Analysis of Human miRNAs, Gene Targets and Diseases Network

Kwan-Yau Cheung*, Kin-Hong Lee, Kwong-Sak Leung
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong.

* Corresponding author. Tel.: 852-39438430; email: kycheung@cse.cuhk.edu.hk
Manuscript submitted March 10, 2015; accepted May 15, 2015.
doi: 10.17706/ijbbb.2015.5.3.140-148

Abstract: Micro-RNAs are small non-coding RNAs having important biological functions such as gene regulation and disease causality. Network analysis on miRNA-related network can help understanding gene regulation mechanism and propose cures for miRNAs related diseases. In this paper, we integrated miRNA-related data from three state of the art databases, miRTarBase, miRBase and HMDD v2.0 to construct a human miRNAs, gene targets and diseases network. We then performed network statistics analysis, disease cluster analysis and gene-disease association analysis on the network. The results show that there are community structures in the network, similar disease are associated with similar miRNAs with enriched biological functions and gene-disease pairs connected by multiple paths in the network and are more likely to have biological association.

Key words: Network analysis, miRNAs, gene targets, diseases.

1. Introduction

Micro-RNAs (miRNAs) is one type of non-coding RNAs that have important biological functions. miRNAs dysregulation is also known to be related to many diseases, e.g. inherited diseases and cancer [1], [2]. Network analysis of miRNA-related network is important for understanding gene regulation mechanism and proposing cures for miRNAs related diseases.

There are three state of the art databases related to miRNAs, which are miRTarBase [3], miRBase [4] and Human microRNA Disease Database v2.0 (HMDD v2.0 in short) [5]. miRTarBase is a database of published miRNA-target interactions. miRBase is a database of published miRNA sequences and annotations. HMDD v2.0 is a database that curated experiment-supported evidence for human miRNA and disease associations.

The motivation of this paper is to integrate the data from different databases and perform network analysis. In order to construct the network, the miRNA-related data are downloaded from these databases. The rest of the paper is organized as follows. Section 2 presents the data pre-processing of human miRNA, gene targets and disease network, and the analysis methods. Section 3 presents the analysis results. Section 4 concludes this paper.

2. Materials and Methods

2.1. Data Preprocessing

The miRNA related data are downloaded from miRTarBase, miRBase and HMDD v2.0. There are 4 types

of data, which are pre-mature miRNAs (pre-miRNAs in short), mature miRNAs (miRNAs in short), gene targets (genes in short) and diseases.

The miRBase database of Release 20 was downloaded. Each entry in the database contains information about a pre-miRNA and miRNA(s), where pre-miRNA creates the miRNA(s). We extracted every pre-miRNA and miRNA pairs from the database. Each pre-miRNA and each miRNA are represented by an miRBase accession ID. Then, we selected pre-miRNA and miRNA pairs from human only and ignored the pairs of the other species. 2794 pre-miRNA and miRNA pairs remained.

The miRBase database also contains information about the pre-miRNAs and the genes with overlapping locations in the human genome. We extracted every pre-miRNA and gene pairs with this feature from the database. Each pre-miRNA is represented by an miRBase accession ID and each gene is represented by an Ensembl ID (Ensembl ID is used in Ensembl Genome [6]). We converted the Ensembl IDs into Entrez IDs (Entrez ID is used in gene-specific database at the National Center for Biotechnology Information (NCBI) [7]). Then, we selected pre-miRNA and gene pairs from human only and ignored the pairs of the other species. 1513 pre-miRNA and gene pairs remained.

The miRTarBase database of Release 4.5 was downloaded. Each entry in the database contains information about an miRNA and a gene, where the miRNA interacts with the mRNA translating into that gene. We extracted every miRNA and gene pairs, and the corresponding miRTarBase ID from the database. Each miRNA is represented by an miRBase accession ID and each gene is represented by a gene symbol. We converted the gene symbols into Entrez IDs. Then, we selected miRNA and gene pairs from human only and ignored the pairs of the other species. 37387 miRNA and gene pairs remained.

The HMDD v2.0 database was downloaded. Each entry in the database contains information about a pre-miRNA and a disease, where the pre-miRNA and the disease have association. We extracted every pre-miRNA and disease pairs from the database. Each pre-miRNA is represented by an miRBase accession ID and each disease is represented by a disease name. The database only contains human related pre-miRNAs so we did not filter out anything. 6427 pre-miRNA and disease pairs remained.

After pre-processing, there are 1872 pre-miRNAs, 2578 miRNAs, 12760 genes and 380 diseases in total. The human miRNAs, gene targets and diseases network are constructed from the integrated data. The network representation is as follows. Each pre-miRNA, miRNA, gene and disease is represented by a node respectively. Two nodes are connected by an edge if the node pair is inside the processed data. Disease and pre-miRNA are connected if disease and per-miRNA are associated. Pre-miRNA and miRNA are connected if pre-miRNA creates miRNA. Pre-miRNA and gene are connected if pre-miRNA and gene have overlapping locations in the human genome. miRNA and gene are connected if miRNA interacts with the mRNA translating into that gene. Fig. 1 shows the network abstraction and Fig. 2 shows the complete network, which are prepared for using Gephi 0.8.2 [8] with Fruchterman Reingold algorithm [9] to generate the layout.

2.2. Network Statistics Analysis

The network statistics of the network are computed by using Gephi. The network statistics of the network help understanding the network structure. The network statistics computed for the complete network are Average Degree, Network Diameter, Graph Density, Modularity Score and Number of Connected Components. The definitions of the network statistics of the network are as follows. Let the network $G = (V, E)$, where V is the set of nodes and E is the set of edges. Let $u, v, w \in V$ be the nodes of the network. Let the sub-network $G_{sub} = (V_{sub}, E_{sub})$ where $V_{sub} \subset V$ and $E_{sub} \subset E$ be the set of nodes and edges in the sub-network. Let A be the adjacency matrix of graph G where $A_{u,v} = 1$ if $(u, v) \in E$, and $A_{u,v} = 0$ otherwise. Let the community assignment $C(u)$, where $C(u)$ is a function that assign community to node u.

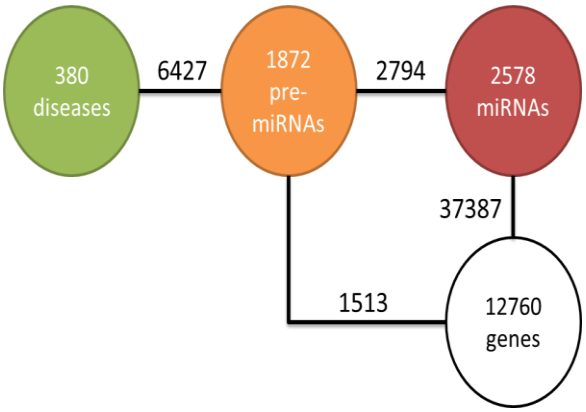

Fig. 1. Network abstraction.

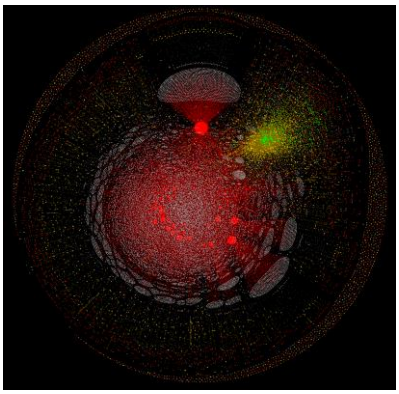

Fig. 2. Picture of the complete network.

Definition 1. Average degree = Average number of edges connected to the node

$$AverageDegree(G) = \frac{2|E|}{|V|}$$

Definition 2. A path between a pair of nodes = A sequence of edges that connect the node pair

$$Path(u,v) = \big((u,t_1),(t_1,t_2),\dots,(t_{n-1},v)\big), \text{where } (u,t_1),(t_1,t_2),\dots,(t_{n-1},v) \in E, \text{and } t_1,t_2,\dots,t_{n-1} \in V$$

Definition 3. Shortest Path(s) between a pair of nodes = A path(s) between a node pair where the number of edges is the smallest among all possible paths of that node pair

$$ShortestPath(u,v) = Path(u,v), \text{where } \forall Path'(u,v), |Path(u,v)| \leq |Path'(u,v)|$$

Definition 4. Network Diameter = Number of edges in the shortest path between the furthest pair of nodes

$$NetworkDiameter(G) = |ShortestPath(u,v)|, \text{where } \forall u', v' \in V, |ShortestPath(u,v)| \geq |ShortestPath(u',v')|$$

Definition 5. Graph Density = Ratio of the number of edges to the number of possible edges

$$Density(G) = \frac{2|E|}{|V|(|V|-1)}$$

Definition 6. Degree of a node = Number of edges connected to the node

$$Degree(v) = |\{u: (u,v) \in E\}|$$

Definition 7. Modularity Score of a given Community Assignment = Fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random

$$Modularity(G,C) = \frac{1}{2|E|}\sum_{u,v}\left[A_{u,v} - \frac{Degree(u)Degree(v)}{2|E|}\right]\sigma\big(C(u),C(v)\big), \text{where } \sigma\big(C(u),C(v)\big)$$
$$= 1 \; if \; C(u) = C(v), \sigma\big(C(u),C(v)\big) = 0 \text{ if otherwise}$$

Definition 8. Number of Connected Components = Number of subgraphs in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph

$$Number\ Connected\ Components(G) = |\{(V_{sub},E_{sub}): \forall u,v \in V_{sub} \; and \; \forall w \in V - V_{sub}, \exists Path(u,v) \; and \; \nexists Path(u,w)\}|$$

Definition 9. Size of the graph or sub-graph = Number of vertices in the graph or sub-graph

$$Size(G) = |V|$$

2.3. Diseases Cluster Analysis

Disease cluster analysis is performed to find out diseases clusters connecting to similar pre-miRNA nodes. A hierarchical clustering is performed on the disease nodes. First, the node distance between all pairs of disease nodes are computed (Def. 11). Then, each disease node is considered as a cluster. The cluster pairs with the smallest cluster distance are combined to form a new cluster. The cluster distances between the new cluster and all the other clusters are computed (Def. 12). The clustering process repeats until only one cluster remained.

The definitions of node distance of a pair of diseases and the cluster distance of a pair of clusters are as follows. Let the network $G = (V, E)$, where V is the set of nodes and E is the set of edges. Let $u, v, w \in V$ be the nodes of the network. Let $S = \{s_1, s_2, \ldots, s_m\}$ and $T = \{t_1, t_2, \ldots, t_n\}$ be the sets of disease nodes, which represents clusters.

Definition 10. Number of common targets of a node pair = Number of nodes connected by both nodes

$$num_{com}(u, v) = |\{w \in V : \{u, w\} \in E \text{ and } \{v, w\} \in E\}|$$

Definition 11. Node Distance of a node pair = Two nodes is close to each other if the number of common targets is high for both nodes

$$NodeDistance(u, v) = 1 - \frac{num_{com}(u, v)}{max\big(W(u) \cup W(v)\big)}, \text{where } W(w) = \{num_{com}(x, w) : x \in V \text{ and } x \neq w\}$$

Definition 12. Cluster Distance of a cluster pair = Average Node Distance between all inter-clusters node pairs

$$ClusterDistance(S, T) = \sum_{u \in S, v \in T} \frac{NodeDistance(u, v)}{|S||T|}$$

After the hierarchical clustering, enrichment analysis is performed for each disease cluster in order to mine potential biological meaning of the disease cluster. This is achieved by finding out the biological meaning from the miRNAs connected to the cluster. First, we extract all pre-miRNAs that are connected to at least two disease nodes in the disease cluster. The list of pre-miRNAs is then inputted to a web-accessible program called TAM [10] to find out the enriched biological functions.

2.4. Gene-Disease Association Analysis

Gene-Disease association analysis is performed to find out the number of gene-disease association in the network. Since the gene nodes and disease nodes are not directly connected in our network, we would like to know if the gene-disease pairs with small distance have biological association. We verified these gene-disease pairs by using DisGeNET database [11], which contains gene-disease associations from the literatures. In the analysis, all gene-disease pairs with distance ≤ 3 in the network are found (Def. 13) first. Then an association score is computed for each gene-disease pair (Def. 14). Finally, each gene-disease pair is checked if the association exists in DisGeNET. The percentages of verified associations are computed for different association score thresholds (Def. 15). The definitions of distance of a node pair, association score of a node pair and percentage of verified associations of the network are as follows. Let the network $G = (V, E)$, where V is the set of nodes and E is the set of edges. Let $u, v \in V$ be the nodes of the network. Let $P = (u, v)$

be the set of gene-disease associations in DisGeNET.

Definition 13. Distance of a node pair = Number of edges in the shortest path between the node pair

$$Distance(u, v) = |ShortestPath(u, v)|$$

Definition 14. Association Score of a node pair = Number of paths between the node pair where the paths have distance ≤ 3

$$Score(u, v) = |\{Path(u, v): |Path(u, v)| \leq 3\}|$$

Definition 15. Percentage of Verified Associations with a given association score threshold = Number of gene-disease pairs exists in DisGeNET divided by the number of gene-disease pairs beyond the threshold

$$VerifiedPercentage(t) = \frac{|\{(u, v): (u, v) \in P \text{ and } (u, v) \in T(t)\}|}{|\{(u, v): (u, v) \in T(t)\}|},$$
$$\text{where } T(t) = \{(u, v): Distance(u, v) \leq 3 \text{ and } Score(u, v) \geq t\}$$

3. Results

3.1. Network Statistics Analysis

Table 1 shows the details of the network statistics of the complete network. The average degree of the network is 5.517 (Def. 1). The network diameter is 16 (Def. 4), which means the furthest pair of nodes has a shortest path with length 16. The graph density is 1.56 e^{-4} (Def. 5) which means the network is very sparse. The number of connected components is 513 (Def. 8). Most connected components are small with sizes ≤ 11 (Def. 9). The size of the largest connected component is 15889. Fig. 3 shows the distribution of number of connected components of different network sizes.

Table 1. Network Statistics of the Complete Network

Statistics	Value	Descriptions from Gephi Wiki [12]
Average Degree	5.517	The average degree of a network is the average degree of the nodes.
Network Diameter	16	The maximal distance between all pairs of nodes.
Graph Density	$1.56e^{-4}$	Measures how close the network is to complete. A complete graph has all possible edges and density equal to 1.
Modularity	0.472	Measures how well a network decomposes into modular communities. A high modularity score indicates sophisticated internal structure. This structure, often called a community structure, describes how the the network is compartmentalized into subnetworks. These sub-networks (or communities) have been shown to have significant real-world meaning.
Connected Components	513	Determines the number of connected components in the network.

The maximum modularity score is 0.472 (Def. 7) and positive values, indicating the possible presence of community structures (in short communities) (-1 ≤ *modularity score* ≤ 1). In the community assignment of the maximum modularity score, all connected components with size ≤ 11 forms a community structure themselves and the largest connected component are divided into 29 community structures. Fig. 4 shows the distribution of sizes of the 29 community structures from the largest connected component. 19 out of 29 communities are larger than > 100 in size. Most communities have similar distribution of numbers of nodes in each data type (pre-miRNA, miRNA, gene and disease). Fig. 5 shows the distributions of numbers

of nodes in each data type of the 19 communities with sizes > 100. The disease nodes are rare. The miRNA nodes are more than pre-miRNA nodes but the numbers are comparable. The gene nodes are much more than the other types of nodes. The community "540" has a different distribution. There are 378 disease nodes. The pre-miRNA nodes are more than disease nodes but the numbers are comparable. The pre-miRNA nodes are more than the miRNA nodes, and the miRNA nodes are more than gene nodes.

To conclude, the network statistics of the complete network shows that the network is sparse and community structures exist in the network. The communities are similar to one another except community "540", where the data types distribution and some network statistics are different.

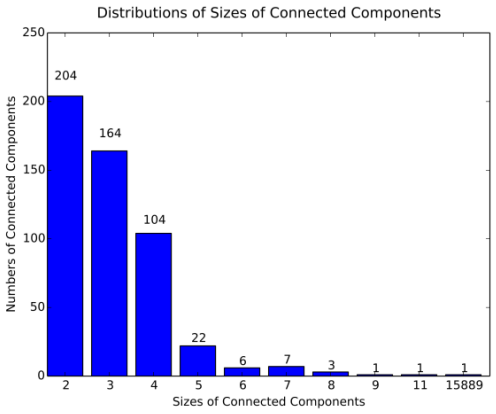

Fig. 3. Distribution of connected components sizes.

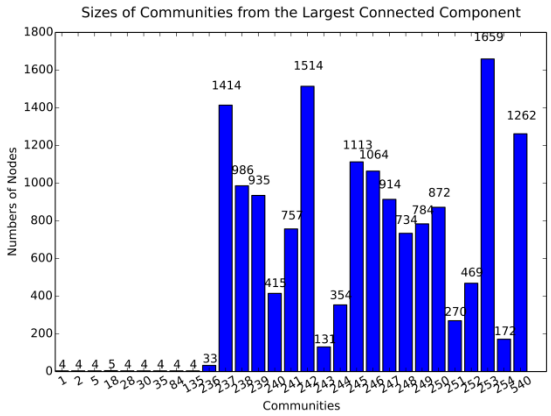

Fig. 4. Communities sizes from the largest component.

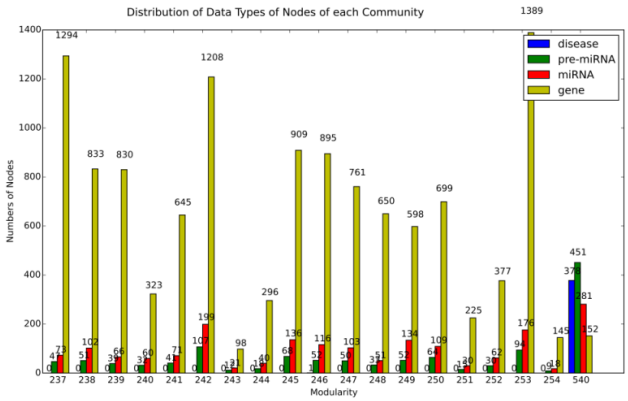

Fig. 5. Distribution of data types of nodes of each community.

3.2. Diseases Cluster Analysis

The hierarchical clustering result will be presented here. Fig. 6 shows the heatmap of hierarchical clustering result. The heatmap shows that some diseases form a cluster, which means they are associated to similar pre-miRNAs. Some clusters contain nodes from similar diseases. Cluster"543" with 13 nodes are related to Lymphoma and Leukemia diseases. Cluster "590" with 29 nodes is related to Neoplasms and Carcinoma diseases. Cluster "490" with 7 nodes is related to Hepatitis disease.

The functional enrichment analysis of the disease clusters will be presented here. Table 2 shows the functional enrichment analysis for three disease clusters "543", "590" and "490". The biological functions found in clusters "543" and "590" are statistically significant in P-values and Bonferroni Values (small value

is more statistically significant).

Table 2. Functional Enrichment of the Disease Clusters

Cluster IDs (Frequently occurred diseases)	Number of Diseases in cluster	Number of miRNAs connected	Biological Functions (with smallest Bonferroni values)	P-values	Bonferroni values
543 (Lymphoma and Leukemia)	13	18	Hematopoiesis	$< 1.00e^{-8}$	$2.90e^{-9}$
			Cell proliferation	$< 1.00e^{-8}$	$7.04e^{-9}$
			Immune response	$< 1.00e^{-8}$	$8.17e^{-9}$
			Apoptosis	$< 1.00e^{-8}$	$1.51e^{-7}$
			HIV latency	$< 1.00e^{-8}$	$1.89e^{-7}$
590 (Neoplasms and Carcinoma)	29	366	Cell cycle related	$< 1.00e^{-8}$	$1.35e^{-8}$
			Apoptosis onco-miRNAs	$< 1.00e^{-8}$	$6.47e^{-7}$
			miRNA tumor suppressors	$1.00e^{-8}$	$1.91e^{-6}$
			Hormones regulation	$2.00e^{-8}$	$6.85e^{-6}$
				$2.00e^{-8}$	$7.01e^{-6}$
490 (Hepatitis)	7	1	Cholesterol biosynthesis	$3.70e^{-3}$	0.0665
			HCV infection	0.0111	0.200
			Carbohydrate metabolism	0.0129	0.233
			Circadian clock	0.0185	0.333
			Circadian rhythm	0.0203	0.366

The majority of the diseases in cluster "543" are related to Lymphoma and Leukemia. The biological functions enriched are also related to these diseases. Hematopoiesis (Bonferroni value = $2.90e^{-9}$) is associated with Leukemia and immune response (Bonferroni value = $8.17e^{-9}$) is associated with Lymphoma.

Another interesting disease cluster is "590". Cluster "590" is related to diseases Neoplasms and Carcinoma. The biological functions enriched are also related to these diseases. Onco-miRNAs (Bonferroni value = $1.91e^{-6}$, which means miRNAs that are associated with cancer, are associated with Carcinoma and miRNA tumor suppressors (Bonferroni value = $6.85e^{-6}$) are associated with Neoplasms. These result shows that the disease clusters found by hierarchical clustering are biological meaningful. As a future work, we can collaborate with the Biologist for further interpretation in the biological meanings and functions of the disease clusters.

For cluster "490", the functions enriched are only marginally significant in P-values and insignificant in Bonferroni values. One possibility is that the number of miRNAs is too small, which affects the enrichment analysis result.

To conclude, the hierarchical clustering results show that some diseases are associated with similar pre-miRNAs. Some disease clusters contain similar diseases and the miRNAs connected to the disease cluster contains enriched biological functions. These biological functions are the biological meanings of these disease clusters.

3.3. Gene-Disease Association Analysis

The percentages of verified association of the gene-disease pairs (Def. 15) in the network will be presented here. Fig. 7 shows the percentages of verified associations with different association score thresholds. Numbers of verified gene-disease associations are presented in blue bar and percentages of verified gene-disease associations are presented in red line. The percentage of verified association increases as the association score threshold increases. The verified percentage > 50% when the association score ≥ 16. This result shows that some gene-disease pairs in the network are associated and the gene-disease pairs with high association scores are likely to be verified in literature. In our network, genes

and diseases are not directed connected but with miRNAs in between. The high association score suggests some gene-disease association is related to miRNAs.

To conclude, some gene-disease pairs in the network are biologically associated and some of the gene-disease pairs association score are related to miRNAs.

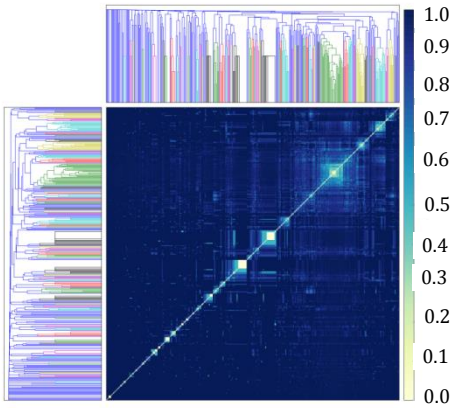

Fig. 6. Heatmap of hierarchical clustering result.

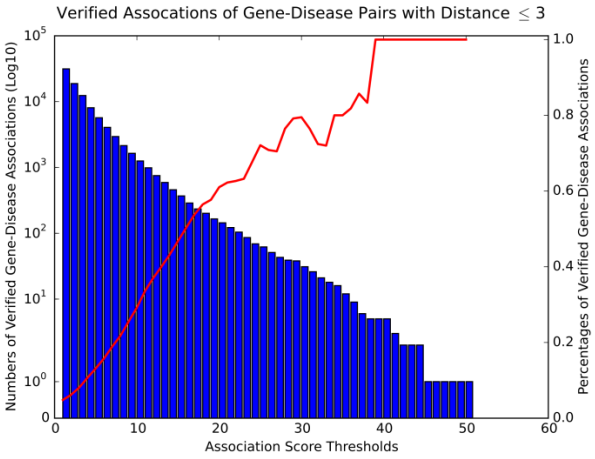

Fig. 7. Percentages of verified associations with different thresholds.

4. Conclusion

In this paper, we have collected and cleansed the data from three state of the art miRNA-related databases, which are miRBase, miRTarBase and HMDD v2.0. A network representing the relationships between miRNAs, genes and diseases have been constructed. We then performed network statistics analysis, disease cluster analysis and gene-disease association analysis to the network. The network statistics analysis shows that community structures exist in the network. The disease cluster analysis shows that similar diseases are associated with similar miRNAs with enriched biological functions. The gene-disease association analysis shows that some gene-disease pairs are biologically associated with miRNAs involved.

Acknowledgment

This research is partially supported by the Direct Grant of CUHK and the General Research Fund (Project ref: 414413) of RGC, Hong Kong SAR, China.

References

[1] Menc´ıa, A., Modamio-Høybjør, S., Redshaw, N., Mor´ın, M., Mayo-Merino, F., Olavarrieta, L., *et al*. (2009). Mutations in the seed region of human mir-96 are responsible for nonsyndromic progressive hearing loss. *Nature Genetics*, *41(5)*, 609-613.

[2] Mraz, M., & Pospisilova, S. (2012). Micrornas in chronic lymphocytic leukemia: From causality to associations and back. *Expert Review of Hematology*, *5(6)*, 579-581.

[3] Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., *et al*. (2014). Mirtarbase update 2014: An information resource for experimentally validated mirna-target interactions. *Nucleic Acids Research*, *42(D1)*, D78-D85.

[4] Kozomara, A., & Griffiths-Jones, S. (2014). Mirbase: Annotating high confidence microrna using deep sequencing data. *Nucleic Acids Research*, *42(D1)*, D68-D73.

[5] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., & Cui, Q. (2014). Hmdd v2. 0: A database for

experimentally supported human microrna and disease associations. *Nucleic Acids Research*, *42*, 1070-1074.

[6] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., *et al*. (2011). Ensembl 2012. *Nucleic Acids Research*, *40*, 84-90.

[7] Maglott, D., Ostell, J., Pruitt, K., D., & Tatusova, T. (2011). Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Research*, *39(suppl 1)*, 52-57.

[8] Bastian, M., Heymann, S., Jacomy, M., *et al*. (2009). Gephi: An open source software for exploring and manipulating networks. *ICWSM*, *8*, 361-362.

[9] Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21(11)*, 1129-1164.

[10] Lu, M., Shi, B., Wang, J., Cao, Q., & Cui, Q. (2010). Tam: A method for enrichment and depletion analysis of a microrna category in a list of micrornas. *BMC Bioinformatics*, *11(1)*, 419.

[11] Bauer-Mehren, A., Rautschka, M., Sanz, F., & Furlong, L. I. (2010). Disgenet: A cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics*, *26(22)*, 2924-2926.

[12] GitHub. (2015). Gephi/gephi. Retrieved May 19, 2015, from https://github.com/gephi/gephi/wiki/Statistics

Cheung Kwan Yau received the BSc degree in computer science from the Chinese University of Hong Kong in 2011, where he is currently working toward the MPhil degree in the Department of Computer Science and Engineering under the supervision of Prof. K. S. Leung and Prof. K. H. Lee. His research interests include bioinformatics and artificial intelligence.

Kin-Hong Lee received the B.S. and M.S. degrees in computer science from the University of Manchester, Manchester, U.K. He was an associate professor with the Department of Computer Science and Engineering, Chinese University of Hong Kong and retired in 2013. His research interests include computer architecture and bioinformatics. He has published over 120 papers in these two fields.

Kwong-Sak Leung received his BSc (Eng.) and PhD degrees in 1977 and 1980, respectively, from the University of London, Queen Mary College. He joined the Computer Science and Engineering Department at the Chinese University of Hong Kong in 1985, where he is currently a professor of computer science & engineering. His research interests are in bioinformatics and soft computing including evolutionary computation, parallel computation, probabilistic search, information fusion and data mining, fuzzy data and knowledge engineering.