Community structure of non-coding RNA interaction network

Jose C. Nacher*

Department of Information Science, Faculty of Science, Toho University, Funabashi, Chiba, 274-8510, Japan

Summary

Rapid technological advances have shown that the ratio of non-protein coding genes rises to 98.5% in humans, suggesting that current knowledge on genetic information processing might be largely incomplete. It implies that protein-coding sequences only represent a small fraction of cellular transcriptional information. Here, we examine the community structure of the network defined by functional interactions between non-coding RNAs (ncRNAs) and proteins related bio-macrolecules (PRMs) using a two-fold approach: modularity in bipartite network and *k*-clique community detection. First, the high modularity scores as well as the distribution of community sizes showing a scaling-law revealed manifestly non-random features. Second, the *k*-clique sub-graphs and overlaps show that the identified communities of the ncRNA molecules of *H. sapiens* can potentially be associated with certain functions. These findings highlight the complex modular structure of ncRNA interactions and its possible regulatory roles in the cell.

1 Introduction

In the twenty-first century, network analysis has rapidly emerged as a promising set of techniques and tools to address the complexity of ubiquitous systems composed by multiple interacting units, from individual cells and biological organisms to large-scale human societies [1]. Cells contain thousands molecules like proteins, metabolites, genes, mRNAs and small RNAs among others that interact with each other by means of pair-wise interactions. As a consequence, complex pathways, communities and networks emerge where biological information is processed and regulated. Although recent advances have allowed us to decipher genome sequences for many organisms, how cells regulate gene expression programs still represents a huge scientific challenge [2]. Transcriptional regulation typically involves a DNA-binding protein (transcription factor) that binds to specific target genes in the genome. A transcriptional regulatory network can be constructed based on these regulatory interactions. In this system, we have two kinds of molecular entities, transcriptional factors and genes, therefore the resulting network can be approximated as a bipartite graph, where transcriptional factors regulate target genes.

It is well-established that genetic information flows from DNA to proteins by means of mRNAs molecules. As a consequence, it is possible to consider a one-to-one correspondence between genes and proteins. While this affirmation seems to be accurate for simple prokaryotes, recent studies have observed that the proportion of protein-coding genes decreases as a function of developmental complexity. In particular, the ratio of non-protein coding genes rises to 98.5% in humans. It implies that protein-coding sequences only represent a small fraction of cellular transcriptional information. The finding that the transcription of non-coding RNA (ncRNA) in higher organism is so abundant raises the

* To whom correspondence should be addressed. Email: nacher@is.sci.toho-u.ac.jp

doi:10.2390/biecoll-jib-2013-217

question on its cellular functionality and suggests that current knowledge on genetic information processing might be largely incomplete [3-5]. If the latter hypothesis was correct, it could imply that RNA molecules were able to evolve and to adapt to transcriptional programs of higher eukaryotes. NcRNA represents a functional RNA molecule that is not translated into a protein. Non-coding RNA comprises introns in protein-coding genes as well as other transcripts that do not seem to encode proteins. Some classifications include transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as snoRNAs, microRNAs, siRNAs and piRNAs as non-coding RNA genes [6].

On the other hand, the discovery of hierarchical signature and modularity in biological networks, from metabolic to protein-protein interaction networks, has led to devise novel and reliable methods based on graph theory for large-scale community detection. Most real-world networks are very heterogeneous and are made of modules and communities consisting of many links within modules and a few links between different modules. First glimpses of hierarchical organization were also observed, revealing nested structures consisting of clusters and modules within modules [1]. Uncovering modules does not only help to understand the structure of the network and dynamic behaviour of the system, but also to uncover functional similarity among nodes. In general, several molecules belong to the same module and together carry out a specific cellular function.

Here, we examine the community structure of the biological network defined by functional interactions between non-coding RNAs (ncRNAs) and proteins related bio-macrolecules (PRMs). Using datasets from NPInter database [6], we collected and retrieved data corresponding to six model organisms, such as *Escherichia coli, Saccharamyces cerevisiae, Caenorhadbitis elegans, Drosophila melanogaster, Mus musculus* and *Homo sapiens.* We define these networks as bipartite graphs, where each link represents a functional interaction between a ncRNA and a protein-related molecule. We then compute the modularity of these bipartite networks using a stochastic optimization algorithm. The computation of the modularity is performed in the bipartite graph. Moreover, for the human ncRNA-protein interaction networks, a *k*-clique community analysis is conducted in the corresponding projected network, which shows specific ncRNA molecules that overlap among communities as well as unique communities with potential common functionality for their memberships. Our findings show that ncRNA- protein interactions have a pervasive community structure with high modularity, far from the random expectation. Community organization of transcriptional regulatory networks corresponding to *E. coli* and *S. cerevisiae* is also computed and compared to the ones calculated using ncRNAs. This comparison highlights the similar complex community architecture of ncRNA mediated interactions with transcriptional networks and that ncRNA interactions display a complex enough organization to address functional and regulatory roles.

2 Methods

2.1 Datasets

The transcriptional regulatory interactions defined by transcriptional factors (TFs) that regulate target genes for *E.coli* were collected from RegulonDB database. The dataset corresponding to the *S. cerevisiae* organism was downloaded from the Uri Alon website, which was originally available in [7]. The pair-wise interactions for the ncNRA-PRMs network were obtained from NPInter database [6]. Datasets from six model organisms, such as *Escherichia coli, Saccharamyces cerevisiae, Caenorhadbitis elegans, Drosophila melanogaster, Mus musculus* and *Homo sapiens* were downloaded and six bipartite graphs were constructed. To increase the statistics of the data analysis, we constructed one more

network that contained all the available interactions among organisms. Hereafter we refer to this network as *all* ncRNa and protein. A total number of 700 collected interactions were constructed between 98 ncRNA molecules and 425 proteins. Although most of PRMs are proteins, there are some interactions that involve protein coding related molecules as mRNAs and genomic DNAs. On the other hand, the collected dataset for human ncRNA-mediated interactions consisted of 34 ncRNAs and 154 protein molecules (see Fig. 1).

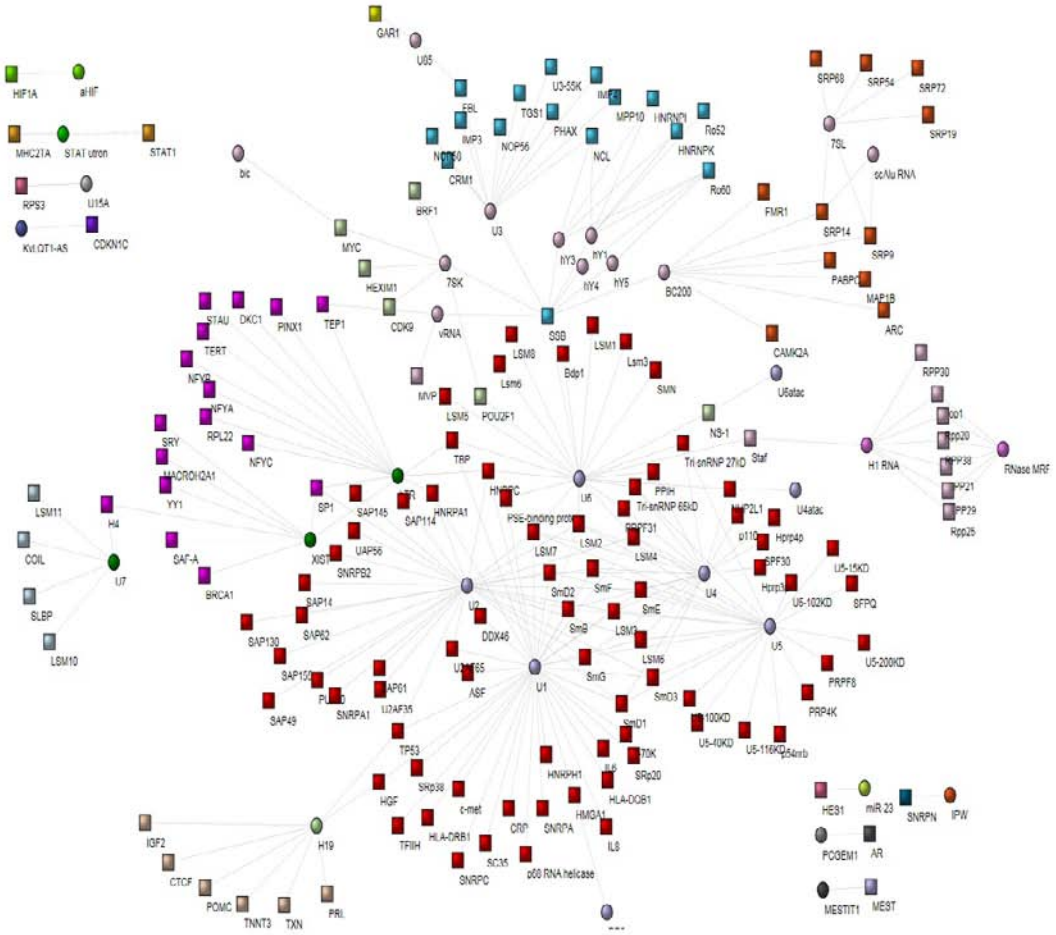

Figure 1: Identified communities for the human bipartite ncRNA-protein network are indicated by color codes. ncRNA (circles) are connected to protein (squares), where each link represents a functional interaction between a ncRNA and a protein according to the NPInter database [6].

2.2 Community structure computation in bipartite network

Each bipartite network composed of two types of nodes can be projected (i.e. transformed) onto two networks, called projections of the original bipartite network. Each projected network is then composed of only one type of nodes. A bipartite graph for TFs/ncRNA and target genes/proteins can be formally defined as $G = (T, P, E)$, where T is a set of TFs/ncRNA molecules, and P a set of target genes/protein and E a set of edges that links two nodes from T and P. $G_T = (T, E_T)$ represents the T-projection of the graph G in which nodes of T are linked together if they have at least one neighbour (P) in common in the graph G. The P-projection G_P is defined dually.

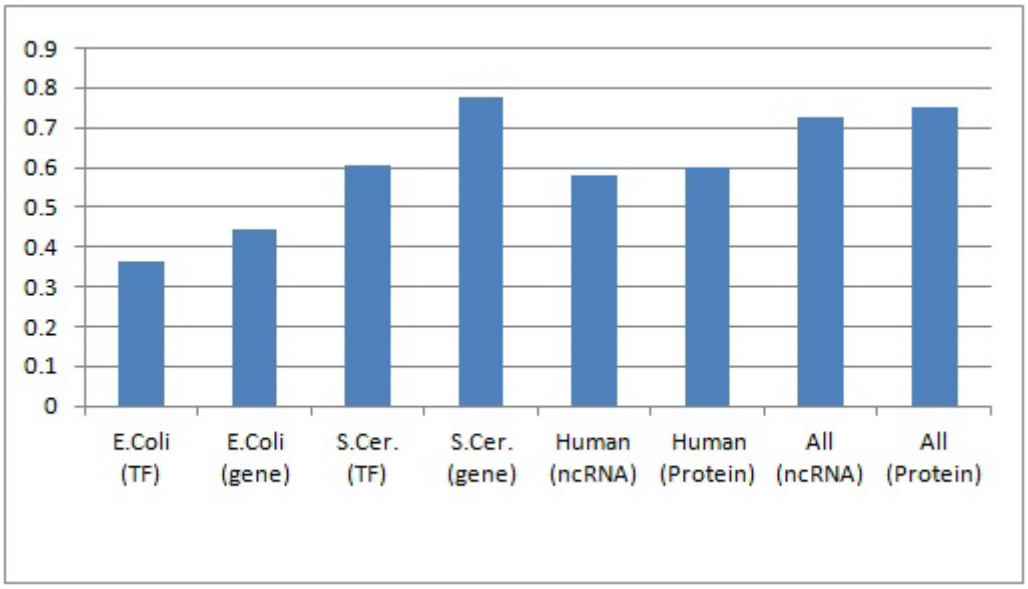

Figure 2: Community modularity values M_B for the analyzed bipartite networks computed using the objective modularity function shown in Eq. 1.

The existing algorithms that identify community structure in networks share a common strategy that is based on the maximization of a modularity function [8]. An objective function that describes modularity is usually based on the concept that the density of edges in the network is highly heterogeneous. Because the projection of bipartite graph significantly increases the number of edges and may not lead to accurate results in the boundary nodes, we compute the modularity directly in the bipartite network itself. This leads to avoid a loss of structural information in the projection process. As in [9], let us define nodes of type T and nodes of type P and consider a modularity functional form as follows:

$$M_B = \sum_{s=1}^{N_M} \left(\frac{\sum_{Pi \neq Pj \in s} c_{PiPj}}{\sum_T m_T (m_T - 1)} - \frac{\sum_{Pi \neq Pj \in s} t_{Pi} t_{Pj}}{\left(\sum_T m_T \right)^2} \right) \qquad (1)$$

where t_{Pi} indicates the total number of TFs(ncRNAs) molecules a target gene/protein P_i interacts with, m_T indicates the number of target genes(proteins) linked to T TFs(ncRNAs), and c_{PiPj} indicates the number of TFs(ncRNAs) that are simultaneously interacting with target genes(proteins) P_i and P_j. Here N_M indicates the total number of modules and s is the module index as in shown in equation 1. This function, after the optimization via simulated annealing algorithm, gives the final modularity score M_B of the bipartite network (see Fig. 2). Note that this objective function is applied to the bipartite networks composed of TF-genes and ncRNA-protein, independently.

It is known that projections may lead to different results since, as mentioned above, a fraction of the information rooted in the bipartite structure may disappear after projection [9]. The computation of modularity in the bipartite network is expected to be more accurate than in the projections. Furthermore, besides the highest accuracy of this algorithm for networks of a few thousands nodes, the method is able to identify not only an optimal partition of the nodes into modules, but also the number of modules N_M and their sizes [10]. Therefore, this algorithm

was selected to investigate the modularity of the transcriptional and ncRNA mediated interactions.

2.3 *k*-clique community computation

A complementary community definition is based on the observation that a specific molecule in a community can be connected to many other molecules. However, it may not be necessarily connected to all other nodes in the community. This is the main difference with the community modularity described above, where highly dense communities tend to have high modularity. That is, a community can be seen as a composition of smaller complete (fully connected) sub-graphs that share nodes. These complete subgraphs are called *k-cliques*, where *k* refers to the number of nodes in the subgraph. Therefore, a *k-clique-community* is defined as the composition of all *k*- cliques that can be reached from each other through a series of *adjacent k-cliques*, where two *k*-cliques are said to be adjacent if they share *k-1* nodes [11].

The method is described in detail in [11], therefore we only summarize the process as follows. A symmetric matrix is created where each row (and column) represents a clique and the matrix elements are equal to the number of common nodes between the corresponding two cliques. The diagonal entries are equal to the size of the clique. Next, the *k*-clique-communities for a given value of *k* are equivalent to such connected clique components in which the neighboring cliques are linked to each other by at least *k-1* common nodes. These components can be obtained by deleting every off-diagonal entry smaller than *k -1*. In addition, every diagonal element smaller than *k* in the matrix will be erased, replacing the remaining elements by one, and then carrying out a component analysis of this matrix. The resulting separate components are equivalent to the different *k*-clique-communities. Next, overlaps between communities (nodes and edges that are shared by more than one community) are also identified. The algorithm is referred as the clique percolation method.

3 Results

3.1 Scaling-law of size of communities

Our findings show that ncRNA- protein interactions have a pervasive community structure with high modularity. Modularity levels of transcriptional regulatory networks corresponding to *E. coli* and *S. cerevisiae* are also computed and compared to the ones calculated using ncRNAs (see Fig. 2). The modularity for the transcriptional factors (target genes) for *S. cerevisiae* and *E. Coli* are 0.604 (0.777) and 0.362 (0.445), respectively. Modularity detection in *H. sapiens* leads to a modularity score of 0.578 (0.598) for ncRNA (proteins), respectively. A computation of the modularity using all the available interactions in all organisms gives a value of 0.720 (0.752) for ncRNA (protein) molecules. Good modularity values typically lie between 0.3-0.7, while higher values are rare. Very low modularity (<0.3) indicates a lack of complex modular structure [9].

Interestingly, the distribution of module sizes shows a heavy tail highlighting a hierarchical organized system formed by interconnected modules of heterogeneous scales (Fig. 3). The figure shows that the probability of finding communities of a given size decays slowly as a power-law with exponent close to 2, for both transcriptional and ncRNA-interactions. The specific degree exponents of each fit are indicated in the caption of Fig. 3. Cumulative distributions are computed using the method described in [12, 13] that gives the degree exponent as well as the lower bound on power-law behaviour $x_{min.}$ In particular, we see that for protein Fig. 3 (b) and regulated genes Figs. 3 (d, f), the power-law behaviour is clearly

observed for several decades. In contrast, this behaviour is less clear in the tail of Figs. 3 (a, e). However, both figures (a, e) display a similar pattern, for ncRNA and TFs, highlighting the similarities between both systems in modularity structure. These findings, in particular, the scaling-law for community sizes, are intriguing and show that ncRNA-mediated interactions have a complex organization, which immediately raises the question on ncRNA modules and communities functionality. This result supports growing evidences suggesting that ncRNAs could be associated with regulatory functions [3-5]. We have also examined the composition of each of the modules and assign biological functions based on database annotation.

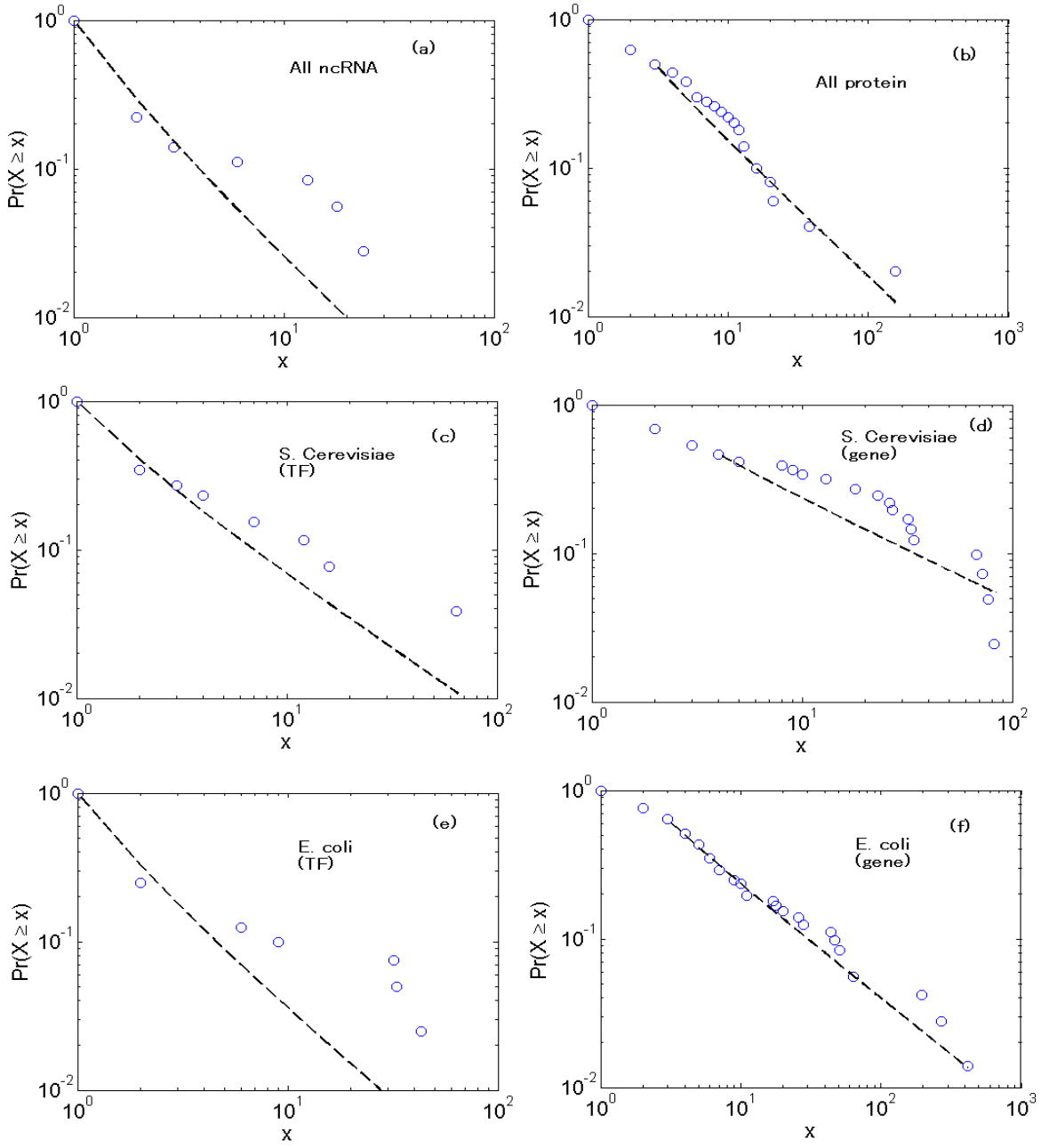

Figure 3: Cumulative degree distributions for the community size *x* of the analyzed bipartite networks. (b) $\gamma = 1.917 \pm 0.459$, **(d)** $\gamma = 1.679 \pm 0.377$, **(c)** $\gamma = 1.021 \pm 0.201$ **(f)** $\gamma = 1.755 \pm 0.102$ **. For (a)** $\gamma = 2.371 \pm 0.321$, **and (e)** $\gamma = 2.212 \pm 0.291$, **the best fit is showed although it clearly deviates from tail of observed experimental data. Note, however, the similar distribution for both ncRNAs and TFs.** x_{min} **is indicated by the starting point of the black dashed-line and corresponds to the point from which the fit is performed.**

The number of identified modules (including single-node module) is 13 (17) for ncRNA (proteins) in *H. sapiens* and 36 (50) when all the available ncRNA-protein interactions are considered. As a comparison, 26 (41) modules are identified for the transcriptional factors (regulated genes) in yeast organism. Fig.1 shows the bipartite network corresponding to the human ncRNA-protein interactions. The ncRNA molecules are indicates by circles and proteins by squares. Each colour indicates a detected community. Four and nine communities with more than one node are observed for human ncRNA and protein molecules, respectively. Of particular interest, is the central community composed of eight ncNRA molecules U1, U2, U4, U5, U6, DD4, U6atac and U4atac. As shown later, some members of this community exhibit overlap with other communities suggesting that their biological roles could also be more relevant. The total number of ncRNA and protein communities for all available interactions, excluding single-node modules, is 8 and 31, respectively. For yeast organism, for example, the transcriptional factor and gene communities with more than one node are 9 and 30, respectively.

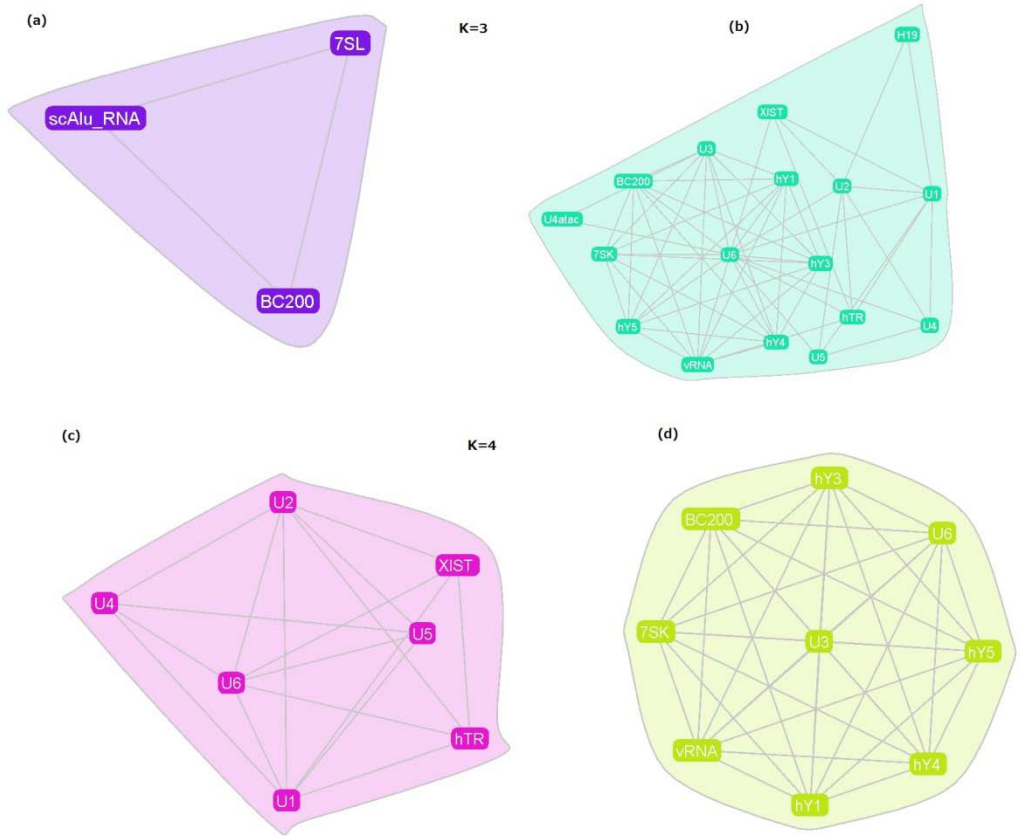

Figure 4: k-clique communities for k=3 and k=4 for human ncRNA network. Each node represents a ncRNA molecule.

3.2 k-cliques and overlapping in human ncRNA interactions

Here we present the results corresponding to the *k*-clique-communities for a given value of *k*. Because the computation of the *k*-cliques requires unipartite graph, the bipartite network of the ncRNA-protein interactions was projected onto a simple network. Therefore, the *k*-cliques identified correspond to sub-graphs composed of only ncRNA molecules. The results show that two *k*-cliques of *k*=3 and *k*=4, respectively, are present in the network (see Fig. 4). In addition, three k-cliques of *k*=5 as well as one clique-community for each *k*=6, 7, 8 and 9,

respectively, were identified (see Fig. 5). By using this k-clique information, the computation of the overlaps between them was performed (Fig. 6). The communities are color coded, the overlapping nodes and are emphasized in black. Large overlaps areas between k-cliques communities are shaded in different color code. Fig. 6 (a, b) display the two communities of the ncRNA molecule U1, and U2, respectively, with *k*=5: the blue and the purple ones. Fig. 6 (c, d) show the two communities of the ncRNA molecule U6 with *k*=4 (c) and three communities of the same molecule with *k*=5 (d).

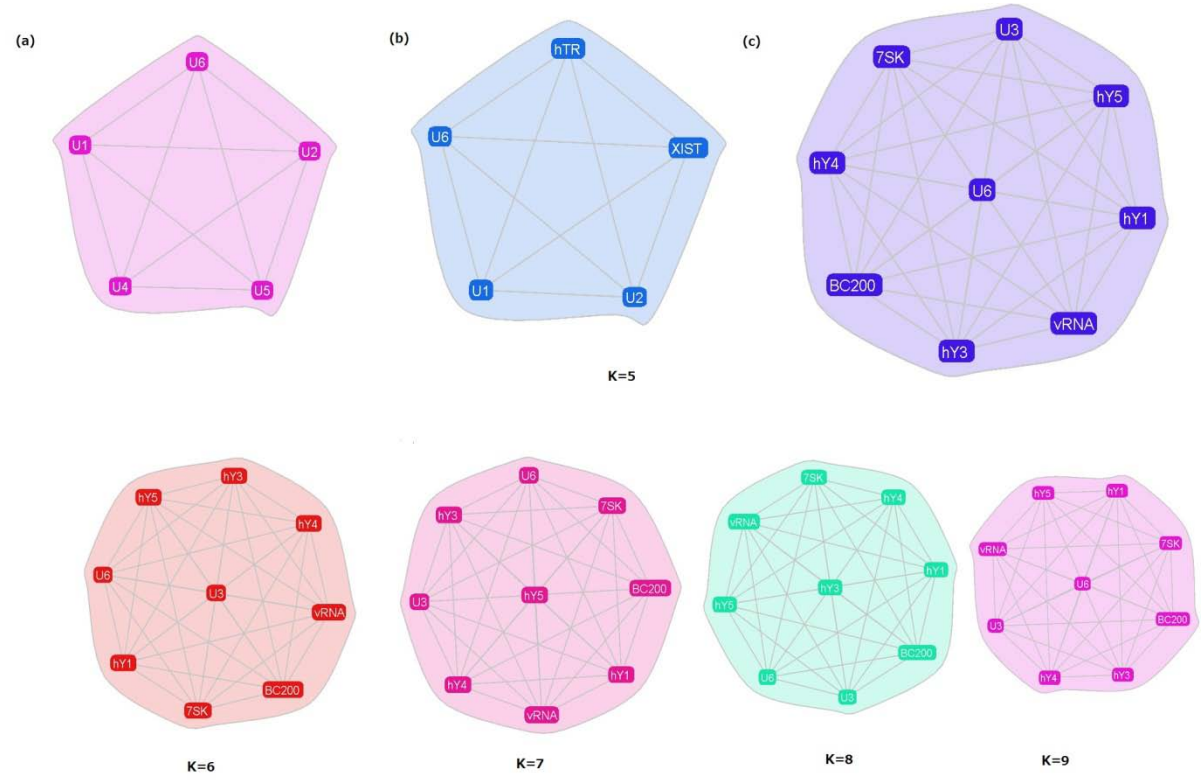

Figure 5: k-clique communities for k=5, 6, 7, 8 and 9 for human ncRNA network. Each node represents a ncRNA molecule.

In eukaryotes the *spliceosome* performs the splicing reactions essential for removing intron sequences. The ncRNA components of the major *spliceosome* are U1, U2, U4, U5, and U6. Interestingly, the *k*-clique community of these molecules is clearly identified in Figs. 4-6. In general, the identified communities of the ncRNA molecules in *H. sapiens* organism could be associated with certain functions.

4 Discussion and Conclusion

Recent works have explored the network of ncRNA molecules from different angles [14-17]. We have examined the community structure of the ncRNA interactions and identified highly cohesive modules in the bipartite network as well as *k*-clique communities in the ncRNA network. The modularity scores as well as the number of meaningful modules at different scales revealed manifestly non-random features. The ncRNA network consists of a high modularity that is organized following a scaling-law for community sizes, similar to those found in other molecular networks. This result supports the idea of a functional system with capability for specific roles in a cell [3, 18]. Recent works have reported that ncRNAs molecules could be implicated in processes related to regulation, cell differentiation and

tumorigenesis. Some analyses have linked ncRNA molecules to other complex diseases like coronary disorders and diabetes [4]. In addition, correlations between microRNA repression and protein interaction using expression data have also been identified [19]. As illustrated above, the *k*-clique communities can be associated to specific functions, therefore an exhaustive mapping of functions available in databases and literature for classified ncRNAs onto the identified clique communities could lead to identify certain functions for yet unclassified ncRNA molecules. Moreover, a complementary analysis could be done by performing a mapping on the presented network using expression data of ncRNA.

In spite of the fact that current collected ncRNA datasets represent a small fraction of all the existing interactions, these findings are promising and encourage the experimental analysis and classification of the non-coding mediated interactions. The clinical potential of ncRNA molecules has not been fully unveiled yet, and it could be possible that non-coding RNAs can play a key role in future personalized therapies [20]. As a future work, we aim to extend the mapping of biological functions on the identified modules and communities. Furthermore, it could be interesting to extend the biological significance of this study by considering the interactions between target genes and those proteins that also interact with ncRNA molecules. Finally, we also work on the development of mathematical models, based on bipartite graphs, that explain the emergence of the scaling-law for community sizes in both transcriptional regulatory networks and ncRNA mediated interactions as well as the *k*-clique community structure.

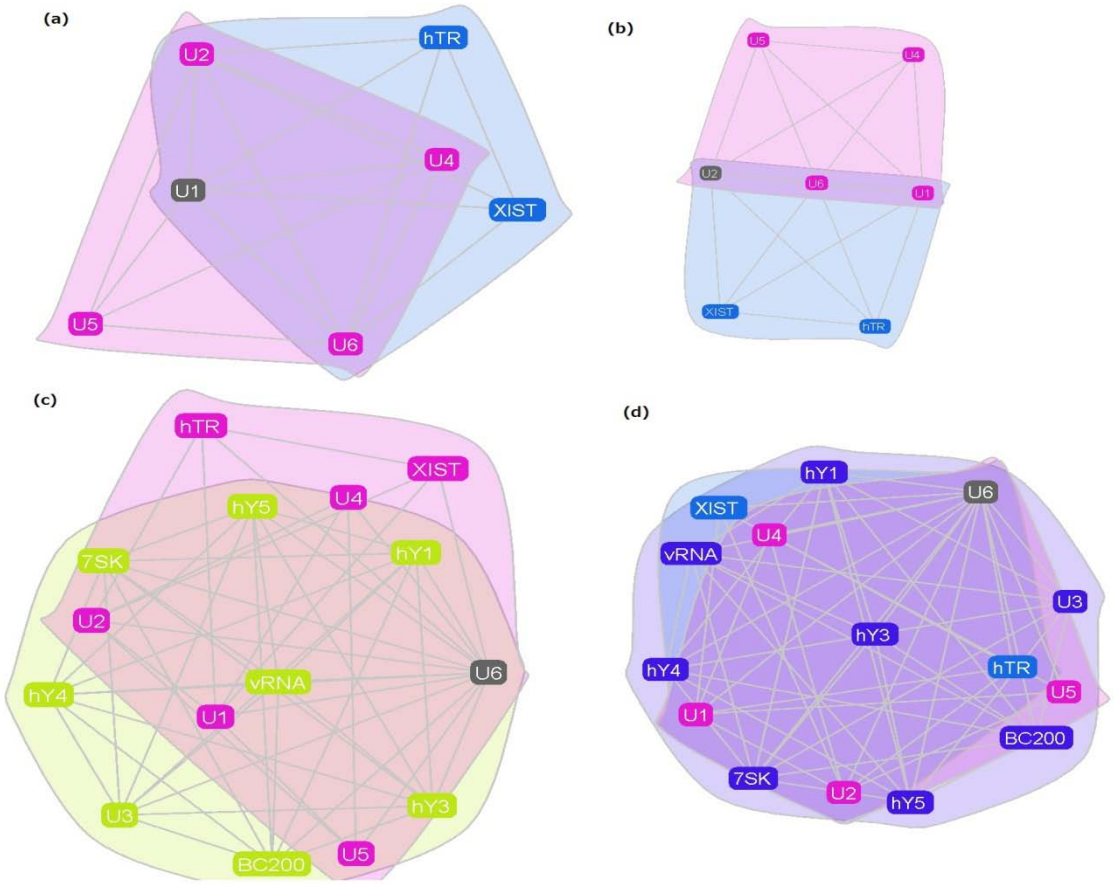

Figure 6: k-clique communities and overlaps for U1, U2, and U6 (black nodes) ncRNA molecules in the human ncRNA network. Each node represents a ncRNA molecule.

Acknowledgements

This work was partially supported by a Grant-in-Aid from MEXT Japan.

References

[1] A-L.Barabási, Z.N. Oltvai ZN Network biology: Understanding the cell's functional organization *Nature Review Genetics*, **5**, 101-113, 2005.

[2] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph et al., Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science* **298**, 799-804, 2002.

[3] J. S. Mattick, "RNA regulation: a new genetics", *Nature Review Genetics*, **5**, 316-323, 2004.

[4] P. P.Amaral, M. E. Dinger, T. R. Mercer and J. S. Mattick, The eukaryotic geneome as an RNA machine, *Science* **319**, 1787-1789, 2008.

[5] J. S. Mattick and I. V. Makunin, Small regulatory RNAs in mammals, *Human Molecular Genetics*, **14**, pp. R121-R132, 2005.

[6] T. Wu, J. Wang, C. Liu, Y. Zhang and B. Shi, NPInter: the noncoding RNAs and protein related biomacromolecules interaction database, *Nucleic Acids Research* **34**, pp. D150-D152, 2006.

[7] M.C. constanzo et al., YPDTM, PombePDTM and WormPDTM: model organisms volumes of BioKnowledgeTM Library, an integrated resource for protein information, *Nucleic Acids Research* **29**, 75-79, 2001.

[8] R. Guimerà., LAN Amaral, Functional cartography of complex metabolic networks. Nature, **433**, 895-900, 2005.

[9] R. Guimerà, M. Sales-Pardo and LAN Amaral, Module identification in bipartite and directed networks. *Physical Review E*, **76**, 036102 , 2007.

[10] L. Danon, A. Díaz-Guilera, J. Duch and A. Arenas Comparing community structure identification. *J Stat Mech*, P09008, 2005.

[11] G. Palla, I. Derényi, I. Farkas and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature **435***, 814-818, 2005.

[12] A. Clauset, C.R. Shalizi and M.E.J. Newman Power-law distributions in empirical data *SIAM Rev.* **51** 661–703, 2009.

[13] http://tuvalu.santafe.edu/ aaronc/powerlaws

[14] U. K. Muppirala,V. G. Honavar and D. Dobbs, Predicting RNA-protein interactions using only sequence information, *BMC Bioinformatics.* **12**: 489, 2011.

[15] L.J. Collins, Advances in Experimental Medicine and Biology: RNA Infrastructure and networks, Vol. **722** of *Advances in Experimental Medicine and Biology, Springer,* 86-102, 2011.

[16] V.P. Zhdanov, Non-coding RNAs and complex distributed genetic networks, *Central European Journal of Physics* **9** (4), 909-918, 2011.

[17] J.C. Nacher and N. Araki Structural characterization and modeling of ncRNA-protein interactions. *Biosystems* **10**, 10-9, 2010.

[18] Y. Shimoni, G. Friedlander, G. Hetzroni, G. Niv et al., Regulation of gene expression by small non-coding RNAs: a quantitative view, *Molecular Systems Biology*, **3**, 1-9, 2007.

[19] H. Liang and W.-H. Li, MicroRNA regulation of human protein-protein interaction n network, *RNA* **13**, 1402-1408 2007.

[20] M. Galasso, M. Elena Sana and S. Volinia Non-coding RNAs: a key to future personalized molecular therapy ?, *Genome Medicine* **2**, 12 2010.