Lumping Protein Complexes to Reduce the Complexity of Human Protein-Protein Interaction Network

Protein Komplekslerini Gruplayarak İnsanda Protein-Protein Etkileşim Ağlarının Karmaşıklığını Azaltmak

Muhammed Erkan Karabekmez

Istanbul Medeniyet University, Department of Bioengineering

Abstract

Aim: For network biology applications in health sciences, most of the time it is impossible to handle the whole protein-protein interaction map at once, and accordingly there have been various approaches to network reduction for computational efficiency. In this study we aimed to use protein complexes as a base for network reduction, proposing a node-lumping procedure.

Materials and Methods: The comprehensive resource of mammalian protein complexes was used to extract protein complex interactions. Human protein-protein interaction map was retrieved from the Agile Protein Interactomes Data (APID) server. A novel lumping procedure was introduced to create a reduced network. Topological analysis of the resulting context-specific network and examination of the highly connected nodes were compared with the original network.

Results: After lumping we obtained a heterogeneous map of 9,888 proteins and 304 lumped nodes with 41,940 interactions. Total number of nodes and interactions were reduced by 9.7% and 16%, respectively. The resulting network preserves the scale-free topology.

Discussion and Conclusion: The results indicated that the procedure was helpful in network reduction without disturbing the biologically relevant structure of the network.

Keywords: computational biology; protein complexes; protein interaction maps

Öz

Amaç: Ağ biyolojisinin sağlık bilimlerindeki uygulamalarında çoğu zaman bütün protein-protein etkileşim haritasını bir kerede ele almak imkansızdır; bu nedenle hesaplama verimliliğini artıran çeşitli ağ küçültme yaklaşımları geliştirilmiştir. Bu çalışmada bir düğüm birleştirme prosedürü önererek, protein komplekslerine dayalı bir ağ küçültme stratejisi kullanmak amaçlanmıştır.

Gereç ve Yöntemler: Protein kompleksi etkileşimlerini çıkarmak için memeli protein komplekslerinin kapsamlı kaynağı kullanıldı. İnsan protein-protein etkileşim haritası *Agile Protein Interactomes Data* (*APID*) sunucusundan alındı. Daha küçük bir ağ oluşturmak için özgün bir gruplama yaklaşımı kullanıldı. Elde edilen bağlama özel ağın ilingesel analizi ve yüksek merkeziliğe sahip düğümlerin incelenmesi, orijinal ağ ile karşılaştırılmalı olarak yapıldı.

Bulgular: Gruplama prosedüründen sonra aralarında 41.940 etkileşimi olan 9.988 protein ve 304 protein grubu içeren heterojen bir etkileşim haritası elde edilmiştir. Toplam düğüm sayısı ve etkileşim sayısı sırasıyla %9,7 ve %16 azalmıştır. Ortaya çıkan ağ, ölçeksiz topolojiyi korumuştur.

Tartışma ve Sonuç: Sonuçlar, yaklaşımın biyolojik olarak anlamlı yapısını bozmadan biyolojik ağı küçültmede işlevsel olduğunu göstermiştir.

Anahtar Sözcükler: hesapsal biyoloji; protein etkileşim haritası; protein kompleksleri

Received/*Geliş* : 18.10.2018
Accepted/*Kabul*: 19.11.2018

DOI: 10.21673/anadoluklin.556987

Corresponding author/*Yazışma yazarı*

Muhammed Erkan Karabekmez
İstanbul Medeniyet Üniversitesi, Kuzey Kampüsü, F-Blok 105, Ünalan, Istanbul, Turkey
E-mail: erkan.karabekmez@medeniyet.edu.tr

ORCID
Muhammed Erkan Karabekmez: 0000-0002-0517-5227

INTRODUCTION

In the post-genomic era rapid advances in molecular biology made both possible and inevitable to evaluate multiple biological entities by using a systematic perspective. "Network biology" emerged as a new subdiscipline to meet this necessity (1). Evaluating biological systems as interacting dynamic systems by using graph theoretical tools widens the research field in life sciences.

The holistic understanding of biological networks can allow researchers to unveil disease pathways and novel drug targets. There is a significant literature around network biology applications in systems biomedicine (2).

However, conventional algorithms in graph theory has high computational complexity as the networks grow. Scientists search relevant network reduction approaches that could reduce computational run-times without disturbing the original structure (3). The need for constructing context-specific networks has also arisen from this aim (4).

There have been some efforts to construct protein complex networks in model organisms (5). The important study of Ruepp et al. compiles mammalian protein complex interactions in a database (6). The work has also shown that the number of reutilization of protein complex subunits in different protein complexes has decreasing frequency, implying that the most of the protein subunits are complex-specific. Co-expression patterns of protein complexes are also investigated for yeast (7). It has been shown that expression levels of complex subunits are slightly more correlated than random pairs and as the sizes of the complexes increase the correlation also increases. It has been shown that even for the protein complexes with low average co-expression correlation among the subunits there can exist tightly correlated sub-complexes (7).

Other co-expression studies on protein complexes unveil that some complexes have a high co-expression pattern under diverse conditions (permanent) and some have a high co-expression pattern only under certain conditions (transient) (8).

All these studies show that protein complexes have strong functional associations between their subunits, other than physical interactions. Therefore, they could be useful as a network reduction base. In this study, human protein-protein interaction map (PPI) was constructed by using interaction data retrieved from the APID database and the comprehensive resource of mammalian protein complexes (CORUM) was used to extract protein complex data. Protein complex subunits are used to reduce the size of the global PPI map without contextualizing the map.

MATERIALS AND METHODS

The comprehensive resource of mammalian protein complexes (CORUM) was used to retrieve protein complex membership data (6) (2.7.2017 CORUM Release).

The APID database (9) that unifies PPIs from primary databases of molecular interactions –BioGRID (10), DIP (11), HPRD (12), IntAct (13), MINT (14)– was used to construct human PPI (downloaded on 5.25.2018).

The MATLAB R2018b was used for lumping procedures and topological analysis of the resulting networks. First, human PPI map and complex membership network were introduced as adjacency matrices. Secondly, redundancies of subunits were calculated. Then, using these data potential groups of subunits to lump were identified. Finally, identified groups were lumped and new indices were saved.

No ethics committee approval was required for this study.

RESULTS

The comprehensive resource of mammalian protein complexes (CORUM) covers 2,126 non-redundant hetero-dimeric protein complexes constituted of 9,802 subunits (2.7.2017 CORUM Release) excluding 89 homo-dimeric complexes. Cumulatively, 3,193 different proteins act as subunits in at least one protein complex (6).

The average protein complex size was calculated as 4.6 and each subunit was found to appear 3.07 times in different protein complexes on average.

Redundancy versus frequency analysis in the original CORUM paper (6) was repeated with the current version of the database and it has been shown that 1400 proteins exclusively appear in only one protein

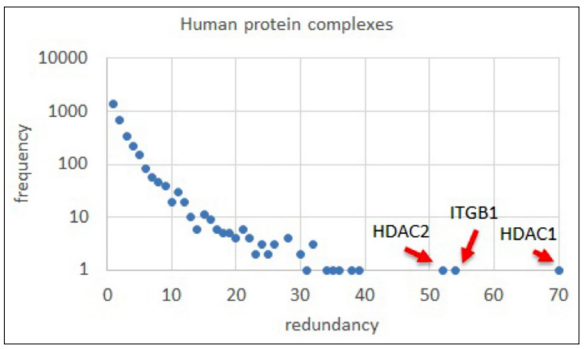

Figure 1. Redundancy vs frequency distribution of the human protein complexes

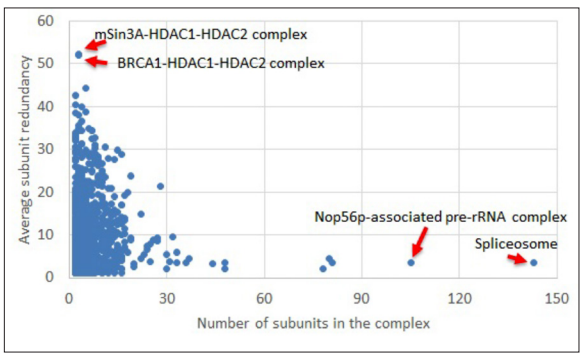

Figure 2. Average subunit redundancy vs the size of the human protein complexes

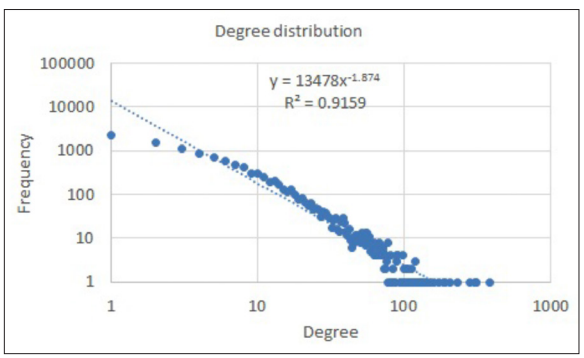

Figure 3. Degree distribution of the original human PPI network

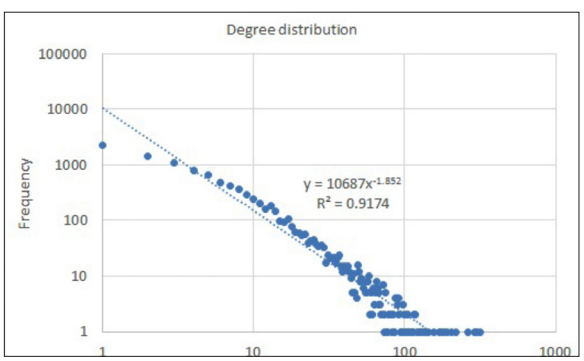

Figure 4. Degree distribution of the reduced human PPI network

complex while 695 and 344 subunits appear in two and three different protein complexes, respectively (Figure 1).

The histone deacetylase 1 (HDAC1) has the highest redundancy with appearance in 70 different protein complexes. The integrin subunit beta 1 (ITGB1) and the histone deacetylase 2 (HDAC2) follow with 54 and 52 appearances, respectively (Figure 1).

The average redundancy of subunits was analyzed with respect to the number of subunits. It was found that mSin3A-HDAC1-HDAC2 and BRCA1-HDAC1-HDAC2 protein complexes with three subunits had the highest average subunit redundancies, not surprisingly (Figure 2).

Totally 124 protein complexes, constituted of 386 novel subunits, had an average subunit redundancy of one. Initially it was decided to merge only 124 protein complexes not to cause structural deformations. The benefit seemed not to be enough. In order to go further, exclusive sub-complexes and sub-complexes that always appear together with at least two subunits were also investigated. It was revealed that, in addition to the 124 protein complexes with exclusive subunits, there were 292 protein sub-complexes made up of 1,067 subunits, either exclusive or always appearing together. For instance, a sub-complex with 48 subunits appears exclusively in 55S and 39S mitochondrial ribosomal protein complexes (Table 1).

As a result, a total of 1,453 protein subunits were decided to be lumped into 416 protein complexes/sub-complexes and the lumping was conducted in MATLAB.

The human PPI network was downloaded from APID (9) with interactions proven by at least two publications. A map of 11,199 proteins with 50,138 interactions was obtained. Degree distribution of this network was investigated and scale-free topology was confirmed (Figure 3).

In this map 248 protein complex subunits obtained from the CORUM database have no interaction while the remaining 2,945 subunits have 17,571 interactions. After eliminating the potential complexes or sub-complexes for lumping without enough interactions in this map the remaining 1,311 subunits were merged into 304 groups.

After lumping we obtained a heterogeneous map

Table 1. Numbers of exclusive or co-existing sub-complexes in human protein complexes

Size of complex /sub-complex	Number of non-redundant complexes	Number of non-redundant sub-complexes	Total number of complexes	Total number of subunits
2	65	170	235	470
3	35	53	88	264
4	11	19	30	120
5	6	18	24	120
6		8	8	48
7	1	4	5	35
8	1	2	3	24
9		3	3	27
10	2	3	5	50
11		1	1	11
12	1		1	12
13		3	3	39
14	1		1	14
15		1	1	15
16	1	1	2	32
18		1	1	18
20		1	1	20
21		1	1	21
30		1	1	30
35		1	1	35
48		1	1	48
TOTAL:	**124**	**292**	**416**	**1453**

of 9,888 proteins and 304 lumped nodes with 41,940 interactions; 78 nodes lost their all interactions and a total of 10,114 nodes remained. Total number of nodes and interactions were reduced by 9.7% and 16%, respectively. The resulting network preserves the scale-free topology (Figure 4).

Hubs with the highest connectivity were also investigated and no significant change was observed.

DISCUSSION AND CONCLUSION

We introduced a strategy to merge proteins that always exist in protein complexes together to simplify the human PPI network to save from computational cost and to highlight pathway-like patterns of the network compared to clique-like patterns. The scale-free, biologically relevant, structure of the network was preserved.

This procedure is also applicable to the drug target studies where small molecules target only a subunit of a complex. By updating drug-polypeptide interactions with lumped network indices these interactions can be preserved in further analysis with the context-specific lumped network.

Our results in the human PPI network show a 16% saving from the number of edges in the network, which corresponds to, for instance, a run-time saving of 31.2%, considering the best algorithm to calculate betweenness centralities with a complexity of O(n^2 log(n)) (15).

When the speed of an algorithm becomes an issue, there are two directions in network biology—through either faster algorithms with approximate solutions or context-specific networks. When approximation or contextualization are not preferable this approach could be useful. Whenever multi-omics data are avail-

Table 2. Change in hubs by the lumping procedure

Hubs of reduced human PPI map		Hubs of original human PPI map	
Genes/Complexes/Sub-complexes	**Degree**	**Genes**	**Degree**
VCAM1	318	VCAM1	384
APP	302	TP53	312
EGFR	295	APP	310
TP53	290	EGFR	304
UBC	263	UBC	281
HDAC1	220	HDAC1	233
HSP90AA1	203	HSP90AA1	204
Nop56p-associated pre-rRNA complex (24)	192	ESR1	189
GRB2	185	GRB2	188
ESR1	179	EP300	174
EP300	172	YWHAZ	159
Spliceosome (32)	170	YWHAG	151
YWHAZ	156	MYC	142
YWHAG	143	BRCA1	136
MYC	138	HNRNPA1	132

able, further lumping co-expressed transient sub-complexes can also simplify the picture.

Comparative studies on topological centrality and biological centrality are also popular in the literature (16). This strategy also provides a novel insight to the centrality concept as new hubs appearing in the hub lists as complexes/sub-complexes add even deeper insight to the PPI network that was hidden before the lumping procedure.

In this study lumping protein complexes/sub-complexes was suggested to reduce the complexity of human protein-protein interaction network in order to speed up the network biology applications in medical researches. The results showed that the approach was useful in building context-specific sub-networks and increasing the efficiency of network algorithms.

Statement of Conflict of Interest

The author has no conflict of interest to declare.

REFERENCES

1. Barabasi A L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature Rev Genet. 2004;5(2):101.
2. Sevimoglu T, Arga KY. The role of protein interaction networks in systems biomedicine. Comput Struct Biotechnol J. 2014;11(18):22–7.
3. Ackermann J, Einloft J, Nöthen J, Koch I. Reduction techniques for network validation in systems biology. J Theor Biol. 2012;315:71–80.
4. Vlassis N, Pacheco MP, Sauter T. Fast reconstruction of compact context-specific metabolic network models. PLoS Comput Biol. 2014;10(1):e1003424.
5. Butland G, Peregrín-Alvarez JM, Li J, Yang W, Yang X, Canadien V, et al. Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature. 2005;433(7025):531.
6. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. 2009;38(suppl. 1):D497–D501.
7. Liu CT, Yuan S, Li KC. Patterns of co-expression for protein complexes by size in Saccharomyces cerevisiae. Nucleic Acids Res. 2008;37(2):526–32.
8. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res. 2002;12(1):37–46.
9. Alonso-Lopez D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res. 2016;44(W1):W529–W535.
10. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID

interaction database: 2015 update. Nucleic Acids Res. 2014;43(D1):D470–D478.
11. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Res. 2004;32(suppl. 1):D449–D451.
12. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database—2009 update. Nucleic Acids Res. 2008;37(suppl. 1):D767–D772.
13. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2011;40(D1):D841–D846.
14. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2011;40(D1):D857–D861.
15. Brandes UA. Faster algorithm for betweenness centrality. J Math Sociol. 2001;25(2):163–77.
16. Karabekmez ME, Kirdar B. A novel topological centrality measure capturing biologically important proteins. Mol Biosyst. 2016;12(2):666–73.