Divergence Pattern of Duplicate Genes in Protein-Protein Interactions Follows the Power Law

Ze Zhang, Z. W. Luo,*† Hirohisa Kishino,‡ and Mike J. Kearsey†*

*School of Biosciences, University of Birmingham, Birmingham, United Kingdom; †Laboratory of Population and Quantitative Genetics, The State Key Laboratory of Genetic Engineering, Fudan University, Shanghai, China; and ‡Graduate School of Agriculture and Life Sciences, University of Tokyo, Tokyo, Japan

The impact of the biological network structures on the divergence between the two copies of one duplicate gene pair involved in the networks has not been documented on a genome scale. Having analyzed the most recently updated Database of Interacting Proteins (DIP) by incorporating the information for duplicate genes of the same age in yeast, we find that there was a highly significantly positive correlation between the level of connectivity of ancient genes and the number of shared partners of their duplicates in the protein-protein interaction networks. This suggests that duplicate genes with a low ancestral connectivity tend to provide raw materials for functional novelty, whereas those duplicate genes with a high ancestral connectivity tend to create functional redundancy for a genome during the same evolutionary period. Moreover, the difference in the number of partners between two copies of a duplicate pair was found to follow a power-law distribution. This suggests that loss and gain of interacting partners for most duplicate genes with a lower level of ancestral connectivity is largely symmetrical, whereas the "hub duplicate genes" with a higher level of ancient connectivity display an asymmetrical divergence pattern in protein-protein interactions. Thus, it is clear that the protein-protein interaction network structures affect the divergence pattern of duplicate genes. Our findings also provide insights into the origin and development of biological networks.

Introduction

Gene duplication, and subsequent divergence, has long been thought to be one of the principal engines powering the evolution of new protein function and facilitating genome complexity (Ohno 1970; Li 1997). Understanding the evolutionary mechanisms of duplicate genes is, therefore, important for evolutionary genomics, functional genomics, and systems biology. Two models have been proposed to characterize the possible mechanisms of divergence of duplicate genes. First, the Dykhuizen-Hartl (Dykhuizen and Hartl 1980) model postulates that, after gene duplication, random mutations are fixed in one daughter gene because of relaxed purifying selection resulting from reduced functional constraint provided by genetic redundancy. These fixed mutations later induce a change in gene function when the environment or the genetic background is altered. This model is neutral and does not involve positive selection. The second model requires positive selection and involves two scenarios. In the first scenario, a few neutral or nearly neutral substitutions occurring after gene duplication may create a new but only weakly active function in one daughter gene. Positive selection then accelerates fixation of the advantageous mutations, enhancing the newly established function (Zhang, Rosenberg, and Nei 1998). The second scenario assumes that the ancestral gene already had dual functions and its duplication provides the opportunity for each daughter gene to adopt different ancestral functions, and further substitutions under positive selection can refine these functions (Hughes 1999).

Although a fast-growing number of case studies have provided evidence to support these respective models (Zhang 2003), little is known about the generic pattern of divergence between two copies of duplicate genes on a genome scale. By analyzing protein-protein interaction data, expression data, and gene knockout data of yeast, Wagner (2002) deduced that divergence patterns of duplicate genes in protein-protein interactions were often asymmetrical; that is, one copy usually has significantly more interacting partners than the other after experiencing some period of divergent evolution. However, the inference was based on the use of synonymous substitutions between two duplicate copies as a proxy of their age since gene duplication occurred. The accuracy in approximating age of duplication by this means is questionable because it can be greatly biased by many factor, such as gene conversion and codon usage bias for many genes in the yeast genome. This may, thus, make it difficult to justify the analysis on the basis that the duplicate pairs under comparison share the same age.

Recently, comparison between genome sequences from two related yeast species has shown that *Saccharomyces cerevisiae* originates from an ancient whole-genome duplication (WGD) that took place about 100 MYA. The duplication event doubled the number of chromosomes in the *Saccharomyces* lineage (Kellis, Birren, and Lander 2004). The polyploid genome returned to functionally normal ploidy, not by chromosomal loss, but instead by a large number of deletion events. Indeed, just 12% of the paralogous gene pairs were retained in each doubly conserved synteny block, and the remaining 88% were lost. Of all the duplicate genes that have been retained in the *S. cerevisiae* genome to date, 457 pairs have been identified as having arisen from the WGD, indicating that these pairs are of the same age (Kellis, Birren, and Lander 2004). These pairs provide a unique opportunity to study the divergence pattern in protein-protein interactions between two copies of many duplicate pairs of genes over the same evolutionary period.

Taking advantage of the most recently updated Database of Interacting Proteins (DIP) together with a data set of the yeast duplicate genes, this report investigates the

Key words: Duplicate genes, protein-protein interactions, divergence, yeast, power law.

E-mail: z.zhang.2@bham.ac.uk; z.luo@bham.ac.uk.

Mol. Biol. Evol. 22(3):501–505. 2004
doi:10.1093/molbev/msi034
Advance Access publication November 3, 2004

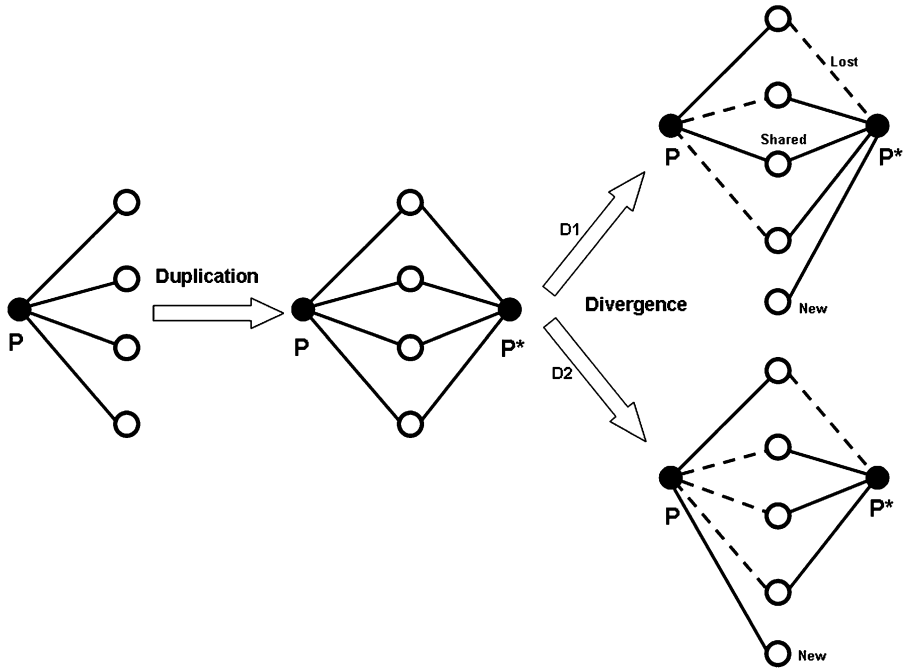

FIG. 1.—A model for divergence in protein-protein interactions between two duplicate genes. Circles stand for proteins. Lines stand for interactions among proteins. Immediately after a gene duplication, the two products P and P* of a duplicate gene have the same partners. Subsequently, divergent evolution results in turnover of interactions. Two duplicates may lose some common interacting partners (dashed lines) and get new partners (path D1). Alternatively, two duplicates may have no common partners and eventually, by complementary loss of partners, get new partners (path D2) (modified from Wagner [2001]).

divergence pattern of duplicated genes in the yeast protein-protein interaction network under the constraint that the duplicates share exactly the same age since duplication took place.

Databases and Analyses

Protein-Protein Interaction Database

The most recently updated version (ver040704) of the protein-interaction data set for the yeast, *S. cerevisiae*, was downloaded from DIP (http://dip.doe-mbi.ucla.edu/dip/) (Salwinski et al. 2004). This data set contains 4,741 proteins and their 15,409 interactions. To validate the robustness of our analyses, the CORE data set of the *S. cerevisiae* protein-protein interactions was also used in the study. The CORE data set includes 2,613 proteins and their 6,574 interactions and is a subset of the entire DIP data. The interactions presented in the data set have been checked by two forms of computational assessments (Deane et al. 2000; Salwinski et al. 2004). This largely reduces the rate of false-positive inferences among the interacting relationships. From these data sets, we counted the number of interacting partners for each copy of duplicates and the number of interacting partners shared between each pair of duplicates. The proteins with only self-interaction information were excluded from the analyses, and the self-interaction was not used to count the number of interacting partners of a protein.

Database of Yeast Duplicate Genes

Kellis, Birren, and Lander (2004) have recently sequenced and analyzed the genome of *Kluyveromyces waltii*, a yeast species that is a closely related to *S. cerevisiae*. They showed that the two yeast species were related by 1:2 mapping, with each region of *K. waltii* corresponding to two regions of *S. cerevisiae*. In the genome sequence database (http://www.broad.mit.edu/seq/YeastDuplication/), there are 457 pairs of the duplicate genes in the *S. cerevisiae* genome, which have been rigorously verified to have the same age since duplication (Kellis, Birren, and Lander 2004). Of the 457 pairs of duplicates, we identified 274 for which both copies have protein-protein interaction information available.

Results and Discussion

Evolution of Duplicate Genes in Protein-Protein Interaction Networks

A model for the divergence of two duplicate genes in protein-protein interactions is illustrated in figure 1. The model assumes that the two copies have an equal number of common interacting partners immediately after gene duplication. Subsequently, divergent evolution between the two duplicates would result in loss of some common interacting partners and gain of some new partners in one or both duplicates. In some cases, given the long period of evolution, the two copies might have no common interacting partners, particularly if the original gene had a few protein interacting partners. We first examine the number of common partners shared by two copies of one duplicate pair. As shown in table 1, after experiencing the same long evolutionary period, 205 of 274 duplicate pairs have no common partners, whereas the remaining 69 duplicate pairs shared some common interacting partners

Table 1
The Numbers for Duplicates Pairs with and Without the Shared Partners and the Estimates for Their Ancient Connectivity Levels

Number of Duplicate Pairs	Average Number of Shared Partners	Average of Estimates for Ancient Gene Connectivity Levels
205	0	4.73
69	2.05	11.30

NOTE.—The Wilcoxon test indicates that the duplicate pairs with the shared common interacting partners have significantly higher average of ancient connectivity level (11.30) than those without any shared common partners (4.73) ($W = 13999$, $P \leq 2.3\text{E}{-15}$).

between their two copies. The number of shared partners ranges from one to 14 with the average being two. This indicates that the rate of interaction turnover is very high, and the yeast-protein interaction network evolves rapidly, which is consistent with one previous study (Wagner 2001).

If the rate of interaction turnover is a constant for all duplicate pairs, the model illustrated in figure 1 will predict a positive correlation between the connectivity level of an ancient gene before duplication and the number of partners currently shared by the two duplicates. It is impossible to know the exact connectivity level of an ancient gene. However, the current average number of interacting partners for two copies of one duplicate pair can be used as its crude estimate under a random model of interaction turnover. With this, we do, in fact, observe such a significantly positive correlation (Pearson correlation: $r = 0.5184$, $P \leq$ 4.4E–20; Spearman rank correlation: $s = 0.5927$, $P \leq 1.2\text{E}{-}22$, $N = 274$ [fig. 2]). Moreover, the duplicate pairs with shared partners between the two copies now have a significantly higher current average number of interacting partners than those without the shared partner (table 1). This suggests that, for the ancient genes with a low level of connectivity, one of the duplicates is likely to evolve toward new interacting partners (new functions), whereas for the ancient genes with a high connectivity level, the two copies are likely to maintain some common interacting partners (functional overlap) during the same evolutionary period. The former tends to provide raw materials for functional novelty, whereas the latter tends to create functional redundancy for a genome.

Divergence Pattern of Duplicate Genes in Protein-Protein Interactions

To investigate how duplicates diverge since the duplication took place, we focus here on the distribution of differences (k) in the numbers of interacting partners between two duplicate copies. The values of k range from 0 to 114. Figure 3 demonstrates that the frequency distribution of k follows the power law with an exponential cutoff at $k_c \approx 11$. By using a least-squares method similar to that in a previous study (Wagner 2001) for log-transformed data, the estimate of the power-law exponent for $p(k) \propto k^{-\tau}$ is $\tau = 1.38$ (fig. 3), which is close to the value of 1.64 of the power-law exponent of the connectivity distribution in one combined protein-interaction network previously constructed from protein complexes (Gavin et al. 2002; Ho et al. 2002; Hahn, Conant, and Wagner 2004). Of the 274 duplicate pairs, 49 (17%) pairs have the same numbers ($k = 0$) of interacting partner, and 93 (34%) have nearly the same partner numbers ($k = 1$ or 2). The duplicate pairs with k values greater than 10 account for only 16% (43 pairs) of all 274 duplicate pairs under question. Moreover, the current average number of interacting partners for two copies of one duplicate pair is significantly and positively correlated with k (Pearson correlation: $r = 0.8786$, $P \leq 3.4\text{E}{-}76$; Spearman rank correlation: $s = 0.7563$, $P \leq 8.1\text{E}{-}36$, $N = 274$ [fig. 4]).

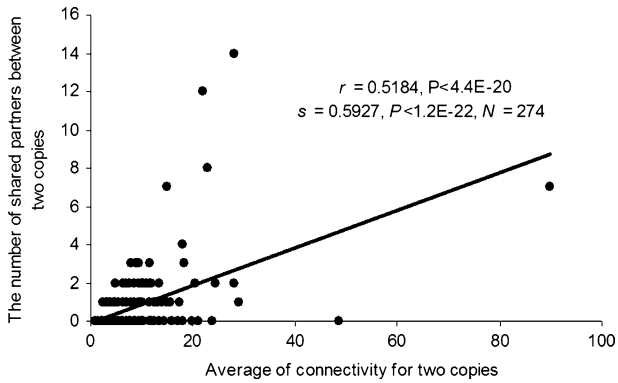

FIG. 2.—The relationship between the average of connectivity for two copies and the number of shared common interacting partners between two copies of one duplicate pair.

The same analyses were carried out with the CORE data set (Deane et al. 2000; Salwinski et al. 2004), the results were quite similar to those presented above and are presented as the Supplementary Material online (http://mbe.oupjournals.org/), confirming the robustness of our analyses across different data sets.

Our observation that ancient genes of most duplicate pairs (~51%) had a low connectivity level, and their two

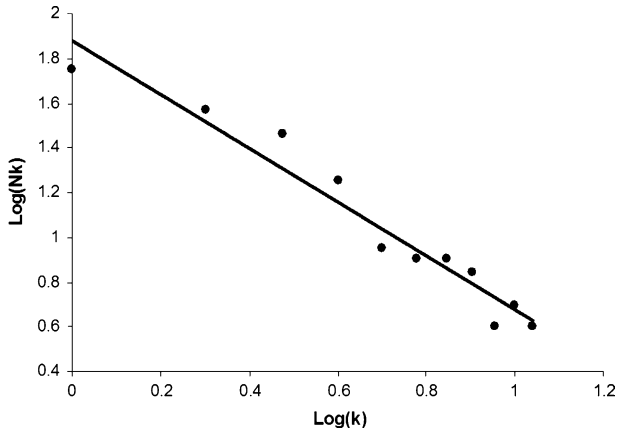

FIG. 3.—The log-log distribution of difference (k) in the numbers of interacting partners between two products of a gene duplication for the power law ($p(k) \propto k^{-1.38\pm0.11}$, $R^2 = 0.9592$, $P < 0.0001$. The data point with 0 of the difference k was not used to estimate the parameter of the power law because logarithm has no definition for zero). N_k is the number of pairs with difference k in the numbers of interacting partners between two duplicates. The exponential cutoff for difference (k) is at $k_c \approx 11$ and indicates that the number (N_k) of duplicated pairs with k more than 11 is slightly less than expected for pure power law.

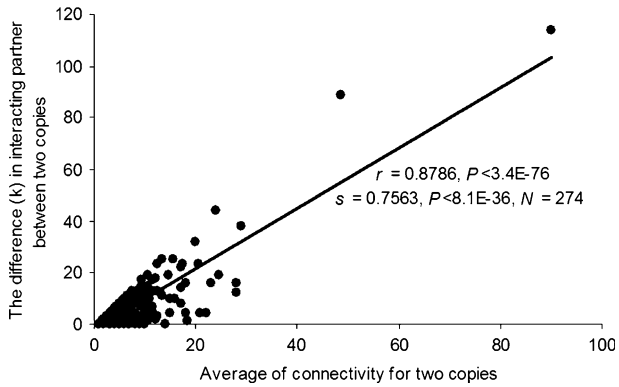

FIG. 4.—The relationship between the average of connectivity for two copies and the difference (k) in interacting partner number between two copies of one duplicate pair.

duplicates follow a symmetric or nearly symmetric divergence pattern supports the random interaction turnover model. This model predicts symmetry in the divergence pattern of the duplicates under the assumption of an equal rate of new partner gains for two copies of one duplicate pair as well as a constant rate of interaction turnover for all duplicate pairs; that is,. two copies of one duplicate pair would maintain a similar number of interacting partners after evolving for a long period. However, asymmetry of divergence in protein interacting number between the two copies increases with the ancient connectivity level of a gene before duplication. A small proportion of the "hub genes" (~16%) start with many partners (the average 16), and their duplicates follow a highly asymmetric divergence pattern (the difference k more than 10); that is, one copy usually has significantly more interacting partners than the other. Wagner (2002) observed only an asymmetric divergence pattern, which is part of a comprehensive pattern uncovered here. The difference in these observations can be explained by the fact that the observation in Wagner (2002) was based on the duplicates inferred from homology search, but it is difficult to prove that the duplicates so inferred have the same age. On the other hand, some of the duplicates included in that analysis may be members of large gene families, and their duplication ages may remain variable. However, one of the distinct features of the present analysis is that all duplicates included in the analysis have been confirmed to share the same duplication age. After having removed the variation, the present analysis should reveal a more comprehensive pattern of divergence of duplicate genes.

The power law is recognized as a universal law defining structure of scale-free networks such as protein-protein interaction and transcriptional and metabolic networks in biology (Jeong et al. 2000, 2001; Rain et al. 2001). The present study reveals that divergence of duplicate genes in protein-protein interactions also follows the power law. Those proteins with few interacting partners are usually on the edges of a protein-interaction network, whereas those with many partners generally reside in the central parts of the networks (Barabasi and Oltvai 2004). Thus, our observations further suggest that the divergence patterns of duplicate genes during the same evolutionary period depend to some extent on their positions in protein-interaction networks: they are governed both by a random process of interaction turnover and by the structure of the protein-interaction networks.

Two fundamental processes have a key role in the origin and development of biological networks. First, most networks are products of a growth process, during which new nodes join the system over an extended time period. Second, nodes prefer to connect to nodes that already have many links, a process that is known as preferential attachment (Barabasi and Albert 1999; Barabasi and Oltvai 2004). These two processes are probably rooted in gene duplication (Bhan, Galas, and Dewey 2002; Pastor-Satorras, Smith, and Sole 2003). Therefore, the comprehensive divergence pattern of duplicate genes uncovered in the present study will shed light on the origin and development of biological networks.

Acknowledgments

The authors are grateful for the comments and criticisms made by two anonymous reviewers, which have helped improve the presentation of this paper. This study was supported in part by a research grant from the United Kingdom Biotechnology and Biological Sciences Research Council and by the Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST). Z.W.L. is also supported by China's National Natural Science Foundation (30430380), the Basic Research Program of China (2004CB518605), and Shanghai S&T Committee.

Literature Cited

Barabasi, A.-L., and R. Albert. 1999. Emergence of scaling in random networks. Science **286**:509–512.

Barabasi, A.-L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**:101–113.

Bhan, C., D. J. Galas, and T. G. Dewey. 2002. A duplication growth model of gene expression networks. Bioinformatics **18**: 1486–1493.

Deane, C. M., L. Salwinski, I. Xenarios, and D. Eisenberg. 2000. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol. Cell. Proteomics **1**:349–56.

Dykhuizen, D., and D. L. Hartl. 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. Genetics **96**:801–817.

Gavin, A. C., M. Bosche, R. Krause et al. (38 co-authors). 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**:141–147.

Hahn, M. W., G. C. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: Does connectivity equal constraint? J. Mol. Evol. **58**:203–211.

Ho, Y., A. Gruhler, A. Heilbut et al. (46 co-authors). 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature **415**:180–183.

Hughes, A. L. 1999. Adaptive evolution of genes and senomes, Oxford University Press, Oxford.

Jeong, H., S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. Nature **411**: 41–42.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi. 2000. The large-scale organization of metabolic networks. Nature **407**:651–654.

Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428**:617–624.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany.

Pastor-Satorras, R., E. Smith, and R. Sole. 2003. Evolving protein interaction networks through gene duplication. J. Theor. Biol. **222**:199–210.

Rain, J.-C., L. Selig, H. De Reuse et al. (13 co-authors). 2001. The protein-protein interaction map of *Helicobacter pylori*. Nature **409**:211–215.

Salwinski, L., C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. **32**:D449–51.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Biol. Evol. **18**:1283–1292.

———. 2002. Asymmetric functional divergence of duplicate genes in yeast. Mol. Biol. Evol. **19**:1760–1768.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. **18**:292–298.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

Takashi Gojobori, Associate Editor

Accepted October 26, 2004