Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 726–749
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

An algorithm for modularization of MAPK and calcium signaling pathways: Comparative analysis among different species

Losiana Nayak, Rajat K. De *

Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India

Received 28 September 2006
Available online 18 May 2007

Abstract

Signaling pathways are large complex biochemical networks. It is difficult to analyze the underlying mechanism of such networks as a whole. In the present article, we have proposed an algorithm for modularization of signal transduction pathways. Unlike studying a signaling pathway as a whole, this enables one to study the individual modules (less complex smaller units) easily and hence to study the entire pathway better. A comparative study of modules belonging to different species (for the same signaling pathway) has been made, which gives an overall idea about development of the signaling pathways over the taken set of species of calcium and MAPK signaling pathways. The superior performance, in terms of biological significance, of the proposed algorithm over an existing community finding algorithm of Newman [Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA 2006;103(23):8577–82] has been demonstrated using the aforesaid pathways of *H. sapiens*.

Keywords: Signal transduction; Systems biology; Biological networks; Graphs; Modules; Community finding algorithm

1. Introduction

Signaling pathways are complex biochemical networks that regulate numerous cellular functions. They are nonlinear, exist as complex webs, and function by serial and successive interactions among large number of vital biomolecules and chemical compounds. Biomolecules are large in size and volume in comparison to the tiny pores present in biological membranes like cell membrane, nuclear membrane, mitochondrial membrane etc. So they cannot possibly travel through these barriers and convey the message to initiate a counter action in all possible circumstances. The information must pass via some other means. Hence the message is signaled from one biomolecule to another in a cascade till it reaches its destination. These cascades are known as signaling pathways.

MAPK signaling pathway is one of the most ubiquitous signal transduction systems [2]. It is characterized by the general path, "*Stimulus* > *MAPKKK* > *MAPKK* > *MAPK* > *Response*", where MAPKK is the kinase of MAPK and MAPKKK is the kinase of MAPKK. The symbol "*A* > *B*" stands for "*A* stimulating *B*". In most of the cases, MAPKKK is activated by small G proteins such as Ras and Rap1 [3,4]. MAPK signaling pathway is conserved in all eucaryotes and plays a key role in regulation of gene expression as well as cytoplasmic activities. They transduce a large variety of external signals; leading to a wide range of cellular responses, including mating, filamentation, high osmolarity responses, cell wall remodelling, sporulation (*S. cerevisiae*), cell growth, differentiation, stress response, T-cell development, inflammation and apoptosis (mammals), morphogenesis, spatial patterning (*D. amoebae*), eye development (*D. melanogaster*), vulva induction (*C. elegans*) [3]. In general MAPK signaling pathways comprise of the conventional MAPK pathway, JNK pathway, p38 pathway and ERK pathway.

* Corresponding author. Fax: +91 33 2578 3357.
E-mail addresses: losiana_t@isical.ac.in (L. Nayak), rajat@isical.ac.in (R.K. De).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.05.007

Calcium signaling pathways are very peculiar in nature. Normal intracellular Ca^{2+} level (10^{-7} M) is much lower from the extracellular concentration of 10^{-3} M. Calcium ions precipitate phosphate of the established energy currency of cells. Also high concentration of intracellular calcium ions lead to cell death. This is the reason why calcium ion concentration must be maintained at low levels in cytoplasm. Hence cells have evolved techniques for free calcium ion binding to reduce its effect towards cytosol, which later is used as well for signal transduction across and inside the cell [5]. Ca^{2+} gradients within cells have been proposed to initiate cell migration, exocytosis, lymphocyte killer cell activity, acid secretion, transcellular ion transport, neurotransmitter release, gap junction regulation and numerous other functions [6].

Ca^{2+} ions affect the cell cycle in more than one way. Depletion of the $InsP_3$ receptor-gated Ca^{2+} pool results in cell cycle arrest at G_0/G_1 and S phases. Calcium is necessary and sufficient for resumption of meiosis in marine eggs and have role in completion of meiosis and initiation of mitosis [7]. It is also found that gene transcription depends on how Ca^{2+} enters into the cell. Entry of Ca^{2+} through voltage-dependent L type Ca^{2+} channels and *N*-methyl-D-aspartic acid (NMDA) receptors initiates gene transcription through distinct DNA-regulatory elements. Intranuclear increase in Ca^{2+} initiates gene expression and cell cycle progression, but also can activate degradative processes in programmed cell death or apoptosis. Prolonged high calcium ion concentration activates nucleases that cleave DNA and degrade cell chromatin. Ca^{2+}-dependent proteases, phosphatases and phospholipases break DNA, resulting in a loss of chromatin structural integrity [8]. Many intracellular signal transduction pathways consider elevated calcium level as an important signal [9].

Modularization is a process which divides a network into smaller units for better understanding and analysis of the original network. The idea is used here to divide calcium and MAPK signaling pathways into smaller simple units called *modules*. There is no single definition available for a module. Hence certain criteria are used to define them. Here we have assumed that a module is a subset of the original biochemical network, which tends to be selfsufficient and have minimal dependency on the rest part of the network. The justification for dividing a network into a number of modules lies in the fact that the complexity of each module is much less than that of the entire pathway and is an easier means of studying the entire network by parts. Thus analyzing all the modules generated from a pathway separately, we can have a better operational view of the whole network.

Methods were developed for defining biochemical network modules in an unbiased fashion. These unbiased network modules were mathematically derived from structure of the whole network under consideration [10]. One way to organize the signaling reactions, might be to separate modules with clearly defined input and output, based on pathway and cellular compartments where relationships among modules may depend on the biological state and cellular context [11]. Another way of studying signaling pathways is to create operational boundaries, which do not exist in a cell. It should be noted that these modules might not correspond to conventional cell biological boundaries such as various membranes. Boundaries of such modules are often defined by functional input–output relationships. Modules may also reflect spatial locations in cytoplasm, as defined by protein scaffolds and anchors [12]. Based on absence of retroactivity, modules can also be defined as done in [13]. Thus division of a biological reaction network into smaller units highly facilitates its investigation.

Here we propose an algorithm for modularization of signaling pathways. Creation of modules starts with a member having maximum number of relations in a given network. The module grows in size by including neighbors of the starting member in successive steps. The neighbors are either included into or excluded from the module depending on the number of their relations being present its inside or outside. That is, if a member has less than or equal to a certain number of relations (known as complexity level c) outside the module, it gets included in the module. The term, complexity level, is specified and can be varied by the user. For comparative analysis, we have to select an appropriate c-value for each pathway. The algorithm is applied to calcium and MAPK signaling pathways of *H. sapiens* for different c-values and an appropriate c-value is selected for further study over different organisms. The effectiveness of the algorithm is demonstrated on two different signaling pathways, *viz.*, MAPK and calcium signaling pathways for different species. Hence we associate biological significance to each of the modules and compare the levels of development of these pathways in different species. The species we have considered for analyzing MAPK signaling pathways are *B. taurus* (cow), *C. familiaris* (dog), *D. melanogaster* (fruitfly), *H. sapiens* (human), *M. musculus* (mouse), *P. troglodytes* (chimpanzee), *R. norvegicus* (rat), *S. cerevisiae* (yeast) and *S. scrofa* (pig), and those for analyzing calcium signaling pathways are all the previous ones except fruitfly and yeast. The superior capability of the algorithm, in partitioning a signaling network into a set of biologically significant modules, over that of an existing community finding algorithm of Newman [1] has been demonstrated using the aforesaid pathways of *H. sapiens*.

The article is organized as follows. First we have given a review on existing methods for partitioning biological networks. The next section describes the proposed algorithm in details, and its application to an example network (Fig. 1). The results section contains thorough analysis of human calcium signaling pathway (Fig. 8), comparative analysis of modules obtained from calcium signaling pathway of different species, analysis of modules of human MAPK signaling pathway (Fig. 14), comparative analysis of modules obtained from MAPK signaling pathway of different species, a section on change of modules with increasing c-value, modules obtained by applying New-

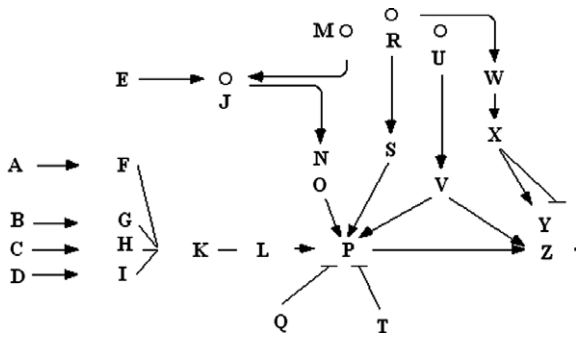

Fig. 1. An example network. Nodes are biomolecules. Ions are denoted with a circle followed by names and the rest (i.e., biomolecules) only by names. Edges are different for different kinds of relations that exist between two biomolecules. That is, → indicates activation, – indicates binding and –| indicates inhibition between two nodes.

man's community finding algorithm [1]. A detailed analysis of comparison between the two methods is included to show superiority of our algorithm over the other. The conclusion section provides a summary of the work explained in the paper.

2. A review on partitioning of biological networks

There exist various approaches for partitioning networks. These include hierarchical clustering techniques [14–16], graph partitioning [17,18], block modelling [19], differential equation based methods [20], cartographic representations [21]. Among them, approaches based on graph partitioning [17,18,22] and community structure detection [23,1,24] are popular. Algorithms based on these two concepts are vastly used to divide, study and analyze networks. Some of these methods have been applied to biological networks. Graph partitioning algorithms were applied mostly to nonbiological networks. These include VLSI (Very Large Scale Integration) [25,26], CAD (Computer Aided Design) [27,28], Hypertext Browsing [29], geographic information services [30], parallel computing [28], integrated circuit designing [31], and for biological networks like physical mapping of DNA (Deoxyribo Nucleic Acid) [32] among others. On the other hand, a huge variety of community detection techniques have been developed based on the notion of centrality measures [33,34], flow models [35], random walks [36,37], resistor networks [24], optimization [38], and many other approaches.

2.1. Graph partitioning techniques

A typical problem in graph partitioning is to divide of a set of tasks among the processors of a parallel computer so as to minimize the necessary amount of interprocessor communication [1]. In such an application, the number of processors is usually known in advance along with an approximate figure of the number of tasks that each processor can handle. Thus we know the number and size of the groups into which the network is to be split. Moreover, the goal is usually to find the best division of the network regardless of the fact whether a good division even exists or not.

Vast and complex biochemical networks have inherent non-local features that require the global structure to be taken into account in the decomposition procedure. It is important to know, the naturally occurring subnetworks of a network, while studying its functionality as a whole. Holme et al. [39] have proposed an algorithm for decomposing biochemical networks into subnetworks based on the global network structure. They have analyzed full hierarchical organization of biochemical networks (metabolic and cellular networks) of 43 organisms taken from the WIT database. The investigation of Jeong et al. [40] suggests the presence of the same topological scaling properties in metabolic networks that show striking similarities to the inherent organization of complex non-biological systems. Identifying recurrent patterns across multiple networks is also an important step to discover biological modules, especially from microarray datasets. Most of the existing algorithms are very costly in time and space for frequent pattern mining as the pattern sizes and network numbers increase. Hu et al. [41] have developed a novel algorithm, called CODENSE, to efficiently mine frequent coherent dense subgraphs across a large number of massive graphs. Unlike the other methods, this algorithm is scalable in the number and size of the input graphs, and adjustable in terms of exact or approximate pattern mining. Graph theoretical algorithms can also be used to identify backbone clusters of residues in proteins. The identified clusters show protein sites with the highest degree of interactions. Patra and Vishveshwara [42] have devised a method for identifying highly interacting centers (clusters) in proteins. This method can be applied to the problems such as identification of domains and recognition of structural similarities in proteins. Pathway analysis of large metabolic networks meets with the problem of combinatorial explosion of pathways. Schuster et al. [43] have developed an algorithm for metabolic pathway decomposition based on local connectivity of the metabolites. Applicability of the method is analyzed with metabolic networks of *M. pneumoniae*. Some studies have also been done on networks of *E. coli* and *C. elegans* by Wagner et al. [44].

2.2. Community finding algorithms

Community structure detection, by contrast, is the best thought of partitioning technique that is used to shed light on the structure of large-scale networks, such as social networks, internet and web data, or biochemical networks [1]. Community structure detection methods normally assume that the network of interest divides naturally into subgroups, if any, and the experimenter's job is to find those groups. The number and size of the subgroups are thus determined by the network itself and not by the experimenter.

In many networks, nodes are joined together in tightly knit groups, between which there are only looser connections. Girvan et al. [33] have proposed a method for detecting such communities. They have used the idea of centrality indices to find community boundaries. Community finding algorithms can also be applied to a network of relations among genes [45]. Wilkinson and Huberman [46] have studied a network of gene co-occurrences for colon cancer from the literature, and partitioned it into communities of related genes. Their method identifies communities where the component genes of each community are related by their functions. They have designed the partitioning procedure to be particularly applicable to large networks in which individual nodes may play a role in more than one community. Biological networks can be of different kinds. A metabolic network represents metabolic substrates and products with directed edges joining them. Protein interaction networks convey mechanistic physical interactions among proteins [47]. Expression of a gene may be controlled by other proteins (activators and inhibitors) in a genetic regulatory network. Hence a genome can be viewed as a switching network with vertices representing the proteins and directed edges representing dependence of protein production on the proteins at other vertices [47]. A robust approach to partition a network involves maximization of a benefit function called ''modularity'' over possible divisions of the network as proposed by Newman [23].

Metabolic and signaling pathways are shaped by the networks of interacting proteins whose production, in turn, is controlled by the genetic regulatory networks. Maslov and Sneppen [48] have quantified correlations among connectivities of interacting nodes and compared them to a null model of a network (a network with all links randomly rewired). They have found that for both protein interaction and gene regulatory networks, links between highly connected proteins are systematically suppressed, whereas those between a highly connected and lowly connected pairs of proteins are favored. This effect decreases the likelihood of cross talk between different functional modules of the cell and increases the overall robustness of a network by localizing effects of some perturbations. Stelling et al. [49] have devised a theoretical method for simultaneously predicting key aspects of network functionality, robustness and gene regulation from network structure alone. They have determined and demonstrated that the non-decomposable pathways are able to operate coherently at steady state by using *E. coli* central metabolism as an illustration.

A gene may have several connections, circuits and pathways that may crosslink and represent connected components. Guelzim et al. [50] have created a network of 909 genetically or biochemically established interactions among 491 yeast genes. After thorough analysis of the interaction network, it has been found that the number of regulating proteins per regulated gene has a narrow distribution with an exponential decay, while the number of regulated genes per regulating protein has a broader distribution with a decay resembling to a power law. As a whole, the yeast transcriptional regulatory network combines a small maximal diameter, an elevated local semi-clustering, a high number of feedback circuits and a global fragmentation. Here each small connected piece indicates towards implementation of a biological function, and the global fragmentation serves to limit inter-functional crosstalk at the transcriptional level.

Clustering properties of the reaction networks can be obtained from maps of known metabolic pathways. Raine and Norris [51] have investigated random connection model, random cluster model and accumulation model for construction of metabolic networks. The random cluster and accumulation models exhibited ''small-world'' (Small worlds are networks that are linked in such a way that they exhibit a high degree of clustering like ordered networks but a relatively short average number of links between any two nodes like random networks) features, in agreement with the structure of real biological networks, while random cluster and accumulation models also depict a long-tailed distribution of nodes of the taken networks.

Milo et al. [52] have defined ''network motifs'' as patterns of interconnections occurring in complex networks. Such motifs were found in networks from biochemistry, neurobiology, ecology and engineering. The motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of *E. coli* and *S. cerevisiae* or from those found in the world wide web. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in *C. elegans*. The authors have used motifs to define universal classes of networks. It is worth to detect and understand network motifs in order to gain insight into their dynamical behavior and to define classes of networks and network homologies. Motif detection in *E. coli* transcription regulation networks is also have been carried out by Shen-Orr et al. [53].

Newman et al. [24,23,1,47,38,45,33] have proposed a series of algorithms to find communities in various kinds of networks. These algorithms are designed to optimize a network's divisions based on the properties of the network itself. We have compared our method with an interesting community finding algorithm of Newman [1], which has already been applied to metabolic pathways along with other kind of networks. Newman's algorithm optimizes a quality function known as ''modularity'' over possible divisions of a given network. Modularity score is directly dependent on the network architecture in terms of adjacency matrix and eigenvalues of a symmetric matrix calculated from the adjacency matrix. Positive value of modularity indicates possible presence of modules in a network. One important aspect of the algorithm is that it refuses to divide a network if no good division exists. In other words, a negative value of modularity indicates no possible division of the given network. Throughout the paper we have referred this algorithm as Newman's com-

munity finding algorithm. One may refer to [1] for its further details.

3. The proposed algorithm

In order to decompose a network into several modules, we have proposed an algorithm which is described in this section. The algorithm views an entire biochemical pathway as a graph having gene products and chemical compounds as vertices, and edges being different kinds of interactions among them. An edge can be a protein–protein interaction or protein-compound interaction or a link to another map. For simplicity, here we have not taken the links to other maps into account. A module is a subset of the original network, which is defined in Section "Introduction". Before describing the algorithm, let us define some useful terms.

- E: Set of all nodes (representing gene products and chemical compounds) present in a network, where each node must have atleast one relation, i.e., isolated nodes are not included in this set
- M: Set of nodes present in a module (a part of network)
- k: *Extension* index (stage of inclusion of immediate neighbors of nodes in a module)
- c: Complexity level (a value fixed by the user determining the inclusion (exclusion) of nodes in (from) a module)
- M^k: Set of nodes present in a module after kth extension
- N_S: Set of succeeding nodes of a given node
- n_s: An individual member of N_S
- N_P: Set of preceding nodes of a given node
- n_p: An individual member of N_P
- r: Type of interaction that exists between n_p and n_s; $r = a$ depicts a relation of activation, $r = b$ depicts a relation of binding or association, $r = i$ depicts a relation of inhibition, $r = d$ depicts a relation of indirect effect between nodes n_s and n_p
- N_R: Set of relations (interactions that can exist between two nodes)
- $n_r = (n_p, n_s, r)$: An individual member of N_R
- R_{np}: Total number of relations that exist with n as the preceding node
- R_{ns}: Total number of relations that exist with n as the succeeding node
- M_P: Set of permanent nodes (nodes having all their relations inside a module)
- *Max*: A function that detects maximum value among all elements present in a given set

The total number of relations with n as either a preceding or succeeding node is given by

$$T_n = R_{np} + R_{ns} \tag{1}$$

Since R_{np} and R_{ns} are outdegree and indegree, respectively, of a node n, T_n is equivalent to total degree of the node n.

Table 1
Relations found among members present in the example network

Sl. No.	Preceding node	Succeeding node	Relation
01	A	F	a
02	B	G	a
03	C	H	a
04	D	I	a
05	F	K	b
06	G	K	b
07	H	K	b
08	I	K	b
09	K	L	b
10	L	P	a
11	Q	P	i
12	T	P	i
13	P	Z	a
14	E	J	a
15	J	N	a
16	M	J	a
17	O	P	a
18	R	S	a
19	S	P	a
20	R	W	a
21	W	X	a
22	X	Y	a
23	X	Y	i
24	U	V	a
25	V	P	a
26	V	Z	a

Capital alphabets represent members of the network. Various types of relation between a pair of members are depicted by small alphabets. There are three kinds of relations (a - activation; b - binding/association; i - inhibition).

Table 2
Calculation of total-relation of members present in the example network

Sl. No.	Node name	Out-relation	In-relation	Total-relation
01	A	1	0	1
02	B	1	0	1
03	C	1	1	2
04	D	1	0	1
05	E	1	0	1
06	F	1	1	2
07	G	1	1	2
08	H	1	1	2
09	I	1	1	2
10	J	1	2	3
11	K	1	4	5
12	L	1	1	2
13	M	1	0	1
14	N	1	1	2
15	O	1	0	1
16	P	1	6	7
17	Q	1	0	1
18	R	2	0	2
19	S	1	1	2
20	T	1	0	1
21	U	1	0	1
22	V	2	1	3
23	W	1	1	2
24	X	2	1	3
25	Y	2	0	2
26	Z	0	2	2

Total number of relations of a node (n) is the sum of relations with n as a preceding node (out-relations) and relations with n as a succeeding node (in-relations).

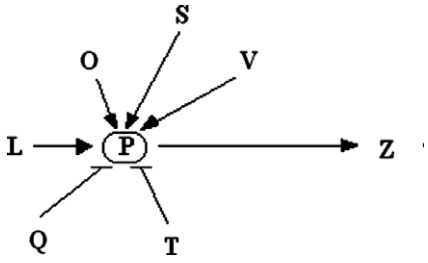

Fig. 2. Initial module of the example network. First steps of the algorithm involves creation of an initial module, taking the immediate neighbors of the starting node into account. For $c = 2$, nodes L, Q, T, Z, V, S, O get included in first extension, as immediate neighbors of node P.

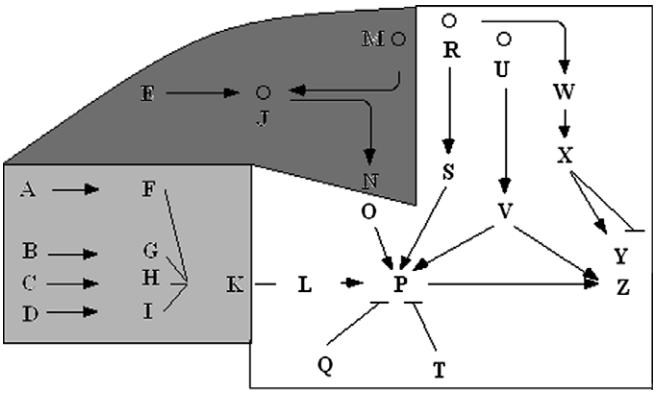

Fig. 4. The modularized example network. After complete modularization of the example network, we get 3 modules, i.e., dark gray colored region—module J, white region—module P and light gray region—module K.

T_n represents the total number of relations associated with node n. T_{n^k} stands for the total number of relations of node n that gets included in a module during kth *extension*. Likewise, T_M represents a set, comprising of T_n-values, where $n \in M$.

3.1. Description

The algorithm starts with detection of a node n having maximum number of relations in the node pool E for a given network. Considering the detected node as the starting point (the starting member being always a permanent member), an initial module is created for relations r, where n is either a *predecessor* or a *successor*. Thus the module is created by including immediate neighboring nodes of n. Here an eventuality may arise where more than one node have maximum number of relations. Then any one of the nodes (having maximum number of relations) that is encountered first by the algorithm is taken as the start point by default (followed by the others).

Once a module is initialized, the total number of relations (T_n) of every individual member is considered. For a node in a module, if the number of relations lying inside the module is equal to the total number of relations associated with the node, the member is considered to be *permanent*. If a node in a module has more than c relations that lie outside the module, it gets *excluded* from the module along with decreasing the previous *non-permanent* nodes'

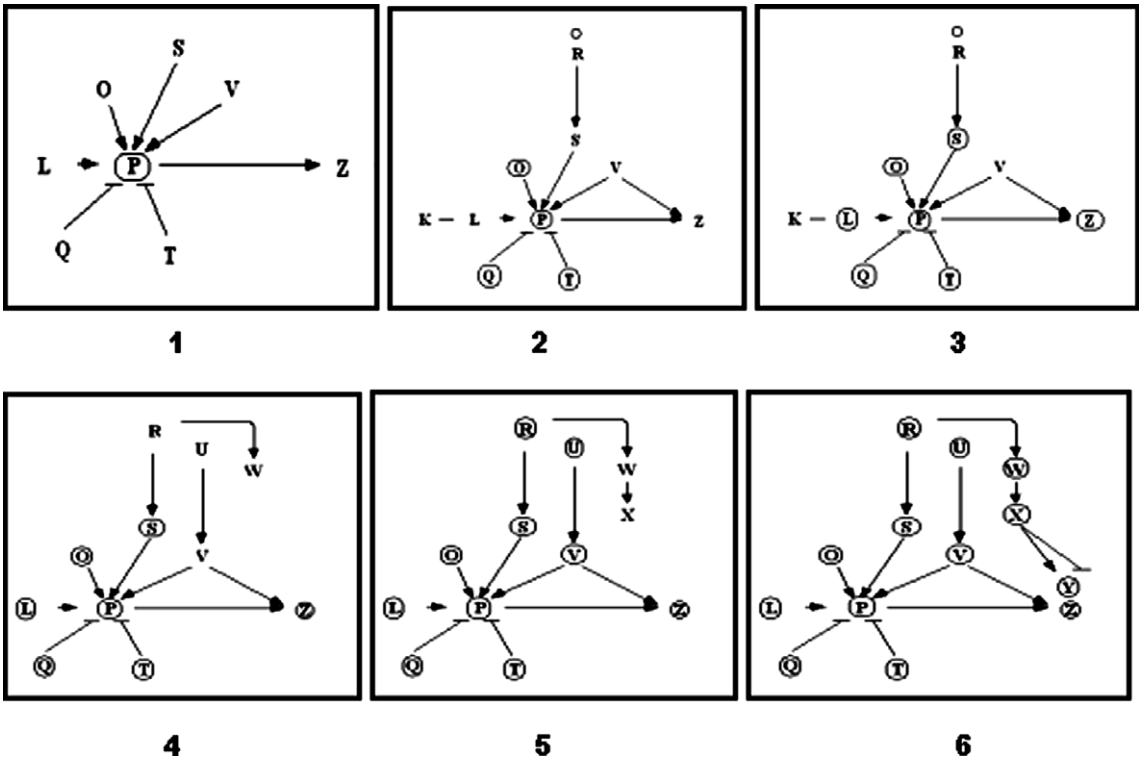

Fig. 3. Stages in construction of Module P. This figure gives stepwise construction of module P for $c = 2$. After each extension, nodes having all their relations inside the module are declared *permanent*. Nodes having more than two out-relations are excluded from the expanding module, and the rest are taken as under consideration members and their immediate neighbors are included during the next phase of extension.

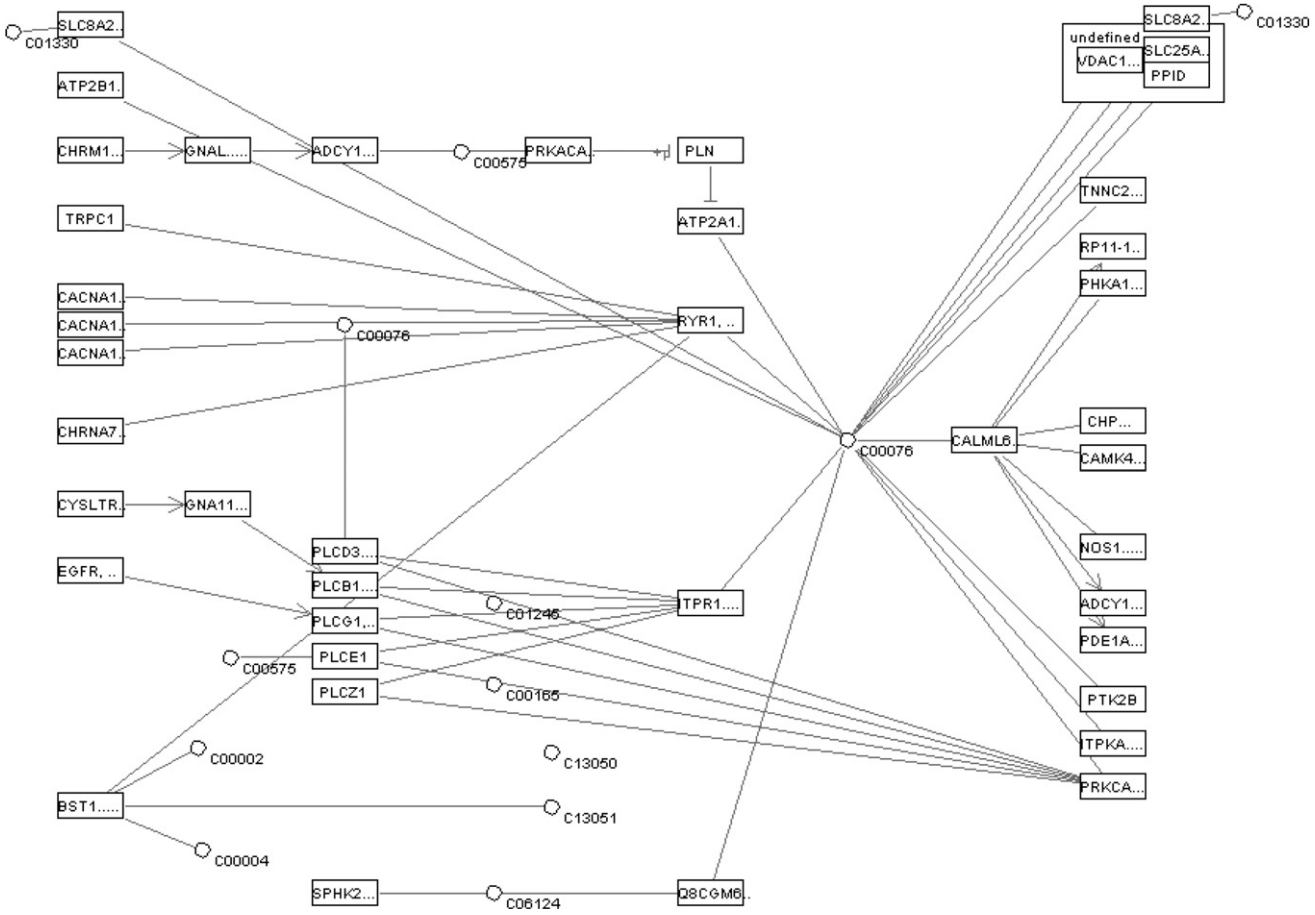

Fig. 5. KGML layout for calcium signaling pathway of *H. sapiens*. KEGG/Pathway database provides information of calcium signaling pathway in xml format that can be represented graphically by java (KGML layout).

total relations T_n by one. These *extension* and *exclusion* processes continue till there is no new immediate neighboring node to be included, or no node is left to be declared *permanent*. It is to be mentioned here that once a member is declared *permanent*, it gets removed from the node pool E. Hence the chance of a single member to be included in more than one module is avoided. Also, if a member appears more than once in a network, its positional significance is taken into account. That is, if a member X is present four times in a network, it will be considered four times as X1, X2, X3 and X4.

After successful completion of the creation of a module, the algorithm will search for another starting point and repeat the above mentioned steps to create another module. This process of creating modules (one by one) will continue till exhaustion of all the nodes present in the node pool E. The basic steps followed to create the modules are presented in a pseudo code (Algorithm 1).

Algorithm for creation of modules from a network

Ensure: $E \neq \phi$
1: Find start/central node
If $(T_n \leftarrow Max\{T_M\})$ **then**
$n \leftarrow$ start point/central node
$M_P \leftarrow M_P \cup \{n\}, E \leftarrow E - \{n\}$
$k \leftarrow 0$
end if
2: Extend module
for $(k \leftarrow k+1)$ **do**
select nodes from N_S and N_P of n and put in M^k
end for
3: Check permanency of nodes
if $N_S \cup N_P \subset M^k$ for a node n^k **then**
$E \leftarrow E - \{n^k\}, M_P \leftarrow M_P \cup \{n^k\}$
end if
4: Exclude node
if $[T_n^k$ - number of nodes in M^k related to $n^k] > c$ **then**
$M^k \leftarrow M^k - \{n^k\}$
for $(n^{(k-1)} \notin M_P)$ **do**
$T_{n^{(k-1)}} \leftarrow [(T_{n^{(k-1)}}) - 1]$
end for
end if
5: Build a complete module
repeat
Step 2-4
until $M^k \subset M_P$
6: Create next module
repeat
Step 1-5
until $E = \phi$

Table 3
List of modules of human calcium signaling pathway for different c-value

Sl. No.	Module name	$c = 1$		$c = 2$		$c = 3, 4$		$c = 5$		$c = 6$		$c = 7$	
		Node	Rel	Node	Rel	Node	Rel	Node	Rel	Node	Rel	Node	Rel
01	(C00076)2	24	23	25	24	29	28	40	43	46	51	54	59
02	CALML6	8	7	8	7	8	7	8	7	8	7		
03	(C00076)1	6	5	7	6	7	6	6	5				
04	C01245	2	1	9	8	10	12						
05	C00165	1	Nil	1	Nil								
06	BST1	4	3	4	3								
07	PLCE1	2	1										
08	PLCG1	2	1										
09	PLCB1	3	2										
10	PLCD3	1	Nil										
11	RYR1	1	Nil										

The column Node indicates number of nodes and column Rel gives number of relations present in a module.

3.2. An example

The hypothetical network in Fig. 1 is considered for generation of modules. The network contains 26 nodes and 26 relations existing among the nodes. The set of relations (N_R) is given in Table 1. The total number of relations (T_n), for all members (n), is considered for selecting the node with maximum number of relations, as the starting point of an originating module. Total number of relations possessed by each node is calculated from Table 2.

Here node P is having the highest number of relations with other nodes of the given network. So P is the starting point of the first module. The module resembles to Fig. 2,

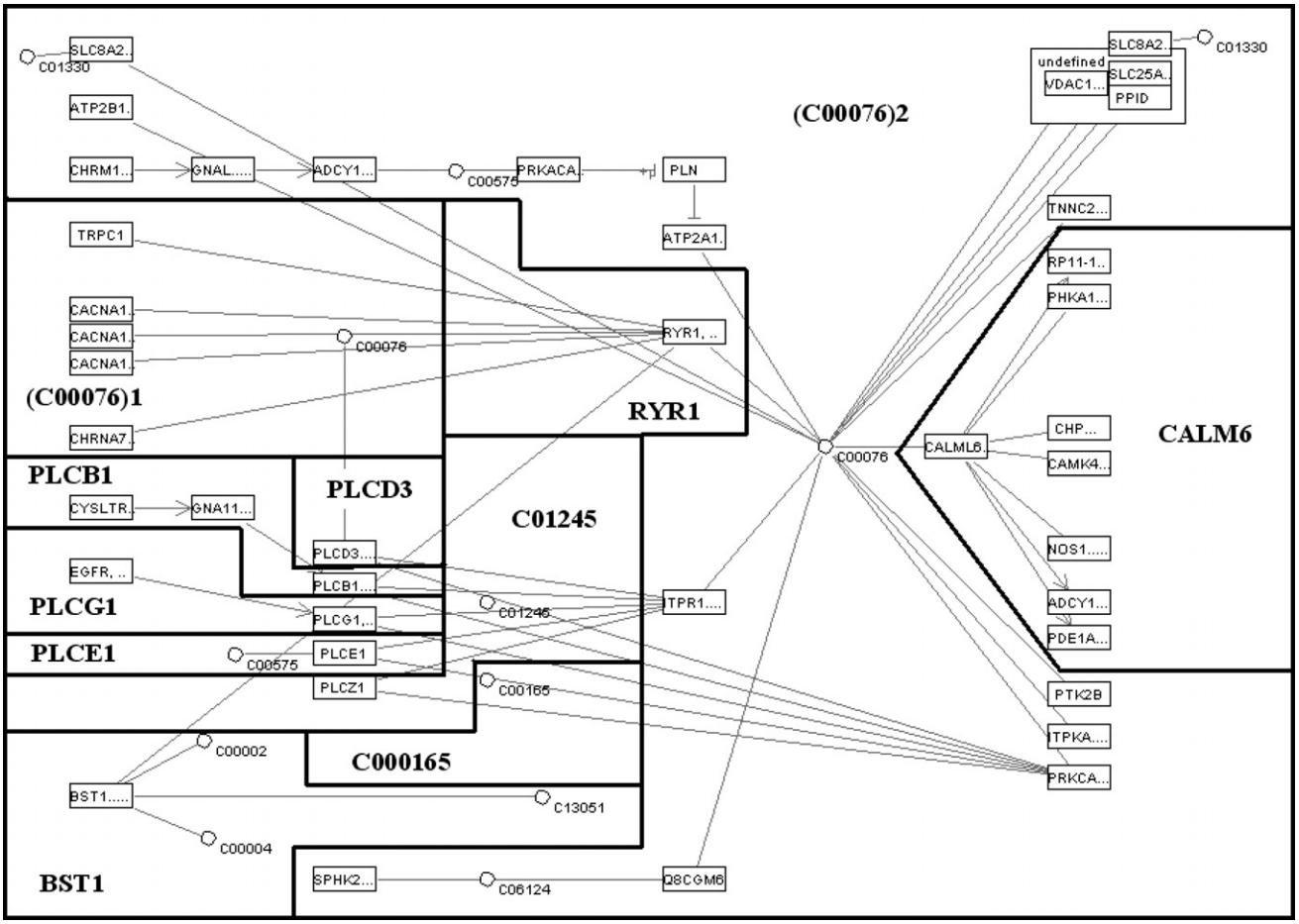

Fig. 6. Modules of calcium signaling pathway of *H. sapiens* for $c = 1$. For $c = 1$, calcium signaling pathway of *H. sapiens* is divided into 11 modules. Here black lines separate the modules from each other. The figure shows that many small modules are resulted in due to low complexity level. This has led to over splitting of the network. Each module is named after its starting node and marked with bold faced letters.

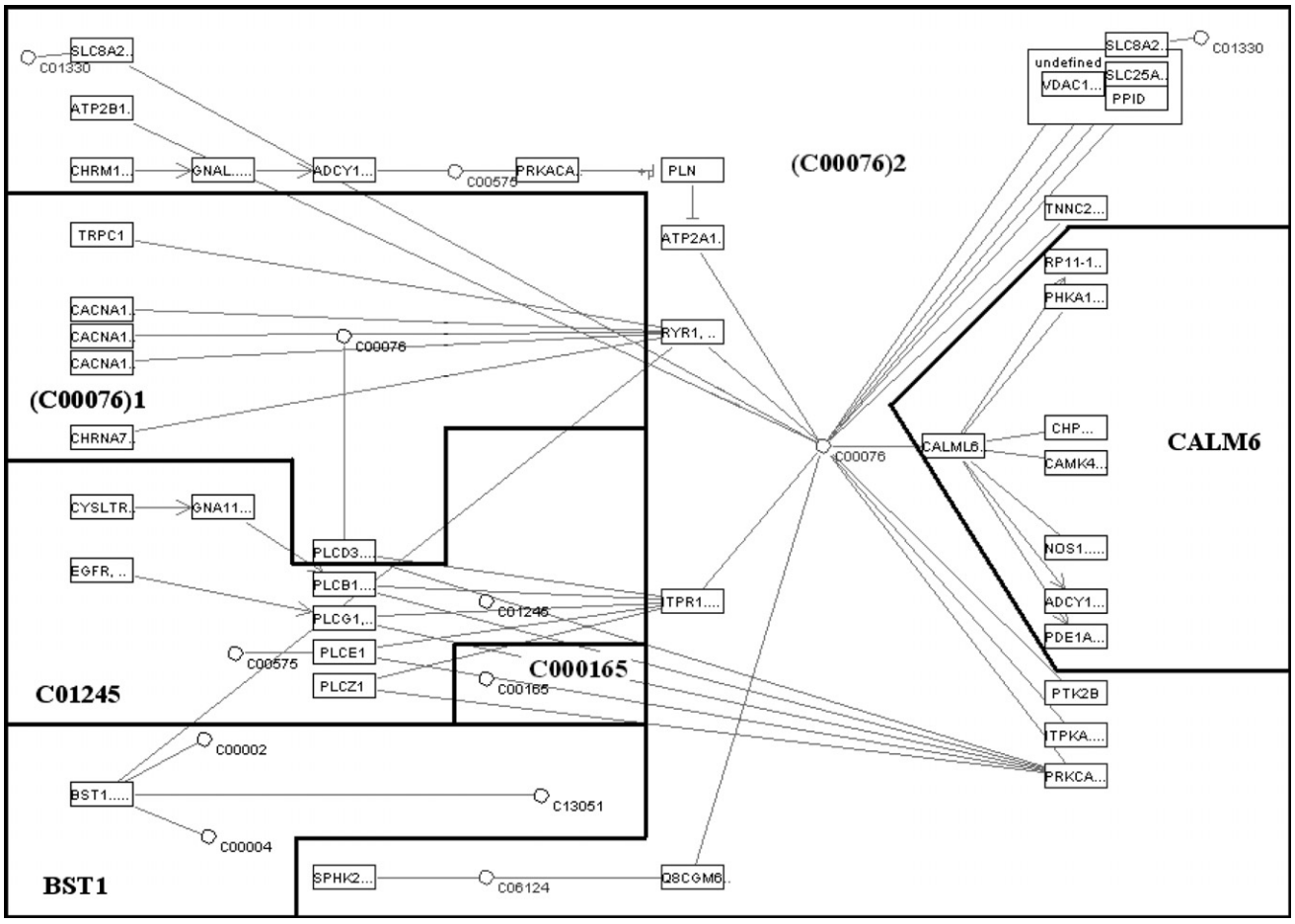

Fig. 7. Modules of calcium signaling pathway of *H. sapiens* for $c = 2$. We have obtained 6 modules from calcium signaling pathway of *H. sapiens* for $c = 2$. These modules are separated from each other by black lines in the figure. Here we are getting less number of modules than in Fig. 6.

after first extension, by including immediate neighbors of node P. The steps for determining the modules of the network are described below.

1. For T_{Q^1}, T_{T^1} and T_{O^1}, it is found that $N_S \cup N_P \subset M^1$. Hence O, Q and T became the permanent members.
2. After second extension, for T_{L^1}, T_{S^1} and T_{Z^1}, it is found that $N_P \cup N_S \subset M^2$. So they were also considered as the permanent members of the module. But T_{K^2}—(number of nodes in M^2 related to K^2) $= 4 - 1 = 3 > 2$. Here c is taken as two. Since node K has more than two out-relations lying outside the present module, K cannot be a member of the module created with P as the starting node.
3. After third extension, for T_{V^1}, T_{R^2} and T_{U^2}, it is again found that $N_S \cup N_P \subset M^3$. So except W, every member present in M^3 in the module is permanent, i.e., they have been excluded from E.
4. Fourth extension has made W permanent. Likewise, after 5th and 6th extension, we have got $M^6 = M_P$. Hence creation of one module is now complete. Here we have named the modules by the name of their starting node. Module P contains 12 permanent members.
5. The whole process is repeated with the node K, the node with maximum relations among the left over nodes of E.
6. After creation of three modules namely P, K and J, node pool E becomes empty.

Fig. 3 shows different stages during construction of module P, and the modularized entire network is given in Fig. 4.

4. Results and comparative analysis

In this section the proposed modularizing algorithm is applied to some real life biological networks, *viz*., calcium signaling pathways belonging to *B. taurus* (cow), *C. familiaris* (dog), *H. sapiens* (human), *M. musculus* (mouse), *P. troglodytes* (chimpanzee), *R. norvegicus* (rat) and *S. scrofa* (pig), and MAPK signaling pathways of *B. taurus*, *C. familiaris*, *D. melanogaster* (fruit fly), *H. sapiens*, *M. musculus*, *P. troglodytes*, *R. norvegicus*, *S. cerevisiae* (yeast) and *S. scrofa*. The data is taken from KEGG/Pathway database (http://www.genome.jp/kegg/pathway.html#environmental) [54–56]. We have considered XML files representing the KGML (KEGG Markup Language) lay-

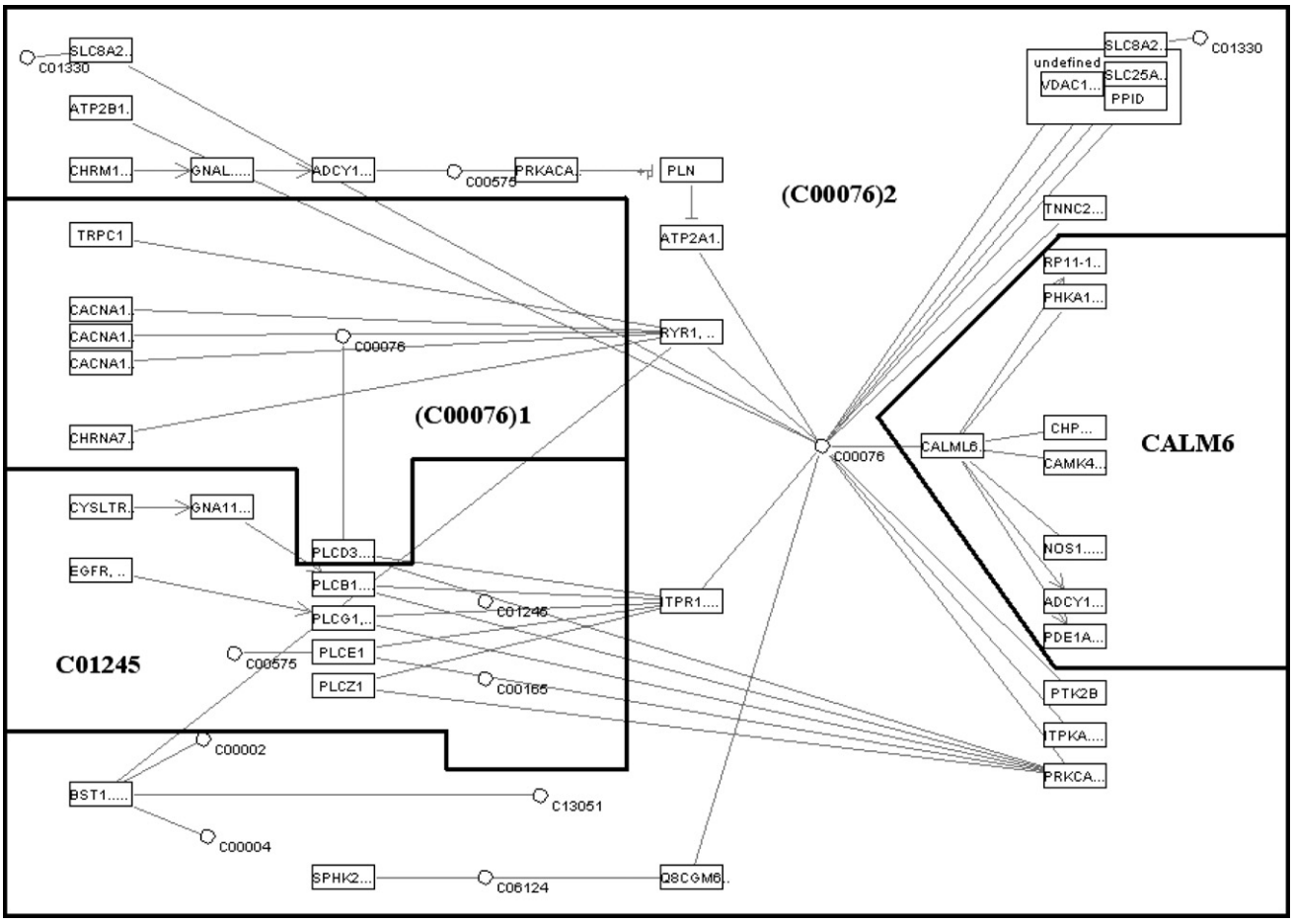

Fig. 8. Modules of calcium signaling pathway of *H. sapiens* for $c = 3$ and 4. Here the network is divided into 4 parts due to modularization. For $c = 3$ and 4, we are getting identical modules.

outs for calcium signaling pathways. Since KGML layout for MAPK signaling pathway is not available in the database except fruitfly, mouse, rat and yeast, we have considered available pictorial representations for these species. Modules have been created from both calcium and MAPK signaling pathways of *H. sapiens* for different values of c. For comparative analysis of a signaling pathway of different species, we have considered a particular c-value, for which the modules appear to be biologically significant. Using these c-values, we have compared these two above mentioned pathways for the aforesaid species in terms of the levels of development. This is followed by the comparison of performance of the proposed algorithm with an existing community finding algorithm of Newman [1]. As mentioned earlier, a module is named with its starting node.

4.1. Calcium signaling pathway of H. sapiens

Calcium signaling pathway of *H. sapiens* contains 55 nodes. One node (C13050) is isolated. So $|E| = 55 - 1 = 54$. These 54 nodes are having 59 relations (i.e., $|N_R| = 59$) among them as shown in Fig. 5. Modules are created from the same pathway for complexity level (c) of 1, 2, 3, 4, 5, 6, 7 and above. The results are given in Table 3.

4.1.1. Modularization for c = 1

For $c = 1$, we get 11 modules. Module *(C00076)2* emphasizes role of plasma membrane, endoplasmic reticulum and mitochondria in calcium ion balance of cells. But some receptors like RYR (Ryanodine receptors) present in ER membrane are not included in this module. Module *CALML6* represents role of calmodulin like proteins (calcium binding proteins) that upon binding with free calcium ions change confirmation and trigger other enzymes and ion channels. *(C00076)1* module contains calcium channels present in plasma membrane for import purpose. Module *BST1* deals with calcium ion flow from outside to inside of bone marrow cells but the way its intracellular balance is maintained is not clear in the module. Moreover for $c = 1$, the network is splitting profusely. Over splitting is giving rise to a lot of small modules as shown in Fig. 6. This is the reason why we are unable to assign any biological significance to the rest seven modules. So modularization of the same network is done for $c = 2$.

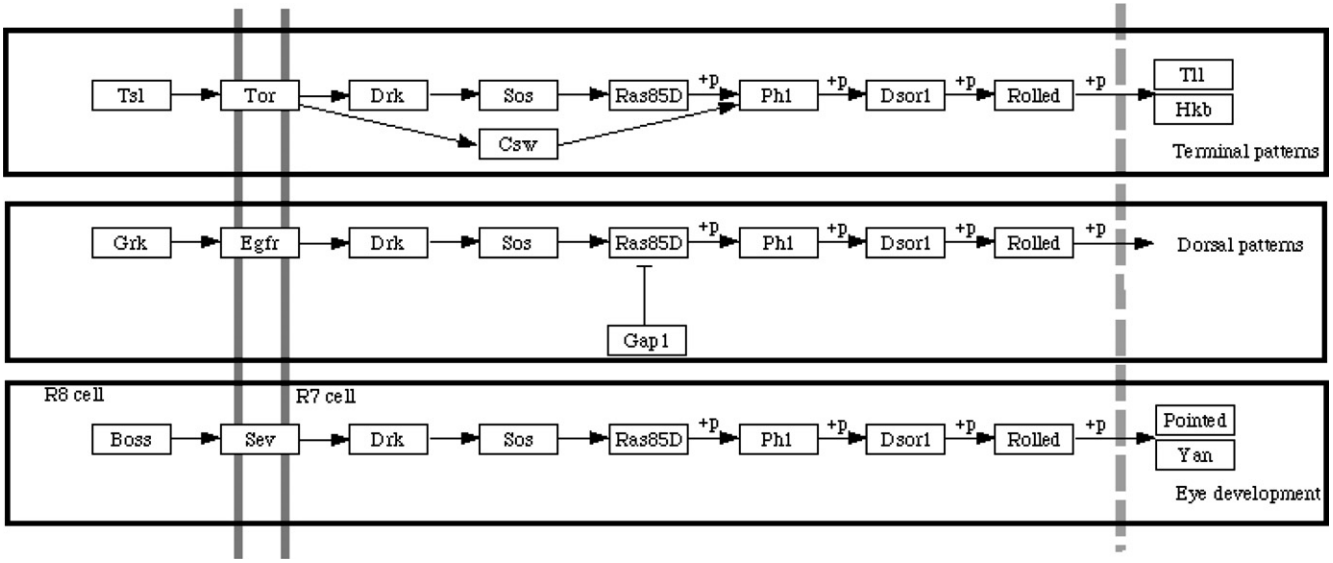

Fig. 9. Modularized MAPK signaling pathway of fruitfly. The simple network itself exists as a combination of three separate independent modules.

4.1.2. Modularization for c = 2

For $c = 2$, 6 modules are obtained as shown in Fig. 7. Modules *CALML6* and *BST1* remain unchanged. In module (*C00076*)*2*, Ryanodine receptors are included, making a clear picture of overall calcium ion flow and balance in a cell. The module (*C00076*)*1* is increased by one node PLCD3. The changed module shows plasma membrane based calcium import channels and interaction of the imported calcium ions with one of the PLC (Phospholipase C) group. *C01245* module includes proteins belonging to PLC family and their relation with C01245. From prior knowledge we know that PLC group members break into C01245 as a result of activation. C01245 molecule is a ligand for ITPR1 (inositol 1,4,5-triphosphate receptor, type 1) present in ER membrane. This module is resulted in due to merging of the modules *C01245*, *PLCB1*, *PLCG1* and *PLCE1* found for $c = 1$. So the problem of over splitting noticed for $c = 1$ is reduced here. But still we are left with the problem of module *BST1* as described above and a singleton module (i.e., a module comprising of a single node) *C00165* to describe. Its difficult to analyze such small modules. This has led to modularization of the network for $c = 3$.

4.1.3. Modularization for c = 3

Four modules are created for $c = 3$ as shown in Fig. 8. Module *BST1* for $c = 2$ is merged with the module (*C00076*)*2* that gives a complete explanation of calcium ion balance in bone marrow cells through Ryanodine receptors present in ER membrane. Module *C00165* for $c = 2$ is

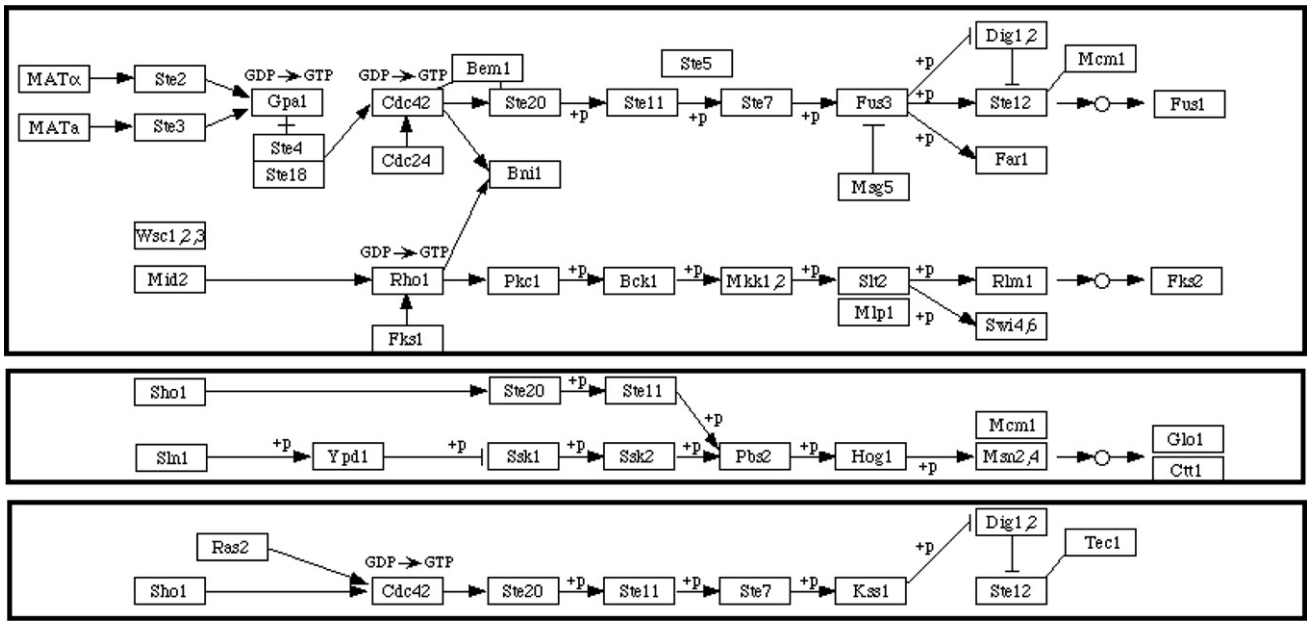

Fig. 10. Modularized MAPK signaling pathway of yeast. The simple network splits into three separate independent modules.

Table 4
Modules obtained from calcium signaling pathway of different species for $c = 3$

Human, rat and mouse		Cow		Pig		Dog		Chimpanzee	
Name	Nodes	Name	Nodes	Name	Nodes	Name	Nodes	Name	Nodes
(C00076)2	29	(C00076)2	13	(C00076)2	9	(C00076)2	5	(C00076)2	4
CALML6	8	C01245	10	CHRM1	2	GNAS	3	469986	3
(C00076)1	7	CALM2	5	(C00076)1	2	PTGER3	2		
C01245	10	CD38	4						
		GNAS1	5						

This table contains information about modules obtained from calcium signaling pathway of *H. sapiens* (human), *R. norvegicus* (rat), *M. musculus* (mouse), *B. taurus* (cow), *S. scrofa* (pig), *C. familiaris* (dog) and *P. troglodytes* (chimpanzee). The column Name gives name of the module and the column Nodes indicates the number of nodes present in a module.

merged with module *C01245*. C00165 is a byproduct when PLC group members break to C01245. Like C01245, it is not a ligand for ITPR1. It binds with PKC (protein kinase C) that takes part in controlling PM based calcium ion channels. But we are able to decipher its role clearly only after its merging with module *C01245*. For $c = 2$, where C00165 is included in another module, it is confusing to decipher and understand this information. Thus it appears that we can associate some biological significances to these modules. We are getting exactly similar modules for $c = 4$.

4.2. Fixing the c-value

Now question arises once we get biologically significant modules at some value of c, whether we should proceed further and continue modularization at higher values of c

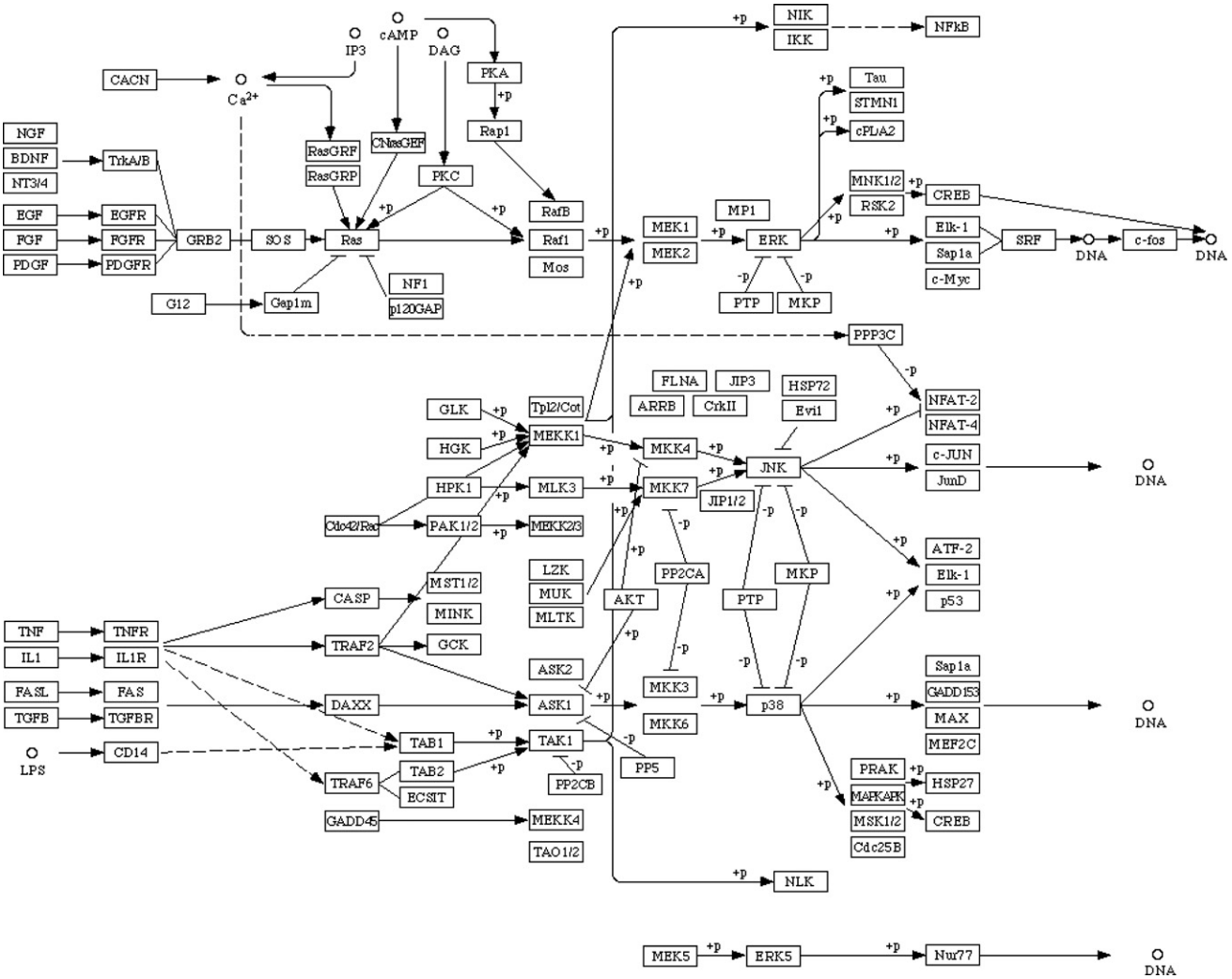

Fig. 11. MAPK signaling pathway of *H. sapiens* as given in KEGG/PATHWAY database. The isolated nodes are omitted from the network as our algorithm cannot consider them.

Table 5
List of modules of human MAPK signaling pathway for different c-values

Sl. No.	Module name	$c=1$		$c=2$		$c=3$		$c=4$		$c=5$	
		Node	Rel	Node	Rel	Node	Rel	Node	Rel	Node	Rel
01	JNK	14	15	14	15	21	23	22	25	22	25
02	p38	13	14	13	15	13	15	40	47	40	47
03	ERK	16	18	17	19	17	19	17	19	17	19
04	Ras	10	9	15	15	16	18	27	29	33	42
05	MEKK1	7	6	7	6	13	14	11	12	7	6
06	TAK1	4	3	4	3	14	16	3	2	3	2
07	MKK4	3	2	3	2	4	3	2	1	2	1
08	MKK7	4	3	4	3	4	3	4	3	4	3
09	MEK1	1	Nil	1	Nil	1	Nil	1	Nil		
10	MEK2	1	Nil	1	Nil	1	Nil	1	Nil		
11	ASK1	2	1	7	6	7	6				
12	TNFR	2	1	2	1	2	1				
13	GRB2	7	6	7	6	11	10				
14	JIP3	1	Nil	1	Nil	1	Nil				
15	MKK3	2	1	2	1	2	1				
16	MKK6	1	Nil	1	Nil	1	Nil				
17	TrkA/B	4	3	4	3						
18	IL1R	2	1	2	1						
19	Ca^{2+}	3	2	3	2						
20	CASP	3	2	3	2						
21	TRAF2	2	1	2	1						
22	TRAF6	2	1	2	1						
23	TAB1	3	2	3	2						
24	RafB	2	1	1	Nil						
25	Raf1	1	Nil	1	Nil						
26	Tp12/cot	1	Nil	1	Nil						
27	MLK3	2	1	3	2						
28	NIK	2	1	2	1						
29	IKK	1	Nil	1	Nil						
30	JIP1/2	1	Nil								
31	PP2CA	1	Nil	1	Nil						
32	DAXX	5	4								
33	PKC	2	1								
34	PKA	1	Nil								
35	Mos	1	Nil								
36	MP1	1	Nil								
37	ERK5	4	3	4	3	4	3	4	3	4	3
38	GADD45	2	1	2	1	2	1	2	1	2	1

The column Node indicates number of nodes and column Rel gives number of relations present in a module.

to get more meaningful modules or stop the process. To get a logical answer, we continue modularization for $c = 5$, 6 and other higher values. We get 3 modules for $c = 5$. Module (*C00076*)*2* is increased by several nodes and relations, which make it large and complex, hence our primary objective of dividing a complex network to simpler units fails here. (*C00076*)*1* module is decreased by one node and one relation that again gives rise to the already discussed problem of calcium ion balance inside the module. Module *CALML6* is identical to that of obtained for lower values of c. For $c = 6$, we have got only 2 modules. The whole network is divided into two parts, i.e., the unchanged *CALML6* module and (*C00076*)*2* module comprising the rest part of the network. In quest of a solution it only aggravated our problem. For $c = 7$ and higher values, the whole network rounds up to a single module.

So in general we can assume after a certain level, modularization with increasing c-value will yield similar results with that of previous complexity level or the modules will be enough larger making their study and analysis difficult. As our objective is to do a simplified study of a network and we are getting approximately biologically significant modules for $c = 3$, we have fixed c-value to 3 for calcium signaling pathways. This value of c is used in the next subsection for comparing calcium signaling pathways of other species.

4.3. Comparative study on modules of calcium signaling pathways of different species

Here the proposed algorithm is applied to calcium signaling pathways of 6 different species available in KEGG/Pathway database for $c = 3$. The original (*C00076*)*2* module that was found in *H. sapiens* exists in two parts namely, ((*C00076*)*2* and *469986*) in *P. troglodytes*. Here (*C00076*)*2* module is under developed compared to that in *H. sapiens*. Role of endoplasmic

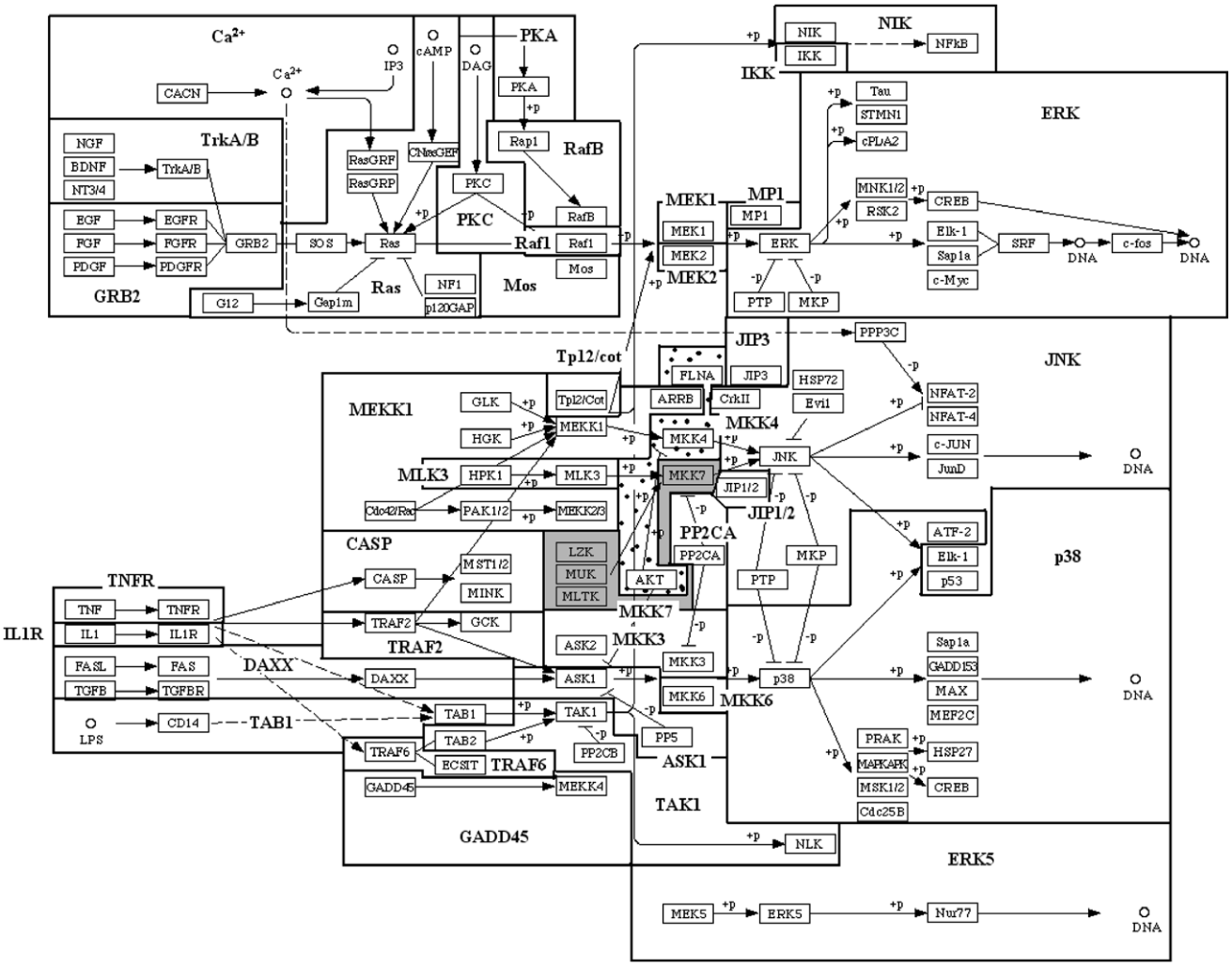

Fig. 12. Modularized MAPK signaling pathway of *H. sapiens* for $c = 1$. Here the network is divided into 38 modules. Twelve of them are singleton and very few are eligible for consideration of biological significance.

reticulum in calcium balance is negligible. However, mitochondria plays significant role in this regard. Absence of coordination among PM based ion channels, ER receptors and mitochondria is evident and indicates little role of calcium ions in signal transduction for *P. troglodytes*. The rest modules that we have found in *H. sapiens* are not present here. In case of *C. familiaris* (dog), module *(C00076)2* is a bit developed. It shows the role of calcium ions in muscle contraction, a fact not being shown in the same module for chimpanzee. In addition, part of module *C01245* is detected.

In *S. scrofa* (pig), for the first time, plasma membrane, endoplasmic reticulum and mitochondria are coordinating with each other to maintain calcium ion balance in module *(C00076)2*. Still it is a far cry from calcium signaling mechanism of *H. sapiens*. Here two members of module *(C00076)1* are also detected. For calcium signaling pathway of *B. taurus* (cow), we have got 5 modules. The original *(C00076)2* module exists in three parts (module *(C00076)2*, module *GNAS1* and module *CD38*) with several members missing. Module *C01245* is fully developed but contains two members of the under developed module *(C00076)1*. Partly developed module *CALM2* shows the role of calcium binding proteins in calcium signaling. As a whole calcium signaling pathway of *B. taurus* shows highest similarity with that of *H. sapiens* among the 6 species we have considered for our study. As calcium signaling pathway of *H. sapiens*, *R. norvegicus* and *M. musculus* are identical, we got similar modules for these 3 species. Table 4 gives data about modules present in calcium signaling pathways of these 7 different species. Analysis of these modules obtained from different species leads us to a conclusion that *H. sapiens*, *R. norvegicus* and *M. musculus* have highly developed calcium signaling mechanisms, *B. taurus* and *S. scrofa* lie as intermediates. But that of *C. familiaris* and *P. troglodytes* is very much under developed. In order to restrict the size of the article, we have included here the figures corresponding to *H. sapiens* only.

It may be mentioned here that in certain species when part of the pathway is only functional, modularized study is helpful. For example, module *(C00076)2* of human calcium signaling pathway is consistent among our taken set

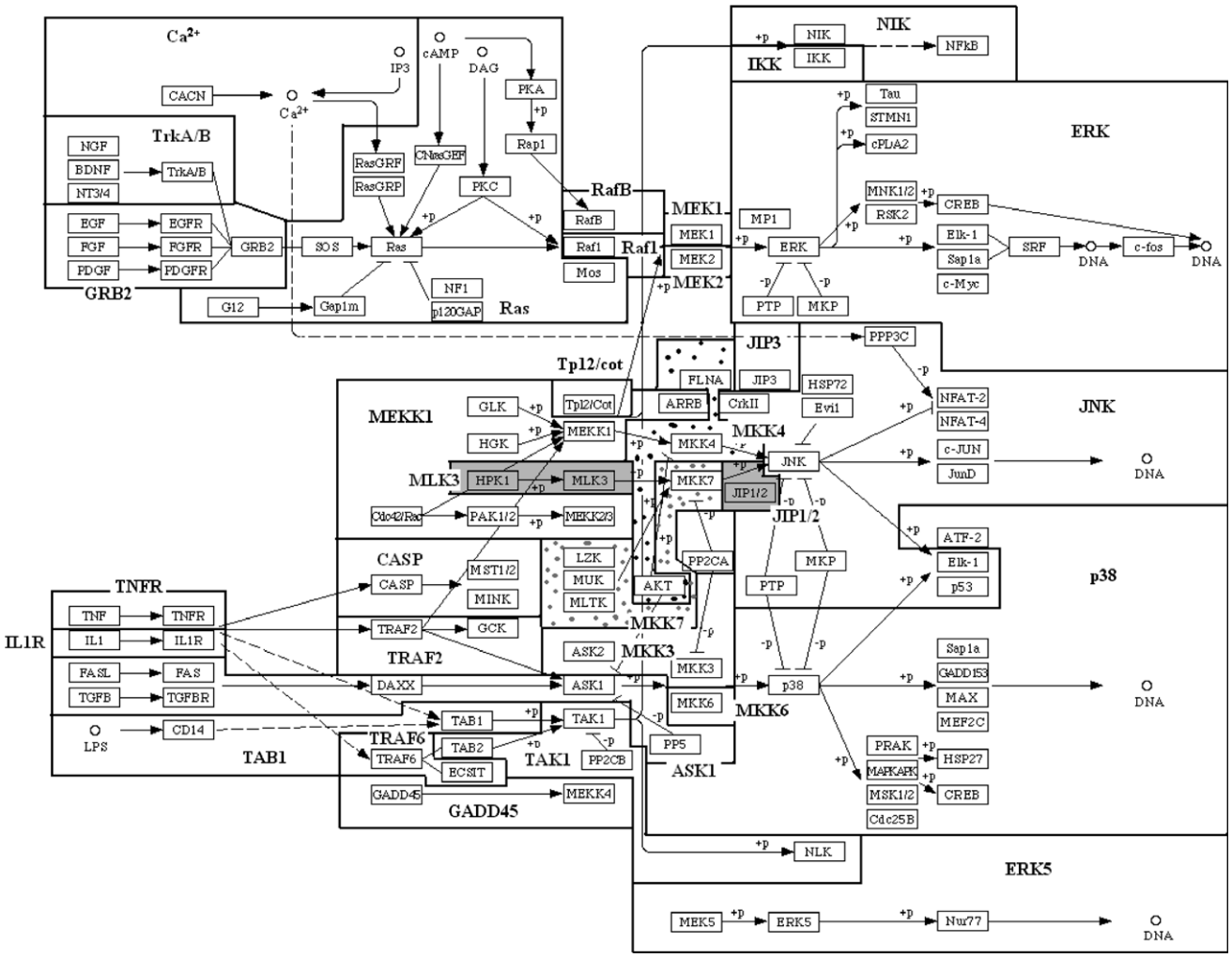

Fig. 13. Modularized MAPK signaling pathway of *H. sapiens* for $c = 2$. The network splits into 32 modules.

of species in varying size (Table 3), while other modules are not. So one can avoid the other modules and compare module *(C00076)2* of human calcium signaling pathway with that of other species instead of comparing the whole pathway to get a comparative view among them.

4.4. MAPK signaling pathway of H. sapiens

MAPK signaling pathway of *H. sapiens* is a complex network of 135 nodes and 182 relations ($|N_R| = 182$). TAO1/2 is the only isolated node present in this network, i.e., $|E| = 135 - 1 = 134$ (Fig. 11). Modules are created from the same pathway at complexity level of 1, 2, 3, 4, 5 and higher. A list of modules created for each *c*-value is given in Table 5.

4.4.1. Modularization for c = 1

For $c = 1$, MAPK signaling pathway of *H. sapiens* gets divided into 38 modules (Fig. 12). Twelve of them are singleton modules. Among the rest modules, only six seem to give any biological significance, i.e., modules *JNK*, *p38*, *ERK*, *Ras*, *MEKK1* and *GRB2*. Module *ERK* along with modules *TrkA/B*, *GRB2* and *Ras* roughly represent the conventional MAPK signaling pathway. Modules *JNK* and *MEKK1* are parts of the JNK pathway. p38 pathway is represented by a module named *p38*. The partly known ERK pathway is represented by module *ERK5*. Here *c*-value is very low. So the network is facing the problem of over splitting and we go for higher complexity values.

4.4.2. Modularization for c = 2

For $c = 2$, 32 modules are obtained as shown in Fig. 13. Module *JNK*, *p38*, *ERK*, *MEKK1*, *GRB2* and *ERK5* remain unchanged except increase by a single node or relation. Module *PKC*, *PKA*, *Mos* and *Ras* merge to give rise to a larger meaningful module. In addition *ASK1* emerges as a major module here. Still 9 singleton modules and 14 small to medium sized modules are left to be explained. So we switch to modularization for $c = 3$.

4.4.3. Modularization for c = 3

We are getting 18 modules for $c = 3$. Module *GRB2*, *Ras* and *ERK* divide effectively the classic

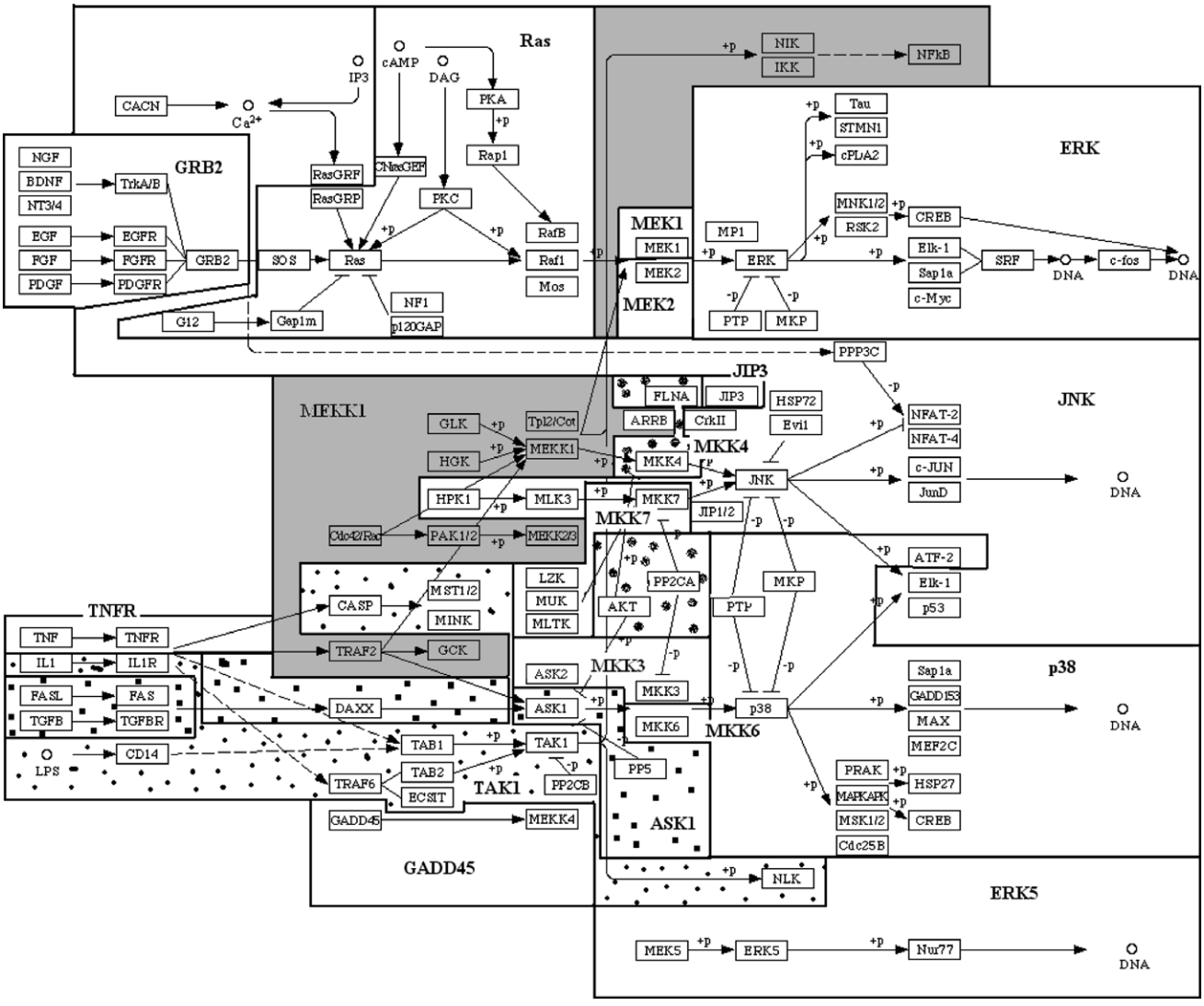

Fig. 14. Modularized MAPK signaling pathway of *H. sapiens* for $c = 3$. Human MAPK signaling pathway gets divided into 18 modules for $c = 3$. Here the problem of over splitting of the network is minimized.

MAPK signaling pathway into 3 parts. JNK pathway is divided into 4 parts namely module *MEKK1*, module *MKK4*, module *MKK7* and the module *JNK*. Modules *p38*, *ASK1* and *TAK1* counter for p38 pathway except 1/2 small modules. Here the problem of over splitting is a lot minimized with only 4 singleton modules and 2 small modules. The details are shown in Fig. 14.

4.4.4. Modularization for $c = 4, 5$

For $c = 4$, the network is neatly separated into 12 modules. Just 2 singleton modules are present. But some modules like *p53* are getting much larger and complex in size. The scenario becomes more difficult for c-value of 5 as the number of modules decreases to 10, each being larger that the previous ones. The modularized networks of human MAPK signaling pathway for $c = 4$ and $c = 5$ are given in Figs. 15 and 16, respectively. For higher values of c, the modules become even more complex.

4.5. Comparative study on modules of MAPK signaling pathways of different species

Keeping the factors, i.e., over splitting of network and complexity of a module in mind, we here provided a comparative view among MAPK signaling pathways of the taken set of 9 species for $c = 3$. MAPK signaling pathways of *D. melanogaster* and *S. cerevisiae* are very simple, and are different in layout from that of the remaining species. Modularized MAPK signaling pathway of *D. melanogaster* is given in Fig. 9. The figure clearly represents three independent modules. The MAPK signaling pathway of *S. cerevisiae* also gives 3 independent modules for $c = 4$ as shown in Fig. 10. We have compared the modules obtained from the rest 7 species for $c = 3$. Details of these modules are given in Table 6. MAPK signaling pathway of *H. sapiens* and *M. musculus* are almost identical. In MAPK signaling pathway of *M. musculus*, two nodes namely RafB and LZK are absent. So their modules have maximum resem-

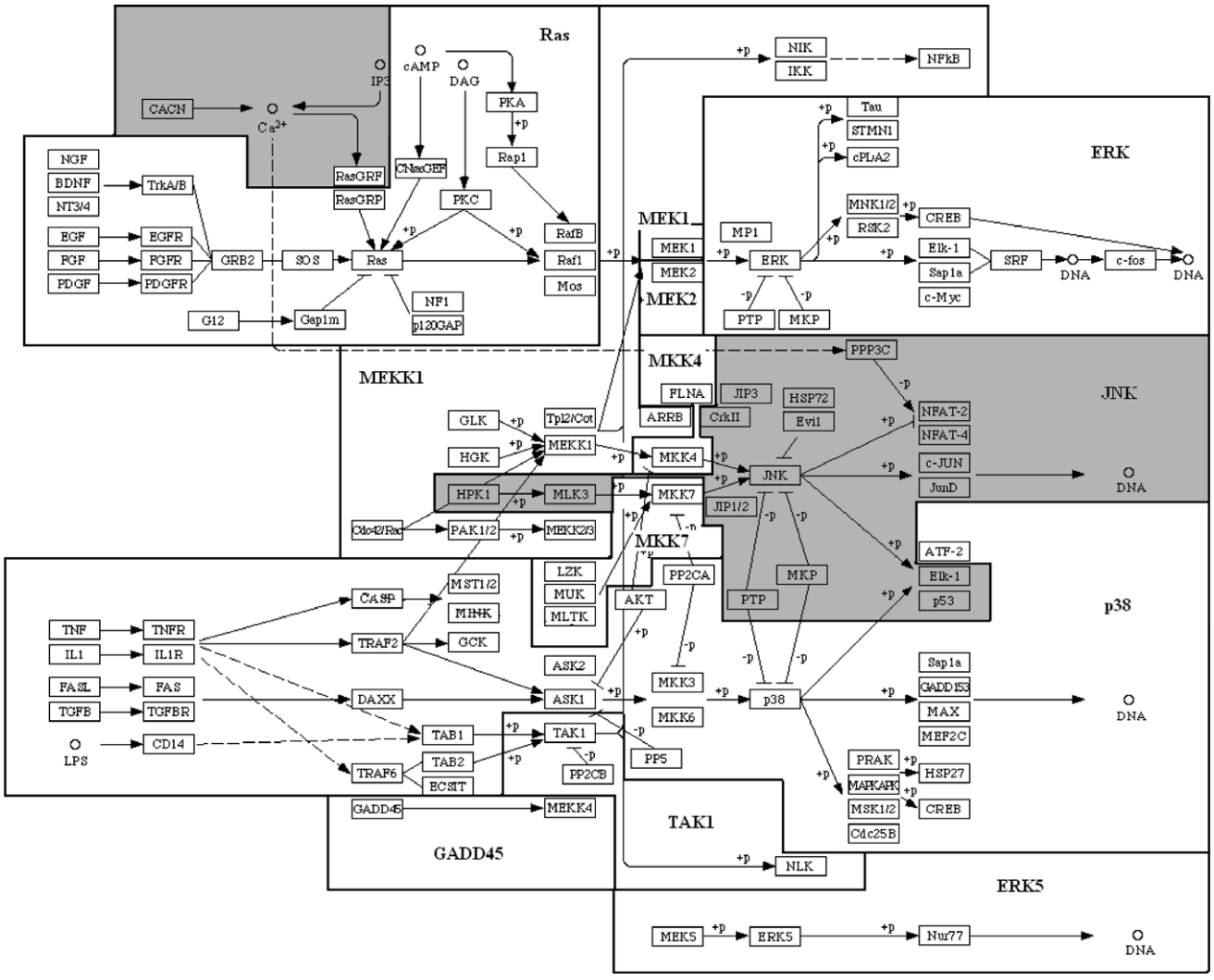

Fig. 15. Modularized MAPK signaling pathway of *H. sapiens* for $c = 4$. Modularization for $c = 4$ divides MAPK signaling pathway of *H. sapiens* into 12 modules. Most of them are eligible for consideration of biological significance.

blance with each other. Next comes the modules of *R. norvegicus*. In *B. taurus*, maximum modules found in *H. sapiens* are in elementary stage and exist in two or three separate modules. MAPK signaling pathway of dog, pig and chimpanzee are least developed. So we are finding a gradual development of the pathway from *C. familiaris* to *H. sapiens*. We have included here the figures corresponding to *H. sapiens*, *D. melanogaster* and *S. cerevisiae* only.

4.6. Changes encountered in modules with increase in c-value

With increasing values of c, the modules decrease by number. In case of a particular module, it is noticed that with each increase in c-value by 1, either new members get added to the module, or the module remains static. Here first we are considering module (*C*00076)2 of human calcium signaling pathway. This module is increasing in terms of members/nodes with increase in c-value till $c = 3$, then remains the same for $c = 4$, again increases in size for $c = 5$ and 6, and finally the whole calcium signaling pathway gets converged into it for $c = 7$. Changes in module (*C*00076)2 with increasing c-values is shown in Fig. 17.

But in some cases certain members get deleted from the module with increase in c-value as seen in the case of module *Ras* of MAPK signaling pathway (Fig. 18). Member RasGRF is inside the module till $c = 2$. But after that it is getting excluded from module *Ras*.

4.7. Comparison of results of the proposed algorithm with that of Newman's community finding algorithm [1]

We have compared the performance of the proposed algorithm with that of Newman's community finding algorithm [1]. For this purpose, we have applied the existing algorithm to our example network (Fig. 1), calcium signaling pathway (Fig. 5) and MAPK signaling pathway (Fig. 11) of *H. sapiens*.

The example network has been divided into 5 modules by Newman's algorithm based on the connectivity of the

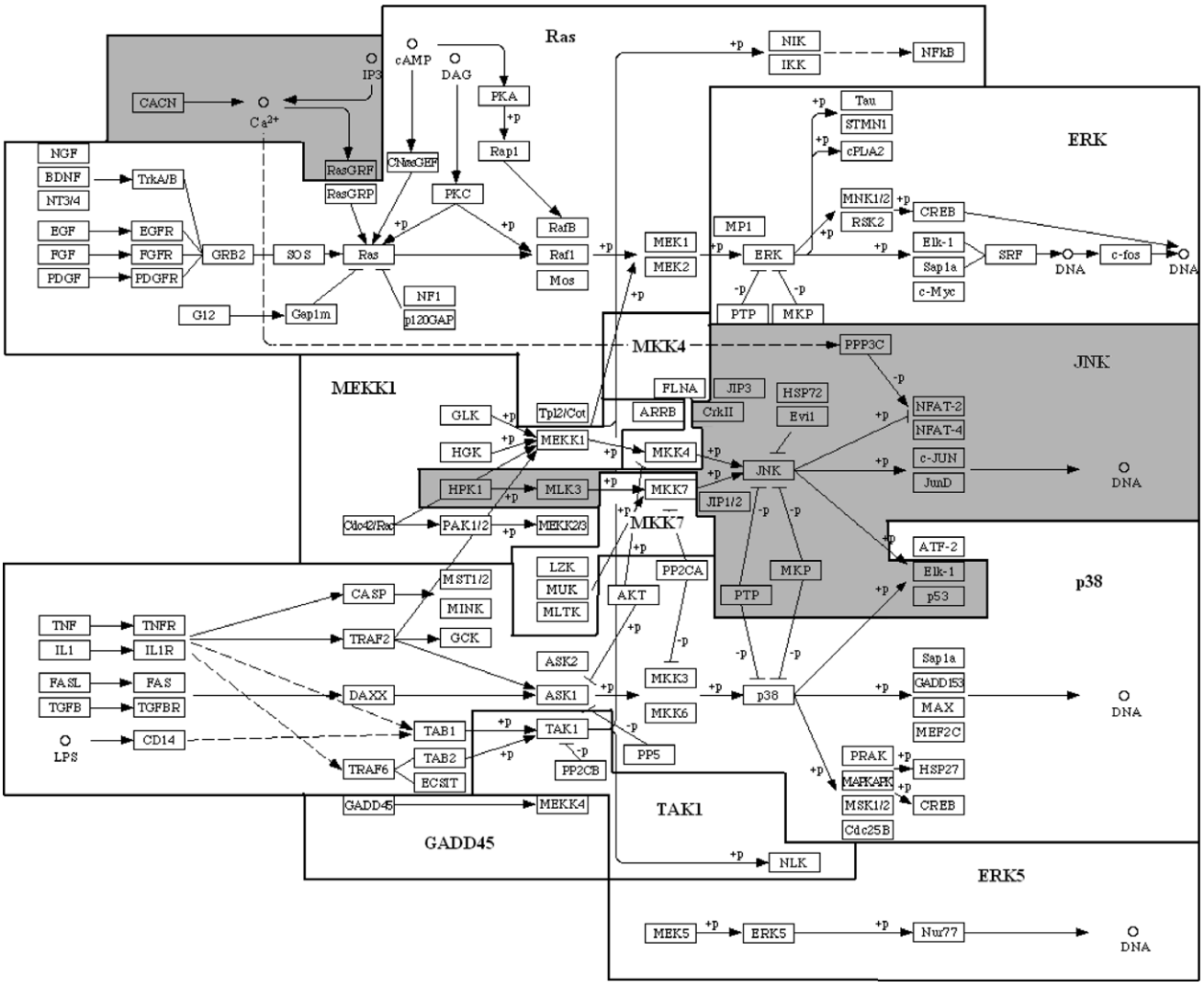

Fig. 16. Modularized MAPK signaling pathway of *H. sapiens* for $c = 5$. Here we get only 10 large complex modules that defy our basic purpose of simplifying a network.

nodes in the network. Our algorithm has created the modules one by one taking a few nodes each time, where there was no competition between two modules to include a particular node. On the other hand, Newman's algorithm has divided the network optimally without considering limitations like size and number of modules that are prevalent in graph separation techniques. Successive divisions that have been done by Newman's algorithm in the example network is given in Fig. 19(a). After first division we have found two subnetworks (subnet1 and subnet2) of 10 and 16 nodes, respectively. Subnet1 is further divided into two parts (module1 and module2). Likewise, subnet2 is further partitioned into 2 parts (ssubnet1 and ssubnet2) of 9 and 7 nodes, respectively. The procedure terminates after division of ssubnet2 into module4 (4 nodes) and module5 (3 nodes). All of these modules are shown in Fig. 19(b).

When Newman's algorithm is applied to calcium signaling pathway of *H. sapiens*, 5 modules have been created of size 15, 13, 9, 5 and 12 nodes, respectively, as shown in Fig. 20. Similarly, human MAPK signaling pathway have been divided into 10 modules (Fig. 21) of different sizes after application of Newman's community finding algorithm. One interesting fact is that sometimes the ΔQ value becomes very less but fails to reach exactly zero. In this case, we have assigned a threshold for the ΔQ value. No significant difference is found in case of the example network and human calcium signaling pathway with different threshold values. But human MAPK signaling pathway has been partitioned into different number of modules when threshold value of ΔQ varies between 0.00000001 and 0.01 (Details are given in Table 7). Considering the number of singleton modules (module that contain single node) found among the modules created from different threshold values of ΔQ, we have settled for analysis of the set of modules obtained for the threshold value of 0.01 as it minimizes the number of singleton modules to a greater extent. Details about all the modules obtained from these three networks described above are given in Table 8. The fact that whether and how much these modules are similar to modules got by our proposed algorithm

Table 6
Modules obtained from MAPK signaling pathways of 7 different species for $c = 3$

Human and Mouse		Cow		Rat		Pig		Chimpanzee		Dog	
Name	*N*	Name	*N*	Name	*N*	Name	*N*	Name	*N*	Name	*N*
JNK	21	Ca^{2+}	3	JNK	5	Ca^{2+}	4	Ca^{2+}	14		
p38	13	ERK	13	p38	17	C-jun	2				
ERK	17	c-fos	2	ERK	14						
Ras	16	G12	2	Ras	25			Ras		Ras	7
MEKK1	13	IKK	2	MEKK1	14			MEKK1	12		
TAK1	14	CD14	2	LPS	2			CASP	2		
MKK4	4	MKK4	5							MKK4	2
MKK7	4	CDC42/Rac	2	MKK7	5					CDC42/Rac	3
MEK1	1	FASL	2	MEK1	1						
MEK2	1					FASL	3				
ASK1	7	TGFB	2	TGFBR	5	TGFB	2	TGFB	2		
TNFR	2	TNFR	3			TNFR	3				
GRB2	11	EGF	2	GRB2	11	EGF	2	EGF	2		
JIP3	1	FGF	2								
MKK3	2	TrkA/B	2								
MKK6	1										
ERK5	4	Nur77	2	ERK5	4					Nur77	2
GADD45	2										

This table contains information about the modules obtained from MAPK signaling pathways of *H. sapiens* (human), *R. norvegicus* (rat), *M. musculus* (mouse), *B. taurus* (cow), *S. scrofa* (pig), *C. familiaris* (dog) and *P. troglodytes* (chimpanzee). Column *N* is giving number of nodes present in the modules.

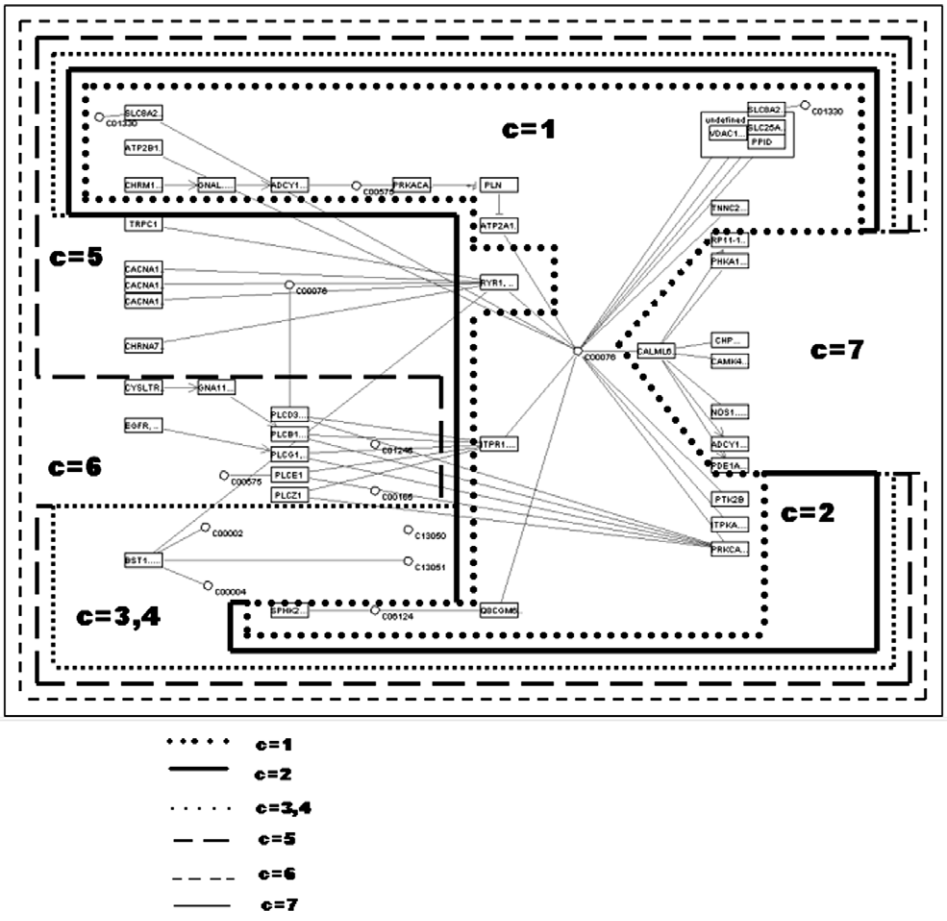

Fig. 17. Module *(C00076)2* of human MAPK signaling pathway for different *c*-values. This diagram shows change in the module *(C00076)2* for *c*-values ranging from 1 to 7.

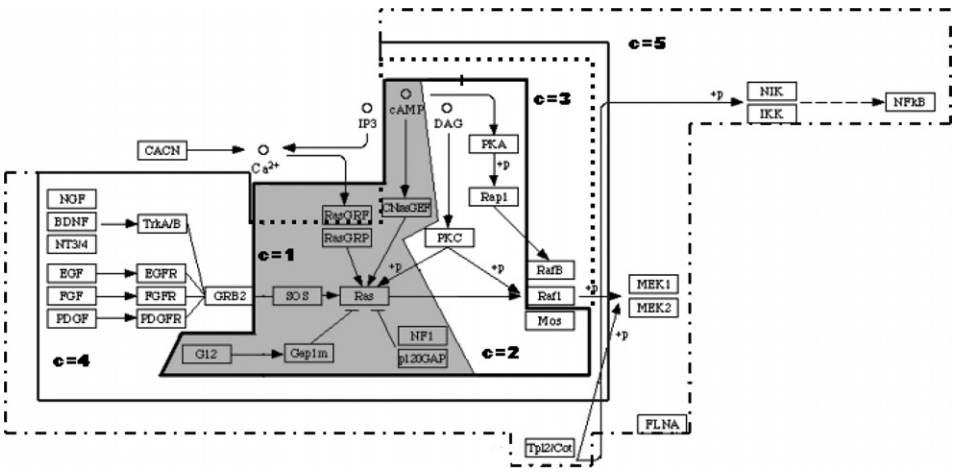

Fig. 18. Module *Ras* of human MAPK signaling pathway for different values of *c*. The diagram shows change in module *Ras* for different *c*-values.

or whether they are important from biological point of view, is an important point of discussion.

The proposed method has divided human calcium signaling pathway into four biologically meaningful modules. It also enables us to get some meaningful modules from the much complex MAPK signaling pathway of *H. sapiens*.

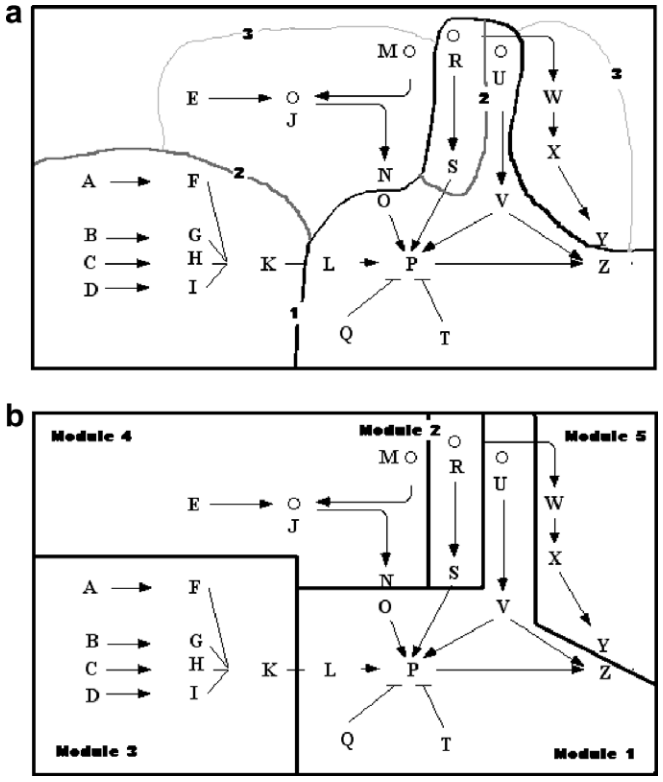

Fig. 19. Example network modularization using Newman's algorithm [1]. (a) Shows successive divisions of the network with lines of decreasing gray value made by Newman's algorithm [1] in the example network. The lines are numbered accordingly to remove any confusion of the reader. Number 1 denotes division one creating two subnetworks, 2 denotes further division of subnetwork1 and 2, and finally number 3 denotes successive divisions of subsubnetwork2. Finally the network gets divided into five modules as shown in (b).

Module1 (of calcium signaling pathway) created by Newman's algorithm contains reactions of calmodulin-like proteins along with IP3, DAG, cADPR, S1P, NCX and Na^{+}. Calmodulin-like proteins help in Ca^{2+} ion balance temporarily in cellular environment and in turn, Ca^{2+} ions effect reactions of proteins from PLC family. Here one event may be indirectly effected by the other, but there is no direct relation between their behavior. With our method, we get module *CALM6* that explains only the behavior of calmodulin like proteins. Module2 and module4, obtained by Newman's algorithm, explain flow of Ca^{2+} ions between plasma membrane, endoplasmic reticulum and mitochondria partly, as some nodes related to this function still lie in module3 and module5. This phenomenon has been fully explained by module *(C00076)2* obtained by our algorithm. Module3 is an unexplainable combination of CACNA (calcium channel, voltage-dependent, L type) proteins and BST1 (bone marrow stromal cell antigen 1) proteins that our algorithm has divided separately in two modules (module *(C00076)2* and *(C00076)1*). Ideally, module4 should be included in module2. Module5 is roughly equivalent to module *C01245* when we do not consider 2/3 nodes that are different between these two modules. Thus, most of the modules obtained by Newman's method are not clearly showing a particular function of the network or behavior of a family of proteins.

The case becomes more complex when we try to analyze the MAPK signaling pathway of *H. sapiens* by comparing modules obtained by application of both these algorithms. As mentioned earlier, MAPK signaling pathway as such is a combination of three specialized pathways, i.e., classic MAPK, JNK and p38. By applying our algorithm, we get all total 18 modules that have separated these pathways optimally without mixing up parts of the specialized pathways in a module for *c*-value of 3. With Newman's method, we have tried to analyze the 10 modules obtained for

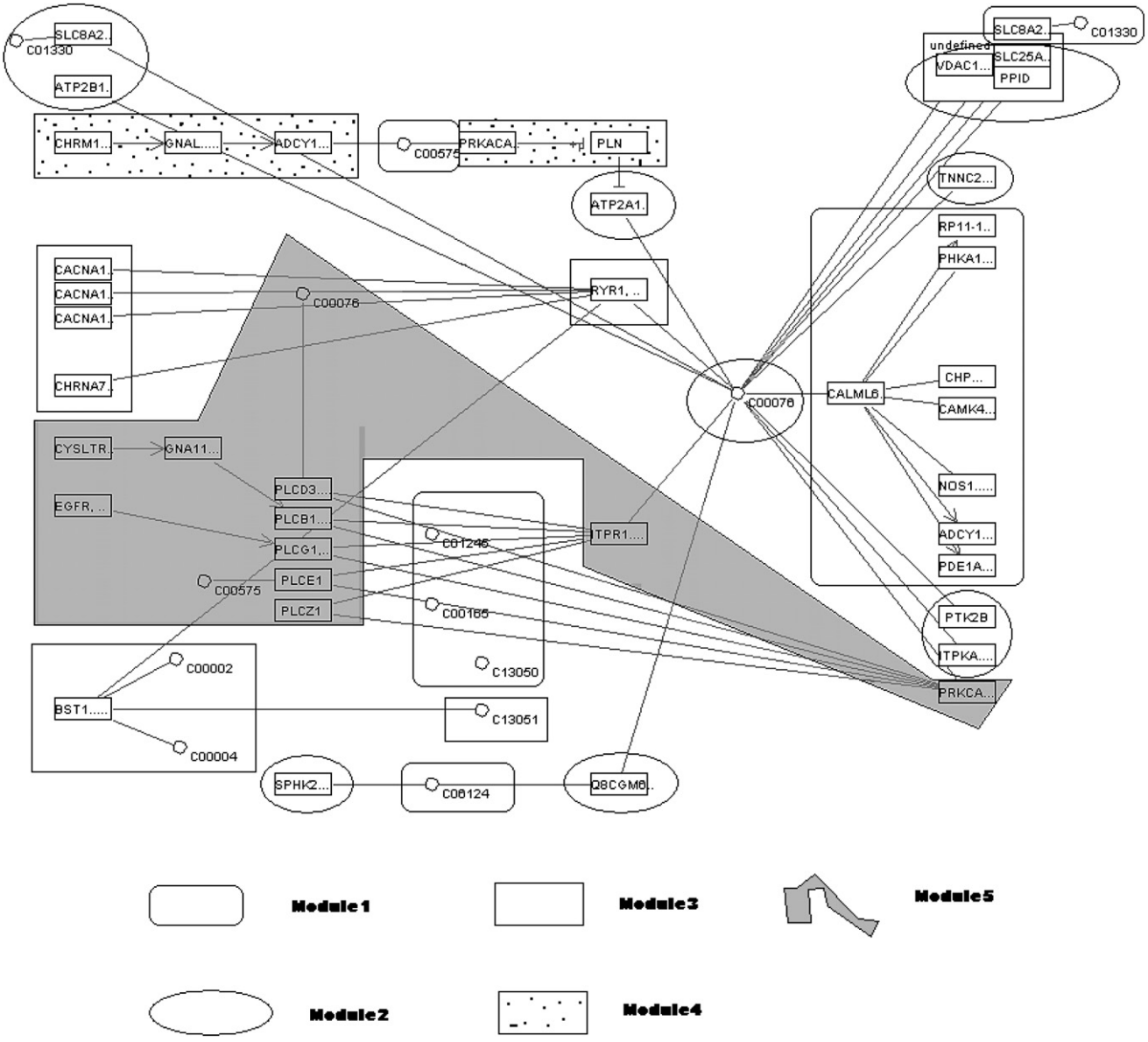

Fig. 20. Human calcium signaling pathway after modularization with Newman's algorithm [1]. Nodes present in different modules are marked differently.

threshold value of 0.01 for ΔQ. Out of them module2, 5, 7 and 8 are singleton modules. It is difficult to assign a function to such modules found in a network where more than one nodes bring forward a function in most of the cases. Module3 is similar to module *Ras*, except the fact that it has some far placed nodes that cannot be explained from biological point of view. The rest modules have nodes that are responsible for more than one functions. Moreover, all the nodes responsible for a single function are not present in a single module, making them unexplainable for a particular biological function.

One interesting fact is that while our algorithm tries to create the modules by centralizing the mostly connected nodes in a network, Newman's algorithm is depicting them as singleton modules, which is not acceptable from biological point of view. Thus the proposed algorithm, unlike Newman's algorithm [1], is able to create biologically significant modules for the aforesaid signal transduction pathways. Moreover, we can think a signal transduction pathway as a black box operating with many layers, where we only know, through laboratory experiments, the input and output of each layer. But what exactly happens inside the black box and the way these layers co-ordinate is difficult to grasp. Probably the task will be easier if we try to understand the mechanism of the black box layer by layer and try to trace a particular input through various layers of the black box till we reach the output. Here our created modules are equivalent to layers of the black box. Now it may happen that a particular input may or may not be involved with all the nodes of intermediate layers of the black box. Likewise, in a signal transduction pathway, the input signal may not involve all the nodes present in

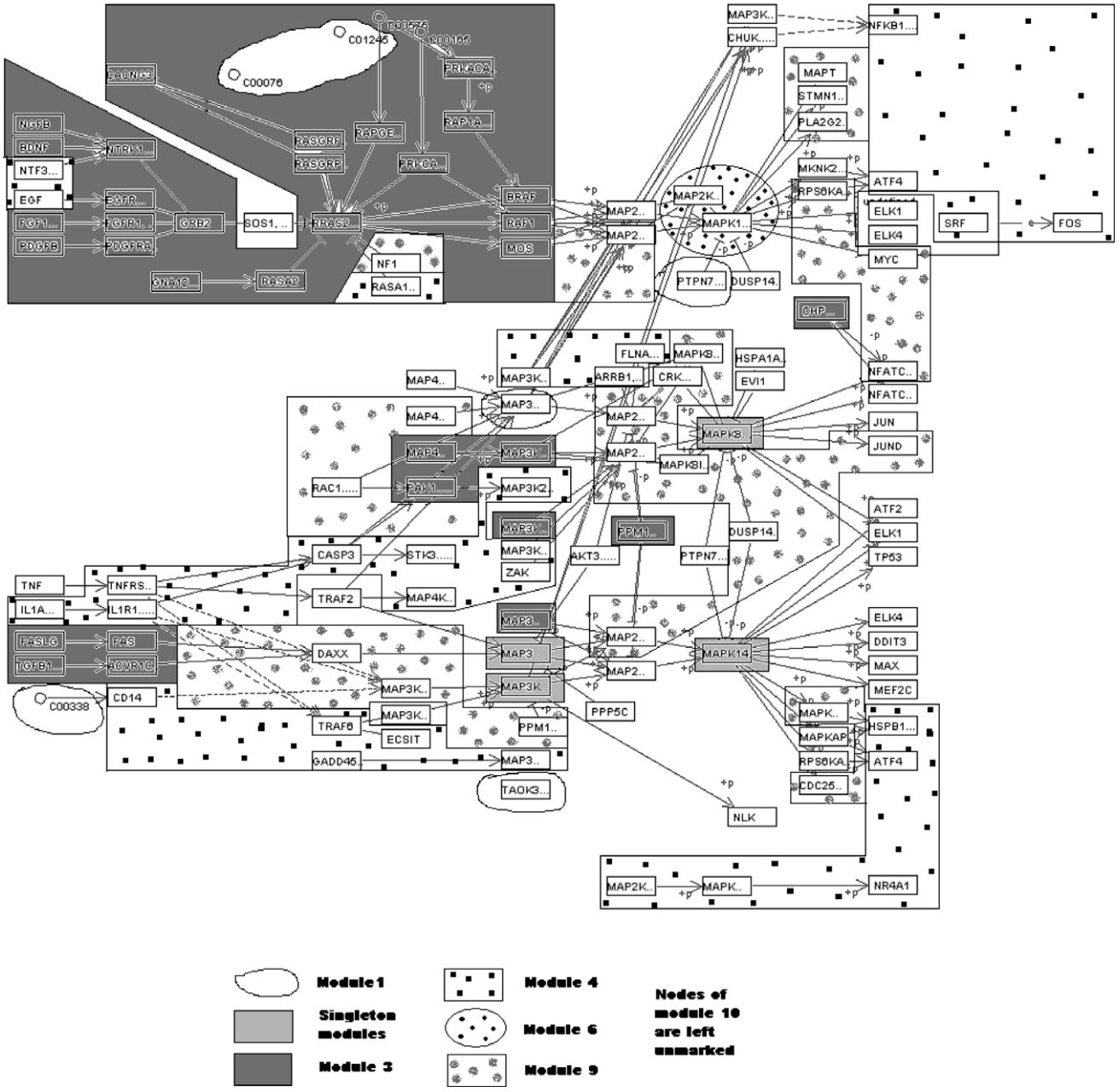

Fig. 21. Human MAPK signaling pathway after modularization with Newman's algorithm [1] for threshold value of 0.01 for ΔQ. Nodes present in different modules are marked differently. Singleton modules (module2, 5, 7 and 8) are represented identically with light gray rectangles. The unmarked nodes belong to module10.

Table 7
Modules obtained from human MAPK signaling pathways by applying Newman's algorithm [1] with varying threshold value of ΔQ

Threshold value for ΔQ	Number of modules	Description
0.00000001–0.00001	86	75 singleton modules, 5 modules each with 2 nodes, 6 modules with 2, 3, 5, 6, 8, 9 and 15 nodes, respectively
0.0001	58	30 singleton modules, 13 modules each with 2 nodes, 8 modules each with 3 nodes, 2 modules each with 4 nodes, 5 modules with 5, 6, 8, 9 and 15 nodes, respectively
0.001	28	14 singleton modules, 2 modules each with 2 nodes, 3 modules each with 8 nodes, 4 modules each with 9 nodes, 2 modules each with 15 nodes, 3 modules with 5, 6 and 12 nodes, respectively
0.01	10	4 singleton modules, 2 modules each with 30 nodes, 4 modules with 2, 6, 23 and 36 nodes, respectively
0.1	2	2 modules with 53 and 78 nodes, respectively

With different values of threshold for ΔQ, different number of modules of varying size are obtained from human MAPK signaling pathway.

Table 8
Modules obtained from different networks by Newman's algorithm [1]

Module name	Module size		
	Example network (Fig. 1)	Human calcium signaling pathway (Fig. 5)	Human MAPK signaling pathway (Fig. 11)
Module1	8	15	6
Module2	2	13	1
Module3	9	9	36
Module4	4	5	30
Module5	3	12	1
Module6			2
Module7			1
Module8			1
Module9			23
Module10			30

The column Module size gives the number of nodes present in each module found in the individual networks.

all the modules. These ideas may lead to a better design of an artificial system that can successfully mimic biological pathways.

5. Conclusions

In this paper we have developed an algorithm for modularizing signal transduction pathways. The algorithm has been applied to calcium and MAPK signaling pathways of various species for comparing the levels of development of these pathways in these species. We have successfully conferred biological significance to the modules obtained from human calcium signaling pathway for complexity level (c-value) of 3. The comparative study indicates gradual increase in development of calcium signaling pathways starting from *P. troglodytes* to *H. sapiens* via *C. familiaris*, *S. scrofa*, *B. taurus*, *M. musculus* and *R. norvegicus*. We have also got some significant modules from MAPK signaling pathway of *H. sapiens* for c-value of 3. The comparative study of MAPK signaling pathways among the taken 7 species shows gradual development of the pathway from *P. troglodytes* to *H. sapiens* via *S. scrofa*, *P. troglodytes*, *B. taurus* and *R. norvegicus*.

When a pathway is difficult to analyze as a whole in certain species, or certain module(s) of it is(are) only functional for a set of species, modularized study is quite helpful. For example, in module *(C00076)2* of human calcium signaling pathway is consistent among our taken set of species in varying size (Table 3), while other modules are not. So one can ignore the other modules and compare module *(C00076)2* of human calcium signaling pathway with that of other species instead of comparing the whole pathway to get a comparative view among them. Given a very large network, these kinds of inferences may save time and cost of wet lab experiments by avoiding less important verifications.

The superior performance, in terms of biological significance, of the proposed algorithm over an existing community finding algorithm of Newman [1] has been analyzed in details for these two pathways. As an extension of our work we are considering to analyze other signaling pathways for different species, including those obtained by combination of various pathways e.g., networks responsible for certain kind of cancer by the algorithm developed here. Moreover, finding an optimal value of c automatically forms a part of further investigation.

References

[1] Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA 2006;103(23):8577–82.
[2] Wilkinson MG, Millar JBA. Control of the eukaryotic cell cycle by map kinase signaling pathways. FASEB J 2000;14(14):2147–57.
[3] Schaeffer HJ, Webber MJ. Mitogen-activated protein kinases: specific messages from ubiquitous messengers. Mol Cell Biol 1999;19(4):2435–44.
[4] Seger R, Krebs EG. The mapk signaling cascade. FASEB J 1995;9(9):726–35.
[5] Clapham DE. Calcium signaling. Cell 1995;80:259–68.
[6] Tsien RW, Tsien RY. Calcium channels, stores and oscillations. Ann Rev Cell Biol 1990;6:715–60.
[7] Means AR. Calcium, calmodulin and cell cycle regulation. FEBS Lett 1994;347:1–4.
[8] Nicotera P, Zhivotovsky B, Orrenius S. Nuclear calcium transport and the role of calcium in apoptosis. Cell Calcium 1994;16:279–88.
[9] Berridge MJ. Elementary and global aspects of calcium signaling. J Physiol 1997;499:291–306.
[10] Papin JA, Reed JL, Palsson BO. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trend Biochem Sci 2004;29(12):641–7.
[11] Eungdamrong NJ, Iyengar R. Modeling cell signaling networks. Biol Cell 2004;96(5):355–62.
[12] Neves RN, Iyengar R. Modeling of signaling networks. BioEssays 2002;24(12):1110–7.
[13] Saez-rodriguez J, Kremling A, Conzelmann H, Bettenbrock K, Gilles ED. Modular analysis of signal transduction networks. IEEE Contr Syst Mag 2004;24(4):35–52.
[14] Costa L da F. The hierarchical backbone of complex networks. Phys Rev Lett 2004;93:098702.
[15] Karypis G, Han EH, Kumar V. Multilevel refinement for hierarchical clustering, Tech. Rep. TR-99-020, Department of Computer Science, University of Minnesota, Minneapolis, 1999.
[16] Ravasz E, Barabasi A. Hierarchical organization in complex networks. Phys Rev E 2003;67(2):026112.
[17] Elsner U. Graph partitioning: a survey, Tech. Rep. 97-27, Technische Universitat Chemnitz, Chemnitz, Germany, 1997.
[18] Fjallstrom PO. Algorithms for graph partitioning: a survey, Tech. Rep. 98-010, Linkoping University, Sweden, Electronic Articles in Computer and Information Science, 1998.
[19] Larue WW, Komp E, Schaffer S, Frost VS, Shanmugan KS, Reznik D. A block oriented paradigm for modeling communication networks. In: Conference Record, 'A New Era', IEEE, 1990. p. 689–95.
[20] Ederer M, Sauter T, Bullinger E, Gilles ED, Allgwer F. An approach for dividing models of biological reaction networks into functional units. Simulation 2003;79(12):703–16.
[21] Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. Nature 2005;433(7028):895–900.
[22] Dutt S. New faster kernighan-lin-type graph-partitioning algorithms, In: Proceedings of international conference on computer-aided design, 1993, p. 370–377.
[23] Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 2006;74:036104.
[24] Newman MEJ. Detecting community structure in networks. Eur Phys J B 2004;38:321–30.

[25] Augeri CJ, Ali HH. New graph-based algorithms for partitioning vlsi circuits. In: Proceedings of circuits and systems (ISCAS '04), 2004. vol. 4. IV-521– IV-524.
[26] Chen CH. Graph partitioning for concurrent test scheduling in vlsi circuit. In: Proceedings of 28th ACM/IEEE design automation conference, 1991, vol. 18.4. p. 287–90.
[27] Cordella LP, Foggia P, Sansone C, Vento M. Fast graph matching for detecting cad image components. In: Proceedings of 15th international conference on pattern recognition; 2000. vol. 2. p. 1034–7.
[28] Pothen A. Parallel numerical algorithms, chapter Graph partitioning algorithms with applications to scientific computing, Kluwer Academic Press; 1997. p. 323–68.
[29] Berry MW, Henderickson B, Raghavan P. The mathematics of numerical analysis, vol. 32 of Lectures in Applied Mathematics, chapter Sparse matrix reordering schemes for browsing hypertext, American Mathematical Society; 1996. p. 99–123.
[30] Karypis G, Kumar V. Parallel multilevel graph partitioning. In: Proceedings of the 10th international parallel processing symposium (IPPS '96), 1996. p. 314–9.
[31] Kedem G, Watanabe H. Graph optimization techniques for ic layout and compaction. In: Proceedings of the 20th conference on design automation, 1983. p. 113–20.
[32] Hendrickson B, Leland R. The chaco user's guide version 2, Tech. Rep. SAND94-2692, Sandia National Laboratories, Albuquerque, NM, 1994.
[33] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci USA 2002;99(12):7821–6.
[34] Fortunato S, Latora V, Marchiori M. A method to find community structures based on information centrality. Phys Rev E 2004;70:056104.
[35] Imafuji N, Kitsuregawa M. Effects of maximum flow algorithm on identifying web community. In: Proceedings of the 4th international workshop on Web information and data management (WIDM '02), 2002. p. 43–48.
[36] Yang S. Exploring complex networks by walking on them. Phys Rev E 2005;71:016107.
[37] Latapy M, Pons P. Computing communities in large networks using random walks. In: Proceedings of the 20th international symposium on computer and information sciences (ISCIS'05), 2005, 3733. p. 284–93.
[38] Newman MEJ. Fast algorithm for detecting community structure in networks. Phys Rev E 2004;69:066133.
[39] Holme P, Huss M, Jeong H. Subnetwork hierarchies of biochemical pathways. Bioinformatics 2003;19(4):532–8.
[40] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature 2000;407(6804):545–658.
[41] Hu H, Yan X, Huang Y, Han J, Zhou XJ. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 2005;21(Suppl. 1):i213–21.
[42] Patra SM, Vishveshwara S. Backbone cluster identification in proteins by a graph theoretical method. Biophys Chem 2000;84(1):13–25.
[43] Schuster S, Pfeiffer T, Moldenhauer F, Koch I, Dandeke T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumonia. Bioinformatics 2002;18(2):351–61.
[44] Wagner A, Fell DA. Small world inside large metabolic networks. In: Proceedings of the Royal Society B, 2001, 268(1478). p. 1803–10.
[45] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004;69:026113.
[46] Wilkinson DM, Huberman BA. A method for finding communities of related genes. Proc Natl Acad Sci USA 2004;101(suppl. 1):5241–8.
[47] Newman MEJ. The structure and function of complex networks. SIAM Rev 2003;45(2):167–256.
[48] Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science 2002;296(5569):910–3.
[49] Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. Nature 2002;420(6912):190–3.
[50] Guelzim N, Bottani S, Bourgine P, Kepes F. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 2002;31(1):60–3.
[51] Raine DJ, Norris V. Network structure of metabolic pathways. J Biol Phys Chem 2001;1(2):89–94.
[52] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science 2002;298(5594):824–7.
[53] Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of escherichia coli. Nat Genet 2002;31(1):64–8.
[54] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999;27:29–34.
[55] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[56] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in kegg. Nucleic Acids Res 2006;34:D354–7.