The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes

Andreas Wagner

Department of Biology, University of New Mexico, and Santa Fe Institute, Santa Fe, New Mexico

In this paper, the structure and evolution of the protein interaction network of the yeast *Saccharomyces cerevisiae* is analyzed. The network is viewed as a graph whose nodes correspond to proteins. Two proteins are connected by an edge if they interact. The network resembles a random graph in that it consists of many small subnets (groups of proteins that interact with each other but do not interact with any other protein) and one large connected subnet comprising more than half of all interacting proteins. The number of interactions per protein appears to follow a power law distribution. Within approximately 200 Myr after a duplication, the products of duplicate genes become almost equally likely to (1) have common protein interaction partners and (2) be part of the same subnetwork as two proteins chosen at random from within the network. This indicates that the persistence of redundant interaction partners is the exception rather than the rule. After gene duplication, the likelihood that an interaction gets lost exceeds 2.2×10^{-3}/Myr. New interactions are estimated to evolve at a rate that is approximately three orders of magnitude smaller. Every 300 Myr, as many as half of all interactions may be replaced by new interactions.

Introduction

Do most gene duplicates retain similar functions for long periods of time? Or do they diverge in function soon after duplication? There is evidence for both possibilities. First, higher metazoan genomes contain scores of genes with overlapping functions (Joyner et al. 1991; Tautz 1992; Thomas 1993; Cadigan, Grossniklaus, and Gehring 1994; Gonzalez-Gaitan et al. 1994; Fromental-Ramain et al. 1996; Wang et al. 1996; Wilkins 1997). In well-studied cases, overlap in gene functions is demonstrated biochemically. Alternatively, genetic evidence, that is, weak phenotypic effects of a synthetic null mutation in one duplicate, has been used to suggest overlapping gene functions. Particularly in vertebrates, many genes with overlapping functions are remnants of ancient (>400 Myr) gene or genome duplications (Joyner et al. 1991; Thomas 1993; Fromental-Ramain et al. 1996; Sharman and Holland 1996; Wang et al. 1996; Bailey et al. 1997). However, even unicellular eukaryotes contain distantly related gene pairs with similar functions. A case in point is that of the CLN genes of budding yeast, a family of three cyclin genes whose products regulate the activity of the yeast cyclin-dependent kinase Cdc28p (Nasmyth 1993). Jointly, they are required for the transition from the G_1-phase to the S-phase of the yeast cell cycle. Individually, null mutants have weak phenotypic effects (Benton et al. 1993). For instance, a null mutation in CLN1 does not exhibit a growth defect on minimal medium (Smith et al. 1996). The most closely related genes in this family are CLN1 and CLN2. Since their duplication, an estimated K_s = 2.4 synonymous nucleotide substitutions have occurred per synonymous site on their DNA (Li 1997). Below, it is estimated that for yeast, a K_s = 1 among duplicated genes corresponds to a duplication age of approximately 100 Myr. By this measure, this duplication may be over 200 Myr old. Another example involves the three TPK genes from yeast. All of them are catalytic subunits of the yeast cyclic AMP-dependent protein kinase. Any two of the three genes are dispensable for growth (Toda et al. 1987). The most closely related pair is TPK1-TPK3, with K_s = 1.31. Based on such examples, it appears that gene duplicates may retain similar functions long after a duplication.

Only indirect evidence is available for rapid functional divergence after gene duplication. A number of case studies suggest that positive selection of advantageous mutations occurs after duplication (Long and Langley 1993; Benton et al. 1997; Cirera and Aguade 1998; Tsaur, Ting, and Wu 1998; Zhang, Rosenberg, and Nei 1998). Unfortunately, it is usually unknown whether such mutations often cause a change in function. An exception is the primate genes for eosinophil cationic proteins (ECPs) and for eosinophil-derived neurotoxin (EDN). They were duplicated an estimated 31 MYA (Zhang et al. 1998). EDN has high RNAse activity and may act as an antiretroviral agent. ECP is an antibacterial toxin exerting its effects independently of RNAse activity by making pores in bacterial cell membranes. A second line of evidence involves the divergence of expression patterns after duplication. If duplicates are expressed in different parts of an organism, then they are likely to have different biological functions, regardless of whether their biochemical activities are identical. Examples include the expression patterns of the *dopa decarboxylase* gene and the α-*methyldopa hypersensitive (amd)* gene from *Drosophila melanogaster*. In the closely related sibling species *Drosophila simulans*, the *amd* gene has evolved a new expression pattern within the last 2–5 Myr (Wang, Marsh, and Ayala 1996).

The relatively few examples of rapidly diverging functions in duplicate genes contrast with the greater documented number of gene duplicates with similar functions. However, it is not valid to conclude that rapid divergence of function is rare. After all, our knowledge in this area relies only on case studies. The purpose of this paper is to address this question more systematically

Key words: protein network, evolution, *Saccharomyces cerevisiae*, yeast, gene duplication.

Address for correspondence and reprints: Andreas Wagner, University of New Mexico, Department of Biology, 167A Castetter Hall, Albuquerque, New Mexico 87131-1091. E-mail: wagnera@unm.edu.

Mol. Biol. Evol. 18(7):1283–1292. 2001

and on a genomewide scale. At what rate does functional divergence occur after gene duplication for a large sample of duplicate genes in a genome? Addressing this question will also shed light on the evolution of a large genetic network.

Before studying the evolution of function in many genes, one has to make a difficult choice. What aspect of function should one focus on? Genes exert their biological roles in many different ways. Some gene products are parts of subcellular structures, others engage in protein-protein interactions, protein-DNA interactions, or catalytic interactions with small molecules. Moreover, genes with the same biochemical activities may be expressed at different times or in different places. A detailed biochemical understanding of the function of all of an organism's genes is not within reach. However, genomic technologies permit the analysis of one aspect of gene function for all genes. For instance, microarray analysis (DeRisi, Iyer, and Brown 1997) can be used to analyze transcriptional regulatory interactions on a genomewide scale. The validity of any functional genomics study rests on the implicit assumption that studying one aspect of gene function allows us to learn about the biology of an organism. The aspect of gene function studied here is protein-protein interaction, i.e., the number and identity of proteins with which the products of duplicate genes in the yeast *Saccharomyces cerevisiae* interact. The required information on these protein-protein interactions comes from a large-scale experiment (Uetz et al. 2000) using the yeast two-hybrid assay (Fields and Song 1989).

The yeast two-hybrid assay (Fields and Song 1989) takes advantage of the fact that many eukaryotic transcription factors have two separate domains, one required for binding of the transcription factor to DNA, and another one required for interaction with RNA polymerase II and for the initiation of transcription. In a two-hybrid assay, a known protein fused to the DNA-binding domain of a transcription factor is transfected into a yeast cell bearing a reporter gene that is under the control of the DNA-binding domain. This fusion protein can then be used as ''bait'' to screen a library of cDNA clones fused to the activation domain of the transcription factor. If the bait protein interacts physically with a fusion protein expressed from the library, then the DNA-binding domain and the activation domain are held in physical proximity, and the reporter gene is expressed. In an effort to elucidate protein interactions on a genomewide basis, this method has been applied to the genomes of viruses (Bartel et al. 1996; McGraith et al. 2000), as well as to the genome of budding yeast, yielding a comprehensive map of protein-protein interactions in yeast (Uetz et al. 2000).

Large-scale two-hybrid assays are tools to identify candidates for interacting proteins. However, they may identify interactions erroneously. First, a protein itself may be able to activate transcription without interacting with an activation domain fusion. Second, because of the extensive use of chimeric proteins, misfolding might occur and yield spurious interactions. Third, some proteins might be toxic when expressed in yeast. Fourth, proteins that are coexpressed during the assay might not normally be expressed at the same time or in the same location. Other interactions might be missed, for example, because they are too transient or because a protein has a strong targeting signal directing it to some compartment other than the nucleus. In sum, the map of protein-protein interactions (Uetz et al. 2000) used as the raw material for this study is best viewed as a statistical estimate of the protein interaction network, an estimate with a possibly considerable number of false-positive and false-negative interactions.

Materials and Methods

Protein Interaction Data and Graph Analysis

Data for pairwise interactions among yeast proteins as reported in Uetz et al. (2000) were obtained from http://depts.washington.edu/sfields/projects/YPLM/Nature-plain.html on February 15, 2000. Utilizing the LEDA library of C++ data types (Mehlhorn and Naher 1999), this list was converted into a graph whose nodes represent proteins and whose edges correspond to protein interactions. There were 43 proteins that were reported to interact with themselves. Before further analysis (except for the analysis of self-interacting gene duplicates shown in fig. 6), all such self-interactions were eliminated. Self-interactions are interactions between two protein products of the same gene, such as might occur for homodimerizing proteins. In earlier large-scale two-hybrid studies that used randomly generated DNA libraries, some reported self-interactions may have been due to intramolecular associations between protein domains (Bartel et al. 1996). This is not likely to be a confounding factor here, because full-length cDNA clones of all yeast open reading frames were used in the analysis (Uetz et al. 2000). The resulting protein interaction graph has $n = 985$ proteins that engage in $k = 899$ pairwise interactions. All graph statistics reported below were calculated by exhaustive enumeration using algorithms implemented in LEDA (Mehlhorn and Naher 1999).

Two proteins v_0 and v_i are connected if there exists a path, i.e., a sequence of adjacent nodes $v_0, v_1, \dots, v_{i-1}, v_i$ from v_0 to v_i. The path length l is defined as the number of edges in the shortest path between v_0 and v_i. The characteristic path length L of a graph is the path length between two nodes averaged over all pairs of nodes. Another quantity (Watts and Strogatz 1998) characterizing a graph is the clustering coefficient $C(v)$ of a node v. Consider all k_v nodes adjacent to a node v, and count the number m of edges that exist among these k_v nodes (not including edges connecting them to v). The maximum possible m is $k_v(k_v - 1)/2$, in which case all m nodes are connected to each other. Let $C(v) := m/(k_v(k_v - 1)/2)$ measure the ''cliquishness'' of the neighborhood of v, i.e., what fraction of the nodes adjacent to v are also adjacent to each other. In extension, the clustering coefficient C of the graph is defined as the average of $C(v)$ over all v. It is very close to 0 for very large random graphs.

Random Graph Comparison

Two types of random graphs are explored here: Erdõs-Rényi (ER) random graphs, and random graphs with a degree distribution that follows a power law (PL). An ER random graph consists of n nodes and k edges, where any pair of nodes is equally likely to be connected by one of the k edges (Bollobás 1985). Finding a random graph with a best fit to the observed protein network is made difficult by the fact that a random graph of the same n and k as the yeast protein network has an expected number of $n_0 = 159$ isolated nodes, i.e., nodes that are not connected to any other nodes ($n_0 \approx n[1 - 2k/(n(n - 1)]^{n-1}$ for a sparse random graph). However, by its very nature, the protein interaction network is defined only for nodes that are not isolated. In order to find a random graph that is a best fit for the yeast protein interaction network but does not have any isolated nodes, a random graph with the same k but a larger number n of nodes than the protein network was chosen. With the aid of the above formula, this n was chosen such that when all isolated nodes were removed from the resulting random graph, the remaining graph had (1) an expected number of $n = 999$ nodes, (2) $k = 899$ edges, and (3) no isolated edges, as does the yeast protein interaction network. The required value of n was calculated numerically as $n = 1{,}325$. Random number generators used were based on Mehlhorn and Naher (1999, section 3.2.2) as described by Knuth (1981, p. 92).

PL random graphs are random graphs whose degree probability distribution $P(d)$ is proportional to $d^{-\tau}$ for some constant τ. Such graphs, with a prespecified number of $n = 6{,}279$ nodes and a number of edges k approximately identical to that of the yeast protein interaction network, were generated following a prescription by M. Newman (personal communication). Briefly, a random graph with $n = 6{,}279$ isolated nodes was generated. A node was then chosen at random from this graph. A random integer $d > 0$ with the desired PL distribution was then assigned to this node in the following way. First, a random number $d = \lceil -\gamma * \log(1 - r) \rceil$ was generated, where r is a random real number uniformly distributed in the interval (0, 1), and $\gamma > 0$ is a constant (see below). The symbol $\lceil x \rceil$ refers to the smallest integer greater than x. Second, this number d was accepted with probability $d^{-\tau}$. If d was not accepted, it was discarded, a new d was generated according to the same prescription, and this process was repeated until a d was accepted. Strictly speaking, the resulting distribution of d is a PL with an exponential cutoff, $P(d) \propto d^{-\tau}\exp(-d/\gamma)$. However, a large value of $\gamma = 1{,}000$ was used here, such that the distortion caused by the cutoff was negligible. Once a d was accepted, it was assigned to the randomly chosen node. Another node was chosen at random (without replacement of the previously chosen node), an integer d was assigned to it in the same way, and this process was repeated until the sum S of all the integers assigned to the chosen nodes first exceeded $2k$. The integers assigned to each node correspond to the node's degree. They may also be thought of as the ''stubs'' of edges emerging from a node. Two such stubs were then chosen at random, and the respective nodes were connected via an edge until the reservoir of stubs was exhausted, i.e., until $S/2$ edges had been placed on the graph.

Gene Duplication Data

Data on yeast gene duplicates were obtained from John Conery (Department of Computer Science, University of Oregon) and were generated as described in Lynch and Conery (2000). Briefly, gapped BLAST (Altschul et al. 1997) was used for pairwise amino acid sequence comparisons of all yeast open reading frames as obtained from GenBank. All protein pairs with BLAST alignment scores greater than 10^{-2} were retained for further analysis. Then, the following conservative approach was followed to retain only unambiguously aligned sequences. Using the protein alignment generated by BLAST as a guide, a sequence pair was scanned to the right of each alignment gap. All sequence from the end of the gap through the first ''anchor'' pair of matched amino acids was discarded. All subsequent sequence (exclusive the anchor pair of amino acids) was retained if a second pair of matching amino acids was found within less than six amino acids from the first. This procedure was then repeated to the left of each alignment gap (see Lynch and Conery [2000] for a more detailed description and justification). The retained portion of each amino acid sequence alignment was then used jointly with DNA sequence information to generate nucleotide sequence alignments of genes. For each gene pair in this data set, the fraction K_s of synonymous (silent) substitutions per silent site, as well as the fraction K_a of replacement substitutions per replacement site, was estimated using the method of Li (1993). Of all aligned gene pairs, only those with $K_s < 5$ were included in the analysis presented here. There were 9,059 such pairs (1,041 pairs with $0 \leq K_s < 1$, 2,378 pairs with $1 \leq K_s < 2$, 3,822 pairs with $2 \leq K_s < 3$, 1,403 pairs with $3 \leq K_s < 4$, and 415 pairs with $4 \leq K_s < 5$). The mean and maximum values of K_a for these 9,059 gene pairs were 0.68 (SE $= 2.8 \times 10^{-3}$) and 1.55, respectively. Both member genes were part of the protein interaction network for 387 among these 9,059 genes. Genes with only one paralog and genes that occurred in multigene families were not distinguished here. All results reported are based on estimates of K_s.

Results

The Protein Contact Network Is Superficially Similar to a Random Graph and Shows a PL Degree Distribution

A description of the protein contact network's global structure is useful. The network is best viewed as a graph, a mathematical object consisting of nodes (vertices) and edges. The nodes in the protein contact graph represent proteins. Two proteins are linked by an edge in this graph if they interact. The yeast protein contact graph has $n = 985$ nodes and $k = 899$ edges (fig. 1). According to the available evidence, only about 16%

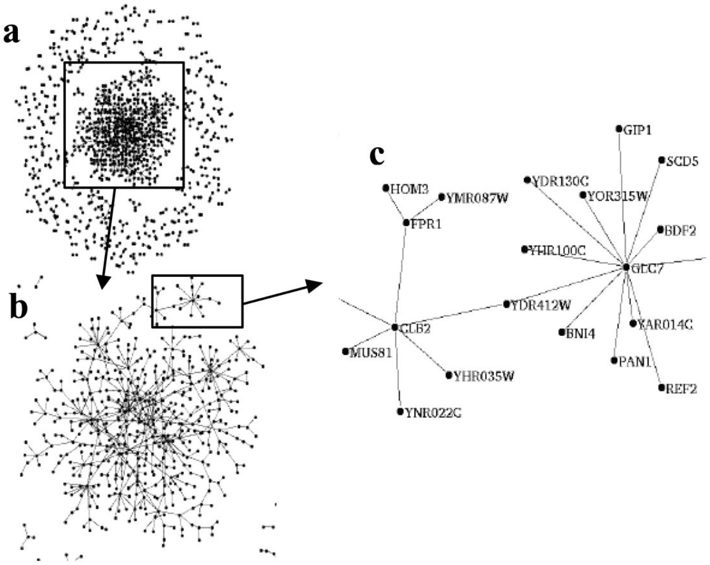


FIG. 1.—Graph representation of the yeast protein contact network. *a,* A two-dimensional drawing of the entire network using a spring embedding algorithm provided by (Mehlhorn and Naher 1999). Each dot corresponds to a protein (node), and each line connecting two proteins corresponds to a contact (edge) between proteins, as reported by Uetz et al. (2000). A group of proteins that interact only with each other and with no other member of the network is called a component, or subnet. Notice the large number of small components surrounding the ''giant'' component in the center. *b,* The giant component of this graph consists of 466 proteins. *c,* A small section of the giant component, with gene or open reading frame names shown next to each node.

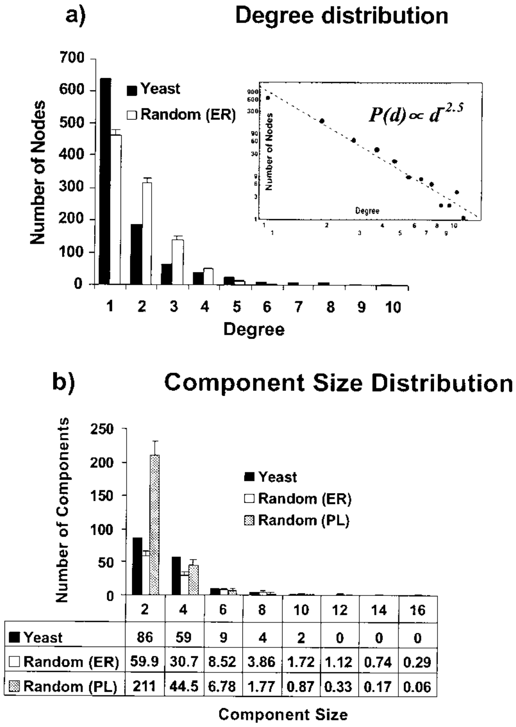
(985/6,279) of all yeast proteins are involved in protein-protein interactions, and the analysis presented below will restrict itself to these proteins. The degree, or connectivity *d*, of a protein is the number of other proteins it interacts with. A component, or subnet, of the graph is a group of proteins that are connected to each other but not to the rest of the network.

Does the protein contact network resemble any graph with known structure? Perhaps the best candidate for such a graph, because of its simplicity, is an ER random graph. An ER random graph is a graph of *n* nodes, where each pair of nodes is equally likely to be connected by one of *k* edges. Indeed, visual inspection of the network shows a feature very typical of random graphs (Bollobás 1985). It has many (163) subnets involving few proteins, and one ''giant'' component with many (466) proteins (fig. 1). Table 1 shows, however, that several descriptors of graph structure differ significantly between the protein interaction network and an ER random graph. Figure 2*a* compares their degree distributions. It demonstrates that the protein network differs in both of these distributions from an ER random graph. A conspicuous feature of the network's degree distribution, $P(d)$, is that it is consistent with a PL, i.e, $P(d) \propto d^{-\tau}$ ($\tau \approx 2.5$; fig. 2*a,* inset), whereas ER random graphs have Poisson-distributed degree (Bollobás 1985). Because the protein interaction network is a small graph, not much statistical confidence can be placed in the ex-

Table 1
Comparison of Statistical Features Between Random Graphs and the Yeast Protein Interaction Network

		RANDOM GRAPHS	
	YEAST	ER	PL (τ = 2.5)
Whole graph			
Nodes	985	984.02 (10.39)	970.7 (81.57)
Degree	1.83	1.85 (0.98)	1.64 (1.76)
No. of components	163	108 (8)*	266.3 (30.6)*
Giant component			
Nodes	466	624.0 (38.7)*	336.9 (86)
Degree	2.3	2.07 (1.05)	2.50 (2.6)
Clustering coefficient ($\times 10^{-3}$)	22	0.59 (0.9)*	4.02 (2.3)*
Characteristic path length	7.14	15.88 (1.76)*	6.01 (1.14)

NOTE.—Numbers in parentheses are the standard deviations of descriptive graph statistics for 100 Erdõs-Rényi (ER) and 100 power law (PL) random graphs. Random graphs were generated and statistics were calculated as described in *Materials and Methods.* Asterisks indicate statistics whose values differ from those of the protein interaction network by more than three standard deviations. The descriptive statistics of no other PL random graph with power $2 < \tau < 3$ fit the protein interaction network better than those displayed for τ = 2.5 (results not shown).



	2	4	6	8	10	12	14	16
■ Yeast	86	59	9	4	2	0	0	0
□ Random (ER)	59.9	30.7	8.52	3.86	1.72	1.12	0.74	0.29
▨ Random (PL)	211	44.5	6.78	1.77	0.87	0.33	0.17	0.06

FIG. 2.—The yeast protein contact network and random graphs. Comparison of the yeast protein contact network (n = 985 nodes, k = 899 edges) with random graphs. Data shown are means and standard deviations over 100 random graphs. *a,* Distribution of the degree (number of contacts or neighbors) in ER random graphs of $n = n_r - n_0$ = 985 nodes, and k_r = 899 edges (see *Materials and Methods*). The protein contact network has an excess of proteins with degree 1, but fewer proteins with a higher degree than the ER random graph. Inset: log-log distribution of degrees in the protein contact network. The degree distribution of power law (PL) random graphs was generated according to this observed exponent. *b,* Histogram of component sizes for protein network, ER random graphs as in *a,* and PL random graphs with power τ = 2.5. Values shown on the y-axis are the numbers of components in each component size bin shown in the uppermost row of values on the x-axis (nodes per component: 1–2, 3–4, 5–6, etc.). The lower rows of values on the x-axis correspond to the (mean) values shown on the y-axis.

act value of the calculated exponent. Also, the statistical structure of the protein network is constrained by more than just its PL degree distribution. This is evident from a comparison of its features with those of a PL random graph with a PL degree distribution ($\tau = -2.5$; fig. 2*b* and table 1). The observed PL is nevertheless remarkable in the context of recent studies showing that metabolic networks (Fell and Wagner 2000; Jeong et al. 2000; Wagner and Fell 2001), as well as a variety of other unrelated graphs (Barabasi and Albert 1999), show such a degree distribution. One key difference of metabolic networks is that the protein interaction network is not a connected graph.

Do Genes in the Protein Network Differ Systematically in Their Propensity to Duplicate?

The effects of a change in gene dosage might be more severe for genes with many interaction partners (high degree). Thus, an attempt was made to determine whether genes in the protein network with paralogs anywhere in the genome had a lower degree than single-copy genes. The mean degrees for the two classes of genes were d = 1.93 (standard error s = 0.086; n = 568) and d = 1.63 (s = 0.074; n = 431), respectively. Thus, although the difference was slight, genes with duplicates appeared more highly connected (F = 4.81; P = 0.03). The reason is unclear. This difference in degree was not significant (F = 0.076; P = 0.78) if more closely related duplicates ($K_s < 1$) were considered (duplicate genes: mean d = 1.87, s = 0.26, n = 53; single-copy genes: mean d = 1.8, s = 0.058, n = 946).

It is conceivable that the size of the subnet a gene is part of influences its propensity to undergo duplications. For genes found outside the largest component of the network, there were mean component sizes of 6.34 (s = 0.62) and 6.49 (s = 0.56) for single-copy genes (n = 191) and duplicated genes (n = 249), respectively, a difference that is not statistically significant (F = 0.03; P = 0.86). This did not change if only duplicates with $K_s < 1$ were considered (results not shown).

Evolution of Shared Interaction Partners Among Closely Related Proteins

Figure 3 illustrates the effect of a gene duplication on gene products involved in protein interactions. Shortly after the duplication, the products of both duplicated genes will have identical interaction partners. Over time, either gene may gain new interaction partners or, perhaps more likely, lose one or more of its interaction partners. The number of shared interactions might be taken as a crude measure of the overlap in the two genes' functions. Eventually, duplicate genes may not interact at all with any of the proteins that they interacted with before duplication, or they may even cease to engage in protein-protein interactions. At least two evolutionary questions can be posed in this context. First, on what timescale does this divergence take place? Second, do most gene duplicates eventually lose all common interactions, or do they retain some common interactions indefinitely? Because absolute duplication time estimates are unavailable for most yeast gene duplications, divergence estimates among duplicates here are based on the fraction K_s of synonymous substitutions per synonymous site (Li 1997). Divergence among synonymous sites is a better indicator of relative times since duplication than many other distance measures between DNA or protein sequences, partly because these sites are under fewer evolutionary constraints than are nonsynonymous sites. The rate of spontaneous mutations per base pair and per round of DNA replication for yeast (2.2×10^{-10}) is similar to those of Drosophila (3.4×10^{-10}) and mice (1.8×10^{-10}; Drake et al. 1998). Based on this observation, and in the absence of measurements of the rate of synonymous substitutions for yeast, it is assumed here that the rate of synonymous substitutions is similar as well. Both mammals and drosophilids are taxa for which rates of synonymous substitutions are known for a variety of genes (Li 1997). Based on this information, the yeast rate is assumed to be the average rate of mammals (3.5 substitutions per site per billion years)

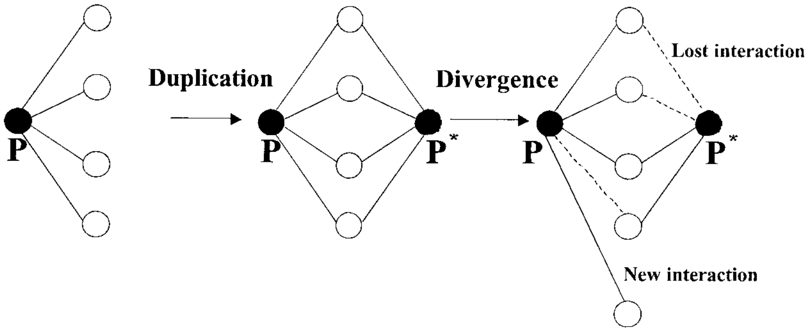


FIG. 3.—The effect of gene duplications on gene products that interact with proteins. Shortly after a gene duplication, the products P and P* of the duplicate genes will interact with the same proteins. Eventually, some or all of the common interactions will be lost, and new interactions may be gained by either protein. In the rightmost panel, protein P has lost one interaction and gained a new interaction partner, whereas protein P* has lost two interactions. If the number of common interaction partners is taken as a measure of functional overlap, then one of the functions of P is also covered by P*, and vice versa.

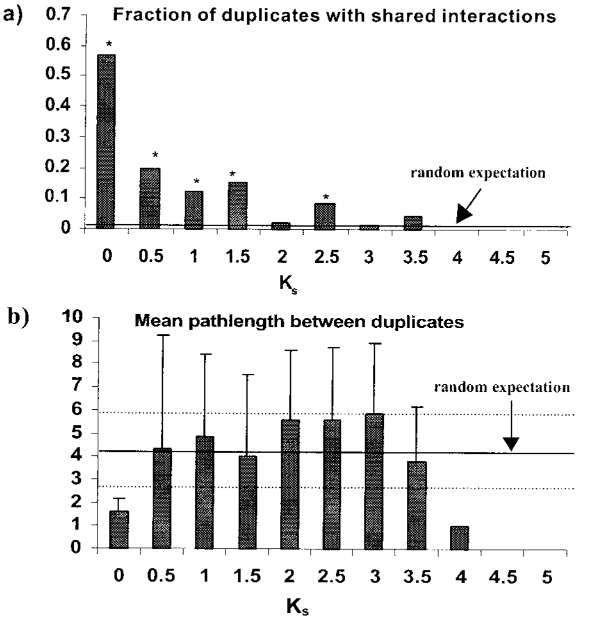


FIG. 4.—Gene duplicates in the same subnet. *a,* Histogram of the fraction of duplicate genes whose products have at least one interacting protein in common as a function of K_s, the fraction of synonymous substitutions per synonymous site (Li 1997). Gene pairs were grouped according to their K_s values into bins of width 0.5 whose lower boundaries are indicated on the x-axis. The horizontal line labeled "random expectation" indicates the estimated probability (0.007; $\sigma \approx 2.6 \times 10^{-3}$) that two proteins chosen at random from the entire network share an interaction partner. This estimate was obtained numerically by randomly choosing 1,000 pairs of proteins from the network. Duplicate gene pairs whose products interact with each other are included in the values shown here. Asterisks above a bar indicate that the number of duplicates with shared interactions is significantly different from the random expectation as assessed by a χ^2 test. The total numbers of gene pairs in each bin are, from left ($K_s < 0.5$) to right ($K_s < 5$), 7, 5, 24, 91, 100, 69, 51, 21, 12, 5, and 2. *b,* Mean and standard deviation of path lengths among products of duplicate genes in the same component as a function of K_s. The solid and dotted horizontal lines indicate mean (4.28) and standard deviation (3.37) in path lengths between two proteins chosen at random from the same subnet within the protein contact graph, as estimated by choosing 1,000 protein pairs at random. Only duplicate genes with $K_s < 0.5$ show path lengths statistically distinguishable from that of two randomly chosen proteins. The total number of gene pairs in each bin is the same as that shown in the legend to figure 5.

and Drosophilids (15.4 per billion years; Li 1997). This yields $K_s = 9.45$ per site per billion years. A $K_s = 1$ between two genes would then indicate that approximately 100 Myr have passed since their duplication. This estimate is consistent with the evolutionary distance (mean $K_s = 1.66$; $\sigma = 0.94$) between 390 genes duplicated in an ancient yeast genome duplication event estimated to have occurred less than 150 Myr ago (Wolfe and Shields 1997).

Two cautionary notes on the use of K_s are in order. Estimates of K_s have substantial margins of error, especially for the large values of K_s studied here, and their interpretation must be approached with great caution. This problem is alleviated by the fact that these estimates are used only for a coarse binning of proteins according to evolutionary distance, and not for any precise estimate of duplication age. While the number of nonsynonymous substitutions at nonsynonymous sites, K_a, is generally much smaller than K_s and could thus be estimated more accurately, it confounds highly conserved ancient duplicates and recent duplicates. This is why it is not used here. A second note of caution concerns the substantial variation in rates of synonymous substitutions across genes. In microbes, the most prominent cause of such variation is codon usage bias of highly expressed genes. Two factors make it unlikely that biased codon usage compromises the analysis carried out here. First, the genes analyzed here have a generally low codon bias index (Bennetzen and Hall 1981; Costanzo et al. 2000) of 0.11 ($\sigma = 0.17$). Only 3.9% of them have codon bias indices greater than 0.4, indicating moderate to high expression. Second, high codon usage bias slows the rate of synonymous substitutions. If genes with high codon usage bias contributed significantly to the evolution of the yeast protein interaction network, they would render the estimates reported here conservative. That is, the network would evolve even faster than estimated below.

How does the fraction of interaction partners shared by duplicate genes evolve? Figure 4*a* shows the fraction of duplicate genes with shared interaction partners as a function of K_s. The line labeled "random expectation" indicates the probability that two proteins share an interaction partner if the two proteins are picked at random from the protein interaction network. Asterisks indicate whether the number of duplicates with shared interactions is significantly different from the random expectation as assessed by a χ^2 test. Strikingly, already for $0.5 < K_s < 1$, only 20% of duplicate gene pairs share an interaction partner. That is, if one applies this criterion of functional overlap, 80% of genes have no functional overlap with their duplicates approximately 100 Myr after the duplication. For $K_s > 2$, the probability that two gene duplicates share an interaction partner approaches the value expected for randomly chosen gene pairs.

A complementary way of studying how fast gene duplicates diverge is to study the length of the path that separates them in the protein contact network. Figure 3 illustrates that either of two proteins with common interaction partners can be reached from the other protein via a path of no more than two edges, i.e., their path length is at most 2. Figure 4*b* shows the mean and standard deviation in path lengths between duplicate proteins as a function of K_s. It shows that the mean path length between products of gene duplicates is less than the mean path length between randomly chosen proteins only for $K_s < 0.5$.

Distribution of Duplicate Gene Products Among Subnets

Proteins that are part of the same subnet may have related biological functions in the sense that they act in the same cellular process. If so, then the function of gene duplicates that are part of the same subnet could be called conserved in this crude sense. Figure 5 shows a histogram of the fraction of duplicates that are part of the same subnet as a function of K_s. For $K_s > 1.5$, the probability that two duplicates are part of the same subnet approaches the probability that two randomly chosen proteins are part of the same subnet. Thus, products of duplicate genes do not remain associated with the same group of interacting proteins.

In sum, duplicate gene products generally do not retain common interaction partners long after duplication. Only 57% (4/7) of the most closely related duplicate gene pairs ($0 < K_s < 0.5$) for which both genes interact with other proteins share any protein interaction partners (fig. 4). For all 380 gene pairs with $K_s > 0.5$, the fraction of duplicate partners with shared interactions is <20%. For $K_s > 1.5$, it dwindles to a value close to the expected number of shared interactions between two proteins chosen at random from within the network. A similar picture emerges for the fraction of proteins that are part of the same subnet. Thus, duplicated gene pairs appear to be reassorted nearly randomly within the protein interaction network. Conserved interactions indicating possible redundancy are the exception rather than the rule.

The Rate of Interaction Loss

Because many characterized protein-protein interactions are responsible for crucial cellular functions, they might not change much over time. However, the rapid divergence of common interaction partners after gene duplication shows otherwise. Under the assumption that this change is caused predominantly by loss of interactions, one can also put a lower bound on the rate at which interactions are lost. There are 127 duplicate gene pairs with $K_s < 2$ where both duplicates engage in protein-protein interactions. Assuming that all of the diversification observed between these duplicates is due to lost interactions, one arrives at a total estimate of 920 interactions immediately after duplication, 429 of which have been lost since. This amounts to a lower bound of $(429/920)(1/200) = 2.3 \times 10^{-3}$/Myr for the probability that a protein interaction is lost. Notice that the actual rate of interaction loss may be much higher, because (1) interactions lost in both duplicate genes cannot be observed and are not accounted for, and (2) many duplicates in this set are younger than 200 Myr. Also, most interactions may be lost shortly after duplication. A very similar estimate is obtained if one corrects for multigene families, admitting only one gene pair per gene family (42 gene pairs, for which 140 of 318 interactions have been lost since duplication, leading to a lower bound of 2.2×10^{-3} lost interactions per million years).

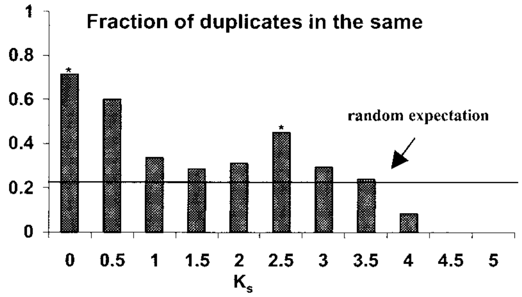


FIG. 5.—Fraction of duplicates in the same subnet. Shown is a histogram of the fraction of duplicate genes whose products are part of the same subnet as a function of K_s, the fraction of synonymous substitutions per synonymous site (Li 1997). Gene pairs were grouped according to their K_s values into bins of width 0.5 whose lower boundaries are indicated on the x-axis. The numbers of gene pairs in each of the bins are (from left to right on the x-axis) 5, 3, 8, 26, 31, 31, 15, 5, 1, 0, and 0. The horizontal line labeled ''random expectation'' indicates the estimated probability (0.231; $\sigma \approx 0.013$) that two proteins chosen at random from the entire network are part of the same subnet. This estimate was obtained numerically by randomly choosing 1,000 pairs of proteins from the network. Asterisks above a bar indicate that the number of duplicates in the same subnet is significantly different from the random expectation as assessed by a χ^2 test. Due to the limited number of genes, interpretation of these results has to be approached cautiously. However, even when bins with fewer than five genes are discounted, it becomes clear that for $K_s > 1.5$, duplicated gene pairs appear to have been reassorted nearly randomly among the subnets of the protein contact graph.

The Evolution of Self-Interactions and New Interactions

Forty-three proteins are reported to interact with themselves (''self-interactors''), 16 of which have one or more paralogs in the yeast genome. There are also 20 paralogous gene pairs whose products interact with each other. Figure 6*a* illustrates the two different routes by which self-interactions and cross-interactions among duplicate genes may evolve. First, a gene product may

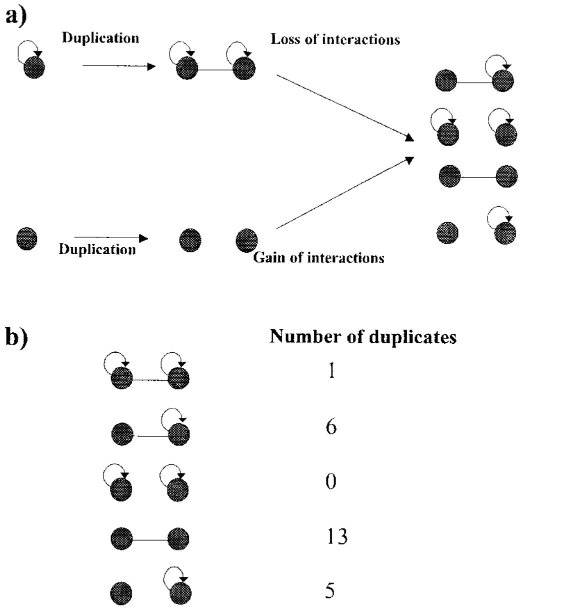


FIG. 6.—Self-interactions and interactions between products of duplicate genes. *a,* Self-interactions of genes with paralogs and interactions between duplicate genes may have evolved by two different routes. First, a gene product may have been a self-interactor before duplication. In this case, observed self-interactions and interactions between paralogs are a reflection of self-interaction before duplication. Second, the interactions may have evolved de novo after the duplication. *b,* Number of paralogous gene pairs observed in the yeast protein interaction networks with the indicated combination of self- and cross-interaction. The last category of five duplicates, in which only one of the paralogs is self-interacting, involves 16 paralogous gene pairs, but 11 of them are redundant. This is because the product of a self-interacting gene, TPK3 (YKL166C), a catalytic subunit of the cyclic AMP-dependent protein kinase, has 12 paralogs that are not self-interactors and thus accounts for 12 of the 16 gene pairs. $K_s < 5$ for all reported gene pairs. Notice the abundance of duplicate pairs without self-interactions (13/25) and the small number of gene pairs (1/25) where both genes are self-interacting.

have been a self-interactor before duplication. In this case, observed self-interactions and interactions between paralogs are most likely a remnant of self-interaction before duplication. Second, the interactions may have evolved de novo after the duplication. Figure 6*b* shows the number of paralogs in yeast with $K_s < 5$ and the indicated combination of self-interactions and cross-interactions. The data set is too small to allow rigorous statistical analysis, but there are at least two conspicuous features. First, there is only 1 out of 25 paralogous pairs for which both proteins show self-interaction. It is not obvious why two duplicates should independently lose the ability to interact with themselves so frequently. Second, for 13 out of 21 paralogous pairs with interactions between the duplicates, neither duplicate shows self-interactions. Thus, there appears to be an abundance of paralogous pairs whose features are more easily explained if one assumes that interactions between duplicates evolve de novo at an appreciable rate.

To obtain a crude estimate of how rapidly new protein interactions might evolve, assume that among the 20 observed interactions between duplicates, only those 13 interactions where neither paralog self-interacts have evolved de novo after the duplication (fig. 6*b*). These 13 genes are among 9,059 duplicate gene pairs with $K_s < 5$. Thus, in the time it took to accumulate five synonymous substitutions per synonymous site, a fraction $13/9{,}059 = 1.44 \times 10^{-3}$ of gene products evolved new interactions. This yields more than 2.88×10^{-6} new interactions per protein pair per million years, if a $K_s = 1$ corresponds to 100 Myr. There are approximately $n = 6{,}280$ open reading frames in the yeast genome, with $n(n - 1)/2 = 1.97 \times 10^7$ possible pairwise interactions. Extrapolating the above estimate to the entire yeast proteome would thus yield $(1.97 \times 10^7)(2.88 \times 10^{-6}) = 56.7$ newly evolved interactions per million years. The multiple caveats to this calculation include uncertainty in the precise number of newly evolved interactions, problems with estimating large K_s values due to biases in correcting for multiple substitutions (Li 1997), and the assumption that all yeast proteins can evolve interactions. Also, many gene pairs in the data set have $K_s < 5$, suggesting that the actual rate of evolution of new interactions is higher. However, even with a possibly large margin of error, the above calculations illustrate that the number of interactions evolving de novo is not negligible.

Discussion

Caveats

Large-scale two-hybrid assays are subject to erroneous identification of protein-protein interactions. However, this will not affect the interpretation of the results as long as the two-hybrid assay is not subject to systematic errors. Whether such errors occur awaits confirmation of the results via other biochemical techniques. Experimental limitations introduce another source of error. To make the large scale study of yeast protein interactions feasible, the number of analyzed interactions was limited to 24 per DNA-binding domain fusion (Uetz et al. 2000). This is, however, not likely to affect the overall result of this analysis, as the mean number of interaction partners per protein is only 1.83 ($\sigma = 1.85$). Extremely ‘‘sticky’’ proteins are thus probably rare. They would not compromise the statistical signal observed here.

Gene functions comprise much more than protein-protein interactions, as is evident from the fact that only 16% of all yeast proteins interact with other proteins. Thus, the value of this analysis would clearly be strengthened if similar results emerged from studies using complementary technologies such as microarray analysis of transcriptional regulation (DeRisi, Iyer, and Brown 1997). The opportunity to use genomic technology to analyze the evolution of many duplicate gene pairs clearly comes at the price of focusing on merely one aspect of function. Consequently, interpretation of the results reported here would have been difficult if protein-protein interactions had been extensively conserved after gene duplication. This is because other aspects of gene function, such as catalytic activity and spatiotemporal expression pattern, might still have diverged. However, exactly the opposite is observed. Even just considering protein-protein interactions, one can conclude that diversification must be extensive after duplication. This might be seen as a contradiction to the observation that up to 40% of synthetic null mutations in yeast open reading frames show weak phenotypic effects (Smith et al. 1996). Many of these mutations involve duplicate genes. However, although widely assumed, redundancy among duplicate genes may not be the cause of this phenomenon (Wagner 2000).

How Do Most Duplicate Gene Products Arrive at Different Subnets?

Most gene duplicates eventually reside in different subnets of the protein interaction graph. If the de novo evolution of new interactions is extremely rare, then only new gene duplications would replenish lost interactions. In this case, duplicate genes would reside in different subnets, because interactions are constantly lost. Originally connected groups of proteins would become disconnected, and the protein interaction network would become increasingly fragmented. However, the observed patterns of interaction between duplicate gene products (fig. 6) suggest that de novo evolution of new interactions does occur. If its rate is sufficiently high, then duplicate gene products may reside within different subnets because new interactions evolved between one of them and gene products in another subnet. The role of interaction loss would then be primarily one of severing existing ties between duplicate gene products. Whether the rate at which new interactions evolve is sufficiently high for this second scenario is an open question.

Interaction Turnover in the Protein Network

Not all protein pairs may be able to engage in physical interactions, and some existing interactions might be absolutely indispensable for cellular function. However, a significant rate of interaction turnover is evident in the protein interaction network. Based on observed interactions between duplicate proteins without self-interaction (fig. 6), it was crudely estimated above that new interactions evolved at a rate of 2.88×10^{-6} per protein pair per million years. This may seem small. However, if extrapolated to all 1.97×10^{7} possible pairwise interactions in the yeast proteome, one arrives at an estimate of 57 newly evolving interactions per million years. Even when restricting oneself to the 985 proteins known to interact with other proteins, one arrives at an estimate of $(2.88 \times 10^{-6})(4.84 \times 10^{5}) \approx 1.4$ newly evolved interactions per million years. The true value is likely to lie somewhere in between.

Based on the assumption that the divergence in protein interactions after gene duplication is largely due to interaction loss, one can put a lower bound on the rate at which interactions get lost at 2.2×10^{-3} per interaction per million years. If a comparable rate holds for interactions between single-copy genes, then 50% of all interactions get lost every 300 Myr. Replenishment of lost interactions would take place through gene duplication and de novo evolution of interactions. Even in the unlikely case that all loss of interaction is restricted to duplicated genes, the rate of interaction turnover would be substantial. This is because the majority (57% for $K_s < 5$) of all genes in the network have one or more paralogs.

Outlook

The observations made here stimulate a multitude of questions regarding their evolutionary significance. Is the change in network structure driven by neutral evolution or by natural selection for advantageous interaction patterns? Similarly, are there many alternative configurations of the network that perform the network's function equally well? Are many protein-protein interactions of little functional significance, or are most of them critical? Does the change in network configurations over time reflect different environments or different adaptations that the organism evolved at different times? Can one understand the network's global structure from a few key parameters, such as the rate of gene duplications and the rate of interaction loss? Perhaps genomic technologies, focusing on the big picture of gene functions, cannot determine the ''blueprint'' of organismal design. At the very least, however, they open new levels of inquiry into the evolution of this design.

Acknowledgments

I thank Drs. Mike Lynch, Bruce Walsh, Rob Miller, and Maggie Werner-Washburne, as well as an anonymous reviewer, for constructive comments and discussion.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. H. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped Blast and Psi-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

BAILEY, W. J., J. KIM, G. P. WAGNER, and F. H. RUDDLE. 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol. Biol. Evol. **14**:843–853.

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

BARTEL, P. L., J. A. ROECKLEIN, D. SENGUPTA, and S. FIELDS. 1996. A protein linkage map of Escherichia coli bacteriophage T7. Nat. Genet. **12**:72–77.

BENNETZEN, J. L., and B. D. HALL. 1981. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

BENTON, B. K., A. TINKELENBERG, I. GONZALEZ, and F. R. CROSS. 1997. Cla4p, a Saccharomyces-Cerevisiae Cdc42p-activated kinase involved in cytokinesis is activated at mitosis. Mol. Cell. Biol. **17**:5067–5076.

BENTON, B., A. TINKELENBERG, D. JEAN, S. PLUMP, and F. CROSS. 1993. Genetic analysis of Cln/Cdc28 regulation of cell morphogenesis in budding yeast. EMBO J. **12**:5267–5275.

BOLLOBÁS, B. 1985. Random graphs. Academic Press, London.

CADIGAN, K. M., U. GROSSNIKLAUS, and W. J. GEHRING. 1994. Functional redundancy: the respective roles of the 2 sloppy paired genes in Drosophila segmentation. Proc. Natl. Acad. Sci. USA **91**:6324–6328.

CIRERA, S., and M. AGUADE. 1998. Molecular evolution of a duplication: the sex-peptide (Acp70a) gene region of Drosophila subobscura and Drosophila madeirensis. Mol. Biol. Evol. **15**:988–996.

Costanzo, M. C., J. D. Hogan, M. E. Cusick et al. (14 coauthors). 2000. The Yeast Proteome Database (YPD) and Caenorhabditis elegans Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res. **28**:73–76.

DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**:680–686.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. Genetics **148**:1667–1686.

Fell, D., and A. Wagner. 2000. The small world of metabolism. Nat. Biotechnol. **18**:1121–1122.

Fields, S., and O. K. Song. 1989. A novel genetic system to detect protein protein interactions. Nature **340**:245–246.

Fromental-Ramain, C., X. Warot, S. Lakkaraju, B. Favier, H. Haack, C. Birling, A. Dierich, P. Dolle, and P. Chambon. 1996. Specific and redundant functions of the paralogous Hoxa-9 and Hoxd-9 genes in forelimb and axial skeleton patterning. Development **122**:461–472.

Gonzalez-Gaitan, M., M. Rothe, E. A. Wimmer, H. Taubert, and H. Jackle. 1994. Redundant functions of the genes knirps and knirps-related for the establishment of anterior Drosophila head structures. Proc. Natl. Acad. Sci. USA **91**:8567–8571.

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.

Joyner, A. L., K. Herrup, B. A. Auerbach, C. A. Davis, and J. Rossant. 1991. Subtle cerebellar phenotype in mice homozygous for a targeted deletion of the En-2 homeobox. Science **251**:1239–1243.

Knuth, D. E. 1981. The art of computer programming. Vol. 2. Seminumerical algorithms. Addison-Wesley, New York.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

———. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

Long, M. Y., and C. H. Langley. 1993. Natural-selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science **260**:91–95.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

McGraith, S., T. Holtzman, B. Moss, and S. Fields. 2000. Genome-wide analysis of vaccinia virus protein-protein interactions. Proc. Natl. Acad. Sci. USA **97**:4879–4884.

Mehlhorn, K., and S. Naher. 1999. LEDA: a platform for combinatorial and geometric computing. Cambridge University Press, Cambridge, England.

Nasmyth, K. 1993. Control of the yeast cell cycle by the Cdc28 protein kinase. Curr. Opin. Cell Biol. **5**:166–179.

Sharman, A. C., and P. W. H. Holland. 1996. Conservation, duplication, and divergence of developmental genes during chordate evolution. Neth. J. Zool. **46**:47–67.

Smith, V., K. N. Chou, D. Lashkari, D. Botstein, and P. O. Brown. 1996. Functional analysis of the genes of yeast chromosome-V by genetic footprinting. Science **274**:2069–2074.

Tautz, D. 1992. Redundancies, development and the flow of information. Bioessays **14**:263–266.

Thomas, J. H. 1993. Thinking about genetic redundancy. Trends Genet. **9**:395–399.

Toda, T., S. Cameron, P. Sass, M. Zoller, and M. Wigler. 1987. Three different genes in S. cerevisiae encode the catalytic subunits of the cAMP-dependent protein kinase. Cell **50**:277–287.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26aa, of Drosophila: divergence versus polymorphism. Mol. Biol. Evol. **15**:1040–1046.

Uetz, P., L. Giot, G. Cagney et al. (20 co-authors). 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature **403**:623–627.

Wagner, A. 2000. Mutational robustness in genetic networks of yeast. Nat. Genet. **24**:355–361.

Wagner, A., and D. Fell. 2001. The small world inside large metabolic networks (in press).

Wang, D. G., J. L. Marsh, and F. J. Ayala. 1996. Evolutionary changes in the expression pattern of a developmentally essential gene in 3 Drosophila species. Proc. Natl. Acad. Sci. USA **93**:7103–7107.

Wang, Y. K., P. N. J. Schnegelsberg, J. Dausman, and R. Jaenisch. 1996. Functional redundancy of the muscle-specific transcription factors Myf5 and Myogenin. Nature **379**: 823–825.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of small-world networks. Nature **393**:440–442.

Wilkins, A. 1997. Canalization: a molecular genetic perspective. Bioessays **19**:257–262.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

Zhang, J. Z., 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.

Simon Easteal, reviewing editor

Accepted March 11, 2001