DNA Research **17**, 261–270, (2010) doi:10.1093/dnares/dsq019
Advance Access Publication: 9 September 2010

Protein Connectivity and Protein Complexity Promotes Human Gene Duplicability in a Mutually Exclusive Manner

Tanusree Bhattacharya, and Tapash Chandra Ghosh*

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

*To whom correspondence should be addressed. Tel. +91 33 23556626. Fax. +91 33 23553886.
E-mail: tapash@boseinst.ernet.in

Edited by Osamu Ohara
(Received 28 January 2010; accepted 26 July 2010)

Abstract

It has previously been reported that protein complexity (i.e. number of subunits in a protein complex) is negatively correlated to gene duplicability in yeast as well as in humans. However, unlike in yeast, protein connectivity in a protein–protein interaction network has a positive correlation with gene duplicability in human genes. In the present study, we have analyzed 1732 human and 1269 yeast proteins that are present both in a protein–protein interaction network as well as in a protein complex network. In the human case, we observed that both protein connectivity and protein complexity complement each other in a mutually exclusive manner over gene duplicability in a positive direction. Analysis of human haploinsufficient proteins and large protein complexes (complex size >10) shows that when protein connectivity does not have any direct association with gene duplicability, there exists a positive correlation between gene duplicability and protein complexity. The same trend, however, is not found in case of yeast, where both protein connectivity and protein complexity independently guide gene duplicability in the negative direction. We conclude that the higher rate of duplication of human genes may be attributed to organismal complexity either by increasing connectivity in the protein–protein interaction network or by increasing protein complexity.

Key words: protein–protein interaction network; protein complexity; haploinsufficient; organismal complexity; human

1. Introduction

Gene duplication is the primary mechanism for generating new genes and biological processes that facilitates complex organisms to evolve more rapidly than the primitive ones.[1–3] In general, eukaryotic genomes contain a large fraction of gene duplicates not only from the single gene or segmental duplications but also from a whole genome duplication (WGD) event. *Saccharomyces cereviceae* is the major example where WGD has occurred ~100 million years ago.[4,5] Genomic instability and massive gene loss promptly followed WGD and purged most of the newly formed gene copies from the yeast genome, retaining ~10% of them.[4] In case of vertebrates, half of all duplicated genes have been maintained.[6] The analysis of the human genome has shown that most of the human genes are duplicated.[7] Nevertheless, what exactly guides gene duplicability for a particular organism still remains unclear. Two of the most well-known factors that guide gene duplication are (i) protein connectivity, defined as the number of links that a protein node has to other nodes in the protein–protein interaction network and (ii) protein complexity, defined as the number of subunits in a protein complex.[8,9]

It has been previously found that protein connectivity is negatively correlated with gene duplicability

(generally defined as the number of paralogs per gene family) in the case of yeast. That is, highly connected proteins (hubs) have a lower rate of duplication compared with lowly connected proteins (non-hub).[10] However, this trend is opposite in the human case, where protein connectivity and gene duplicability are positively correlated.[8] The reason behind this difference is that due to higher protein dosage in yeast, retention of gene duplicates might have a deleterious effect on the organism. But, in the case of mammals, due to a higher rate of diversification in the function of the gene duplicates, it becomes more robust against dosage increase.[11,12]

Protein complexity might also be another important factor that influences gene duplicability. According to their fates, duplicated genes found within protein complex networks belong to one of the three categories: (i) intra-complex paralogs that remain within the same protein complex, (ii) bi-complex paralog when resulting genes function within two separate complexes and (iii) overhang, when the resulting genes posses no general association with a known protein complex.[13] Duplication of a subunit in a protein complex might also cause dosage imbalance if rapid sub-functionalization or neo-functionalization does not occur to the newly arisen genes.[14] Therefore, duplication of a subunit of a protein complex is less likely to be successful than the duplication of a monomer.[14] The proportion of unduplicated genes is high ($\geq$65%) for both monomers and multimers in yeast, whereas it is less ($\leq$30%) in humans[9] since complex organisms are more robust against dosage increase. Detailed studies on yeast protein complexes showed that most of the participating subunits of multiprotein complexes are under tightly regulated gene dosage. These proteins are encoded mainly by haploinsufficient (wild type recessive) genes.[14] This is in agreement with the central prediction of the physiological theory which states that, genes encoding proteins whose functions tend to be insensitive to protein dosage should typically be haplosufficient (dominant wild type).[15] Eventually, human haploinsufficient genes have, on the average, more paralogs than haplosufficient ones because additional products of these genes probably lead to increased fitness.[15]

Recently it has been observed that protein connectivity and protein complexity complement each other in guiding the evolutionary rates of human proteins.[16] Moreover, it has been reported that gene duplicability decreases with the increase in the number of subunits in protein complex, though this rate of decrease is fairly slow for humans compared with yeast.[8,9] In the present study, we investigated the interrelationship among the three features viz. gene duplicability, protein connectivity and protein complexity, in proteins that are present both in protein complexes network as well as in the protein–protein interaction network. Our analysis reveals that the human protein–protein interaction network has a larger influence on a gene to be duplicated. Moreover, we found that non-hub proteins in the complex are duplicated as well as hub proteins, indicating that protein complexity can also increase gene duplicability, in contrast to previous reports.[9] Analyses of haploinsufficient proteins of both humans and yeast suggest that organismal complexity (defined as a measure of the number of different cell types in an organism), as well as the higher rate of functional divergence of human protein complexes, has a large influence on higher duplicability.

2. Materials and methods

2.1. *Protein–protein interaction data*

All human protein–protein interaction data were taken from the Human Protein Reference Database (HPRD) release 7 (http://www.hprd.org/download), which is essentially a database covering thousands of protein–protein interactions, posttranslational modifications, enzyme/substrates relationships, disease associations, tissue expression and subcellular localization. The data were extracted from the literature manually by biologists who read and interpreted $>$300 000 published articles during the annotation process[17] where a total of 9386 unique protein interactors were found. Proteins with more than five interactions were considered as hub proteins and proteins with one or two interacting partners were considered as non-hub proteins.[16,18] All yeast protein–protein interaction data were collected from Li *et al.*,[10] where protein–protein interaction pairs were collected from various high throughput experiments and databases. This collection was combined with another high throughput dataset of Bader *et al.*[19] Subsequently, many small-scale experiments were performed to obtain a larger high confidence dataset.[10] This high confidence dataset (http://mbe.oxfordjournals.org/content/vol0/issue2005/images/data/msi249/DC1/msi249supp2.txt) is taken for our analysis of yeast proteins. Finally, we considered 1732 human (Supplementary Table S1) and 1269 yeast proteins (Supplementary Table S2) for which both protein–protein interaction data as well as protein complex data are available.

2.2. *Identification of protein complexes*

We obtained the list of human protein complexes from http://mips.gsf.de/genre/proj/corum. A total of 1345 number of protein complexes[20] were found (Supplementary Table S3). The protein complex data for yeast was collected from Gavin *et al.*[21]

(http://www.nature.com/nature/journal/v440/n7084/extref/nature04532-s3.pdf; Supplementary Table S4).

The protein complexes having >10 subunits were classified as large complexes and complexes having less than 10 subunits were considered as small complex.

2.3. Identification of haploinsufficient genes

Human haploinsufficient gene sequences were taken from Dang *et al.*[22] where haploinsufficient genes have mainly been collected from PubMed and OMIM search results with one of the major keyword being haploinsufficient and spurious hits were gradually removed by comparing against known human gene names and symbols[22] (http://www.nature.com/ejhg/journal/v16/n11/suppinfo/ejhg2008111s1.html?url=/ejhg/journal/v16/n11/abs/ejhg2008111a.html, Yeast haploinsufficient genes are collected from Deutschbauer *et al.*[23] (http://www.genetics.org/cgi/content/full/genetics.104.036871/DC1). These haploinsufficient proteins are generally collected from yeast deletion strain experiments on YPD and minimal media.

2.4. Identification of paralogs

The paralogs for both humans and yeast were taken from ensemble paralog database version 52. The general steps followed by the ensemble[24] for identifying paralogs sequences are based mainly on the construction of the gene tree reconciled with the species tree formed by the cluster of aligned sequences obtained from BLASTP. True paralogy for every pair of genes in the gene tree has been inferred by calculating the d_N/d_S ratio. This method not only enables the detection of true paralogs but also helps discard misleaded paralogs obtained from domain fusion. Finally, true paralog sets in humans were downloaded from the ensemble using 30% similarity between two sequences, and the alignable region between two sequences is >80% of the longer protein.[19] Thus, a total set of 9609 number of duplicated proteins in humans and 1604 duplicated proteins in yeast were identified. Among these 9609 human paralogs only 664 genes (Supplementary Table S5) having paralogs are present both in the protein–protein interaction network as well as protein complex, network and in the case of yeast, 182 genes (Supplementary Table S6) having paralog have been identified, present in both networks.

2.5. Statistical analysis

All statistical analyses were performed by the software SPSS and TANAGRA. We measured Spearman rank correlation for both bivariate and partial correlations, as our data values were mostly repetitive, and Spearman rank correlations have been calculated after finding rank of the values. As a result, extreme variations in values have less control over the correlation. Simple bivariate correlation may be biased due to the presence of other factors and therefore Spearman's partial correlation coefficient was calculated after eliminating the confounded factors. Additionally, we performed Principal Component Analysis (PCA) through SPSS.

3. Results

3.1. Protein connectivity, protein complexity and gene duplicability

In order to examine how the gene duplicability in humans as well as in yeast are influenced by protein connectivity and protein complexity, we collected 1732 human and 1269 yeast proteins for which both protein–protein interaction data and protein complex annotation were available. Non-parametric Spearman's correlations for both protein connectivity and protein complexity with the gene duplicability of human genes reveal that, gene duplicability is negatively correlated (Spearman rank test; $R = -0.111$, $P < 0.001$) with protein complexity and positively correlated (Spearman rank test; $R = 0.169$, $P < 0.001$) with protein connectivity. In order to examine whether these two factors independently influence gene duplicability, we computed Spearman's partial correlation analysis between gene duplicability and one of the two factors by controlling the other. We observed that correlation between protein complexity and gene duplicability diminished drastically when protein connectivity was kept controlled ($\rho = -0.080$, $P < 0.001$), whereas positive correlation existed between protein connectivity and gene duplicability when protein complexity was controlled ($\rho = 0.150$, $P < 0.001$). The same trend was also observed by computing the Pearson partial correlation coefficient which showed that protein complexity did not have any correlation with gene duplicability when protein connectivity was controlled ($\rho = 0.035$, $P = 7.6 \times 10^{-2}$), but the effect of protein connectivity remained when protein complexity was controlled ($\rho = 0.158$, $P < 0.001$). These results suggest that protein connectivity is the more important factor for gene duplicability in humans, although the effect of protein complexity over gene duplicability is guided mainly by protein connectivity. Moreover, a significant negative correlation ($R = -0.201$, $P < 0.001$) was found between protein connectivity and protein complexity. This correlation holds even when gene duplicability is controlled ($\rho = -0.186$, $P < 0.001$).

To test the independence of the sample points we removed those paralogs which are present in the protein complex and performed the same analysis. We obtained the same trend, that is, gene duplicability is negatively correlated with protein complexity ($R = -0.121$, $P < 0.001$) whereas it is positively correlated with protein connectivity ($R = 0.065$, $P = 4.7 \times 10^{-2}$) and protein connectivity and protein complexity are negatively correlated with each other ($R = -0.225$, $P < 0.001$). This indicates that shared gene duplicability does not have any bias on our analysis.

All the above analyses have been applied in the case of yeast genes. Non-parametric Spearman's correlation between protein complexity and gene duplicability revealed that gene duplicability is negatively correlated ($R = -0.236$, $P < 0.001$) with protein complexity. The same trend (though with reduced correlation value) was observed when Spearman's correlation was determined between protein connectivity and gene duplicability ($R = -0.074$, $P = 2.8 \times 10^{-2}$). However, there is no significant correlation between protein connectivity and protein complexity. Spearman's partial correlation analysis demonstrates that both protein connectivity ($\rho = -0.077$, $P = 2.37 \times 10^{-3}$) and protein complexity ($\rho = -0.237$, $P < 0.001$) affect gene duplicability independently. This has been further supported by Pearson partial correlation which also finds independent correlation for protein connectivity ($\rho = -0.136$, $P < 0.001$) as well as for protein complexity ($\rho = -0.172$, $P < 0.001$).

In order to evaluate how the position of a protein in the protein–protein interaction network influences gene duplicability, we correlated gene duplicability with the protein's centrality which is measured in terms of closeness (average number of nodes connecting a protein to all other proteins) and betweenness (measures the frequency with which a node lies on the shortest path between all other nodes) of a protein in the protein–protein interaction network.[26] In both the human and yeast genes, we found that protein connectivity is positively correlated with both these parameters ($R = 0.791$, $P < 0.001$) and ($R = 0.901$, $P < 0.001$), respectively, in case of humans and ($R = 0.779$, $P < 0.001$) and ($R = 0.560$, $P < 0.001$), respectively, in case of yeast. These results indicate that the relationship between protein connectivity and gene duplicability is in the same direction as that of protein's centrality, that is, closeness and betweenness ($R = 0.220$, $P < 0.001$) and ($R = 0.160$, $P < 0.001$), respectively, in case of humans and ($R = -0.118$, $P = 9 \times 10^{-3}$, $R = -0.183$, $P < 0.001$), respectively, in case of yeast.

At this point, it is worthwhile to mention that the protein–protein interaction data come from a variety of methods and the percentage of interactions determined by various methods may differ between yeasts and human. Therefore, it is important to make sure that the differing trends are not simply due to different types of methods used in determining protein connectivity in protein–protein interaction network. In order to test this we have collected all the 7328 interaction data of humans as well as 4127 interaction data of yeast from yeast two-hybrid experiments of which 561 human proteins and 1032 yeast proteins are part of both protein complex and the protein–protein interaction network. We found the same trend when we performed our analysis on genes from yeast two-hybrid experiments (Supplementary Table S7).

Moreover, it has been reported that human protein complexes are more robust against dosage than yeast protein complexes due to their high organismal complexity. Therefore, we have investigated the effect of gene dosage over human and yeast protein complexes.

3.2. *Combinatorial effects of protein dosage, gene duplicability, protein connectivity and protein complexity*

The role of protein complex network on gene duplication can only be estimated by evaluating the known dose sensitive complex proteins, i.e. the haploinsufficient proteins[27] present in the protein complex. Haploinsufficiency occurs when a diploid organism only has a single functional copy of a gene (with the other copy inactivated by mutation) and the single functional copy of the gene does not produce enough gene product (typically a protein) to bring about a wild type condition. Studies show that mutation or loss of a single allele may be sufficient to exert diseased cellular phenotypes.[28] This gene dosage effect results in haploinsufficiency. Thus, haploinsufficient gene needs both of its alleles to be functional in order to express the wild type.

Analyzing 299 human haploinsufficint genes obtained from Dang *et al*.[22] it has been found that 59 proteins are present in both the protein complex and the protein–protein interaction network, of which 21 (36%) are duplicated. A detailed study of these proteins shows that protein connectivity is not at all correlated with gene duplicability whereas protein complexity has a positive correlation ($R = 0.228$, $P = 3.5 \times 10^{-2}$) with gene duplicability. Again we have also collected all the haploinsufficient data for humans from the OMIM to get a larger dataset. In that dataset where we have collected a total of 410 human haploinsufficient proteins of which 32 duplicated haploinsufficient proteins that are present in both the protein complex as well as in

the protein–protein interaction network among the 83 (~39%) proteins present in both protein–protein interaction as well as protein complex network. In this dataset also we have obtained the same result as the previous, that is, protein complexity is positively correlated ($R = 0.213$, $P = 2.18 \times 10^{-2}$) whereas no correlation exists between protein connectivity and gene duplicability ($R = 0.008$, $P = 9.24 \times 10^{-1}$). This trend has also been found in the protein–protein interaction dataset verified by the yeast two-hybrid experimental method where protein complexity is also positively correlated with gene duplicability ($R = 0.247$, $P = 2.7 \times 10^{-2}$). However, the trend is the opposite in the case of *Saccharomyces cerevisiae*, where out of 184 haploinsufficient proteins collected from Deutschbauer *et al.*,[23] 79 are found to be present in both the protein–protein interaction network as well as in the protein complex, of which only 12 (15%) are duplicated. After analysis it has been found that protein complexity is negatively correlated ($R = -0.183$, $P = 2.8 \times 10^{-2}$) with gene duplicability whereas protein connectivity is not correlated with gene duplicability ($R = 0.0661$, $P = 4.32 \times 10^{-1}$). The same trend has been observed for protein–protein interaction data obtained by the yeast two-hybrid experiment case of protein complexity ($R = -0.154$, $P = 3.8 \times 10^{-2}$).

But in both the cases of humans and yeast, the correlation between protein complexity and the number of paralogs per gene exists when connectivity is controlled ($\rho = 0.247$, $P = 1.15 \times 10^{-2}$) for human and ($\rho = -0.1775$, $P = 3.4 \times 10^{-2}$) for yeast whereas, protein connectivity does not have any significant correlation when complexity is controlled ($\rho = 0.1917$, $P = 7.05 \times 10^{-2}$). Partial correlation with our new data also suggest the same trend, that is, protein complexity is positively correlated ($\rho = 0.231$, $P = 1.27 \times 10^{-2}$) and protein connectivity does not have any correlation at all ($\rho = 0.0941$, $P = 3.17 \times 10^{-2}$). Even Pearson partial correlation also shows the positive trend of protein complexity towards gene duplicability even when protein connectivity is controlled ($\rho = 0.068$ $P = 5.34 \times 10^{-1}$). Moreover, a significant negative correlation ($R = -0.269$, $P = 1.2 \times 10^{-2}$) was found between protein connectivity and protein complexity in human haploinsufficient proteins which remains while gene duplicability is controlled ($\rho = -0.307$, $P = 4.16 \times 10^{-3}$). This negative correlation is even increased in the analysis with the new data ($R = -0.361$, $P < 0.001$). The above results from humans suggest that, if haploinsufficient genes are present in the protein complex, protein connectivity is not the most influencing factor, but it helps protein complexity to play. However, the inverse correlation between gene duplicability and protein complexity in yeast haploinsufficient proteins, might be due to their differential organismal complexity between two organisms.[9]

The data from the yeast haploinsufficient proteins supports the 'balance hypotheses' as well as the existence of strong dose sensitivity of this organism.[14]

However, in the analysis of complex proteins without any haploinsufficient genes it has been observed that protein complexity is negatively correlated with gene duplicability in both yeast and humans as observed earlier (Supplementary Table S8). This result suggests that in protein complexes where haploinsufficient genes are absent, the effect of protein connectivity within the protein–protein interaction network is more predominant than that of protein complexity on gene duplication.

3.3. *Effect of protein connectivity, and protein complexity on gene duplicability in humans*

The differences in the correlations of the two different types of connectivities with gene duplicability prompted us to study the differences of additional parameters namely, protein complexity, protein connectivity and gene duplicability in these two groups of proteins (Table 1). From the Table 1 it is evident that there is no significant difference in average protein connectivity as well as in average number of paralogs per gene in haploinsufficient proteins and in proteins where haploinsufficient proteins are absent. However, in one case protein connectivity is positively correlated (where haploinsufficint proteins are absent) and in the other case protein connectivity does not have any correlation (in haploinsufficient proteins present in both the protein–protein interaction network and protein complex network) with gene duplicability. At the same time the differences between complex sizes in these two groups are also apparent (Table 1). But, in both the cases protein connectivity and protein complexity are negatively correlated.

Thus, when we have estimated the average number of paralogs for both highly connected (hub) (445 number of proteins) and lowly connected (non-hub) (96 number of proteins) proteins present in the protein complex (excluding 117 proteins which have an intermediate connectivity between 3 and 5), we found, surprisingly, that both hub and non-hub proteins present in protein complex show nearly equal (Mann–Whitney U-test; $P = 0.128$) average number of paralogs (Fig. 1). Again we have also verified this trend by remodeling hub and non-hub data, that is, taking hub proteins of connectivity >9 and non-hub proteins as connectivity 1 and we have obtained the same result, that is, both hub and non-hub proteins present in the complex show nearly the same gene duplicability (Mann–Whitney U-test; $P = 0.533$). But

Table 1. Connectivity, complexity and average number of paralogs per gene across haploinsufficient proteins and after removal of haploinsufficient proteins in human

	Haploinsufficint proteins present in both the protein–protein interaction network and protein complex network	After removing haploinsufficient proteins from proteins present in both the protein–protein interaction network and protein complex network	Level of significance
Average complexity	8.6511	12.9321	$P < 0.0001$
Average interaction	26.3720	27.2717	$P = 0.261$
Average number of paralogs per gene	1.8720	2.2857	$P = 0.464$

Note. Average complexity means average number of subunits per complex in a particular group of proteins and average interactions indicate average protein connectivity in the protein–protein interaction network of a particular group of proteins.

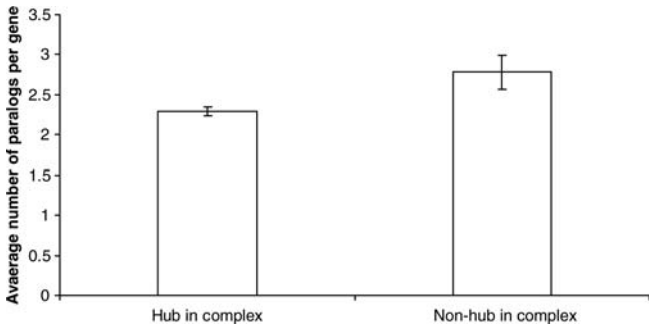

Figure 1. Average number of paralogs per gene of highly connected (hub) and lowly connected (non-hub) protein present in the protein complex.

the significant difference in average number of paralogs between hub and non-hub proteins in both the methods reappear (Mann–Whitney U-test; $P = 0.012$, $P = 0.015$) after removing both of the hub and non-hub proteins present in the protein complex. These results suggest that protein connectivity influences more protein complexity that guides gene duplicability. But when this factor fails to correlate directly, then the other factor, i.e. protein complexity, plays its role.

This explains the higher duplicability of non-hub proteins in the protein complex. In this case since these proteins are poorly connected, the protein's complexity guides the duplicability. Since most of the non-hub proteins are part a of large complex (average complexity 26.75–27) we aimed to analyze what exactly happens to protein connectivity and gene duplicability in large and small protein complexes and, interestingly, we found that there is a significant difference in average connectivity as well as average complex size between the large and small complexes [average connectivity of large complex is 18 and average connectivity of small complex is 32 (Mann–Whitney U-test; $P < 0.001$] and the average complexity of the large complex is 30 and average complexity of the small complex is 5 (Mann–Whitney U-test; $P < 0.001$)). However, there is no significant difference in the paralog number of the large and small complexes (average paralog number of large complex is 2.17 while average paralog number of the small complex is 2.31, respectively). From the above result it is evident that though there is a significant difference in protein connectivity between large and small protein complexes there exist no significant differences in average paralog numbers between these two groups of proteins. So, if protein connectivity is the most effective factor for the duplication of a gene, then the average paralog number of small complexes has to be greater than that of large complexes. But this is not the case. This indicates that there is a significant role of protein complexity over gene duplicability. The results indicate that protein complexity has a positive role in increasing gene duplicability. This result has further been validated by analyzing proteins present in the large complex (complex size is ≥ 10). In the large complex group, we surprisingly found that, both protein connectivity and protein complexity are positively correlated with gene duplicability. These correlations hold true even when both the connectivity and complexity are controlled separately (Table 2). The actual distribution of gene duplicability with respect to protein complexity and protein connectivity in the case of human is shown in Fig. 2.

We used the PCA using a correlation coefficient matrix in order to disentangle the contributions of protein complexity and protein connectivity to gene duplicability. The dominant eigen vectors (taken as equal or >1) that emerge from this analysis can be interpreted as the one of the contributors guiding gene duplication. The first principal component accounts for 59.37% of the total variance and both protein complexity and protein connectivity have equal contributions on this factor.

We have also checked the results for the protein–protein interaction data collected by the yeast two-hybrid experiment and we found that both protein connectivity and protein complexity are positively correlated with gene duplicability ($R = 0.132$, $P < 0.001$; $R = 0.208$, $P < 0.001$), respectively.

Table 2. Correlation between protein connectivity and gene duplicability as well as between gene duplicability and protein complexity in large complexes in human genes

	Spearman rank correlation for Gene duplicability	Level of significance	Partial correlation for gene duplicability with control	Level of significance
Protein complexity	0.150	$P < 0.001$	0.179 (Connectivity control)	$P < 0.001$
Protein connectivity	0.075	$P = 3.30 \times 10^{-2}$	0.124 (Complexity control)	$P < 0.001$

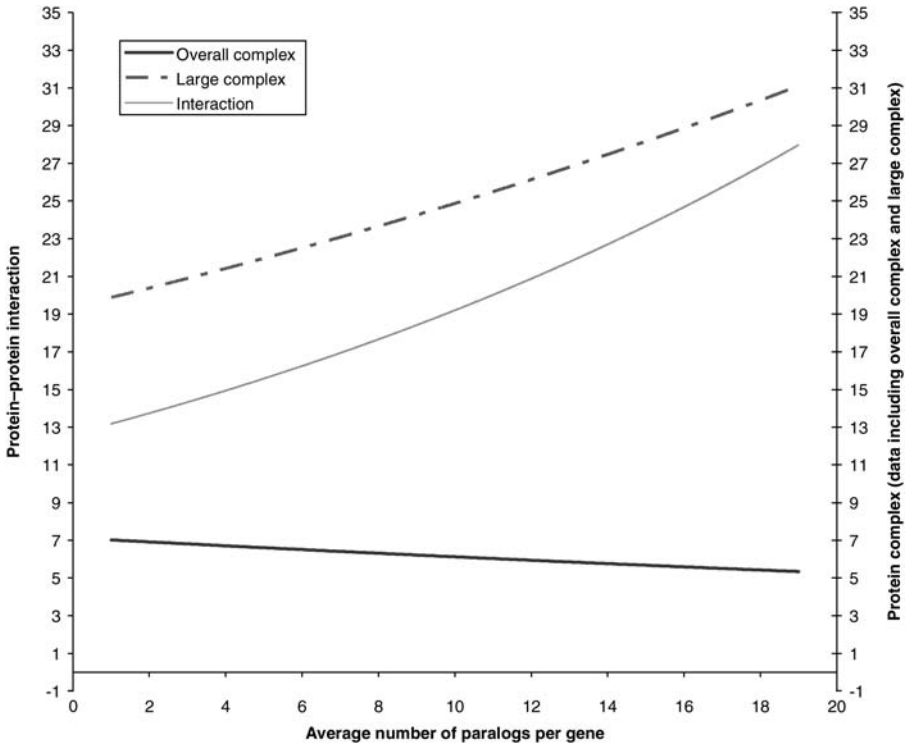

Figure 2. The distribution showing the relation of gene duplicability in overall complex, large complex and protein connectivity in human.

Increase in one component of protein complex can change the stoichiometry of other participating proteins of that complex and hence the net effect may be deleterious for that complex. Therefore, the fixation of the duplicate gene can only happen if rapid sub-functionalization or neo-functionalization can occur in the duplicated genes.

To find what actually happens to the duplicated genes of large complexes, we have analyzed two large complexes separately namely, Splicesome complex and Nop56p-associated pre-rRNA complex. It has been found that for the Spliceosome complex out of 148 subunits only 36 subunits are duplicated, i.e. 24.34% (36/148) subunits are duplicated which gives rise to total 47 unique number of paralogs, of which 18 remain in the complex and most of these proteins are bi-complex paralogs whereas, 14 out of these 18 proteins have a tendency to remain within the same complex. So, due to the duplication 61.70% of overhangs are produced from the Spliceosome complex (Table 3).

In the case of Nop56p-associated pre-rRNA complex the same trend is followed. Out of 104 subunits of the complex 50 subunits are duplicated to give a total 134 unique number of paralogs, of which 14 are retained in the complex (4 retain in same complex). So only 10.44% paralogs are part of the protein complexes and the rest remain as overhang outside the complex (Table 3).

Moreover, analysis of the duplicated genes of the whole human protein complex reveals that most of the duplicated complex proteins are overhangs $(926/1290) \times 100 = 71.78\%$, i.e. they remain outside the complex (Table 3).

This result is just the opposite of yeast, where very few overhangs are produced.[13] This may be the effect of higher rate of functional divergence of duplicate genes in case of humans, which is not so rapid for a simple organism like yeast. Moreover, human protein complexes are more robust against dosage than yeast protein complexes.[9] All these results lead

Table 3. Analysis of human Spliceosome, Nop56p-associated pre-rRNA and overall complex to estimate the percentage of overhang generated in each case

Name	Total subunit	Total subunit duplicated	Number of paralogs produced	Percentage of overhang produced (%)
Spliceosome	148	36	47	61.70
Nop56p-associated pre-rRNA	104	50	134	89.6
Overall complex	2106	720	1290	71.78

us to conclude that organismal complexity is the important factor which causes higher duplicability of mammalian genes sometimes by protein connectivity within protein–protein interaction network or by proteins belongingness to complexes.

4. Discussions

In our earlier studies, while investigating the role of evolutionary rate and intrinsic disorder on protein connectivity in the protein–protein interaction network as well as in protein complex assembly in humans, we have found that both protein connectivity and protein complexity complement each other in a mutually exclusive manner.[16] But, the role of protein connectivity in the protein–protein interaction network as well as in protein complex assembly on gene duplicability was yet to be understood. In earlier studies it has been reported that in the case of humans, gene duplicability is positively correlated with protein connectivity in the protein–protein interaction network, but in the case of yeast gene duplicability is negatively correlated with protein connectivity.[8] These may be the reasons for rapid functional divergence of the duplicate genes in the case of humans, which is not true in a simple organism like yeast as proposed by Yang *et al*. (2003)[9] where they have shown a strong positive correlation between proportion of duplicate gene pairs with the divergence of gene expression and both synonymous and non-synonymous divergence.[14] Higher duplicability of essential genes in the case of humans may promote organismal complexity, as gene essentiality is positively correlated with protein connectivity.[8] But, in the case of yeast, higher duplicability of less-important genes was found.[29] These may cause duplicability to be increased with protein connectivity in case of human. If this phenomenon is true for humans, then protein connectivity within protein complex can also promote higher gene duplicability in humans. We have obtained equal rate of duplication for both hub and non-hub proteins present in protein complex. From the higher duplicability of non-hub proteins we may say that protein's belongingness to a complex may also promote gene duplicability in the positive direction. But till date it has been concluded that the higher the belongingness in a complex, lesser is the duplication. This is true for both humans and yeast. This motivated us to study the combined effect of protein connectivity and protein complexity over gene duplicability in the case of humans. Our study reveals that when combined, although apparently it seems that protein connectivity has a significant positive correlation with gene duplicability and protein complexity has a significant negative correlation with gene duplicability, detailed studies show that protein connectivity has an immense effect on protein complexity and hence acts as a modulating factor for increasing gene duplicability. So when protein connectivity increases, protein complexity decreases, and both of these phenomena will lead to higher duplicability of a gene which is schematically explained further by Fig. 3A. But, eventually what happens to that case where protein connectivity fails to control duplicability directly? That is, in the case of haploinsufficient genes which are

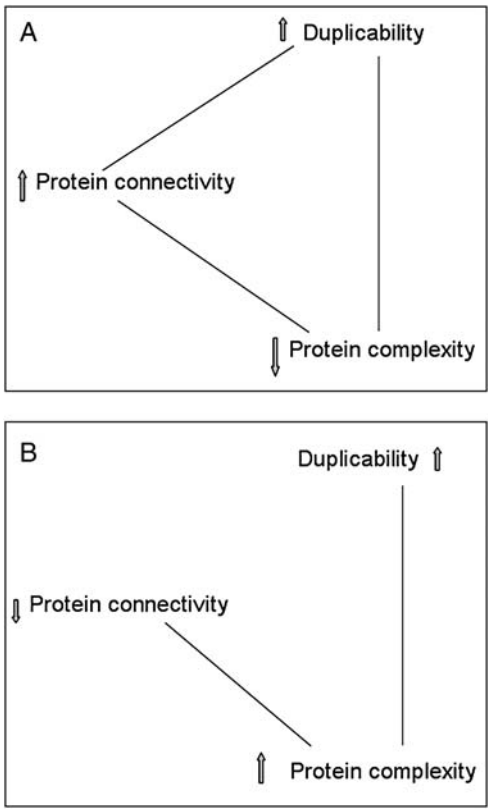

Figure 3. Comparison of relationship between protein connectivity, protein complexity over gene duplicability (A) when both are influencing gene duplication; (B) when only complexity influencing gene duplication.

present in the protein complex. It has been found that in case of human haploinsufficient genes protein complexity is positively correlated with gene duplicability whereas no significant correlation has been found with protein connectivity in the protein–protein interaction network. This signifies when the connectivity within the protein–protein interaction network is absent; then connectivity within the complex plays the same role of connectivity within protein–protein interaction network, which is schematically explained further by Fig. 3B. This positive correlation becomes more prominent when we consider the paralogs which are produced from proteins of large complexes.

This also confirms our earlier findings that non-hub proteins in the complex are as duplicated as hub proteins in the complex. Eventually, we have found that non-hub proteins are mostly part of large complexes. So, this also supports the idea that protein connectivity within the protein complex plays the same role as protein connectivity within the protein–protein interaction network, and the negative correlation enables them to work in a mutually exclusive manner. But, the theory can only be supported if the protein product produced from the participating gene of the complex undergoes the following fates. Firstly, if rapid functional divergence occurs to these proteins so that the duplicated protein is not part of the complex at all, then this would not affect the dosage of the participating complex. Secondly, if the duplicated protein enhances the topology of the complex which may enhance the total complex assembly, then the enhanced dosage of the particular gene would have no effect over the protein complex.[8,30] Thirdly, a multimer protein might tend to be involved in more functions than a monomer. So the rapid functional divergence of the multimer protein may promote functional divergence more rapidly. So, if the duplicated gene does undergo through the above fates, then the enhanced dosage of that particular gene may have a deleterious effect on the whole complex. Essentially, our analysis reveals that most of the proteins produced from complexes are not part of the complex and this higher rate of diversification is not found in case of yeast. That is why, in the case of yeast the higher the association within complex lesser is the duplicability.

So, in case of both humans and yeast, protein connectivity in the protein–protein interaction network as well as in the protein complex behaves in the same manner. But, in addition to this, in the case of human there exists a negative correlation between these two sorts of connectivities, which is absent in the case of yeast. In yeast these two connectivities act as two different forces that control in the same negative direction. But in case of a complex organism like humans the higher rate of duplication for a particular gene may be the cause of either connectivity in the protein–protein interaction network or perhaps connectivity within the protein complex. Hence, organismal complexity promotes protein complexity as well as protein connectivity to behave differently due to different dosage effect on these two organisms. In both the cases the higher divergence of the duplicated protein does not affect network assembly. This may be the effect of organismal complexity which has a major effect on gene duplicability.

Acknowledgements: We thank the Department of Biotechnology, Govt. of India for financial help. We are also thankful to Mr Sanjib K. Gupta and Mrs Sujata Roy for their technical help. We are grateful to Professor Pradip Kumar Parrack for his critical reading of the manuscript. We are thankful to two anonymous referees for their helpful comments in improving the manuscript.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

Department of Biotechnology, Govt of India (Sanction no 102/IFD/SAN/PR-1860/2008-2009).

References

1. Ohno, S. 1970, *Evolution by Gene Duplication*, Springer: Berlin, pp. 160.
2. Samonte, R.V. and Eichler, E.E. 2002, Segmental duplications and the evolution of the primate genome, *Nat. Rev. Genet.*, **3**, 65–72.
3. Li, W.H. 1997, *Molecular Evolution*, Sinauer Associates: Sunderland, MA, pp. 432.
4. Wolfe, K.H. and Shields, D.C. 1997, Molecular evidence for an ancient duplication of the entire yeast genome, *Nature*, **387**, 708–13.
5. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. 2003, Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature*, **423**, 241–54.
6. Nadeau, J.H. and Sankoff, D. 1997, Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution, *Genetics*, **147**, 1259–66.
7. Li, W.H., Gu, Z.L., Wang, H.D. and Nekrutenko, A. 2001, Evolutionary analyses of the human genome, *Nature*, **409**, 847–9.
8. Liang, H. and Li, W.H. 2007, Gene essentiality, gene duplicability and protein connectivity in human and mouse, *Trends Genet.*, **23**, 375–8.
9. Yang, J., Lusk, R. and Li, W.H. 2003, Organismal complexity, protein complexity, and gene duplicability, *Proc. Natl. Acad Sci. USA*, **100**, 15661–5.

10. Prachumwat, A. and Li, W.H. 2006, Protein function, connectivity, and duplicability in yeast, *Mol. Biol. Evol.*, **23**, 30–9.
11. Qian, W.F. and Zhang, J.Z. 2008, Gene dosage and gene duplicability, *Genetics*, **179**, 2319–24.
12. Kondrashov, F.A. and Koonin, E.V. 2004, A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications, *Trends Genet.*, **20**, 287–91.
13. Papp, B., Pal, C. and Hurst, L.D. 2003, Dosage sensitivity and the evolution of gene families in yeast, *Nature*, **424**, 194–7.
14. Makova, K.D. and Li, W.H. 2003, Divergence in the spatial pattern of gene expression between human duplicate genes, *Genome Res.*, **13**, 1638–45.
15. Prince, V.E. and Pickett, F.B. 2002, Splitting pairs: the diverging fates of duplicated genes, *Nat. Rev. Genet.*, **3**, 827–37.
16. Manna, B., Bhattacharya, T., Kahali, B. and Ghosh, T.C. 2009, Evolutionary constraints on hub and non-hub proteins in human protein interaction network: insight from protein connectivity and intrinsic disorder, *Gene*, **434**, 50–5.
17. Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. 2009, Human Protein Reference Database–2009 update, *Nucleic Acids Res.*, **37**, 767–72.
18. Ekman, D., Light, S., Bjorklund, A.K. and Elofsson, A. 2006, What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*?, *Genome Biol.*, **7**, R45.
19. Bader, J.S., Chaudhuri, A., Rothberg, J.M. and Chant, J. 2004, Gaining confidence in high throughput protein interaction networks, *Nat. Biotechnol.*, **22**, 78–85.
20. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., et al. 2008, CORUM: the comprehensive resource of mammalian protein complexes, *Nucleic Acids Res.*, **36**, D646–D650.
21. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., et al. 2006, Proteome survey reveals modularity of the yeast cell machinery, *Nature*, **440**, 631–6.
22. Dang, V.T., Kassahn, K.S., Marcos, A.E. and Ragan, M.A. 2008, Identification of human haploinsufficient genes and their genomic proximity to segmental duplications, *Eur. J. Hum. Genet.*, **16**, 1350–7.
23. Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., et al. 2005, Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast, *Genetics*, **169**, 1915–25.
24. Vilella, A.J., Severin, J., Ureta-Vidal, A., et al. 2009, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates, *Genome Res.*, **19**, 327–35.
25. Li, W.H., Gu, Z., Cavalcanti, A.R. and Nekrutenko, A. 2003, Detection of gene duplications and block duplications in eukaryotic genomes, *J Struct. Funct. Genomics*, **3**, 27–34.
26. Hahn, M.W. and Kern, A.D. 2005, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Mol. Biol. Evol.*, **22**, 803–6.
27. Veitia, R.A. 2002, Exploring the etiology of haploinsufficiency, *Bioessays*, **24**, 175–84.
28. Fodde, R. and Smits, R. 2002, Cancer biology. A matter of dosage, *Science*, **298**, 761–3.
29. He, X.L. and Zhang, J.Z. 2006, Higher duplicability of less important genes in yeast genomes, *Mol. Biol. Evol.*, **23**, 144–51.
30. Oberdorf, R. and Kortemme, T. 2009, Complex topology rather than complex membership is a determinant of protein dosage sensitivity, *Mol. Syst. Biol.*, **5**, 253.