J Mol Evol (2005) 61:758–764
DOI: 10.1007/s00239-005-0037-z

JOURNAL OF MOLECULAR EVOLUTION

Gene Duplication and the Properties of Biological Networks

Austin L. Hughes, Robert Friedman

Department of Biological Sciences, University of South Carolina, Coker Life Sciences Building, 700 Sumter Street, Columbia, SC 29205, USA

Received: 18 February 2005 / Accepted: 12 July 2005 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. Patterns of network connection of members of multigene families were examined for two biological networks: a genetic network from the yeast *Saccharomyces cerevisiae* and a protein–protein interaction network from *Caenorhabditis elegans*. In both networks, genes belonging to gene families represented by a single member in the genome (''singletons'') were disproportionately represented among the nodes having large numbers of connections. Of 68 single-member yeast families with 25 or more network connections, 28 (44.4%) were located in duplicated genomic segments believed to have originated from an ancient polyploidization event; thus, each of these 28 loci was thus presumably duplicated along with the genomic segment to which it belongs, but one of the two duplicates has subsequently been deleted. Nodes connected to major ''hubs'' with a large number of connections, tended to be relatively sparsely interconnected among themselves. Furthermore, duplicated genes, even those arising from recent duplication, rarely shared many network connections, suggesting that network connections are remarkably labile over evolutionary time. These factors serve to explain well-known general properties of biological networks, including their scale-free and modular nature.

Key words: Gene duplication — Biological networks — Scale-free networks — Genome evolution

Correspondence to: Austin L. Hughes; *email:* austin@biol.sc.edu

Introduction

With increasing knowledge of gene regulation, protein–protein interactions, and metabolic processes, it has become possible to assemble these and similar sorts of biological information in the form of networks, that is, graphical representations of intermolecular interactions (Kanehisa 2000). Networks have been constructed from information on metabolic pathways, signal transduction, transcriptional regulation, and other cellular processes (Kanehisa 2000). One important generalization regarding biological networks is that they tend to be scale-free (Barabási and Albert 1999). Unlike a random network, a scale-free network has the property that a small proportion of nodes have a large number of connections, while the other nodes have smaller numbers of connections (Barabási and Albert 1999). In intuitive terms, in a scale-free network, nodes are divided into ''hubs'' (having many connections) and ''spokes'' (having few connections and connected with one another mainly through hubs). More formally, the scale-free property occurs when $P(k)$, the probability that a node in the network is connected to k other nodes, decays as a power law, following $P(k) \sim k\ k^{-\gamma}$, where γ is a positive real number (often about 2.0 in a wide variety of networks known from both the biological and the social sciences) (Barabási and Albert 1999).

An apparent paradox of biological networks is that within many such networks there are numerous small modules of densely interconnected nodes, while connections between modules are sparser (Ravasz et al. 2002). A modular organization would seem to contradict the scale-free property, since in such a network nodes would tend to have roughly equal

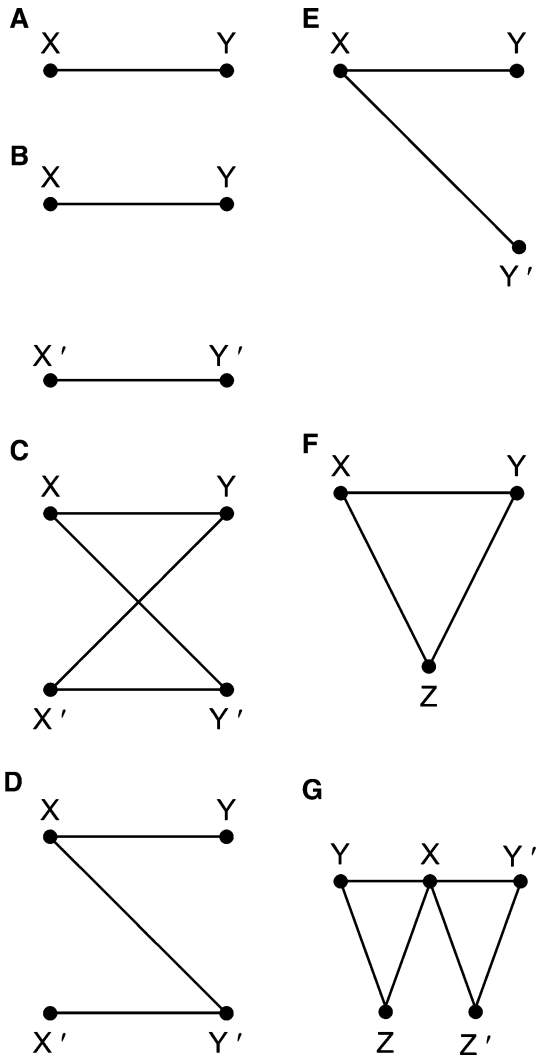

Fig. 1. Hypothetical scenarios of gene duplication in biological networks. (For explanation, see text.)

numbers of connections. However, it has been shown in the case of metabolic networks from a variety of species that the network consists of many small highly connected modules that combine in a hierarchical manner, as a result of a small number of nonrandom intermodule links that connect modules in a nested fashion (Ravasz et al. 2002). This hierarchical mode of organization explains the observation that biological networks are scale-free and yet consist of functionally distinct modules.

It has been hypothesized that repeated gene duplication over evolutionary time can account for the properties of biological networks (Wagner 2001). However, this will only be true if the duplication process has certain characteristics. Figure 1A illustrates the simplest possible protein–protein interaction network: a network consisting of just two interacting proteins, X and Y. This network is not scale-free. Suppose that the genes encoding X and Y are both duplicated, giving rise to two new genes encoding two new proteins (X′ and Y′). If X retains its interaction with Y, while X′ interacts only with Y′ (Fig. 1B), the network will not be any more scale-free than was the original network. Similarly, if X interacts with both Y and Y′ while X′ interacts with both Y and Y′ (Fig. 1C), the network will not be any more scale-free than was the original network.

On the other hand, the scale-free property is increased if, after duplication, interactions are lost differentially. An example of such a process is illustrated in Fig. 1D. Here, after duplication of both the genes encoding X and Y, X retains the capacity to interact with both Y and Y′, whereas X′ interacts only with Y′ (Fig. 1D). Likewise, the scale-free property of the network will be increased if only one of the two genes is duplicated, but both duplicates retain the capacity to interact with the unduplicated gene (Fig. 1E). The same effect would be produced if both genes were duplicated, but one duplicated gene was subsequently deleted.

These simple examples show that gene duplication will increase the scale-free property of networks if genes involved in networks are duplicated differentially, are deleted differentially after duplication, and/or if interactions are retained differentially after duplication. A similar effect would also be produced if different interactions were acquired by each gene independently after duplication, especially if duplicates differed with respect to the number of interactions acquired.

Figure 1F represents a simple "module" of three interacting proteins (X, Y, and Z). If the genes encoding Y and Z are duplicated, and the proteins encoded by the duplicates continue to interact with X (Fig. 1G), the result will be a network with two hierarchically combined modules. This simple example shows that differential duplication can, under appropriate circumstances, yield a network having the property of a modular and hierarchical organization. Again, the same effect would be produced if both genes were duplicated, but one duplicated gene was subsequently deleted. Consistent with this reasoning, eukaryotic genomes include substantial numbers of unduplicated genes ("singletons"), in spite of the existence of numerous multigene families (Friedman and Hughes 2001a, b).

These theoretical considerations lead to the predictions that biological networks will be characterized by three phenomena: (1) After gene duplication, new interactions will be differentially acquired by duplicates and/or ancestral interactions will be differentially retained by duplicates. The latter will be the case particularly when the duplication involves genes constituting multiconnected nodes ("hubs") in a gene interaction network. (2) Genes corresponding to network hubs will consist disproportionately of singletons. (3) When genes corresponding to network hubs are duplicated, it will often happen that one

copy is quickly deleted. Here we test these predictions with data on gene interaction networks from two model organisms: the genetic interaction network of yeast *Saccharomyces cerevisiae* (Tong et al. 2004) and protein–protein interaction network of the nematode worm *Caenorhabditis elegans* (Li et al. 2004).

Methods

The complete sets of predicted protein translations for the following organisms were downloaded from the euGenes Web site, http://iubio.bio.indiana.edu:8089: yeast *Saccharomyces cerevisiae* (version 06/24/2002) and *Caenorhabditis elegans* (version 06/24/2002). For each of these genomes, gene families were assembled using the BLASTCLUST computer program available in the BLAST tools (Altschul et al. 1997). This program establishes families by BLASTP homology search and the single-linkage method (i.e., if a match is scored between A and B and between B and C, then A, B, and C are placed in the same family). We used a value of 10^{-6} for the E parameter (representing the probability that a score as high as that observed between two sequences will be found by chance in a database of the size examined) of the BLAST algorithm. To score a match between two proteins, we further required that 30% of amino acids be identical and 50% of aligned amino acid sites be shared. These criteria have been shown to assemble multigene families whose members show evidence of homology throughout the length of the sequence, thus making them suitable for phylogenetic analysis and estimation of evolutionary distances (Hughes and Friedman 2004).

In the case of selected duplicate gene pairs, sequences were aligned at the amino acid level using the CLUSTALW program (Thompson et al. 1994), and this alignment was imposed on the DNA sequences. The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) were estimated by a maximum likelihood method (Yang and Nielsen 2000) using the software package PAML (Yang 1997). Since most synonymous mutations are selectively neutral or nearly so (Kimura 1977), d_S is expected to be correlated with the amount of time since duplication of the two genes compared. By contrast, d_N reflects the extent to which the two genes are subject to purifying selection arising from functional constraint on the amino acid sequence (Kimura 1977; Nei 1987).

Information for the yeast genetic interaction network was obtained from Tong et al. (2004), who determined about 4000 interactions for about 1000 genes using computational analysis of data from 132 synthetic gene array screens. Information for the *C. elegans* protein–protein interaction was obtained from Li et al. (2004), who obtained data on over 4000 interactions using high-throughput yeast two-hybrid screens and combined these results with previously described interactions and in silico predictions for a total of about 5500 interactions. Genes included in these networks (820 genes from yeast and 2606 genes from *C. elegans*) were matched with the gene families determined by homology search. We examined the relationship between the number of connections a gene had in the network and the size of the family to which it belonged. Family sizes were based on the complete sets of predicted proteins rather than on the sets included in the networks. The clustering coefficient for node i with k_i links was defined as $C_i = 2n_i/[k_i(k_i - 1)]$, where n_i is the number of links between the k_i neighbors of i (Ravasz and Barabási 2003). Note that C_i is undefined for nodes with only one connection.

Information on duplicated segments in the yeast genome was obtained from Seoighe and Wolfe (1999). In order to obtain a conservative estimate of duplicated regions, for purposes of our analyses we did not include duplicated regions designated "possible" or "low-scoring" by those authors.

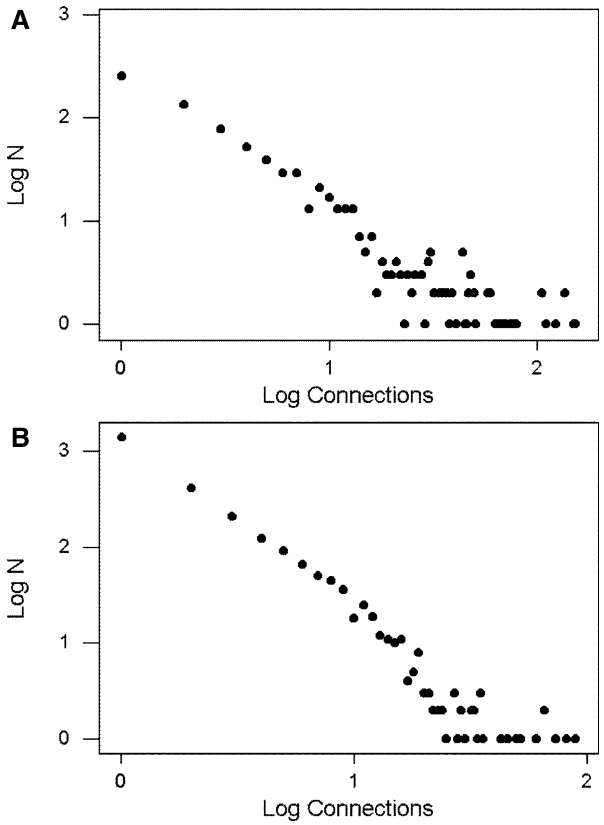

Fig. 2. Plots of the log number of nodes (N) having a given number of connections vs. the log number of connections (Connections) for (**A**) the yeast genetic interaction network and (**B**) the *C. elegans* protein–protein interaction network. In the yeast network, the relationship between N and Connections was described by the regression equation $Y = 162.18\ X^{-1.16}$ ($R^2 = 83.9\%$). In the *C. elegans* network, the relationship between N and Connections was described by the equation $Y = 1288.25\ X^{-1.84}$ ($R^2 = 91.4\%$).

Phylogenetic analyses were conducted by the following methods: (1) the maximum parsimony (MP) method, implemented in the PAUP* program (Swofford 2002); (2) the quartet maximum likelihood method (QML), implemented in the PUZZLE 5.0 program (Strimmer and van Haeseler 1996); and (3) the neighbor-joining (NJ) method (Saitou and Nei 1987), implemented in the MEGA2 program (Kumar et al. 2001). The NJ trees were based on the gamma-corrected amino acid distance, with the shape parameter estimated by the PUZZLE 5.0 program. The reliability of clustering in the MP and NJ trees was assessed by bootstrapping (Felsenstein 1985); 1000 bootstrap samples were used. In QML trees, the proportion of puzzling steps supporting a branch provided a similar index of the reliability of clustering patterns. Since all phylogenetic methods produced essentially identical trees, only the MP tree is shown in the following.

Results

Network Properties

Both the yeast gene interaction network and the *C. elegans* protein–protein interaction network showed patterns characteristic of a scale-free network, with a negative exponential relationship between the number of connections and the number of

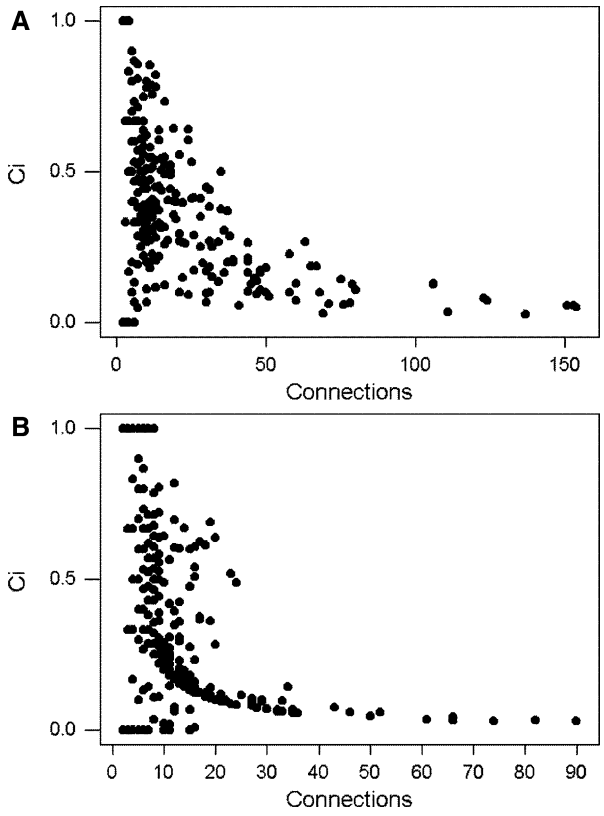

Fig. 3. Plots of the clustering coefficient (C_i) of a node vs. the number connections at the node for (**A**) the yeast genetic interaction network (r_S = –0.588; $p < 0.001$) and (**B**) the *C. elegans* protein–protein interaction network (r_S = –0.500; $p < 0.001$).

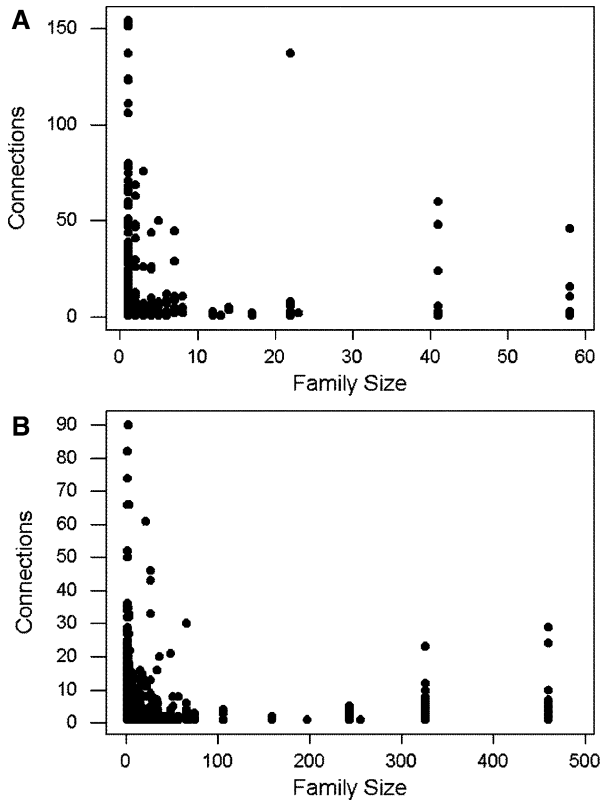

Fig. 4. Plots of the number of connections of a node vs. family size for (**A**) the yeast genetic interaction network (r_S = –0.124; $p < 0.001$) and (**B**) the *C. elegans* protein–protein interaction network (r_S = –0.053; $p = 0.007$).

nodes in the network having that number of connections. In each case, a regression relating the log number of genes to the log number of connections produced a highly significant linear relationship with a negative slope (Fig. 2). The slope of the relationship had a much higher absolute value in the *C. elegans* network (slope = −1.84; Fig. 2B) than in the yeast network (slope = −1.16; Fig. 2A). This difference is explained by a higher frequency of nodes with a small number of connections in the former network than in the latter network. In the *C. elegans* network, 1406 of 2606 nodes (54.0%) had a single connection, whereas in the yeast network only 254 of 820 (31.0%) of nodes had a single connection.

In spite of this difference, clustering coefficients for the two networks were similar. The median clustering coefficient (C_i) for the yeast network was 0.500, while that for the *C. elegans* network was 0.533. These medians were not significantly different (Mann–Whitney test). In both networks, there was a strong negative correlation between C_i and the number of connections at a node (Fig. 3). The Spearman rank correlation coefficient (r_S) between C_i and the number of connections was –0.588 ($p < 0.001$) in the case of the yeast network (Fig. 3A) and –0.500 ($p < 0.001$) in the case of the *C. elegans* network (Fig. 3B).

Family Size and Connections

In the yeast network, there was a negative rank correlation between the number of connections a gene had and its family size (r_S = –0.124; $p < 0.001$; Fig. 4A). This negative correlation was explained by the fact that most genes with large numbers of connections were singletons (Fig. 4A). Nine of 10 genes with 100 or more connections were singletons, and 64 of 83 (77.1%) of genes with 25 or more connections were singletons (Fig. 4A). In the *C. elegans* network, there was a similar, though weaker, negative rank correlation between the number of connections a gene had and its family size (r_S = –0.053; $p = 0.007$; Fig. 4B). The protein with the highest number of connections (90) was a member of a two-member family, while the proteins with the next three highest numbers of connections (82, 74, and 66) were all encoded by singletons (Fig. 4B). Of 27 proteins with 25 or more connections, 15 (55.6%) were proteins encoded by singletons.

Because of the negative correlation between C_i and the number of connections in both networks (Fig. 3), we further examined the relationship between family size and number of connections using rank partial correlation, controlling for the effect of C_i (Table 1). Since C_i is not defined for nodes with only one con-

Table 1. Rank partial correlations among three variables describing network nodes, in each case controlling for the other variable

	C_i	p	No. connections	p
	Yeast network (566 nodes)			
Family size	−0.033	N.S.	−0.099	0.019
C_i			−0.588	< 0.001
	C. elegans network (1200 nodes)			
Family size	−0.027	N.S.	−0.076	0.008
C_i			−0.501	< 0.001

Table 2. Summary statistics for variables describing 34 two-member families in the *C. elegans* protein–protein interaction network

Variable	Mean ± SE	Median	Range
d_N	0.539 ± 0.047	0.595	0.016–1.095
d_S	2.500 ± 0.212	2.284	0.039–4.942
No. shared connections	0.471 ± 0.128	0.000	0–3
Percentage connections shared	7.8 ± 2.7	0.0	0.0–66.7

nection, such nodes were not included in the partial correlation analyses. Even excluding such nodes, there was a significant negative partial rank correlation between family size and number of connections in both the yeast network and the *C. elegans* network (Table 1). Likewise, in both networks there was a highly significant negative rank partial correlation between number of connections and C_i, controlling for the effect of family size (Table 1). On the other hand, in neither network was there a significant partial rank correlation between family size and C_i, controlling for the number of connections (Table 1).

Gene Duplication and Network Connections

There is evidence of extensive ancient segmental duplication in the yeast genome, which has been attributed to an ancient polyploidization event which occurred about 200 million years ago (Wolfe and Shields 1997; Seoighe and Wolfe 1999; Friedman and Hughes 2001a; Hughes and Friedman 2003; Kellis et al. 2004). Of 68 single-member yeast families with 25 or more network connections, 28 (44.4%) were located in duplicated blocks believed to have originated from polyploidization (Seoighe and Wolfe 1999). The fact that, in spite of their location in duplicated regions, these families contain a single member implies that, after segmental duplication, one duplicate member of each of these 28 families was deleted from the genome.

In *C. elegans*, we compared the connections of 34 two-member families, both members of which were included in the protein interaction network (Table 2). Most of the duplication events giving rise to the 34 pairs of paralogues were ancient, as indicated by high mean values of d_S and d_N (Table 2). In general, the paralogous gene pairs showed little tendency to share network connections; the mean number of connections shared was less than one, and the median number of connections shared was zero (Table 2). Furthermore, there was no significant correlation between either d_S or d_N and either the number of connections shared or the percentage of connections shared (data not shown). In only 3 of the 34 gene pairs was d_S less than 1.0, and in all 3 of these pairs no network connections were shared between pair members. C38C10.4 and F22B7.13 were the protein pair with the lowest d_S (0.038); these two proteins shared none of the five connections of the former protein or of the six connections of the latter protein.

Phylogenetic analyses of families with multiple members in a network were used to examine the relationship between phylogenetic relatedness and sharing of network connections. Figure 5A shows the phylogenetic tree of MAP kinases from the yeast network; this was the family showing the greatest within-family contrast in numbers of connections in either network. The two genes in this family with the highest numbers of connections, YPL031C (with 62 connections) and YHR030C (with 60 connections), were not sisters (Fig. 5A). There was strong (100%) bootstrap support for clustering of YHR030C with YLR113W, which had only 24 connections (Fig. 5A). The same clustering pattern received strong support in QML and NJ trees (data not shown). When sharing of connections among these genes was examined, YHR030C was found to share only a single connection with the closely related YLR113W (Fig. 5B). On the other hand, YHR030C shared 14 connections with YPL031C (Fig. 5B). All other members of this family included in the yeast network shared at most a single connection (Fig. 5B).

Discussion

A number of general patterns emerged from the evolutionary analysis of two biological networks with rather different properties, a genetic interaction network of yeast and a protein–protein interaction network of *C. elegans*. First, in both networks, genes belonging to gene families represented by a single member in the genome (“singletons”) were disproportionately represented among the nodes having large numbers of connections. Furthermore, in the case of yeast, there was evidence that when singletons with large numbers of connections have been duplicated, one of the two duplicate copies has frequently been deleted. Of 68 single-member yeast families with

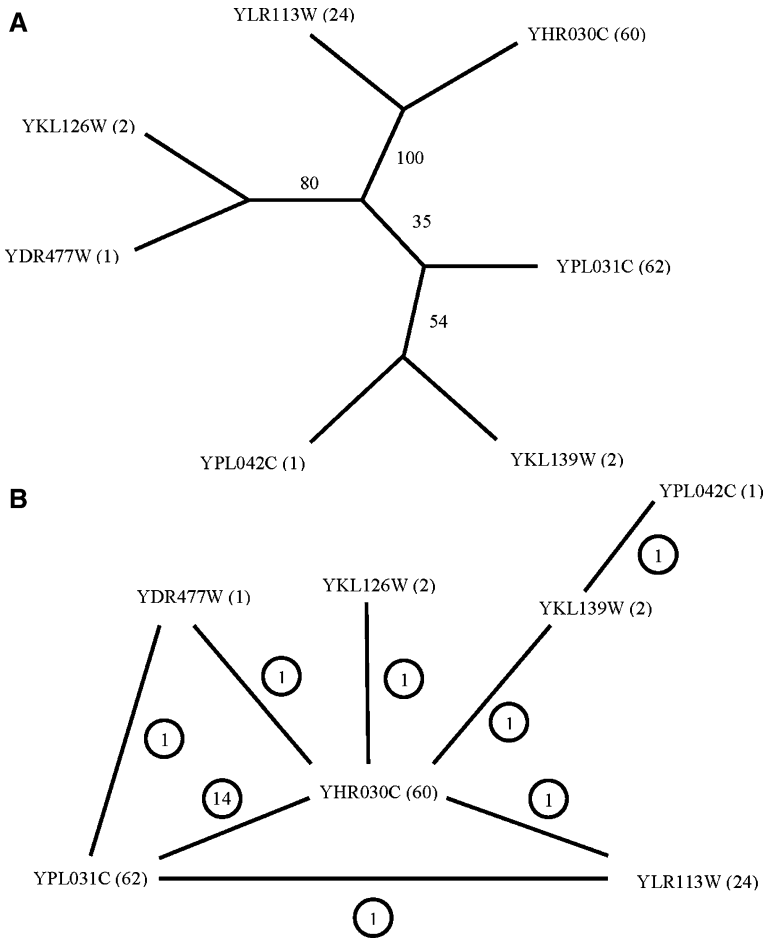

Fig. 5. **A** MP tree of yeast MAP kinases included in the genetic interaction network. Numbers in parentheses after each gene name are numbers of network connections. Numbers on the branches show the percentage of 1000 bootstrap samples supporting the branch. **B** Network indicating numbers of network connections shared by yeast MAP kinases. Numbers in parentheses after each gene name are numbers of network connections.

25 or more network connections, 28 (44.4%) were located in duplicated genomic segments believed to have originated from an ancient polyploidization event (Seoighe and Wolfe 1999). Each of these 28 loci was thus presumably duplicated along with the genomic segment to which it belongs, but one of the two duplicates has subsequently been deleted.

A second property shared by both networks was the strong negative correlation between the clustering coefficient (C_i) of a node and the number of connections at the node. This relationship held even when the effect of family size was controlled for statistically. Nodes connected to major "hubs" with a large number of connections tended to be sparsely interconnected among themselves.

Finally, there was evidence that network connections are remarkably labile over evolutionary time. Immediately after gene duplication, it seems reasonable to suppose that gene duplicates have the same network connections, unless one duplicate is partial or an exon-shuffling event or other recombinational event has accompanied gene duplication. But our results suggest that duplicated genes generally have quite distinct sets of connections, and that such changes can happen soon after duplication, as indicated by paralogous gene pairs in *C. elegans* with relatively low synonymous divergence.

Taken together, these observations paint a picture of the evolutionary process underlying the known characteristics of biological networks, namely, the properties of being scale-free and modular/hierarchical in organization. A multiply connected node is a hierarchical hub if the nodes connected to it have relatively little connection among themselves, whereas a module within a network would consist of a small set of mutually interconnected nodes. Therefore, evidence of a negative correlation between the clustering coefficient and the number of connections at a node provides an insight into at least one mechanism maintaining the modular/hierarchical nature of biological networks.

There was evidence that duplicated copies of multiply connected genes are frequently deleted, as evidently happened with such genes in duplicated segments of the yeast genome. Such deletion of duplicate genes has happened frequently enough in yeast to suggest that it may result from natural selection against duplication of multiply connected hubs. Moreover, the evidence that network connections are highly labile over evolutionary time suggests

that even when multiply connected genes are duplicated and both duplicates are retained, one duplicate may lose numerous connections, while the other duplicate retains ancestral connections. This process would result in widely different numbers of connections within multigene families, as in the MAP kinase family of yeast (Fig. 5).

Acknowledgments. This research was supported by Grant GM 043940 to A.L.H. from the National Institutes of Health.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Friedman R, Hughes AL (2001a) Gene duplication and the structure of eukaryotic genomes. Genome Res 11:373–381

Friedman R, Hughes AL (2001b) Pattern and timing of gene duplication in animal genomes. Genome Res 11:1842–1847

Hughes AL, Friedman R (2003) Parallel evolution by gene duplication in the genomes of two unicellular fungi. Genome Res 13:794–799

Hughes AL, Friedman R (2004) Differential loss of ancestral gene families as a source of genomic divergence in animals. Pro R Soc Lond B (Suppl) 271:S107–S109

Kanehisa M (2000) Post-genome informatics. Oxford University Press, Oxford

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain P-O, Han J-D J, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual J-F, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berritz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mange SE, Saxton WM, Strome S, van den Heuvel S , Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303:540–543

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. Phys Rev E 67:026112

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Seoighe C, Wolfe KH (1999) Updated map of duplicated regions in the yeast genome. Gene 238:253–261

Strimmer K, van Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Swofford DL (2002) PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland, MA

Thompson JD, Higgins DG, Gibson T (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesan H, Goldberg DS, Haynes J, Humphries C, He G, Levinson JN, Lu H, Ménard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu A-M, Shapiro J, Sheikh B, Suter B, Wong SL, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303:808–813

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

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43