Topological Characterization of Housekeeping Genes in Human Protein-Protein Interaction Network

Pei Wang
Sch. Math. Inf. Sci.
Henan University
Kaifeng 475004, China
Email: wp0307@126.com

Yuhuan Zhang
Sch. Math. Inf. Sci.
Henan University
Kaifeng 475004, China
Email: baiyigecun@163.com

Jinhu Lü
Acad. Math. Syst. Sci.
Chinese Academy of Sciences
Beijing 100190, China
Email: jhlu@iss.ac.cn

Xinghuo Yu
Sch. Elect. Comput. Eng.
RMIT University, Melbourne
VIC 3001, Australia
Email: x.yu@rmit.edu.au

***Abstract*—Human housekeeping genes (HKGs) are widely expressed in various tissues, which involve in cell maintenance or sustaining cell function, and are often taken as experimental control and normalization references in gene expression experiments. Based on literature curation and up-to-date databases, we construct a large-scale human protein-protein interaction network (HPIN) and a HKGs subnetwork. Through the topological features of HKGs in the HPIN, we characterize the topological features of human HKGs. Our results indicate HKGs are with very large average degree, k-shell, betweeness, semi-local and eigenvector centralities, clustering coefficient, closeness, PageRank and motif centrality, which are all higher than that of the HPIN. Among the nine indexes, HKGs are with the average betweeness about 7 times larger than that for the HPIN, but they are also with the largest coefficient of variant (CV). The closeness of HKGs is with the smallest CV and very large median. Based on ROC analysis, we find most of the indexes and their compositions can be used to predict HKGs, with prediction accuracy around 80%. Especially, the prediction accuracy of the closeness can achieve as high as 82.36%. The investigations shed some lights on the characterization and identification of human functional genes, which have potential implications in systems biology and networked medicine.**

I. Introduction

With the development of high-throughput technologies, such as the yeast two-hybrid and mass spectrometry technique, large-scale protein-protein interaction networks have been facilitated to be constructed. Many databases have been established to provide the binary interactions among proteins for various organisms, such as the Human Protein Reference Database (HPRD) [1], the BIOlogical General Repository for Interaction Datasets (BioGRID) [2], the Online Predicted Human Interaction Database (OPHID) [3]. The increasingly accumulated datasets and the arrival of the era of big data facilitate the exploration of the structural characteristics of large-scale HPIN.

Human proteins are encoded by genes, therefore, proteins in the HPIN correspond to human genes. From different perspectives, human genes can be classified into different categories. From the perspective of whether a gene is widely expressed in various human tissues or not, human genes can be classified into HKGs and tissue-enriched genes (TEGs) [4]- [8]. From the perspective of whether mutations on a gene is lethal or not, human genes can be classified into essential genes and viable genes [9]. In this paper, we mainly consider the HKGs. HKGs are genes that broadly express in various tissues, which involve in some way in processes necessary to the survival of a cell. Some HKGs may involve in sustaining cell function, while others may involve in cell maintenance. HKGs tend to produce proteins at steady rates, and errors in their expression can lead to cell death. Like a housekeeper, they keep a cell running smoothly so that it can continue to function, and they also contribute to the overall function of larger organisms [4]- [8]. TEGs only express and function in a few specific human tissues. It is reported that HKGs are more conserved than the other genes and evolve more slowly [7]. Therefore, HKGs have been widely used as experimental controls and normalization references in gene expression experiments [4]. Since TEGs are highly expressed in one or a few specific tissues or cell types, they can serve as biomarkers of particular tissues or biological processes, some of them may act as drug targets [4].

Characterization and identification of HKGs and TEGs have attracted an increasing attention over the last decades [4], [5], [8], which are mainly based on microarray gene expression profiling analysis. For example, in 2008, Zhu et al. [5] reported 1206 HKGs from microarray data, which are widely expressed in 18 human tissues. In 2009, She et al. [4] found 1522 HKGs and 975 TEGs from 18149 human genes, where the HKGs are highly expressed in 42 human tissues. However, few results have been reported on the topological features of HKGs in the HPIN. With the rapid development of complex network theory, it is feasible to explore the topological characterization of HKGs in the HPIN.

Motivated by the above problems, we will construct an up-to-date large-scale HPIN and investigate the topological characterization of HKGs in the HPIN. The rest of the paper is organized as follows. We construct the HPIN and the subnetwork of HKGs in Section II, simultaneously, we investigate the structural characteristics of these networks. Section III explores the topological characterization of HKGs in the

2014 The 8th International Conference on Systems Biology (ISB)
978-1-4799-7294-4/14/$31.00 ©2014 IEEE

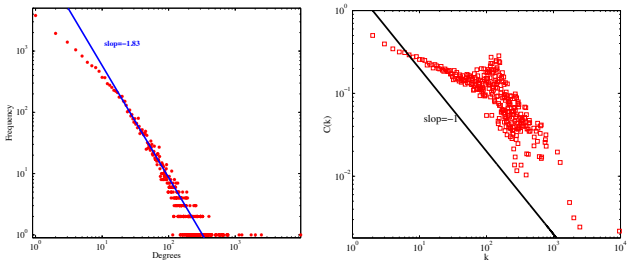
Qingdao, China, October 24–27, 2014

HPIN. Discussions and conclusions are in the last Section IV.

II. Human Protein-Protein Interaction Network And HKGs

A. Human protein-protein interaction network

Based on up-to-date data collected in the BioGRID, HPRD and curated from the literature [10], we constructed the HPIN. The BioGRID covers literature-curated data from the year 1970 to Jan. 2014. Till January 2014, the BioGRID database has collected 153379 interactions among 16287 proteins. The HPRD has included 39008 interactions. Except the data from the BioGRID and HPRD, binary protein-protein interactions have been also widely reported in many literature [10], [11]. For example, in 2005, based on the Y2H high throughput technology and literature curation, Rual et al. [10] constructed a connected HPIN with 2784 nodes and 6438 interactions. By integrating the data from BioGRID, HPRD and literature [10], we derive a raw HPIN, which contains 17423 proteins and 178469 interactions. The giant connected component (GCC) contains 17311 nodes and 151412 interactions. The average degree $\langle k\rangle$ of such connected network is 17.4932. It is estimated that the complete human protein interactome contains about 25000 gene-encoding proteins and more than 375000 interactions among them [12], [13]. Therefore, though the considered HPIN may still suffer from the sampling effect [14], they cover almost 70% of the complete human interactome, the investigations on such large-scale network can provide hints for the understanding of the complete interactome.



(a) (b)

Fig. 1. (a) Degree distribution of the HPIN. (b) The average clustering coefficient $C(k)$ as a function of degree k indicates hierarchical modularity of the HPIN.

Statistical indexes for the GCC of the HPIN are shown in Tab.1, where we have shown its average degree $\langle k\rangle$, maximum and minimum degrees k_{max}, k_{min}, network diameter D, average path length APL, clustering coefficient C^*, small-world index SW, Pearson correlation coefficient PCC, Power-law exponent PLE. Here, C^* is defined as the ratio of number of triangles to number of paths of length 2 [15]. SW is defined as follows [17].

$$SW = \frac{C^*/C^*_{rand}}{APL/APL_{rand}}, \tag{1}$$

where C^*, C^*_{rand} are the clustering coefficients for the HPIN and the randomized networks. APL_{rand} is the average path length for the randomized networks. For Erdös-Rényi (ER) random networks with n nodes and average degree $\langle k\rangle$, the clustering coefficient C^*_{rand} can be approximated by $C^*_{rand} = \langle k\rangle/n$. APL_{rand} can be approximated by $APL_{rand} = ln(n)/ln(\langle k\rangle)$ [17], [18]. $SW > 1$ indicates the small-worldness of the network. PCC is defined in Ref. [15], [16], $PCC < 0$ indicates the disassortativity of the network.

TABLE I
Statistical characteristics of the HPIN, the HKGs and TEGs subnetworks.

		HPIN	HKGs	TEGs
Raw data	Nodes	17423	1389	697
	Interactions	178469	10306	246
GCC	Node	17311	1346	138
	Edge	151412	10306	179
	$\langle k\rangle$	17.4932	15.3135	2.5942
	k_{max}	9638	1216	27
	k_{min}	1	1	1
	D	11	6	11
	APL	2.7736	2.1520	4.6149
	CC	0.2281	0.4163	0.1548
	C^*	0.0070	0.0366	0.0334
	SW	8.5168	3.9471	1.9900
	PCC	-0.0637	-0.1338	-0.1751
	PLE	-1.8300	-1.4900	-1.6810

From Tab.I, we can conclude that the HPIN is sparse, with connection density 0.1011%. The APL, C^* and SW in Tab.I indicate the HPIN is small-world. Fig.1(a) shows the degree distribution of the HPIN, where we can conclude that it is power-law, and with $PLE = -1.83$. Moreover, the PCC of the HPIN is -0.0637, which indicates the disassortativity. To verify whether the HPIN is with modularity, we draw the average clustering coefficient $C(k)$ as a function of degree k, as shown in Fig.1(b). Here, $C(k)$ is the average clustering coefficient of nodes with degree k [19], where the clustering coefficient C_i for node i is defined as $C_i = 2n_i/(k_i(k_i-1))$ [20]. k_i is the degree of node i, n_i represents the number of links among the k_i neighbors. From [19], if $C(k)$ versus k distributes along the line with slop -1 in log-log scale, then the network is with modular structure. From Fig.1(b), we can conclude that the HPIN is hierarchical modular. Existing investigations have illustrated that the yeast protein-protein interaction network is sparse, small-world, scale-free, disassortative and with modularity [21], [22], our investigations clarify that the HPIN has similar properties as that for the yeast.

B. Housekeeping genes

In the following, the HKGs and TEGs predicted in Ref. [4] will be used to investigate the topological characteristics of the HKGs. Firstly, we construct the connection network for HKGs. Among the 1522 HKGs in Ref. [4], 1389 HKGs and 10306 interactions are in the HPIN, where 1346 proteins are connected. 697 of the 975 TEGs are nodes in the HPIN, and 138 nodes are largely connected through 179 links. The network for the HKGs are shown in Fig.2. The statistical indexes for the networks of HKGs and TEGs are shown in Tab.I, where we can conclude that the two subnetworks are with similar

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

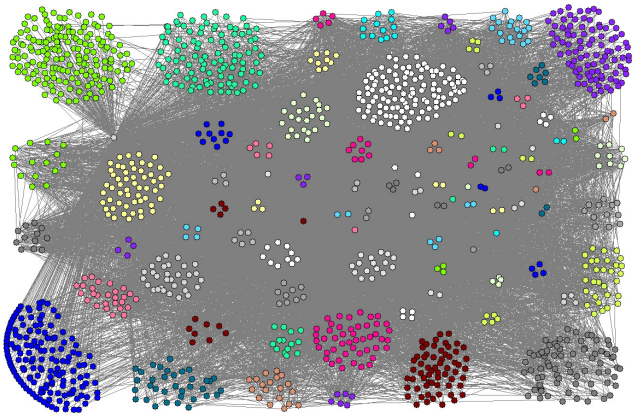


Fig. 2. The connection network of the 1346 HKGs in the HPIN. Clusters with different colors are with different degrees.

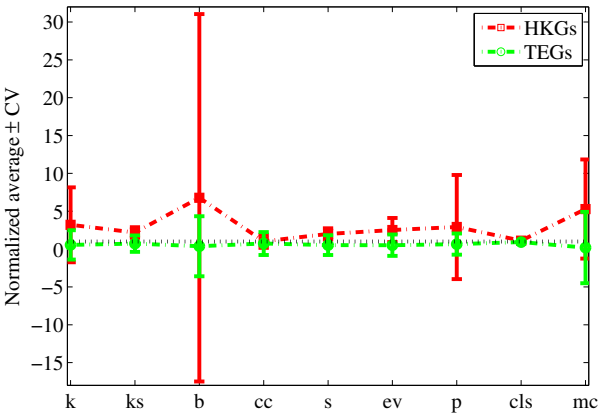


Fig. 3. Normalized and CV of the nine indexes for the 1389 HKGs and 697 TEGs.

properties as the HPIN. The APL for the network of HKGs is 2.1520, which means that any two HKGs can connect with each other through about two interactions. Whereas, there is a two-fold increase in the APL for the network of the TEGs. Moreover, the HKGS and TEGS are both small-world, with SW indexes 3.9471 and 1.9900, respectively. Furthermore, negative PCC values indicate the two subnetworks are all disassortative.

III. Topological Characterization of HKGs in the HPIN

A. Structural features of HKGs in the HPIN

To characterize the HKGs in the HPIN, we compute the statistical features of each node in the HPIN and extract the features for the HKGs. Generally speaking, degree [15], betweeness [23], k-shell [24], semi-local centrality [25], closeness [25], PageRank [26], eigenvector centrality [15], clustering coefficient [15] and network motif centrality [22], [27] can characterize the structural importance of nodes in a complex network. In the following, we compute the nine indexes for the HPIN. It is noted that for the network motif centrality, we have only considered the 3-node fully connected motif [22]. Hereinafter, we derive the average indexes for all nodes in the HPIN, and the average indexes for the 1389 HKGs and 697 TEGs in the HPIN, as summarized in Tab.II. Here, $k, ks, b, cc, s, ev, p, cls, mc$ denote the degree, k-shell, betweeness, clustering coefficient, semi-local centrality, eigenvector centrality, PageRank, closeness and motif centrality, respectively. Std denotes standard deviation. CV denotes the coefficient of variation, which is defined as the ratio of the standard deviation to the mean. The CV is a normalized measure of dispersion of a probability distribution or frequency distribution. To compare among different indexes, we normalized the average indexes for the HKGs and TEGs by the corresponding average values for the overall HPIN. The normalized indexes and CV of the nine indexes are shown in Fig.3.

From Tab.II and Fig.3, we can see that all the average indexes for the HKGs are larger than that for the HPIN. Whereas, the average indexes for the TEGs are all lower than that for the HPIN. Especially, the average betweeness and motif centrality for the HKGs are more than five times larger than that for the HPIN. Moreover, the HKGs are with average degree more than 3 times larger than the average of the HPIN. The clustering coefficient and closeness of the HKGs are a little larger than that for the HPIN. However, among the nine indexes, the betweeness values of the HPIN and HKGs have the largest CV, which indicates that although the average betweeness of the HKGs are very high, its distribution is more decentralized than the other indexes. Additionally, for the HKGs, the standard deviation Std and CV of the degree, PageRank and motif centrality are all very large. Whereas, for the HKGs, the closeness, semi-local centrality, clustering coefficient and k-shell have very small CV. The median of the closeness for the 1389 HKGs is $2.380e-5$, which is much larger than the average; while the median of the other indexes are all smaller than their averages. The large average betweeness of the HKGs indicate that they tend to act as bottlenecks or bridges in the HPIN. The large motif centrality for the HKGs reveal that the HKGs tend to frequently involve in the triangle motifs, and act as building blocks of the HPIN. The large average degree indicate the HKGs tend to be hubs of the HPIN. For the HKGs, the large average and median of the closeness and the smallest CV indicate most of the HKGs are with much larger closeness than the other nodes. In summary, the HKGs are with marked structural features, which facilitate the characterization and identification of them through topological structures of the HPIN.

B. ROC analysis

ROC (Receiver Operating Characteristic) curves are widely used in the area of medical tests and signal processing, which can evaluate the performance of a new test [28], [29] . The general idea of such analysis is as follows. Suppose the

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

TABLE II
STATISTICAL CHARACTERISTICS OF NINE STRUCTURAL INDEXES FOR THE HPIN, HKGS AND TEGS.

	HPIN				HKGs				TEGs			
Index	Average	Std	CV	Median	Average	Std	CV	Median	Average	Std	CV	Median
k	17.493	88.019	5.032	5	56.038	277.054	4.944	22	9.6399	18.7116	1.9411	4
ks	8.930	10.017	1.122	5	19.596	13.783	0.7034	17	6.4548	6.995	1.08437	4
b	3.155e4	1.48e6	46.889	407.292	2.138e5	5.187e6	24.264	7.856e3	1.184e4	4.693e4	3.9652	246.763
cc	0.228	0.278	1.220	0.1429	0.2341	0.201	0.8587	0.182	0.1673	0.253	1.5117	0.0635
s	7.973e7	7.782e7	0.976	9.814e7	1.595e8	1.180e8	0.7396	1.361e8	4.222e7	5.514e7	1.3062	1.073e7
ev	0.0039	0.0065	1.6596	0.0033	0.0099	0.0157	1.5875	0.006	0.002	0.0028	1.4121	7.339e-4
p	1.000	6.072	6.072	0.424	2.9127	20.0326	6.8776	1.1697	0.6583	0.9193	1.3964	0.4004
cls	2.099e-5	3.147e-6	0.150	2.33e-5	2.325e-5	1.946e-6	0.0837	2.380e-5	1.937e-5	2.964e-6	0.153	1.840e-5
mc	82.622	865.939	10.481	3	437.217	2.863e3	6.549	44	16.463	77.654	4.7167	1

concerned n subjects can be classified into positive (normal or important) or negative (abnormal or unimportant), and the actual classification of the n subjects has been known, which is called gold standard. We have a new measure, and we want to evaluate the performance of such measure on the classification of all the subjects. If we take a threshold value T for the new test, then the n subjects can be classified into positive (above T) or negative (below T) through such threshold. By comparing between the classifications from the new test and the gold standard, we can derive a contingency table, as shown in Tab.III. Based on Tab.III, we can define several measures to evaluate the performance of the new test, which are defined as follows.

$$fpr = \frac{fp}{fp + tn}, \tag{2}$$

$$tpr = \frac{tp}{tp + fn}, \tag{3}$$

$$acc = \frac{tp + tn}{n}, \tag{4}$$

where fp denotes the number of false positive nodes, which are actually the number of nodes which are positive in the new test but negative in the gold standard. Similarly, tn represents the number of true negative nodes, tp and fn denote the number of true positive and false negative nodes, respectively. fpr, tpr are therefore called false positive rate and true positive rate, respectively. acc defined in eq.(4) is called the accuracy of the new measure under the threshold value T.

TABLE III
CONTINGENCY TABLE OF THE FOUR OUTCOMES FROM THE NEW TEST AND THE GOLD STANDARD.

		Gold standard	
		Condition positive	Condition negative
Test outcome	Test positive	True positive: tp	False positive: fp
	Test negative	False negative: fn	True negative: tn

Given a threshold value T, one can obtain a coordinate point (fpr, tpr). When the threshold value T is taken over the range of the new measure and we plot the corresponding points in two dimensional coordinate system, we can derive the ROC curve. Suppose the range of the new measure is $[A, B]$, when $T = A$, then all the subjects are treated as positive and $tn = 0, fn = 0, fpr = 1, tpr = 1$; when $T = B$, all the subjects will be treated as negative from the new test, and $tp = 0, fp = 0, tpr = 0, fpr = 0$. Therefore, the ROC curve must locate in the area $[0, 1] \times [0, 1]$, and $(0, 0), (1, 1)$ are two extreme points of the ROC curve. The area under the curve (AUC) of ROC can measure the accuracy the the new test. The bigger AUC, the better the new test will distinguish between the positive cases and negative ones. If $AUC = 1$, the new test can act as a perfect classifier. If $0.5 < AUC < 1$, the new test is better than a random classifier. For the tests with very high AUC, with properly chosen threshold values, one can use them to predict the abnormal subjects.

C. Characterization of the HKGs via ROC curves

In the above subsections, we have discussed the structural features of HKGs in the HPIN and reviewed the idea of ROC analysis, in the following, through ROC curves, we discuss the characterization and identification of HKGs in the HPIN.

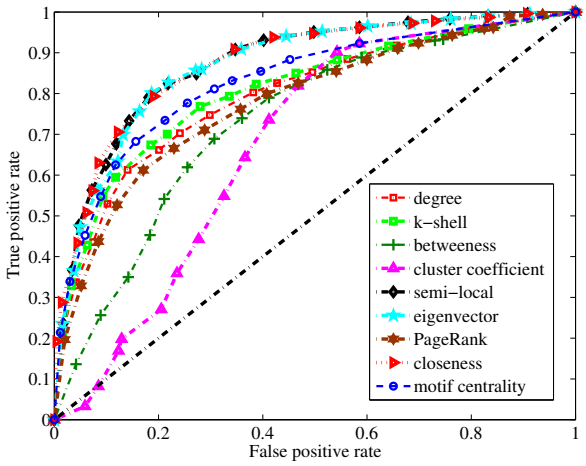


Fig. 4. ROC curves for the nine measures. The nine indexes for the 1389 HKGs and 697 TEGs in the HPIN are used to plot the ROC curves.

Since there are 15225 nodes with unknown types in the HPIN, we only consider the HKGs and TEGs, and take the 1389 HKGs as positive and the 697 TEGs as negative ones. Based on the nine indexes for the 2086 genes in the HPIN, we evaluate the performance of each index in characterizing the

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

HKGs. For each index, we take 21 threshold values T to derive the ROC curve. Each index under each T acts as a classifier and classified the 2086 nodes into positive and negative. The T is taken as top $0\%, 5\%, ..., 95\%, 100\%$ of a index. Fig.4 shows the ROC curves for the nine indexes.

From Fig.4, all the ROC curves for the nine indexes are with $AUC > 0.65$. Especially, the AUC for the closeness, semi-local centrality and eigenvector centrality are the largest, which are actually 0.8763, 0.8741 and 0.8706, respectively. This indicates that they can more effectively predict the HKGs. Except the three indexes, the AUC for the motif centrality and k-shell are 0.8301 and 0.8085, and indicate that they are another two effective indexes to identify the HKGs. The clustering coefficient is with the lowest AUC, which are 0.6778. The AUC for the betweeness is 0.7333, which is only bigger than that for the clustering coefficient. For the closeness, when $T = 60\%$, acc achieves its maximum value 0.8236. That is, when the closeness is used to predict HKGs in the HPIN, the highest accuracy rate will be 82.36%.

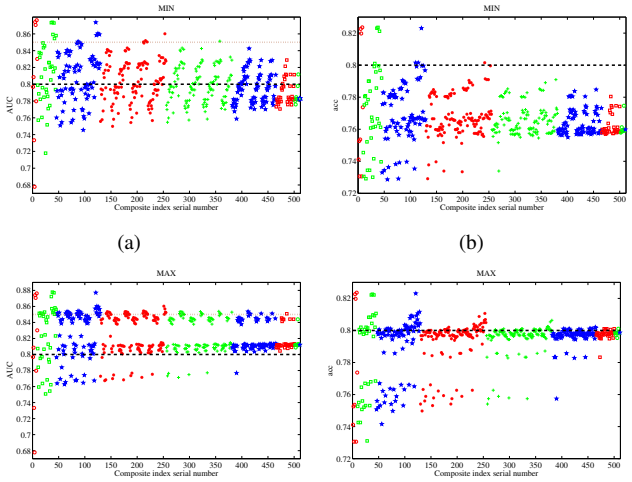


Fig. 5. (a) AUC and (b) acc under composite index with the MIN operation. (c) AUC and (d) acc under composite index with the MAX operation.

Following, we investigate whether the compositions of multiple indexes can enhance the prediction accuracy of HKGs. We transform the nine indexes for all the nodes in HPIN as fractional ranks. Suppose the nine indexes for the n nodes consist of the matrix $X = (X_1, X_2, ..., X_9)$. For X_i, if n_1 nodes are with the highest value, then the fractional ranks for the n_1 nodes are n_1/n; for the subsequent n_2 nodes are with the second largest value, their fractional ranks are $(n_1+n_2)/n$. For the n_k nodes with the $k'th$ largest value, their fractional ranks are $(n_1+n_2+...+n_k)/n$. Obviously, for the nodes with the lowest value in X_i, their fractional ranks are 1. Denote the corresponding fractional ranks of matrix X as X^*, then the values in X^* must be in the interval $(0, 1]$. Based on all the possible compositions of the nine columns in X^*, we consider two cases. Firstly, given a composition with j indexes, for each node, we take the minimum of the j indexes as a composite index, and compare the classification of genes according to this new index with the gold standard. For the second case, we consider the maximum of the j indexes as a composite index. For simplicity, we call them as the MIN case and MAX case. For the nine indexes, there are totally 511 compositions, which include indexes ranging from one to nine. For each composition, we obtain the AUC of ROC and the corresponding highest acc. Fig.5 shows the AUC and acc for all the 511 compositions under the two cases.

From Fig.5, we can see that the AUC for most of the compositions under the MAX case are above 0.8, and the corresponding acc are all around 0.8, which indicate the MAX case is more effective in predicting HKGs than the MIN case. Moreover, under both cases, with the increasing of indexes included in the composite index, less AUC can achieve 0.85, and less acc values are higher than 0.8. This indicates that with more indexes considered, the prediction abilities can not be effectively enhanced, but the differences on the prediction abilities of different compositions tend to be smaller. The nine hallow circles on the left of each panel correspond to the nine single indexes, where the closeness, semi-local and eigenvector centralities are with almost the highest AUC and acc. Therefore, Fig.5 also indicates that some single indexes can well predict the HKGs. Though the compositions of several indexes can not effectively enhance the prediction ability, the AUC and acc for many of the composite indexes can achieve 80%.

IV. Discussions and Conclusions

In this paper, we constructed a large-scale HPIN. Based on the statistical analysis of the HPIN and HKGs, we find the HKGs in the HPIN are characterized with much higher average degree, k-shell, betweeness, clustering coefficient, semi-local centrality, eigenvector centrality, PageRank and motif centrality than the other nodes. Based on ROC analysis, we find the closeness, semi-local and eigenvector centralities are with the highest prediction accuracy.

The average betweeness of the HKGs is more than 7 times larger than the HPIN, but it has very large CV and very small median. Whereas, the closeness is with the smallest CV and very large median, the median is larger than its mean. The prediction accuracy of the betweeness is very low, while the closeness is with the highest prediction accuracy. This indicates that the prediction accuracy of an index is not only related to its mean, but also related to its standard deviation, CV and median. Fig.6 shows the betweeness, clustering coefficient, semi-local centrality and closeness for the 1389 HKGs and 697 TEGs in the HPIN. In order to facilitate the observation, we show these figures in log-log scale. From Fig.6, we can intuitively see that the betweeness and clustering coefficient of the HKGs are very dispersive, while the semi-local centrality and closeness of HKGs tend to be clustered, and can be easily distinguished from the TEGs.

The related investigations facilitate the identification of HKGs in the HPIN. For example, by setting the threshold value $T = 0.6$, from fractional ranks of the closeness X_8^* for

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

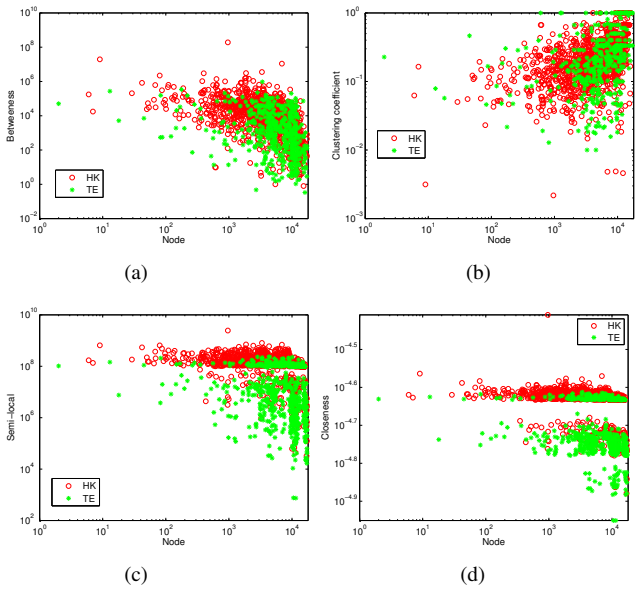


Fig. 6. Indexes for the 1389 HKGs and 697 TEGs in the HPIN. (a) Betweeness. (b) Clustering coefficient. (c) Semi-local centrality. (d) Closeness.

the rest 15225 nodes in HPIN, we can further identify many HKGs, and the accuracy will be about 82.36%. It is noted that, we have only considered nine statistical indexes of the HPIN, one can extend the related discussions to some other indexes. It is also noted that, for the composite indexes, we only consider two cases. It is intriguing to consider some other composite indexes, such the indexes based on the principle component analysis [22], [30]. We will discuss these problems in our future works. The investigations shed some lights on the characterization and identification of human functional genes, which have potential implications in systems biology and networked medicine [31].

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Major Project of China under Grants 2014ZX10004-001-014; the National Natural Science Foundation of China under Grants 61304151,11172215,61174028 and 11105040, the Australia ARC Discovery Grants DP130104765, the Science Foundation of Henan University under Grants 2012YBZR007 and 2013YBRW005.

REFERENCES

[1] S. Peri, J.D. Navarro, R. Amanchy, et al., "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Res.*, vol.13, no.10, pp.2363-2371, Oct. 2003.

[2] C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol.34, pp.D535-D539, Ian. 2006.

[3] K. R. Brown and I. Jurisica, "Online predicted human interaction database," *Bioinformatics*, vol.21, no.9, pp.2076-2082, Jan. 2005.

[4] X. She, C.A. Rohl, J.C. Castle, A.V. Kulkarni, J.M. Johnson and R. Chen, "Definition, conservation and epigenetics of housekeeping and tissue-enriched genes," *BMC Genomics*, vol.10, art.no. 269, June 2009.

[5] J. Zhu, F. He, S. Song, J. Wang and J. Yu, "How many human genes can be defined as housekeeping with current expression data? " *BMC Genomics*, vol.9, art.no. 172, April 2008.

[6] A.J. Butte, V.J. Dzau, S.B. Glueck, "Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues," *Physiol Genomics*, vol.7, no.2, pp. 95-96, Dec. 2001.

[7] L. Zhang, W. Li, "Mammalian housekeeping genes evolve more slowly than tissue-specific genes," *Mol. Biol. Evol.*, vol.21, no.2, pp.236-239, Feb. 2004.

[8] H.J. de Jonge, R.S. Fehrmann, E.S. de Bont, et al., "Evidence based selection of housekeeping genes," *PLoS One*, vol. 2, no.9, art.no. e898, Sep. 2007.

[9] J.E. Dickerson, A. Zhu, D.L. Robertson and K.E. Hentges, "Defining the role of essential genes in human disease," *PLoS One*, vol.6, no.11, art.no.e27368, Nov.2011.

[10] J. Rual, K. Venkatesan, T. Hao, et al., "Towards a proteome-scale map of the human proteinCprotein interaction network," *Nature*, vol. 437, pp.1173-1178, Oct. 2005.

[11] U. Stelzl, U. Worm, M. Lalowski, et al., "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol.122, no. 6, pp.957-968, Sep. 2005.

[12] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics*, vol.22,no.22,pp.2800-2805, 2006.

[13] A.L. Barabási, N. Gulbahce and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nat. Rev.*, vol.12, pp.56-68, Jan.2011.

[14] M.P.H. Stumpf, C. Wiuf and R.M. May, "Subnets of scale-free networks are not scale-free: sampling properties of networks," *Proc. Natl. Acad. Sci. USA*, vol.102, no.12, pp. 4221-4224, 2005.

[15] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol.45, no. 2, pp. 167-256, 2003.

[16] P. Wang, C. Tian and J. Lu, "Identifying influential spreaders in artificial complex networks," *J. Syst. Sci. Complex.*, vol.27, pp. 650-665, 2014.

[17] M. D. Humphries and K. Gurney, "Network 'small-world-ness': a quantitative method for determining canonical network equivalence," *PLoS One*, vol.3, no.4, art.no. e0002051, 2008,

[18] M. E. J. Newman, S. H. Strogatz and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Phys. Rev. E*, vol.64, art.no. 026118, 2001.

[19] E. Ravasz and A. L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, art.no. 026112, 2003.

[20] D. J. Watts and S. H. Strogatz,"Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-442, 1998.

[21] X. Wan, S. Cai, J. Zhou and Z. Liu,"Emergence of modularity and disassortativity in protein-protein interaction networks," *Chaos*, vol.20, art.no. 045113, 2010.

[22] P. Wang, X. Yu and J. Lü, "Identification and evolution of structurally dominant nodes in protein-protein interaction networks," *IEEE Trans. Biomed. Circuits Syst.*, vol.8, no.1, pp.87-97, Feb. 2014.

[23] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no.2, pp.163-177, 2001.

[24] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt and E. Shir, "A model of Internet topology using k-shell decomposition," *Proc. Natl. Acad. Sci. USA*, vol.104, pp. 11150-11154, 2007.

[25] D. Chen, L. Lü, M. Shang and T. Zhou, "Identifying influential nodes in complex networks," *Physica A*, vol. 391, pp. 1777-1787, 2012.

[26] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Networks ISDN Syst.*, vol. 30, pp. 107-117, 1998.

[27] D. Koschützki, H. Schwöbbermeyer and F. Schreiber, "Ranking of network elements based on functional substructures," *J. Theor. Biol.*, vol. 248, pp. 471-479, 2007.

[28] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol.27, pp. 861C874, 2006.

[29] P. Wang, J. Lü and X. Yu, "Identification of important nodes in directed biological networks: a network motif approach," *PLoS One*, vol. 9, no.8, art.no. e106132, 2014.

[30] W.K. Härdle and L. Simar, *Applied multivariate statistical analysis*, Springer-Verlag, Berlin Heidelberg, 2012.

[31] P. Wang and J. Lü, "Control of genetic regulatory networks: opportunities and challenges," *Acta Automatica Sinica*, vol.39, no.12, pp.1969-1979, 2013.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE