

# Cluster analysis of the p53 genetic regulatory network: Topology and biology

Jennifer S. Hallinan

**Abstract**—In this paper we describe a network module detection approach which combines a rapid and robust clustering algorithm with an objective measure of the coherence of the modules identified. The approach is applied to the network of genetic regulatory interactions surrounding the tumor suppressor gene p53. This algorithm identifies ten clusters in the p53 network, which are visually coherent and biologically plausible.

**Index Terms**—Genetic regulatory networks, cancer, robustness, cluster analysis.

## I. INTRODUCTION

Cancer is a systemic disease. It has different manifestations in different tissues—idiosyncrasies which are of fundamental importance to patients and their clinicians—but underlying all forms of cancer is a shared network of genetic regulatory interactions. The 30,000-odd genes of the human genome probably produce several hundred thousand proteins; the activities of thousands of genes have been observed, in DNA microarray experiments, to change in cancerous cells [1]. And yet it appears that mutations in only four to six genes are sufficient to trigger the process of malignant transformation[2].

Research into cancer has, so far, primarily taken the form of intensive investigation of individual genes and a small number of interactions between genes. Recently, however, it has been suggested that cancer can only be fully understood, and hence combated, at the systems level. Several researchers have published small genetic regulatory networks (GRNs) for cancer genes (e.g. [1], [3]) and our laboratory is currently in the process of compiling a comprehensive database of regulatory interactions surrounding the key tumor suppressor gene p53.

We selected p53 as the starting point of the network because it is widely recognized as one of the most important cancer genes. p53 and a gene known as Mdm2 engage in a complex feedback loop in which the Mdm2 protein inhibits the p53 protein by binding to it and marking it for degradation. p53 in turn promotes the transcription of Mdm2. This delicately balanced loop keeps p53 present at low levels in the cell.

When the cell is subjected to stress or DNA damage the loop is disrupted and p53 levels rise, triggering a cascade of molecular events via a complex web of genetic regulatory interactions. This cascade eventually leads to either a delay in the cell cycle, giving time for the DNA repair machinery to operate, or, if the damage is too severe, to programmed cell death, known as apoptosis. If p53 is incapacitated the ability of the cell to cope with damage to the DNA is severely impaired; p53 has been found to be mutated in more than 50% of solid tumours, and inactivated in many others [3]. Consequently there has been considerable research into p53, its nature and interactions. Tens of thousands of papers have been published on this gene since its discovery in 1979; a search of the PubMed literature database in August 2004 yielded 32,454 hits for "p53". Restricting the search to "p53 regulation" reduced the number of hits returned to a (slightly) more manageable 7,567.

The computational modelling of genetic regulatory networks has been of interest to researchers since Kauffman developed the concept of Random Boolean Networks (RBNs) in the 1960s [4]. In an RBN genes are modelled as nodes in a network, with interactions between genes or gene products represented as directed arcs between nodes. Time is discreet, and at any time point a node may be in one of two states—on or off, represented as 1 or 0. Each node updates itself using a random Boolean function of its inputs

RBNs have been widely studied as GRN models because, when appropriately parameterized, they exhibit similar gene expression dynamics to living cells. Depending upon factors such as average connectivity, mode of update, and proportion of inhibitory links, the network may collapse rapidly to a stable state (a point attractor), cycle through a small number of states (limit cycle behaviour) or exhibit no discernable pattern of expression ('chaotic' behaviour)<sup>1</sup> [5].

From the point of view of biologists interested in a specific disease, such as cancer, RBNs are too general a model. In particular, the random manner in which interactions between nodes are assigned, and the use of random node update tables, means that while RBNs may model large scale GRN behaviour adequately, they do not provide deep insight into specific systems. One alternative which has been used by some researchers is to model GRNs in exquisite detail, including

Manuscript received June 19, 2004 This work was supported in part by UQ ECR grant no. 2003001500.

J. S. Hallinan is with the ARC Centre for Bioinformatics at The University of Queensland, Australia 4072 (phone: +61 7 3346 2615 fax: +61 7 3346 2103; e-mail: j.hallinan@imb.uq.edu.au).

<sup>1</sup> A network of finite size cannot exhibit truly chaotic behaviour, since the number of states which it can visit are finite. The term 'chaotic' is, however, widely used for networks with no discernable expression pattern.

factors such as DNA and RNA transcription and translation rates, mRNA and protein degradation times, and reaction kinetics (e.g.[6]). This approach is only computationally feasible for small numbers of genes, and detailed information about reaction and degradation rates is not available for large numbers of genes and proteins.

Eukaryotic gene regulation is a complex and delicately balanced process; multiple proteins are often required, binding to each other and to the promoter regions of the DNA, to initiate or inhibit DNA transcription. The details of the regulatory process are not fully understood for most genes, making it impossible to model large genetic regulatory networks in detail.

The type of information which is available on a large scale is at a somewhat higher level of abstraction, the level of promotion and inhibition of gene transcription. By modelling the p53 GRN at an intermediate level of abstraction we hope to be able to use analytical approaches developed using abstract models such as RBNs while retaining a sound biological basis for the topology of the networks.

## II. METHODS

### A. The p53 Genetic Regulatory Network

The p53 network is built from a database of genetic regulatory interactions culled manually from the biomedical literature. In order to build the database the biomedical literature was scanned from individual regulatory interactions between genes and proteins known to be important in protecting cells from malignant transformation, starting with those proteins known to interact with p53. Individual interactions are then assembled into a regulatory network.

In the network each node represents a protein. In most cases each gene produces a single protein, but there are several instances of a single gene giving rise to multiple proteins, usually by alternative splicing of exons. In such a case, each protein is represented as an individual node. Arcs between nodes represent regulatory interactions, either positive or negative.

The p53 network to date consists of 91 interactions between 67 proteins. Because of the manner in which the data has been collected, starting from a single protein and adding relationships between that protein and others known to interact with it, the network consists of a single connected component. Of the regulatory relationships 72 (79%) are excitatory (Figure 1). Inhibitory interactions do not appear to be scattered randomly throughout the network, as tends to be the case in RBNs; instead they are centred around a small number of genes. Since some nodes in the network represent oncogenes (which are usually inactive), while some represent tumour suppressor genes (which are usually active) the concentration of inhibitory links around specific genes does make sense.



Figure 1. The p53 network. Nodes are genes and arcs are regulatory interactions between genes or gene products. Dark lines indicate inhibitory interactions, while light lines are excitatory interactions.

### B. Network statistics

Interaction networks from domains as distinct as ecology, economics, sociology and molecular genetics have been found to have a number of characteristics in common. These include a scale-free connectivity distribution. A scale-free network has no characteristic number of connections per node as does a randomly constructed network; the probability  $P(k)$  of finding a node with  $k$  connections follows a power law:

$$P(k) \propto k^{-\gamma} \quad (1)$$

where the scaling exponent,  $\gamma$ , varies with the degree distribution of the network. When the degree,  $k$ , of the nodes of a scale-free network is plotted against the probability of occurrence of that degree,  $P(k)$ , on a log-log scale the data forms a straight line, the slope of which is  $\gamma$  [7].

Many interaction networks have also been shown to exhibit “small-world” properties [8]. A small-world network has a small but significant number of short-cut connections between otherwise widely separated nodes. This organization leads to characteristic topological features, including a small diameter (defined as the longest of the shortest paths between every pair of nodes in the network). Small-world networks also have a large average cluster coefficient,  $C$ , compared with randomly connected networks with the same number of nodes and links. Cluster coefficient is a measure of the extent to which the neighbours of a node are linked to each other:

$$C = \frac{1}{n} \sum_{i=1}^n \frac{C_i}{N_i(N_i - 1)/2} \quad (2)$$

where  $n$  is the number of nodes in the network,  $C_i$  is the number of connections between neighbours of node  $i$  and  $N_i$  is the number of neighbours of node  $i$ .

We used the network analysis package Pajek to analyze the p53 network in terms of the above statistics [9].

### C. Iterative Vector Diffusion

Many networks appear to be organized into a number of modules. A module is generally defined as a subnetwork of a graph, the nodes of which have more connections to other nodes within the module than to external nodes [10]-[13]. We would expect the p53 network to be modular; several different types of intracellular interaction networks have already been found to be have a modular organization [14]-[16]. More specifically, Thieffry & Romero have shown in the case of RBN models that there is a functional requirement for feedback loops in GRNs. They found that "the fraction of the total number of consistent combinations of [network] parameter values that make a circuit functional decreases geometrically with the circuit length. From a biological point of view, this suggests that regulatory networks could be decomposed into small and relatively independent feedback circuits or 'regulatory modules'" [17].

We identified clusters of nodes in the network using a two-step process. The similarity matrix for the network was generated using an iterated vector diffusion algorithm [14], and clusters detected using a standard  $k$ -means clustering algorithm.

For the purpose of running the iterative vector diffusion algorithm, the directed p53 network was converted to an undirected network. "Closeness" between nodes is thus calculated based only the pattern. of links, and not the direction of those links.

The algorithm is initialized by assigning to each vertex a binary vector of length  $n$ , initialized to

$$v_{i,j} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \quad (3)$$

where  $i$  is an index into the vector and  $j$  is the unique number assigned to a given node. This generates an initial set of  $n$  orthogonal vectors.

The algorithm proceeds iteratively. At each iteration an edge from the network is selected at random and the vectors associated with each of its nodes are moved towards each other by adding a small amount,  $\delta$ , to each element of the vector. This vector diffusion process is iterated until a stopping criterion is met. We chose to compute a maximum number of iterations as the stopping criterion. This number,  $iter$ , is dependant upon both the number of connections in the network,  $c$ , and the size of  $\delta$ , such that

$$iter = c * \left( \frac{\alpha}{\delta} \right) \quad (4)$$

where  $\alpha$  is the average amount by which a vector is changed in the course of the run. A value for  $\alpha$  of 0.1 was selected empirically in trials on artificially generated networks. The stopping criterion was not found to significantly affect the clustering results.

### D. Clustering

At the end of the vector diffusion process the vectors, initially mutually orthogonal, are clustered in  $n$ -dimensional

space. To reduce the dimensionality of the data set, the vectors are then subjected to clustering using the  $k$ -means clustering algorithm implemented by [19]. Since the iterative vector diffusion algorithm is stochastic, generating a slightly different set of vectors on each run, the algorithm was run ten times on the p53 network to generate ten sets of  $n$ -dimensional vectors. Each set of vectors was clustered independently

Most clustering algorithms,  $k$ -means included, require the user to specify in advance the number of clusters to be identified in the data. The algorithm will then proceed to identify that number of clusters, whether or not they have any meaning in the context of the data. When analyzing a genetic regulatory network, it is important not to have spurious clusters on the one hand, or to erroneously merge independent clusters, on the other. In order to objectively identify the number of clusters inherent in the network, we used a measure we call cluster coherence,  $\chi$  [14]. Cluster coherence is a measure of the relative proportion of links between nodes within and external to a previously identified module, and is defined as

$$\chi = \left( \frac{2k_i}{n(n-1)} \right) - \frac{1}{n} \sum_{j=1}^n \left( \frac{k_{ji}}{(k_{jo} + k_{ji})} \right) \quad (5)$$

where  $k_i$  is the total number of edges between nodes in the module,  $n$  is the number of nodes in the network,  $k_{ji}$  is the number of edges between node  $j$  and other nodes within the module, and  $k_{jo}$  is the number of edges between node  $j$  and other nodes outside the module.  $\chi$  takes a value between -1 (no modularity) and 1 (a fully connected, stand-alone module).

The cluster coherence measure was used to assess the quality of the clusters generated by the  $k$ -means clustering. For each set of vectors we ran multiple cluster analyses, each with a different target number of clusters, with the number of clusters varying between two and twenty. The maximum value of twenty clusters was chosen because with twenty clusters in a 68-node network, each cluster would be expected to contain around four nodes on average. Clusters of fewer than four nodes are unlikely to be biologically significant.

For each clustering the number of clusters and the minimum, maximum and average cluster coherence was recorded. The number of clusters used in the final analysis was selected by inspection of the scatterplot of number of clusters against average cluster coherence. The number of clusters was chosen to correspond to the point at which maximum average cluster coherence was generated.

Each clustering generated a slightly different set of clusters. A consensus clustering was produced by assigning each node to the cluster to which a majority of the runs had assigned it. All further analysis was carried out on the consensus clustering.

## III. RESULTS

### A. Connectivity

Many naturally occurring networks have been found to have

a scale-free pattern of connectivity. The p53 network appears to be no exception, although there are still too few nodes in the network to make such a statement unequivocally (Figure 2). The degree distribution of the nodes, both for in-degree and out-degree closely fits a power law, with an  $R^2$  value of 0.98 for input connectivity and 0.96 for output connectivity.

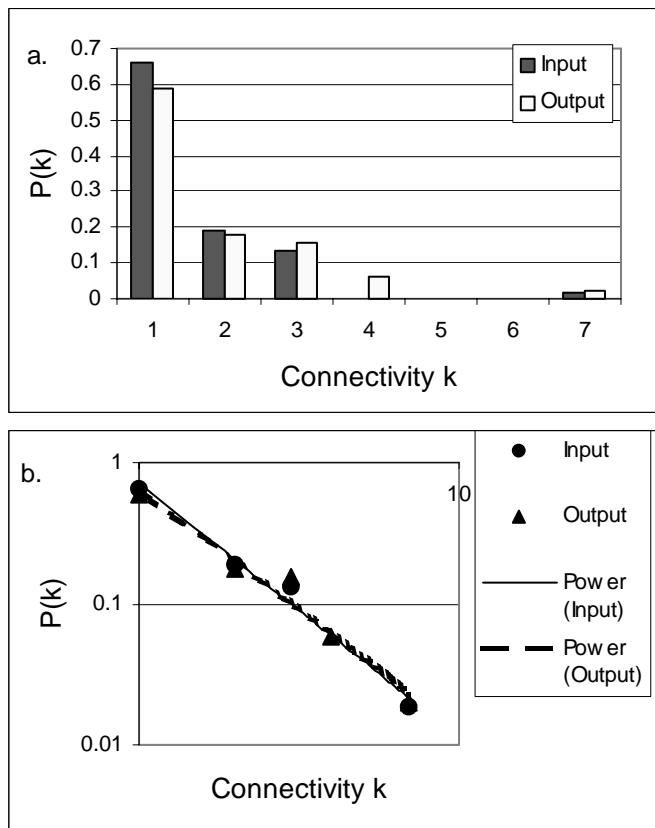


Figure 2. Connectivity of the p53 network. a. Bar chart of input and output connectivity. b. Log-log plot of data in 1a, with lines of best fit for power laws superimposed. The  $R^2$  value for the input connectivity is 0.98, that for output connectivity is 0.96.

The diameter of the network is 9, and the cluster coefficient is 0.0087. The equivalent values for a randomly connected network with the same number of nodes and links would be diameter 14 and cluster coefficient 0.0115. The network does not therefore appear to have small world characteristics.

### B. Cluster identification

As discussed in the Methods section, the number of clusters to be analyzed was selected by inspection of the scatterplot of cluster number against average cluster coherence (Figure 3). Because the iterative vector diffusion algorithm is stochastic, the target number of clusters was not always achieved. The concept of coherence does not make sense for single-node "clusters", so such clusters were eliminated from the analysis.

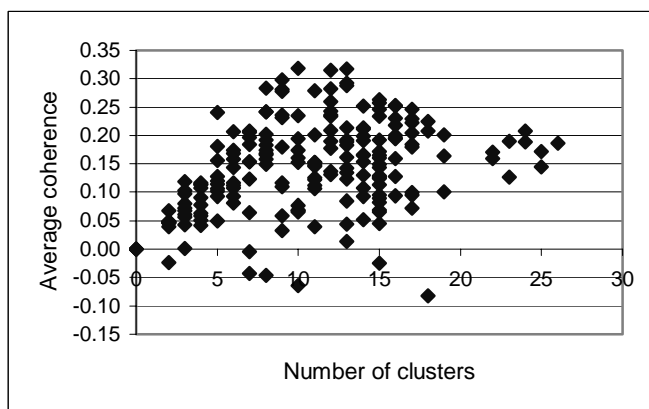


Figure 3. Scatterplot of cluster number against average cluster coherence. Coherence peaks at around ten clusters.

The resulting network, with the clusters indicated by the numbering of the nodes, is shown in Figure 4.

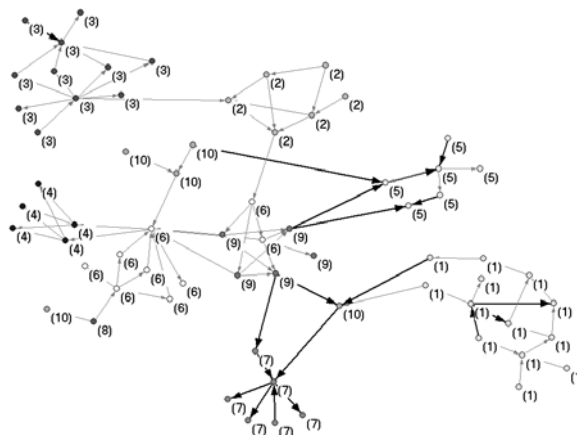


Figure 4. Consensus clustering of the p53 network. The number in brackets beside each node is the cluster to which the node is assigned.

For purposes of comparison a standard graph-partitioning algorithm was also applied to the network. The algorithm used was a  $k$ -core partitioning algorithm, which produces subnetworks in which each vertex has at least  $k$  neighbors in the same core according to the number of input links, output links, or all links. Partitioning on the basis of all links, regardless of direction, as was done for the  $k$ -means based algorithm, produced the network shown in Figure 5.

Three partitions were identified using this algorithm. One of the partitions corresponds closely to cluster 9 in Figure 4, with the addition of two nodes from cluster 6. The members of the other two partitions are scattered throughout the rest of the network, and do not form tightly connected clusters.

Because the algorithm operates on the basis of number of edges per node, it tends to assign low-connectivity peripheral nodes to a one partition and the higher-connectivity internal nodes to another. The clustering arising from the two-step iterative vector diffusion /  $k$ -means clustering appears to give more biologically relevant results.



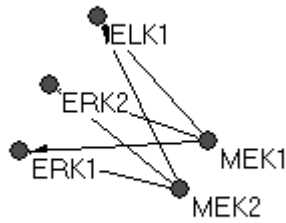


Figure 8. Cluster 4

Cluster 5 contains the conceptual, if not the visual, heart of the network; the bidirectional feedback loop between p53 and Mdm2, as discussed in the Introduction. Also included in this cluster are genes directly regulated by p53. This cluster clearly reflects the incompleteness of the network to date; the p53 protein is known to bind directly to numerous other proteins, none of which have yet been incorporated into the database. Also shown in this cluster is the Human Papilloma Virus E6, which acts to inhibit p53.

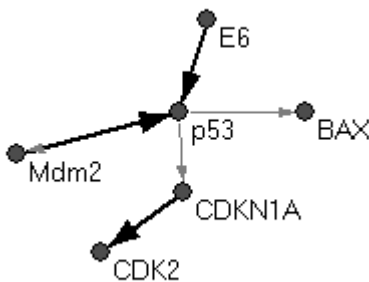


Figure 9. Cluster 5

Cluster 6 is based around the oncogene RAF1, which functions downstream of the RAS family of oncogenes, of which K-RAS is included in this cluster. RAF1 activates MEK1, which triggers a cascade of gene regulation involved in, amongst other things, cell division and apoptosis, defects in both of which are crucial to carcinogenesis. Also in this cluster are the Janus kinases (JAK1, 2 and 3) and their activators.

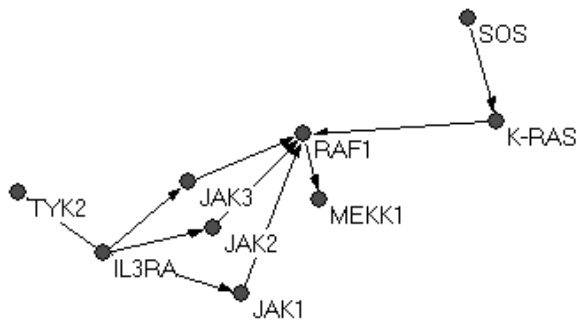


Figure 9. Cluster 6

Cluster 7 consists of the retinoblastoma gene, Rb, and associated transcription factors (E2F1, E2F2 and E2F3).

Retinoblastoma mutations predispose individuals to multiple retinal tumours in childhood. The transcription factors are also known to be important in cell cycle progression, as is the final member of the cluster, CDK6.

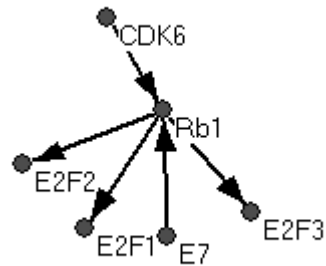


Figure 10. Cluster 7

Clusters 8, 9 and 10 are less clear-cut and convincing as biological modules than the previous ones. The network at this point is quite highly connected, with a number of long-distance links which make cluster membership hard to determine.

Cluster 8 consists solely of the gene IL3. Cluster 9 comprises two related oncogenes, H-RAS and N-RAS, both of which upregulate the cyclin dependant kinase inhibitors INK4A and ARF. Like cluster 4, this cluster is a classic network motif. The addition of the B-RAF oncogene, about which almost nothing is known, is questionable. Apart from the fact that B-RAF, like H-RAS and N-RAS, is an oncogene, there appears to be biological evidence that it is functionally related to the other members of this group. Apparently anomalous cluster memberships such as this will be of particular interest to observe as the network grows and more details of the biological relationships between these proteins are incorporated.

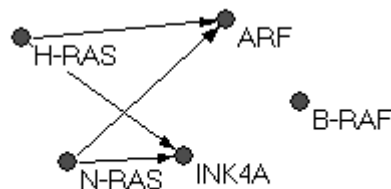


Figure 11. Cluster 9

Cluster 10 is a less-than-convincing unit, comprising a triad representing the fusion product of the BCR gene on chromosome 22 and the ABL gene on chromosome 9. A reciprocal translocation between the two chromosomes at points within these genes produces the Philadelphia chromosome, often found in patients with chronic myelogenous leukemia. The inclusion within this cluster of cyclin dependant kinase 4 (CDK4) and the colony stimulating factor 4 receptor beta gene (IL3RB4) is erroneous.

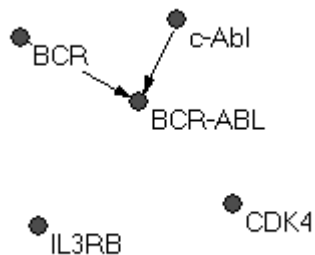


Figure 12. Cluster 10.

#### IV. DISCUSSION

The p53 GRN appears so far to have a scale-free pattern of connectivity, as has been found for several other intracellular interaction networks [24], [14]. The average connectivity is about 1.4, but there are a small number of highly connected hubs which give the network its scale-free character. Interestingly, the network is not a small world network [25], since while its diameter of 9 is very close to the expected value for an equivalent random network (14), the cluster coefficient of this network is also very similar to that expected of a random network (0.0087 and 0.0115 respectively). A small world network is characterized by a large cluster coefficient, which is hypothesized to facilitate local information transfer. It is possible that the strongly directed nature of a GRN means that efficient local information transfer is not a selective advantage to the organism. Alternatively, the low cluster coefficient may be an artefact of the currently relatively small size of the network.

Another striking feature of the network is the manner in which inhibitory links tend to be concentrated around a small number of genes. The majority of the interactions (79%) in the network are positive. The negative interactions are focused around three genes—the tumor suppressor Rb1, the p53/Mdm2 loop and the oncogene c-Abl.

Simulation of GRN dynamics with RBN models has shown that the networks exhibit stable limit cycle dynamics only when correctly parameterized. For an RBN this means an average connectivity of about 2, a high proportion of nodes with canalizing Boolean update functions, or a proportion of inhibitory links of about 0.4 [25]. Networks with sparse connectivity tend to freeze to a point attractor, while those with high connectivity become chaotic. In an RBN inhibitory links tend to be assigned at random and hence are scattered throughout the network. The effect of the concentration of inhibitory links observed in the p53 network upon the dynamics of the network is yet to be examined. The likely effect of inhibitory interactions is further complicated by the existence within the network of both oncogenes and tumor suppressor genes. Inhibition of an oncogene is functionally equivalent to the activation of a tumor suppressor gene such as Rb.

The dynamic behaviour of the network will also be affected by its organization into a number of modules. The clustering

approach used here combines a simple and robust clustering algorithm with an objective measure of the "goodness" of the resulting clustering. Using the combined algorithm we identified ten clusters within the network, most of which are visually coherent and appear to have biological plausibility. Several of them, notably clusters 8 and 9, are typical of the type of network motifs identified by automated algorithms [26],[27], which have been identified as recurrent functional motifs.

The p53 GRN is probably inherently modular, as most biological interaction networks appear to be. However, the network as it stands is incomplete, and no firm conclusions can yet be drawn about the modularity of the whole network. Individual biologists tend to work upon small groups of related genes, an approach which probably biases the generated data in favour of modularity; and there are an unknown number of as-yet unidentified interactions between known genes. Despite these caveats, the approach developed upon this relatively small data set has produced biologically plausible clustering, and will be directly applicable to larger networks as they are accumulated. It is interesting to note in this context that the  $k$ -mode graph partitioning algorithm did not

As the p53 network becomes larger, the question of scalability of the clustering algorithm arises. One advantage of the use of binary vectors is that even very large vectors are easily manipulated by desktop computers. A modification of the current algorithm, using hierarchical rather than  $k$ -means clustering (which is much more computationally intensive) has previously been applied to a network of 1,955 nodes and 1,416 edges and has proven computationally tractable [28]. Although the p53 network may eventually exceed this size, it does not appear that scalability will be a major problem.

The incomplete nature of the current network is not necessarily a drawback. As well as facilitating the development of analytical algorithms, starting with a small network permits monitoring the characteristics of the network as the database grows. This approach should permit us to predict the consequences of incomplete information (and large-scale biological datasets are always incomplete!) upon the topology of the resulting network; for example, if modularity decreases as the network size increases we may be able to predict the degree of modularity of the complete network of several thousand genes well in advance of actually collecting the data.

We have chosen to model the p53 network at an intermediate level of abstraction. There are several reasons for this. We wish to retain as much biological information as possible in the network structure and update rules, in order that analyses of the network structure and dynamics yield useful information which can be applied to the real p53 GRN, producing hypotheses which can be tested *in vivo*. However, detailed biological information simply does not exist for many of the thousands of genes and proteins involved. The choice then becomes whether to build highly detailed models of very small networks, or more abstract models of very large ones.

We have chosen to pursue the second option, largely because of the considerable body of research which has accumulated over the last several decades about the behaviour of large, complex networks. Insights such as the analogy of a cell type with an attractor in gene expression state space and the concept of studying state transitions and limit cycles have arisen from research into large, abstract models of GRNs, and will be readily applicable to our more focused model as it becomes larger and more complete.

A long-term aim of the current modelling work is the identification of gene relationships which may be potential therapeutic targets. The current approach to drug development focuses primarily upon the identification of individual genes which contribute to cancer, and the development of compounds which bind to or in otherwise affect the functioning of the gene. Using a GRN approach should permit the identification of vulnerable interactions, which may prove to be easier to disrupt. Furthermore, a systems biology approach should make it easier to predict potential side-effects of altering a node or an interaction. Systems biology approaches to the study of cancer such as the one described here will build upon, and add value to, traditional reductionist molecular genetics.

## V. CONCLUSIONS

The relationship between network topology and dynamics is of considerable interest to modelers of genetic regulatory networks. In this paper we describe an algorithm for the detection of modularity in an undirected network, and apply it to the GRN surrounding the tumor suppressor gene p53. The modules so detected are biologically plausible, and some of them appear to be self-contained functional units. Future work on this model will examine the relationship between the network topology and gene expression dynamics.

## REFERENCES

- [1] Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C. & Spellman, P. (2000). "Systematic variation in gene expression patterns in human cancer cell lines". *Nature Genetics* Vol. 24, pp. 227 - 235.
- [2] Hahn, W. C. & Weinberg, R. A. 2002. 'Modelling the molecular circuitry of cancer', *Nature Reviews Cancer* vol. 2, p.331 - 341.
- [3] Vogelstein, B., Lane, D. & Levine, A. J. 2000. 'Surfing the p53 network', *Nature* vol. 408, p.307 - 310.
- [4] Kauffman, S. 1969. 'Metabolic stability and epigenesis in randomly constructed genetic nets', *Journal of Theoretical Biology* vol. 22, p.437 - 467.
- [5] Wuensche, A. 1998. 'Genomic regulation modeled as a network with basins of attraction', In Altman, R. B., Dunker, A. K., Hunter, L. & Klien, T. E. (eds.) *Pacific Symposium on Biocomputing '98* Singapore: World Scientific vol. 1998,.
- [6] Mendoza, L. & Alvarez-Buylla, E. R. 1998. 'Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis', *Journal of Theoretical Biology* vol. 193, p.307 - 319.
- [7] Barabasi, A.-L., Albert, R. & Jeong, H. 2000. 'Scale-free characteristics of random networks: The topology of the world-wide web', *Physica A* vol. 281, p.69 - 77.
- [8] Watts, D. J. 1999. *Small Worlds: The Dynamics of Networks between Order and Randomness* Princeton, NJ: Princeton University Press.
- [9] Batagelj, V. & Mrvar, A. 1998. 'Pajek - Program for Large Network Analysis', *Connections* vol. 21, p.47 - 57.
- [10] AnceL, L. W. & Fontana, W. 2002. 'Evolutionary lock-in and the origin of modularity in RNA structure', In Callabaut, W. & Rasskin-Gutman, D. (eds.) *Modularity. Understanding the development and evolution of complex natural systems* Cambridge, MA: MIT Press.
- [11] Calabretta, R., Nolfi, S., Parisi, D. & Wagner, G. P. 1998. 'A case study of the evolution of modularity: Towards a bridge between evolutionary biology, artificial life, neuro- and cognitive science', In Adami, C., Belew, R., Kitano, H. & Taylor, C. (eds.) *Proceedings of the Sixth International Conference on Artificial Life* Cambridge, MA: MIT Press p.275 - 284.
- [12] Csete, M. E. & Doyle, J. C. 2002. 'Reverse engineering of biological complexity', *Science* vol. 295, p.1664 - 1669.
- [13] Ravasz, E., Somera, A. L., Oltvai, Z. N. & Barabasi, A. -L. 2002 'Hierarchical organization of modularity in metabolic networks', *Science* vol. 297, p.1551 - 1555.
- [14] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A.-L. 2000. 'The large-scale organization of metabolic networks', *Nature* vol. 407, p.651 - 654.
- [15] Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. 1999. 'From molecular to modular cell biology', *Nature* vol. 402, no. Supp, p.C47 - C52.
- [16] Rives, A. W. & Galitski, T. 2003. 'Modular organization of cellular networks', *Proceedings of the National Academy of Sciences of the USA* vol. 100, no. 3, p.1128 - 1133.
- [17] Thieffry, D. & Romero, D. 1999. 'The modularity of biological regulatory networks'. *Biosystems* vol. 50, p. 49 - 59.
- [18] Hallinan, J. 2004. 'Gene duplication and hierarchical modularity in intracellular interaction networks', *Biosystems* vol. 74, no. 1 - 3, p.51 - 62.
- [19] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. 1998. 'Cluster analysis and display of genome-wide expression patterns', *Proceedings of the National Academy of Sciences* vol. 95, p.14863 - 14868.
- [20] The Gene Ontology Consortium (2000). 'Gene ontology: Tool for the unification of biology:'. *Nature Genetics* Vol. 25, pp. 25 - 29.
- [21] Entrez Gene database (2004). <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene>. Downloaded 11 June 2004.
- [22] Stewart, S. A., Hahn, W. C., O'Connor, B. F., Banner, E. N., Lundberg, A. S., Modha, P., Mizuno, H., Brooks, M. W., Fleming, M., Zimonjic, D. B., Popescu, N. C. & Weinberg, R. A. 2002. 'Telomerase contributes to tumorigenesis by a telomere length-independent mechanism', *Proceedings of the National Academy of Sciences of the USA* vol. 99, no. 20, p.12606 - 12611.
- [23] Arcaro, A, Zvelebil, M. J, Wallasch, C, Ullrich, A, Waterfield, M. W & Domin, J (2000). 'Class II Phosphoinositide 3-Kinases are downstream targets of activated polypeptide growth factor receptors'. *Molecular & Cellular Biology* Vol, 20(11), pp. 3817.
- [24] Farkas, I. J., Jeong, H., Vicsek, T., Barabasi, A.-L. & Oltvai, Z. N. 2002. 'The topology of the transcription regulatory network in the yeast *S. cerevisiae*', *Physica A* vol. 318, no. 3-4, p.601 - 602.
- [25] Kauffman, S. 1995, *'At Home in the Universe'* London: Penguin..
- [26] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. 2002. 'Network motifs: Simple building blocks of complex networks', *Science* vol. 298, p.824 - 827.
- [27] Yegeer-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U. & Margalit, H. 2004. 'Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction', *Proceedings of the National Academy of Sciences of the USA* vol. 101, no. 16, p.5934-5939.
- [28] Hallinan, J. 2003©. 'Self-organization leads to hierarchical modularity in an internet community', In Palade, V., Howlett, R. J. & Jain, L. (eds.) 'Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems': *Lecture Notes in Artificial Intelligence* vol. 2773, p.914 - 917.