is not a great deal relevant to my research because GO terms describe a lot of generic processes and bear very little relationship to the actual pathophysiological data described in the literature.
4	Spontaneously, I would say that the GO clusters are not very accessible. It requires a certain amount of dedication to understand them. We use microarrays in several ways to understand gene regulation on the post-transcriptional level. Several projects have come up with gene lists in which GO terms are significantly enriched.
5	The 'normal' diagram was more difficult to read than the 'disease' diagram because of the increase in the number of lines causing congestion in the diagram. It would be good to have a function where clicking on a gene cluster changes the colour of the lines to the GO terms. I haven't used a data visualisation program on my data yet, so this looks good and will be helpful.
6	Summarize the results nicely by showing all the relationships between gene groups and the related GO terms. However when there are a lot of interactions between groups it's a little hard to read the results.
7	The lines were a bit vague when interpreting the correlations but the visualization could be relevant in the future for interpreting data.
Group: Block matrix	
8	These GO clusters are quite easy to understand and analyse. It can get a bit tricky when looking for a gene ontology that comes up not only in one cluster but multiple ones. If the dataset of the normal tissue is bigger than what has been shown, it will be even more difficult to do task D. Yes, it is relevant. I need to use gene ontology annotation in all my microarray experiments. At present, I have to compare the gene ontology terms between two datasets by visually inspecting them on Excel spreadsheets.
9	It was difficult to relate GO terms that were represented more than once in the diagram. The visualization is very relevant. We are doing a lot of microarray work and have large datasets which we would like to relate to functional outcomes.
10	A bit confusing trying to understanding the difference between co-expressed and co-regulated, and being able to compare disease to normal tissue using the cluster patterns. However, it provided a good way to visually inspect the groups of genes regulated in each tissue, as well as between tissue types, I think a spreadsheet would also have to be provided for the comparison between disease and normal.
11	I understood the concepts behind the visualisation graphics. It was quite easy to follow and then draw conclusions from what is being presented.
12	Easy to visualise. Perhaps might even be better if each of the bioprocesses had a different colour code.
13	A few questions were difficult without any experience in microarrays.
14	The graphs were easy enough to understand for a person who has no previous experience with gene arrays.

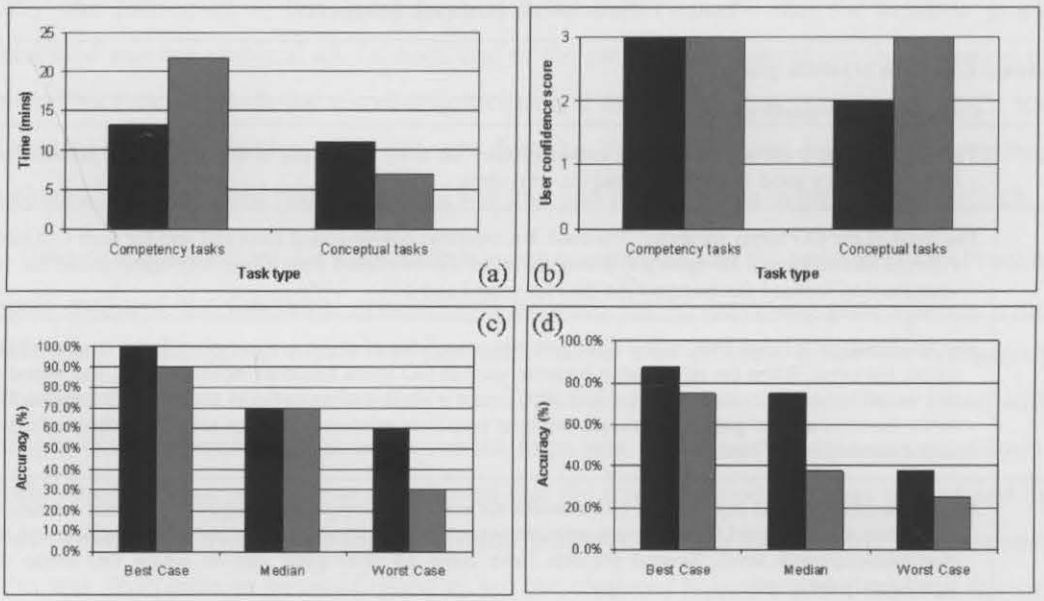


FIGURE 3.14. Summary of evaluation results. The data for the bipartite graph group and its block matrix counterpart are shown in *dark* and *grey* bars respectively. (a) Median completion time for each task type. (b) Median confidence score per task type. (c) Overall median accuracy of competency tasks. (d) Overall median accuracy of conceptual tasks.

However, some participants found that the edge crossings in the normal sample set were interfering with graph reading (see TABLE 3). This might explain why it took the bipartite graph group a longer median completion time on the normal sample set than that on the disease sample set, except for tasks A and D. Edge crossings should have very little impact on tasks A and D since task A did not require edge traversal. With task D, most of the gene clusters containing two genes or fewer were located at the far right end of the screen where the distribution of edge crossings was sparse (see FIGURE 3.6(a)).

With the conceptual tasks, the median accuracy in the bipartite graph group was twice of the block matrix group (see FIGURE 3.14(d)). However, the bipartite graph group took 36% longer to complete the conceptual tasks than the block matrix group (see FIGURE 3.14(d)). This was because several participants in the bipartite graph group seemed to realize that some of the competency and conceptual tasks are related. They were re-examining not only the representation but also their answers made to the competency tasks repeatedly.

The median user confidence score showed that the bipartite graph group had a less positive user experience than the block matrix group (see FIGURE 3.14(b)). This was suggesting a case of *perception/performance mismatch*. In other words, participants in the bipartite graph group did not realize that they were giving a higher number of correct responses than its block matrix counterpart. A task-to-task analysis revealed that perception/performance mismatch occurred not only with tasks G and H, but also with the competency tasks B and D. More importantly, it occurred only with questions that were

based on the normal sample set suggesting that visual complexity could be the cause. Taken together, the advantage of the bipartite graph over block matrix laid in enhancing the biologist's analytical accuracy but in its current form was perceptually less usable than the block matrix. The participants' post-task comments reflected this.

Another important observation was that the number of correct responses for task F was comparable to that of task C, and the number of correct responses for task G was comparable to that of task D (questions 7 and 8). Because tasks F and G were designed to complement tasks C and D respectively, the participant's correctness in answering the conceptual tasks should depend on his/her correctness in answering the complementary competency tasks. Therefore, the results indicated that reading accuracy could influence the biologist's analytical accuracy.

Of all the conceptual tasks, tasks I and J were the most challenging. The poor number of correct responses and the poor user confidence score given by both groups were a reflection of this (see FIGURES 3.14(b) and (c)). Both tasks challenged the participants to make deductions based on their expertise in biology. This might suggest that the GO Process labels presented were not informative enough for the participants to draw an accurate conclusion to either task. During the performance of task I, several participants in the block matrix group had verbally expressed that the co-existence or the absence of certain GO Process labels was in conflict with their knowledge on liver physiology. However, the same did not occur with the bipartite graph group. This difference could be a result of the different layout of GO Process labels in each representation. The display of GO Process labels beneath each cluster in the block matrix might give the participants a stronger impression that the biological processes of a particular cluster were functionally related as compared to the single level layout of GO Process labels in the clustered bipartite graph.

The conflict between what was being perceived and what was the participant's own knowledge precept could be happening to all the participants, because the answers given seemed to be deduced from a selected few rather than the entire set of GO Process labels (see TABLES 2 and 3). It was possible that the participants were exhibiting *cognitive bias* towards the biological processes that they were most familiar with. In the best case, one participant based his deduction for task I on four GO Process labels and another participant based his deduction for task J on three GO Process labels. This would suggest that a systems-level understanding of HCC was achievable only if the dataset has been cross examined by biologists from different areas of expertise.

3.6. Remarks

This chapter exposed the different strength and limitations of the two visual representations, i.e. the block matrix and the clustered bipartite graph, as visual analysis methods. The strength of the block matrix representation laid in its visual simplicity and its gene centric semantics. However, this apparent advantage over the clustered bipartite graph did not translate into real performance enhancement in either the task completion time or in analytical accuracy. The underlying reason could be its redundant representation of the GO Process labels. Readability was further compromised when the redundant GO Process labels were scattered throughout the visualization. In its current form, block matrix was only suitable for pairwise comparison between clusters for their functional differences.

The strength of the clustered bipartite graph representation laid in its graphical semantics which emphasized on connectivity between two sets of nodes. It was a faithful representation of the *m*:*n* gene_cluster-GO relationship. The better performance of the clustered bipartite graph group in reading and analytical accuracy might imply that capturing the network view was more relevant to biologists than preserving their gene-centric view using the block matrix, even though the usability of the former representation decreased with the increase in edge crossings. This finding was especially relevant in the present day when experimental biologists increasingly needed to adopt the network view in order to make a better use of high-throughput data for hypotheses deduction.

Based on the evaluation results, several design requirements for visualizing GO-annotated gene clusters could be recommended. Listed in their order of priority are:

1. The *m*:*n* gene_cluster-GO relationship has to be faithfully represented. The usability evaluation showed that the biologist's analytical reasoning cannot be enhanced by emphasizing the functional partitioning of genes. The representation of the *m*:*n* gene_cluster-GO relationship is of equal importance if not more.
2. The representation of GO Process labels in the block matrix has to be non-redundant. This will allow the biologist to quickly deduce the biological differences between phenotypes.
3. Edge crossing minimization in the clustered bipartite graph representation is necessary for enhancing graph readability.
4. It may be important to represent the co-expressed genes explicitly to preserve the biologist's mental model of a gene cluster even though this will increase the ink-to-information ratio.

Most participants indicated that the visualization was relevant to their research despite that only a few were able to deduce correctly the tissue of origin from which the dataset was derived and the pathological condition of the tissue based solely on the visual representations.

Finally, the usability evaluation exposed two challenges facing today's biologists when analyzing high-throughput data. The first one was *cognitive bias*. The biologist would tend to deduce hypothesis based on those GO Process terms that he/she was familiar with. The most appropriate solution for alleviating this problem was to increase collaboration within the biological research community. The second one was the informational scope of GO. It is a controlled vocabulary for representing what biologists have historically studied about the functions of certain genes. The knowledge about these genes can be biased by the fact that their functionality has only been studied by experts within a particular domain, say embryonic development, and not in any other domains, e.g. metabolic diseases. An added concern is that computational prediction of gene function based on comparative genomics could result in misleading GO annotations. The basis of comparative genomics is specie orthogonality. For example, gene *a* found in mouse should have a similar function to gene *A* in human since their 70% or more of their DNA sequences are identical. They are known as orthologs. By this rationale, it should be able to predict the function of an unknown human gene if we know its mouse ortholog. However, the function of a gene is defined by what other genes or proteins it interacts with and its position in the gene regulatory and the protein-protein interaction networks. These could vary from specie to specie. Therefore GO terms sourced from computational predictions are highly hypothetical. Their use in GO cluster visualizations should be restricted. In the light of these, when investigating biomedical questions, GO annotation on high-throughput data might not be informative enough for the purpose of hypothesis deduction and the use of some other biomedical ontologies such as the OMIM Morbid Map [136] or NCI Thesaurus [136] might be necessary.

{End of Chapter 3}

CHAPTER 4

Visualization and Analysis of Gene Ontology-Defined Protein Interaction Networks

"No proteins, No work"

4.1. Introduction

In the previous chapter, the visualization of Gene Ontology (GO)-annotated gene clusters showed the modular organization of gene expression [3]. This modularity is defined by the biological processes in which the co-expressed genes are involved in. A functional module can be understood as a network path or a sub-network of the single cell molecular network [116]. Such a module has a defined biological function which is comprised of one or more biological processes. Based on this understanding, we seek to explore molecular networks as the means for biological research in the coming chapters. Since co-regulated biological processes require protein-protein interactions to function, visualizing gene co-expression in the context of a protein interaction network (PIN) becomes the second step of our visual analysis framework (see Chapter 1, section 1.2).

In this chapter, we explore the problem of PIN visualization. The challenges to this problem are two-fold. The first one is *scale*. The single cell PIN of a particular organism is cognitively challenging when visualized. For example, the latest version of the human PIN contains approximately 25,000 proteins and the number of interactions could exceed 40,000 [15]. The second one is to capture the *functional modularity* of PIN. As mentioned in Chapter 3 (see section 3.1), proteins belonging to a functional module can be defined by their membership in a cellular component or in a biological process. Furthermore, a PIN corresponding to a biological process can itself be subdivided into smaller modules (or sub-networks) if its proteins co-exist in multiple cellular components. In other words, if we extract a sub-network as a functional module out of the single cell PIN, that sub-network can be made of smaller sub-networks or functional modules. This property of '*network within network*' or '*module within module*' is universal to all molecular networks [3].

In an attempt to meet these challenges, we present two methods for visualizing the human PIN, i.e. the non-clustered PIN and the clustered PIN visualizations. We implemented them as part of a visualization system that allows the biologist to select the criterion, i.e. GO Biological Process (also known as the GO Process) or the GO Cellular Component (also

known as the GO Component), for filtering the single cell human protein network to a smaller network. Each network contains only the protein-protein interactions that correspond to the selected GO Process or GO Component. We called this network a *GO-defined PIN*. If it is a result of using the GO Process as the filtering criterion, we called the network a *GO_Process-defined PIN*. If it is a result of using the GO Component as the filtering criterion, we called the network a *GO_Component-defined PIN*.

A GO-defined PIN can be visualized using two methods. In the non-clustered visualization, the PIN is visualized using the force-directed layout which has been the conventional method [44, 52]. This is a generic layout which does not take any biological context into account. In the clustered visualization, the PIN is visualized as a set of inter-connected clusters using a circular layout. The clustering criterion applied is a GO category complementary to that for the filtering criterion. For example, if the GO Process is applied as a filtering criterion, the GO Component will be applied as a clustering criterion. The resulting visualization will consist of clusters labeled with GO Component terms. This should capture the functional modularity of the GO_Process-defined PIN.

The use of filtering or clustering by domain knowledge has been used in PIN visualization before. For example, in ProViz, the user can filter the entire PIN by GO categories or other ontologies [178]. In PATIKAweb [42], a PIN is being drawn as a compound graph using a force-directed layout. Each sub-network is visually confined in a rectangular partition representing a biological process. In turn, each partition is being superimposed on a grid where the partitions represent various cellular compartments. This approach is in effect a kind of clustering and is closest to the ideal of exposing the nested modularity of PIN. The limitation is that PATIKAweb requires pre-defined pathway data as the input. It does not provide network filtering like ProViz.

Our approach is different in a way that we applied the '*filter-and-cluster*' combination with the specific aim of not only exposing the nested modularity of a PIN [116] but also preserving the biologist's analytical approach of "*filter first, zoom and details, overview if necessary*". This has been the motivation behind using one GO category as a filtering criterion and another as a clustering criterion. The filtering step allows the user to pre-select a particular functional module whereas the clustering step further exposes its modular structure.

To evaluate the merits of using the non-clustered and the clustered PIN visualizations as visual analysis methods, we employed the hepatocellular carcinoma (HCC) gene expression dataset [58] as the input for each PIN visualization in our case study. The objective is to identify which biological processes are the most affected in HCC. In the visual analysis, we also addressed some of the HCC-specific biological processes found in the previous chapter.

To evaluate if the visualizations meet the biologist's expectation in terms of usability, we conducted a domain expert evaluation with an expert biologist who has research interests in proteomics and bioinformatics. We asked him to evaluate the visualizations based on a set of evaluation criteria derived from published pathway visualization heuristics [134].

The rest of this chapter is divided into five sections. The graph-theoretic models of three types of PINs are defined in section 4.2. The drawing algorithms for the non-clustered PIN and clustered PIN visualizations are presented in section 4.3. The HCC case study is introduced in section 4.4. This is the most important section since the strength and limitations of each visualization are tested here. Furthermore, some of the protein-protein interactions and biological processes mentioned here later become the subjects for further analysis using the two- and three-overlapping networks (see Chapters 5 and 6). The domain expert evaluation conducted with an expert biologist is elaborated in section 4.5. This includes the evaluation criteria, the biologist's background, and the evaluation results. Finally, the advantage of applying two complementary visualizations on the same PIN is discussed in section 4.6 as a conclusion to this chapter.

4.2. Representation of Protein Interaction Network

4.2.1. General Protein Interaction Network

The graph theoretic model of a PIN is an undirected graph in which the node set represents the proteins and the edge set represents the physical interactions between proteins. The graph-theoretic definition of the PIN is defined as the following:

Definition 4.1. *A protein interaction network is an undirected network* $G_P = (V_P, E_P)$ *in which* V_P *denotes the node set of proteins and* E_P *denotes the edge set of protein-protein interactions. The edge* $e = (v_1, v_2)$ *represents the pairwise interaction between two proteins* v_1 *and* v_2 *where* $e \in E_P$, $v_1 \in V_P$ *and* $v_2 \in V_P$.

4.2.2. Gene Ontology-defined Protein Interaction Network

Because the single cell PIN often exceeds 10,000 nodes, there is a need to filter it to a smaller scale. Furthermore, the single cell PIN does not have any biological context in itself. To provide biological context to the filtered PIN, ontology identifiers from either the GO Process category or the GO Component category are used as the filtering criterion. The GO Process category is often used as the abstraction for biological processes. The GO Component category is used as the abstraction for cellular components [60]. It should be noted that the GO Component category provides ontology at four different levels of abstraction. At the highest level, the GO Component category provides ontologies for describing sub-cellular regions, e.g. intracellular (GO:0005622). At the middle level, it provides ontologies for describing organelles, e.g. nucleus (GO:0005634). The next level is

the ontologies for describing sub-organelles, e.g. nucleolus (GO:0005730). At the lowest level, it provides the ontologies for protein complexes, e.g. ribosome (GO:0005840).

The graph-theoretic definition of the GO-defined PIN is defined as the following:

Definition 4.2. *A GO-defined PIN is an undirected network* $G_F = (V_F, E_F)$ *where* $V_F \subseteq V_P$, $E_F \subseteq E_P$. *Each* $v_f \in V_F$ *has two node attributes, i.e.* `BP_ID` *and* `CC_ID`. *Each attribute contains a tuple of unique GO identifiers.* `BP_ID` *contains identifiers for the GO Process terms.* `CC_ID` *contains identifiers for the GO Component terms.*

If `BP_ID` is used as the filtering criterion, all nodes in V_F will have the same GO identifier for the node attribute `BP_ID`. If `CC_ID` is used as the filtering criterion, all nodes in V_F will have the same GO identifier for the node attribute `CC_ID`.

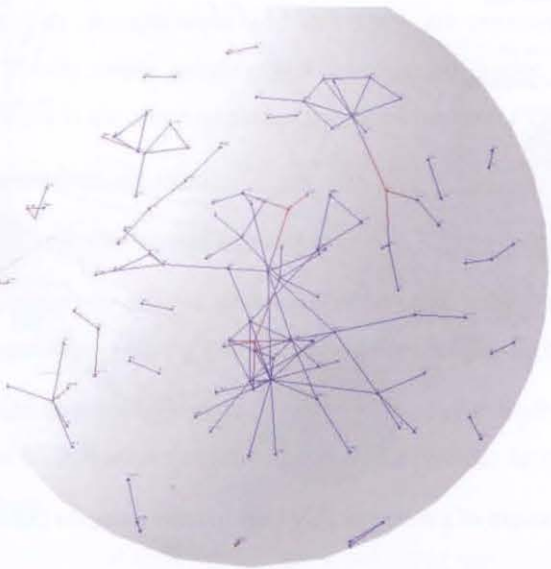


FIGURE 4.1. An example of a non-clustered PIN visualization using the force-directed layout.

4.2.3. Clustered Protein Interaction Network

After filtering, the node set V_F can be grouped to multiple clusters using the GO category that is complementary to the one used for filtering. For example, if the GO Process is applied as a filtering criterion, the GO Component will be applied as a clustering criterion. The graph-theoretic definition of a clustered PIN is defined as the following:

Definition 4.3. *A clustered PIN is an undirected network* G_C *containing a cluster set C of subgraphs, i.e.* $C = \{G_1, G_2, \ldots, G_k\}$ *where* $G_k = (V_k, E_k)$. V_k *denotes the node set of proteins.* E_k *denotes the intra-cluster edge set of protein-protein interactions. Furthermore,* E_C *denotes the inter-cluster edge set of protein-protein interactions. Given two subgraphs* $G_i = (V_i, E_i)$ *and* $G_j = (V_j, E_j)$, *where* $i \neq j$, *if there are edges between* V_i *and* V_j, *then there are edges between* G_i *and* G_j.

If `CC_ID` is used as the clustering criterion, all nodes in a given subgraph G_k will share the same GO identifier for `CC_ID`, and similarly for `BP_ID`. It should be noted that some subgraphs can share some common nodes and edges. Given two subgraphs G_i and G_j where $i \neq j$, their intersection is non-empty, i.e. $G_i \cap G_j \neq 0$. Therefore, proteins that belong to multiple clusters are redundantly represented in the clustered PIN visualization in order to avoid overlapping clusters.

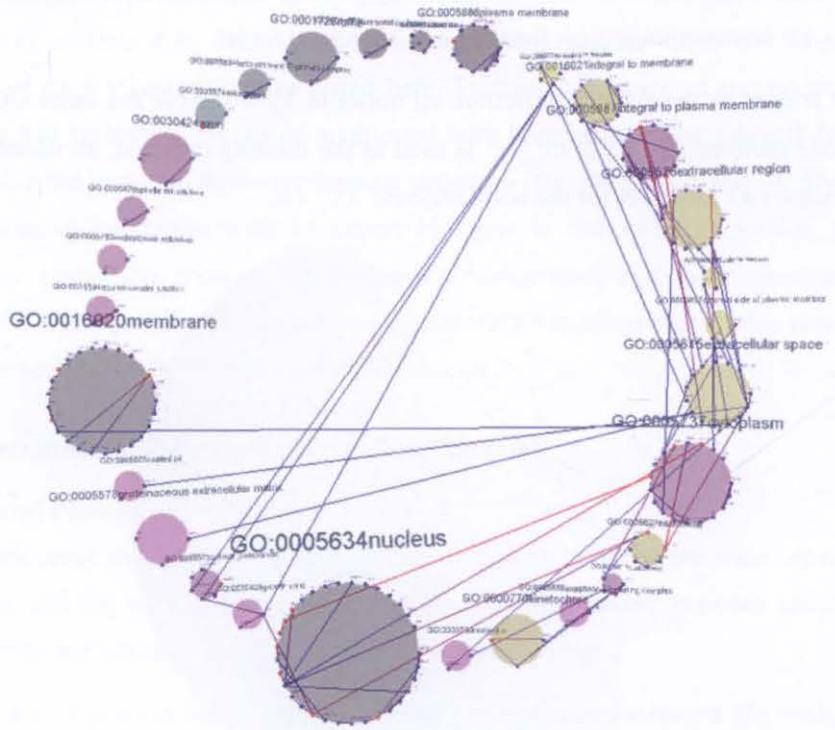


FIGURE 4.2. An example of a clustered PIN visualization using the clustered circular layout.

4.3. Visualization of Gene Ontology-defined Protein Interaction Network

4.3.1. Non-Clustered PIN Visualization

The non-clustered PIN visualization is an undirected network drawn in the force-directed layout [44]. The design criteria for this visualization are (1) to display the interaction between proteins, and (2) to highlight the proteins of co-expressed genes.

The second criterion is to support the rationale that a pair of proteins should be similar in their molecular abundance if they are co-expressed. If they are interacting neighbours at the same time, their protein-protein interaction is likely to be functional. The distance between each pair of nodes is determined by the equilibrium between the spring and the repulsive forces.

Each protein is represented by a spherical node. To satisfy the second criterion, nodes that represent co-expressed proteins are coloured red. Nodes representing proteins that are not co-

expressing are colored blue (see FIGURE 4.1). Edges are represented by straight lines. An appropriate color gradient is being applied to each edge with its ends sharing the colours of the end nodes. For example, an edge between an expressed protein and a protein that did not co-express will be half in red and another half in blue.

4.3.2. Clustered PIN Visualization

The design criteria of this layout are (1) to display the modular structure of the GO-defined PIN with each module defined by a GO Process or a GO Component, (2) to provide a fixed layout of the nodes (protein and cluster nodes), (3) to minimize intra- and inter-cluster edge crossings, and (4) to highlight the proteins of co-expressed genes.

Each cluster is represented by a circular node and is coloured with a different hue to differentiate between clusters of individual GO terms. Within each cluster node, the protein nodes are arranged along its circumference and the edges are represented by straight lines (see FIGURE 4.2). The cluster nodes are arranged in a circular layout. The colour coding of the protein nodes and edges is the same as that in the non-clustered PIN.

The drawing algorithm involves six steps.

Algorithm 4.1. Clustered circular layout algorithm

1. Compute the ordering of the cluster nodes such that the inter-cluster edge crossing is minimized. This is achieved with the use of the circular shifting algorithm [10].
2. Compute the Cartesian coordinates of each cluster node. For the i-th cluster node c_i, it has the polar coordinates $(r, \Delta\theta)$ where r is the radius of the circular layout and $\Delta\theta$ is the polar angle in radians. Hence,

$$r = \frac{S}{2\pi}$$

$$\Delta\theta \approx \frac{d_i}{S} \times 2\pi$$

$$S \approx \varepsilon + \sum_{i=0}^{n} d_i$$

where ε is the spacing factor, S is the circumference of the circular layout, and d_i is the diameter of c_i.

The Cartesian coordinates (x_i, y_i) for c_i can be computed as:

$$x_i = x_0 + r\cos(\Delta\theta)$$
$$y_i = y_0 + r\sin(\Delta\theta)$$

where (x_0, y_0) is the centre of the drawing area.

3. Compute the ordering of the protein nodes in each cluster such that the intra-cluster edge crossing is minimized. This is achieved with the use of the circular shifting algorithm [10].

4. Compute the Cartesian co-ordinates of the protein nodes within each cluster. This step is similar to step 2. The difference is that the $\Delta\theta$ of the polar co-ordinates $(r, \Delta\theta)$ for each member node is constant. The Cartesian coordinates for each protein nodes are computed relative to the centre coordinates of the cluster node.

5. Draw the cluster nodes in the circular layout.

6. Draw the protein nodes within each cluster node in the circular layout.

7. Draw all the edges.

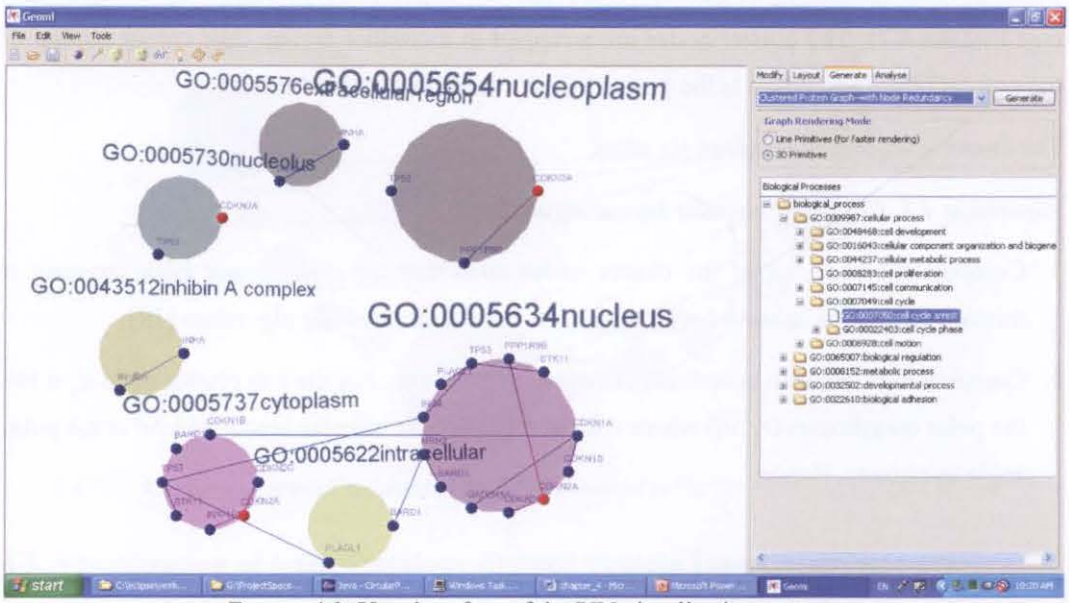


FIGURE 4.3. User interface of the PIN visualization system.

4.3.3. Implementation

In order to generate the user interface for selecting individual GO terms under the GO Process or GO Component category, a new plug-in is added to the network visualization and analysis tool GEOMI [2]. A right-hand panel is added that contains a drop-down menu for selecting either the GO Process and GO Component category. Beneath the drop-down menu box is a tree menu for selecting individual GO terms (see FIGURE 4.3). The tree menu also reveals the parent-child relationship between GO terms. The algorithms for generating the PIN visualizations are implemented using the Java3D library as new plug-ins to GEOMI.

4.4. Case Study: Proteomics of Hepatocellular Carcinoma

To generate insights into the biological implications of gene co-expression, in the context of PIN visualizations, we overlaid the HCC dataset used in the previous chapter onto the human PIN. We then investigated PINs that were filtered by selected GO Process terms. The objective was to identify the biological processes that are the most affected in HCC cells.

4.4.1. Network Construction

4.4.1.1. Datasets

Human protein interaction data. The human protein interaction data was collected from the BIOGRID download version 2[1].0.20 [15]. Many of the protein interactions in the BIOGRID data had been verified by more than one laboratory technique. Hence it was more reliable than data generated solely by the yeast two- hybrid method.

Gene expression data. The gene expression data for HCC was identical to that used in Chapter 3 section 3.4.1. Only the list of co-expressed genes found in the 95 HCC samples provided by Gamberoni *et al.* [58] is being applied to the PIN visualizations.

Gene Ontology. The three categories of GO—Component, Process, and Function, were obtained from the Gene Ontology Consortium [60].

4.4.1.2. Data mapping

The human protein interaction data was downloaded from the BIOGRID database as a tab-delimited file. Each record contained a pair of gene symbols representing the interacting protein partners. Each gene symbol was mapped to a node. If a pair of gene symbols belonged to the same record, they were mapped to the nodes of an edge. For the clustered PIN visualization, a GO term was mapped to each cluster node.

4.4.2. Visualization and Analysis

In order to ensure that the biological processes selected for visual analysis was of relevance to our subject of study, i.e. HCC, we used the latest understanding in cancer biology as the conceptual guide. According to Hanahan and Weinberg [68], cancer cells have six characteristics. These are (1) self-sufficiency in growth signals, (2) insensitivity to anti-growth signals, (3) evasion of apoptosis, (4) limitless replicative potential, (5) sustained angiogenesis, and (6) tissue invasion and metastasis.

We then selected the biological processes that were thought to give rise to each characteristic. We performed visual analysis on each GO_Process-defined PIN in a way that was close to the biologist's preferred practise. Given a non-clustered PIN visualization, biologists were likely to perform the following tasks.

1. Identify co-expressed proteins based on node colour. If two red nodes co-exist in the same PIN, they represent co-expressed proteins and are likely to be co-regulated. Blue nodes represent proteins that are not co-expressed. It is possible that some of these proteins did not co-express because their level of expression detected are below the sensitivity of the microarrays used by Chen *et al.* [29].

2. Identify protein-protein interactions between co-expressed proteins. If two red coloured nodes are connected to each other with a red coloured edge, it implies that the protein-protein interaction is functioning.

3. Identify unique node topologies in the PIN. This step is also optional. It may be used by systems biologists who make use of node topologies to deduce the molecular organization of a protein complex, and the probable gene expression dynamics of proteins that have unique topologies. For example, biologists often call the proteins that form a highly connected sub-network as *party hubs* and the protein that form the centre of a star shape sub-network as *date hubs* [66]. A protein that connects the two hubs is called a *bottleneck* protein [168]. These three types of proteins have been known to exhibit different interaction dynamics. The date hub interacts with its neighbours dynamically whereas the party hub tends to interact with theirs for a longer time period [66].

Given a clustered PIN visualization, biologists were likely to perform the following tasks:

1. Identify the cellular distribution of co-expressed proteins. This step involves identifying red coloured nodes in different cluster nodes. Each cluster node has a GO Component label representing a sub-cellular component.

2. Identify protein-protein interactions between co-expressed proteins that are distributed in different clusters. This step involves identifying inter-cluster edges that are coloured red. It is most likely performed by systems biologists as a follow up to task 3 in the non-clustered PIN analysis. In other words, after using node topologies to infer the molecular organization of a protein complex, they will use clustered PIN visualization to identify the subunits of the protein complex and the probable functionality of each subunit.

We divided our case study into four sections. In the first three sections, we described in each section the visual experimentation and analysis results of one or more GO_Process-defined PINs that were thought to give rise to three of the six characteristics of HCC. They were evasion of apoptosis (section 4.4.2.1), self-sufficiency in growth signal (section 4.4.2.2), and limitless replicative potential (section 4.4.2.3). In the fourth section, we described the GO_Process-defined PINs that were shared by the last three characteristics of HCC, i.e.

sustained angiogenesis, tissue invasion and metastasis (section 4.4.2.4). Where available, the Gene Ontology [60] identifier was given in parentheses for every biological process mentioned. Similarly, the Entrez Gene [99] identifier was given in parenthesis for every human gene mentioned.

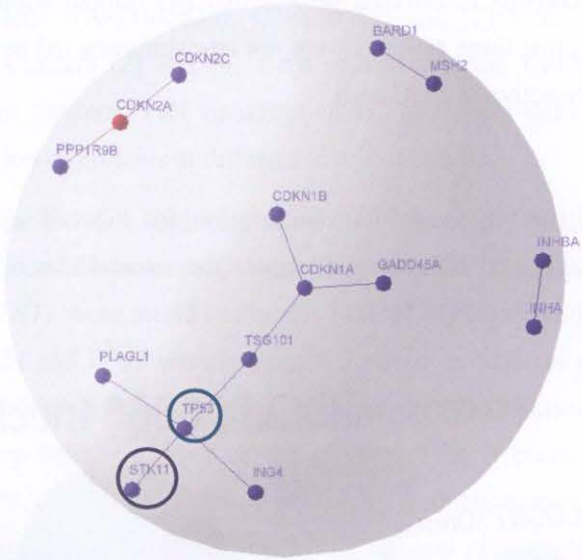


FIGURE 4.4. Non-clustered PIN visualization of the cell cycle arrest (GO:0007050) biological process in the force-directed layout. *TP53* and *STK11* which are discussed in section 4.4.2.1 are circled in green and blue respectively.

4.4.2.1. Evasion of apoptosis

I. Non-clustered PIN

Cell cycle arrest (GO:0007050) was the biological process relating to the evasion of apoptosis and had been found to be HCC-specific in Chapter 3. FIGURE 4.4 showed the non-clustered PIN visualization for the cell cycle arrest biological process. It consisted of 15 proteins and 11 interactions in four connected components. The largest connected component consisted of 8 proteins and 7 interactions.

We observed that the only node coloured red is the one labeled *CDKN2A* indicating that it was the only expressed protein in the cell cycle arrest PIN. The rest of the nodes were coloured blue indicating that they did not co-express with *CDKN2A*. This implies that there were no functioning protein-protein interactions in the cell cycle arrest (GO:0007050) biological process. There was a complete loss of functional protein interactions that could initiate cell cycle arrest. We further examined the blue coloured nodes to see if we could deduce the likely cause of this complete shutdown of cell cycle arrest.

First we noticed that one neighbour of node *CDKN2A* was the node labelled *CDKN2C*. Both proteins belong to the family of cyclin-dependent kinase inhibitors. Their known function is to arrest cell cycle progression [157]. In the largest connected component, the

node labeled *TP53* (GeneID: 7157) had the highest degree and was also a tumour suppressor protein well-studied by cancer biologists. It had been known to be involved not only in cell cycle arrest but also in the initiation of apoptosis [129]. Apoptosis (GO:0006915) is a biological process that leads to cell death. One neighbour of node *TP53* was the node labeled *STK11*. This was a recently discovered protein that has tumour suppressing functionality [72]. Our observation that these proteins were not co-expressing led us to the deduction that apoptosis might not be active in HCC cells.

II. Clustered PIN

FIGURE 4.5 showed the clustered PIN visualization for the cell cycle arrest biological process. We used this visualization to investigate the intracellular distribution of the cell cycle arrest proteins (see FIGURE 4.5(b)).

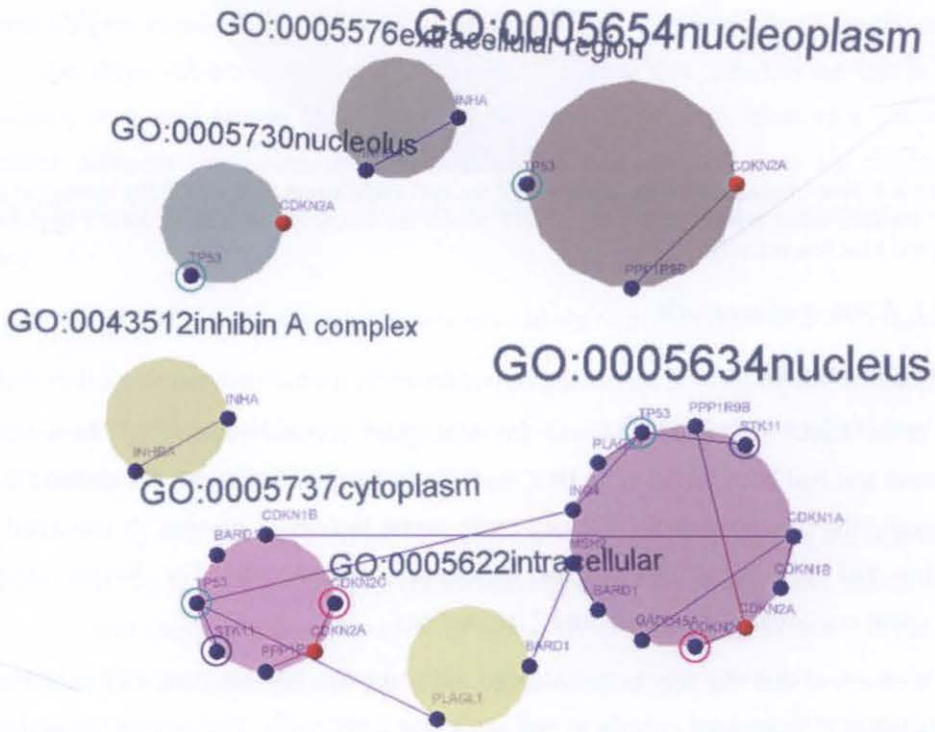


FIGURE 4.5. Clustered PIN visualization of the cell cycle arrest (GO:0007050) biological process in the clustered circular layout. *CDKN2C*, *TP53*, and *STK11* which were discussed in section 4.4.2.1 are circled in magenta, turquiose and blue respectively.

The visualization consisted of 41 nodes and 37 edges in seven clusters. The cluster nodes were arranged in a circular layout and the protein nodes within each cluster node were also arranged in a circular layout. The seven cluster nodes were labeled with GO Component terms with one label per cluster. Each cluster therefore represented a sub-cellular component. The seven GO Component labels were: (1) '*GO:0005634 nucleus*', (2) '*GO:0005622 intracellular*', (3) '*GO:0005737 cytoplasm*', (4) '*GO:0043512 inhibin A complex*', (5)

'*GO:0005730 nucleolus*', (6) '*GO:0005576 extracellular region*', and (7) '*GO:0005654 nucleoplasm*'.

Out of the above, clusters (1), (3), and (5) were organelles. An organelle was an intracellular cell compartment that performed specific cellular functions [83]. The largest cluster was the nucleus (GO:0005634). Clusters (5) and (7) were compartments of the nucleus (cluster 1). Clusters (2) and (6) were cellular regions. Cluster (4) was a protein complex. As such, the clustered PIN visualization in FIGURE 4.5 did effectively present the intracellular distribution of proteins at different levels of details.

We searched for the proteins identified in our non-clustered PIN analysis and found that *CDKN2A* and *TP53* were the most ubiquitous. They were found in clusters (1), (3), (5) and (7). *CDKN2C* and *STK11* were found in clusters (1) and (3) (see FIGURE 4.5). We therefore deduced that *CDKN2A* and *TP53* were physically located in different organelles. Biologists found that many proteins involved in cell cycle arrest are functionally inactive in the cytoplasm unless they are re-localized in the nucleus. That is because they interact only within the nucleus [50, 152, 156]. Combined with our deduction gained from the clustered PIN visualization, we proposed that, maybe in HCC, many proteins involved in cell cycle arrest were abnormally sequestrated in the cytoplasm rather than in the nucleus. If this happened, it would further cripple their ability to arrest cell cycle progression in HCC. This hypothesis is yet to be verified by biologists.

III. Summary

For this analysis on the cell cycle arrest-defined PIN, we found that the non-clustered PIN visualization was able to support the following analytical tasks.

1. Identify co-expressed proteins based on node colour.
2. Identify protein-protein interactions that are likely to be functioning.

In the second analytical task, we used the current biological knowledge about *TP53* in combination with visual analysis to deduce that apoptosis (GO:0006915) could also be affected, even though what had been visualized was the cell cycle arrest-defined PIN. This was possible because apoptosis and cell cycle arrest are inter-connected biological processes. They have been known to share a common subset of protein-protein interactions [129]. Based on the non-clustered PIN visualization of a certain biological process, it was possible to deduce hypothesis on its related biological process.

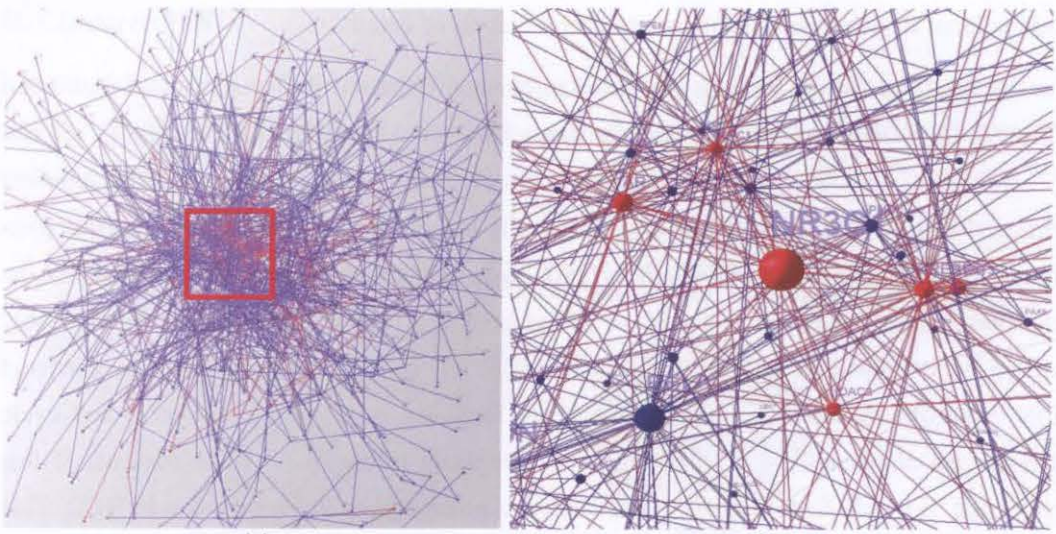
The clustered PIN visualization allowed us to identify the cellular distribution of co-expressed proteins. When interpreted in the context of the current biological knowledge, we were able to deduce a biologically meaningful hypothesis.

4.4.2.2. Self sufficiency in growth signals

Self sufficiency in growth signals involved at least two biological processes, i.e., the regulation of transcription, DNA-dependent (GO:0006355) and signal transduction (GO: 0007165).

1. Regulation of transcription, DNA-dependent (GO:0006355)

I. Non-clustered PIN



(a) (b)

FIGURE 4.6. Non-clustered PIN visualization of the regulation of transcription, DNA-dependent (GO:0006355) biological process in the force directed layout. (a) Overview. (b) Zoom-in view of the bound area.

FIGURE 4.6(a) showed the non-clustered PIN visualization for the regulation of transcription biological process. It consisted of 577 proteins and 1211 interactions. At this scale, the force-directed layout algorithm generated the 'hair ball' effect typically seen in many PIN visualizations [148]. The aggregation of red-coloured nodes at the high density centre of the PIN became our visual focus (FIGURE 4.6(a)).

We zoomed into this region and identified a set of red-coloured high degree nodes (node degree > 20) that represented a group of six high degree protein hubs. They were labeled *AR*, *CREBBP*, *HDAC1*, *HDAC3*, *NR3C1*, and *STAT3*. (FIGURE 4.6(b)). These nodes were coloured red indicating that they co-expressed in HCC. We traversed only the red coloured edges originating from each of these hubs and found that they were inter-connected to one another. That meant they physically interact with one another. We deduced from their node degrees and connectivity that these hub proteins could be the *kernel* of the regulation of transcription (GO:0006355) biological process. A biological *kernel* was a set of master proteins which 'on/off' expression states collectively influenced the states of all other proteins, thereby controlling the activity level of one or even multiple biological processes.

In order to make biologically meaningful deductions, we first examined the known biological function of each hub protein described in the biological literature. *AR* and *NR3C1* were signaling receptors that could also function as gene regulators. *AR* (androgen receptor; GeneID: 367) is a nuclear receptor for the androgen receptor signaling pathway (GO:0030521). It had been known to induce the transcription of androgen responsive genes. Up-regulation of *AR* had been known to associate with poor prognosis in prostate cancer [46] but its role in HCC was unknown.

NR3C1 (GeneID: 2908) is a receptor for glucocorticoids that can act as a gene regulator. Some of its target genes are also gene regulators. This protein is typically found in the cytoplasm until it binds a ligand, which induces transport into the nucleus. Its role in cancer biology was unknown.

CREBBP (GeneID: 1387) is a master co-activator which had been known to interact with more than 50 proteins of different biological functions. Some examples are the (1) hepatic master gene regulators *ONECUT1*, *HNF1A*, and *HNF4A*, (2) tumour suppressors *KLF4*, *BRCA1* and *TP53*, (3) nuclear receptors *AR* and *NR3C1*, (4) proto-oncogenes *E2A*, *FOS* and *JUN*, (5) the cytokine-induced signal transducers *STAT1*, *STAT2*, and *STAT3*, and (6) the cAMP-regulated enhancer specific gene regulator *CREB1*. It also interacts with the cyclin *CCND1* and the cyclin-dependent inhibitor *CDKN1A*, and the signal transducers, *SMAD2* and *SMAD3*, for the *TGFB1* signaling pathway.

HDAC1 (GeneID: 3065) and *HDAC3* (GeneID: 8841) are members of the histone deacetylase complex. Histone deacetylation catalyzed by this complex could enhance cancer survival [139]. The histone deacetylase complex was also known to inhibit *TP53* and its negative effect on cell growth.

STAT3 (GeneID: 6774) is a gene regulator which expression has been known to be active in response to cytokines and growth factors, such as *IFN*s, *EGF*, *IL5*, *IL6*, *HGF*, *LIF* and *BMP2*. It had been known to involve in many cellular processes such as cell growth and apoptosis.

With their known biological function in mind, we attempted to identify all the co-expressed proteins that were neighbours to each of the above protein hubs in FIGURE 4.6(b). However, our effort was hampered by the time-consuming navigation through the large force-directed layout using pointer-directed panning. Edge traversal was disrupted frequently by edge crossings. We only managed to identify three red coloured nodes labeled *CCND1*, *JUN*, and *SMAD2* that are connected to the node *CREBBP*. That implies that they were co-expressing neighbours of *CREBBP*.

Given the difficulty in visual exploration, we limited our deduction to our previous observation that the six hub proteins were connected among themselves in the context of their known biological functions. Half of the hubs seen in FIGURE 4.6(b) that made up the kernel were proteins with the dual functionality of signal transduction and gene regulation, i.e. *AR*, *NR3C1*, and *STAT3*. Their expression could be activated by growth factors. The other half consisted of proteins that initiate gene regulation, i.e. *CREBBP*, *HDAC1* and *HDAC3*. We suspected that in HCC cells, the kernel was in a permanent 'on' state in order to confer increased sensitivity towards growth factors. This could be one mechanism that led to self-sufficiency in growth signals in HCC cells.

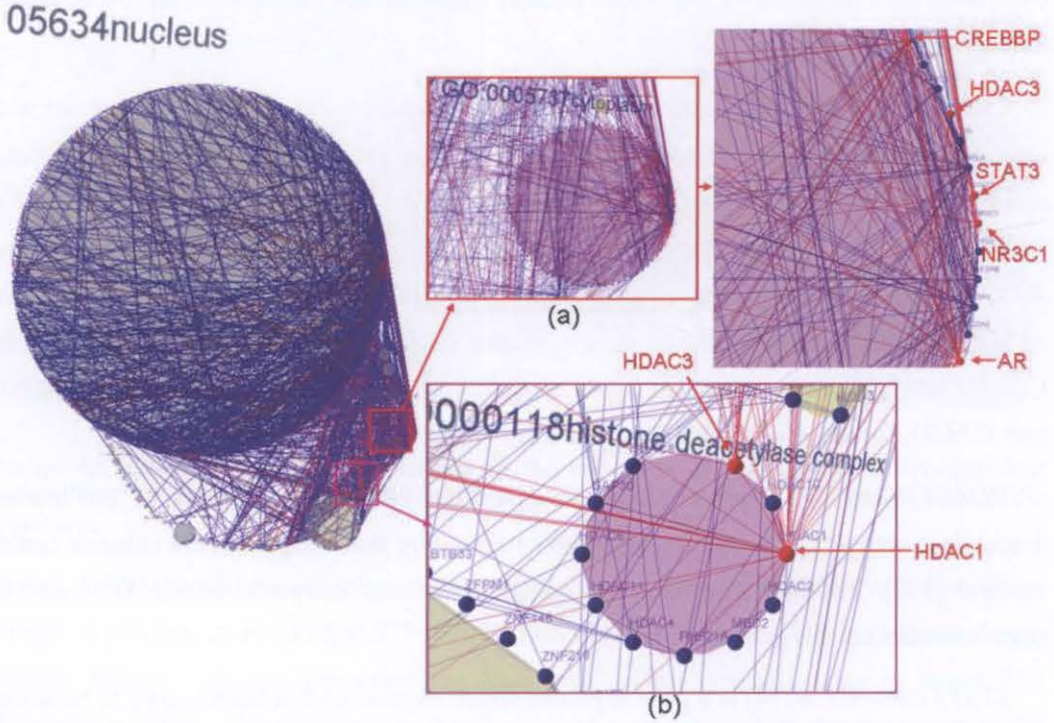


FIGURE 4.7. Clustered PIN visualization of the regulation of transcription, DNA-dependent (GO:0006355) biological process in the clustered circular layout. The overview of the clustered PIN is shown on the left. (a) A zoom in view of the high-density region in the '*GO:0005737 cytoplasm*' cluster (red bounded area). (b) A zoom in view of the high-density region in the '*GO:0000118 histone deacetylase complex*' cluster (magenta bounded area).

II. Clustered PIN

FIGURE 4.7 showed the clustered PIN visualization for the regulation of transcription (GO:0006355) biological process. The redundant representation of the same nodes and edges in different clusters had inflated the network size by 3.5 times. The resulting visualization consists of 2019 proteins and 4039 interactions in 67 clusters. The high number of node and edge crossings made the visualization very cluttered in especially the intra-cluster edge cluttering on the right hand side of the visualization.

The first feature we noticed from the visualization is that the largest cluster node had the GO Component label '*GO:0005634 nucleus*'. It contained most of the intra-cluster edges. From this observation, we deduced that most protein-protein interactions for the regulation of transcription (GO:0006355) biological process occurred exclusively within the cell nucleus (GO:0005634). Therefore the regulation of transcription was one specific biological function of the cell nucleus, a deduction that aligned with the current domain knowledge [83].

There were two red-coloured high-density regions near the lower right corner of the '*nucleus*' cluster that also drew our attention. These two regions seemed to be connected to the '*nucleus*' cluster by inter-cluster edges. The region closest to the '*nucleus*' cluster contained the five high degree hubs previously identified in the non-clustered PIN (see FIGURE 4.7(a)). We found that they were also members of the cluster labeled '*GO:0005737 cytoplasm*'. Hence, we deduced that these hub proteins may be transported in between the nucleus and the cytoplasm organelles. The region at a further distance from the '*nucleus*' cluster contained two high degree nodes labeled *HDAC1* and *HDAC3*. We found that they were members of the cluster labeled '*GO:0000118 histone deacetylase complex*' (see FIGURE 4.7(b)). This finding was supported by the current knowledge about their function as histone deacetylases [99]. At this point, we could not make any more deductions from the clustered PIN visualization because of the difficulty in traversing inter-cluster edges from any of the high degree nodes identified.

III. Summary

For this analysis on the regulation of transcription-defined PIN, we found that the non-clustered PIN visualization was able to support the following analytical tasks to varying degrees of success.

1. Identify co-expressed proteins based on node colour.
2. Identify protein-protein interactions that are likely to be functioning.
3. Identify unique node topologies in the PIN.

We had no difficult performing the first task but had only limited success with the second and the third tasks. At a scale of 577 proteins and 1211 interactions, we could identify only those co-expressed proteins that resemble date hubs and the edges connecting them together. At the scale of 2019 proteins and 4039 interactions, the clustered PIN visualization supported the following analytical tasks poorly.

1. Identify the intracellular distribution of co-expressed proteins.
2. Identify protein-protein interactions between co-expressed proteins.

We could visually detect that there are co-expressed proteins being displayed in different cluster nodes, but we could neither identify each individual protein nor the inter-cluster edges connecting the co-expressed proteins. The visual complexity encountered hampered these tasks. In this case, the clustered PIN visualization failed to provide more information than its non-clustered counterpart especially when the node-edge distribution skewed towards a single cluster.

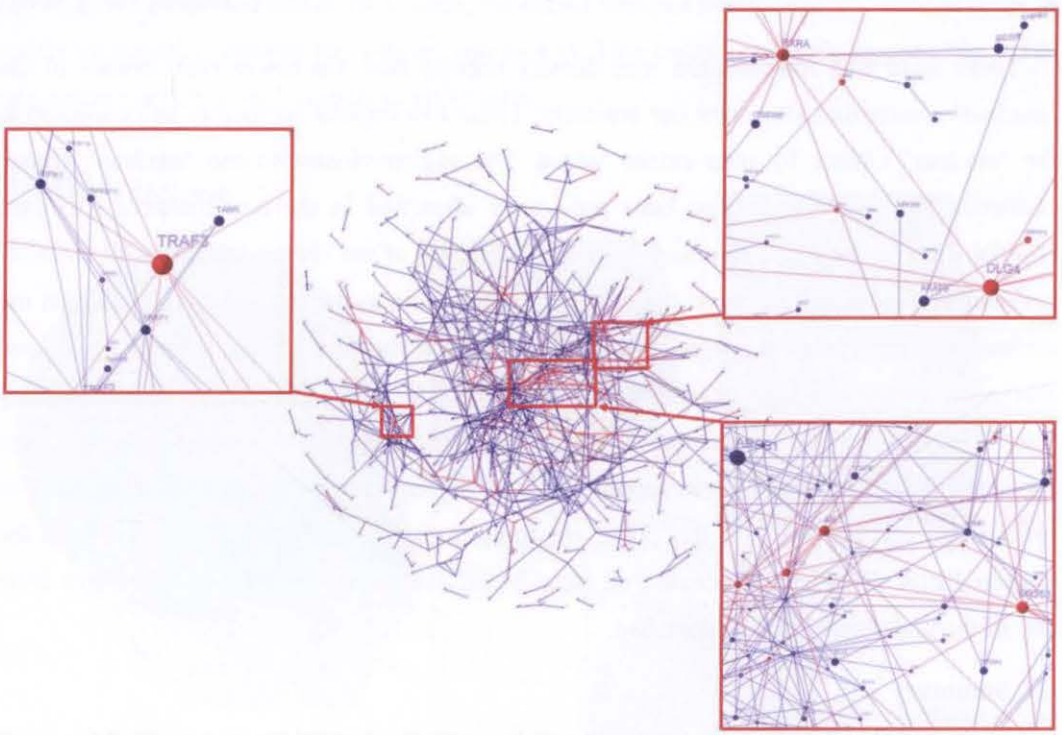


FIGURE 4.8. Non-clustered PIN visualization of the signal transduction (GO:0007165) biological process in the force-directed layout. The overview of the non-clustered PIN is at the centre. Insets of bounded areas (red boxes) show examples of nodes representing the co-expressed proteins.

2. Signal transduction (GO:0007165)

I. Non-clustered PIN

FIGURE 4.8 showed the non-clustered PIN visualization for the signal transduction biological process. It consisted of 563 proteins and 832 interactions. An attempt to identify all the co-expressed proteins was hampered by the time-consuming navigation through the PIN visualization using pointer-directed panning. We found that the only nodes that were easy to identify in this layout are those that were coloured red and resembled date hubs (node degree $\geq$ 10). They were labeled *AR*, *CREBBP*, *STAT3*, *NR3C1*, *DLG4*, *RXRA*, and *TRAF4*. This was suggesting that they co-expressed. We traversed the red coloured edges originating from each of these hubs and found that only *AR*, *CREBBP*, *STAT3*, and *NR3C1* were connected to each other. This was suggesting that they were actively interacting with each other to perform certain biological functions.

We examined the known biological function of each hub protein documented in the Entrez public database [136]. The biological functions of *AR*, *CREBBP*, *STAT3*, and *NR3C1* were mentioned in the regulation of transcription-defined PIN. Their appearance in the signal transduction-defined PIN further confirmed their dual functionality as a signaling protein and a gene regulator.

We examined the known biological function of the protein node labeled *DLG4* (GeneID: 1742) and found that it was a signaling protein of the N-methyl-D-aspartate glutamate receptor (GO:0032281). It was found in human nerve cells and was involved in fast synaptic transmission (GO:0007268) [102]. We found that the node labeled '*HGS*' was also coloured red and was a neighbour of node *DLG4* (see FIGURE 4.8, upper right inset). Hence, *HGS* co-expressed and interacted with *DLG4*. The known biological function of *HGS* (Gene ID: 9146) was an endosomal ATPase protein that regulated endosome transport (GO:0016197), and *DLG4* was somehow involved in this process. However, the interaction between *HGS* and *DLG4* had only been known to occur in nerve cells [30]. We were puzzled to find that a protein-protein interaction of nerve cell origin to be actively functioning in HCC cells, especially biologists found that *DLG4* was not highly expressed in normal hepatocytes [119]. We were uncertain on the biological significance of this finding. *RXRA* (GeneID: 6256) was a nuclear receptor that mediated the retinoic-acid induced gene expression. Thus *RXRA* was also a gene regulator. Although its up-regulation in the hepatoma cell line had been shown to associate with cell growth [166], the exact role of *RXRA* in regulating cell cycle progression was still poorly understood.

TRAF4 (GeneID: 9618) was a protein that interacted with neurotrophin receptors. It had been known to inhibit apoptosis. Recently, cancer biologists found that the over-expression of *TRAF4* due to gene amplification was very common in different types of cancer [21]. Gene amplification meant a cell having more than two copies of the same gene in its genomic DNA. We therefore hypothesized *TRAF4* may be another oncogenic protein that accelerated HCC proliferation.

II. Clustered PIN

FIGURE 4.9 showed the clustered PIN visualization for the signal transduction biological process. It consists of 1689 proteins and 2041 interactions in 62 clusters. We found that identifying co-expressed proteins using the clustered PIN visualization is much easier than using its non-clustered counterpart. That is because, in the clustered circular layout, the cluster nodes define specific areas in the visualization for containing the protein nodes. The circular layout of the protein nodes within each cluster node reduces their inter-node distances. Instead of having to navigate throughout the visualization, we simply zoomed in to those clusters that contain red-coloured nodes and red-coloured intra-cluster edges.

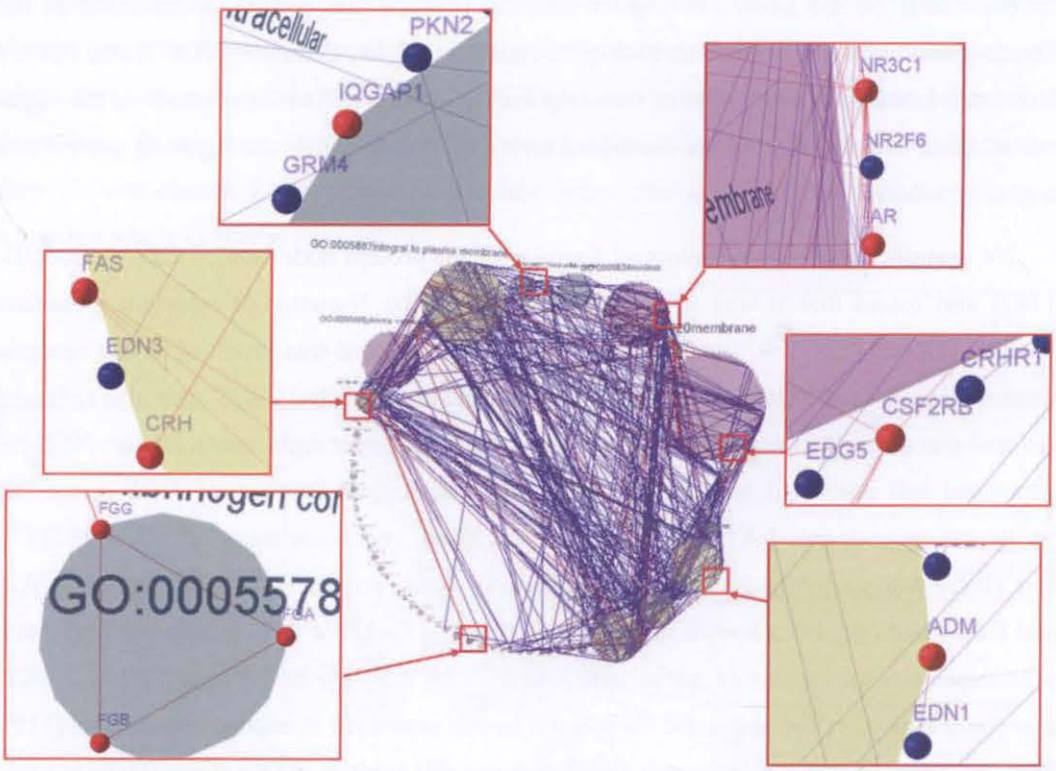


FIGURE 4.9. Clustered PIN visualization of the signal transduction (GO:0007165) biological process in the clustered circular layout. The overview of the clustered PIN is at the centre. Insets of bounded areas (red boxes) show examples of nodes representing the co-expressed proteins.

In this way, we did not have to rely on node degree to identify the co-expressed proteins. Using the previously mentioned interactivity, we identified twenty six co-expressed proteins. Including the five proteins identified in the non-clustered PIN visualization, a total of thirty co-expressed proteins had been identified altogether. We examined the known biological functions of each co-expressed proteins documented in the Entrez public database [136] and found that they could be divided into five groups. They were (1) nine growth factors and their receptors *FAS*, *CCL15*, *CXCR4*, *CXCL12*, *CSF2RB*, *IL2RG*, *IL15RA*, *TGFBR3*, and *VEGFB*; (2) eight signal mediators which were either signal transducers *STAT3*, *RANGAP1*, *RAP1A*, *RAP2A*, or signal amplifiers *HGS*, *IQGAP1*, *MAPK1*, and *PIK3CD*; (3) five inflammation-induced proteins such as fibrins *F2R*, *FGA*, *FGB*, *FGG*, and a complement factor *C3AR1*; (4) four peptide hormones and their receptors *ADM*, *AR*, *CRH*, and *RXRA*; (5) and the rest were the extracellular matrix protein *ACTL6*, the cyclin-dependent kinase *CDK4*, the apoptotic factor *TNFSF10*, and the hypoxia-induced gene regulator *HIF1A*. At least thirteen of these proteins were known to be induced by inflammation. We deduced that HCC development or progression might involve the activation of inflammation-related signaling pathways such as those activated by cytokines and interleukins.

III. Summary

For this analysis on the signal transduction-defined PIN, we found that the non-clustered PIN visualization suffered the same limitations seen with the analysis on the regulation of transcription-defined PIN. We could identify only those co-expressed proteins that resemble date hubs and the edges that connect them together. The visual complexity of the non-clustered PIN visualization severely limited its use for visual analysis, and hence biological deduction.

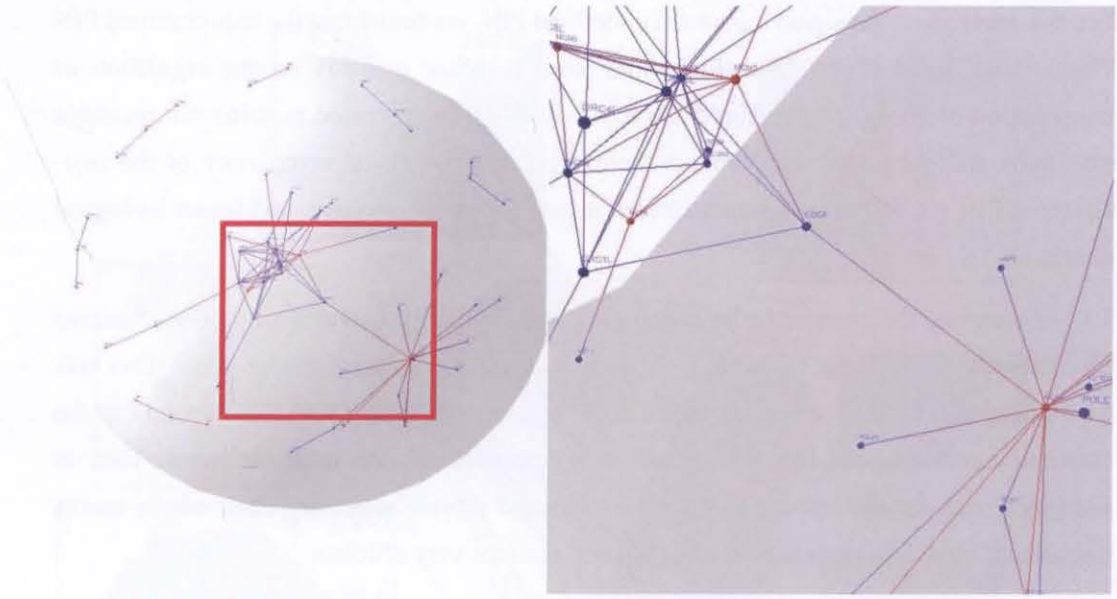
To our surprise, the clustered PIN visualization proved more useful than its non-clustered counterpart when coming to the identification of non-hub co-expressed proteins. This task was supposed to be well supported by the non-clustered PIN visualization. We found that the clustered circular layout has the advantage of constraining inter-node distances. This, in addition to the distinct colouring of the co-expressed protein nodes and their edges, makes the tasks of identifying interacting co-expressed proteins very efficient.

4.4.2.3. Limitless replicative potential

GO Process terms relating to limitless replicative potential are several. These included cell cycle (GO:0007049), regulation of cell growth (GO:0001558), chromatin remodeling (GO:0006338) and DNA replication (GO:0006260). We decided to investigate only the DNA replication (GO:0006260) process since the cost of limitless replicative potential to HCC cells is replication stress. The consequence of which is increased genome instability that has been suspected to enhance the invasiveness of HCC [19].

I. Non-clustered PIN

FIGURE 4.10 showed the non-clustered PIN visualization for the DNA replication biological process. It consisted of 55 proteins and 83 interactions distributed in seven connected components. The largest connected component consisted of 40 proteins and 45 interactions. We found in the zoom-in view that the largest connected component was made of two connected sub-networks, each with a distinct network topology (see FIGURE 4.10(a)). One sub-network consisted of a group of nine highly inter-connected nodes. In biological terms, each of these nodes was called a *date* hub [66]. The other was a star shape sub-network with the centre node being labeled *PCNA*. In biological terms, this node was called a *party* hub [66]. The two sub-networks were connected by the node labeled *CDC6*. At the same time, *CDC6* was also a member of the party hub. Therefore biologists called such type of protein a *hub-bottleneck* protein [168].

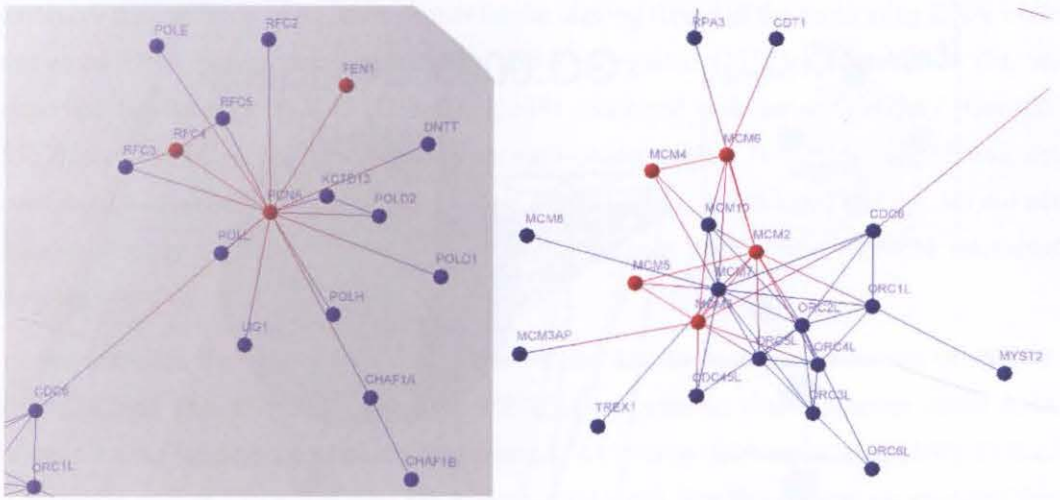


(a) (b)

FIGURE 4.10. Non-clustered PIN visualization of the DNA replication (GO:0006260) in the force-directed layout. (a) Overview. (b) Zoom in view of the bound area.

Biologists who studied PINs had been suggesting that the hub-bottleneck protein tended to serve as connectors for holding functionally different complexes together to form a large protein complex [168]. When we interpreted FIGURE 4.10(b) in this context, the large connected component should represent a single protein complex. Within this protein complex were two connected sub-complexes with each being represented by a sub-network. In the star-shape sub-network, the node topology of the date hub *PCNA* (see FIGURE 4.10(a)) suggests that it served as an exchange point for various proteins when different functionalities are required [137]. In the highly connected sub-network, the inter-connectivity among *MCM* proteins (*MCM2* to *MCM8*) and *ORCL* proteins (*ORC1L* to *ORC6L*) suggested that they served a common function fundamental to the DNA replication biological process.

In terms of expression dynamics, we found that the date hubs labeled *FEN1*, *PCNA*, *RFC4*, are coloured red (see FIGURE 4.11(a)). We therefore deduced that only two proteins *FEN1* and *RFC4* were co-expressed with *PCNA*, and likely to be interacting with one another. *FEN1* (GeneID: 2273) is a protein required for Okazaki fragment maturation during replication of lagging DNA strand and excising base-mismatch in DNA repair [96]. We also found the party hubs labeled *MCM2*, *MCM3*, *MCM4*, *MCM5*, and *MCM6* are coloured red (see FIGURE 4.11(b)). We therefore deduced that these proteins are actively interacting with one another.



(a) (b)

FIGURE 4.11. A zoom in view on the two sub-networks forming the largest connected component seen in FIGURE 4.10. (a) the *PCNA* complex. (b) Pre-replicative (pre-RC) complex.

According to the documented biological functions of the party hubs displayed in FIGURE 4.11(b), they form the subunit known as the *pre-replicative complex* (pre-RC) [64]. In the pre-RC subunit, the *MCM* proteins initiate the unwinding of the DNA double strand and the *ORCL* proteins interact with specific DNA sequences known as the *origin of replication*. At the same time, the pre-RC subunit provides a platform for the *PCNA* hub and its neighbours. We expected that all the party hubs should be co-expressed but according to FIGURE 4.11(b), the colour coding indicated that only some of the *MCM* proteins were co-expressed and the *ORCL* proteins were not expressed at all. Thus, any protein-protein interactions involving *ORCL* proteins may be inactive. From this, we deduced that part of the DNA replication complex was poorly formed and could be functionally deficient.

We searched the biological literature for interpreting the significance of *MCM* protein co-expression. We found that this phenomenon has been reported in many forms of cancer [64]. Cancer biologists also recognized that an up-regulated *MCM7* was usually associated with a high proliferation rate [65, 126]. While we do not have enough information to verify *MCM7* up-regulation, we observed that the *MCM7* node was coloured blue (see FIGURE 4.11(b)). That indicated *MCM7* did not co-express with *MCM2*, *MCM3*, *MCM4*, *MCM5*, and *MCM6* proteins. Hence, we suspected that the progression of HCC might not rely on a high proliferation rate.

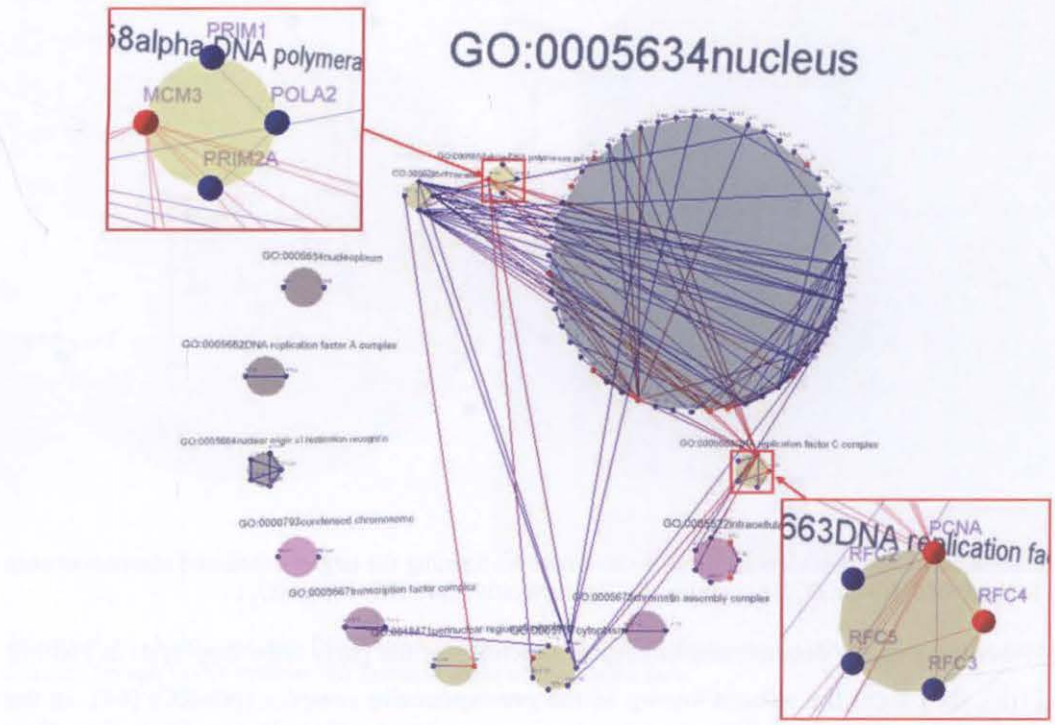


FIGURE 4.12. Clustered PIN visualization of the DNA replication (GO:0006260) in the clustered circular layout. The zoom in view of cluster '*GO:0005658 alpha DNA polymerase:primase complex*' cluster (cluster (2)) is shown at the top left corner. The zoom in view '*GO:0005663 DNA replication factor C complex*' cluster (cluster (3)) is shown at the lower right corner.

II. Clustered PIN

FIGURE 4.12 showed the clustered PIN visualization for the DNA replication biological process. It consisted of 148 nodes and 153 edges in 13 clusters. From the GO Component labels, we recognized that the thirteen clusters represented two types of sub-cellular components, i.e. organelles and protein complexes. Six of them were highly inter-connected. They were (1) '*GO:0005634 nucleus*', (2) '*GO:0005658 alpha DNA polymerase:primase complex*', (3) '*GO:0005663 DNA replication factor C complex*', (4) '*GO:0000785 chromatin*', (5) '*GO:0005622 intracellular*', and (6) '*GO:0005737 cytoplasm*'. The largest cluster was the nucleus. The GO Component labels (2) and (3) represented subunits of the larger DNA replication protein complex. The label for cluster (4) represented a type of organelle in the cell nucleus. Chromatin was the ordered and organized complex of genomic DNA and protein that formed the chromosome (GO:0000785). We therefore deduced that proteins represented by protein nodes in cluster (4) interact directly with the genomic DNA in chromatin.

We identified the node *MCM3* not only in cluster (1) but also in cluster (2) as well (see FIGURE 4.12). For cluster (2), its GO Component label represented the alpha DNA polymerase:primase complex. It is a subunit of the DNA replication complex which

catalyzes the synthesis of an RNA primer on the lagging strand of the replicating DNA while the alpha DNA polymerase is catalyzing DNA elongation [51]. In cluster node (2), we observed that the node *POLA2* (GeneID: 23649) connected with the node *PRIM1* (GeneID: 5557) and *PRIM2A* (GeneID: 5558) with intra-cluster edges. However, *MCM3* was not connected to the nodes *PRIM1*, *PRIM2A*, and *POLA2*. Thus, we deduced that *MCM3* did not interact directly with *PRIM1*, *PRIM2A*, and *POLA2*, but *PRIM1* and *PRIM2A* interacted directly with *POLA2*.

We searched the Entrez public database for the known biological function of *PRIM1*, *PRIM2A*, and *POLA2* [99]. *PRIM1* and *PRIM2A* are proteins that synthesize small RNA primers on the lagging strand of the replicating DNA during DNA replication [160]. *POLA2* is a protein required for catalyzing DNA elongation [160]. For this reason, we expected that *PRIM1*, *PRIM2A*, and *POLA2* should co-express with *MCM3* in order to form a stable alpha DNA polymerase:primase complex. However, the blue node colour of *PRIM1*, *PRIM2A*, and *POLA2* informed us that they did not co-express with *MCM3*. Therefore, it is possible that there were functional alpha DNA polymerase:primase complexes being formed but they may be unstable.

We found that the two red coloured nodes labeled *PCNA* and *RFC4* were members of cluster (3). For cluster (3), its GO Component label represented the DNA replication factor C complex. It is another subunit of the DNA replication complex which loads the protein *PCNA* onto the DNA to initiate DNA synthesis catalyzed by DNA polymerases [64]. In cluster node (3), the nodes labeled *RFC2*, *RFC3*, *RFC4*, and *RFC5* represent the core proteins that had been known to form the DNA replication factor C complex. The intra-cluster edges between them showed that they interacted with one another. We expected that, for this protein complex to function, the above proteins should co-express with *PCNA*. However, we found that only *RFC4* and *PCNA* are coloured red (see FIGURE 4.12). This observation led us to the deduction that the replication factor C complex might be defective.

III. Summary

For this analysis on the DNA replication-defined PIN, we found that the non-clustered PIN visualization supported the following analytical tasks very well.

1. Identify co-expressed proteins based on node colour.
2. Identify protein-protein interactions between co-expressed proteins.
3. Identify unique node topologies in the PIN.

The clustered PIN visualization also provided us with more detailed information on the molecular organization of the DNA replication complex. This allowed us to make deductions

complementary to those made with the non-clustered PIN visualization. In short, the clustered PIN visualization supported the following analytical tasks well.

1. Identify the intracellular distribution of co-expressed proteins.
2. Identify protein-protein interactions between co-expressed proteins that are distributed in different clusters.

4.4.2.4. Angiogenesis, Tissue invasion, Metastasis

Angiogenesis (GO:0001525) is defined as a biological process that mediates blood vessel formation when new vessels emerge from the proliferation of pre-existing blood vessels. Cancer biologists had been suggesting that tissue invasion and metastasis probably depended on the same protein-protein interactions. Their rationale was that tissue invasion and metastasis had often been observed in parallel to angiogenesis [59]. For this reason, we decided to investigate only the angiogenesis (GO:0001525) biological process.

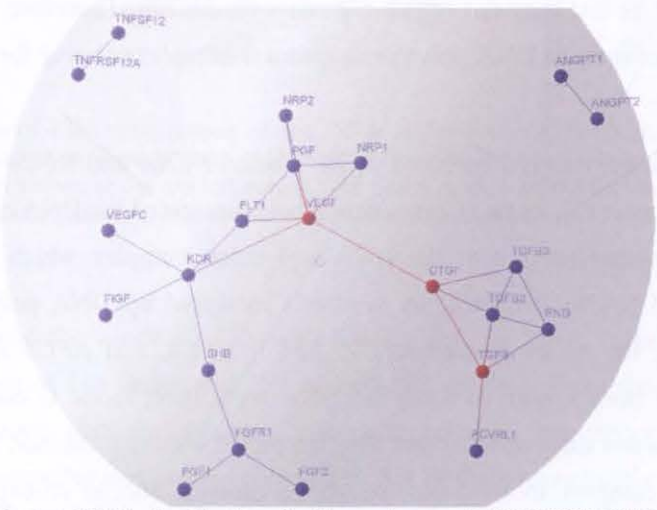


FIGURE 4.13. Non-clustered PIN visualization for the angiogenesis (GO:0001525) biological process using the force-directed layout.

I. Non-clustered PIN

FIGURE 4.13 showed the non-clustered PIN for the angiogenesis biological process. It consists of 24 proteins and 28 interactions in three connected components. Eighteen of them formed a large connected component. Within which, we identified two sets of party hubs. One set consisted of nodes labeled *CTGF*, *TGFB1*, *TGFB2*, *TGFB3*, *ENG*, and *ACVRL1*. Another set consisted of nodes labeled *VEGF*, *PGF*, *FLT1*, *NRP1*, *NRP2*, *FIGF*, *VEGFC*, and *KDR*. The two sets of party hubs were connected together by the nodes *CTGF* and *VEGF*. Therefore, we identified them as hub-bottleneck proteins. The node *SHB* seemed to be a non-hub bottleneck protein which sat in between *KDR* and *FGFR1*. Using the same

rationale applied in the DNA replication-defined PIN analysis, we deduced that the party hubs could be subunits of a larger protein complex.

We searched the Entrez public database and found that eleven of the proteins in the large connected component were known angiogenic growth factors [136]. This information in combination with its network topology observed led us to hypothesize that the large connected component could be representing a protein complex that mediated angiogenesis. However, the biological literature informed us that angiogenic growth factors had only been known to interact with each other as protein pairs or triplets [80, 154]. Furthermore, the functioning of angiogenesis as a biological process had not been known to involve an eighteen-member protein complex either. Therefore our hypothesis was rejected by the current biological research. The large connected component shown in FIGURE 4.13 could merely represent a collection of pairwise interactions that shared the same set of proteins, not a representation of a protein complex. Therefore, the non-clustered PIN visualization did not represent the biological reality currently known to biologists.

In the large connected component, we found the nodes labeled *CTGF*, *TGFB1* and *VEGF* were coloured red indicating that they were co-expressed in HCC. This co-expression could be explained by the finding that *CTGF* (also known as *CCN2*; GeneID: 1490) is an early intermediate gene induced by *TGFB1* signaling and itself could induce *VEGF* expression [8]. Furthermore, we noticed that *CTGF* was connected to *TGFB1* (GeneID: 7040) and *VEGF* (GeneID: 7422) with red-coloured edges. Therefore, we deduced that *CTGF* interacted with *TGFB1* and *VEGF*.

According to the biological literature [80], *VEGF* is the major angiogenic growth factor which signals through *VEGFR2* (GeneID: 3791), the main signaling *VEGF* receptor that mediates neo-angiogenesis. The interaction between *VEGF* and *VEGFR2* can trigger multiple signaling paths that lead to (1) the induction of DNA replication (GO:0006260), (2) cell growth for endothelial cells, (3) actin cytoskeletal re-modeling and eventually (4) endothelial cell migration. Biologists had also found recently that *CTGF* is also involved in angiogenesis, probably through endothelial cell growth, cell migration (GO:0016477), and cell-cell adhesion (GO:0007155). They further suspected that *CTGF* might act as a co-factor of *TGFB1* in mediating the same processes [108]. The biological significance of the interaction between *CTGF* and *VEGF* is unknown. *CTGF* might interact with multiple signaling proteins and gene regulators that regulate cell growth [108]. We speculated that *CTGF* and *VEGF* might act as cofactors to each other in order to amplify their influence of angiogenesis, but this had not been supported by any biological research so far.

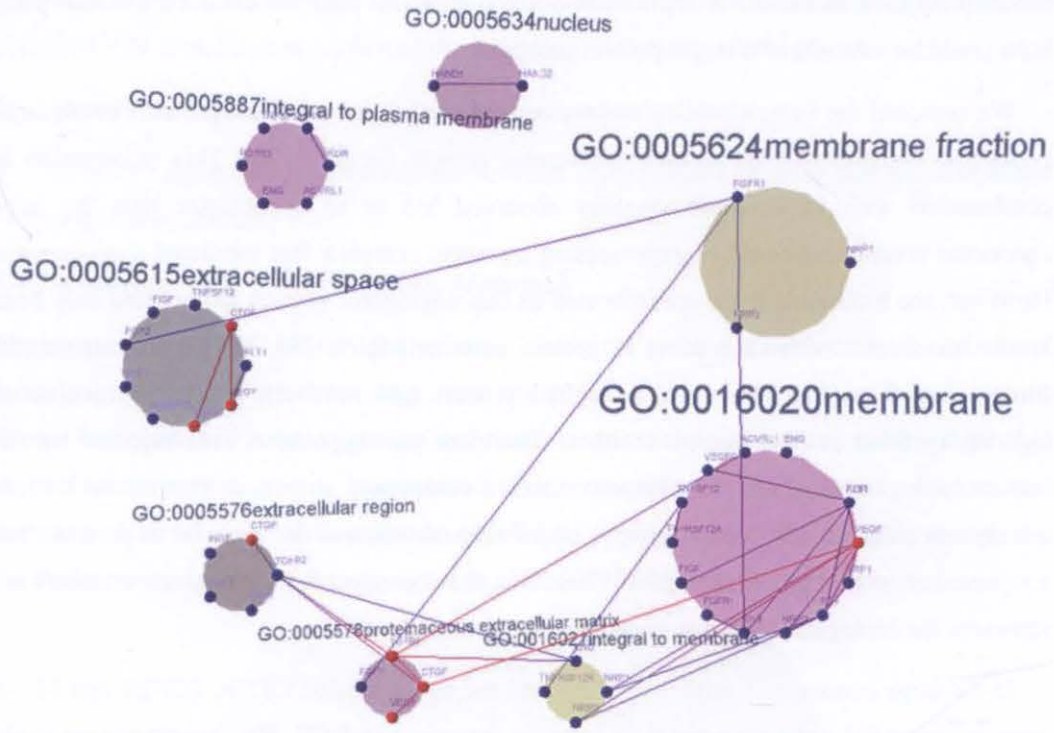


FIGURE 4.14. Clustered PIN visualization of the angiogenesis (GO:0001525) biological process using the clustered circular layout.

II. Clustered PIN

FIGURE 4.14 showed the clustered PIN visualization for the angiogenesis biological process. It consisted of 21 nodes and 21 edges in 8 clusters. The GO Component labels shown were (1) '*GO:0005634 nucleus*', (2) '*GO:0005624 membrane fraction*', (3) '*GO:0016020 membrane*', (4) '*GO:0016021 integral to membrane*', (5) '*GO:0005578 proteinaceous extracellular matrix*', (6) '*GO:0005576 extracellular region*', (7) '*GO:0005615 extracellular space*', and (8) '*GO:0005887 integral to plasma membrane*'.

We noticed that the red coloured nodes labeled *CTGF* and *VEGF* were connected by a red coloured intra-cluster edge in clusters (5) and (7). The node *CTGF* in cluster (5) is also connected to the node *VEGF* in cluster (3) by a red coloured inter-cluster edge. This observation implied that the proteins *CTGF* and *VEGF* interacted with each other in the extracellular space and also in the cell membrane. As mentioned in the last section, *VEGF* interacts with its receptor protein *VEGFR2* on the cell membrane to trigger angiogenesis [80] but the receptor protein for *CTGF* is unknown. Based on the clustered PIN visualization, we further refined our previous deduction about the biological significance of the interaction between *CTGF* and *VEGF*. *CTGF* could be a co-factor of *VEGF* and both might interact with the receptor protein *VEGFR2* on the cell membrane to amplify the rate of angiogenesis.

III. Summary

For this analysis on the angiogenesis-defined PIN, we found that the non-clustered PIN visualization supported the following analytical tasks effectively.

1. Identify co-expressed proteins based on node colour.
2. Identify protein-protein interactions that are likely to be functioning.
3. Identify unique node topologies in the PIN.

However, we also discovered two limitations with the non-clustered PIN visualization. The first problem was that even though task 3 was supported by the visualization, it did not lead to a biologically meaningful deduction. The visualized large connected component led us to a false deduction because we tried applying the same node topology-based deduction as in the DNA replication-defined PIN analysis. The second problem was that the non-clustered PIN visualization was not as informative as its clustered PIN counterpart. In the present analysis, we relied on the clustered PIN visualization to provide information on the distribution of proteins in different cell components. By this means, we refined any deductions made with the non-clustered PIN visualization. In short, the clustered PIN visualization supported the following analytical tasks effectively.

1. Identify the intracellular distribution of co-expressed proteins.
2. Identify protein-protein interactions between co-expressed proteins that are distributed in different clusters.

4.4.2.5. Conclusion

For each biological process, the non-clustered PIN visualization supported the four analytical tasks mentioned at the beginning of this section to a varying degree of success. These tasks were:

1. Identify co-expressed proteins based on node colour.
2. Identify protein-protein interactions that are likely to be functioning.
3. Identify unique node topologies in the PIN.

Generally speaking, non-clustered PIN visualization in the force-directed layout did not support task 2 for GO Process-defined PIN that approached 500 nodes or more. The visualized non-clustered PINs for the regulation of transcription (GO:0006355) and signal transduction (GO: 0007165) biological processes were comparable. Both had more than 500 nodes and more than 1000 edges. At this scale, it was difficult to identify the protein-protein interactions specified in task 2 because edge traversal was frequently hampered by edge crossing.

When coming to task 3, the high degree nodes that resembled date hubs and were coloured red attracted our visual attention. However, our biological deductions were also limited to these date hubs because of the difficulty in carrying out task 2. This meant all the ink for drawing the non-clustered PINs for the regulation of transcription (GO:0006355) and signal transduction (GO: 0007165) biological processes was wasted. Therefore non-clustered visualization in the force-directed layout was not effective as a visual analysis method at a large scale. In this regard, our finding that it was easier to identify non-hub co-expressed proteins in the clustered PIN visualization of the signal transduction-defined PIN might provide a hint that visual clustering may overcome the limitation of force-directed layout.

We found that interpreting clustered PIN visualizations was cognitively less challenging because of the additional information provided by the GO Component labels. With the non-clustered PIN visualization, we relied almost completely on the current biological knowledge on either a particular set of proteins and/or a particular set of protein-protein interactions to make deductions. Our analysis on the DNA replication-defined PIN was a good illustration. Although we could deduce from the non-clustered PIN visualization that the DNA replication protein complex might have two sub-networks of distinct functionalities, it was the clustered PIN visualization that informed us on the structural organization of the same protein complex. Clustered PIN visualization was therefore more informative than its non-clustered PIN counterpart. With the exception of the regulation of transcription-defined PIN, the clustered PIN visualization in the clustered circular layout supported the following analytical tasks effectively.

1. Identify the cellular distribution of co-expressed proteins.
2. Identify protein-protein interactions between co-expressed proteins that are distributed in different clusters.

In all analyses, we found that the use of both visualizations were necessary for making biological deductions. Biologists who were interested in studying PIN from the systems biology viewpoint would especially find the clustered PIN visualization indispensable.

Of the seven cancer-related biological processes we had investigated so far, the cell cycle arrest-defined PIN showed the least number of co-expressed genes. Only *CDKN2A* was shown to be expressed but with no co-expressing neighbours. Hence there might not enough functioning protein-protein interactions for initiating cell cycle arrest. While exploring the signal transduction-corresponding PIN visualizations, we noticed that almost one-third of the co-expressed genes were growth factors and inflammatory cytokines, and another one-third of them were signal mediators. These observations were suggesting that HCC resisted cell

cycle arrest but was sensitive to growth signals. Hence we concluded that both signal transduction and cell cycle arrest were the most affected biological processes.

In the next section, a user evaluation conducted with an expert biologist was being presented. The purpose was to examine the usability of each layout and how it affected biological reasoning. A domain expert evaluation was used since it did not involve response time measurement and hence, could be conducted within an hour.

4.5. Domain Expert Evaluation

To evaluate the usability of the two PIN visualizations, we conducted a user evaluation with an expert biologist who had research interests in proteomics and bioinformatics. The biologist was an expert in *E. coli*, yeast and human PINs who regularly used computer-generated visualization in his research.

The design of the evaluation tasks was based on four criteria, i.e. *information overlay*, *spatial information*, *inter-connectivity*, and *interactivity*. They were based on a selection of published molecular pathway visualization heuristics [134]. The PIN visualizations corresponding to the DNA replication biological process were used as the test case since its network size was moderate even for the clustered PIN visualization (see FIGUREs 4.10 and 4.12).

In regards to information overlay, the expert biologist had no difficulty distinguishing between co-expressed and non co-expressed proteins based on node colour. The co-expressed proteins were coloured red whereas their non co-expressed counterparts were coloured blue. Nor did the biologist have difficulty identifying edges between co-expressed and non co-expressed proteins. He did this by first identifying the red coloured node and then traversed the edge originated from the node. However, the biologist suggested using more contrasting colour hues for the cluster nodes to facilitate the identification of individual cluster as a unique cell component.

In proteomics, a *functionally essential* protein is defined as a protein that when deleted will abrogate the biological process of concern, and cannot be restored by an alternative protein. In this evaluation, that biological process was DNA replication. When asked to identify functionally essential proteins in the non-clustered PIN visualization, the expert biologist identified *MCM2*, *MCM3*, *MCM7*, *ORC2L*, and *PCNA*. He was thinking aloud while performing this task indicating verbally that he was visually searching for nodes that had a node degree greater than five.

When performing the same task on the clustered PIN visualization, the biologist also identified *PCNA* as one of the essential proteins but he was not certain whether *MCM2*,

MCM3, *ORC1L*, and *ORC2L* were essential proteins. He approached this task by first identifying nodes with a node degree > 5 in each cluster and then examined if the same protein nodes had inter-cluster edges that connected them with those in other clusters. His rationale was that protein nodes that were highly connected inside and outside its own cellular component must be functionally essential. The biologist mentioned that the replicated nodes made it difficult to understand the connectivity between proteins as interacting neighbours. Therefore, he preferred using the non-clustered PIN visualization for identifying essential protein because it informed him better on how highly connected each protein really is. These results suggested that the redundant representation of protein nodes can affect biological reasoning.

When asked to identify which protein was a bottleneck between two or more protein complexes in the non-clustered PIN visualization, the biologist identified *CDC6* immediately. His rationale was that it is the only protein commonly connected to two complexes. This showed that the non-clustered PIN visualization presented could clearly reveal the direct interactions between proteins. He was then asked to interpret the biological implication of co-expression between *PCNA* and *MCM3*. The biologist stated that the up-regulation *PCNA* and *MCM3* might be not mediated by *CDC6* given that it was positioned between *PCNA* and *MCM3*. His response indicated that he was associating the bottleneck protein with its potential gene regulatory role in the PIN.

Identifying inter-cluster connections was often important in deducing the functional relationship between cellular components. The biologist correctly identified from the clustered PIN visualization that '*GO:0005634 nucleus*' and '*GO:0000785 chromatin*' were the most highly inter-connected clusters. His deduction was based on the inter-cluster edge density observed. The biologist also suggested that it would be more helpful if the clusters were arranged according to the descending order of their inter-cluster connectivity.

In terms of interactivity, the expert biologist commented that visual zooming by pointer dragging while keeping the middle button of the pointer device pressed down was not user-friendly. He preferred a dial-like device for controlling the zooming function. He also found it tedious to use pointer motion for navigating through the visualization. In short, he found the interactivity provided short of his expectation.

Lastly, the biologist commented that the clustered PIN did not preserve his mental model of a PIN because its original topology had not been preserved by the clustered circular layout. However, the clustered PIN visualization was still a good complementary visualization to its non-clustered counterpart. The fixed layout of the clusters would facilitate biological analysis when the same visualization had to be examined repeatedly. In the non-clustered PIN visualization, each rendering gave a different layout. He cautioned that the

clustered PIN visualization could mislead users into thinking that the clusters are non-intersecting and the inter-cluster edges are biologically meaningful only if the clusters represent different protein complexes.

4.6. Remarks

In the HCC case study, we demonstrated the merit of applying two different visualization methods on the same GO_Process-defined PIN to gene expression analysis. Our approach allowed us to deduce the implication of gene co-expression in seven cancer-related biological processes. More importantly, it allowed us to identify which biological process(es) is/are the most affected. Our approach also demonstrated the feasibility of making biological deductions by following the approach of "*filter first, zoom and details, overview if necessary*". Visualizing a selective biological process-corresponding PIN gave us a good starting point for exploring gene expression in the context of the single cell PIN.

When applying the non-clustered PIN visualization to visual analysis, we found that the network topology in the force-directed layout allowed the identification of the various protein hubs and bottlenecks, i.e. party hubs, date hubs, hub-bottlenecks, and non-hub bottlenecks. Their topologies informed us of their network properties and allowed us to deduce their probable interaction dynamics with their neighbours. However, it was the clustered PIN visualization that informed us of their role in a biological process. Often it was the combination of GO Component clusters representing the collection of sub-cellular organelles and protein complexes that were the most informative. That was because an organelle functions as a compartment for localizing a specific set of biological processes whereas a protein complex is a functional module in its own right within the whole cell PIN.

In terms of usability, both visualizations had their strength and limitations. The greatest strength of the clustered PIN visualization was the *fixed* circular layout of the cluster and protein nodes. Fixed layout reduced the user's cognitive load more than the free layout since the biologist did not need to re-adapt to a new layout generated in every rendering of the same PIN. That could explain why the expert biologist commented during the user evaluation, "*The fixed positioning of clusters is important especially knowing where a particular group of proteins for an ontology term is positioned.*" We also found the clustered circular layout more effective for identifying co-expressed proteins than the force-directed layout when the visualized network approached 500 nodes. That was because the circular layout of protein nodes within each cluster reduces their inter-nodal distance.

However, the clustered PIN visualization had one limitation that hampered its usability. The redundant representation of the same protein nodes but in different clusters had increased the size of the visualized PIN. In addition, some intra-cluster edges were being

redundantly represented as inter-cluster edges. As the number of nodes increased, the number of edge crossings also increased. When the visualized PIN approached 1000 nodes and 1000 edges or more, the edge crossing made the visualization un-readable. The non-clustered PIN visualization also had its limitation. As the size of the PIN increases, the force-directed layout became increasingly unreadable eventually giving the 'hair ball' effect [148]. The identification of party hubs became increasingly difficult whereas the high degree date hubs continued to attract visual attention as the size of the PIN increased.

Our domain expert evaluation showed that non-clustered and the clustered PIN visualizations gave the biologist a different perception on protein hubs and subsequently their identification of essential proteins. Thus different visual design could affect biological reasoning.

Finally, the PIN visualizations could be applied as the follow-up step to the GO-annotated gene cluster visualization (see Chapter 3, section 3.4). Our deductions in the HCC case study could further explain how the functional organization of a living cell might emerge from the probable protein-protein interactions underlying the various biological processes. In particular, we could explain how the loss of functional protein-protein interactions in the cell cycle arrest (GO:0007050) biological process could result in the constitutive activation of oncogenic proteins in the regulation of transcription, DNA-dependent (GO:0006355) and the signal transduction (GO:0007165) biological processes. However, our analysis of the DNA replication (GO:0006260) process suggested that the activation of these oncogenic proteins did not seem to result in high mitotic cycles, a finding that agrees well with the observation that the frequency of replication errors in HCC is low [45]. Rather, the co-expressed signaling proteins found in the signal transduction (GO:0007165) network are predominantly involved in mediating inflammation or are inflammation-induced. This observation raises suspicion that prolonged tissue-level injury caused by the de-regulation of inflammation may be the most likely cause of HCC development.

The appearance of co-expressing fibrin genes shows that fibrosis was present in HCC. Indeed, published epidemiology has revealed a strong correlation between liver cirrhosis and the heightened risk of HCC development [20]. Furthermore, the persistence expression of inflammatory cytokines can spur venous metastasis, therefore amplifying the invasiveness of HCC [182]. The co-expression of the *TGFB1*, *VEGF* and *CTGF* in the angiogenesis (GO:0001521) network and the expression of *HIF1A* in the signal transduction network suggest that angiogenesis did take place in the hypoxic microenvironment within the HCC tumour [26] and would therefore further enhance metastasis. In conclusion, our deduction had been supported by recent publications, thus validating our visual analysis.

CHAPTER 5

Visualization and Analysis of Two-Overlapping Heterogeneous Biological Networks

"Connectivity creates Complexity"

5.1. Introduction

In Chapter 1, we mentioned that the routine application of high-throughput technologies in biological research has generated copious amount of data that were challenging to decipher. This situation is not exclusive to biology. Recent advances in computing technologies have produced huge datasets, and as a result large and complex network models emerged in many other application domains, e.g. finance and sociology. Visualization can be an effective analysis tool only if it can reveal the intricate structure of the networks. Otherwise, it cannot amplify human understanding; let alone leading to new insights and hypotheses deduction. The current challenges remain to be the *scalability* and *complexity* issues.

Life emerges out of complex molecular interactions and highly orchestrated biological processes. In fact, one can view the single-cell network as a system of multiple sub-networks, with each serving a specific biological process, e.g. gene regulation, signal transduction, or metabolism [3]. These biological processes can be treated as three distinct but inter-connected (or *overlapping*) networks. Each network has its own interaction types. For example, the *metabolic network* (MN) consists of protein-metabolite interactions. Its purpose is to transform metabolites to biomass [142]. The *signal transduction network* (STN) consists of fast and transient protein-protein interactions that operate on the time scale of seconds to minutes. Its purpose is to propagate the activation/de-activation signal originated from a few signaling proteins throughout the entire STN. The *gene regulatory network* (GRN) consists of protein-gene interactions that operate on the time scale of minutes to hours. Its purpose is to control the rate of protein expression when necessary [3]. The MN, STN, and GRN are considered to be overlapping because they share a common set of proteins or genes between one another.

Because a single cell molecular network consists of a system of networks, the limitation of focusing only on one network with a single interaction type becomes apparent. For example, protein interaction networks (PIN) can be used for inferring the probable protein-protein interactions involved in the various biological processes, but they cannot fully

explain how the functional organization of a cell emerges out of co-regulated biological processes. Does the physical interaction occur for the purpose of signal transduction or for the purpose of metabolic reaction? Do certain biological processes require more than just protein-protein interactions? For this reason, we require novel methods for visualizing overlapping networks as the third step of our visual analysis framework (see Chapter 1, section 1.2). This brought us to the new problem of visualizing *overlapping networks*.

As the first step towards investigating this new problem, we introduce in this chapter the *two*-overlapping network representations and their corresponding visualization methods. They are really a type of multi-plane layout [183]. The rationale behind our choice is that integrated analysis, hence systems-level insight, should be better supported by good visualizations of two inter-connected heterogeneous networks, rather than a separate visualization of each network. This is being achieved by visually highlighting the inter-connections between the two networks while exposing their differences in interaction types and network topologies. The two-overlapping network can come in the two-plane or the three-plane representation. The difference between them lies in the way the inter-connections are being highlighted. It is noteworthy to mention that, thus far, multi-plane layouts have only been experimented on metabolic networks [14, 187]. The multi-plane layout methods presented in this chapter and chapter 6 represent our first experimentation on heterogeneous biological networks.

To evaluate their merits as visual analysis methods, we perform two case studies using the previously mentioned biological networks (MN, PIN, GRN and STN) as the input. The first case study involves molecular networks found in the bacterium *Escherichia coli*. *E. coli* has been one of the best studied organisms in biology. It has been studied by biologists coming from the domains of medicine and biotechnology for the last two decades. Hence, it has been used as a model organism for studying prokaryotic biological networks on the systems scale. For this reason, *E. coli* is a suitable case study for evaluating the potential of overlapping networks as a concept model visualization. *Concept model visualization* means that the visualization represents human knowledge [76]. Visual experimentation is done on two combinations of networks, the MN-PIN and the GRN-PIN. The objective is to evaluate their effect on biological reasoning and see whether any deductions made are supported by the recent or current biological literature. We also use the GRN-PIN combination to fulfill another objective, i.e. to evaluate the visualizations of the two-plane and the three-plane representations for their readability when networks of over 500 nodes are used.

The second case study involves human molecular networks. The visual experimentation and analysis are done on the STN-PIN combination. Instead of using the single-cell cancer STN [35] and the canonical human PIN [15], only the *TGFB1* (transforming growth factor

beta) sub-network and the nuclear PIN are being applied, since both have been known to be active in different forms of cancer. The objective is to evaluate the effectiveness of the STN-PIN combination in hypothesis deduction especially as a follow-up step to the PIN visualizations presented in the previous chapter (see Chapter 4, section 4.4.2). The domain application remains to be hepatocellular carcinoma (HCC). The analytical objective is to study the probable effect of *TGFB1* signal transduction on the human cell cycle. In both case studies, we also evaluate the merits of the two-plane and three-plane visualizations as visual analysis methods.

Because these visualizations are very new to biologists and their interpretation can take several hours, it is difficult to design a user evaluation. For this reason, we do not include user evaluation as part of our investigation. Instead, we use the case studies to provide anecdotal evidence on the usability of each two-overlapping network visualization.

The rest of this chapter is divided into five sections. The representations of the MN, GRN and STN are defined in section 5.2. The two representations of the two-overlapping network are defined in section 5.3. The drawing algorithm for the layout of each representation is presented in section 5.4. The *E. coli* case study is introduced in section 5.5 followed by the human case study in section 5.6. Finally, the strength and limitations of each layout and the suitability of the two-overlapping network visualization for biological analysis are discussed in section 5.7.

5.2. Representation of Three Molecular Networks

5.2.1. Metabolic Network

The original graph theoretic model of a metabolic network (MN) is a hypergraph in which the node set represents the metabolites and the hyper-edge set represents the metabolic reactions [11]. In practice, metabolic networks are often represented as directed bipartite graphs in which one node set represents the metabolites and the other node set represents the metabolic proteins called *enzymes* [14]. For the applications discussed in this chapter, the graph-theoretic definition of the metabolic network is defined as the following:

Definition 5.1. *A metabolic network is a directed bipartite network* $G_M = (V_1, V_2, E_M)$ *in which* V_1 *denotes the node set of enzymes,* V_2 *denotes the node set of metabolites, and* E_M *denotes the edge set of unidirectional reactions. The pair of edges* e_1 *and* e_2 *where* $e_1 = (u, v)$ and $e_2 = (v, w)$ *denote the single directional reaction catalyzed by the enzyme* v *which converts metabolite* u *to metabolite* w *where* $v \in V_1$, $u, w \in V_2$, *and* (u, v), (v, w) *are ordered pairs. If the same reaction involving metabolites* u *and* w *is bi-directional, then it is represented by an additional pair of edges* e_3 *and* e_4 *where* $e_3 = (w, v)$ and $e_4 = (v, u)$.

Note that V_1 is a subset of the protein nodes in a PIN because metabolic enzymes are proteins that specialize in catalyzing metabolic reactions [83]. A metabolic reaction often generates a number of small molecules. These molecules are often known as *currency* metabolites and are usually omitted in the graph-theoretic representation and the visualization of MN. Examples of currency metabolites are ATP, ADP, H_2O, NO_2, CO_2, H^+, and inorganic phosphate.

5.2.2. Signal Transduction Network

Strictly speaking, a signal transduction network (STN) should be a directed bipartite graph in which one node set represents the signaling proteins and the other represents the energy molecules cyclic-ATP, ATP, cyclic-GTP and their metabolites cyclic-ADP, ADP, cyclic-GDP, and inorganic phosphate. However, the STN is designed for propagating the activation/de-activation signals from a subset of proteins called *receptors* to the rest of the network rather than metabolizing energy molecules. Therefore, it can be reduced to a set of protein-protein interactions which are directional in order to account for the flow of signals. For the applications discussed in this chapter, the graph-theoretic definition of the STN is defined as the following:

Definition 5.2. *A signal transduction network is a directed network* $G_{ST} = (V_{ST}, E_{ST})$ *in which* V_{ST} *denotes the finite node set of signaling proteins and* E_{ST} *denotes the finite directed edge set of directional signaling. The edge* $e = \langle v_1, v_2 \rangle \in E_{ST}$ *denotes the signal relay from signaling protein* v_1 *to signaling protein* v_2 *where* $v_1, v_2 \in V_{ST}$ *and* $\langle v_1, v_2 \rangle$ *is an ordered pair.*

Note that G_{ST} is a subset of the PIN because interactions. Similar to the MN, the currency metabolites generated by the signaling interactions are omitted. It should also be noted that the STN visualization presented in this chapter does not provide information on the activation/de-activation state of each protein as a consequence of phosphorylation or acetylation. Some proteins become active while others become inactive when phosphorylated or acetylated.

5.2.3. Gene Regulatory Network

A gene regulatory network (GRN) is often mistaken as a sub-network of the PIN. However, it is really a network of two interaction types, protein-DNA and RNA-RNA interactions. Hence the original theoretic model of the GRN should be a k-partite graph where $k = 4$. In this representation, one node set represents the proteins known as *gene regulators* or *transcription factors*. The second node set represents the non-coding RNAs. The third node set represents messenger RNAs (mRNA) [92]. The fourth node set represent the DNA

sequence that codes for a protein or RNA, and the directed edge set represents interactions between members of the four node sets.

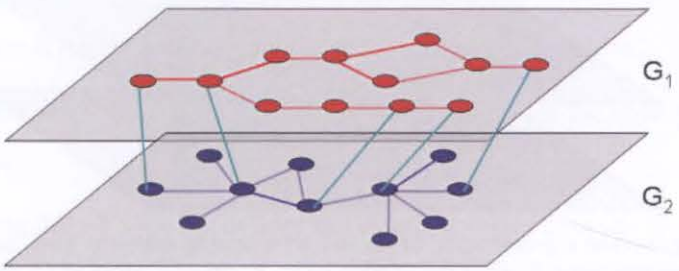


FIGURE 5.1. Two-plane representation of the two-overlapping network. Each network G_i is enclosed by a two-dimensional plane P_i. G_1 and G_2 are the heterogeneous biological networks. Nodes and edges of G_1 are coloured red within the top plane P_1. Nodes and edges of G_2 are coloured blue within the bottom plane P_3. The inter-plane edges E_{12} connecting the corresponding nodes in different planes are coloured green.

Because all four node sets represent gene products, GRN is often simplified as a network of gene-gene interactions. Classically, a *gene* is defined as the information coded by a DNA sequence in the genome [83]. In the protein-DNA interaction, the protein translated from a protein-coding gene interacts physically with the promoter sequence of another gene thereby inducing or repressing its transcription. In the RNA-RNA interaction, the microRNA transcribed from an RNA-coding gene or simply RNA gene interacts physically with its target mRNA, thereby preventing it from being translated to proteins. This process is known as *gene silencing* [120]. For the application described in this chapter, the graph-theoretic definition of the GRN is defined as the following:

Definition 5.3. *A gene regulatory network is a directed network* $G_R = (V_R, E_R)$ *in which* V_R *denotes the finite node set of genes and* E_R *denotes the edge set of directed control. The edge* $e = (v_a, v_b) \in E_R$ *denotes the induction or repression of gene* v_b *transcription (or expression) by gene* v_a *where* $v_a, v_b \in V_R$ *and* (v_a, v_b) *is an ordered pair.*

Note that the protein gene regulators are a subset of the protein nodes in a given PIN. In the human and mammalian GRNs, they interact with one another to form a protein complex and at the same time, some interact with the promoter DNA sequence upstream of a gene.

5.3. Representation of the Two-Overlapping Network

In general, k networks overlap if they share a subset of nodes and edges. This chapter describes only the overlapping networks with $k = 2$. They are heterogeneous networks with each representing a different type of interaction. For example, one network can be a PIN and the other can be an STN or a GRN. Yet these networks are inter-connected because they share a subset of common nodes and in some cases common edges as well.

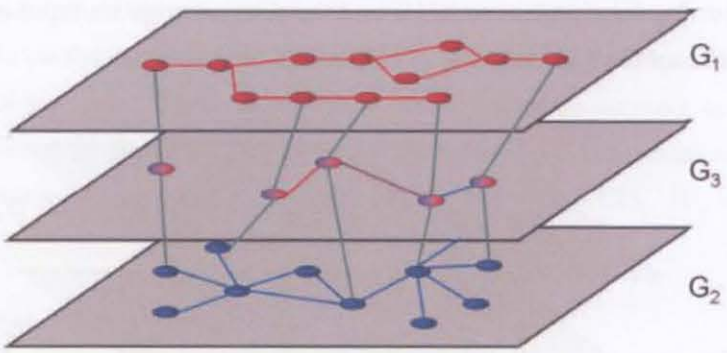


FIGURE 5.2. Three-plane representation of the two-overlapping network. Each network G_i is enclosed by a two-dimensional plane P_i. G_3 is the overlap layer. Nodes and edges of G_1 are coloured red within the top plane P_1. Nodes and edges of G_2 are coloured blue within the bottom plane P_3. Nodes in G_3 are coloured in a red/blue blend. Overlap edges in G_3 that correspond to G_1 are coloured red, those corresponds to G_2 are coloured blue, and those corresponds to both G_1 and G_3 are coloured magenta. The inter-plane edges connecting the corresponding nodes in different planes are coloured green.

Here, we defined the two representations of the two-overlapping network. In the *two-plane* representation, G_1 and G_2 are the two heterogeneous biological networks with layouts L_1 and L_2 respectively (see FIGURE 5.1). The inter-plane edge set E_{12} is added to connect nodes commonly shared between G_1 and G_2. In the *three-plane* representation, an additional layer G_3 is added (see FIGURE 5.2). This network contains nodes shared by G_1 and G_2 and edges found in either G_1 or G_2 or both. We called this the *overlap* layer within which the nodes and edges are called *overlap nodes* and *overlap edges* respectively. The inter-plane edge set E_{13} is added to connect the G_1 nodes with the overlap nodes (or G_3 nodes). Similarly, E_{23} is added to connect the G_2 nodes with the overlap nodes.

5.3.1. Two-Plane Representation

To generate the two-plane representation, the following inputs are required:

- Two networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where V_1, V_2 are the node sets and E_1, E_2 are the edge sets.
- A 1-to-1 mapping $M_V : V_{11} \leftrightarrow V_{22}$ defines the nodes common to G_1 and G_2, where $V_{11} \subseteq V_1$ and $V_{22} \subseteq V_2$.

The outputs are as the follows:

- The layouts L_1 and L_2 of G_1 and G_2 respectively, including the edge set E_{12} that connects the corresponding nodes between G_1 and G_2.

Note that G_1 and G_2 are drawn in layouts L_1 and L_2 respectively. The layouts may not be given based on the drawing convention of the specific network. For example, PINs are

usually drawn in the force-directed layout [44] whereas MNs are usually drawn in the hierarchical layout or the KEGG layout which is pre-defined [79].

5.3.2. Three-Plane Representation

To generate the three-layer representation, the following inputs are required:

- Two networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where V_1, V_2 are the node sets and E_1, E_2 are the edge sets.
- A 1-to-1 mapping $M_V : V_{11} \leftrightarrow V_{22}$ defines the common nodes between G_1 and G_2, where $V_{11} \subseteq V_1$ and $V_{22} \subseteq V_2$.

The outputs are as the follows:

- Construction of the overlap layer $G_3 = (V_3, E_3)$, where the overlap node set V_3 is defined as those nodes common to G_1 and G_2, i.e.

$$V_3 = V_1 \cap V_2$$

and the overlap edge set E_3 is defined as the edges found in G_1 and/or G_2, i.e.

$$E_3 = \{(v,v') \mid (v,v') \in E_1 \cup E_2, v,v' \in V_3\}$$

- The layouts L_1, L_2, and L_3 of G_1, G_2, and G_3 respectively, including the two edge sets E_{13} and E_{23} that connects the corresponding nodes between G_1 and G_3, and between G_2 and G_3 respectively.

5.4. Visualization of the Two-Overlapping Network

In this section, we present the drawing algorithms for the visualizations of the two-overlapping networks, based on the representations discussed in the previous section. We use the notations in the previous section for denoting the various planes, networks, nodes and edges in the drawing algorithms. All the visualizations mentioned in this chapter have one network drawn in a fixed (or given) layout. The purpose is to preserve the conventional layout which most biologists will prefer. For example, MNs are preferably drawn in the KEGG layout [79]. Therefore the layout L_1 is given by the KEGG layout if G_1 is an MN. We then draw the other network using a variation of the force-directed layout.

5.4.1. Two-Plane Visualization

The two-plane representation is visualized in a two-parallel plane layout in which networks G_1 and G_2 are drawn on separate planes P_1 and P_2. The design criteria are: (1) to achieve drawing aesthetics for G_1 and G_2, and (2) to minimize the total edge length of E_{12} between parallel planes in order to minimize occlusion in the 2.5-dimensional visualization.

The drawing algorithm involves four steps:

Algorithm 5.1. Two-parallel plane layout

1. Draw G_1 with a given layout L_1 on P_1;
2. Assign the position of node v_1 in L_1 as the initial position of its corresponding node v_2 in G_2;
3. Add the inter-plane edge set E_{12} and model each inter-plane edge as a zero-length natural spring, i.e. attraction force only. Note that such spring force does not change the inter-plane distance;
4. Draw G_2 on P_2 and the edge set E_{12} with a force-directed layout [44] with the previous initial positions;

At step 2, by assigning a good initial position based on L_1, it can help the force-directed layout of G_2 at step 4 to converge quickly. Furthermore, the corresponding nodes v_1 in L_1 and v_2 in L_2 have similar x, y-coordinates. At step 3, the zero-length natural spring for the inter-plane edges is being added in order to reduce their total edge length. Note that at step 4, this force competes with other forces in G_2 in order to produce a readable layout when equilibrium is reached.

5.4.2. Three-Plane Visualization

The three-plane representation is visualized in a three-parallel plane layout in which networks G_1, G_2 and G_3 are drawn on three parallel planes P_1, P_2 and P_3. The design criteria are: (1) to display the overlap network G_3 on a separate plane P_3, (2) to achieve drawing aesthetics for G_1, G_2 and G_3, and (3) to minimize the total edge length of E_{13} and E_{23} between the parallel planes.

Given that G_1 has a fixed layout L_1, L_2 and L_3 are being computed by taking L_1 into account. A variation of a force-directed layout [44] can be used to produce a readable layout for G_2 and G_3 while reducing the total inter-plane edge length.

The drawing algorithm contains six steps:

Algorithm 5.2. Three-parallel plane layout

1. Draw G_1 with a given layout L_1 on P_1;
2. Assign the position of node v_1 in L_1 as the initial position of its corresponding node v_2 in G_2;
3. Add inter-plane edge sets E_{13} and E_{23} and model each inter-plane edge as a zero-length natural spring (i.e. attraction force only);

4. Draw G_2 on P_2 and the edge sets E_{13} and E_{23} with a force-directed layout with the previous initial positions;
5. Assign the position of node $v_3 \in G_3$ using the barycenter of v_1 in L_1 and v_2 in L_2;
6. Draw G_3 on P_3.

The effect of steps 2 to 4 are the same as that described in the two-parallel plane layout drawing algorithm. At step 6, the barycenter of L_1 and L_2 is being used to draw G_3.

5.4.3. Implementation

The drawing algorithms were implemented as new plug-ins to GEOMI [2]. This network visualization tool has the Java3D™ package embedded, thus allowing for three-dimensional computer graphics implementation. Data for constructing the networks presented in the following case studies could be loaded into the plug-in as tab-delimited files, either as a Pajek [9] output or downloaded from a public database beforehand. The mappings required for constructing the overlap layer G_3 is computed automatically based on common node identifiers.

5.5. Case Study: Inter-connected Networks in *Escherichia coli*

5.5.1. Network Construction

5.5.1.1. Datasets

Protein interaction data. The protein interaction data for the organism was obtained from the Database of Interacting Proteins (DIP) Release July 2007 [132]. It stored protein-protein interactions as rows of protein pairs. Each protein was identified by a unique identifier, and several other identifiers such as the UniProt ID [151] are given in the file. The data contained 1846 proteins and 8013 interactions. From this, a subset containing the largest connected component consisted of 1440 nodes with 7279 edges was extracted.

Glycolytic pathway data. The data for glycolysis was obtained from the KEGG Pathway database Release 43.0 (June 2008) [79]. The data contained 52 nodes and 62 interactions. Of which, 29 nodes represented proteins (also known as enzymes) and the rest represented metabolites.

Gene regulatory interaction data. The gene regulatory data for *E. coli* was obtained from the public database RegulonDB version 6 [57] from which the largest connected component was extracted. Each row described the pairwise interaction between two proteins. The source nodes were listed on the left column and end nodes on the right column. The data contained 1371 genes and 2030 interactions.

5.5.1.2. Data mapping

Because proteins and enzymes are annotated with a common UniProt identifier [151], inter-connections between the PIN and the glycolytic pathway could be established. In total, nine enzymes from the glycolysis pathway had corresponding proteins in the PIN.

To construct an integrated two-overlapping network, the *E. coli* PIN and the glycolytic pathway were combined. In the two-plane representation, proteins common to the PIN and the MN were connected by inter-plane edges. In the three-plane representation, proteins common to the MN and the PIN were added to G_3 as new nodes. To reduce the visual complexity of the two- and the three-parallel plane visualizations, the PIN was reduced to the *1-neighbourhood network* (see definition 5.4). Finally, labels for the proteins in the interaction network were replaced by the corresponding gene name, if known.

Because UniProt identifiers are nomenclatures for proteins not genes, the inter-connections between PIN to GRN were established based on common gene identifiers [99]. A total of 160 proteins in the PIN had corresponding genes in the GRN. In the three-plane representation, proteins in the PIN having corresponding genes in the GRN were connected to new G_3 nodes. The PIN was reduced to the 1-neighbourhood for all proteins connected to the GRN.

The graph-theoretic definition of the 1-neighbourhood network was defined as the following:

Definition 5.4. *Given a network* $G = (V, E)$ *and a subset of nodes* $V' \subseteq V$. *The neighbourhood network* $G_{V'} = (V_1, E_1)$ *where* $V_1 \subseteq V$ *and* $E_1 \subseteq E$. *Also,* $v_1 \in V_1$ *and* $v' \in V'$ *such that* $V_1 = \{v_1 \mid v_1 \in V, v' \in V', (v', v_1) \in E\}$ *and* $E_1 = \{(v_1, v_1') \mid (v_1, v_1') \in E, v_1, v_1' \in V_1\}$.

5.5.2. Visualization and Analysis

Where available, the Gene Ontology [60] identifier was given in parentheses for every biological process mentioned. Similarly, the Entrez Gene [99] identifier was given in parentheses for every *E. coli* gene mentioned.

5.5.2.1. MN-PIN-overlapping network

FIGURE 5.3 showed the visualization of the MN-PIN overlapping network. Here, G_1 ($|V_1| = 52$; $|E_1| = 62$) represented the glycolytic pathway within MN (blue nodes; blue edges). G_2 ($|V_2| = 89$; $|E_2| = 85$) represented the 1-neighbourhood PIN (green nodes; green edges) which was a single connected component. G_1 was laid out using the fixed coordinates obtained from KEGG whereas G_2 was laid out using the force-directed method stated in the *three-parallel layout* algorithm (see algorithm 5.1). G_3 ($|V_3| = 8$; $|E_3| = 1$) represented the $G_1 \cap G_2$ (light blue nodes). The nodes in G_1 and G_2 were connected to their corresponding nodes in G_3 by inter-plane edges. We found in G_3 that only eight but not all glycolytic enzymes were

present in G_2 (PIN). One would expect that every enzyme should be represented in the PIN, as enzymes were really proteins. This was not the case, because protein interaction databases, including DIP, usually refer to proteins for which at least one interaction was known.

We noticed from the top view of G_1 in the visualization that G_2 contains of a high degree node (node degree > 50) that resembled a date hub (see FIGURE 5.3(a)). Using a slight rotation on the x-axis revealed that the node label of the hub is '*DIP:9040N*'. Except for the node *DIP:9040N* itself, we noticed that none of its neighbours have inter-plane edges connecting to G_1 nodes. This strongly suggested that the date hub *DIP:9040N* seen in G_2 does not interact solely with glycolytic enzymes. It could be a junction point between multiple metabolic pathways and would be an ideal target for metabolic control by gene regulation. Its topology also implied that the date hub forms a multi-protein complex (metabolon) with many other proteins.

We traversed the inter-plane edge from the date hub *DIP:9040N* in G_2 to G_1 and found that the inter-plane edge connects to the G_1 node labeled *aceF* (GeneID: 956812). This meant that the node *aceF* is the corresponding node to the date hub in G_2. A closer inspection of G_1 revealed that the node *aceF* had two outgoing blue coloured edges pointing to two other nodes. One edge points to the node labeled '*acetyl-CoA*' and the other edge points to the node labeled '*5-Acetyldihydrolipoamide*'. There was also an incoming edge originated from the node labeled '*Dihydrolipoamide*' pointing to the node *aceF*. The direction of these edges informed us that *aceF* catalyzes the reaction that converted Dihydrolipoamide to 5-Acetyldihydrolipoamide and acetyl-CoA.

We then traversed from the G_1 node *Dihydrolipoamide* to the other G_1 node labeled '*aceE, aceE1*' via three G_1 edges and two nodes. The *aceE* node had one incoming edge originating from the node labeled '*Pyruvate*' and three outgoing edge pointing to the nodes labeled '*Lipoamide*', '*2-alpha-Hydroxyethylene*', and '*Thiamine diphosphate*'. This informed us that *aceE* catalyse the reaction that converted pyruvate to lipoamide, 2-alpha-Hydroxyethylene, and thiamine diphosphate. If we followed the G_1 edges back to the node *Lipoamide*, we eventually came back to the node *acetyl-CoA*. This G_1 path is as follows:

Pyruvate→ *aceE* →*Lipoamide*→ *lpd* →*Dihydrolipoamide*→ *aceF*→ *acetyl-CoA*

From this path, we deduced that the metabolic enzyme *aceE* (GeneID: 7062918) in conjunction with *lpd* (GeneID: 944854) and *aceF* (GeneID: 956812) catalyse the reaction that converts pyruvate to acetyl-CoA. When we compared the above G_1 path with the current knowledge on glycolysis as visualized in FIGURE 5.5, we found that the G_1 path displayed in our MN-PIN-overlapping network visualization is in fact a detailed view of the *Pyruvate*→

acetyl-CoA metabolic reaction. Hence our deduction was supported by the current biological knowledge on *E. coli* metabolic network [110].

Next, we noticed that the node G_1 labeled '*pykF*' had an outgoing edge to the node *Pyruvate* and an incoming edge from the node labeled '*phospho-enol-pyruvate*'. This observation informed us that *pykF* (GeneID: 946179) catalyses the reaction that converts phospho-enol-pyruvate to pyruvate. The node *pykF* also connects to its overlap node in G_3 and then its corresponding node labeled '*DIP:10622N*' in G_2 by two inter-plane edges (FIGURE 5.3(a)). When taking the side view (FIGURE 5.3(b)), we found that node *DIP:10622N* was another hub (node degree = 7) which was less connected than its *aceF* counterpart. This was more obvious when taking the top view of G_2 (see FIGURE 5.4). This view showed that some of its neighbours were not connected to any overlap nodes in G_3 and any G_1 nodes via inter-cluster edges. Hence we deduced that, similar to *aceF*, *pykF* does not interact solely with glycolytic enzymes. Therefore, it could be a junction point between glycolysis and other metabolic pathways and likely to be subjected to metabolic control by gene regulation. As shown in FIGURE 5.5, both *aceF* and *pykF* are junction points to glycolysis, the tricarboxylic acid (TCA) cycle and the erythrose dehydrogenase (ED) pathway. Again, our deduction was supported by the current biological knowledge on *E. coli* metabolic network [110].

As the final step in our analysis, we took the oblique view in order to identify the pair of overlap nodes in G_3 that are connected to each other with an overlap edge. We identified that one node is labeled '*pykF-DIP:10622N*', the other is labeled '*ptsI*', and the overlap edge between them is coloured green. This observation informed us that *pykF* has protein-protein interaction with *ptsI*. We traversed their inter-plane edges from G_3 to their corresponding nodes in G_2 and confirmed that *ptsI* is indeed a neighbour of *pykF*. We then traverse the inter-plane edge from the overlap node *ptsI* to its corresponding node in G_1. We found in G_1 that node *ptsI* has one incoming edge from the node labeled '*D-Glucose*' and one outgoing edge to the node labeled '*alpha-D-Glucose-6-phosphate*'. This observation informed us that *ptsI* catalyzes the metabolic reaction that converts D-glucose to alpha-D-glucose-6-phosphate.

We further made the interesting observation that nodes *ptsI* and *pykF* were distant to each other in G_1 even though they were neighbours that interacted directly with each other in G_2.

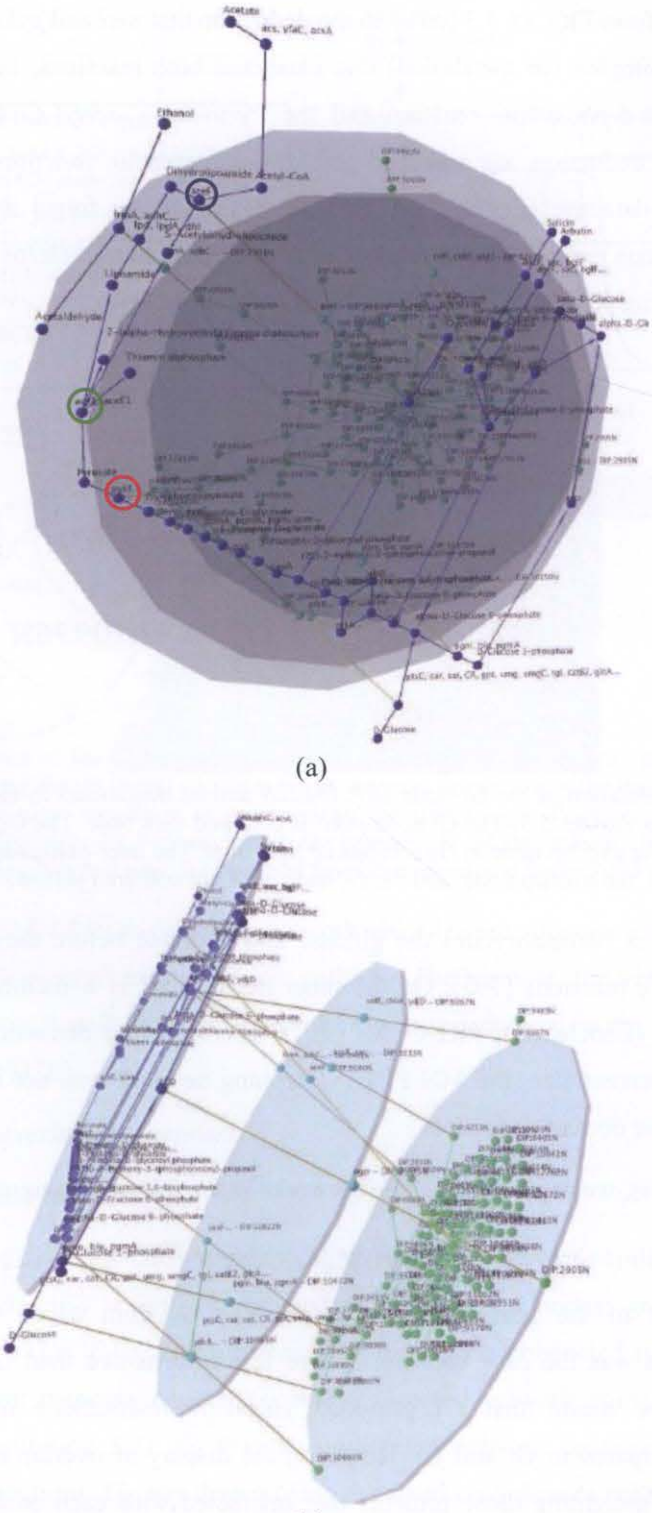


FIGURE 5.3. Visualization of the *E. coli* MN-PIN-overlapping network in the three-parallel plane layout. The G_1 network represents the MN (dark blue nodes, dark blue edges), the G_2 network represents the PIN (green nodes, green edges) and the G_3 network is the middle layer which represents the overlap layer (light blue nodes, green edge). The glycolytic pathway is presented as the MN. (a) Top view. The G_1 nodes *aceF*, *aceE*, and *pykF* are circled in blue, green, and red respectively. (b) Oblique view.

This finding from FIGURE 5.3 led us to the deduction that *ptsI* and *pykF* might be part of a larger protein complex (or metabolon) that catalyzed both reactions, i.e. the *D-Glucose→ alpha-D-Glucose-6-phosphate* reaction and the *Pyruvate→ acetyl-CoA* reaction. To find support for our deduction, we searched the known metabolic function of *pykF* from the EcoCyc public database (EcoCyc ID: PKI-COMPLEX). We found that both previously metabolic reactions require the phosphoenolpyruvate-sugar phosphotransferase system (PTS) to function. Within which, *ptsI* was one of the protein members.

FIGURE 5.4. Visualization of the G_2 node *DIP:10622N* and its neighbours in the three-parallel plane layout as shown in FIGURE 5.3. The G_1 node *pykF* is coloured dark blue. The G_2 node *DIP:10622N* is coloured green. The overlap node in G_3 is coloured light blue. The inter-plane edges connecting the G_2 node *DIP:10622N*, the overlap node, and the G_1 node *pykF* are coloured yellow.

The PTS complex phosphorylates the glucose and pyruvate before they can participate in further metabolic reactions [105]. On the other hand, *pykF* is a member of the pyruvate kinase complex (EcoCyc ID: PKI-COMPLEX). Therefore our deduction was not entirely correct. At its current size, the MN-PIN-overlapping network was not informative enough for us to make the correct deduction.

In this analysis, we explored the three networks in the following sequence:

$$G_2 \rightarrow G_1 \rightarrow G_2 \rightarrow G_3 \rightarrow G_1$$

We spent most of the analytical time on G_1 and G_2 from which we generated three deductions. This was the case because G_3 was less informative than G_1 and G_2. By '*less informative*', we meant that G_3 provided visual representations of fewer molecular interactions compared to G_1 and G_2. However, the display of overlap edges in G_3 had the advantage of highlighting those proteins that interacted with each other directly but were catalyzing different metabolic reactions in the pathway. In this way, the overlap edges in conjunction with the intra-cluster edges within G_1 and G_2 provided with us information on the spatial organization of the metabolic reactions.

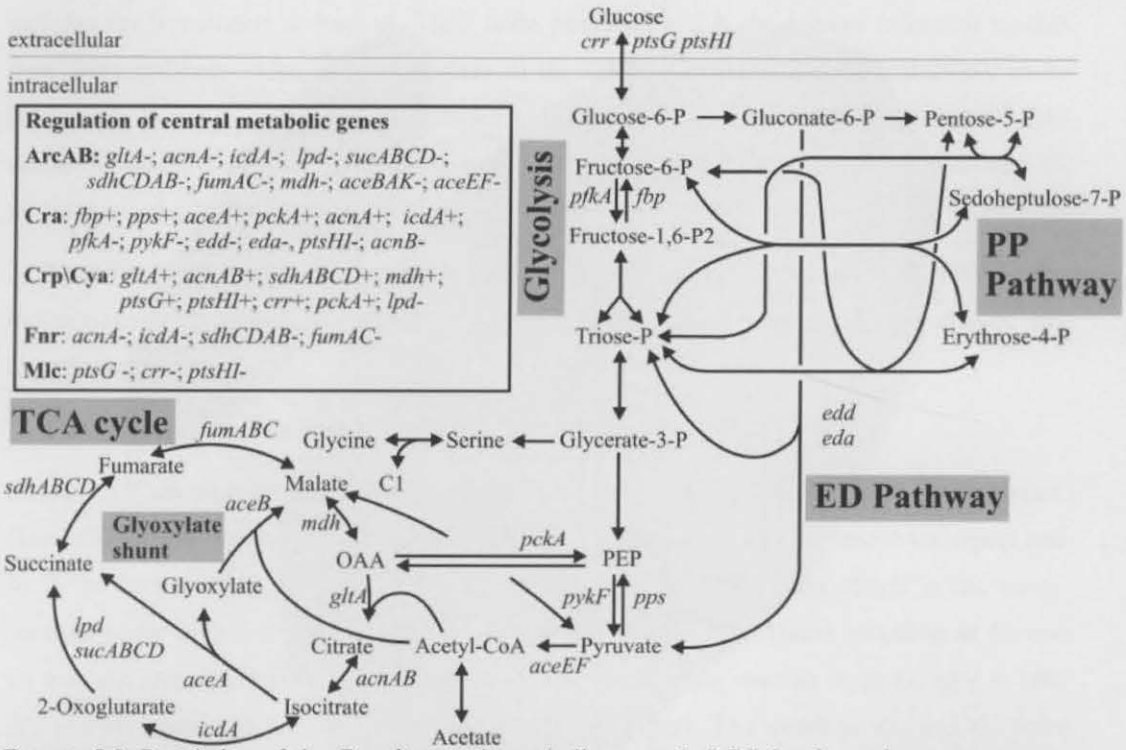


FIGURE 5.5. Regulation of the *E. coli* central metabolic network (MN) by the various master gene regulators. The gene regulators (bold) and their target enzymes are shown in the inset box at the upper left corner. Negative gene regulation is indicated by the minus sign. Positive gene regulation is indicated by the positive sign. PP Pathway stands for pentose phosphate pathway. ED Pathway stands for erythrose dehydrogenase pathway. Reproduced from Perrenoud and Sauer 2005 [110].

In summary, the MN-PIN-overlapping network visualization supports analysis relating to the role of date hub proteins in a metabolic pathway. Furthermore, the visualization was a good visual model on *E. coli* glycolysis since most of our deductions were supported by the biological literature [110].

5.5.2.2. PIN-GRN-overlapping network

I. Two-parallel plane layout

FIGURE 5.6 showed the PIN-GRN-overlapping network in the two-parallel plane layout. Here, G_1 ($|V_2|$ = 451; $|E_2|$ = 730) represented the PIN (green nodes; green edges). The network G_2 ($|V_1|$ = 1371; $|E_1|$ = 2030) represented the largest connected component of the GRN (yellow nodes; magenta edges). G_2 was a directed network containing gene-gene interactions (see definition 5.3). G_1 was an undirected network containing the physical protein-protein interactions. G_2 was drawn using the fixed co-ordinates computed according to the Kamada-Kawai layout [78] while G_1 was drawn using the force-directed method stated in the *two-parallel plane layout* drawing algorithm (see algorithm 5.1). The nodes in G_1 were directly connected to their corresponding nodes in G_2 by the inter-plane edge set E_{12} (yellow edges). The resulting visualization contained 1822 nodes and 2920 edges.

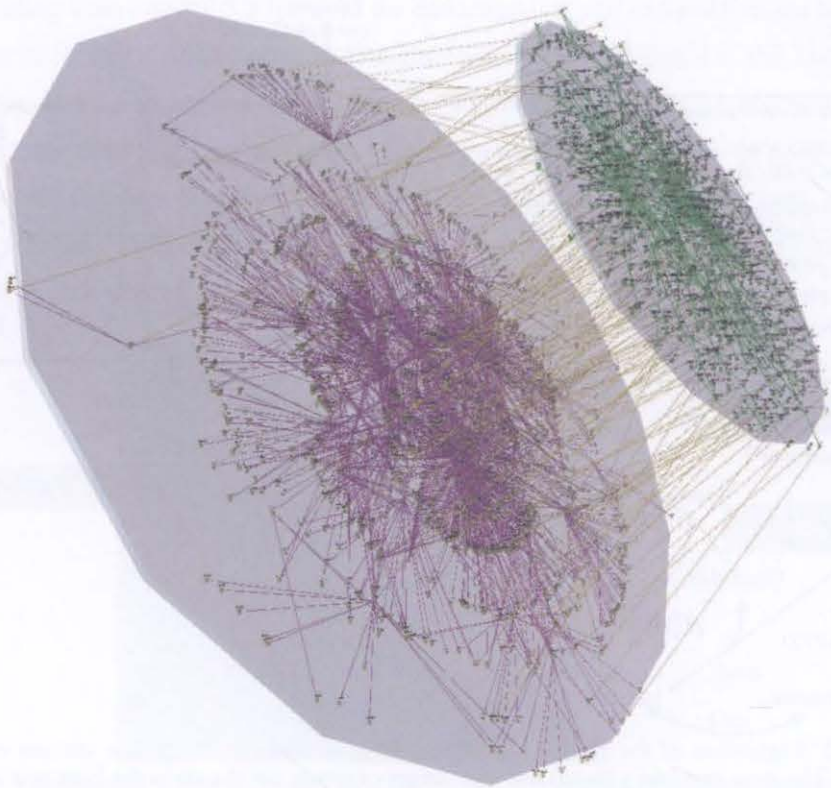


FIGURE 5.6. Visualization of the *E. coli* PIN-GRN-overlapping network in the two-parallel plane layout. The G_1 network represents the PIN (green nodes, green edges) and the G_2 network represents the GRN (magenta nodes, magenta edges). The inter-plane edge set E_{12} represents node correspondence between G_1 and G_2. The oblique view is shown here.

With a large two-overlapping network, the first task we performed was to identify any visual feature that could serve as a visual focus. Our intention was to use this visual focus as a starting point for our network exploration. Using the top view, we identified a high density area in G_1. We zoomed into the area in search for high degree nodes (node degree > 30) that resemble date hubs. Immediately, we found ourselves facing two challenges. The first challenge was to decide which high degree date hub should be our first object of investigation. This was because the hubs had similar node degrees or connectivity visually. We later decided to search for the G_1 node labeled '*aceF*' since it was found to be a high degree date hub in the MN-PIN-overlapping network visualization before. At this step, we faced our second challenge, the aggregation of high degree nodes and their node labels give rise to occlusion that hampered our effort. We therefore decided to switch to the oblique view in an attempt to locate readable inter-plane edges.

In the oblique view (see FIGURE 5.6), we identified several inter-plane edges that connected the G_1 and G_2 nodes which were located at the periphery of their respective planes. These nodes have node degrees ranging from 1 to 3. We did not take further steps to identify their node labels because they were not likely to provide us enough biological

insights for hypothesis deduction. Their node positions and node degrees informed us that they were unlikely to be gene regulators in the GRN, and they were also unlikely to be bottleneck proteins in the PIN which are subjected to tight regulation. Substantial edge cluttering made it difficult for us to identify inter-plane edges that connected the G_1 and G_2 nodes positioned near the centre of their respective planes.

Thus far, the PIN-GRN-overlapping network visualized in the two-parallel plane layout did not seem to be an effective visual analysis method. We were not able to deduce any hypotheses from it.

II. Three-parallel plane layout

FIGURE 5.7 showed the PIN-GRN-overlapping network in the three-parallel plane layout. Here, G_1 and G_2 represented the same networks as in its two-parallel plane counterpart and so is the layout of G_1. G_2 was drawn using the force-directed layout stated in the *three-parallel plane layout* drawing algorithm (see algorithm 5.2). The visual encoding of G_1 and G_2 was the same as that in the two-parallel plane layout. The overlap layer G_3 ($|V_3| = 160$; $|E_3| = 154$) represented $G_1 \cap G_2$ (red nodes; green edges). The nodes in G_1 and G_2 were connected to the overlap nodes in G_3 by the inter-plane edge sets E_{13} and E_{23} respectively (yellow edges). The resulting visualization had a total network size of 1982 nodes and 3234 edges.

The inclusion of the overlap layer G_3 increased the size of the PIN-GRN-overlapping network visualization by 16% when compared to the two-parallel plane layout. A large part of this increase was due to the additional inter-plane edges required for connecting G_1 and G_2 to G_3. As a consequence, we observed more occlusions due to node aggregation and more edge cluttering in the three-parallel plane layout. Despite this visual complexity, we found that the red coloured overlap nodes in G_3 served as a visual focus for us. These overlap nodes informed us that there were proteins commonly represented in G_1 (PIN) and G_2 (GRN) (see FIGURE 5.7). In this context, we called these proteins as *common proteins*.

Based on their positions in G_3, we located two overlap nodes whose inter-plane edges connects two G_1 nodes located in the high density region of G_1 and two G_2 nodes located near the periphery of the G_2 plane. In order to identify what the common proteins were, we zoomed into the two overlap nodes and found that one was labeled '*rpoD* - *rpoD*' and the other was labeled '*uvrD* - *uvrD*' (see FIGURE 5.8(a)). We traversed the inter-plane edges to G_2 and found that the nodes labeled '*rpoD*' and '*uvrD*' had magenta coloured incoming intra-plane edges originated from the same G_2 node (see FIGURE 5.8(a)). From this observation, we deduced that the expression of proteins from the genes *rpoD* and *uvrD* were regulated by the same gene regulator. We then traversed the inter-plane edges to G_1 and

found that the nodes labeled '*rpoD*' and '*uvrD*' were date hubs with node degrees of approximately 20 (see FIGURE 5.8(b)). Despite their proximity in G_1, we found the two hubs were not connected to each other by any edges. Therefore, we reasoned that *rpoD* and *uvrD* did not interact with each other. We had difficulty locating their neighbours by traversing G_1 edges originating from '*rpoD*' and '*uvrD*' due to the edge crossings. However, when taking the top view at a distance from G_1, the nodes *rpoD* and *uvrD* seemed to have edges connecting two other date hubs within G_1 (see FIGURE 5.8(c)).

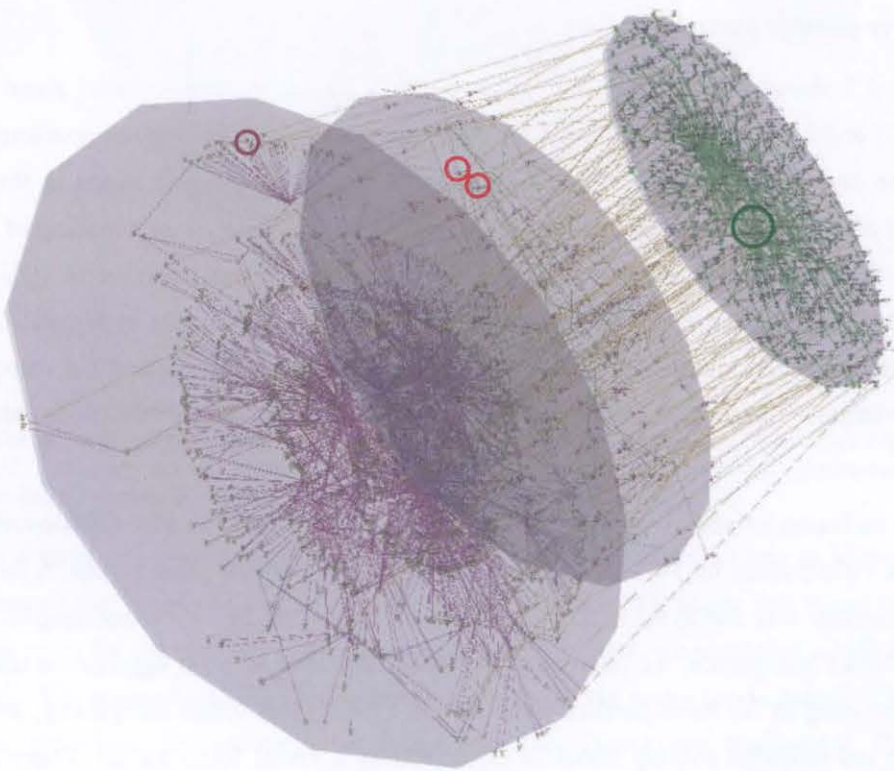


FIGURE 5.7. Visualization of the *E. coli* PIN-GRN-overlapping network in the three-parallel plane layout. The G_1 network represents the PIN (green nodes, green edges), the G_2 network represents the GRN (magenta nodes, magenta edges) and the G_3 network is the middle layer which represents the overlap layer (red nodes, green edges). The inter-plane edge sets E_{12} and E_{23} represents node correspondence between G_1, G_2 and G_3. The two overlap nodes in G_3 that connect to the high density area of G_1 are circled in red. Their neighbours in G_2 are circled in magenta. The oblique view is shown here. Also see FIGURE 5.8.

With all our observations made so far, we hypothesized that *rpoD* was a member of one protein complex and *uvrD* was a member of another. Each protein could be a bottleneck protein that held the subunits of its corresponding protein complex together. Their co-regulation by the same gene regulator implied that *rpoD* and *uvrD* might be functioning co-operatively in two related biological processes. To see whether our hypothesis was supported by the current biological knowledge, we searched the above protein labels in the public database Entrez [99] for their biological function.

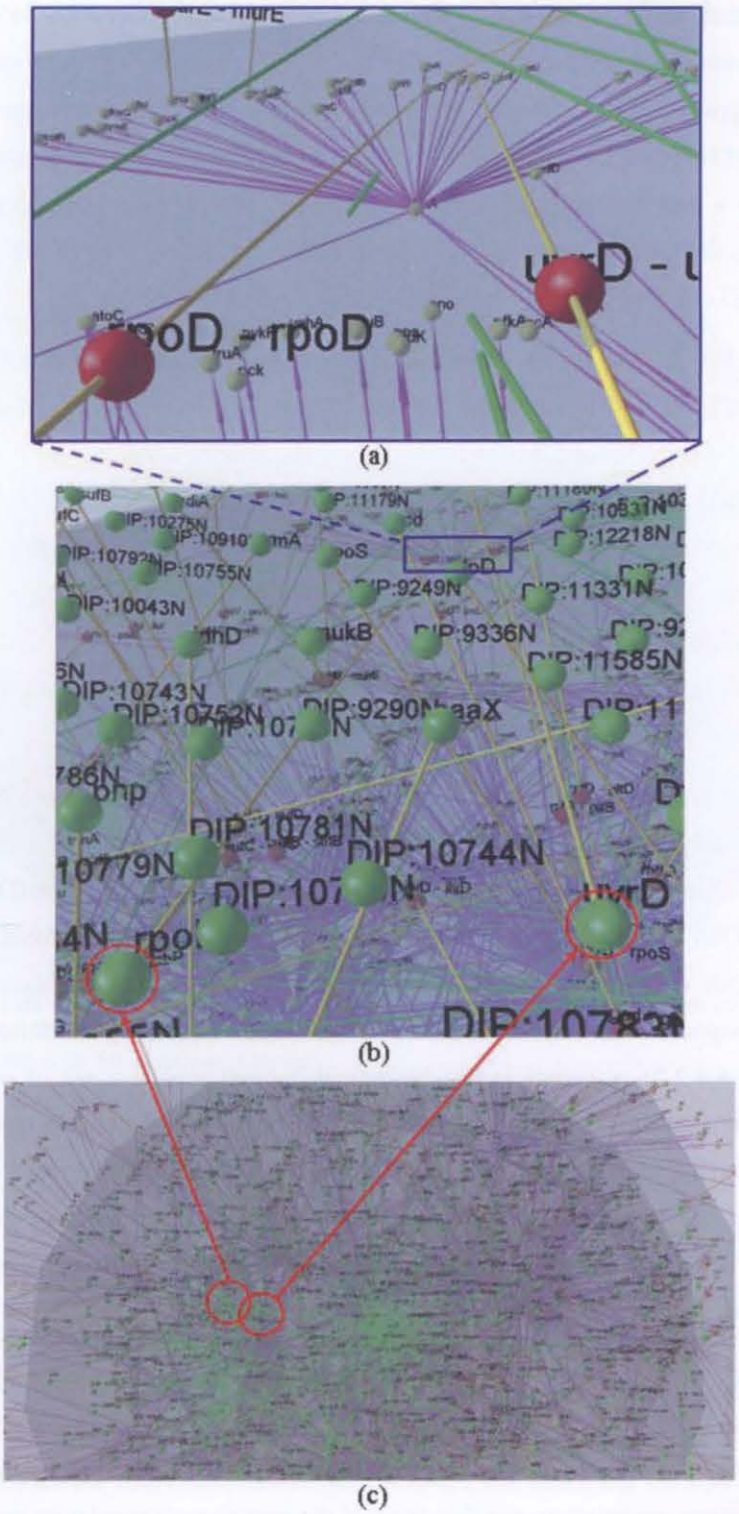


FIGURE 5.8. A fly-through sequence from the G_2 (GRN) network to the G_1 (PIN) traversing the inter-plane edges originating from the two overlap nodes *rpoD* and *uvrD*. (a) A zoom-in view of the overlap nodes *rpoD* and *uvrD* and their neighbouring nodes at G_2. (b) The inter-plane edges that connect the overlap nodes in G_3 (boxed in blue) to the G_1 nodes *rpoD* and *uvrD*. (c) Their positions in the G_1 are circled in red. This figure is derived from FIGURE 5.7.

We found that *rpoD* is an RNA polymerase sigma factor (GeneID: 4492022). It is an initiation factor for promoting the interaction of RNA polymerase with specific initiation sites on the genomic DNA. This step is essential to the initiation of gene transcription. *uvrD* (GeneID: 948347) is a DNA helicase required for the unwinding of the genome DNA double strands. It is a member of the *E. coli* pre-replicative complex. During the process of gene transcription, the unwinding of the DNA strands by *uvrD* is required before the synthesis of RNA by *rpoD* can take place.

After the above analysis, we examined G_3 again and noticed that it contained only green coloured overlap edges. This was most obvious after we hid the G_1 layer (see FIGURE 5.9). We therefore deduced that none of the common proteins represented were involved in any gene-gene interactions but only protein-protein interactions. In that case, they were unlikely to be gene regulators. Rather, the overlap nodes could be representing *effector* genes which function as the output layer of the GRN. They code for the *response* proteins which execute the biological processes required to form a biological response, e.g. increased glucose metabolism or terminal cell differentiation [38]. The G_3 nodes representing effector genes usually had one incoming G_3 edge and no outgoing G_3 edges.

We further noticed from FIGURE 5.9 that G_2 contained six high degree date hubs. They were more easily identified from the top view over G_2 (see FIGURE 5.9). The hubs were labeled '*crp*' (GeneID: 947867), '*ihfAB*' (GeneID: 6062397), '*hns*' (GeneID: 945829), '*fis*' (GeneID: 947697), '*arcA*' (GeneID: 948874), and '*fnr*' (GeneID: 945908). We deduced from their node degrees that they could be representing the master regulators of the GRN. Hence, we suspected that they formed the *kernel* of the *E. coli* GRN. As defined in Chapter 4 (see section 4.4.2.2), a kernel in the biological context means a set of master genes or proteins which 'on/off' states collectively influence the state of all other genes or proteins, thereby controlling the activity levels of multiple biological processes. Our deduction was supported by the biologist's recent finding that half of the *E. coli* genes were directly regulated by seven master regulators [101]. Note that the hub *ihfAB* is an operon containing two master regulators, *ihfA* and *ihfB*.

Of interest, we found that none of the high degree date hubs representing the master gene regulators had inter-plane edges (see FIGURE 5.10). This meant that none of the master gene regulators were represented by any overlap nodes in G_3 or any corresponding nodes in G_1. The most obvious explanation was that none of the master gene regulators needed to interact with one another or with any protein co-factors in order to mediate gene regulation. Such a biological systems design greatly reduces the amount of protein-protein interactions required for mediating gene regulation. This allows the bacterium to fine tune its functional organization rapidly in response to any external environmental challenges.

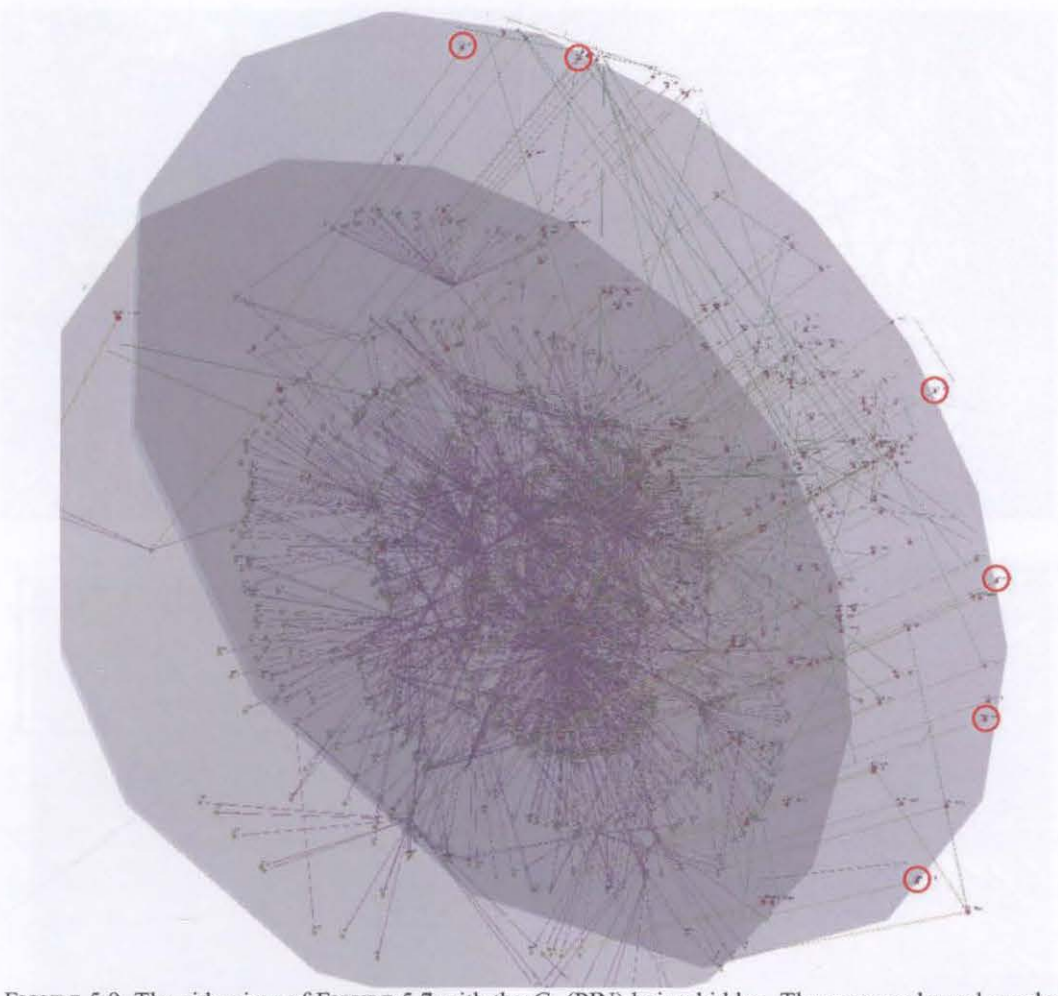


FIGURE 5.9. The side view of FIGURE 5.7 with the G_1 (PIN) being hidden. The green coloured overlap edges can be seen. The overlap nodes circled red are examples of nodes representing effector genes. These are genes that code for response proteins.

In this analysis, we begun our network exploration with the G_3 network and finished at the G_2 network. Our exploratory path through the networks could be summarized as follows:

$$G_3 \rightarrow G_2 \rightarrow G_1 \rightarrow G_3 \rightarrow G_2$$

We found that the inclusion of G_3 has its benefits for a two-overlapping network of the current size. The red-coloured overlap nodes provided a visual focus for us and served as a starting point for network exploration. The overlap nodes also helped us to identify inter-plane edges that were most likely to be biologically interesting whereas the overlap edges allowed us to deduce the functional relationship between the GRN and the PIN in *E. coli*. In this PIN-GRN-overlapping network which contained approximately 1000 nodes, we found that being able to prioritize which inter-plane edges to investigate first was crucial to our success in making biologically interesting deductions. In this case study, we demonstrated that the overlap layer G_3 served such a purpose well.

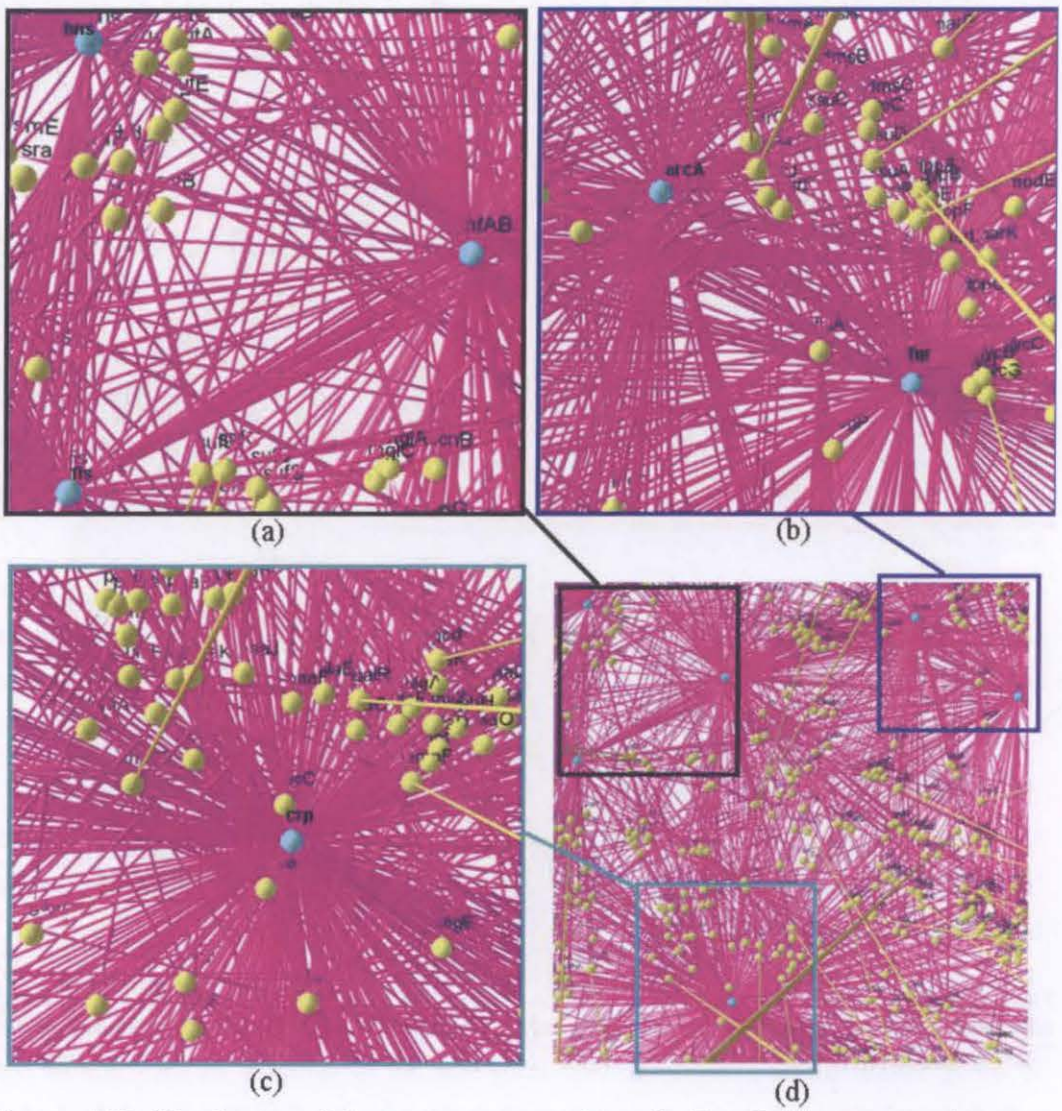


FIGURE 5.10. Visualization of the master gene regulators in *E. coli*. The gene regulators are highlighted in light blue. (a) *hns*, *ihfAB* and *fis*. (b) *arcA* and *fnr*. (c) *crp*. (d) Overview. This view is derived from the top view of FIGURE 5.9.

5.5.2.3. Conclusion

In summary, for the PIN-GRN-overlapping network visualization, the three-parallel plane layout was visually more complex than its two-parallel plane counterpart. However, the explicit visualization of the overlap layer in the three-parallel plane layout was of great assistance in the process of hypothesis deduction. For a large network, we found that the three-parallel plane layout was a more effective visual analysis method than its two-parallel plane counterpart.

5.6. Case Study: Human *TGFB1* Signaling in Hepatocellular Carcinoma

In the previous chapter, the protein *TGFB1* and its neighbours were shown in the angiogenesis (GO:0001525)-defined PIN visualizations. *TGFB1* activates the *TGFB1* signaling pathway (GO:0007179) upon interacting with its receptor protein *TGFBR1*.

TGFB1 signaling is of particular interest to cancer biologists. In normal epithelial cells, it suppresses cell survival by inducing cell cycle arrest (GO:0006917), and its inactivation contributes to oncogenesis. However, during cancer progression, *TGFB1* changes its function from being a tumour suppressor to a growth promotor in epithelial cells [154]. Cancer biologists suspected that the signaling proteins activated by *TGFB1* may be interacting with a set of proteins in cancer cells different from that in the normal cells.

To investigate the probable effect of *TGFB1* signaling on the development of HCC, we used the overlapping networks to visualize the connection between the *TGFB1* signaling pathway (*TGFB1*-STN) and the PIN in the sub-cellular organelle known as the nucleus (GO:0005634).

5.6.1. Network Construction

5.6.1.1. Datasets

Nuclear protein interaction data. The human nuclear protein interaction data was extracted from two datasets downloaded from the BioGRID [15] and the ECHO databases [74]. The ECHO database provided a list of HCC-specific proteins. A subset of the canonical human protein interaction data, in which every protein node shared the GO Component term of 'GO:0005634 nucleus', was extracted from the BioGRID data using the appropriate database transaction. The resulting dataset which contained 1748 proteins was then queried against the list of HCC-specific proteins. The final dataset contained 605 protein nodes with 787 protein interactions that were not only found in the cell nucleus but were also HCC-specific. This data came in the form of a tab-delimited file which was an output from the MySQL™ Database Management System.

***TGFB1* signal transduction data.** The *TGFB1* signal transduction interaction data was manually curated from two publications [35; 154]. This data contained 48 proteins and 46 interactions.

5.6.1.2. Data mapping

As gene symbols had been used as node labels in both datasets, inter-connections between the nuclear PIN and the *TGFB1*-STN could easily be established. A total of 20 signaling proteins had corresponding nodes in the nuclear PIN. Not all signal transduction proteins have a corresponding node to the nuclear PIN because some of them were localized within -

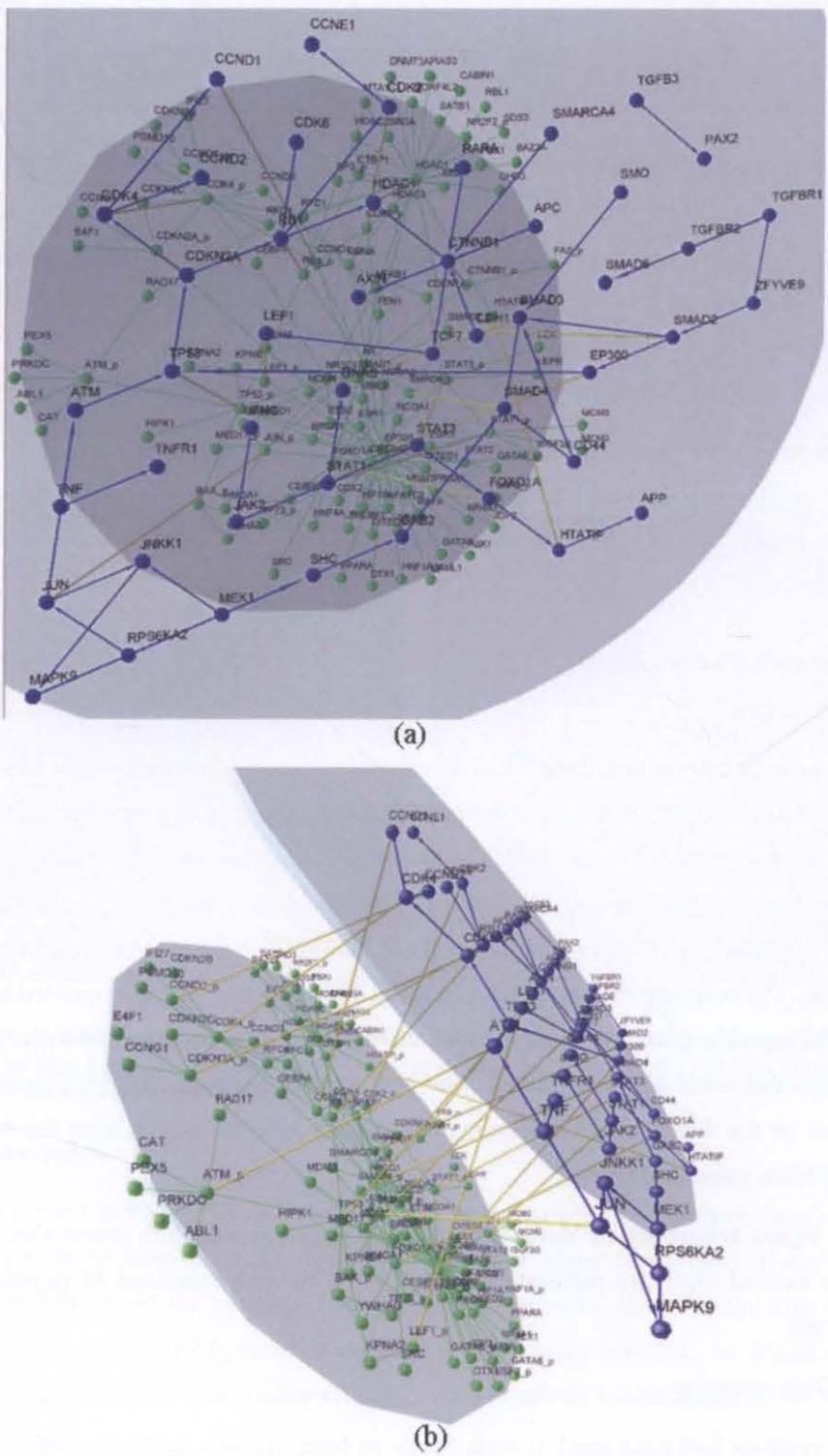


FIGURE 5.11. Visualization of the human STN-PIN-overlapping network in the two-parallel plane layout. The G_1 network represents the *TGFB1*-STN (blue nodes, blue edges) and the G_2 network represents the PIN (green nodes, green edges) in human cell nucleus. The inter-plane edge sets E_{12} represents node correspondence between G_1 and G_2 (yellow edges). (a) Top view. (b) Oblique view.

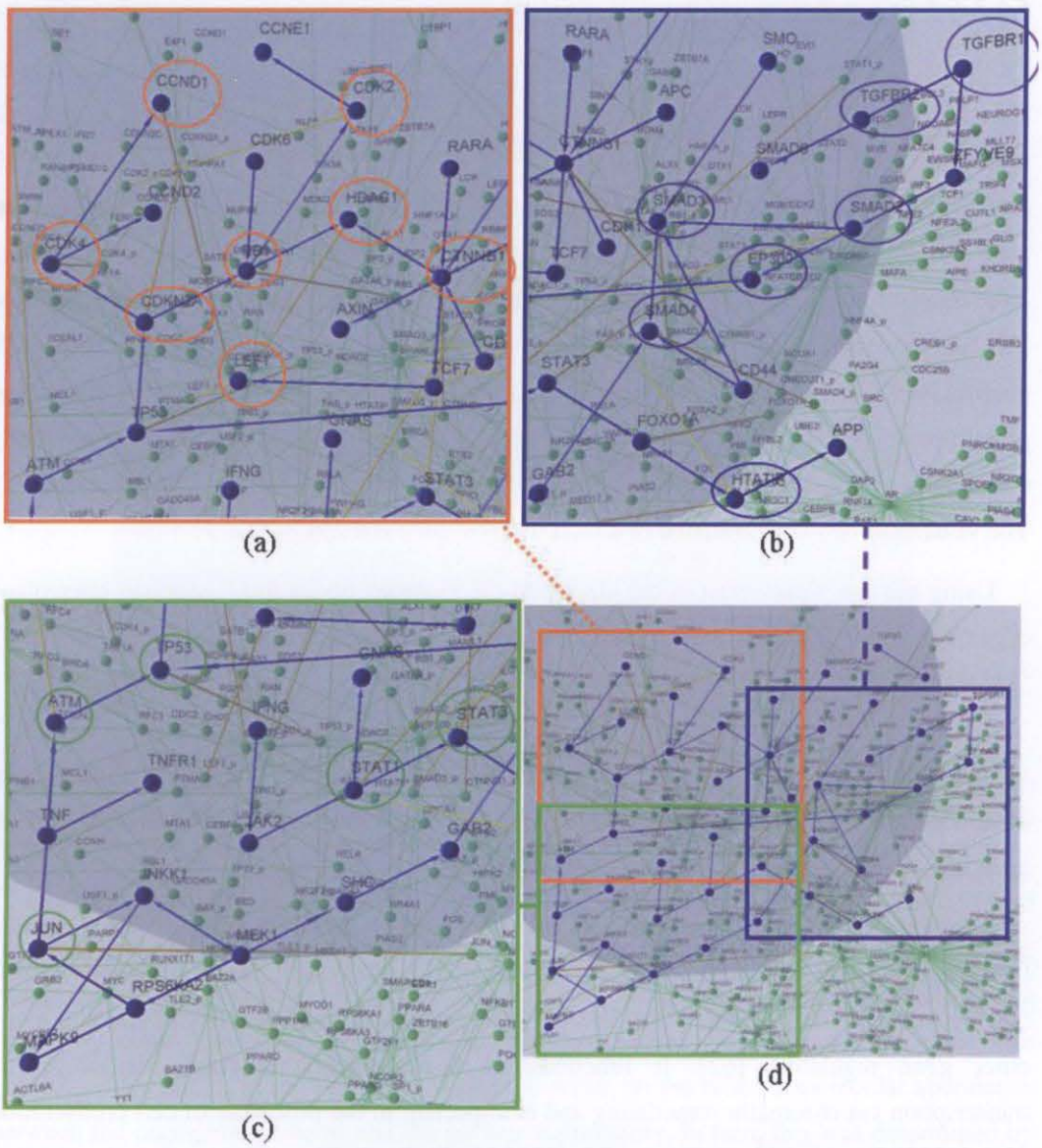


FIGURE 5.12. Zoom-in views of the G_1 network in the human STN-PIN-overlapping network shown in FIGURE 5.11. The signaling proteins discussed in section 5.6.2 are circled in ovals in different views of G_1. (a) Region that contains mostly cell cycle proteins. (b) Region that contains *TGFB1* receptors and their transducers. (c) Region that contains the *IFNG* signaling path. (d) Overview.

the cytosol and the cell membrane compartments. When reduced to the 1-neighbourhood for all proteins connected to the STN, the resulting PIN being visualized had 108 nodes and 188 edges.

5.6.2. Visualization and Analysis

Where available, the Gene Ontology [60] identifier was given in parentheses for every biological process mentioned. Similarly, the Entrez Gene [99] identifier was given in parentheses for every human gene mentioned.

5.6.2.1. Two-parallel plane layout

FIGURE 5.11 showed the STN-PIN-overlapping network in the two-parallel plane layout. Here, G_1 ($|V_1| = 48$; $|E_1| = 46$) represented the *TGFB1*-STN (blue nodes; blue edges). G_2 represented the nuclear PIN after being reduced to the 1-neighbourhood network ($|V_2| = 108$; $|E_2| = 188$). G_2 was laid out using the force-directed method stated in the *two-parallel plane* algorithm (see algorithm 5.1). G_1 was a directed network containing protein-protein interactions that served the sole purpose of propagating the phosphorylation signal from the *TGFBR1* receptor protein to the rest of the network. This happened only when the peptide growth factor *TGFB1* interacted with *TGFBR1*. G_2 was an undirected network containing protein-protein interactions that occurred in the cell nucleus. The nodes in G_1 were directly connected to their corresponding nodes in G_2 by the inter-plane edge set E_{12} (yellow edges). The visualized network consisted of a total of 156 nodes and 234 edges.

Using the top view over G_1, we identified the nineteen nodes in G_1 that had inter-plane edges connecting to their corresponding nodes in G_2. These are nodes labeled '*SMAD2*', '*SMAD3*', '*SMAD4*', '*EP300*', '*HTATIP*', '*CTNNB1*', '*FOXO1A*', '*STAT3*', '*CDK2*', '*HDAC1*', '*RB1*', '*STAT1*', '*LEF1*', '*CCND1*', '*CCND2*', '*CDK4*', '*CDKN2A*', '*TP53*', '*ATM*', and '*JUN*'. We examined the known biological function of each signaling protein documented in the Entrez public database [136] and found that they could be divided into six groups.

The six groups were (1) transducers for the *TGFBR1-TGFBR2* receptor protein dimer [154], e.g. *SMAD2* (GeneID: 4087), *SMAD3* (GeneID: 4088), and *SMAD4* (GeneID: 4089) (see FIGURE 5.12(b)); (2) *EP300* (GeneID: 2033) which is a cofactor of *SMAD*s and many other gene regulators [62]. It functions as histone acetyltransferase that regulates transcription via chromatin remodeling and is important in the processes of cell proliferation (GO:0008283) and differentiation (GO:0030154) [95]; (3) The adherens junction protein *CTNNB1* (GeneID: 1499) (see FIGURE 5.12(a)). Adherens junctions (AJs; also called the zonula adherens) are critical for the establishment and maintenance of epithelial layers, such as those lining organ surfaces; (4) The cell cycle proteins that are involved cell division, e.g. *CDKN2A* (GeneID: 1029), *CDK2* (GeneID: 1017), *CDK4* (GeneID: 1019), *CCND1* (GeneID: 595), and *CCND2* (GeneID: 894) (see FIGURE 5.12(a)); (5) The oncogenic proteins that activate the above cell cycle proteins, e.g. *JUN* (GeneID: 3725), *HDAC1*, *LEF1* (GeneID: 51176), *STAT1* (GeneID: 6772) and *STAT3* (GeneID: 6774) (see FIGURE 5.12(a) and (c)); (6) The well studied tumour suppressors, *ATM*, *RB1* (GeneID: 19645), and *TP53* (GeneID: 7157), that inactivate the cell cycle proteins by phosphorylation (see FIGURE 5.12(c)). A lesser known tumour suppressor is *HTATIP* (GeneID: 10524) is a histone acetylase that has a role in DNA repair and apoptosis (see FIGURE 5.12(a)).

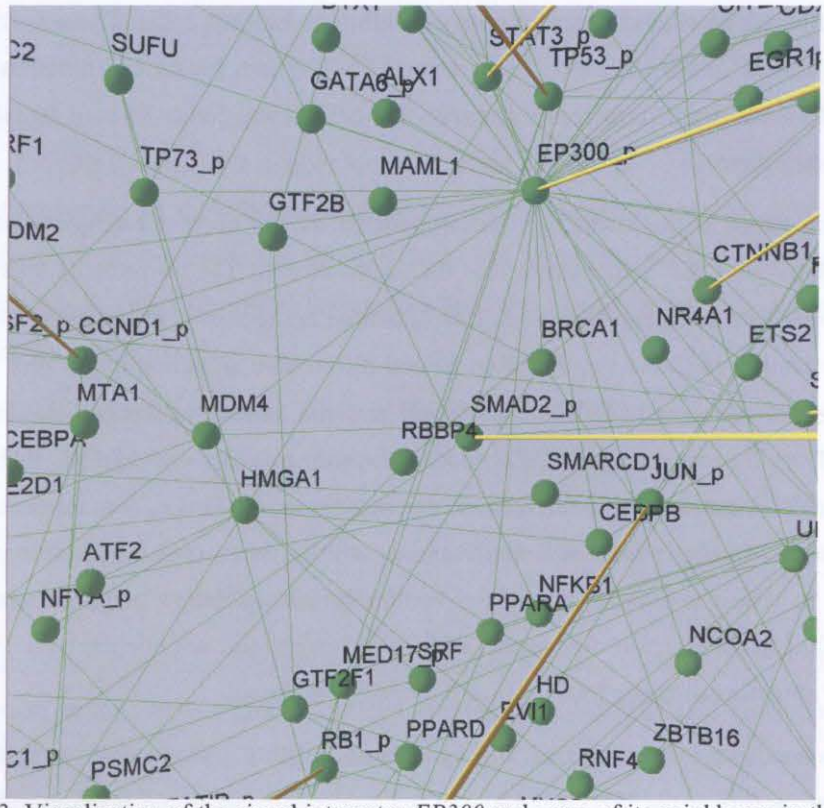


FIGURE 5.13. Visualization of the signal integrator *EP300* and some of its neighbours in the G_2 (PIN) network. Nodes with yellow edges signify that they have corresponding nodes in the G_1 (STN) network. This view is derived from two-parallel plane layout shown in FIGURE 5.11.

Out of the six groups of signaling proteins, we noticed that groups (4) and (5) had functions that promote cell growth and proliferation. Group (6) had the function of arresting cell proliferation. Their co-existence in G_1 led us to deduce that *TGFB1* signaling was a secondary control point of the cell cycle. In other words, *TGFB1* signaling alone was inadequate for initiating the cell cycle. Rather, it relied on the relative molecular abundance between the oncogenic proteins and the tumour suppressors. In turn, this was determined by the active/inactive states of other signaling proteins prior to the onset of *TGFB1* signaling. This is known as *differential signaling* [39].

We reasoned that for *differential signaling* to happen, there should be a protein that could interact with signaling proteins that had antagonistic functions like the ones in groups (5) and (6). To identify such as a protein, we traversed the inter-plane edges out of the nineteen G_1 nodes to see if any of their corresponding proteins were high degree date hubs in G_2. The result was that we identified the G_2 nodes labeled '*HDAC1*' (node degree = 23) and '*EP300*' (node degree = 35) as two such hubs. We then visually inspected the neighbours of each hub to identify if any of them had inter-plane edges originating out of them. This would indicate to us that a particular neighbour had a corresponding node in G_1 and therefore functioned as a signaling protein. We found that only the node *EP300* was connected to four other nodes

that had inter-plane edges originating out of them (see FIGURE 5.13). These G_2 nodes were labeled '*CCND1*', '*SMAD2*', '*STAT3*', and '*TP53*'. Their biological function had been mentioned previously. We therefore deduced that *EP300* was likely to be a signal integrator shared among them.

We then asked if our deduction was supported by the current biological literature. A group of biologists did demonstrate that the abundance of *EP300* proteins was limited in human fibroblasts and gene regulators that activated antagonistic activities had to compete for its availability [62]. Hence, *TGFB1*-induced differential signaling might rely heavily on the *differential affinity* [162] between *EP300* and the variety of tumour suppressors e.g. *TP53*, cell cycle proteins e.g. *CCND1*, and oncogenic proteins e.g. *STAT3*. If the last two types of proteins out-competed the tumour suppressors in HCC cells, then its cancerous state might become firmly entrenched. Differential affinity is a type of protein interaction dynamics in which a protein interacts dynamically with different neighbours at different times [162]. Its interaction frequency with each neighbour will depend on the relative molecular abundance among its various competing neighbours.

In this analysis, we navigated through the G_1 and G_2 plane following the simple exploratory path $G_1 \rightarrow G_2$. We spent most of our analytical time on G_1 because of the need to identify G_1 nodes that had inter-plane edges. This step was the most tedious in the entire analytical process. To shorten the time taken for hypothesis deduction, we used the current biological knowledge about the proteins represented by the G_1 nodes to decide which of their corresponding G_2 nodes would be of use to our hypothesis deduction. Otherwise, we would have to examine all the network paths in G_1 and then used all the inter-plane edges to G_2 to deduce the functional relationship between the edges in G_1 and the edges in G_2.

5.6.2.2. Three-parallel plane layout

FIGURE 5.14 showed the STN-PIN-overlapping network in the three-parallel plane layout. Here, G_1 ($|V_1| = 48$; $|E_1| = 46$) represented the *TGFB1*-STN (blue nodes; blue edges). G_2 represented the nuclear PIN after being reduced to the 1-neighbourhood network ($|V_2| = 108$; $|E_2| = 188$). The fixed co-ordinates for G_1 were manually assigned while G_2 was laid out using the force-directed method stated in the *three-parallel plane* algorithm (see algorithm 5.2). The overlap layer G_3 ($|V_3| = 20$; $|E_3| = 22$) represented $G_1 \cap G_2$ (red nodes; blue and green edges). The blue overlap edges represented the protein-protein interactions shared with G_1. The green overlap edges represented the protein-protein interactions shared with G_2. The nodes in G_1 and G_2 were connected to their corresponding nodes in G_3 by the inter-plane edge sets E_{13} and E_{23} respectively (yellow edges). The resulting overlapping network visualization contained a total of 176 nodes and 256 edges.

We performed the same analysis as in the two-parallel plane layout. We found that using the top view of the three-parallel plane layout (see FIGURE 5.14(a)) to identify those G_1 nodes that had inter-plane edges were more effective than using that of the two-parallel layout (see FIGURE 5.11(a)). We simply located the red colored G_3 nodes and then traversed the inter-plane edges to their corresponding G_1 nodes. However, the top view in FIGURE 5.14(a) suffered from the limitation of edge occlusion where the overlap edges in G_3 were obscured by the G_1 edges on top. We therefore used the oblique view to continue with our analysis (see FIGURE 5.14(b)).

The oblique view showed the obvious benefits of visualizing G_3. The red coloured overlap nodes informed us on the protein nodes that were common to G_1 (STN) and G_2 (PIN). The colour hues of the overlap edges informed us on the interaction types that the common proteins were involved in. Some common proteins were engaged in both protein-protein interactions and signaling interactions.

An example of a common protein was represented by the overlap node *TP53*. In G_3, it was connected to *RB1* with a green coloured overlap edge but was also connected to node *LEF1* with a blue coloured overlap edge (see FIGURE 5.15). We deduced from this observation that the *TP53* physically interacted with *RB1* in the PIN, and also engaged in a signaling interaction with the protein *LEF1* in the *TGFB1*-STN. As will be shown later, it was this visual display of overlap edges in G_3 that helped us to deduce a hypothesis on how the growth factor *TGFB1* could change from a tumour suppressor to a growth promoter in HCC cells.

We identified the overlap node *EP300* and its neighbours simply by examining G_3. Although occlusion had been observed near the *EP300* node (see FIGURE 5.14(b)), it was resolvable by z-axis rotation. In G_3, we found that the overlap node *EP300* was connected to four overlap nodes labeled '*CCND1*', '*SMAD2*', '*STAT3*', and '*TP53*' with overlap edges of different colours. The overlap edge between node *EP300* and nodes *SMAD2*, *TP53* were coloured blue. This informed us that *EP300* had signaling interactions with *SMAD2* and *TP53*. The overlap edge between node *EP300* and nodes *CCND1*, *STAT3* were coloured green. This informed us that *EP300* had protein-protein interactions with *CCND1* and *STAT3*. We therefore reasoned that *SMAD2* and *TP53* should be neighbours of *EP300* in G_1 whereas *CCND1*, *STAT3*, *SMAD2*, and *TP53* should be neighbours of *EP300* in G_2.

To confirm our reasoning, we traversed the inter-plane edges from the overlap nodes *EP300*, *CCND1*, *STAT3*, *SMAD2*, and *TP53* to G_2. We found that the G_2 node *EP300* did connect to its neighbouring nodes *CCND1*, *STAT3*, *SMAD2*, and *TP53* with green coloured edges, thus informing us on their protein-protein interactions.

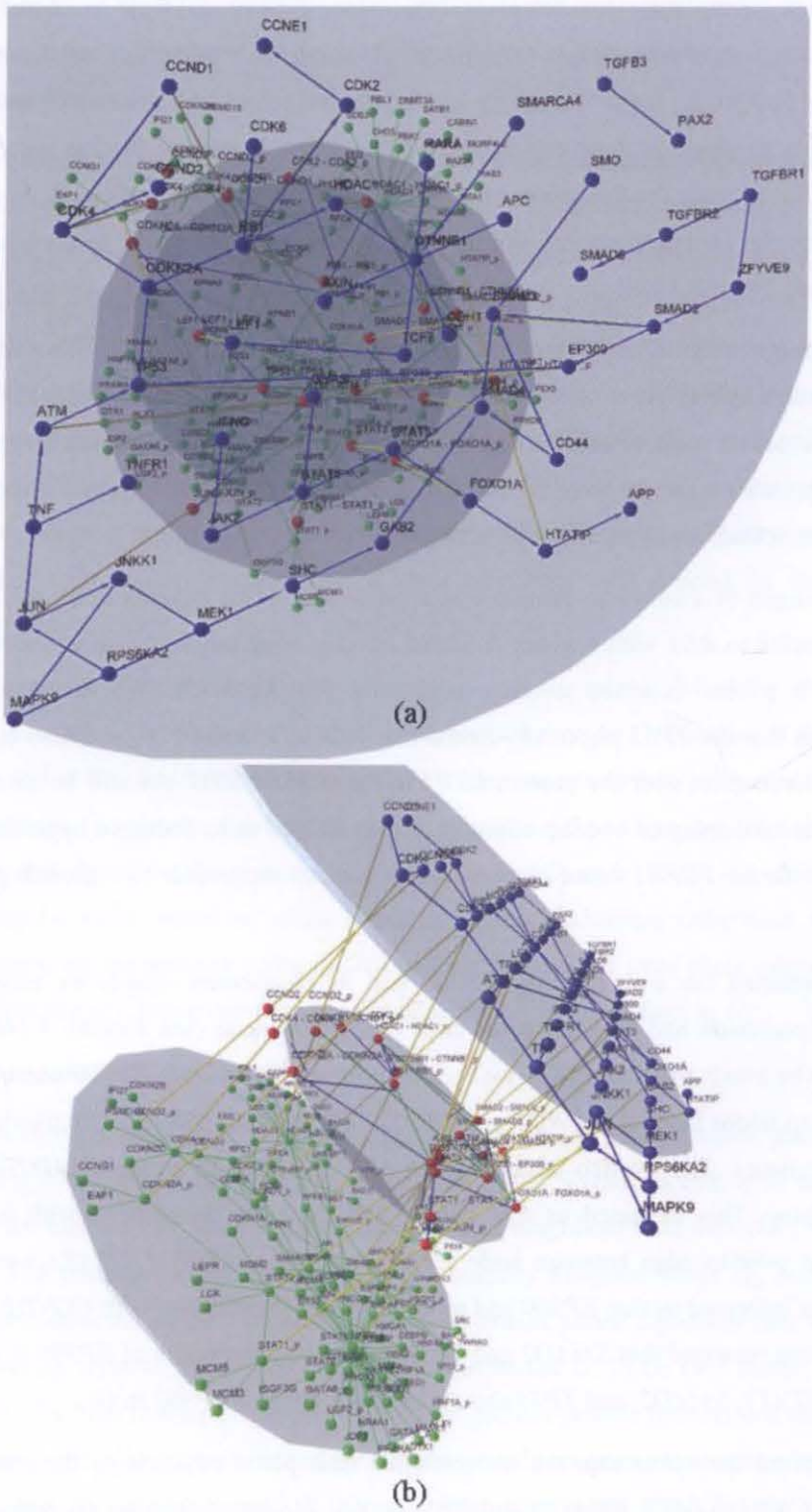


FIGURE 5.14. Visualization of the human STN-PIN-overlapping network in the three-parallel plane layout. The G_1 network represents the *TGFB1*-STN (blue nodes, blue edges), the G_2 network represents the PIN (green nodes, green edges) in human cell nucleus, and the G_3 network is the middle layer which represents the overlap layer (red nodes, blue and green edges). The inter-plane edge sets E_{12} and E_{23} represents node correspondence between G_1, G_2 and G_3. (a) Top view. (b) Oblique view.

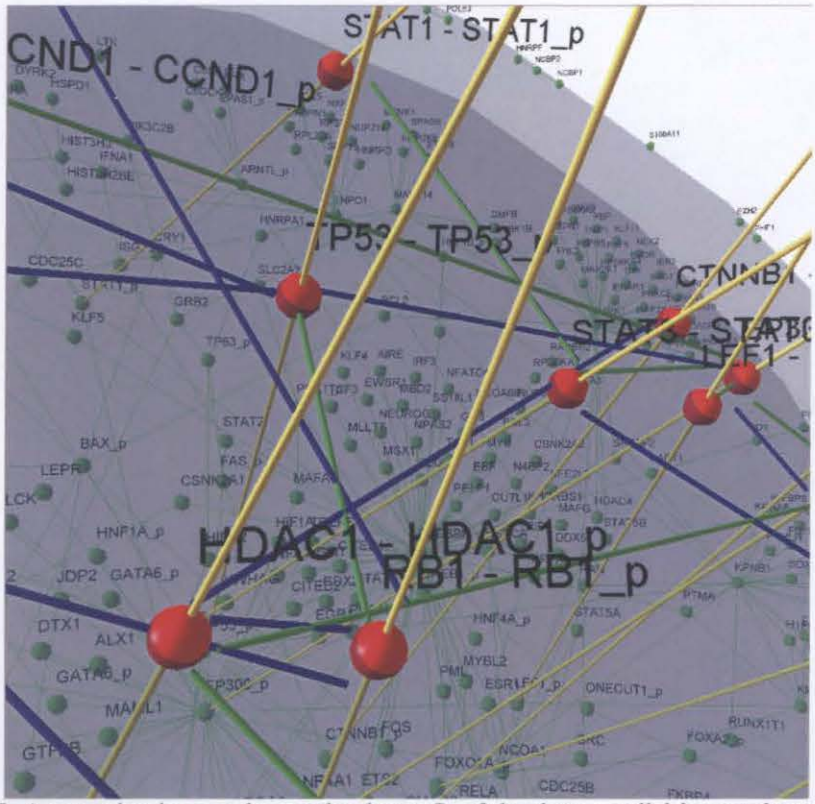


FIGURE 5.15. A zoom-in view on the overlap layer G_3 of the three-parallel layout shown in FIGURE 5.14. The red coloured overlap nodes represent proteins common to G_1 and G_2. The blue overlap edges represent signaling interactions between the overlap nodes. The green overlap edges represent protein-protein interactions between the overlap nodes. The same overlap node, e.g. *TP53*, can connect to other overlap nodes with different edge types.

We also found that the G_2 node *EP300* was a high degree hub (node degree = 35).

Apart from the previously mentioned neighbours, *EP300* was also connected to other G_2 nodes that to our knowledge had cancer related biological functions. Some of them represented tumour suppressors, e.g. *TP73* (GeneID: 7161) and *BRCA1* (GeneID: 672). Others represented oncogenic proteins, e.g. *GATA6* (GeneID: 2627) and *ETS2* (GeneID: 2114) (see FIGURE 5.16).

Next, we traversed the inter-plane edges from the overlap nodes *EP300*, *CCND1*, *STAT3*, *SMAD2*, and *TP53* in G_2 to their corresponding nodes in G_1. Here, we found that the G_1 node *EP300* had a blue-coloured incoming edge from the node *SMAD2* and a blue-coloured outgoing edge to its neighbouring node *TP53* (see FIGURE 5.14(a)). The three nodes therefore formed the following network path:

$$SMAD2 \rightarrow EP300 \rightarrow TP53$$

When we viewed G_1 from the rightmost node *TGFBR1*, we found that the above path was part of the following network path:

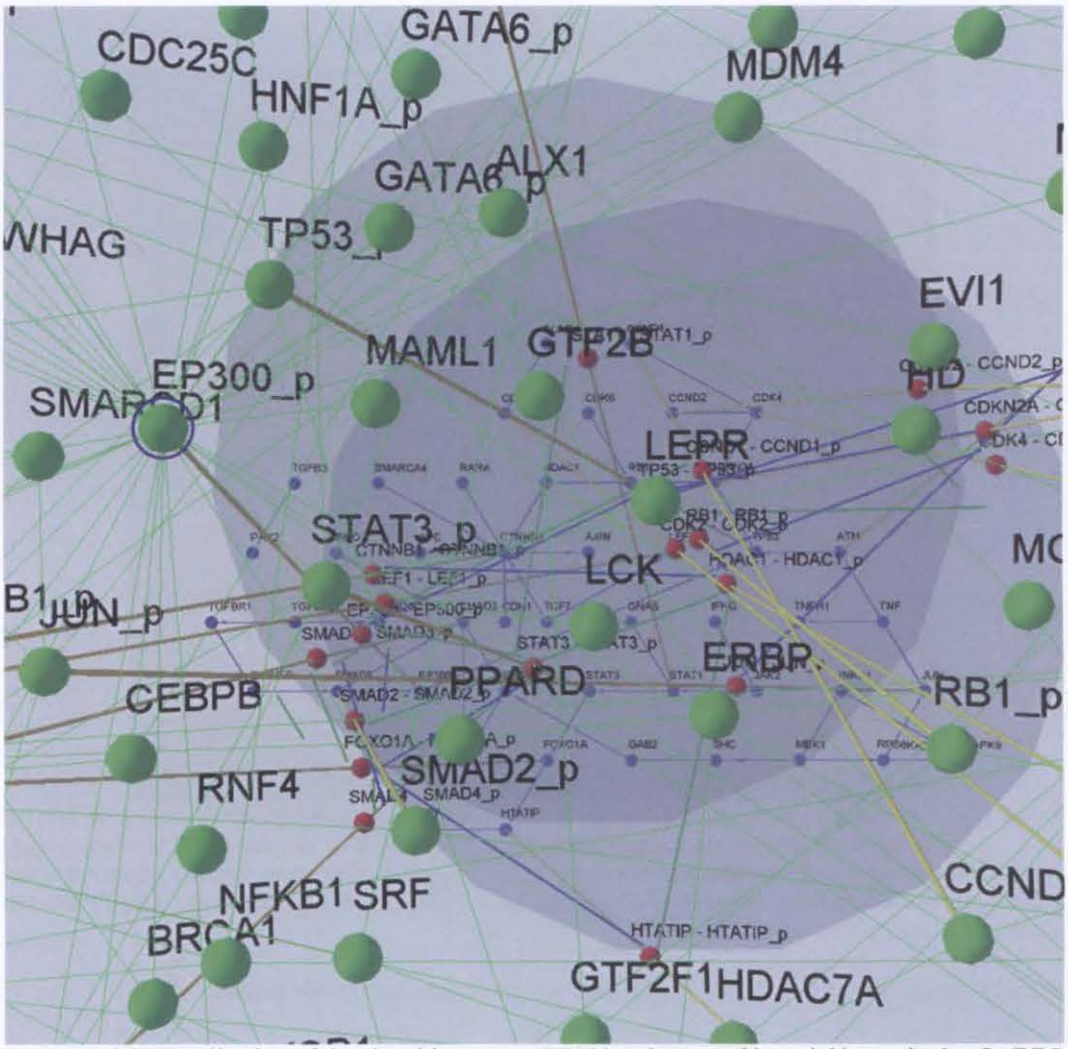


FIGURE 5.16. Visualization of the signal integrator *EP300* and some of its neighbours in the G_2 (PIN) network. The node *EP300* is circled blue. This view is derived from three-parallel plane layout shown in FIGURE 5.14.

$$TGFBR2 \rightarrow TGFBR1 \rightarrow ZFYVE9 \rightarrow SMAD2 \rightarrow EP300 \rightarrow TP53 \quad (1)$$

The biological meaning of path (1) was that, upon interacting with the growth factor *TGFB1*, *TGFBR2* phosphorylated *TGFBR1* and so forth. This phosphorylation signal was being relayed through path (1) until *TP53* was phosphorylated. From the biological literature, we found that path (1) represented the signaling path that led to *TP53* activation by phosphorylation [77]. The other two G_1 nodes, *CCND1* and *STAT3*, were found to locate in different network paths (see FIGURE 5.14(a)). The node *CCND1* was found in the following network path in G_1:

$$TP53 \rightarrow CDKN2A \rightarrow CDK4 \rightarrow CCND1 \quad (2)$$

This path represented the signaling path in the *TGFB1*-STN which led to the inactivation of the cell cycle protein *CCND1* by phosphorylation [150]. This was also the path that led to

cell cycle arrest (GO:0006917). The node *STAT3* was found in the following network path in G_1:

$$IFNG \rightarrow JAK2 \rightarrow STAT1 \rightarrow STAT3 \rightarrow FOXOA \rightarrow HTPATIP \quad (3)$$

This path represented the *IFNG* (Gene ID: 3458) signaling path in the *TGFB1*-STN which led to the inactivation of *HTPATIP* by phosphorylation.

After exploring the networks in all three planes, we attempted to construct our hypothesis on how the growth factor *TGFB1* could change from a tumour suppressor to a growth promoter in HCC cells. To achieve this objective, we needed to recall two deductions made in Chapter 4. In the analysis of the cell cycle arrest-defined PIN, we deduced that there were no functional protein-protein interactions in the cell cycle arrest (GO:0006917) biological process (see Chapter 4, section 4.4.2.1). We also found that the tumour suppressor protein *TP53* was not expressed in HCC cells. In the analysis of the angiogenesis-defined PIN, we deduced that *TGFB1* was up-regulated and co-expressed with the growth factors *VEGF* and *CTGF* in HCC cells (see Chapter 4, section 4.4.2.4).

In the event that *TP53* was not expressed and *TGFB1* was up-regulated, we deduced that the signaling path represented by network path (1) would lead to an accumulation of *EP300* proteins that are activated by phosphorylation, and path (2) would not be functioning. In other words, there would be no inactivation of *CCND1* by phosphorylation. With the absence of *TP53* in paths (1) and (2), there would be more *EP300* available for interactions with *CCND1* and *STAT3*, and more bio-active *CCND1* available for interaction with *EP300*.

According to the current medical literature, the *IFNG* signaling path represented by path (3) would also be functional in HCC cells. *IFNG* was produced by virally-infected cells. It had been known that HCC could be pre-disposed by the chronic infection of hepatitis A, B, and C viruses [20]. The expression of *STAT1* and *STAT3* had been known to be induced by *IFNG* which was important to cell growth and division [36]. Therefore, it was likely that there were more *STAT3* available for interaction with *EP300* in HCC cells. At this point, we deduced that the loss of *TP53* would lead to the concomitant increase in protein-protein interactions between *EP300* and the growth-promoting proteins *CCND1* and *STAT3*. As a result, the *TGFB1*-activated signaling path represented by network path (1) would more likely be promoting cancer cell growth rather than initiating cell cycle arrest.

To go a step further, we recalled our previous finding that *EP300* also interacted with other tumour suppressors e.g. *BRCA1*, and oncogenic proteins, e.g. *ETS2*. We hypothesized that in the course of disease progression, the progressive loss of functional tumour suppressors in HCC cells would eventually lead to the loss of differential signaling. The only

functional interactions with *EP300* would be from the oncogenic proteins. Once that happened, we predicted that the cancerous state of HCC cells would be irreversible.

In this analysis, we explored the networks in all three planes in order to develop our hypothesis on how the growth factor *TGFB1* could change from a tumour suppressor to a growth promoter in HCC cells. In the process, we explored the networks in the following sequence:

$$G_3 \rightarrow G_1 \rightarrow G_3 \rightarrow G_2 \rightarrow G_1$$

We spent most of our analytical time on deducing the interactions between overlap nodes. The visualization of G_3 was crucial to our success in hypothesis deduction. Its display of overlap edges in different colour hues effectively assisted us in deciding which set of G_1 and G_2 edges will be of use to our hypothesis deduction. Furthermore, the overlap edges not only provided spatial information on the signaling interactions but also the molecular complexity underlying the *TGFB1*-STN. By '*molecular complexity*', we meant the non-signaling type protein-protein interactions required for the *TGFB1*-STN to function properly.

5.6.2.3. Conclusion

In conclusion, we found that the STN-PIN-overlapping network visualized in the three-parallel plane layout was more efficient than its two-parallel plane counterpart as a visual analysis method and as a knowledge discovery tool. This was because the overlap layer provided with us a starting point for analyzing the functional relationship between *TGFB1*-STN and the human nuclear PIN.

5.7. Remarks

In this chapter, our case studies demonstrated that the use of the two-overlapping network visualizations as visual analysis methods. The *E. coli* case study demonstrated that they are effective visual knowledge representations since most of our deductions made in section 5.5.2 were supported by the biological literature. The human case study demonstrated that they were useful for deducing novel hypotheses.

The *E. coli* GRN-PIN-overlapping network example showed that the overlap layer G_3 in the three-parallel plane layout could serve as visual focus for the biologist. He/she could use the overlap layer as a starting point for exploring large networks. This benefit was also seen with the human STN-PIN-overlapping network even though it was of a smaller size than the *E. coli* GRN-PIN-overlapping network.

Our case studies also suggested what combinations of molecular networks were good choices for designing effective two-overlapping network visualizations. The first choice was complementary visualization of two interaction types for the same biological function or

process. The *E. coli* MN-PIN-overlapping network was such a case. The G_1 (MN) layer displayed enzyme-metabolite interactions but could not display the protein-protein interactions required for catalyzing each metabolic reaction. This limitation was compensated by having the PIN as the G_2. The second choice was visualization of a particular subset of interactions in relation to the complete set. The human STN-PIN-overlapping network was such a case. The G_1 (STN) layer displayed signaling interactions which were really a subset of the larger human nuclear PIN in G_2. The third choice was visualization of two different interaction types, each mediated a different biological function or process, but was functionally inter-dependent. The *E. coli* GRN-PIN-overlapping network was such a case. The G_1 (GRN) layer displayed the gene-gene interactions which controls the induction and repression of gene expression. This process controlled the availability of proteins for the subsequent protein-protein interactions shown in G_2.

Finally, the two-overlapping network visualization was also suitable for use as a follow-up visual analytical step to the PIN visualization discussed in Chapter 4. It provided the biologists with an integrated visualization of multiple interaction types. The different interaction types were being visualized as individual networks enclosed in two-dimensional planes, thus achieving visual separation between interaction types. If the biologist found a set of GO-defined PINs relevant to one's research question, he/she could investigate the functional relationship between any pairwise combinations using the two-overlapping network visualization.

{End of Chapter 5}

CHAPTER 6

Visualization and Analysis of Three-Overlapping Heterogeneous Biological Networks

"Everything is Connected"

6.1. Introduction

In the previous chapter, it has been demonstrated that the two-overlapping network visualization is useful as a concept model visualization and even knowledge discovery. However, the two-overlapping network visualization provides only a limited view of a single-cell molecular network. It does not provide with us an integrated view of the metabolic network (MN), gene regulatory network (GRN), and the protein interaction network (PIN) in *E. coli*, or an integrated view of the signal transduction network (MN), gene regulatory network (GRN), and the protein interaction network (PIN) in human. That is why, in this chapter, we introduced the problem of three-overlapping network visualization. Again, the networks in the visualization were *heterogeneous*. Our research problem was without doubt inspired by our understanding on biological networks, which considers a molecular network as a system of heterogeneous but inter-connected networks [3].

Good visualization of the three-overlapping networks should provide a systems-level view on the functional role of a pathway in the context of systems architecture. Such a biological question cannot be investigated simply by visualizing each network independently. The benefit of integrating three heterogeneous networks in a single visualization is the highlighting of important inter-connections between different networks, while emphasizing their different biological functionalities. As a follow-up step to the two-overlapping network visualization (see Chapter 5) and as the final analysis step in our visual analysis framework (see Chapter 1, section 1.2), we experiment with two visualization methods for the three-overlapping network representation, i.e. the *parallel plane layout* and the *circular plane layout*.

To evaluate their merit as visual analysis methods, we applied the *E. coli* and human networks to the three-overlapping networks. Visual experimentation on two different layouts, *parallel plane* and *circular plane*, was applied to each case study. The objective is to evaluate the effect of each layout on biological reasoning. For the *E .coli* case study, the objective is to evaluate the potential of the three-overlapping network as a visual knowledge

representation. The purpose of integrating the metabolic network (MN), protein interaction network (PIN), and gene regulatory network (GRN) is to provide an explicit view on the inter-connectivity among these molecular networks (see FIGURE 6.1). This should provide an overview on how the GRN in *E. coli* regulates the MN by influencing the physical organization of the PIN. We also use the *E. coli* case study to test the readability of a three-overlapping network when two of the networks were large and were highly inter-connected.

For the human case study, the objective is to evaluate the merits of the GRN-STN-PIN overlapping network in hypothesis deduction especially as a follow-up step to the two-overlapping network visualization. The domain application remains to be hepatocellular carcinoma (HCC). Because many interactions in the human GRN are yet to be discovered, some of the gene regulatory interactions visualized were projected from those found in other organisms. This provides us with ample scope for deducing new hypotheses. For this reason, the human case study is suitable for evaluating the potential of the three-overlapping network as a knowledge discovery method.

The rest of this chapter is divided into five sections. The representation of the three-overlapping network is defined in section 6.2. The drawing algorithms for the two layouts in two variations were presented in section 6.3. The *E. coli* case study is elaborated in section 6.4 followed by the human case study in section 6.5. Finally, the strength and limitations of each layout and the role of the three-overlapping network in biological analysis were discussed in section 6.6.

6.2. Representation of the Three-Overlapping Network

A three-overlapping network contains a set of three heterogeneous networks with each representing a different type of interactions. For example, the first can be a signal transduction network (STN), the second can be a PIN and the third can be a GRN. They are inter-connected because they share a subset of common nodes.

The three-overlapping network comes in two representations. In the *parallel plane* representation (see FIGURE 6.2), G_1, G_2 and G_2 were the three heterogeneous biological networks with layouts L_1, L_2 and L_3 respectively. Two inter-plane edge sets are added. The edge set E_{12} is added to connect G_1 nodes with their corresponding G_2 nodes, and the edge set E_{23} is added to connect G_2 nodes with their corresponding G_3 nodes. In the *circular plane* representation (see FIGURE 6.3), an additional inter-plane edge set E_{13} is added to connect G_1 nodes with their corresponding G_3 nodes. In either representation, overlap layers are not included as in the two-overlapping network in order to avoid added visual complexity.

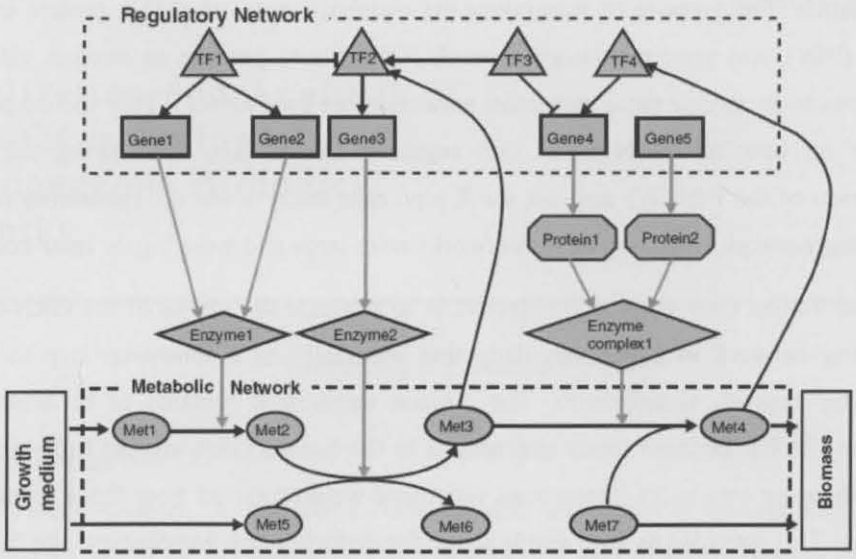


FIGURE 6.1. The schematic representation of the inter-connected metabolic, proteomic, and gene regulatory networks in *E. coli*. TF stands for transcription factor (also called gene regulators). Met stands for metabolite. Reproduced from Shlomi *et al.* 2007 [142].

6.2.1. Parallel Plane Representation

To generate the parallel plane representation, the following inputs are required:

- Three networks $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ and $G_3 = (V_3, E_3)$, where V_1, V_2, V_3 are the node sets and E_1, E_2, E_3 are the edge sets.
- A 1-to-1 mapping $M_{V1} : V_{11} \leftrightarrow V_{22}$ defines the common nodes between G_1 and G_2, where $V_{11} \subseteq V_1$ and $V_{22} \subseteq V_2$.
- A 1-to-1 mapping $M_{V2} : V_{23} \leftrightarrow V_{32}$ defines the common nodes between G_2 and G_3, where $V_{23} \subseteq V_2$ and $V_{32} \subseteq V_3$.

Thus the generated output is:

- The networks G_1, G_2, and G_3 respectively, including the two edge sets E_{12} and E_{23}. E_{12} connects the corresponding nodes between G_1 and G_2. E_{23} connects the corresponding nodes between G_2 and G_3 respectively (see FIGURE 6.2).

6.2.2. Circular Plane Representation

To generate the cyclical representation, the following inputs are required:

- Three networks $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ and $G_3 = (V_3, E_3)$, where V_1, V_2, V_3 are the node sets and E_1, E_2, E_3 are the edge sets.
- A 1-to-1 mapping $M_{V1} : V_{11} \leftrightarrow V_{22}$ defines the common nodes between G_1 and G_2, where $V_{11} \subseteq V_1$ and $V_{22} \subseteq V_2$.

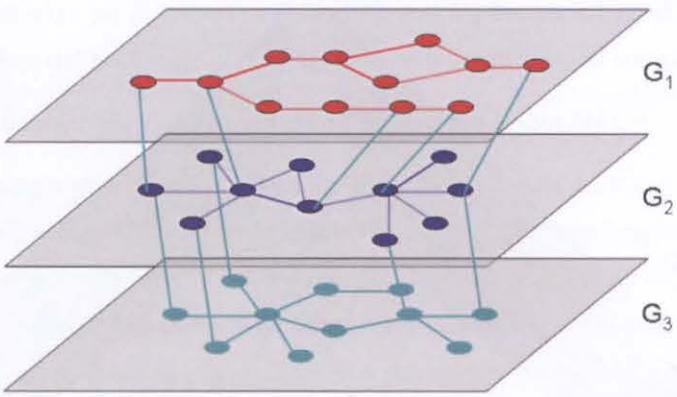


FIGURE 6.2. Parallel plane representation of the three-overlapping network. The networks G_i are drawn on three parallel planes P_i. Nodes and edges of G_1 are coloured red within the top plane P_1. Nodes and edges of G_2 are coloured blue within the middle plane P_2. Nodes and edges of G_3 are coloured green within the bottom plane P_3. The inter-plane edges connecting the corresponding nodes in different planes are coloured green.

- A 1-to-1 mapping $M_{V2} : V_{23} \leftrightarrow V_{32}$ defines the common nodes between G_2 and G_3, where $V_{23} \subseteq V_2$ and $V_{32} \subseteq V_3$.
- A 1-to-1 mapping $M_{V3} : V_{13} \leftrightarrow V_{31}$ defines the common nodes between G_1 and G_3, where $V_{13} \subseteq V_1$ and $V_{31} \subseteq V_3$.

Thus the generated output is:

- The networks G_1, G_2, and G_3 respectively, including the three edge sets E_{12}, E_{23} and E_{13}. E_{12} connects the corresponding nodes between G_1 and G_2. E_{23} connects between G_2 and G_3 respectively. E_{13} connects the corresponding nodes between G_1 and G_3 respectively (see FIGURE 6.3).

6.3. Visualization of the Three-Overlapping Network

In this section, we present algorithms for drawing the visualizations of the three-overlapping networks. In each visualization, the networks G_1, G_2 and G_3 are drawn on separate planes P_1, P_2 and P_3. Each representation mentioned in section 6.2 is being visualized in either of the two cases, i.e. *fixed-free-fixed* or *free-fixed-free*. If the layouts L_1 and L_3 are fixed (or given), the visualization is of the *fixed-free-fixed* case. If only the layout L_2 is fixed, the visualization is of the *free-fixed-free* case.

The two representations differ mainly in the arrangement of the planes in the three-dimensional space. The parallel plane layout arranges the planes in parallel within the 2.5-dimensional space. The circular plane layout arranges the planes in a triangular formation. The parallel-plane layout is more suitable for representing the three-overlapping network as a path, i.e. G_1 overlaps G_2 and G_2 overlaps G_3.

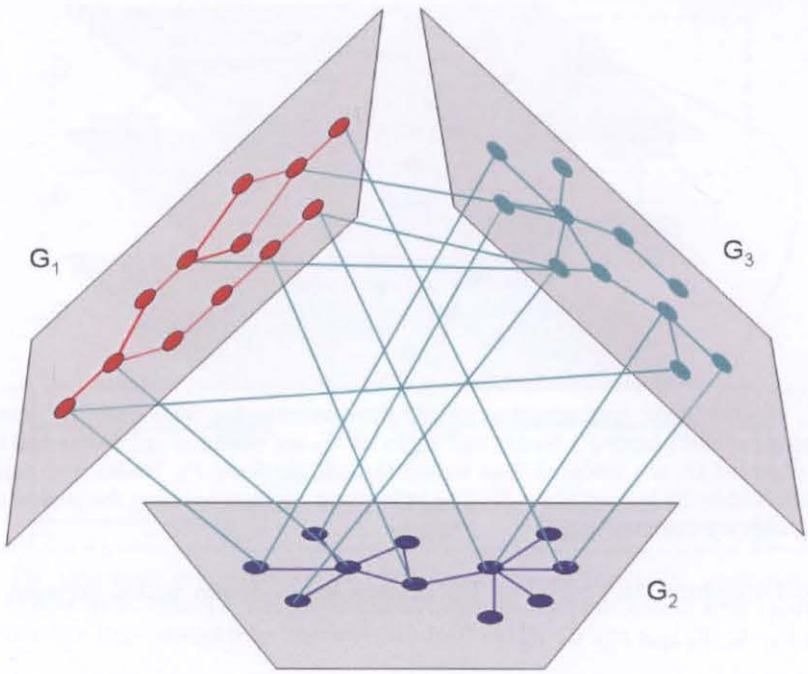


FIGURE 6.3. Circular plane representation of the three-overlapping network. The networks G_i are drawn on three planes P_i. Nodes and edges of G_1 are coloured red within the top plane P_1. Nodes and edges of G_2 are coloured blue within the middle plane P_2. Nodes and edges of G_3 are coloured green within the bottom plane P_3. The inter-plane edges connecting the corresponding nodes in different planes are coloured green.

The circular-plane layout is more suitable for representing the three-overlapping network as a cycle, i.e. G_1 overlaps G_2, G_2 overlaps G_3, and G_3 overlaps G_1.

6.3.1. Three-Parallel Plane Visualization

6.3.1.1. Fixed-free-fixed case

Given that G_1 and G_3 have fixed layouts L_1 and L_3, the layout L_2 of G_2 is being computed by taking L_1 and L_3 into account. The design criteria are: (1) to achieve drawing aesthetics for G_1, G_2 and G_3, and (2) to minimize the total edge length of E_{13} and E_{23} between parallel planes in order to minimize occlusion in the 2.5-dimensional visualization. The drawing algorithm involves six steps:

Algorithm 6.1. Fixed-free-fixed parallel layout

1. Draw G_1 with a given layout L_1 on P_1;
2. Draw G_3 with a given layout L_3 on P_3;
3. Arrange the planes P_1, P_2, and P_3 in parallel;
4. Assign the initial position of node v_2 in G_2 using the barycenter position of its mapped nodes v_1 in L_1 and v_3 in L_3;

5. Add inter-plane edge set E_{12} between P_1 and P_2, and E_{23} between P_2 and P_3 to represent mappings, and model each inter-plane edge as a zero-length natural spring;
6. Draw G_2 and the inter-plane edges using a force-directed layout [44].

At step 4, by assigning a good initial position for G_2 based on L_1 and L_3, it can help the force-directed layout of G_2 at step 6 to converge quicker. At step 5, the zero-length natural spring for the inter-plane edges are added in order to reduce the total edge length of the inter-plane edges. In step 6, this force competes with other forces in G_2 to try producing a readable layout for G_2. As a result, the inter-plane edges may not always perfectly align themselves in parallel.

6.3.1.2. Free-fixed-free Case

Given that G_2 has a fixed layout L_2, then the new layouts L_1 and L_3 of G_1 and G_3 are computed by taking L_2 into account. The design criteria applied to the fixed-free-fixed variation also apply here. The algorithm involves seven steps:

Algorithm 6.2. Free-fixed-free parallel layout

1. Draw G_2 with a given layout L_2 on P_2;
2. Assign the initial position of node v_1 in G_1 using the barycenter of its mapped nodes in L_2;
3. Add inter-plane edge set E_{12} between planes P_1 and P_2, and model each inter-plane edge as a zero-length natural spring;
4. Draw G_1 on P_1 and inter-plane edge set E_{12} using a force-directed layout [44];
5. Assign the initial position of node v_3 in G_3 using the barycenter of its mapped nodes in L_2;
6. Add inter-plane edge set E_{23} between planes P_2 and P_3, and model each inter-plane edge as a zero-length natural spring;
7. Draw G_3 on P_3 and inter-plane edge set E_{23} using a force-directed layout [44].

The purpose of steps 2 and 5 are alike. By assigning good initial positions of nodes in G_1 and G_3 based on L_2, it will help the force-directed layouts of G_1 and G_3 to converge quickly. At steps 3 and 6, the zero-length natural spring for the inter-plane edges are added in order to reduce the total edge length of the inter-plane edges. For both G_1 and G_3, the zero-length natural spring force competes with other forces to produce readable layouts (see steps 4 and 7).

6.3.2. Three-Circular Plane Visualization

6.3.2.1. Fixed-free-fixed case

Like the fixed-free-fixed case of the parallel plane layout, G_1 and G_3 have fixed layouts; then the new layout L_2 of G_2 is computed by taking into account L_1 and L_3. The design criteria are: (1) to achieve drawing aesthetics for G_1, G_2, and G_3, and (2) to minimize the total edge length of E_{12}, E_{23}, and E_{13} in order to reduce visual complexity. The drawing algorithm involves seven steps:

Algorithm 6.3. Fixed-free-fixed circular layout

1. Arrange the planes P_1, P_2, and P_3 in the circular layout (see FIGURE 6.3);
2. Draw G_1 with a given layout L_1 on P_1;
3. Draw G_3 with a given layout L_3 on P_3;
4. Add inter-plane edge set E_{13} between P_1 and P_3;
5. Assign the initial position of node v_2 in G_2 using the barycenter positions of its mapped nodes v_1 in L_1 and v_3 in L_3;
6. Add inter-plane edge set E_{12} between P_1 and P_2, and E_{23} between P_2 and P_3 to represent mappings, and model each inter-plane edge as a zero-length natural spring (i.e. attraction force only);
7. Draw G_2 and the inter-plane edges E_{12} and E_{23} using a force-directed layout [44].

At step 5, by assigning a good initial position based on L_1 and L_3, it can help the force-directed layout of G_2 at step 7 to converge quicker. At step 6, the zero-length natural spring for the inter-plane edges are added in order to reduce the total edge length of the inter-plane edges. In step 7, this force competes with other forces in G_2 to try producing a readable layout for G_2.

6.3.2.2. Free-fixed-free case

In this case, G_2 has a fixed layout L_2. Thus the new layouts L_1 and L_3 of G_1 and G_3 are being computed by taking L_2 into account. The design criteria applied to the fixed-free-fixed case (see section 6.3.2.1) also apply here. The algorithm involves nine steps:

Algorithm 6.4. Free-fixed-free circular layout

1. Arrange the planes P_1, P_2, and P_3 in the circular layout (see FIGURE 6.3);
2. Draw G_2 with a given layout L_2 on P_2;
3. Assign the initial position of node v_1 in G_1 using the position of its mapped node v_2 in L_2;

4. Add inter-plane edge set E_{12} between planes P_1 and P_2, and model each inter-plane edge as a zero-length natural spring;
5. Draw G_1 on P_1 and inter-plane edge set E_{12} using a force-directed layout [44];
6. Assign the initial position of node v_3 in G_3 using the barycenter positions of its mapped nodes in L_1 and L_2;
7. Add inter-plane edge set E_{23} between planes P_2 and P_3, and model each inter-plane edge as a zero-length natural spring;
8. Draw G_3 on P_3 and inter-plane edge set E_{23} using a force-directed layout;
9. Add inter-plane edge set E_{13} between planes P_1 and P_3.

The effect of the above steps is quite similar to that mentioned in the three-parallel plane drawing algorithm for the free-fixed-free variation (see section 6.3.1.2).

6.3.3. Implementation

The drawing algorithms are implemented as plug-ins to GEOMI [2]. Data for constructing the networks presented in the following case studies could be loaded into the plug-in as tab-delimited files, but they had to be created either as a Pajek [9] output or downloaded from a public database beforehand.

6.4. Case Study: Systems Architecture of *Escherichia coli*

6.4.1. Network Construction

6.4.1.1. Datasets

The *E. coli* MN, GRN and the PIN data applied were described in Chapter 5 (see section 5.5.1). In the MN, only the glycolytic pathway was presented.

6.4.1.2. Data mapping

To construct the three-overlapping network, the MN, GRN and the PIN networks were integrated into one. Protein nodes in the PIN that had corresponding nodes in either the glycolytic pathway or the GRN, were connected by inter-plane edges. The largest connected component of the GRN was used. To reduce the largest connected component in the PIN, only the proteins that had corresponding nodes in either the glycolytic pathway or the GRN, and their neighbours were used. The final PIN being visualized contained 514 proteins and 807 interactions. This was the *1-neighbourhood* PIN defined in Chapter 5 (see section 5.5.1.2). A total of 250 proteins in the PIN had corresponding genes in the GRN. A total of seven enzymes in the glycolytic pathway had corresponding proteins in the PIN. An enzyme was a protein specialized in catalyzing metabolic reactions (see Chapter 5, section 5.2.1).

In the circular plane representation, metabolic enzymes that had corresponding nodes in the GRN were also connected by inter-plane edges. Four glycolytic enzymes had corresponding genes in the GRN. Because the GRN had not been reduced to a 1-neighbourhood network, it allowed the mapping of any proteins that had dual functionalities to both the glycolytic pathway and the GRN. The resulting representation would help us to deduce the control points through which the GRN regulated the flow of metabolites through the glycolytic pathway and even its neighbouring pathways, e.g. amino acid biosysnthesis and tricarboxylic acid cycle.

6.4.2. Visualization and Analysis

Where available, the Gene Ontology [60] identifier was given in parentheses for every biological process mentioned. Similarly, the Entrez Gene [99] identifier was given in parentheses for every *E. coli* gene mentioned.

6.4.2.1. Parallel plane layout

FIGURE 6.4 showed the MN-PIN-GRN-overlapping network in the fixed-free-fixed parallel plane layout (see algorithm 6.1). Here G_1 ($|V_1| = 52$; $|E_1| = 62$) represented the MN (blue nodes; blue edges) which showed only the glycolytic pathway. G_2 ($|V_2| = 514$; $|E_2| = 807$) represents the 1-neighbourhood network of PIN and G_3 ($|V_3| = 1371$; $|E_3| = 2030$) represented the GRN (yellow nodes; magenta edges). The complete network consisted of 1937 nodes and 3156 edges of which 257 are inter-plane edges. The nodes common to G_1 and G_2 were connected by the inter-plane edge set E_{12} (yellow edges), and those common to G_2 and G_3 are connected by the inter-plane edge set E_{23} (yellow edges). G_1 was laid out using the fixed coordinates obtained from KEGG. G_3 was laid out using fixed coordinates from the Kamada-Kawai layout generated by Pajek [78]. Because the layouts of G_1 and G_3 were fixed, the resulting overlapping network was of the *fixed-free-fixed* case.

The parallel plane layout captured the biological concept model known as the *cascade* model [3]. The cascade model depicted a clear functional ordering of the three networks. The organization of the PIN (G_2) was subjected to regulation by GRN (G_3) which would affect the interactions in the MN (G_1). In return, the operation of MN would influence the organization of the PIN which would alter the feedback regulation from the GRN. Thus, in the cascade model, the PIN was the control target.

With the parallel plane layout, we found that only G_1 (glycolytic pathway) and the inter-plane edges between G_1 (MN) and G_2 (PIN) were readable. With 250 inter-plane edges between G_2 and G_3 (GRN), edge cluttering within the inter-plane edge set E_{23} seriously hindered their readability (see FIGURE 6.4). Therefore, large and highly inter-connected networks were not good choices for the parallel plane layout.

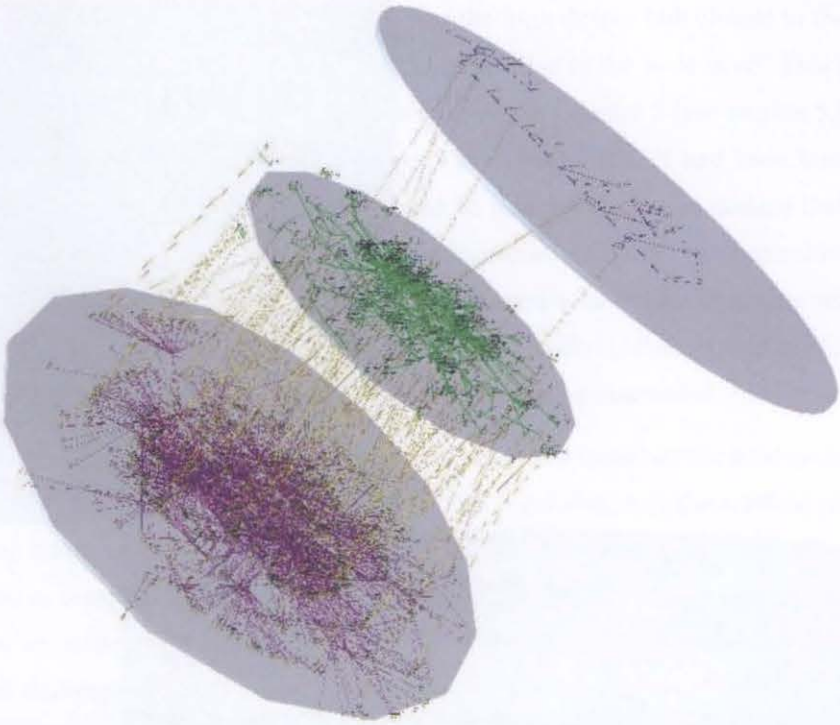


FIGURE 6.4. Visualization of the *E. coli* MN-PIN-GRN-three-overlapping network in the fixed-free-fixed parallel plane layout. The G_1 network represents the MN (dark blue nodes, dark blue edges), the G_2 network represents the PIN (green nodes, green edges), and the G_3 network represents the GRN (magenta nodes, magenta edges). The glycolytic pathway is presented as the MN.

It had been shown in the MN-PIN-overlapping network visualization (see Chapter 5, section 5.5.2.1) that the node representing the metabolic enzyme *aceF* was a high degree hub (node degree > 50). Together with another enzyme *aceE* (DIP:9039N), *aceF* catalyzed the metabolic reaction *Pyruvate→acetyl-CoA*. Because they were known to biologists as the junction point between glycolysis, the tricarboxylic acid cycle, and amino acid biosynthesis [167], we would expect *aceE* and *aceF* to be targets of one or more master gene regulators in *E. coli*. Therefore we decided to examine the connectivity of nodes labeled '*aceE*' and '*aceF*' in the three-overlapping network visualization in FIGURE 6.4. Since neither enzyme was represented in all three networks, the parallel plane layout was sufficient for the purpose of biological analysis.

We explored the parallel plane layout by drilling through the planes starting from the *aceE* nodes in G_1 (MN) and tried to identify the corresponding node for *aceE* in G_2 (PIN) by traversing the inter-plane edges (see FIGURE 6.5(a)). In G_2 (PIN), both *aceE* (DIP:9039N) and *aceF* were high degree hubs (see FIGURE 6.5(b)). We noticed that only *aceE* had a corresponding node in G_2 but not *aceF*. We also found that only G_2 node *aceF* had a corresponding node in G_3 (GRN) but not the G_2 node *aceE* (see FIGURE 6.5(c)). We therefore deduced that only the expression of *aceF* is being regulated by certain gene regulators in the GRN.

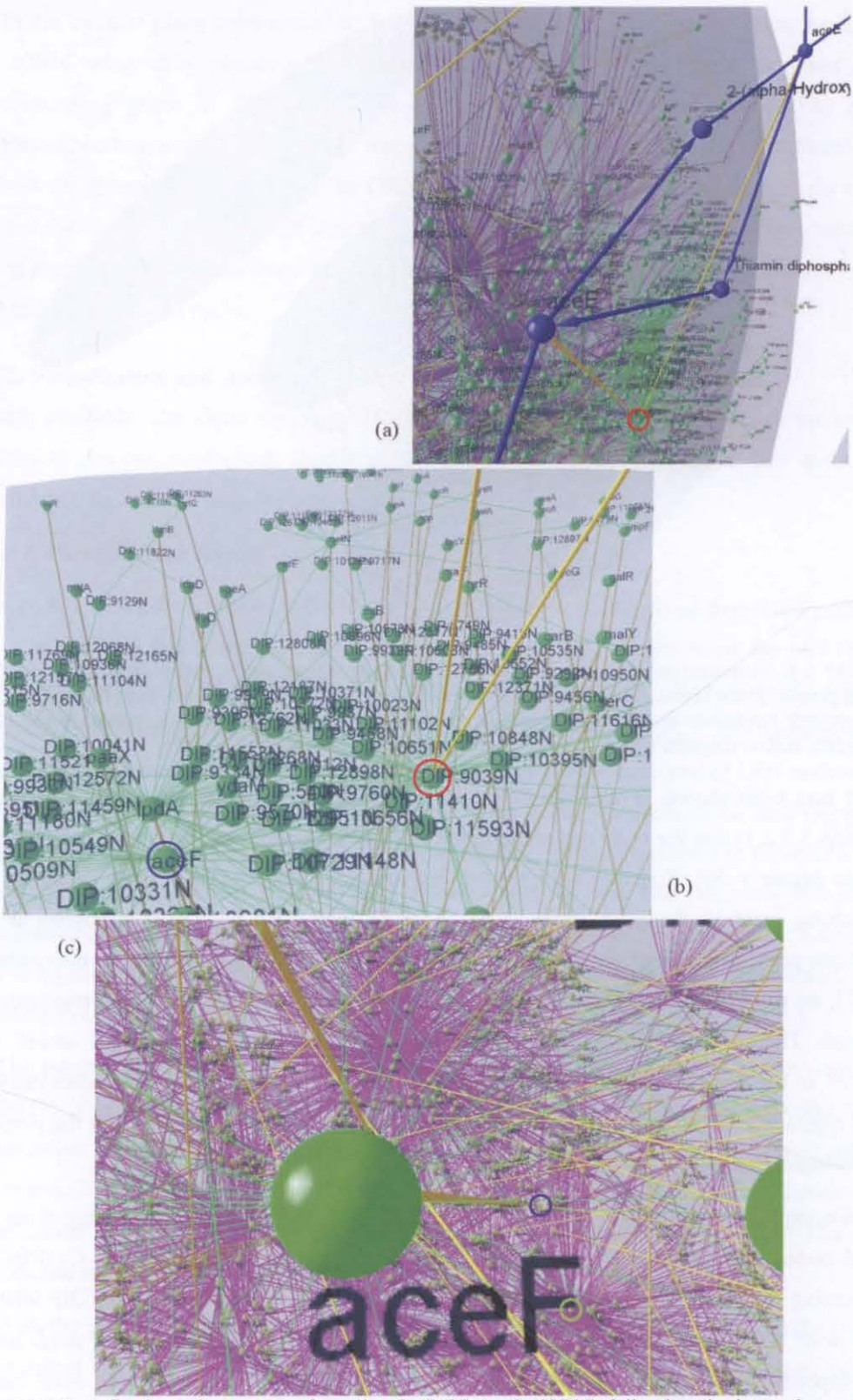


FIGURE 6.5. A fly-through sequence from the G_1 (MN) to G_3 (GRN) following the inter-plane edges originated from the glycolytic enzyme *aceE*. (a) A zoom-in view of *aceE* in the glycolytic pathway. (b) Its corresponding node DIP:9039N (circled red) and its neighbour *aceF* (circled blue) in the PIN. (c) The corresponding node of *aceF* in the GRN layer is (circled blue) and its regulator *arcA* (circled green). This series of diagrams was derived from the parallel plane layout shown in FIGURE 6.4.

In G_3, we found that the node labeled '*arcA*' was the high degree hub closest to the node *aceF*. The node *arcA* had an outgoing magenta coloured edge to the node *aceF*. This means that *aceF* was the regulatory target of *arcA*. As mentioned in Chapter 5 (see section 5.5.2.2), *arcA* was one of the seven master gene regulators in *E. coli* GRN. It had been known to biologists that *arcA* represses *aceF* [110]. It would be reasonable then to deduce that *arcA* controlled the formation of the pyruvate dehydrogenase complex by controlling the interaction between *aceE* and *aceF*. Hence, they were known as *scaffold* proteins within a protein complex [168]. The importance of our deduction was that if the *scaffold* proteins are not expressed, their corresponding protein complex would not be assembled.

Biologists had discovered recently that some enzymes in the tricarboxylic acid cycle were redundantly expressed but not *aceE* and *aceF*. [142]. By regulating only the scaffold proteins while allowing others to be redundantly expressed, the need to produce every member of a metabolic protein complex just-in-time was effectively minimized. This eventually allowed the bacterium to achieve energy efficiency while ensuring rapid responses to changing environmental challenges.

In this analysis, the three-parallel plane layout allowed us to discover the indirect regulatory relationship between the gene regulator *arcA* and the metabolic enzyme *aceE*. We achieved this by exploring the three networks in a linear sequence as the follows:

$$G_1 \rightarrow G_2 \rightarrow G_3$$

We relied on the inter-plane edges to guide us from one plane to another, and relied on the G_2 edge between the nodes *aceE* and *aceF* to visually guide us from the former node to the latter. In G_3, we relied on the node degree of the hub *arcA* and its outgoing edge to node *aceF* to visually guide us from the latter node to the former.

The small size of G_1 was important for initiating our visual analysis process. With G_2 and G_3 exceeding 500 nodes each and over 200 inter-plane edges between the two planes, the smaller and more readable G_1 provided us a starting point for network exploration. The choice of G_1 is therefore an important design consideration for the three-overlapping network in the parallel plane layout. Since our deduction made was supported by the current biological literature, the visualization was a good visual knowledge representation on the regulation of the metabolic reaction *Pyruvate→ acetyl-CoA* by the *E. coli* gene regulator *arcA*.

6.4.2.2. Circular plane layout

FIGURE 6.6 showed the MN-PIN-GRN-overlapping network in the fixed-free-fixed circular plane layout (see algorithm 6.3). The datasets used were the same as section 6.4.2.1.

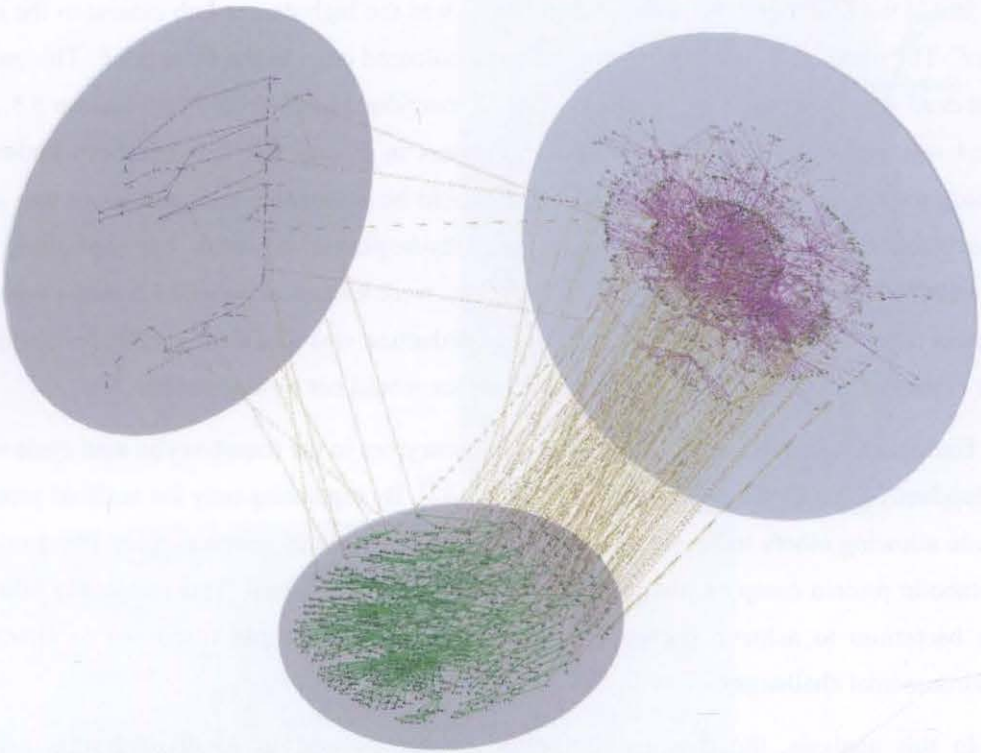


FIGURE 6.6. Visualization of the MN-PIN-GRN-three-overlapping network in the fixed-free-fixed circular plane layout. The G_1 network represents the MN (dark blue nodes, dark blue edges), the G_2 network represents the PIN (green nodes, green edges), and the G_3 network represents the GRN (magenta nodes, magenta edges). The glycolytic pathway is presented as the MN.

The networks for G_1, G_2, and G_3 are MN, PIN, and GRN respectively and layouts of G_1 and G_3 were fixed. G_1 was laid out using the fixed coordinates obtained from KEGG. G_3 was laid out using fixed coordinates from the Kamada-Kawai layout generated by Pajek [78]. The complete network consisted of 1937 nodes and 3160 edges of which 261 were inter-plane edges. The nodes common to G_1, G_2 and G_3 were connected by the inter-plane edge sets E_{12}, E_{13}, and E_{23} (yellow edges) respectively.

As shown in FIGURE 6.6, the planes P_1 and P_3 were arranged at incidence angles of $1/4\pi$ and $3/4\pi$ radians to the x-axis respectively. There was little edge cluttering seen within the inter-plane edge sets E_{12} and E_{13}, but substantial edge cluttering with the inter-plane edge set E_{23} was observed. The inter-plane edges in the circular plane layout appear longer than those in the parallel plane layout.

The circular plane layout captures another biological concept model known as the *systems* model (see FIGURE 6.1). In this model, the three networks MN, PIN, and GRN did not form a path. Rather, they formed a circle. It suggested that GRN (G_3) and PIN (G_2) can co-operatively influence the interactions in the MN (G_1). In return, MN could influence the organization of the PIN and equally influenced the regulation of the PIN via the GRN.

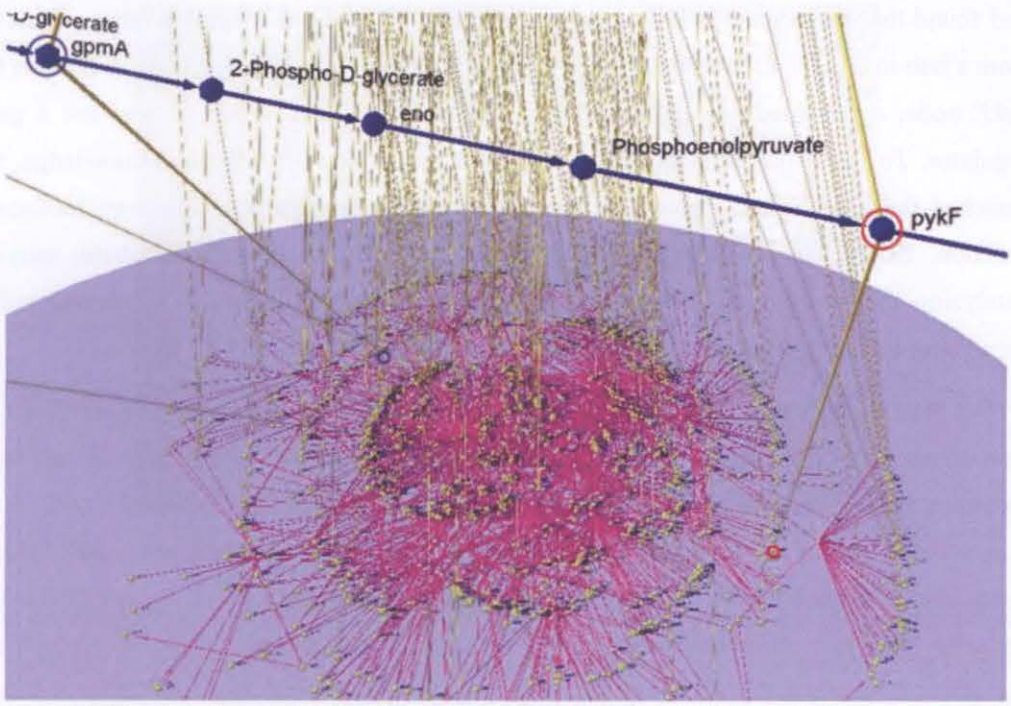


FIGURE 6.7. Visualization of the glycolytic enzymes *gpmA* and *pykF* (blue nodes) in relation to their corresponding G_3 (GRN) nodes. *pykF* in G_1 and its corresponding node in G_3 are circled red. *gpmA* in G_1 and its corresponding node in G_3 are circled blue. This view was derived from FIGURE 6.6.

The degree of influence exerted by one network on another depends how inter-connected they are. Of the three sets of inter-plane edges, E_{23} was the largest. It is the one connecting the PIN with the GRN, thus showing that the organization of the PIN was tightly regulated by the GRN. The inter-plane edge set E_{13} simply suggested that some glycolytic enzymes can be found in the PIN and they interacted with proteins that are non-glycolytic enzymes.

We made a closer examination on G_1 and found that the three nodes labeled '*pykF*', '*gpmA*' (GeneID: 945068), and '*ptsG*' (GeneID: 945651) had inter-plane edges projecting towards G_2 and G_3. Because only corresponding nodes in different planes were connected by inter-plane edges, we reasoned that these three G_1 nodes should also have corresponding nodes in both G_2 and G_3. In other words, the G_1 nodes *pykF*, *gpmA*, *ptsG* represent the *common proteins* shared by the MN, PIN, and the GRN. FIGURE 6.7 showed that the G_1 node *pykF* is connected to its corresponding node in G_3 with an inter-plane edge. We traversed this inter-plane edge to G_3, and found that the corresponding *pykF* node in G_3 is a neighbour to a hub. Since it had only one incoming edge, we deduced that the *pykF* node in G_3 represented one of the *effector* genes that formed the final output of the GRN [38] and was itself not a gene regulator. This line of reasoning was explained in the analysis of the PIN-GRN-overlapping network visualization (see Chapter 5, section 5.5.2.2).

We examined the second G_1 node *gpmA* in FIGURE 6.7 which also had an inter-plane edge connected to its corresponding node in G_3. We traversed this inter-plane edge to G_3,

and found that the corresponding *gpmA* node in G_3 also had one incoming edge originated from a hub in G_3 (see FIGURE 6.7). Therefore, we deduced that the *gpmA* node in G_3, like the *pykF* node, represented one of the *effector* genes in the GRN. Thus, it was not a gene regulator. To see if our deduction was supported by the current biological knowledge, we searched the protein label '*gpmA*' in the Entrez public database for its known biological function. Indeed, *gpmA* (GeneID: 6971780) had been known to be a metabolic enzyme catalyzing the reaction that converted 3-phospho-glycerate to 2-phospho-glycerate rather than being a gene regulator.

We then examined the third G_1 node *ptsG* in the circular plane layout. We traversed its inter-plane edge to its corresponding node in G_3 and found that the node *ptsG* in G_3 had three incoming edges from the nodes labeled '*fis*', '*crp*', and '*dgsA*' (see FIGURE 6.8). Furthermore, two of the G_3 nodes were high degree party hubs (node degree > 30). One of them, *crp*, had been known to be one of the master gene regulators in *E. coli* [101]. We deduced from this observation that *ptsG* could be an effector gene with very important biological function in *E. coli*. That was why it was being directly regulated by *crp*. It was very likely that *ptsG* was essential to multiple metabolic pathways or even non-metabolic biological processes.

We traversed the inter-plane edge originated from the node *ptsG* in G_3 to its corresponding node in G_2. In G_2 (PIN), the node *ptsG* had two neighbours labeled '*dgsA*' and '*DIP:6179N*'. This meant that the protein *ptsG* interacted with one of its gene regulators *dgsA* and another protein DIP:6179N. We searched for the accession ID '*DIP:6179N*' from the DIP public database and found that it was a record for the protein *crr* (GeneID: 3828900). We re-examined the node *ptsG* in G_1. It had one incoming edge from the G_1 node labeled '*D-glucose*' and one outgoing edge to the G_1 node labeled '*alpha-D-glucose-6-phosphate*'. This meant that *ptsG* is involved in the conversion of D-glucose to alpha-D-glucose-6-phosphate. Together, these strongly suggested that *ptsG* was both a gene regulatory co-factor and a metabolic enzyme that could be involved in multiple biological processes.

We therefore searched the biological literature and the Entrez public database for the known function(s) of protein *ptsG*. As was currently known, *ptsG*, a membrane-bound protein, together with three other cytoplasmic proteins, i.e. *ptsI* (GeneID: 3828981), *ptsH* (GeneID: 6970555), and *crr* (GeneID: 3828900) form the bacterial phosphoenolpyruvate sugar phosphor-transferase system (PTS). It was primarily involved in the concomitant phosphorylation and transmembrane glucose uptake into the cytoplasm [105]. *ptsG* was also known to interact with *mlc*, a gene regulator that mediated the glucose induction of other PTS subunits and glycolytic proteins [25]. It functioned as a *pts* operon repressor [110].

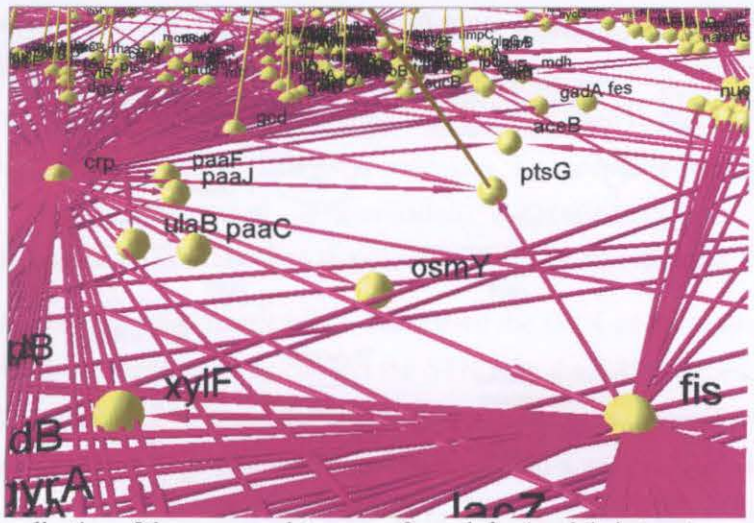


FIGURE 6.8. Visualization of the gene regulators *crp*, *fis*, and *dgsA* and their target gene *ptsG* in the G_3 (GRN) of the circular plane layout shown in FIGURE 6.6.

The dephosphorylated form of *ptsG* sequestrated *mlc* thus de-repressing the *pts* operon and allowed the transcription of PTS subunits and glycolytic enzymes to be switched on by other gene regulators (see FIGURE 6.9). Hence, *ptsG* did function as a metabolic protein and also as a gene regulatory co-factor in de-repressing the *pts* operon.

In this analysis, we used the circular mapping among the G_1, G_2, and G_3 networks to uncover proteins that were common to the MN, PIN, and the GRN. We explored all the three networks in the following sequence:

$$G_1 \rightarrow G_3 \rightarrow G_2 \rightarrow G_1$$

The most important process in our analysis was to single out the common proteins that were likely to be have the dual function of being a metabolic enzyme and a gene regulator. We achieved this by first locating the corresponding nodes in G_3 by traversing the inter-plane edges from the G_1 nodes. Following this, we deduced whether they were likely to represent gene regulators based on their intra-plane node degrees. In this way, we successfully identified *ptsG* as the glucose metabolic enzyme that also had a gene regulatory function.

6.4.2.3. Conclusion

In conclusion, the parallel plane layout of the MN-PIN-GRN-overlapping network visualization was good for vertical and sequential exploration. In the present case study, the parallel plane layout helped us to identify the indirect biological relationship between the metabolic enzyme *aceF* and the master gene regulator *arcA* in *E. coli*. On the other hand, the circular plane layout was good for cyclical exploration because the biologist could visualize the direct mappings among all the three networks.

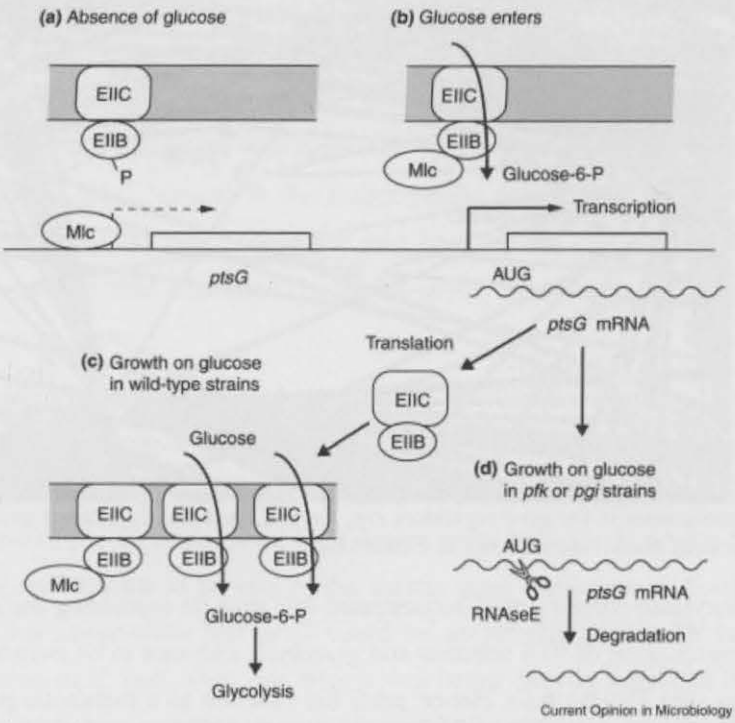


FIGURE 6.9. A schematic representation of transmembrane glucose transport regulated antagonistically by *ptsG* and *mlc*. (a) In the absence of glucose, $EIICB^{Glc}$ (*ptsG*) is present in the membrane in its phosphorylated form and the *ptsG* gene is repressed by *mlc* upstream of the promoter. (b) Glucose enters and is phosphorylated by $EIICB^{Glc}$. *Mlc* interacts with de-phosphorylated $EIICB^{Glc}$ and is sequestered away from its target operons to the membrane. *Mlc*-controlled genes are de-repressed, and *ptsG* is expressed. (c) In wild-type *E. coli* strains growing on glucose, the newly synthesized *ptsG* proteins are inserted into the inner membrane for glucose transport. *Mlc* remains attached to the predominately de-phosphorylated $EIICB^{Glc}$. The glucose-6-phosphate formed enters into the glycolytic pathway. (d) In strains unable to complete glycolysis (such as *pfkA* or *pgi* mutants), RNAseE cuts near the AUG initiation codon. The *ptsG* mRNA is then degraded and *ptsG* protein levels are not increased [85]. Reproduced from Plumbridge 2002 [114].

From the same *E. coli* networks, the circular plane layout helped us to identify the metabolic enzyme *ptsG* that had biological functions in the MN and the GRN. For the *E. coli* dataset, both visualization methods had their place in visual analysis and in biological deductions.

6.5. Case Study: microRNA Regulation of *TGFB1* Signaling

The human two-overlapping network visualization discussed in Chapter 5 (see section 5.6.2) was obviously insufficient for understanding the biology of HCC. Many of the proteins in the nuclear PIN were gene regulators that regulated the expression of many *TGFB1*-induced signaling proteins. Therefore, to gain a further understanding on how the human hepatocytic GRN regulated the nucleus PIN and influenced signaling interactions in the *TGFB1* STN, a three-overlapping network had to be used.

In contrast to *E. coli* GRN, only a few human GRN interactions had been known. Some of the GRN interactions included in the three-overlapping network visualizations were inferred from cross-species homologs but the most novel type of interactions was the

microRNA-gene interactions. Homologues are two genes from two different species that had almost identical DNA sequences and coded for proteins of the same molecular function [163]. The reason for their inclusion was because microRNAs had recently been discovered to be regulators of cancer-related biological processes, i.e. apoptosis (GO:0006915), cell development (GO:0048468), cell differentiation (GO:0030154), cell proliferation (GO:0008283), metabolism (GO:0008152), and immunity (GO:0006952) [84, 170].

If one applied the *cascade* model which considered the GRN as the output of the STN [38], the protein gene regulators that connect the STN with the GRN seemed to be heavily regulated by microRNAs [34]. There had also been suggestions that microRNAs might even co-operate with certain protein gene regulators in regulating a common set of target genes thereby allowing a co-ordinated fine-tuning of gene expression [89, 138]. The three-overlapping network described here was the first visualization that included microRNA-gene interactions as part of the GRN, allowing us to deduce novel hypotheses on the probable impact of microRNAs on HCC development.

6.5.1. Network Construction

6.5.1.1. Datasets

Nuclear protein interaction data. The canonical human protein interaction data used had been described in section 5.6.1.1.

***TGFB1* signal transduction data.** The *TGFB1* signal transduction interaction data used had been described in section 5.6.1.1.

Gene regulatory interaction data. Because known gene regulatory interactions in the human hepatocyte were few, the dataset was a combination of the TRANSPATH public database and three publications [16; 37; 104]. This data contained 108 genes and 100 interactions. Out of these, nine are non-protein coding genes which encoded for a new class of gene regulators called microRNA. These RNAs had recently been found to exert gene silencing by post-transcriptionally inhibiting the translation of messenger RNAs and later led to their degradation [153].

6.5.1.2. Data mapping

To construct the three-overlapping network, every protein in the *TGFB1*-STN was mapped to its corresponding node in the GRN or the PIN if they shared an identical gene symbol. A total of 20 *TGFB1* STN proteins had corresponding nodes in the nuclear PIN. Another nine STN proteins had corresponding nodes in the human GRN. A total of 29 nuclear proteins had corresponding nodes in the GRN. Four protein-coding genes were shared by all three networks.

6.5.2. Visualization and Analysis

Where available, the Gene Ontology [60] identifier was given in parentheses for every biological process mentioned. Similarly, the Entrez Gene [99] identifier was given in parentheses for every human gene mentioned.

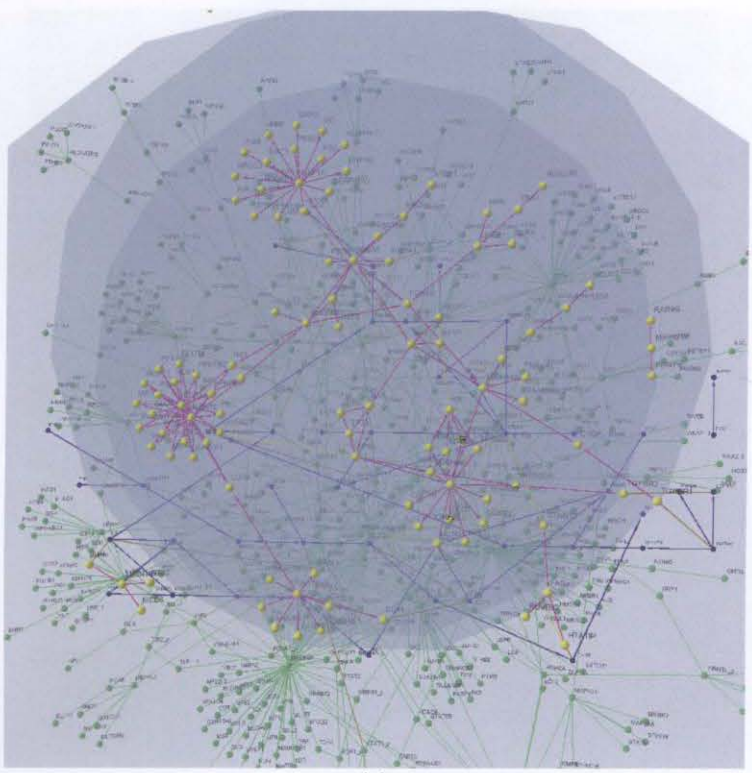
6.5.2.1. Parallel plane layout

FIGURE 6.10 showed the GRN-STN-PIN-overlapping network in the free-fixed-free parallel plane layout (see algorithm 6.2). Here, G_1 ($|V_1| = 108$; $|E_1| = 100$) represented the GRN (magenta nodes; magenta edges) and G_2 ($|V_2| = 48$; $|E_2| = 46$) represented the *TGFB1*-STN (blue nodes; blue edges). G_3 ($|V_3| = 605$; $|E_3| = 787$) represented the PIN (green nodes; green edges) within the cellular organelle known as the nucleus. The complete network consisted of 761 nodes and 962 edges of which 29 were inter-plane edges. The nodes common to G_1 and G_2 were connected by the inter-plane edge set E_{12} (yellow edges), and those common to G_2 and G_3 were connected by the inter-plane edge set E_{23} (yellow edges). Only G_2 had a fixed layout which co-ordinates were manually assigned to give a grid layout. Therefore the overlapping network was of the *free-fixed-free* case.

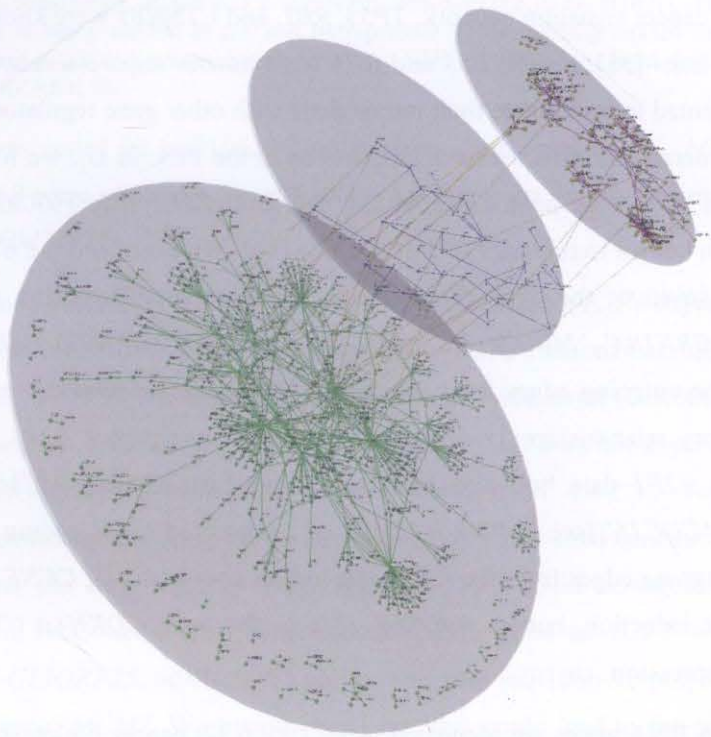
This layout captured the *signaling cascade* model which was the conventional view held by most biologists. This model depicted G_1 (GRN) as the output layer of the G_2 (STN) while some STN proteins also interacted with other nuclear proteins in the PIN. As such, this view regarded STN as the controller of the PIN and GRN.

We could explore the overlapping network using two different approaches. The first approach was by drilling through the planes starting from the top of G_1 and tried to identify corresponding nodes in each network using the inter-plane edges (see FIGURE 6.10(a)). Alternatively, we could use the oblique view (see FIGURE 6.10(b)) to identify nodes common to all three networks. This was achieved by identifying those nodes in G_2 that had inter-plane edges connected to their corresponding nodes in G_1 and G_3 (see FIGURE 6.11). With only 29 inter-edges in the visualization, we found the latter method was more effective because of the ease in identifying G_2 nodes that have inter-plane edges in E_{12} and E_{23}.

Using FIGURE 6.10, we identified five G_2 nodes that had inter-plane edges in E_{12} and E_{23}. These nodes were labeled '*TP53*', '*JUN*', '*RB1*', '*CTNNB1*' and '*HTATIP*'. Because these G_2 nodes had corresponding nodes in G_1 and G_3, we reasoned that these nodes represented the common proteins shared by the GRN, STN, and the PIN. We therefore deduced that they were the signaling proteins that were also gene regulators and also interacted with proteins in the cell nucleus. This agrees with the assertion that cancer-associated genes were enriched in nuclear proteins which formed the output layer of the signaling network [35]. The biological functions of the five common proteins were discussed in Chapter 5 (see section 5.6.2).



(a)



(b)

FIGURE 6.10. Visualization of the human GRN-STN-PIN overlapping network in the free-fixed-free parallel plane layout. The G_1 network represents the GRN (magenta nodes, magenta edges), the G_2 network represents the *TGFB1*-STN (dark blue nodes, dark blue edges), and the G_3 network represents the PIN (green nodes, green edges). The inter-plane edge sets E_{12} and E_{23} represents node correspondence between G_1, G_2 and G_3. (a) Top view. (b) Oblique view.

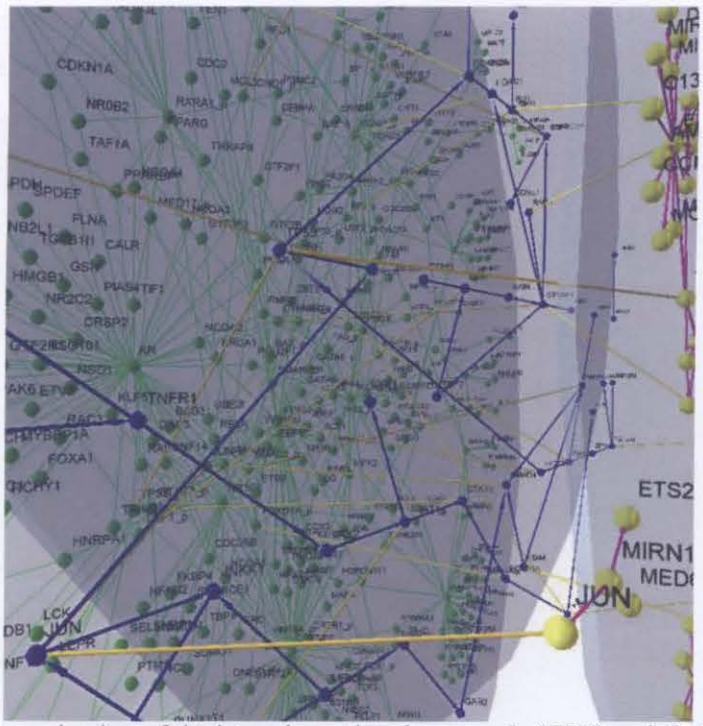


FIGURE 6.11. A zoom-in view of the inter-plane edges between G_1 (GRN) and G_2 (STN) in FIGURE 6.10.

In the human cancer signaling network, *TP53*, *RB1*, and *CTNNB1* were known to be three of the signaling hubs [35]. Since *TP53* and *RB1* were tumour suppressors well studied by biologists, we wanted to investigate their interactions with other gene regulators in the GRN and how these interactions affected their interactions in the PIN. In G_1, we found that both nodes *TP53* and *RB1* were neighbours of the date hub *E2F1* (see FIGURE 6.12). Both proteins had been known to repress *E2F1* expression [16]. We examined the *E2F1* party hub and found that seven of the 16 nodes represent RNA genes. They were nodes labeled '*C13ORF25*', '*MIRN18A*', '*MIRN20A*', '*MIRN25*', '*MIRN92*', '*MIRN93*', '*MIRN106A*', and '*MIRN106B*'. The outgoing edges from the node *E2F1* to the above nodes represented the positive regulatory relationship. This meant that the gene regulator *E2F1* induced their expression. The *E2F1* date hub also had four neighbours labeled '*MCM5*', '*CCNE1*', '*CDKN1A*' and '*CDC16*' (see FIGURE 6.12). They represented genes coding for cell cycle proteins. The outgoing edges from the *E2F1* date hub to nodes *MCM5*, *CCNE1*, and *CDC16* represented their induction, but its outgoing edge to the node *CDKN1A* (GeneID: 1026) represented its repression.

Of the sixteen RNA genes we identified in FIGURE 6.12, cancer biologists had recently found that the RNA genes *MIRN20A*, *MIRN25*, *MIRN92* and *MIRN106A* were expressed in different types of cancer cells [155]. Of interest, *MIRN20A* and *MIRN92* had been found to be coded by *C13ORF25*.

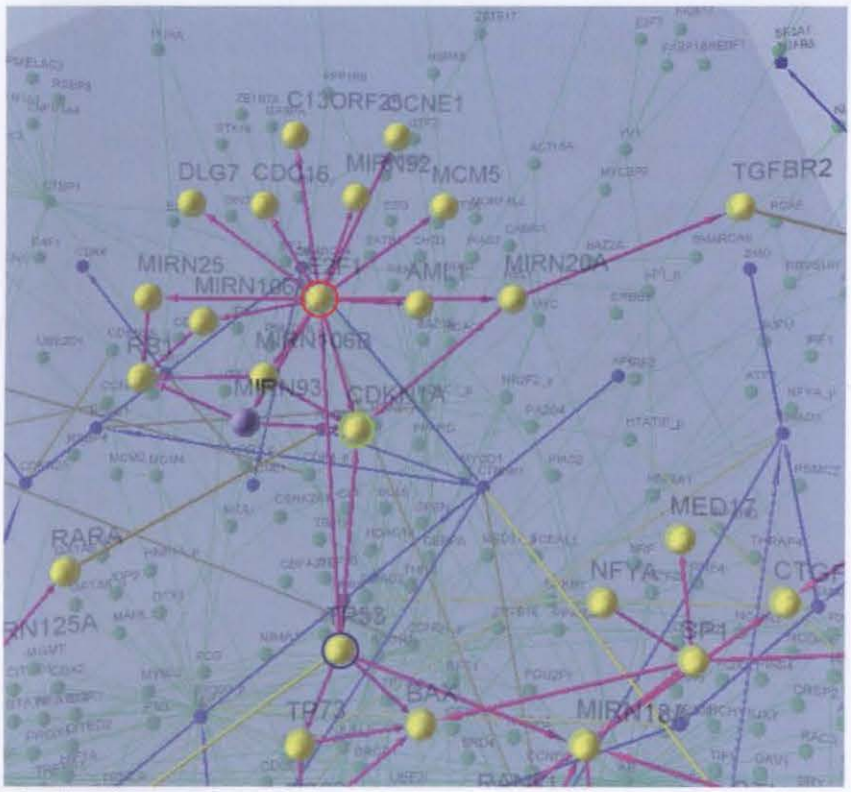


FIGURE 6.12. Visualization of the *E2F1* regulatory sub-network in G_3 (GRN). The oncogenic gene regulator *E2F1* is being circled in red and its repressor *TP53* is being circled in blue. The cyclin-dependent kinase *CDKN1A* which is repressed by *E2F1* is circled in light green. This view was derived from FIGURE 6.10.

MIRN106A was coded by *miR*106a-92 polycistron in chromosome X. *MIRN25*, *MIRN93*, and *MIRN106B* were coded by the *miR*106b-92 polycistron within the intron13 of the cell cycle gene *MCM7* [111].

As mentioned before, the tumour suppressor *TP53* represses *E2F1* expression. Therefore *TP53* could also indirectly repress the expression of *E2F1*-induced microRNAs. In Chapter 4, we deduced that *TP53* was not expressed in the cell cycle arrest (GO:0006917) biological process (see section 4.4.2.1). We also deduced that *MCM7* was not expressed in the DNA replication (GO:0006260) biological process (see section 4.4.2.3). Together with the deductions we made so far using the parallel plane layout, we tried constructing a hypothesis to suggest that the inappropriate inactivation of *TP53* could de-regulate *E2F1*-induced microRNAs in HCC cells. The concomitant up-regulation of *E2F1* would induce the expression of *C13ORF25*, *miR*106a-92 polycistron, and *miR*106b-92 polycistron in parallel with its host gene, *MCM7*. We hypothesized that this could lead to the repression of *RB1* expression due to gene silencing by multiple microRNAs (*MIRN25*, *MIRN93*, *MIRN106A*, and *MIRN106B*).

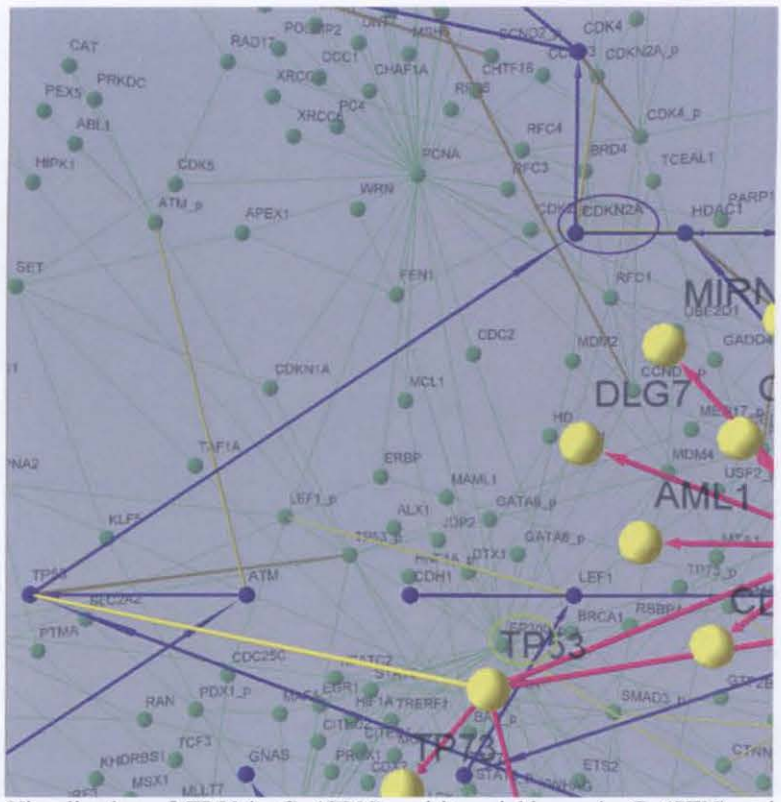


FIGURE 6.13. Visualization of *TP53* in G_1 (GRN) and its neighbours in G_2 (STN) and G_3 (PIN). Its neighbour *CDKN2A* in G_2 (circled blue) and its neighbour *EP300* in G_3 (circled light green). This view was derived from FIGURE 6.10.

FIGURE 6.12 showed that the node *CDKN1A* had incoming edges from multiple nodes *MIRN20A*, *MIRN93*, and *MIRN106B*. These edges represented the biological event that cyclin-dependent kinase inhibitor, *CDKN1A*, could be silenced by multiple microRNAs and is also repressed by *E2F1*. We moved from G_1 (GRN) to G_2 (STN) following the inter-plane edge originated from node *TP53* and found that it had signaling interaction with the node *CDKN2A*. It represented another cyclin-dependent kinase inhibitor like *CDKN1A* (see FIGURE 6.13). In turn, *CDKN2A* interacted with *CDK4*. *CDKN2A* had been known to inactivate *CDK4* by phosphorylation [64]. In G_3 (PIN), *CDK4* was shown to interact directly with cyclins, e.g. *CCND1* and *CCND2*, and DNA replication complex proteins, e.g. *RFC1*, *RFC2*, *RFC3*, *RFC4*, and *RFC5* (see FIGURE 6.13). *CDK4* had been known to interact with the above cell cycle proteins in different phases of the cell cycle [64]. The deduction that could be drawn from these observations was that the loss of *TP53* would lead to the abrogation of cell cycle arrest (GO:0006917).

A novel mechanism was the increase in *E2F1*-induced microRNAs which silenced the expression of cyclin-dependent kinase inhibitors e.g. *CDKN1A*, *CDKN2A*, and the tumour suppressor *RB1* [16]. More importantly, these two mechanisms co-operatively abrogated cell cycle arrest in HCC.

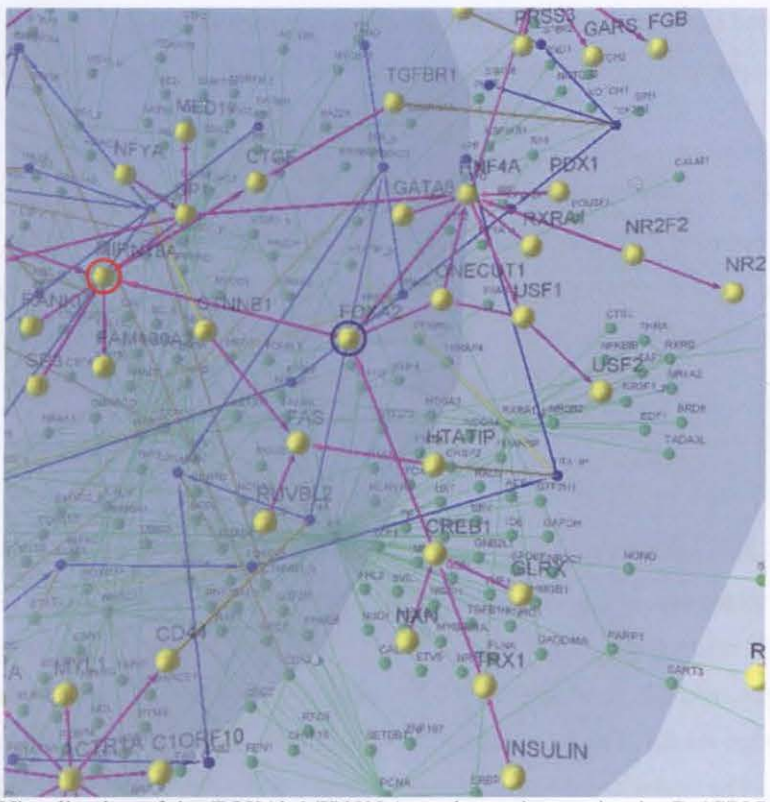


FIGURE 6.14. Visualization of the *FOXA2-MIRN18A* regulatory interaction in G_1 (GRN). *MIRN18A* is circled in red and *FOXA2* is circled in dark blue. This view was derived from FIGURE 6.10.

Our deduction further explained why there were no operating protein-protein interactions seen in the cell cycle arrest biological process (see Chapter 4, section 4.4.2.1). In Chapter 4 (see section 4.4.2.3), the PIN visualization for the DNA replication biological process (GO:0006260) showed that *MCM7* did not co-express with *MCM3*, *MCM4*, and *MCM5*, suggesting that *MCM7* was not as highly expressed. Thus the expression of *MCM7*-linked *miR*106b-92 polycistron in HCC should be more moderate compared to cancers with *MCM7* over-expression, e.g. prostate carcinoma [126] and gastric carcinoma [112]. However, a group of biologists found that *MIRN18A* had been actively expressed in 60% of Japanese HCC patients and was hitherto known to be liver- and cancer-specific [104]. Both findings suggested that *C13ORF25* expression was up-regulated in a subset of HCC cases but not the expression of other *E2F1*-induced microRNAs. FIGURE 6.14 showed that the node *MIRN18A* in G_1 could also be induced by the gene regulator *FOXA2*. It maybe that genome instability could reduce the number of functional *TP53* proteins in some HCC cases due to haplo-insufficiency and hence exerted a moderate repression of *E2F1* and *E2F1*-induced microRNAs.

In this analysis, we started our network exploration by identifying G_2 nodes that had two inter-plane edges connecting corresponding nodes in G_1 and G_3. We reasoned that these nodes must be common to all three networks and should have important biological functions.

We then started our visual analysis with one of the common nodes *TP53* which represented a well studied tumour suppressor protein. Our exploratory path followed the sequence:

$$G_2 \rightarrow G_1 \rightarrow G_2 \rightarrow G_3 \rightarrow G_1$$

In the process, we spent most of our analytical time in exploring the gene regulatory interactions between *TP53*, *E2F1* and *E2F1*-induced microRNAs in G_1. This was because the influence of microRNAs on the development of HCC was still largely unknown. The inter-connection between the GRN, *TGFB1*-STN and PIN visualized in the parallel plane layout provided with us ample information for deducing novel hypothesis on the probable biological role of microRNA in HCC progression.

6.5.2.2. Circular plane layout

FIGURE 6.15 showed that the same GRN-STN-PIN-overlapping network in the free-fixed-free circular plane layout (see algorithm 6.4). The networks for G_1, G_2, and G_3 were GRN, STN, and PIN respectively. The complete network consisted of 761 nodes and 991 edges of which 58 were inter-plane edges. The nodes common to G_1 and G_2 were connected by the inter-plane edge set E_{12}. The nodes common to G_1 and G_3 were connected by the inter-plane edge set E_{13}. The nodes common to G_2 and G_3 were connected by the inter-plane edge set E_{23}.

This layout captured the *systems* model which depicted the three networks as a tightly inter-connected system. As such, it represented the latest understanding in biology [93]. The comparable sizes of E_{13} and the E_{23} suggested that both GRN and STN were an integral part of the human nuclear PIN. The systems model was increasingly supported by recent data that many human signaling proteins could also act as gene regulators [113].

Although we attempted to identify nodes common to all three networks, the occlusions in G_1 and G_2 hindered this task. In G_1, the occlusion was caused by inter-plane edges obscuring one another in E_{13}. In G_2, the occlusion was mainly caused by nodes obscuring parts of the inter-plane edges in E_{12}, and poor angular resolution between the G_2 edges and the inter-plane edges in E_{12}. We found that only node *JUN* in G_1 and node *HTATIP* in G_2 were readily identifiable. Such occlusions were resolvable by rotating the x-axis and then the z-axis.

We also identified the node labeled *TGFBR1* in G_1 and the inter-plane edge to its corresponding node in G_2 readily. This was because the *TGFBR1* node was positioned at the upper side of P_1 where there was no occlusion (see FIGURE 6.15). We traversed the magenta coloured outgoing edge originated from the node *TGFBR1* in G_1 that led us to the node *CTGF*. We therefore deduced that the gene *CTGF* is a regulatory target of *TGFBR1*. This deduction was supported by the current biological knowledge that *CTGF* was induced by *TGFB1* [8].

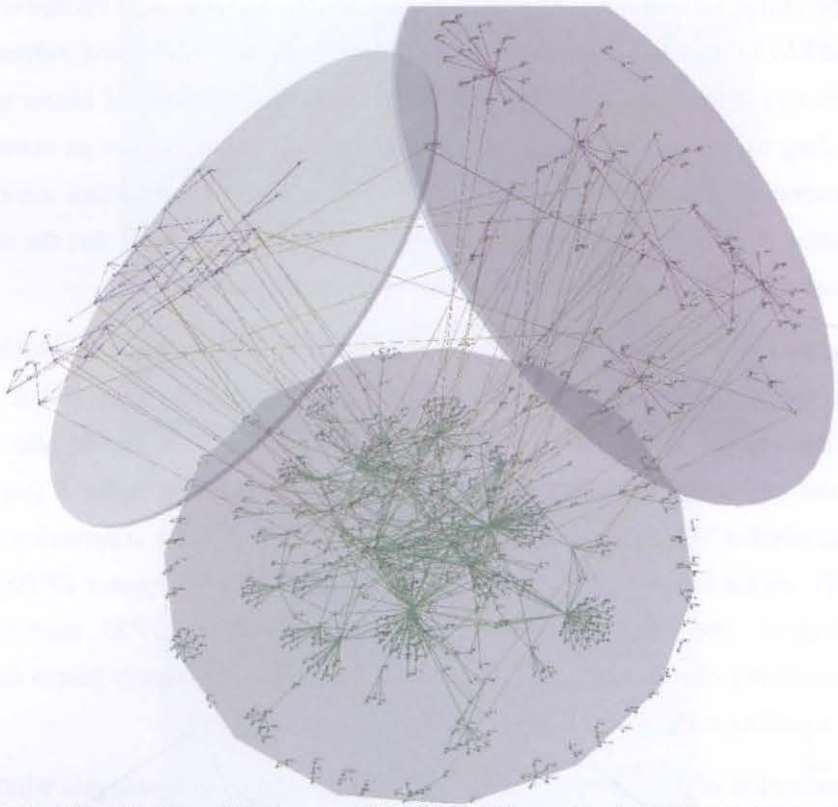


FIGURE 6.15. Visualization of the human GRN-STN-PIN-overlapping network in the circular plane layout. The G_1 network represents the GRN (magenta nodes, magenta edges), the G_2 network represents the *TGFB1*-STN (dark blue nodes, dark blue edges), and the G_3 network represents the PIN (green nodes, green edges). The inter-plane edge sets E_{12}, E_{13}, and E_{23} represents node correspondence between G_1, G_2 and G_3.

Biologists had observed that *CTGF* could promote angiogenesis and cell migration [108]. It had been shown to co-express and interact with *TGFB1R* in the angiogenesis-defined PIN (GO:0007155) (see Chapter 4, section 4.4.2.4).

Of interest, we also found in FIGURE 6.15 that the node *CTGF* had an incoming edge originated from the node labeled *MIRN18A*. *E2F1*-induced microRNA, *MIRN18A*, had a gene regulatory interaction with *CTGF* which meant that its expression could be silenced by the former. In angiogenesis, *CTGF* induced the secretion of collagen and fibronectin from cancer cells which formed the scaffolding of the extracellular matrix, a crucial step in the formation of a new vascular system [28]. The up-regulation of *MIRN18A* in some HCC cases could lead to the poor formation of the extracellular matrix due to the repressed translation of *CTGF* mRNAs. One probable consequence could be excessive endothelial cell migration but inadequate cell anchorage due to a poorly formed extracellular matrix and hence poor vascular formation.

The above deduction was supported by the latest biological knowledge that the vascular system in the cancer cell mass was known to be structurally defective with excessive leakage [81] and *MIRN18A* expression could be a contributing factor. This could enhance HCC metastasis in two ways. The first could be the increased dissemination of cancer cells into the surrounding liver tissue because of vascular leakage, a process known as extravasation [81]. The second could be the amplification of tissue invasion by *MMP*s which were induced by the hepatitis B viral oncoprotein *HBX* in cancer cells [33]. It seemed that the impact of *E2F1*-induced microRNAs may extend well beyond cell cycle control.

We suggested in section 6.5.2.1 that the de-regulation of *E2F1*-induced microRNAs may had altered the interaction dynamics within the GRN and the STN in favour of anti-apoptosis. The change in the interaction dynamics within the STN should also have an impact on the interaction dynamics in the PIN. We suggested in Chapter 5 (see section 5.6.2.1) that, whether *TGFB1* would promote cancer growth or tumour suppression relies, at least partially, on the differential affinity [162] between the signal integrator *EP300* and the variety oncogenic proteins and tumour suppressors. Therefore *TP53* inactivation in conjunction with the silencing of tumour suppressor genes should in theory hasten the loss of differential signaling in the *TGFB1*-STN.

The visualization of the inter-plane edge set E_{13} prompted us to investigate which of the gene regulators in G_1 was the neighbour of *EP300* in G_3. *EP300* was also a signaling protein represented in G_2 (STN). We started from the node *TGFBR1* in G_2 and traversed the network path:

$$TGFBR2 \rightarrow TGFBR1 \rightarrow ZFYVE9 \rightarrow SMAD2 \rightarrow EP300$$

The biological meaning of this path had been delineated in Chapter 5 (see section 5.6.2.2). At G_2, we traversed the inter-plane edges originating from node *EP300* to its corresponding node in G_3 (see FIGURE 6.16(a)). In G_3, we visually searched for the neighbours of *EP300* that had inter-plane edges pointing towards the direction of G_1. We found one such neighbour labeled *HNF1A* (see FIGURE 6.16(b)).

To identify the neighbours of *HNF1A* in G_1 (GRN), we traversed the inter-plane edges originating from the G_3 node *HNF1A* to its corresponding node in G_1. We found that *HNF1A* (Gene ID: 6927) was a date hub that had sixteen neighbours (see FIGURE 6.16(c)). With the exception of *HNF4A*, the topology of node *HNF1A* shows that it had outgoing edges toward its neighbours. We noticed that the intra-plane edge connecting *HNF1A* and *HNF4A* was bi-directional. From this observation, we deduced that the *HNF1A* gene regulates all its neighbours but was also a regulatory target of *HNF4A*. Therefore *HNF1A* and *HNF4A* form a regulatory loop.

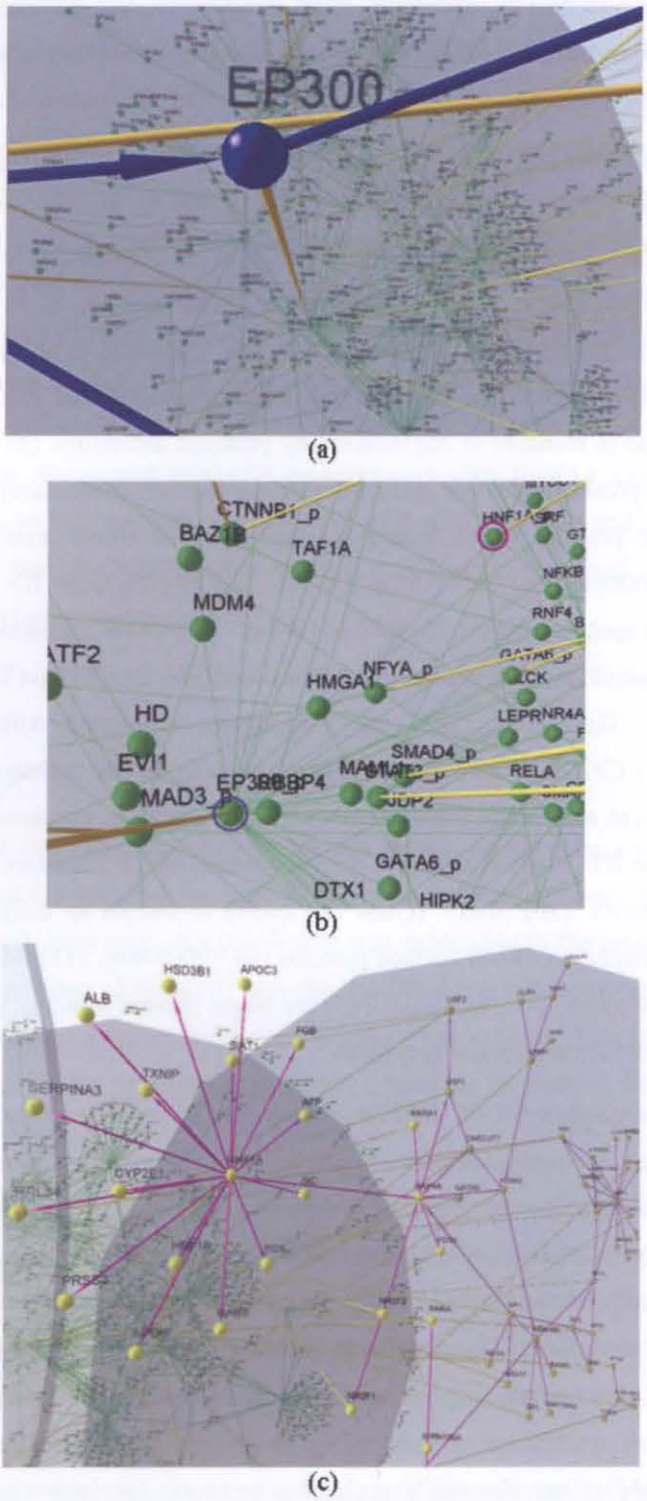


FIGURE 6.16. A fly-through sequence from G_2 (STN) to G_1 (GRN) via G_3 (PIN). (a) A zoom-in view of the node *EP300* in G_2 with an inter-plane edge connected to its corresponding G_3 node. (b) The node *EP300* in G_3 (circled blue) is connected to one of its neighbours *HNF1A* (circled magenta). (c) The node *HNF1A* and its neighbours form a date hub in G_1. This figure is derived from FIGURE 6.15.

Furthermore, from our observations made in G_3, *HNF1A* needed to interact physically with *EP300* in order to induce the expression of its neighbours. We then searched the Entrez database [99] for their known biological functions and found that they could be divided into twelve groups.

These groups were (1) liver-specific gene regulators, i.e. *HNF1B* (Gene ID: 6928) and *HNF4A* (Gene ID: 3172). Both had recently been known to regulate liver-specific genes [107]. (2) An enzyme crucial to the synthesis of tRNA, i.e. *GARS* (Gene ID: 2617). It is a glycyl-tRNA-synthetase that attaches the amino acid glycine to a tRNA molecule (GO:0006426). The glycyl-tRNA is then used by the ribosome in the process of mRNA translation. (3) *PRSS3* (Gene ID: 5646) is a serine protease associated with the inflammation of the pancreas and is resistant to degradation by protease inhibitors. (4) *SERPINA3* (Gene ID: 12) is a serine protease inhibitor and its deficiency had been associated with liver disease. (5) Lipid transport proteins which transport lipids from the blood vessels to the liver for catabolism (GO:0006869). *APOC3* (Gene ID: 345) and *APOH* (Gene ID: 350) are two such proteins known as apolipoproteins. (6) *RPL34* (Gene ID: 6164) is a member of the ribosome protein complex which mediates protein biosynthesis (GO:0006412). (7) *FGB* (Gene ID: 2244) and *FGA* (Gene ID: 2243) are fibrinogens associated with wound healing (GO:0042060). (8) *CYP2E1* (Gene ID: 1571) is a member of the cytochrome P450 protein family. It is known to metabolize ethanol and carcinogens such as nitrosamines and benzene. (9) *HSD3B1* (Gene ID: 3283) is a steroid dehydrogenase which catalyzes the conversion of cholesterol to steroids. (10) *SIAT1* (Gene ID: 6480) is known as sialytransferase which catalyzes the glycosylation of cell surface proteins (GO:0006486). (11) *ALB* (Gene ID: 213), *GC* (Gene ID: 2638) and *AFP* (Gene ID: 147) are blood plasma proteins. (12) *TXNIP* (Gene ID: 733688) is thioredoxin interacting protein.

Given that the neighbours of *HNF1A* shown in G_1 were involved in such diverse biological processes, we suspected that it could be a master regulator in human hepatocytes. Our deduction was supported by the latest computer model of liver-specific GRN [107]. The functional diversity seen among neighbours of *HNF1A* might also explain how multiple risk factors could pre-dispose the onset of HCC, e.g. hepatitis B viral infection or liver damage due to alcohol addiction [47]. The proteins in groups (3), (7), and (8) had been suspected by clinicians and biologists to pre-dispose carcinogenesis [45; 47]. We therefore proposed that if these risk factors could lead to the prolonged up-regulated expression of *HNF1A* and *TGFB1*, they could prime the cell's molecular network for later changes in network dynamics required for cellular transformation. Cellular transformation meant the phenotypic shift of a normal cell to a cancer cell [161].

In Chapter 5, we postulated from the two-overlapping network that for cellular transformation to occur, there needed to be a total loss of differential signaling in the *TGFB1*-STN. This could be due to the progressive loss of functional tumour suppressors (see section 5.6.2.2). Visual analysis of the three-overlapping network (see FIGURE 6.16) gave us the added insight that the prolonged up-regulation of *HNF1A* expression might have pre-disposed the loss of differential signaling in the *TGFB1*-STN. The reason was that some of the neighbours of *HNF1A* suspected to pre-dispose to carcinogenesis were also persistently up-regulated. A marker of *HNF1A* up-regulation was the expression of *AFP* found in group (11) in the above list. *AFP* was considered by clinicians to be a conventional diagnostic and prognostic biomarker in HCC. High expression level of *AFP* was often associated with poor prognosis due to cell proliferation, high angiogenesis, and low apoptosis [103]. Recently, a group of biologists discovered that *AFP* co-expressed with *HNF1A* and with some of its neighbours shown in FIGURE 6.16, i.e. *HNF4A*, *ALB*, *GC*, *APOC3*, and *APOH* [118]. This finding indicated that the *HNF1A* sub-network was highly expressed in HCC cells and provided some support to our deduction.

In this analysis, we explored all the networks in the circular plane layout in the following sequence:

$$G_1 \rightarrow G_2 \rightarrow G_3 \rightarrow G_1$$

We spent most of our analytical time in identifying the nodes were common to all three networks and found that this task was better done with the parallel plane layout than with the circular plane layout. However, this conclusion was applicable to the present human dataset only. The other limitation of the latter was the need to traverse lengthy inter-plane edges as compared to its parallel plane counterpart. However, the explicit visualization of the inter-plane edge set E_{13} in the circular plane layout helped us to deduce the probable role of *HNF1A* in pre-disposing HCC.

6.5.2.3. Conclusion

In summary, both methods for visualizing the human GRN-STN-PIN-overlapping network have their strength and limitations. In the present case study, the parallel plane layout helped us to identify signaling proteins represented in the *TGFB1*-STN that were also present in the GRN and the PIN. On the other hand, the circular plane layout was good for cyclical exploration because the biologist could visualize the direct mappings among all the three networks. From the same human networks, the circular plane layout helped us to identify the liver-specific gene regulator *HNF1A* that interacted with the signal integrator *EP300* and had biological functions in the GRN. For the human dataset, both visualization methods had their place in visual analysis and in biological deductions.

6.6. Remarks

Through the case studies, we demonstrated the use of the three-overlapping network visualizations for biological analysis. In the *E. coli* case study, our deductions were supported by recent publications, thus showing that the visualizations of concern were effective visual knowledge representations. In the human case study, we made deductions on the biology of HCC that further explained those made in the previous chapters. Therefore, the three-overlapping network visualizations did support knowledge discovery.

So far, our case studies showed that the readability and usability of the circular plane layout and its parallel plane layout were dependent on the dataset applied, even though the former was probably more aligned to the latest domain knowledge in biology. In the *E. coli* case study, it was more difficult to identify nodes common to all three networks using the parallel plane layout as compared to the circular plane layout. When using the parallel plane layout, this task involved identifying G_2 nodes that had inter-plane edges connected to their corresponding nodes in G_1 and G_3. The occlusion seen between the planes P_2 and P_3 in the *E. coli* case made the above task difficult. The opposite was true in the human case. That was because there was less occlusion seen within each inter-plane edge set in the parallel plane layout. However, in the human case, occlusion seen in the circular plane layout was resolvable using rotation.

With the circular plane layout, the above task required identifying G_1 or G_3 nodes that have two inter-plane edges. The rationale was that, if a node in G_1 has corresponding nodes in both G_2 and G_3, then the node must be common to all three networks. This task was achievable in the *E. coli* case because there was no occlusion seen within the inter-plane set E_{13}. However, the opposite was true in the human case. Serious occlusion was seen within the plane P_2 and the inter-plane edge set E_{13} making the identification of common nodes tedious.

Because their explicit mapping was not available, the parallel-plane visualization required the use of *transitivity* to imply the relationship between the G_1 and G_3 networks. Transitivity meant that if a G_1 node was connected with its corresponding node in G_2, and the G_2 node was connected with its corresponding node in G_3, then the G_1 node must corresponded to the G_3 node. Apart from biology, we knew that transitivity applied to financial transaction network consisting of heterogeneous transaction types. However, transitivity might not apply to every domain. Therefore the generalization of this property was yet to be tested extensively.

Biologists could use the three-overlapping network visualization as a follow-up visual analysis step to its two-overlapping network counterpart. The human case study

demonstrated the effectiveness of this approach. We first used the STN-PIN overlapping network visualizations to investigate how *TGFB1* differential signaling might work in relation to the differential affinity of *EP300* with nuclear proteins (see Chapter 5, section 5.6.2.1). We then used the GRN-STN-PIN overlapping network visualizations to investigate how GRN regulated the signaling proteins within the STN, and how those two networks may together influence the organization of the nuclear PIN. When used sequentially, the two types of overlapping networks allowed us to deduce novel hypotheses on how HCC developed robustness against any repression on cell growth and metastasis.

{End of Chapter 6}

CHAPTER 7

Conclusion and Future Work

"This is just the Beginning"

Throughout this thesis, we demonstrated the relevance of network visualization in supporting visual analysis on molecular biology networks. We also demonstrated that biological concept models map very well to networks. To go a step further, we demonstrated that our visual analysis framework could assist the biologist in the incremental investigation of gene expression in the context of bio-molecular networks. Such a framework contained a series of network visualizations in decreasing level of abstraction. The objective was to assist the expert biologist in exercising analytical reasoning.

Because our focus was on visual analysis of biological networks, we studied how different methods for visualizing networks with different biological focus, different size, and different layout support biological analysis.

7.1. Summary

7.1.1. Visual Analysis Framework

Our visual analysis framework (see FIGURE 1.2) was a novel approach towards visual analytics for molecular biology. Each step in the framework had a different focus. We found that the deductions made in one visual analysis step could often be further explained by subsequent steps. It did support the biologist's practice of incremental investigation. The progressive use of increasingly complex visualizations did increase the explanatory power of each visual analysis step. For example, we deduced the impact of *TP53* inactivation on the progression of hepatocellular carcinoma (HCC) using the three visual analysis steps in the framework as follows.

Step 1– Co-expressed gene clusters. We found that the GO Process label '*GO:0007050|cell cycle arrest*' in the clustered bipartite graph representation was exclusive to the 'disease' sample (see Chapter 3, section 3.4.2.2). This finding suggested that the cell cycle arrest biological process was dysfunctional in HCC.

Step 2– GO_Process-defined PIN. We then proceeded to the cell cycle arrest-defined PIN analysis and deduced from the non-clustered PIN visualization that there was a complete loss of functional protein interactions that can initiate cell cycle arrest. We also identified that one of the nodes labeled '*TP53*' represented a protein not only involved in cell cycle arrest but

also in the initiation of apoptosis [129]. This protein was a tumour suppressor well studied by cancer biologists (see Chapter 4, section 4.4.2.1). We also deduced from the non-clustered visualization of the angiogenesis-defined PIN that *TGFB1* expression was up-regulated. From the clustered PIN visualization, we proposed that, maybe in HCC, many proteins involved in cell cycle arrest were abnormally sequestrated in the cytoplasm rather than in the nucleus. If this happens, it would further cripple their ability to arrest cell cycle progression in HCC.

Step 3– Integrated molecular network. With the two deductions in step 2, we examined the connectivity between the *TGFB1*-STN and the PIN in the STN-PIN overlapping network visualization in the three-parallel plane layout to deduce a hypothesis that explained how the loss of *TP53* expression might have led to the loss of differential signaling in the *TGFB1*-STN (see Chapter 5, section 5.6.2). Finally we deduced from the GRN-STN-PIN-overlapping network visualization in the parallel plane layout that the loss of *TP53* expression could have de-regulated *E2F1*-induced microRNAs that silenced tumour suppressor genes. This might hasten the loss of differential signaling in the *TGFB1*-STN thereby promoting cancer cell growth (see Chapter 6, section 6.5.2).

Using the visual analysis framework, we demonstrated the feasibility of using a series of model visualization for generating biological insight. On hindsight, our visual analysis framework had a similar emphasis to that proposed by Amar and Stasko [4]. Their framework proposed that a user has to perform three cognitive tasks in sequence when using a visual representation of data for assisting his/her high-level analytical reasoning. This sequence is as the follows:

Analyst perceptual processes→Perceiving useful relationships→Explaining relationships

They argued that a visualization system is effective if it can present relationships among data clearly and also indicate useful visual representations and their limits. In our framework, the use of network visualizations for assisting the above cognitive tasks was the main concern. That was because explaining biological relationships was the leading step to biological deduction. That was why evaluating network visualizations for their merits in supporting biological deduction in each visual analytical step became the focus of this thesis.

We could define the meaning of a network visualization being 'cognitively challenging' in the context of Amar and Stasko's framework [4]. If its visual complexity was preventing the biologist from perceiving relationships, the network visualization was considered to be '*cognitively challenging*' and was therefore ineffective in supporting the user's (or analyst's) analytical reasoning. Furthermore, we could also define the term '*biological insight*' in the

context of our visual analysis framework as a rational explanation on the perceived biological relationships displayed in a biological network visualization.

7.1.2. Visualization of the GO_Process-Annotated Co-expressed Gene Clusters

In Chapter 3, we investigated the merits of two different visual representations of GO Process-annotated co-expressed gene clusters in supporting analytical reasoning. We contributed two visual representations, the block matrix and the clustered bipartite graph, with the intention to capture two different biological concept models. We used the *block matrix* representation to capture the *gene-centric* model which had been the biologist's conventional view on gene co-expression. It assumed that by knowing which biological process(es) each gene cluster was involved in, the biologist should be able to make deductions on which biological processes were co-regulated. On the other hand, we used the *clustered bipartite graph* representation to capture the *network* model. This model assumes that the biological processes are inter-connected because they share the same co-expressed gene clusters. The network model is increasingly being adopted by systems biologists who study molecular biology as a network.

We experimented with each visual representation using the datasets on normal hepatocytes and hepatocellular carcinoma (HCC) as input [58]. We then performed visual analysis on each representation as a series of analytical tasks that biologists are most likely to perform. Our user experience in the visual analysis brought us to the conclusion that the block matrix representation is more suitable for examining the biological function of a selected gene cluster or comparing a cluster pair for functional relatedness. The clustered bipartite graph representation is more suitable for comparing between sample sets (normal *vs.* disease) for their biological differences. Since biologists often need to perform sample set comparisons in biological analysis, the clustered bipartite graph representation is better suited to gene expression analysis than the block matrix.

Our conclusion was further supported by the results from the usability evaluation. We noticed that our participants stalled when they were asked to deduce the biological differences between the 'normal' sample and the 'disease' sample (see Chapter 3, section 3.5.5.1). To our surprise, the user evaluation results also suggested that visual simplicity does not necessarily enhance user performance, in either the task completion time or in analytical accuracy. Although the block matrix representation is visually simpler than the clustered bipartite graph representation, our participants in the former did not give a better performance in task completion time and in analytical accuracy. Rather, the redundant representation of the GO Process terms and their distribution throughout the visualization reduced the readability of the block matrix and hampered its use for visual analysis. The same problem was not seen with the clustered bipartite graph representation which displayed

the *m:n* gene_cluster-GO_Process relationship as inter-level edges. However, the readability of the clustered bipartite graph representation deteriorated with the increase in edge crossing.

By far the most valuable insight gained from our research in Chapter 3 was that capturing the *network* concept model is more relevant to biologists, in the context of biological analysis, than preserving their *gene-centric* model. This is especially relevant in the present day when experimental biologists increasingly need to adopt the network view in order to make a better use of high-throughput data for hypotheses deduction. Another important lesson we gained was that emphasizing on visual simplicity at the expense of visual representational accuracy will hamper analytical accuracy rather than enhancing it.

7.1.3. Visualization of the GO_Process-defined Protein Interaction Networks

In Chapter 4, we investigated the merits of two different visualizations of a GO_Process-defined PIN in supporting analytical reasoning. By using the GO hierarchy to filter down the complete human PIN, we obtained a filtered PIN in which all the nodes share the same GO label as an attribute. We called such a filtered PIN as a *GO-defined PIN*. If the GO Process category was used as the filtering criterion, the resulting PIN was a GO_Process-defined PIN.

We visualized the GO_Process-defined PIN using two methods, i.e. the *non-clustered PIN* visualization and the *clustered PIN* visualization. The non-clustered PIN is simply the GO_Process-defined PIN being visualized in the force-directed layout which has been the conventional method [44, 148]. The alternative option is the clustered PIN visualization. This is the GO_Process-defined PIN being visualized as a set of inter-connected clusters in a clustered circular layout. The visual clustering of the protein node set was done by using GO Component as a criterion. The purpose is to allow the biologists to add complementary information to a GO_Process-defined PIN at their discretion, thus enhancing their analytical reasoning. The clustered circular layout provided a novel way for visualizing clustered PIN. The layout was easy to compute and captured the nested modularity of the PIN effectively.

We implemented a visualization system that allowed biologists to navigate from one GO Process label to the other. We then experimented with each PIN visualization by overlaying the HCC co-expression dataset [58] on the network nodes. We performed visual analysis on each visualization as a series of analytical tasks that biologists are most likely to perform. During the visual analysis, we found that to make biologically meaningful deductions, the non-clustered PIN visualizations had to be interpreted in the context of the current biological knowledge. The GO_Process-defined PIN has no meaning to a user who is not an expert biologist. Although the non-clustered PIN allowed us to identify protein nodes with unique topologies, e.g. bottleneck proteins and party hubs, the biological meaning of these

topologies might not apply well to every GO_Process-defined PIN. Therefore, we cautioned biologists to be careful in interpreting the biological meaning of node topologies in any non-clustered PIN visualization.

By comparison, we found that interpreting clustered PIN visualizations was cognitively less challenging because of the added information on the cellular distribution of proteins represented by the GO Component labels. They alleviated our cognitive loading by providing biological knowledge in visual form which was useful to biological deduction. Therefore our user experience suggested that the clustered PIN visualization is more informative than its non-clustered PIN counterpart.

We made two important findings in the domain expert evaluation. The first finding was that the fixed layout of the clustered PIN visualization has the advantage of reducing the biologist's cognitive load on repeated usage since the person does not need to re-adapt to a new layout in every rendering of the same network. This visual feature should be an important design consideration in molecular network visualizations. The second finding was that, when asked to identify *functionally essential* protein from each visualization, the expert biologist was relying on node degrees to achieve this task. In the non-clustered PIN visualization, he identified nodes with a node degree greater than five as representations of essential proteins. In the clustered PIN visualization, the expert biologist took inter-cluster edges into account when trying to identify essential proteins. This observation aligned with our own experience when performing visual analysis, we found that node degrees in a PIN visualization do have biological meaning and needed to be accounted for in visual analysis.

The most important achievement we made in this study was demonstrating the feasibility of making biological deductions without following the conventional information visualization mantra of "*overview, zoom and filter, details on demand*" [141]. Our approach was to use the biological concept model of cancer as a rational guidance for our visual analysis. We then began with a selected GO_Process-defined PIN as the starting point. In short, our approach can be summed up as "*filter first, zoom and details, overview if necessary*". We argued that this approach is better for guiding those expert biologists who come from the 'reductionist' school of biology into systems exploration. These biologists need PIN visualization at a scale that they can cognitively handle. In contrast, the conventional mantra is more suitable for biologists who come from an engineering background.

7.1.4. Visualization of the Two-Overlapping Networks

In Chapters 5 and 6, we introduced the novel problem of visualizing *overlapping networks*. In Chapter 5, we introduced two visualization methods of the *two*-overlapping network, i.e.

the two-parallel plane layout and the three-parallel plane layout. Their difference laid in the display of the overlap layer in the three-parallel plane layout.

To investigate the merits of the two visualization methods in supporting analytical reasoning, we performed two case studies. In the *E. coli* case study, we experimented with each visualization method using the public datasets for protein interaction network (PIN), metabolic network (MN) or gene regulatory network (GRN). The objective was to use well studied biological networks to evaluate the effectiveness of the two-overlapping network as a concept model visualization. If it does, our deductions should be supported by the current biological literature. In the human case study, we experimented with each visualization methods using the public datasets for the *TGFB1* signal transduction network (STN) and the human PIN that are not only found in the cell nucleus but are also HCC-specific. The objective was to evaluate the effectiveness of the two-overlapping network in supporting hypothesis deduction.

In both case studies, we consistently found that the three-parallel plane layout supported analytical reasoning better than its two-parallel plane layout. Even for a large two-overlapping network representation that exceeded 1000 nodes, such as the *E. coli* PIN-GRN-overlapping network (see Chapter 5, section 5.5.2.2), the three-parallel plane layout could still help us to make limited deductions whereas the two-parallel plane layout could not. We attributed the strength of the three-layer parallel plane layout to its displaying of the overlap nodes and edges within a separate plane and highlighting the overlap nodes with a distinct colour hue. These visual features were crucial to our success in making biologically interesting deductions because they helped us to prioritize which of the two heterogeneous biological networks to investigate first.

In the *E. coli* use case, we demonstrated that the two-overlapping network visualizations were good concept model visualization in general. The deductions we made were supported by the recent biological literature. In the human use case, we demonstrated that the two-overlapping network visualization is effective in assisting hypothesis deduction (see Chapter 5, section 5.5.2.3).

7.1.5. Visualization of the Three-Overlapping Networks

In Chapter 6, we introduced two visualization methods of the *three*-overlapping network, i.e. the parallel plane layout and the circular plane layout. We used the *parallel plane layout* to capture the biological concept model known as the *cascade* model [3]. The cascade model depicts a clear functional ordering of three heterogeneous networks. We used the *circular plane layout* to capture the biological concept model known as the *systems* model. The systems model depicts the functional co-operation of three heterogeneous networks. Again,

we used *E. coli* networks and human networks as case studies to evaluate the merits of the two visualization methods in supporting analytical reasoning.

In both case studies, we were able to use the parallel plane layout and circular plane layout to make certain biological deductions. In the *E. coli* case study, the parallel plane layout helped us to identify the indirect biological relationship between a metabolic enzyme and a gene regulator, thus helping us to deduce the regulatory targets in the *E. coli* MN. On the other hand, the circular plane layout helped us to identify the metabolic enzymes that had biological functions in the MN and the GRN. In the human case study, the parallel plane layout helped us to identify signaling proteins represented in the *TGFB1*-STN that were also present in the GRN and the PIN, thus allowing us to hypothesize their influence on the progression of HCC. On the other hand, the circular plane layout helped us to identify the indirect biological relationship between a gene regulator and a signaling protein. The usability of each layout seemed to depend on the chosen data set. Therefore we could only conclude that the both visualization methods had their place in visual analysis and in biological deductions.

7.1.6. Usability Evaluation

In Chapter 3, we introduced the first set of benchmark tasks for evaluating Gene Ontology-annotated gene clusters (see section 3.5). These analytical tasks emphasized on identifying active and co-regulated biological processes based on inter-cluster comparison. The list of tasks was by no means exhaustive and there could be other analytical tasks that could be added to the evaluation in future. Our benchmark tasks could be modified to suit any non-biology domains where there was a need to analysis a set of ontology-annotated clusters visually.

In the evaluation process, we measured the dependent variables: task completion time, accuracy, and user confidence score for each task. They were conventionally used for measuring participant's performance in user evaluation studies [134, 164]. In practice, we found that task completion time was not a good measure on user performance for the conceptual tasks, i.e. tasks that require analytical reasoning for deducing a solution. The reason was that the biologist's personality and domain expertise could affect his/her thinking path and approach towards the same conceptual task, not just the design of the visualization.

By comparison, accuracy is a better measurement on user performance because every participant's solution is being compared against a well-defined solution. However, in a knowledge-intensive domain like biology, a well-defined solution is only possible if the evaluator shares the same domain knowledge with the participants. In our case, we shared the same understanding on the biological implication of gene co-expression with our

participants. Furthermore, we chose a set of up-regulated and positively co-expressed genes for the evaluation, so that the participants are likely to interpret gene co-expression in a similar way.

7.2. Challenges and Future Work

Through this thesis, we provided the interested experts from the fields of bioinformatics, information visualization, and visual analysis, with a starting point for investigating visualization-related problems in molecular biology. For the biologists, we hoped that this thesis would lead them to consider combining visual analysis with quantitative analysis methods in studying their research problems.

However, this thesis also raised prospects for future research regarding bioinformatics visualization. In this section, we listed future work that provided possible directions for further research into the visualization and analysis of biomolecular network.

7.2.1. Visual Analysis Framework

So far, we had only emphasized on the effectiveness of model visualization in our visual analysis framework. In Amar and Stasko's vocabulary [4], we were attempting to bridge the *worldview gap* but not the *rationale gap*. The definition of the term '*worldview gap*' is the gap between what is being displayed and what actually needs to be displayed in order to deduce a straightforward conclusion for making a decision. The definition of the term '*rationale gap*' is the gap between perceiving a relationship and actually being able to explain confidence in that relationship and the usefulness of that relationship.

Amar and Stasko [4] asserted that visualization systems built to bridge the worldview gap often failed in elucidating the strengths of the perceived relationships and the confidence in these relationships. This was the limitation of our visual analysis framework. The failure to bridge the rationale gap could lead the biologist to the wrong analytical path. Such a scenario appeared during our analysis of the angiogenesis-defined PIN using the non-clustered PIN visualization (see Chapter 4, section 4.4.2.4). We had incorrectly deduced that the large connected component in the PIN visualization represented a protein complex with subunits. Yet the current biological knowledge informed us that those perceived protein-protein interactions are more likely to be pairwise. Therefore the confidence of those protein-protein interactions being part of a protein complex is almost zero, so is their usefulness in HCC research. To bridge the rationale gap, we needed to clearly label the edges in the angiogenesis-defined PIN as '*pairwise only*'. How to use visual encoding to close the analytical gap seen in our visual analysis framework will be our next objective.

Although we applied co-expressed gene clusters as our first visual analysis step, the framework can also be modified according to the biologist's requirement. For example, some biologists may prefer visualizing differential gene clusters as the first step. The future work will be to design and implement the framework as a full visualization system equipped with molecular network visualization of different organisms, e.g. yeast, bacteria, fungi, human, mouse, and plants.

7.2.2. Visualization of Co-expressed Gene Clusters

With our research indicating that the clustered bipartite graph representation was visually more complex more than its block matrix counterpart; several options could be taken to improve the usability of the former. Feasible options were listed as follows:

1. To make the clustered bipartite graph more informative, interactivity should be added to the gene nodes such as drop down menu on brushing to provide hyperlinks to public databases such as ENTREZ.
2. To further reduce the negative impact of edge crossings on readability, edge highlighting on pointer brushing can be added to the clustered bipartite graph representation.
3. For comparing the cluster patterns of two sample sets (e.g. normal *vs.* disease), a tripartite graph with the GO Process nodes as the intermediate layer could be an option. This should help the biologist to compare two sample sets for their differences in GO Processes.
4. Another user evaluation can be performed by comparing a published visualization method with clustered bipartite graphs in various designs. This should provide more information on their usability in microarray analytics.

7.2.3. Clustered PIN Visualization in Biological Context

In general, we found that for PIN networks that were approaching 1000 nodes, the force-directed layout algorithm produced the 'hair ball' effect on the network visualization [148]. We found that such a visualization is poor in supporting analytical reasoning. At this scale, the high-degree nodes (or hubs) automatically became our visual focus because of their star-shaped formation and of the high edge density around them.

In the analysis of the signal_transduction-defined PIN (see Chapter 4, section 4.4.2.2), we experienced difficulty in using the non-clustered PIN visualization to identify the low node degree neighbours of the hubs in the force-directed layout. This layout contained 563 nodes and 832 edges. Instead we had to use the clustered PIN visualization to achieve the same task. That was because, in the clustered circular layout, the protein nodes colour-coded for co-expression were confined to specific areas demarcated by the cluster nodes. Furthermore

the label of each cluster node provided information on the cellular component in which the co-expressed proteins were contained, thus making the visualization more informative.

The only limitation of the clustered PIN visualization was that node redundancy reduced the readability of the visualization by increasing the size of the original GO Process-defined PIN, and the number of redundant edges. This limited the size of the clustered PIN that can be visualized using the circular layout. Taking into account our experience with the GO-annotated gene cluster visualizations (see Chapter 3, section 3.4.2.2, subsection I), redundant node representation should be minimized if possible.

7.2.4. Visualization of Overlapping Networks

In the analysis of the *E. coli* PIN-GRN-overlapping network (see Chapter 5, section 5.5.2.2), the high degree hubs in the PIN of the two-parallel plane layout became our visual focus. This PIN was also visualized in the force-directed layout and contained 451 nodes and 730 edges. In this visualization, the aggregation of high degree nodes and their node labels gave rise to occlusion that made it difficult for us to identify the high degree node within our interest. However, the overlap layer in the three-parallel plane layout provided a visual representation for the subset of nodes and edges commonly shared between the PIN and the GRN. Furthermore, the overlap layer was visually confined to a separate plane from those confining the PIN and the GRN.

In the above case studies, the force-directed layout limited our ability in perceiving biological relationships that were useful to biological deductions, even though the size of the PINs visualized may be very moderate. Our user experiences in these cases studies provided a strong argument for using layouts that factor in some kind of biological context. Visual experiments with biological context-based layouts had been conducted recently [7, 71, 87]. So far, only the *betweenness fast-layout algorithm* had been tested on a PIN with 10^4 nodes [71]. Furthermore, none of them had been tested on a network integrated with multiple interaction types. It would be interesting to see what visual effect would be generated from an overlapping network that employed one or more of these biological context-based layouts. The interesting question would be: could we omit the overlap layer in the two-overlapping network visualization simply by using different biological context-based layouts for different networks?

7.2.5. Usability Evaluation

There is a need to explore the suitability of using heuristics evaluation for assessing the usability of bioinformatics visualization. It is not as time-consuming to conduct as controlled experiments and is usually conducted with a few domain experts. The challenge is that there is no benchmark heuristics defined for evaluating biological network visualizations.

Although there were new heuristics being contributed for the evaluation of pathway visualization systems [134], the question as to whether other published heuristics [4, 141, 172] are suitable for the same purpose is yet to be answered. In particular, it will be interesting to test whether the knowledge and task-based framework [4] is suitable for evaluating network visualizations that integrate multiple molecular interaction types, since it is the only framework that coupled relationship recognition with high-level analytical tasks.

7.2.6. Biology of Hepatocellular Carcinoma

In Chapters 5 (see section 5.6.2.2) and 6 (see section 6.5.2.2), we made two inter-related hypotheses to explain how the loss of *TP53* could lead to the loss of differential signaling in the *TGFB1*-STN, thus explaining how *TGFB1* could change from a tumour suppressor to a growth promoter in HCC cells.

Our hypotheses could provide directions to biologists in their research into HCC. Our first hypothesis suggested that the loss of *TP53* expression would increase *EP300* availability to oncogenic proteins. To verify this hypothesis, the biologist would need to measure the molecular abundance of all proteins in HCC cells using a quantitative high-throughput technology that had a sensitivity of 100 protein molecules and then verifed all the protein neighbours of the signal integrator *EP300*. Our second hypothesis suggested that *E2F1*-induced microRNAs might silence many of the tumour suppressors that interacted with the signal integrator *EP300*. The verification of this hypothesis would require a technology platform that could measure both microRNA and protein-coding gene expression in parallel. At present, microRNA expression data were seriously under-sampled. With more than 200 human microRNA genes being discovered today, the existing published dataset was collected from fewer than 40 patients [104]. A larger survey on a few hundred patient samples would be necessary. If our hypotheses could be verified, the new knowledge might provide new diagnostic methods for the detection of HCC or new drug targets for the treatment of HCC.

7.2.7. Future Direction for Bio-informatics Visualization

As mentioned in section 7.2.1, our focus has been on the metaphor and effectiveness of model visualization. In our visual framework, we assumed that a biologist will use the framework as a tool for carrying out data analysis based on a linear sequence of visualization. In reality, systems biology research is a complex interplay between information visualization, statistical and mathematical analysis, and high-throughput technologies. Examples of such can be found in references [179-181]. For this reason, the 'user and tool' model breaks down almost completely because there is no 'one tool fits all'. As such, systems biology research needs a federated model which allows the user to design

the entire project involving a network of high-throughput technologies, visualization methods, statistical algorithm, mathematical algorithms, and databases. This federated model will have to be evaluated using the Knowledge Precept Model proposed by Amar and Stasko [4] based on two types of knowledge precepts. The first is the *worldview precept*. Elements of this include (1) exposing important domain parameters, (2) exposing multivariate explanation, and (3) facilitating hypothesis testing. The second is the *rationale precept*. Elements of this include (1) exposing uncertainty, (2) concretizing relationships discovered in the data, and the usefulness of these relationships. The incorporation of network statistics to network visualization should provide the biologist a signpost for gauging the trustworthiness of the knowledge precepts. At the time of writing, there remains a research project that investigates the aforementioned.

{End of Chapter 7}

Bibliography

[1] A. Adai, S. Date, S. Wieland, and E. Marcotte, "LGL: creating a map of protein function with an algorithm for visualizing very large biological networks", *Journal of Molecular Biology*, vol. 340, pp. 179-190, 2004.

[2] A. Ahmed, T. Dwyer, M. Forster, et al., "GEOMI: Geometry for maximum insight", *Lecture Notes in Computer Science*, vol. 3843, Springer, pp. 468-479, 2006.

[3] U. Alon, "Network Motifs in Developmental, Signal Transduction, and the Neuronal Networks", *An Introduction to Systems Biology: Design principles of biological circuits*, Chapman & Hall/CRC Mathematical and Computational Biology Series, pp. 97-134, 2007.

[4] R. A. Amar and J. T. Stasko, "Knowledge precepts for design and evaluation of information visualizations", *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, pp. 432-442, 2005.

[5] E. H. Baehrecke, N. Dang, K. Babaria, and B. Shneiderman, "Visualization and analysis of microarray and gene ontology data with treemaps", *BMC Bioinformatics*, vol. 5, pp. 84-96, 2004.

[6] A. L. Barabási and Z. Oltvai, "Network biology: understanding the cell's functional organization", *Nature Reviews Genetics*, vol. 5, pp. 101-113, 2004.

[7] A. Barsky, J. L. Gardy, R. E. W. Hancock, and T. Munzner, "Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation", *Bioinformatics*, vol. 23, pp. 1040-1042, 2007.

[8] L. Bartholin, L. L. Wessener, J. M. Chirgwin, and T. A. Guise, "The human Cyr61 gene is a transcriptional target of transforming growth factor beta in cancer cells", *Cancer Letters*, vol. 246, pp. 230-236, 2006.

[9] V. Batagelj and A. Mrvar, "Pajek-analysis and visualization of large networks", *Lecture Notes in Computer Science*, vol. 2265, Springer, pp. 477-478, 2001.

[10] U. Baur and U. Brandes, "Crossing reduction in circular layouts", *Lecture Notes in Computer Science*, vol. 3353, Springer, pp. 332-343, 2004.

[11] M. Y. Becker and I. Rojas, "A graph layout algorithm for drawing metabolic pathways", *Bioinformatics*, vol. 17, pp. 461-467, 2001.

[12] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, W. H. Freeman, 2002.

[13] Y. Blat and N. Kleckner, "Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region", *Cell*, vol. 98, pp. 249-259, 1999.

[14] U. Brandes, T. Dwyer, and F. Schreiber, "Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions", *Journal of Integrative Biology*, pp. 119-132, 2004.

[15] B. J. Breitkreutz, C. Stark, T. Reguly, et al., "The BioGRID Interaction database: 2008 update", *Nuclei Acids Research* (database issue), vol. 36, pp. D637-640, 2008.

[16] R. Brosh, R. Shalgi, A. Liran, et al., "p53-repressed miRNAs are involved with E2F in a feed-forward loop promoting proliferation", *Molecular Systems Biology*, vol. 4, pp. 229, 2008.

[17] M. Bruls, K. Huizing, and J. J. van Wijk, "Squarified Treemaps", *Proceedings of the joint Eurographic and IEEE TVCG Symposium on Visualization*, Eurographics Association, pp. 33-42, 2000.

[18] J. A. Burger and T. J. Kipps, "CXCR4: a key receptor in the crosstalk between tumor cells and their microenvironment", *Blood*, vol. 107, pp. 1761-1767, 2006.

[19] W. C. Burhans and M. Weinberger, "DNA replication stress, genome instability, and aging", *Nuclei Acids Research*, vol. 35, pp. 7545-7556, 2007.

[20] D. Y-. K. But, C-. L. Lai, and M-. F. Yuen, "Natural history of hepatitis-related hepatocellular carcinoma", *World Journal of Gasteroenterology*, vol. 14, pp. 1652-1656, 2008.

[21] S. Camilleri-Broët, I. Cremer, B. Marmey, et al., "TRAF4 overexpression is a common characteristic of human carcinomas", *Oncogene*, vol. 26, pp. 142-147, 2007.

[22] S. K. Card, J. D. MacKinlay, B. Shneiderman, and M. Card, *Readings in Information Visualization: Using Vision to Think*, Academic Press, 1999.

[23] S. Carpendale and A. Agarawala, "PhylloTrees: Harnessing nature's phyllotactic patterns for tree layout", *IEEE Symposium on Information Visualization 2004*, IEEE Computer Society Press, pp. 215.3, 2004.

[24] A. Castro, C. Bernis, S. Vigneron, J-. C. Labbé, and T. Lorca, "The anaphase-promoting complex: a key factor in the regulation of cell cycle", *Oncogene*, vol. 24, pp. 314-325, 2005.

[25] B. Charpentier, V. Bardey, N. Robas, and C. Branlant, "The EIIGlc protein is involved in glucose mediated activation of Escherichia coli gapA and gapB pgk transcription", *Journal of Bacteriology*, vol. 180, pp. 6476-6483, 1998.

[26] D. Chattopadhyay, D. M. Manas, and H. L. Reeves, "The development of targeted therapies for hepatocellular cancer", *Current Pharmaceutical Design*, vol. 13, pp. 3292-3300, 2007.

[27] C. Chen and M. Czerwinski, "Empirical evaluation of information visualizations: an introduction", *International Journal of Human-Computer Studies*, vol. 53, pp. 631-635, 2000.

[28] P. P. Chen, W. J. Li, Y. Wang, et al., "Expression of Cyr61, CTGF, and WISP-1 correlates with clinical features of lung cancer", *PLoS ONE*, vol. 2, article no. e5, 2007.

[29] X. Chen, S. T. Cheung, S. So, et al., "Gene expression patterns in human liver cancers", *Molecular Biology of the Cell*, vol. 13, pp. 1929-1939, 2002.

[30] D. M. Chetkovich, R. C. Bunn, S. H. Kuo, Y. Kawasaki, M. Kohwi, and D. S. Bredt, "Postsynaptic targeting of alternative postsynaptic density-95 isoforms by distinct mechanisms", *Journal of Neuroscience*, vol. 22, pp. 6415-6425, 2002.

[31] T. L. Chmielewski, D. F. Dansereau, and J. L. Moreland, "Using common region in node-link displays: The role of field dependence/independence", *Journal of Experimental Education*, vol. 66, pp. 197-207, 1998.

[32] C. Christensen, J. Thakar, and R. Albert, "Systems-level insights into cellular regulation: inferring, analyzing, and modeling intracellular networks", *Institution of Engineering and Technology Systems Biology*, vol. 1, pp. 61-77, 2007.

[33] T. W. Chung, Y. C. Lee, and C. H. Kim, "Hepatitis B viral HBx induces matrix metalloproteinase-9 gene expression through activation of ERKs and PI-3K/AKT pathways", *FASEB Journal*, vol. 18, pp. 1123-1125, 2004.

[34] Q. Cui, Z. Yu, E. O. Purisma, and E. Wang, "Principles of microRNA regulation of a human cellular signaling network", *Molecular Systems Biology*, vol. 2, article no. 46, 2006.

[35] Q. Cui, Y. Ma, M. Jaramillo, et al., "A map of human cancer signaling", *Molecular Systems Biology*, vol. 3, article no. 152, 2007.

[36] C. J. Dafonesca, F. Shu, and J. J. Zhang, "Identification of two residuals in MCM5 critical for the assembly of the MCM complexes and Stat1-mediated transcription

activation in response to IFN-γ", *Proceedings of the National Academy of Sciences USA*, vol. 98, pp. 3034-3039, 2001.

[37] D. Das, Z. Nahlě, and M. Q. Zhang, "Adaptively inferring human transcriptional subnetworks", *Molecular Systems Biology*, vol. 2, article no. 2006.0029, 2006.

[38] E. H. Davidson, "The regulatory genome for animal development", *The Regulatory Genome*, Academic Press, pp. 22, 2006.

[39] S. de la Fuente van Bentem, W. I. Mentzen, A. de la Fuente, and H. Hirt, "Towards functional phosphoproteomics by mapping differential phosphorylation events in signaling networks", *Proteomics*, vol. 8, pp. 4453-4465, 2008.

[40] N. C. Denko, "Hypoxia, HIF1, glucose metabolism", *Nature Reviews Cancer*, vol. 8, pp. 705-713, 2008.

[41] S. W. Doniger, N. Salomonis, K. D. Dahlquist, et al., "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data", *Genome Biology*, vol. 4, pp. R7, 2003.

[42] U. Dogrusöz, E. Z. Erson, E. Giral, et al., "PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization", *Bioinformatics*, vol. 22, pp. 374-375, 2006.

[43] P. Duesberg, R. Stindl, R. Li, R. Helhmann, and D. Rasnick, "Aneuploidy versus gene mutation as cause of cancer", *Current Science*, vol. 81, pp. 490-500, 2001.

[44] P. Eades, "A heuristic for graph drawing", *Congressus Nemerantium*, vol. 42, pp. 149-160, 1984.

[45] H. B. El-Serag and K. L. Rudolph, "Hepatocellular carcinoma: epidemiology and molecular carcinogenesis", *Gastroenterology*, vol. 132, pp. 2557-2576, 2007.

[46] A. Ergün, C. A. Lawrence, M. A. Kohanski, T. A. Brennan, and J. J. Collins, "A network biology approach to prostate cancer", *Molecular Systems Biology*, vol. 3, pp. 82, 2007.

[47] P. A. Farazi and R. A. DePinho, "Hepatocellular carcinoma pathogenesis: from genes to environment", *Nature Reviews Cancer*, vol. 6, pp. 674-687, 2006.

[48] D. G. Feitelson and M. Treinin, "The blueprint for life", *Computer*, vol. 35, pp. 34-40, 2002.

[49] R. Fleischer and C. Hirch, "Graph drawing and its applications", *Lecture Notes in Computer Science*, vol. 2025, Springer, pp. 1-22, 2001.

[50] V. Fogal, M. Gostissa, P. Sandy, et al., "Regulation of p53 activity in nuclear bodies by a specific PML isoform", *EMBO Journal*, vol. 19, pp. 6185-6195, 2000.

[51] D. N. Frick and C. C. Richardson, "DNA primases", *Annual Review in Biochemistry*, vol. 70, pp. 39-80, 2001.

[52] A. Frick, et al., "A fast adaptive layout algorithm for undirected graphs", *Lecture Notes in Computer Science*, vol. 894, Springer, pp. 388-403, 1994.

[53] B. Fry, "Computational information design", *Doctor of Philosophy Dissertation*, Massachusetts Institute of Technology, pp. 14, 2004.

[54] D. C. Y. Fung, S-. H. Hong, K. Xu, and D. Hart, "Visualizing the Gene Ontology-annotated clusters of co-expressed genes: a two-design study", *Proceedings of the Fifth International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics (MediVis 2008)*, IEEE Computer Society Press, pp. 9-14, 2008.

[55] D. C. Y. Fung, S-. H. Hong, D. Koschützki, F. Schreiber, and K. Xu, "2.5D visualization of overlapping biological networks", *Journal of Integrative Biology*, pp. 90, 2008.

[56] D. C. Y. Fung, S-. H. Hong, D. Koschützki, F. Schreiber, and K. Xu, "Visual analysis of overlapping biological networks", *Proceedings of the 13th International Conference Information Visualization (InfoVis 2009)*, IEEE Computer Society Press, pp. 337-342, 2009.

[57] S. Gama-Castro, et al., "RegulonDB version 6.0: gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation", *Nuclei Acids Research*, vol. 36, pp. D120-124, 2008.

[58] G. Gamberoni, S. Storari, and S. Volinia, "Finding biological process modifications in cancer tissues by mining gene expression correlations", *BMC Bioinformatics*, vol. 7, pp. 6-15, 2006.

[59] R. A. Gatenby and R. J. Gillies, "Perspectives: A microenvironmental model of carcinogenesis", *Nature Reviews Cancer*, vol. 8, pp. 56-60, 2008.

[60] Gene Ontology Consortium, "The Gene Ontology (GO) project in 2006", *Nuclei Acids Research* (database issue), vol. 34, pp. D322-326, 2006.

[61] N. Gershon and S. G. Eick, "Information Visualization", *IEEE Computer Graphics and Applications*, vol. 14, pp. 29-31, 1997.

[62] A. K. Ghosh and J. Varga, "The transcriptional coactivator and acetyltransferase p300 in fibroblast biology and fibrosis", *Journal of Cell Physiology*, vol. 213, pp. 663-671, 2007.

[63] A. Goesmann, M. Haubrock, F. Meyer, J. Kalinowski, and R. Giegerich, "PathFinder: reconstruction and dynamic visualization of metabolic pathways", *Bioinformatics*, vol. 18, pp. 124-129, 2002.

[64] M. A. Gonzalez, K. K. Tachibana, R. A. Laskey, and N. Coleman, "Control of DNA replication and its potential clinical exploitation", *Nature Reviews Cancer*, vol. 5, pp. 135-141, 2005.

[65] T. Guida, G. Salvatore, P. Faviana, et al., "Mitogenic effects of the up-regulation of minichromosome maintainence proteins in anaplastic thyroid carcinoma", *Journal of Clinical Endocrinology and Metabolism*, vol. 90, pp. 4703-4709, 2005.

[66] J. D. Han, N. Bertin, T. Hao, et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network", *Nature*, vol. 430, pp. 88-95, 2004.

[67] D. Hancock, M. Wilson, G. Velarde, et al., "maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination", *BMC Bioinformatics*, vol. 6, pp. 264, 2005.

[68] D. Hanahan and R. Weinberg, "The hallmarks of cancer", *Cell*, vol. 100, pp. 57-70, 2000.

[69] H. Hartson and D. Hix, *Developing User Interfaces: Ensuring Usability through Product and Process*. John Wiley and Sons, 1993.

[70] L. H. Hartwell, J. J. Hopefield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology", *Nature*, vol. 402(suppl), pp. C47-C52, 1999.

[71] T. B. Hashimoto, M. Nagasaki, K. Kojima, and S. Miyano, "BFL: a node and edge betweenness based fast layout algorithm for large scale network", *BMC Bioinformatics*, vol. 10, pp. 19, 2009.

[72] A. F. Hezel and N. Bardeesy, "LKB1: linking cell structure and tumor suppression", *Oncogene*, vol. 27, pp. 6908-6919, 2008.

[73] M. Holford, N. Li, P. Nadkarni, and H. Zhao, "VitaPad: visualization tools for the analysis of pathway data", *Bioinformatics*, vol. 21, pp. 1596-1602, 2005.

[74] C-. N. Hsu, J-. M. Lai, C-. H. Liu, et al., "Detection of the inferred interaction network in hepatocellular carcinoma from ECHO (Encyclopedia of hepatocellular carcinoma genes online)", *Bioinformatics*, vol. 8, pp. 66, 2007.

[75] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks", *Nature*, vol. 411, pp. 41-42, 2001.

[76] W. Johnston, "Model Visualization", *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kauffman, pp. 223-228, 2001.

[77] E. Kalo, Y. Buganim, K. E. Shapira, et al., "Mutant p53 attenuates the SMAD-dependent transforming growth factor beta1 (TGF-beta1) signaling pathway by repressing the expression of TGF-beta receptor type II", *Molecular and Cellular Biology*, vol. 27, pp. 8228-8242, 2007.

[78] M. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs", *Information Processing Letters*, vol. 31, pp. 7-15, 1989.

[79] M. Kanehisa, S. Goto, M. Hattori, et al., "From genomics to chemical genomics: new developments in KEGG", *Nuclei Acids Research*, vol. 34, pp. D354-357, 2006.

[80] R. S. Kerbel, "Molecular origins of cancer—tumor angiogenesis", *New England Journal of Medicine*, vol. 358, pp. 2039-2049, 2008.

[81] R. S. Kerbel, "Supplementary to Molecular origins of cancer—tumour angiogenesis", *New England Journal of Medicine*, vol. 358, pp. 2039-2049, 2008.

[82] H. A. Kestler, A. Müller, T. M. Gress, et al, "VennMaster: Area- proportional Euler diagrams for functional GO analysis of microarrays", *BMC Bioinformatics*, vol. 9, pp. 67, 2008.

[83] E. Klipp, R. Herwig, A. Kowald, H. Wierling, and H. Lehrach, *Systems Biology in Practice. Concepts, Implementation and Application*, Wiley-VCH, Weinheim GmbH, 2006.

[84] V. N. Kim, "MicroRNA biogenesis: coordinated cropping and dicing", *Nature Reviews Molecular Cell Biology*, vol. 6, pp. 376–85, 2005.

[85] K. Kimata, Y. Tanaka, T. Inada, and H. Aiba, "Expression of the glucose transporter gene, *ptsG*, is regulated at the mRNA degradation step in response to glycolytic flux in *Escherichia coli*", *EMBO Journal*, vol. 13, pp. 3587-3595, 2001.

[86] A. Kobsa, "An empirical comparison of three commercial information visualization systems", *Proceedings of IEEE Conference on Information Visualization (InfoVis 2001)*, pp. 123-130, 2001.

[87] K. Kojima, M. Nagasaki, E. Jeong, M. Kato, and S. Miyano, "An efficient grid layout algorithm for biological networks utilizing various biological attributes", *BMC Bioinformatics*, vol. 8, pp. 76, 2007.

[88] G. Kumar and M. Garland, "Visual exploration of complex time-varying graphs", *IEEE Transactions of Visualization and Computer Graphics*, vol. 12, pp. 805-812, 2006.

[89] J. Lee, Z. Li, R. Brower-Sinning, and B. John, "Regulatory circuit of human microRNA biogenesis", *PLoS Computational Biology*, vol. 3, pp. e67, 2007.

[90] E. C. R. Lee and L. A. Megeney, "The yeast kinome displays scale free topology with functional hub clusters", *BMC Bioinformatics*, vol. 6, pp. 271, 2005.

[91] I. Ladunga, "Finding homologs to nucleotide sequences using network BLAST searches", *Current Protocol in Bioinformatics*, Chapter 3 Unit 3.3, 2002.

[92] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, and Z. Bar-Joseph, "Transcriptional regulatory networks in Saccharomyces cerevisiae", *Science*, vol. 298, pp. 799-804, 2004.

[93] S. Legewie, N. Blüthgen, R. Schäfer, and H. Herzel, "Ultrasensitization: switch-like regulation of cellular signaling by transcriptional induction", *PLoS Computational Biology*, vol. 1, pp. e54, 2005.

[94] B. H. Liu, C. Goh, L. L. Ooi, and K. M. Hui, "Identification of unique and common low abundance tumor-specific transcripts by suppression subtractive hybridization and oligonucleotide probe array analysis", *Oncogene*, vol. 27, pp. 4128-4136, 2008.

[95] X. Liu, L. Wang, K. Zhao, et al., "The structural basis of protein acetylation by the p300/CBP transcriptional coactivator", *Nature*, vol. 451, pp. 846-850, 2008.

[96] Y. Liu, H. I. Kao, and R. A. Bambara, "Flap endonuclease 1: a central component of DNA replication", *Annual Review of Biochemistry*, vol. 73, pp. 589-615, 2004.

[97] D. J. Lockhart, H. Dong, M. C. Byrne, et al., "Expression monitoring by hybridization to high-density oligonucleotide arrays", *Nature Biotechnology*, vol. 14, pp. 1675-1680, 1996.

[98] J. Loscalzo, I. Kohane, and A-. L. Barabási, "Perspective: Human disease classification in the postgenomic era: A complex systems approach to human pathobiology", *Molecular Systems Biology*, vol. 3, pp. 124, 2007.

[99] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI", *Nuclei Acids Research*, vol. 33, pp. D54-58, 2005.

[100] P. M. Magwene and J. Kim, "Estimating genomic coexpression networks using first-order conditional independence", *Genome Biology*, vol. 5, pp. R100, 2004.

[101] A. Martinez-Antonio and J. Collado-Vides, "Identifying global regulators in transcriptional regulatory networks in bacteria", *Current Opinion in Microbiology*, vol. 6, pp. 482-489, 2003.

[102] M. Migaud, P. Charlesworth, M. Dempster, et al. "Enhanced long-term potentiation and impaired learning in mice with mutant postsynaptic density-95 protein", *Nature*, vol. 396, pp. 433-439, 1998.

[103] N. Mitsuhashi, S. Kobayashi, T. Doki, et al., "Clinical significance of alpha-fetoprotein: involvement in proliferation, angiogenesis, and apoptosis of hepatocellular carcinoma", *Journal of Gasteroenterology and Hepatology*, vol. 23, pp. e189-197, 2008.

[104] Y. Murakami, T. Yasuda, K. Saigo, et al., "Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues", *Oncogene*, vol. 25, pp. 2537-2545, 2006.

[105] T. W. Nam, S-. H. Cho, D. Shin, et al., "The Escherichia coli glucose transport enzyme IICBGlc recruits the global repressor Mlc", *EMBO Journal*, vol. 20, pp. 491-498, 2001.

[106] C. North, T. M. Rhyne, and K. Duca, "Bioinformatics visualization: introduction to the special issue", *Information Visualization*, vol. 4, pp. 147-148, 2005.

[107] T. Odom, R. D. Dowell, E. S. Jacobsen, et al., "Core transcriptional regulatory circuitry in human hepatocytes", *Molecular Systems Biology*, vol. 2, article no. 2006.0017, 2006.

[108] B. Perbel, "CCN proteins: multifunctional signalling regulators", *Lancet*, vol. 363, pp. 62-64, 2004.

[109] A. Perer and B. Shneiderman, "Balancing systematic and flexible exploration of social networks", *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 693-700, 2006.

[110] A. Perrenoud and U. Sauer, "Impact of global transcriptional regulation by arcA, arcB, cra, crp, cya, fnr, and mlc on glucose catabolism in *Escherichia coli*", *Journal of Bacteriology*, vol. 187, pp. 3171-3179, 2005.

[111] F. Petrocca, A. Vecchione, and C. M. Croce, "Emerging role of *miR-106b-25/miR-17-92* clusters in the control of transforming growth factor β signaling", *Cancer Research*, vol. 68, pp. 8191-8195, 2008.

[112] F. Petrocca, R. Visone, M. R. Onelli, et al., "E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer", *Cancer Cell*, vol. 13, pp. 272-286, 2008.

[113] N. Planque, "Nuclear trafficking of secreted factors and cell-surface receptors", *Cell Communication and Signaling*, vol. 4, pp. 7, 2006.

[114] J. Plumbridge, "Regulation of gene expression in the PTS in Escherichia coli: the role and interactions of Mlc", *Current Opinion in Microbiology*, vol. 5, pp. 187-193, 2002.

[115] G. Rigaut, A. Shevchenko, B. Rutz, et al., "A generic protein purification method for protein complex characterization and proteome exploration", *Nature Biotechnology*, vol. 17, pp. 1030-1032, 1999.

[116] A. W. Rives and T. Galitski, "Modular organization of cellular network", *Proceedings of the National Academy of Sciences USA*, vol. 100, pp. 1128-1133, 2003.

[117] J. D. Saffer, V. L. Burnett, G. Chen, and P. van der Spek, "Visual analytics in the pharmaceutical industry", *IEEE Computer Graphics and Applications*, vol. 24, pp. 10-15, 2004.

[118] S. Saito, H. Ojima, H. Ichikawa, S. Hirohashi, and T. Kondo, "Molecular background of alpha-fetoprotein in liver cancer cells as revealed by global RNA expression analysis", *Cancer Science*, vol. 99, pp. 2402-2409, 2008.

[119] D. G. Stathakis, K. B. Hoover, Z. You, and P. J. Bryant, "Human postsynaptic density-95 (PSD95): location of the gene (DLG4) and possible function in nonneural as well as in neural tissues", *Genomics*, vol. 44, pp. 71-82, 1997.

[120] L-. X. Qin, "An integrative analysis of microRNA and mRNA expression—a case study", *Cancer Informatics*, vol. 6, pp. 369-379, 2008.

[121] J. Rachlin, D. D. Cohen, C. Cantor, and S. Kasif, "Biological context networks: a mosaic view of the interactome", *Molecular Systems Biology*, vol. 2, article no. 66, 2006.

[122] G. Rao and D. Mingay, "Reports on usability testing of Census Bureau's dynamaps CD-ROM product", http://infovis.cs.vt.edu/cs5764/papers/dynamapsUsability.pdf, 2001.

[123] S. Raza, K. A. Robertson, P. A. Lacaze, et al., "A logic-based diagram of signaling pathways central to macrophages", *BMC Systems Biology*, vol. 2, pp. 36, 2008.

[124] C. Reas and B. Fry, *Processing:A Programming Handbook for Visual Designers and Artists,* MIT Press, Massachusetts, 2007.

[125] D. M. Reif, S. M. Dudek, C. M. Shaffer, J. Wang, and J. H. Moore, "Exploratory visual analysis of pharmacogenomic results", *Pacific Symposium on Biocomputing*, vol. 10, pp. 296-307, 2005.

[126] B. Ren, G. Yu, G. C. Tseng, et al., "MCM7 amplification and overexpression are associated with prostate cancer progression", *Oncogene*, vol. 25, pp. 1090-1098, 2006.

[127] T. M. Rhyne, "Visualization Viewpoints: Does the difference between information and scientific visualization really matter?", *IEEE Computer Graphics and Applications*, vol. 23, pp. 6-8, 2003.

[128] J. Rieman, "A field study of exploratory learning strategies", *ACM Transactions on Computer-Human Interaction*, vol. 3, pp. 189-218, 1996.

[129] F. Rodier, J. Campisi, and D. Bhaumik, "The two faces of p53: aging and tumor suppression", *Nuclei Acids Research*, vol. 35, pp. 7475-7484, 2007.

[130] J. Rougemont and P. Hingamp, "DNA microarray data and contextual analysis of correlation graphs", *BMC Bioinformatics*, vol. 4, pp. 15, 2003.

[131] N. Salomonis, K. Hanspers, A. C. Zambon, et al., "GenMAPP 2: new features and resources for pathway analysis", *BMC Bioinformatics*, vol. 8, pp. 217, 2007.

[132] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. "The Database of Interacting Proteins: 2004 update", *Nuclei Acids Research*, vol. 32, pp. D449-451, 2004.

[133] R. Santamaria, R. Theron, and L. Quintales, "A visual analytics approach for understanding biclustering results from microarray data", *BMC Bioinformatics*, vol. 9, pp. 247, 2008.

[134] P. Saraiya, C. North, and K. Duca, "Visualization for biological pathways: requirements analysis, systems evaluation and research agenda", *Information Visualization*, vol. 4, pp. 191-205, 2005.

[135] P. Saraiya, C. North, V. Lam, and K. A. Duca, "An insight-based longitudinal study of visual analytics", *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 1511-1522, 2006.

[136] E. W. Sayers, T. Barrett, D. A. Benson, et al., "Database resources of the National Center for Biotechnology Information", *Nuclei Acids Research*, vol. 37, pp. D5-15, 2009.

[137] R. A. Sclafani and T. M. Holzen, "Cell cycle regulation of DNA replication", *Annual Review on Genetics*, vol. 41, pp. 237-280, 2007.

[138] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel, "Global and local architecture of the mammalian microRNA–transcription factor regulatory network", *PloS Computational Biology*, vol. 3, pp. e131, 2007.

[139] S. Shankar and R. K. Srivastava, "Histone deacetylase inhibitors: mechanisms and clinical significance in cancer: HDAC inhibitor-induced apoptosis", *Advances in Experimental Medicine and Biology*, vol. 615, pp. 261-298, 2008.

[140] P. Shannon, A. Markiel, O. Ozier, et al., "Cytoscape: A software environment for integrated models of biomolecular interaction networks", *Genome Research*, vol. 13, pp. 2498-2504, 2003.

[141] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations", *Proceeding of the IEEE Symposium on Visual Languages*, pp. 336-343, 1996.

[142] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin, "A genome-scale computational study of the interplay between transcriptional regulation and metabolism", *Molecular Systems Biology*, vol. 3, article no. 101, 2007.

[143] T. Soukup, *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, John Wiley and Sons Inc., 2002.

[144] R. Spence, "A framework for navigation", *International Journal of Human-Computer Studies*, vol. 51, pp. 919-945, 1999.

[145] D. Stekel, "Analysis of relationships between genes, tissues or treatments", *Microarray Informatics*, Cambridge University Press, pp. 139-182, 2003.

[146] R. Stevens, C. Goble, P. Baker, and A. Brass, "A classification of tasks in bioinformatics", *Bioinformatics*, vol. 17, pp. 180-188, 2001.

[147] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene co-expression network for global discovery of conserved genetic modules", *Science*, vol. 302, pp. 249-255, 2003.

[148] M. Suderman and M. Hallett, "Tools for visually exploring biological networks", *Bioinformatics*, vol. 23, pp. 2651-2659, 2007.

[149] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for visual understanding of hierarchical system structures", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, pp. 109-125, 1981.

[150] R. Taylor and G. R. Stark, "Regulation of the G2/M transition by p53", *Oncogene*, vol. 20, pp. 1803-1815, 2001.

[151] The UniProt Consortium, "The universal protein resource (UniProt)", *Nuclei Acids Research*, vol. 35, pp. D193-197, 2007.

[152] V. S. Tompkins, J. Hagen, A. A. Frazier, et al., "A novel nuclear interactor of ARF and MDM2 (NIAM) that maintains chromosomal stability", *Journal of Biological Chemistry*, vol. 282, pp. 1322-1333, 2007.

[153] M. A. Valencia-Sanchez, J. Liu, G. J. Hannon, and R. Parker, "Control of translation and mRNA degradation by miRNAs and siRNAs", *Genes and Development*, vol. 20, pp. 515-524, 2006.

[154] J. M. G. Vilar, R. Jansen, and C. Sander, "Signal processing in the TGF-β ligand-receptor network", *PLoS Computational Biology*, vol. 2, article no. e3, 2006.

[155] S. Violinia, G. A. Calin, C-. G. Liu, et al., "A microRNA expression signature of human solid tumors defines cancer gene targets", *Proceedings of the National Academy of Sciences U.S.A.*, vol. 103, pp. 2257-2261, 2006.

[156] M. Vivo, R. A. Calogero, F. Sansone, et al., "The human tumor suppressor *arf* interacts with spinophilin/neurabin II, a type 1 protein-phosphatase-binding protein", *Journal of Biological Chemistry*, vol. 276, pp. 14161-14169, 2001.

[157] G. J. Walker and N. K. Hayward, "p16INK4A and p14ARF tumour suppressors in melanoma: lessons from mouse", *Lancet*, vol. 359, pp. 7-9, 2002.

[158] Y. Ward, S. Gupta, P. Jensen, M. Wartmann, R. J. Davis, and K. Kelly, "Control of MAP kinase activation by the mitogen-induced threonine/tyrosine phosphatase PAC1", *Nature*, vol. 367, pp. 651-654, 1994.

[159] C. Ware, *Information Visualization: Perception for Design*, Morgan Kauffman, 2000.

[160] T. S-. F. Wang, "DNA replication in eukaryotic cells", *Annual Review in Biochemistry*, vol. 60, pp. 513-552, 1991.

[161] D. Wigle and I. Jurisica, "Cancer as a system failure", *Cancer Informatics*, vol. 2, pp. 10-18, 2007.

[162] M. Wilkins and S. K. Kummerfeld, "Sticking together? Falling apart? Exploring the dynamics of the interactome", *Trends in Biochemical Sciences*, vol. 33, pp. 195-200, 2008.

[163] D. M. Williams and M. C. Ebach, "Homologues and Homology", *Foundations of Systematics and Biogeography*, Springer, pp. 126-138, 2008.

[164] P. C. Wong, H. Foote, G. Chin, P. Mackey, and K. Perrine, "Graph signatures for visual analytics", *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 1399-1413, 2006.

[165] S. Wu, N. M. Lourette, N. Tolić, et al., "An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications", *Journal of Proteome Research*, vol. 8, pp. 1347-1357, 2009.

[166] Y. Wu, Y. Cai, J. Aquilo, T. Dai, Y. Ao, and Y. J. Wan, "RXRalpha mRNA expression is associated with cell proliferation and cell cycle regulation in Hep3B cell", *Experimental and Molecular Pathology*, vol. 76, pp. 24-28, 2004.

[167] C. –H. Yeang and M. Vingron, "A joint model of regulatory and metabolic networks", *BMC Bioinformatics*, vol. 7, pp. 332, 2006.

[168] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics", *PloS Computational Biology*, vol. 3, article no. e59, 2007.

[169] B. R. Zeeberg, H. Qin, S. Narasimhan, et al., "High-Throughput GOMiner, an industrial strength integrative gene ontology tool for interpretation of multiple microarray experiments, with applications to studies of Common Variable Immune Deficiency (CVID)", *BMC Bioinformatics*, vol. 6, pp. 168, 2005.

[170] B. Zhang, X. Pan, G. P. Cobb, and T. A. Anderson, "microRNAs as oncogenes and tumor suppressors", *Developmental Biology*, vol. 302, pp. 1-12, 2007.

[171] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy, "GOTree Machine (GOTM) a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies", *BMC Bioinformatics*, vol. 5, pp. 16, 2004.

[172] T. Zuk and M. S. T. Carpendale, "Theoretical analysis of uncertainty visualizations", *Proceedings of SPIE and IS&T Conference on Electronic Imaging, Visualization and Data Analysis*, vol. 6060, pp. 606007, 2006.

[173] M. C. F. de Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: a survey", *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 378-394, 2003.

[174] D. A. Keim, "Information visualization and visual data mining", *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 1-8, 2002.

[175] I. Herman, "Graph visualization and navigation in information visualization: a survey", *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, pp. 24-43, 2000.

[176] I. G. Tollis, G. di Battista, P. Eades, and R. Tamassia, *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice Hall, 1998.

[177] U. Brandes, "Drawing on physical analogies", *Lecture Notes in Computer Science*, vol. 2025, Springer, pp. 71-86, 2001.

[178] F. Iragne, M. Nikolski, B. Mathieu, D. Auber, and D. Sherman, "ProViz: protein interaction visualization and exploration", *Bioinformatics*, vol. 21, pp. 272-274, 2005.

[179] K. C. Gunsalus, H. Ge, A. J. Schetter, et al., "Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis", *Nature*, vol. 436, pp. 861-865, 2005.

[180] M. A. Pujana, J. D. Han, L. M. Starita, et al., "Network modeling links breast cancer susceptibility and centrosome dysfunction", *Nature Genetics*, vol. 39, pp. 1338-1349, 2007.

[181] E. M. Schmid and H. T. McMahon, "Integrating molecular and network biology to decode endocytosis", *Nature*, vol. 448, pp. 883-888, 2007.

[182] A. Budhu, M. Forgues, Q. H. Ye, et al., "Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment", *Cancer Cell*, vol. 10, pp. 99-111, 2006.

[183] S. H. Hong, "MultiPlane: a new framework for drawing graphs in three (2½) dimensions", *AVI'06*, Venice, Italy, May 2006.

[184] C. J. Alpert and A. B. Kahng, "Recent developments in netlist partitioning: A survey", *Integration: the VLSI Journal*, vol. 19, pp. 1-81, 1995.

[185] G. Even, J. Naor, S. Rao, and B. Schieber, "Fast approximate graph partitioning algorithms", *SIAM Journal on Computing*, vol. 28, no. 6, pp. 2187-2214, 1999.

[186] M. Kaufmann and D. Wagner, (ed), "Drawing graphs: methods and models", *Lecture Notes in Computer Science Tutorial 2025*, Springer Verlag, 2001.

[187] J. Ho and S. Hong, "Drawing clustered graphs in three dimensions", *Proceedings of Graph Drawing*, 2005.