EM Clustering Based Approach to Decipher Functional Modules in Cross-Talking Signaling systems

Thanh-Phuong Nguyen, Adaoha E. C. Ihekwaba, and Corrado Priami

***Abstract*—It has become increasingly clear that signalling pathways are extensively interconnected and are embedded in networks with common protein components. These components do not exist in isolation but may gather together to form crosstalk modules. Constructing these crosstalk modules has emerged as a good method to understand the mechanisms underlying the propagation of transduction signals in cell. In this paper, we have presented an advancement of the method, which is chiefly used to integrate multiple topological and functional data to detect crosstalk modules between nuclear factor kappa B (NF-kB), p53 and theG1/S phase of the cell cycle systems. Applying the Expectation Maximization (EM) clustering algorithm, our results were comparable to the *k*-means algorithm. The EM algorithm as a soft clustering method is able to distinguish overlapping parts among clusters, and here we show that it is potentially more effective than the *k*-means algorithm in detecting the cross talking modules involved in the network interactions between the two systems NF-kB and p53. In addition, the biological analyses support our findings, and propose testable hypotheses to which the functional networks are involved in along with their associated human diseases.**

***Index Terms*—Cross talk, Multi-relational Clustering, Functional modules, Signaling network.**

I. Introduction

In this post-genomic era, understanding the complex networks of interacting genes, proteins, hierarchical connection across space, and small molecules that give rise to biological form and function is a major challenge since data on molecular interactions are increasing exponentially. Measurement of this flood of information can be undertaken in a high-throughput, multivariate manner using various kinds of array technologies. Since this multivariate technology, data is then relatively intractable to hypothesis generation, computational algorithms for mining the data to generate hypotheses concerning the potential interpretation of these datasets is necessary. Consequently, in order to develop new predictions for experimental test or design, computational modelling is required for similar reasons [1].

Signal transduction systems are complex biological networks that allow the cell to receive, transmit, and act upon molecular signals. Although these networks are essential for the correct functioning of the cell, they have also been reported to themselves result in abnormal cellular transformation or differentiation, often producing a pathological disease outcome [2]. NF-kB and p53 systems have attracted a widespread interest because of their central role in several human diseases [3], [4]. NF-kB exists in most cell types as a homodimer or heterodimer of a family of structurally related proteins, and is particularly important in modulating the expression of immune-regulatory genes. NF-kB has been shown to interact with other signalling networks resulting in complex non-linear responses to different combinations of stimuli [5], [6]. One of the most relevant networks interacting with NF-kB is p53, a tumour suppressor that upon DNA damage recognition mediates the activation of defense mechanisms, such as DNA repair, cell cycle arrest and eventually apoptosis [7]. The critical role of p53 in tumour suppression is underscored by the observation that over 50% of all cancers involve a disruption of this system [8]. Deciphering the interaction between NF-kB and p53 systems is appealing since it would be useful in gaining a deeper understanding on the cellular response to external stimuli. Most of the work on uncovering the interaction between these two networks has been done by traditional biological experiments [6], [9].

To improve the understanding of signalling networks, some groups have taken advantage of the published data on protein-protein interaction (PPI) networks of signalling systems (assuming that signal transduction is mainly dependent on specific protein-protein interactions [10], [11], and have tried to computationally reconstruct those networks. Some of the methods used for this purpose are the Markov chain Monte Carlo method [12], the computational algebra method [2] the cost search functions method [13], the shortest path profiles computation [14] and others. Even though, the aforementioned works achieved interesting results, they mostly considered signalling networks as independent linear cascades. Nevertheless, we know from literature that signalling networks do not exist in isolation, but are likely to interact or influence each other to perform a complex signalling task in cell [15]-[19]. For this reason, associating the NF-kB and p53 system with some cell cycle proteins, in particular, the G1/S phase cell cycle proteins [20]-[23] have also been investigated in this study. Cell cycle proteins were considered since previously published literature showed some of its proteins to be activated by one pathway and to be relevant for the regulation of another [20], [22]-[24]; and thus

Manuscript received April 17, 2011. (Write the date on which you submitted your paper for review.) This work has been partially funded by the NOBEL project.

Thanh-Phuong Nguyen is with The Microsoft Research - University of Trento Centre for Computational and Systems Biology, Piazza Manci 17, 38123, Povo (Trento), Italy. (phone: +39 0461 282811, fax: +39 0461 282814; e-mail: nguyen@ cosbi.eu).

Adaoha E. C. Ihekwaba is with Cambridge Centre for Brain Repair, University of Cambridge, Forvie Site, Robinson Way, Cambridge, CB2 0PY, UK. (e-mail: aeci3@cam.ac.uk).

Corrado Priami is with The Microsoft Research - University of Trento Centre for Computational and Systems Biology, Piazza Manci 17, 38123, Povo (Trento), Italy and University of Trento, via Belenzani, 12 I-38122 Trento, Italy. (e-mail: priami@ cosbi.eu).

may be useful in showing a level of complexity not visible by looking at the NF-kB and p53 systems alone.

Our objective is to model functional modules of network interaction between NF-kB, p53 systems, and then between these two systems and the G1/S phase cell cycle systems. We did not consider proteins as individual elements as it was done in previous work [11]-[14], but we analyzed their interplay in performing signalling crosstalk both topologically and functionally. To this end, the integration of multiple data was carried out to achieve a comprehensive view of how these systems interact with each other.

In this paper, we present a new computational method to detect functional modules cross talking between NF-kB and p53 systems that combine and mine multiple data. First, from the large-scale human protein interaction networks, our method computes all the connecting proteins (CPs) that are likely to be shared between two signalling networks. Then we proceed to extract numerous topological and functional data relating to the connections from multiple data sources – such as the Universal Protein Resource (Uniprot) [25], the Interologous Interaction Database (i2d) [26], the Reactome [27] and the InterPro database [28] – and integrated them in a multi-relational scheme. Finally, we applied the Expectation Maximization (EM) algorithm to mine and detect the functional modules between the two systems NF-kB, p53. To estimate the performance of the method we calculated its case likelihood and compared it to that obtained for the *k*-means algorithm, a hard clustering method. The better case likelihood shows that our soft clustering, EM algorithm is more suitable to discover crosstalk in the propagation of transduction signals between NF-kB and p53 networks. This work was extended to unveil the functional modules of the network interactions between NF-kB, p53 and cell cycle systems. The obtained results of this extension confirmed the better performance of the EM algorithm than the *k*-mean algorithm. In addition, we also analyzed the biological functions of the clustered networks to assess the biological relevance of the findings. The analysis gave interesting insights on the biology of NF-kB and p53 signalling networks that could be the starting point for new studies on the regulation mechanisms underlying the pathophysiology of various diseases.

The remainder of the paper is organized as follows. In Section II, we present our proposed method to detect crosstalk modules of the p53 and NF-kB networks. The results of the proposed methods are given in Section III. Some biological relevance is analysed in Section IV. Section V provides some concluding remarks.

II. Materials and Methods

In this section, we present our proposed method for the detection of the signalling crosstalk modules. Algorithm 1 describes our method that consists of three main tasks: *(i)* Computing the CPs between pathway networks, *(ii)* Manipulating and combining multiple data, and *(iii)* Constructing crosstalk modules using the EM algorithm.

Algorithm 1: Constructing functional modules from cross talking signalling systems based on multiple data.

Input:
Set of protein-protein interactions Φ.
Set of mulitple features corresponding to the extracted data $F = \{f_{dg}, f_{cc}, f_{func}, f_{pw}, f_{dm}\}$.

Output:
Set of crosstalk modules Ψ.

Step 1: Identify proteins joining the systems under investigation

Step 2: Model protein interaction networks of from the set Φ Φ.
Nf κB :=

Step 3: Find the shortest paths p between every pairs of two proteins (p_i, g_j) (p_i, g_j are proteins belonging to two different systems). $P = P \cup p$.
(P is the set of all shortest paths)

Step 4: Compute connecting proteins c_i joining path p, $\forall p \in P$. $\Omega := \{c_i\}$

Step 5: For each connecting protein c_i ∈

Step 5.1 : Calculate two topological features, degree f_{dg} and cluster coefficient f_{cc} from the constructed networks

Step 5.2: Extract functional data for three features, abstract function f_{func} from the Uniprot database, biological process f_{pw} from the Reactome database, and protein domain f_{dm} from the InterPro database

Step 5.3: Combine both topological data and functional data by a multi-relational scheme.

Step 6: Run EM algorithm to construct functional modules from cross talking systems s_i based on combined

data. $\Psi := \{s_i\}$.

Step 7: return Ψ.

In Algorithm 1, Step 1 is to identify proteins joining the two systems NF-kB and p53; then in Step 2, the PPI network of these proteins is extracted in Step 2. Steps 3 and 4 are to compute connecting proteins between pathways. Steps 1 to Step 4 are explained in greater detail in SectionII-A, and were partially reported in our previous paper [29]. In Step 5, we carry out the data extraction and the data combination by using a multi-relational scheme. The reference databases are showed in TABLE I. Step 6 is for constructing the crosstalk modules between the NF-kB and p53 by applying the EM clustering algorithm. The output is the set of the crosstalk modules.

A. Computing Connecting Proteins Between Networks

The complete protein network of NF-kB and p53 systems were modeled based on binary interactions. First, we identified the proteins joining the two systems. There are five proteins belonging to the NF-kB family known in mammalian cells: RelA (also known as p65), c-Bel, RelB, NF- B1 (p50/p105), and NF-kB2 (p52/p100). Additionally, NF- B exists in the cytoplasm in an inactive form associated with inhibitory proteins termed I- B, of which the most important ones are I-kB , I-kB , and I-kB . In the p53 network, p53 protein binds to the regulatory region of the MDM2 gene and promotes its transcription [30]. TABLE II lists protein members of the systems considered in the study (highlighted proteins are reported to be activated in one system and involved in the regulation of another).

We investigated the binary protein interaction network extracted from the i2d database as an undirected graph $G(E,V)$ consisting of a set of nodes (proteins) V and a set of edges (interactions) E between them. An edge e_{ij} connects vertex v_i

with vertex v_j. Proteins that act as intermediates in the signal transduction can be uncovered by searching the shortest paths between networks.

TABLE I. REFERENCE DATABASES USED FOR DATA RETRIEVAL DURING THE INVESTIGATION.

Database	*Description*	*URL*	*Statistics*	*Data extracted*
Uniprot [25]	comprehensive, high-quality and freely accessible resource of protein sequence and functional information.	http://www.uniprot.org	220,325 entries	function,post-translation modification, location, developmental stage, etc.
I2d [26]	on-line database of known and predicted mammalian and eukaryotic protein-protein interactions	http://ophid.utoronto.ca/	424,066 entries (92,561 for human)	protein interaction
Reactome [27]	curated resource of core pathways and reactions in human biology.	http://www.reactome.org	928 pathways for human	pathway
Interpro [28]	an integrated database of predictive protein "signatures" used for the classification and automatic annotation of proteins and genome	http://www.ebi.ac.uk/interpro/		protein domain

TABLE II. LIST OF PROTEINS AND NETWORKS CONSIDERED (THE PROTEINS HAVE BEEN LISTED ACCORDING TO THEIR UNIPROT NAMES).

Network	Uniprot accession	Uniprot entry name	Alternative name
p53 pathway	P04637	P53_HUMAN	p53
	Q00987	MDM2_HUMAN	mdm2
	P38936	CDN1A_HUMAN	p21
	Q8N726	CD2A2_HUMAN	p14ARF
NF-κB pathway	O00221	IKBE_HUMAN	NF-κB inhibitor epsilon
	O14920	IKKB_HUMAN	IKK2
	O15111	IKKA_HUMAN	IKK1
	P19838	NFKB1_HUMAN	Nuclear factor NF-κB p105 subunit
	P25963	IKBA_HUMAN	IκB-alpha
	Q00653	NFKB2_HUMAN	Nuclear factor NF-κB p100 subunit
	Q01201	RELB_HUMAN	Transcription factor RelB
	Q04206	TF65_HUMAN	Transcription factor p65 (RelA)
	Q04864	REL_HUMAN	C-Rel protein
	Q14164	IKKE_HUMAN	Inhibitor of nuclear factor κB kinase subunit epsilon
	Q15653	IKBB_HUMAN	NF-kappa-B inhibitor beta
	Q96HD1	CREL1_HUMAN	Crel1
	Q6UXH1	CREL2_HUMAN	Crel2
	Q9Y6K9	NEMO_HUMAN	IKKγ
G1/S phase cell cycle proteins	P24385	CCND1_HUMAN	Cyclin D1
	Q01094	E2F1_HUMAN	E2F -1
	P06400	RB_HUMAN	Rb
	P46527	CDN1B_HUMAN	P27

If the PPI networks here constitute an unweighted graph, the weight function *f* can be considered as a path length *l* (the number of edges in path *p*). In this case, the shortest path problem consists in finding a path *p* having the minimal path length. A Breadth-First Search algorithm [31] has been employed to find the shortest paths between two nodes. The shortest paths may have different path lengths ($l = 1$, $l = 2$, $l = 3$, $l = 4$, etc.). The nodes which belong to the shortest paths, except the starting nodes and the ending nodes, are called connecting proteins (CPs). Playing the roles of intermediate molecules, the CPs themselves do not act independently, but functionally group into signalling crosstalk modules used in connecting the two systems.

B. Manipulating and Combining Multiple Data

The crosstalk modules in the NF-kB and p53 systems were identified by both topological and functional properties. As a result, in order to construct the modules we investigated two types of data: *(1)* topological data representing the relationship between a CP and its network neighbours, and *(2)* functional data representing biological information of a CP.

The topological data was used in form of two key measures based on the protein interaction network obtained from the i2d database. The first one is the degree of a CP, which is the number of its neighbours. Node degree is one of the principal measures used to study the topology of a network. The neighbourhood *N* for a node v_i is defined as its immediately connected neighbours as $N = \{v_i\} : e_{ij} \in E$ [32]. The second one is the average clustering coefficient $C(k)$, which characterises the overall tendency of nodes to form clusters or groups; and $C(k)$ the average clustering coefficient of all nodes with k links is an important measure of the network structure [32].

The three different categories of functional data that were incorporated include: abstract function, biological process and protein domain. The abstract functions are the protein keywords documented in the Uniprot database including protein function, subcellular localization, structure, relevant mutations, and others. Fig. 1 depicts the occurrence of protein keywords of investigated CPs. Since the proteins in the crosstalk modules are likely to participate in the same cellular processes, data on the biological process from the

Reactome database also was combined. The protein domain data obtained from the InterPro database gives information on the different domains present in the CP. Protein domains are defined as structural or functional elements within a protein and affect the way that the proteins interact with each other and compose the crosstalk modules. Since different databases have different names for each entry, the Uniprot accession number for identifying a protein was used as our standard and thus all CPs' names were converted and mapped accordingly to their Uniprot accession numbers.

To store and integrate the acquired data, a multi-relational scheme was appropriately structured in form of tables (with columns and rows) and relationships between tables in the SQL Server Database Management. The data types are heterogeneous, since the abstract function data, the process data and the domain data are in form of a categorical free text, while the node degree and the cluster coefficient are numerical values. For example, protein IKBA_HUMAN (NF-kappa-B inhibitor alpha) with node degree equal to 60 and cluster coefficient equal to 0.05826, contains many keywords, such as Phosphoprotein, 3D-structure, ANK repeat, Cytoplasm, Disease mutation, e.t.c.; takes part in two processes, such as signalling by NGF, signalling in Immune system; and has two domains ANK and NF-kB inhibitor. In this regard, the heterogeneity in the data makes the use of traditional clustering methods unsuitable in clustering multiple relational data and underlines the importance of using a multi-relational clustering approach in the detection of CPs propagating transduction signals across systems being investigated based on their relational data.

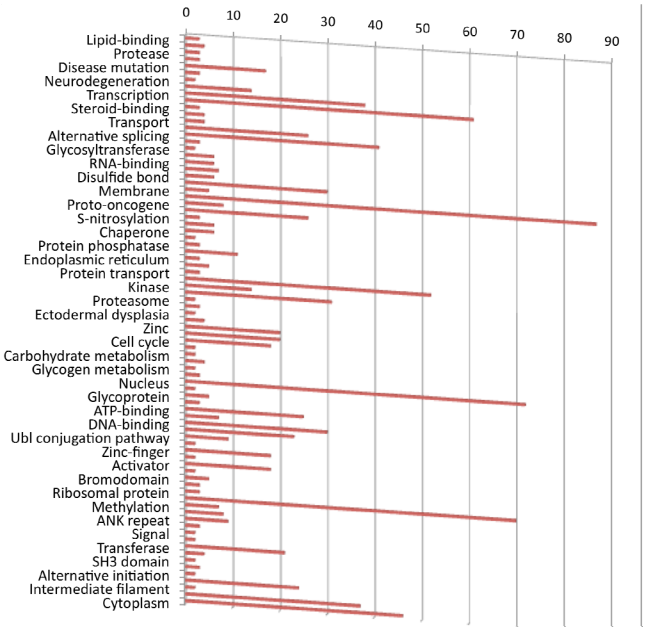

Fig. 1 Occurrence of protein keywords extracted for connecting proteins. The horizontal axis is the list of protein keywords and the vertical axis is the occurrence of keyword.

The network was computed using the COSBILab Graph software [33]. The data extraction was implemented in the Python programming language[1] and was derived from the library BioPython[2].

C. Detecting Crosstalk Using Multi-Relational Clustering

Multi-relational data mining (MRDM) involves pattern search in multiple data tables with multiple relations from a relational database, unlike most existing data mining approaches that restrict search in a single data table [34]. With the rapid growth of public biological databases, MRDM can be widely applied to discover complex patterns through the rich relational structure and the mixed-up types of data. Multi-relational clustering is the process of partitioning data objects into a set of clusters based on their similarity, utilising the information in multiple relations [35]. Due to the notable characteristics of multi-relational clustering, we employed this technique for inferring crosstalk modules.

The crosstalk modules do not separate but share some common CPs and their interactions. For this reason, the EM algorithm was used to perform the clustering task. The EM algorithm is a soft clustering method, this means that a data point always belongs to multiple clusters, and that a probability is calculated for each combination of a data point and a cluster. We applied the EM algorithm in the SQL Server 2008 Analysis Services (SSAS) since it allows the user to explore data in multiple relational tables[3]. In the EM algorithm, an initial cluster model is iteratively refined to fit the data and the probability that a data point exists in a cluster is produced. The fitness function is the *log̃likelihood* of the data given the model. The process ends when the probabilistic model fits the data.

III. Results

A. Functional modules from cross talking NF-kB and p53 systems

After computing the shortest paths between the p53 and NF-kB networks, we found 112 CPs joining at least one of these paths. The multiple data of the 112 CPs were extracted from four databases (the i2d database, the Uniprot database, the Reactome database, and the InterPro database) and integrated by the multi-relational scheme and manipulated in a corresponding relational database. This database consists of 2,086 data records of the abstract function feature, of which 265 represents the biological process feature, and 761 for the protein domain feature.

Based on the combined data, the clusters of CPs were produced by running the EM algorithm. The number of cluster were set to 5 and other parameters were set to default values. Fig. 2 shows the graphical view of five produced clusters. Each cluster is presumably a crosstalk module of CPs between the NF-kB and p53. Clusters 1 and 3 have the largest population with 28 CPs, followed by cluster 2 with 21 CPs, cluster 4 with 19 CPs and finally cluster 5 with 16 CPs. The thickness of the line connecting the clusters indicate the degree of similarity of the links.

It is apparent from 0, that cluster 3 is strongly related to clusters 1, 2 and 5 and loosely to cluster 4. There is no

[1]http://www.python.org

[2]http://biopython.org

[3]http://technet.microsoft.com/en-us/library/ms174879.aspx

apparent relation between cluster 4 and clusters 2 and 5.

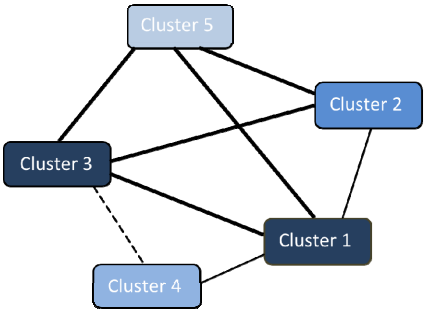

Fig 2. The graphical view of five clusters for the combined network of NF-kB and p53. Each cluster is visualized by a rectangle/a node and the edge represents the association between two clusters

To evaluate the performance of the EM algorithm, we also performed the *k*-means clustering of the same data set and compared the goodness of the two methods. The k-means clustering is a popular hard clustering method predominantly used to assign cluster membership. It works by minimizing the differences among items in a cluster while maximizing the distance between clusters.

The goodness of the two methods was calculated in term of case likelihood. Case likelihood is defined as the sum of cluster likelihood scores for each case (a case here is simply defined as a record of a CP in the Microsoft SQL Server database), divided by the number of cases in the partition.

The case likelihood indicates how likely it is that a case belongs to a particular cluster. The case likelihood values range between 0 to 1 with a higher value indicating a higher probability of occurrence of a case in the model. We performed a 3-fold cross validation for both the algorithms. In 3-fold cross-validation, the original set data set is randomly partitioned into three subsets. Of the three subsets, a single subset is used as the validation data for testing the model, and the remaining three subsets are used as training data. The cross-validation process is then repeated three times (the folds), with each of the three subsets used exactly once as the validation data. The final estimation is the average of three results from the folds [36]. Table III shows the results of the 3-fold cross validation done for the EM algorithm and the k-means algorithm respectively.

The EM algorithm achieved a considerable case likelihood of 81.7%, meaning that 81.7% of cases were most likely clustered appropriately. This was relatively higher than the case likelihood of 74.6% obtained for the *k*-means algorithm. The EM algorithm, being a soft clustering method, was able to distinguish overlapping parts among clusters, and here we show that it is potentially more effective than the *k*-means algorithm in detecting the cross talking modules involved in the network interactions of the two systems NF-κB and p53.

B. Functional modules from cross talking p53, NF-κB and the G1/S phase of the cell cycle systems

Since it has been suggested, that some cell cycle proteins are activated by one pathway and are relevant for the regulation of another [36], [36]-[37], it was of interest to investigate the interaction between the NF-κB, p53 and the cell cycle systems. For this study, only events leading to the G1/S transition phase of the cell cycle, the point where NF-kB and p53 signal transduction events are active the most [38] were considered.

TABLE III. CASE LIKELIHOOD OF THE EM ALGORITHM AND THE K-MEANS ALGORITHM BY 3-FOLD CROSS VALIDATION. THE FIRST COLUMN IS THE ID OF THE PARTITION AND THEN THE NUMBER OF PROTEIN IN THE PARTITION IS IN THE PARTITION SIZE. THE THIRD COLUMN IS THE CLUSTERING LIKELIHOOD CORRESPONDING TO EACH THE PARTITION. THE FINAL CLUSTERING LIKELIHOOD OF THE METHOD IS THE AVERAGE OF THREE ABOVE LIKELIHOODS.

EM algorithm			***k*-means algorithm**		
#Partition	*Partition Size*	*Clustering likelihood*	*#Partition*	*Partition Size*	*Clustering likelihood*
1	38	0.720	1	38	0.724
2	37	0.971	2	37	0.808
3	37	0.760	3	37	0.707
	Average	**0.817**		Average	0.746
	Standard Deviation	0.110		Standard Deviation	0.044

We repeated the analysis (shown in above section) to include interactions between the rest of the G1/S cell cycle proteins (RB_HUMAN, CCND1_HUMAN, CDN1B_HUMAN, and E2F1_HUMAN) and the members of the p53 and NF-κB networks. We obtained 106 CPs between these three systems. Please note that this set of CPs is different from the set of CPs computed for the combined networks of p53 and NF-κB.

The number of cluster was set to five, and the rest of the parameters were set to default values. Fig. 3 shows the graphical view of the five produced clusters corresponding to five functional modules. Cluster 1 had the largest population with 39 CPs, followed by cluster 2 with 34 CPs. Cluster 3 and cluster 4 had the same number of CPs, which is 14. The smallest cluster was cluster 5 with 5 CPs. Cluster 2 was identified to associate with all the other clusters, and even loosely with cluster 4. Yet, cluster 4 seemingly separates from the other clusters. The diagram shows the complicated relationship among clusters that might not be unveiled by using traditional clustering methods.

We also did the procedure of 3-fold cross validation for evaluating the goodness. Table IV shows the results of 3-fold cross validation obtained by the EM algorithm and the *k*-means algorithm respectively. The EM algorithm achieved a higher case likelihood (81.3%) than k-means algorithm did (75.6%). In addition to the result of study on the NF-kB and p53 systems, this result confirmed that our proposed methods could discover the functional modules more effectively than hard clustering method.

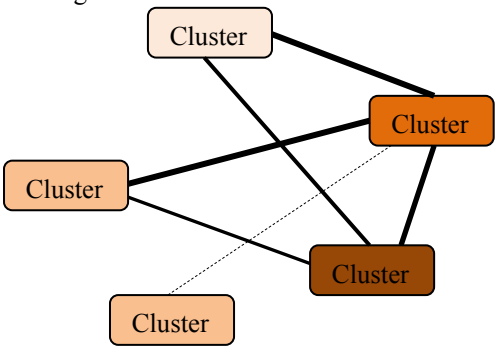

Fig 3. The graphical view of five clusters for the combined network of p53, NF- κ B and the G1/S phase of the cell cycle system.

TABLE IV. CASE LIKELIHOOD OF THE EM ALGORITHM AND THE K-MEANS ALGORITHM BY 3-FOLD CROSS VALIDATION. THE FIRST COLUMN

IS THE ID OF THE PARTITION AND THEN THE NUMBER OF PROTEIN IN THE PARTITION IS IN THE PARTITION SIZE. THE THIRD COLUMN IS THE CLUSTERING LIKELIHOOD CORRESPONDING TO EACH THE PARTITION. THE FINAL CLUSTERING LIKELIHOOD OF THE METHOD IS THE AVERAGE OF THREE ABOVE LIKELIHOODS.

EM algorithm			***k-means algorithm***		
#Partition	*Partition Size*	*Clustering likelihood*	*#Partition*	*Partition Size*	*Clustering likelihood*
1	36	0.798	1	36	0.680
2	35	0.874	2	35	0.855
3	35	0.766	3	35	0.732
	Average	**0.813**		Average	0.756
	Standard Deviation	0.055		Standard Deviation	0.0899

IV. DISCUSSION

The validation of the method was performed by analysing the biological relevance of clustered networks. We examined the characteristics of each network described by the probability of 'features data' of CPs in the network.

Studying the biological processes involved in network 3, we found a high number of CPs for three processes 'Signalling in Immune system' (15.58%), 'Gene Expression' (12.27%), and 'Cell Cycle Checkpoints' (10.27%). On further investigation of network 2, we found three main processes: 'Signalling by NGF' (20.41%), 'Cell Cycle, Mitotic' (16.28%), and 'Signalling in Immune system' (15.58%). These results on the functional modules of CPs suggest that the two systems (NF-κB and p53) to essentially effect the signalling and cell cycle processes of the cell, a finding which is in agreement with known characteristics of the transcription factors [3], [6] – that is to say our method was able to correctly mine data that correctly represents the investigated factors.

We then reconstruct the five produced modules. We extracted the interactions between two CPs inside the clustered network; that is to say, other interactions with proteins outside the cluster were excluded (in Fig. 4). Network 3 (with 28 nodes and 92 interactions) was the most complex one with high-connected nodes. More than half of proteins in network 3 have a degree value more than 5, three proteins with the highest degree were P04637- Cellular tumor antigen p53 (degree = 21), Q09472 - Histone acetyltransferase p300 (degree = 15) and Q92793 - CREB-binding protein (degree =14). Most of the proteins share several common GO terms, such as signal transduction, interspecies interaction between organisms, cell-cycle, many others. In network 2, more than 87% of the proteins are annotated as transcription proteins and involved in the regulation of the transcription process.

Furthermore, we studied the association between the detected networks and diseases. For network 3, 32.05% of the proteins that are annotated to mediate interactions between a host-virus and to influence resistance and recovery from viral infections. About 31.70% of CPs in this network were associated with diseases causing mutation. For example, CREB-binding protein in network 3 had histone acetylase activity. andchromosomal aberrations involving the gene coding for CREBBP, were shown to be linked to acute myeloid leukemias. Additionally, histone acetyltransferase p300 as the name implies has histone acetyltransferase activity through which it regulates transcription via chromatin remodelling. Defects in EP300 may be associated to epithelial cancer. Chromosomal aberrations involving the gene coding for EP300 may be a cause of acute myeloid leukemias. This brief observation offers the chance to look into the relationship between NF-kB and p53 pathways and reinstates their role in the pathophysiology of various diseases as a consequence of their dis-regulation.

Repeating the same evaluation for the modules between p53, NF-kB and the G1/S phase of the cell cycle systems, the five networks were reconstructed. Fig. 5 visualizes these networks. We found that the biggest network is network 1 due its large number of CPs, in contrast to network 5 with 5 proteins and 3 interactions. Network 4 appears much more dense than network 3, even though they have the same number of nodes. Analyzing the functionality of network 1, there are 35.90% of proteins involved in the regulation of the transcription process. We also found a high number of CPs in network 1 involving in three pathways 'Signalling in immune system' (17.95%), 'Nerve growth factor receptor signaling pathway' (15.39%), and 'Insulin receptor signaling pathway' (12.82%). Attracted by the high-connected topology of network 4, there is a dominant number of proteins performing their functions in transcription regulation (81.82%); and 45.46% of proteins are activators. In network 4, 36.36% of proteins are showed to having at least one variant, responsible for a disease and there is the same number of proteins that are considered involved in direct interaction between the host cell macromolecular machinery and viral proteins. These above analysis show that these functional modules probably play a crucial role in causing diseases.

An important advantage of computational analysis is that it allows the detection of such modules by integrating complex and heterogeneous data. However, it is worthy of note that using computational techniques alone we could not verify the reactions underlying the crosstalk, however, in this report, we provide a theoretical foundations for targeted experimental studies toward potential functional modules. Our method and results suggest some testable biological hypotheses to focus future studies on, and reveal an essential functionality of cross talking networks and their biological relevance.

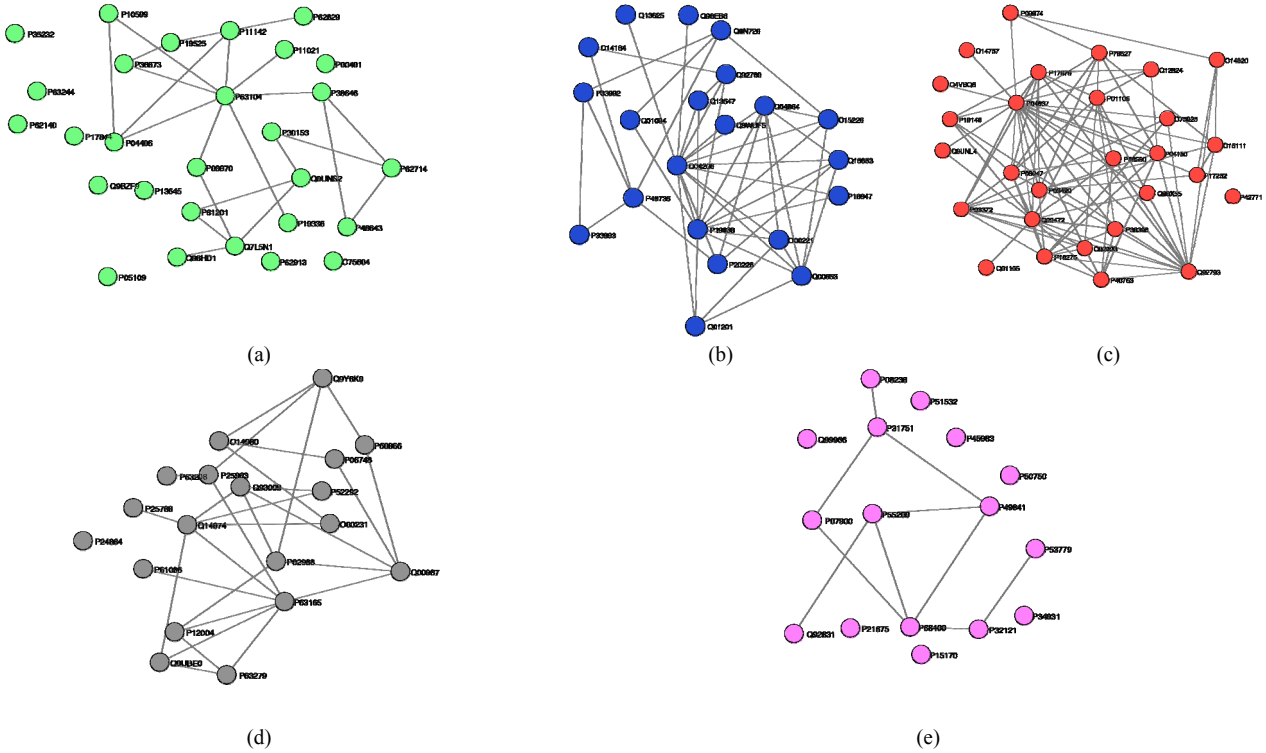

Fig 4. The protein network constructed for the networks p53, NFkB: cluster 1 (a), cluster 2 (b), cluster 3 (c), cluster 4 (d) and cluster 5 (e). In the network, the node is the connecting protein of the corresponding cluster with their protein ID and the edge is the interaction between two proteins.

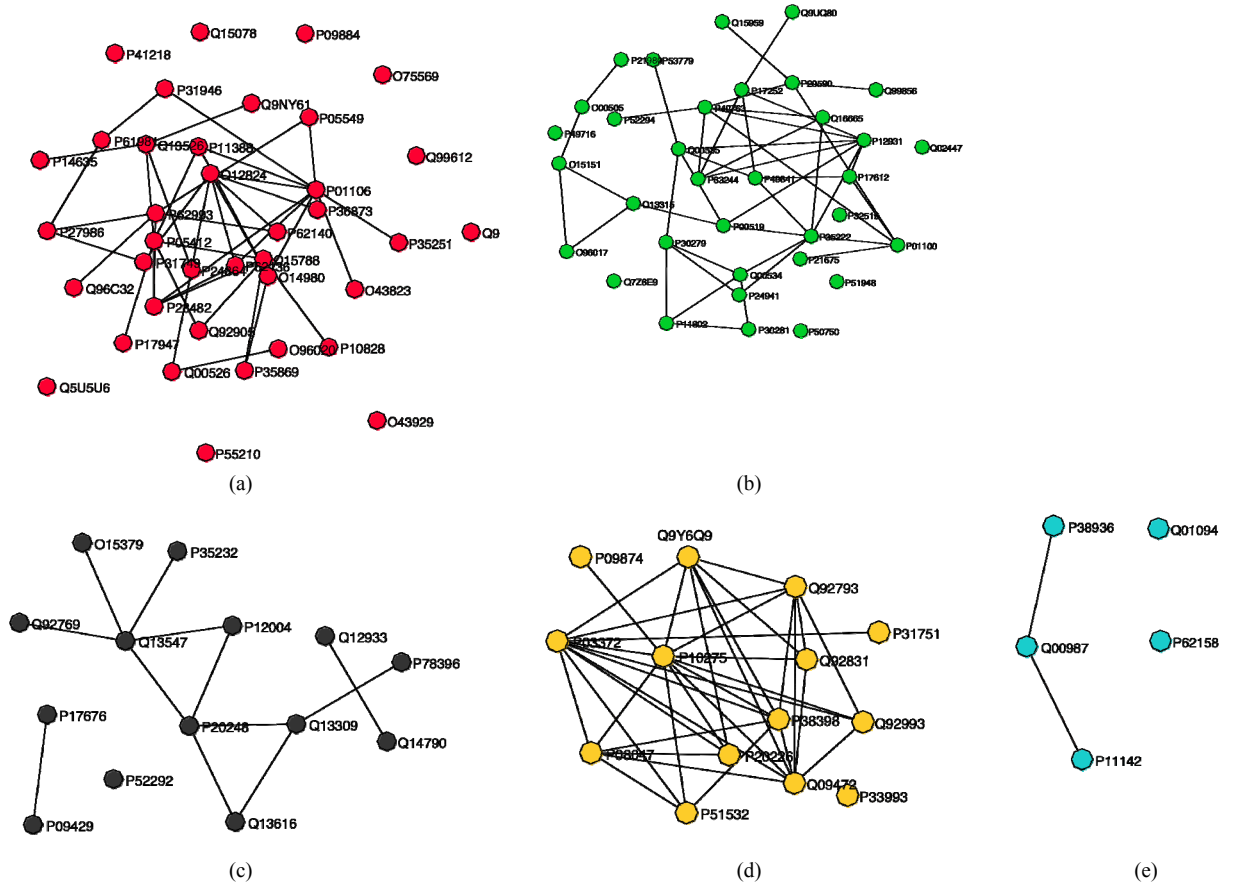

Fig 5. The protein network constructed for the networks between cell cycle systems and p53, NF-kB cluster 1 (a), cluster 2 (b), cluster 3 (c), cluster 4 (d) and cluster 5 (e). In the network, the node is the connecting protein of the corresponding cluster with their protein ID and the edge is the interaction between two proteins.

V. Conclusion

We have proposed a computational method to detect signalling crosstalk modules that potentially function in transducing signals between the NF-κB and p53 pathways, and the G1/S phase of the cell's cycle. The advantage of this method is its ability to efficiently combine various relational data extracted from different data sources. The experimental results demonstrated that our proposed method performs well, especially when the EM algorithm is used. The biological analyses also showed the plausibility of these findings, and put forward several testable hypotheses to which the functional networks are involved in along with their associated human diseases. For future work, we would like to integrate other relational data into the scheme and study further the relationship between the functional modules, diseases and therapeutic targets. The larger experiments with more pathways are prospective to reveal the completely complex mechanisms in systems biology.

Acknowledgment

This research has been partially funded by the NOBEL project. The authors wish to thank Danish Memon (Department of Chemical Engineering, Indian Institute of Technology Bombay, India) for valuable discussions and editorial suggestions. We are grateful to two Anonymous Referees for constructive critics improving the quality of the paper.

References

[1] C. Priami, "Algorithmic systems biology," *Commun. ACM*, vol. 52, pp. 80–88, May 2009.

[2] E. Allen, J. Fetrow, L. W. Daniel, S. Thomas, and D. John, "Algebraic dependency models of protein signal transduction networks from time-series data," *J. of Theoretical Biology*, vol. 238, no. 2, pp. 317–330, 2006.

[3] J. Albert S. Baldwin, "The transcription factor NF-kappaB and human disease," *J Clin Inves*, vol. 101, no. 1, pp. 3–6, 2001.

[4] Y. Yang and A. M. Weissman, "Regulating the p53 system through ubiquitination," *Oncogene*, vol. 23, no. 11, pp. 3485–2106, 2004.

[5] M. Natarajan, K. Lin, R. Hsueh, P. Sternweis, and R. Ranganathan, "global analysis of cross-talk in a mammalian cellular signalling network," *Nat Cell Biol*, vol. 8, no. 6, pp. 571–580, 2006.

[6] G. A. Webster and N. D. Perkins, "Transcriptional Cross Talk between NF-kappa B and p53," *Mol. Cell. Biol.*, vol. 19, no. 5, pp. 3485–3495, 1999.

[7] Gerard Evan and Trevor Littlewood, "A Matter of Life and Cell Death", *Science*, 281 (5381), 1317-1322, 1998.

[8] K. H. Vousden and D. P. Lane, "p53 in health and disease," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 4, pp. 275–283.

[9] K. Schumm, S. Rocha, J. Caamano, and N. Perkins, "Regulation of p53 tumour suppressor target gene expression by the p52 NF-kappaB subunit," *The EMBO Journal*, vol. 25, p. 4820–4832, 2006.

[10] N. J. Eungdamrong and R. Iyenga, "Modeling cell signaling networks," *Biology of the Cell*, vol. 96, no. 5, pp. 355–362, 2004.

[11] Y. Liu and H. Zhao, "A computational approach for ordering signal transduction pathway components from genomics and proteomics data." *BMC Bioinformatics*, vol. 5, no. 158, 2004.

[12] S. M. Gomez, S. Lo, and A. Rzhetsky, "Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks," *Genetics*, vol. 159, no. 3, pp. 1291–1298, 2001.

[13] X. Zhao, R. Wang, L. Chen, and K. Aihara, "Automatic modeling of signal pathways from protein-protein interaction networks," in *The Sixth Asia Pacific Bioinformatics Conference*, 2008, pp. 287–296.

[14] Y. Li, P. Agarwal, and D. Rajagopalan, "A global pathway crosstalk network," *Bioinformatics*, vol. 24, no. 12, pp. 1442–1447, 2008.

[15] Gagneur J, Casari G: From molecular networks to qualitative cell behavior. *FEBS Lett*, 579(8):1867-1871, 2005.

[16] Gagneur J, Krause R, Bouwmeester T, Casari G: Modular decomposition of protein-protein interaction networks. *Genome Biol* 2004, 5(8):R57.

[17] Bhalla US: Understanding complex signaling networks through models and metaphors. *Prog Biophys Mol Biol*, 81(1):45-65, 2003.

[18] Kell DB: Metabolomics, machine learning and modelling: towards an understanding of the language of cells. *Biochem Soc Trans*, 33(Pt 3):520-524, 2005.

[19] Yaffe MB: Signaling networks and mathematics. *Sci Signal*, 1(43):eg7, 2008.

[20] Webster GA, Perkins ND: Transcriptional cross talk between NF-kappaB and p53. Mol Cell Biol 1999, 19(5):3485-3495.

[21] Dotto GP: p21(WAF1/Cip1): more than a break to the cell cycle? *Biochim Biophys Acta*, 1471(1):M43-56, 2000.

[22] Sheahan S, Bellamy CO, Treanor L, Harrison DJ, Prost S: Additive effect of p53, p21 and Rb deletion in triple knockout primary hepatocytes. *Oncogene*, 23(8):1489-1497, 2004.

[23] Kamijo T, Weber JD, Zambetti G, Zindy F, Roussel MF, Sherr CJ: Functional and physical interactions of the ARF tumor suppressor with p53 and Mdm2. *Proc Natl Acad Sci USA*, 95(14):8292-8297, 1998.

[24] Pomerantz J, Schreiber-Agus N, Liegeois NJ, Silverman A, Alland L, Chin L, Potes J, Chen K, Orlow I, Lee HW et al: The Ink4a tumor suppressor gene product, p19Arf, interacts with MDM2 and neutralizes MDM2's inhibition of p53. *Cell*, 92(6):713-723, 1998.

[25] T. U. Consortium, "The Universal Protein Resource (UniProt) in 2010," *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D142–D148, 2010.

[26] K. Brown and I. Jurisica, "Unequal evolutionary conservation of human protein interactions in interologous networks," *Genome Biology*, vol. 8, no. 5, pp. R95+, May 2007.

[27] Croft D. et al., "Reactome: a database of reactions, pathways and biological processes." *Nucleic acids research*, November 2010.

[28] S. Hunter et al., "InterPro: the integrative protein signature database," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D211–D215, 2009.

[29] A. E. Ihekwaba, P. T. Nguyen, and C. Priami, "Elucidation of functional consequences of signalling pathway interactions," *BMC Bioinformatics*, vol. 10, no. 370.

[30] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network." *Nature*, vol. 408, no. 6810, pp. 307–310, November 2000.

[31] C. Cormen, T.H.and Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed.. Cambridge, MA The MIT Press, 2001.

[32] J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[33] R. Valentini and F. Jordan, "CoSBiLab Graph The network analysis module of CoSBiLab," *Environmental Modelling and Software*, vol. 25, no. 7, pp. 886 – 888, 2010.

[34] S. Dzeroski and N. Lavrac, Eds., *Relational Data Mining*. Springer, 2001.

[35] Han J, Kamber M: *Data Mining: Concepts and Techniques* (The Morgan Kaufmann Series in Data Management Systems). San Francisco: Morgan Kaufmann, 2000.

[36] GJ, McLachlan, K.A. Do, C. Ambroise. *Analyzing microarray gene expression data*. Wiley, 2004.

[37] Stott FJ, Bates S, James MC, McConnell BB, Starborg M, Brookes S, Palmero I, Ryan K, Hara E, Vousden KH et al: The alternative product from the human CDKN2A locus, p14(ARF), participates in a regulatory feedback loop with p53 and MDM2. *Embo J*, 17(17):5001-5014, 1998.

[38] Kaltschmidt B, Kaltschmidt C, Hehner SP, Droge W, Schmitz ML: Repression of NF-kappaB impairs HeLa cell proliferation by functional interference with cell cycle checkpoint regulators. *Oncogene*, 18(21):3213-3225, 1999.

Thanh-Phuong Nguyen was born in Hanoi, Vietnam. Thanh-Phuong Nguyen received her B.S. (2003) and M.S. (2005) in Information Technology from Hanoi University of Technology (HUT), Vietnam. In September 2008, she received a PhD in Bioinformatics from the School of Knowledge Science at the Japan Institute of Science and Technology (JAIST). From 2003 to 2010, she was a lecturer at the Faculty of Information Technology, Hanoi University of Technology. Her current research interests are formal model, data mining and machine learning applied to biological and medical data in general, and molecular networks and molecular networks

related to diseases in particular. Since October 2008, she has worked at The Microsoft Research - University of Trento Centre for Computational and Systems Biology, Trento, Italy.

Adaoha Elizabeth C. Ihekwaba received a PhD in 2006 from The University of Manchester, UK. She was awarded a Wain International post-doctoral Fellowship from the BBSRC (2006) to establish a methodology for developing computational models of cellular biochemistry based on genome-wide molecular profiling data at Virginia Bioinformatics Institute. In 2007, she joined The Microsoft Research – University of Trento Centre for Computational and Systems Biology (CoSBi) as a researcher, to explore the effectiveness of numerical simulation techniques to better understand the implications of complex and kinetics of cellular signal transduction events – with the intention of using existing data and state-of the art technologies to generate data sets that would be useful in building and testing computer models. She is currently being supported by John Van Geest research fellowship (2010) to focus on the systematic analysis of the movement of non-coding RNAs (ncRNAs) in secreted micro vesicles (MVs) shuttled between neural / precursor stem cells (NPCs) and their neighbouring recipient cells.

Corrado Priami obtained his Laurea and PhD degrees in Computer Science at the University of Pisa, visited as associate researcher at the laboratory LIX, École Politechnique, Paris (1995) and the École Normale Supérieure, Paris under an EC Marie Curie TMR grant (1996). He was a researcher and associate professor at the University of Verona (1997-2001). Currently, he is a professor of Computer Science at the University of Trento. The results of his PhD thesis on stochastic pi- calculus were the basis for the foundation of the Microsoft Research - University of Trento Centre for Computational and Systems Biology (COSBI), of which he is the President and CEO. Those same results, besides constituting the scientific foundation of COSBI, are recognized as fundamental in the field of systems biology by an expanding international community, which is using them to model the behavior of biological systems (the CMSB conference is a milestone of this). He was member of the expert group on the EU 7th FP of the CRUI and has participated in many projects promoted by the European Commission for the advancement of emerging areas of research. He regularly serves on the evaluation committees for projects presented by the European Commission, is an anonymous reviewer for many international journals, and serves in the review panels of the Science Foundation Ireland for institutes of systems biology and of the Netherlands Organisation for Scientific Research. His research covers computational methods for the modelling, analysis, and simulation of biological systems, programming languages, and formal computational theories. He published over 130 papers in international journals and conferences, given more than 40 invited talks and lectures at conferences and universities around the world, participated in the program committees for 21 international conferences (ten of which he was chair), is a member of three steering committees of international conferences (of which one he is president). He is editor-in-chief of the international journal Transactions on Computational Systems Biology and member of the editorial board of the International journal Bioinformatics Research and Applications. He founded the international conferences "Computational Methods in Systems Biology (CMSB)," which is continuing to grow, and "Converging Sciences," whose success has been described by many international journals. He was a member of ISTAG-FET (Information Society Technologies Advisory Group - Future and Emerging Technologies) of the European Commission.