Mutational Robustness and Asymmetric Functional Specialization of Duplicate Genes

Andreas Wagner

SFI WORKING PAPER: 2002-02-006

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu

SANTA FE INSTITUTE

Mutational robustness and asymmetric functional specialization of duplicate genes.

Andreas Wagner

University of New Mexico

University of New Mexico
Department of Biology
167A Castetter Hall
Albuquerque, NM 817131-1091
Phone: +505-277-2021
FAX: +505-277-0304
Email: wagnera@unm.edu

Abstract: Most duplicate genes are eliminated from a genome shortly after duplication, but those that remain are an important source of biochemical diversity. Much of their diversification arises via functional “specialization”, loss of some functions of the duplicates remaining in the genome. I here present evidence from genome-scale protein-protein interaction data, microarray expression data, and large-scale gene knockout data that this diversification is often asymmetrical: one duplicate usually shows significantly more molecular or genetic interactions than the other. I propose a model that can explain this divergence pattern if duplicate gene pairs are less likely to suffer deleterious mutations when having diverged asymmetrically. The data may provide the first evidence that natural selection has increased mutational robustness in genetic networks.

Soon after a gene duplication, degenerative mutations are likely to eliminate duplicate genes from the genome (Li 1997; Lynch and Conery 2000). However, gene duplications occur continuously and at high rates in eukaryotes, which accounts for the fact that up to 50% of a eukaryotic genome may consist of duplicate genes (Lynch and Conery 2000; Rubin *et al*. 2000). These persisting duplicate genes are perhaps the most prominent source of biochemical innovation of gene products. However, little is known about how this innovation occurs or about how gene duplicates diverge in general.

Studying functional divergence among duplicate genes requires a definition of gene function, but no universal such definition is possible. The reason is that there are several complementary ways of classifying gene functions. For instance, gene products can be characterized biochemically, e.g, as enzymes or transcription factors. Second, they can be characterized through their time and locus of expression, e.g., expression during a cell cycle stage, in the cytoplasm, or during brain development. Third, they can be characterized genetically through mutations and through other genes that these mutations affect. This list is not necessarily complete.

Functional genomics has added much information to each of these categories, especially in model organisms like the yeast *Saccharomyces cerevisiae.* First, monitoring expression through microarrays (Chu *et al*. 1998; Eisen *et al*. 1998; Gasch *et al*. 2000; Spellman *et al*. 1998) provides spatiotemporal expression information for thousands of genes at once. This information is indicative of the biological process a gene is involved in. Second, genome-wide protein-protein interactions can characterize physical interactions among thousands of gene products (Bartel *et al*. 1996; Fromont-Racine *et al*. 1997; Ito *et al*. 2001; Uetz *et al*. 2000). Third, large-scale gene knockout screens in combination with microarray experiments indicate which genes' expression level are affected by a mutated gene. Thus, even in the absence of a detectable phenotype – all too frequent in knock-out experiments – a putative function can sometimes be assigned using genetic interactions with known genes.

Attempts to identify gene functions according to any of the above criteria, whether they use genomic or pre-genomic techniques, yield one key message: most genes have more than one, if not many functions. They are expressed at multiple times and in multiple places, they affect multiple biological processes when mutated, or they interact with proteins with diverse biochemical and biological roles (Bender *et al*. 1983; Gerhart and Kirschner 1998; Jack and Delotto 1995; Kirchhamer *et al*. 1996; Li and Noll 1994; Schwikowski *et al*. 2000; Slusarski *et al*. 1995; Wagner 2001). This multi-functionality has important implications for the divergence of duplicate genes: duplicate genes often diverge through functional specialization, loss of complementary (sub)functions in each duplicate (Force *et al*. 1999; Lynch and Force 2000; Wagner 2000b). Examples abound. To name but two, the *ZAG1* and *ZMM2* genes are paralogues in the maize genome. They are orthologues of the Arabidopsis *AGAMOUS* gene which is involved in carpel and stamen development. Each of them appears to have largely lost one of their ancestral expression domains: *ZAG1* is expressed at high levels in developing carpels and *ZMM2* is expressed in developing stamens. A null-mutation in *ZAG1* affects only early carpel development. (Coen and Meyerowitz 1991; Mena *et al*. 1996; Schmidt *et al*. 1993). Force and collaborators (Force *et al*. 1999) report on the zebrafish *engrailed* genes *eng1* and *eng1b*, the likely results of a teleost-specific gene duplication of the tetrapod *En1* gene. In mice and chicken, *En1* is expressed in the developing pectoral appendage bud and in specific neurons of the developing hindbrain and spinal cord. In zebrafish, *eng1* retained expression in the pectoral appendage bud, whereas *eng1b* is only expressed in the hindbrain and the spinal cord. Similar patterns of divergence may be quite common in zebrafish (Ekker *et al*. 1997; Ekker *et al*. 1995; Lee *et al*. 1996).

Studies focussing on individual gene pairs fall short of identifying general divergence patterns of many duplicate genes. At first sight, analyzing functional divergence of many duplicate genes may seem like a hopeless task. Because it is not even straightforward to *classify* one gene’s function, how would one *compare* the functions of many divergent duplicates? Functional genomic experiments provide a crude remedy for this problem. Despite their disadvantage of providing largely qualitative information about genetic and molecular interactions of genes, their great advantage is that they do so for thousands of genes at once. They thus yield insight about one aspect – however minute – of gene function, such as the protein interaction partners of a gene, gene expression patterns affected through mutating a gene, or the response of gene expression to environmental challenges. It is this aspect of gene function I will focus on.

Methods

Gene duplication data. Data on yeast gene duplicates was kindly provided by John Conery (University of Oregon, Department of Computer Science) and was 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 a BLAST alignment score 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 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 were estimated using the method of Li (Li 1993).

Protein interaction data and analysis. Data for 899 pairwise interactions among 985 yeast proteins, as reported in (Uetz *et al.* 2000), was obtained from http://depts.washington.edu/sfields/projects/YPLM/Nature-plain.html on February 15, 2000. There are 43 proteins that were reported to interact with themselves. Before further analysis 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.) The resulting protein interaction network was then represented as a graph using LEDA (Mehlhorn and Naher 1999). Within this graph representation, common and different protein interactions among gene family members are easily analyzed (Wagner 2001). To analyze protein interaction data not generated by two-hybrid experiments, I used information on physical interactions among yeast proteins obtained from the MIPS database (Mewes *et al*., 1999, http://mips.gsf.de/proj/yeast/CYGD/db/index.html). I eliminated from this data all protein interactions confirmed by two-hybrid experiments. The remaining 899 interactions involve 680 proteins. I did not distinguish between genes with only one paralogue and genes that occur in multigene in the analysis of either data set.

I tested the observed pattern of the number of interactions for products of paralogous yeast genes against a null-hypothesis of symmetric divergence through *loss-of-interactions*. According to this hypothesis, for two proteins P and P* that (i) are products of duplicate genes, that (ii) have d_1 and d_2 protein interactions, respectively, and that (iii) share b of these interaction partners (Fig. 1a), the d_1-b and d_2-b non-shared interactions of P and P* are due only to loss of interactions after the duplication. In other words, immediately after the gene duplication leading to P and P*, each protein had $d_{1+}d_2$-b interactions with the same proteins. P lost d_2-b of these interactions, whereas P* lost d_1-b interactions. Central to this null-hypothesis is that both proteins have equal probability to lose an interaction, that is, that the total number of lost interactions is partitioned between the two proteins according to a binomial distribution B($d_{1+}d_2$-$2b_,$ ½). The many paralogous gene pairs to be analyzed have different and usually small values for $d_{1,}$ d_2, and b. Thus, it is most expedient to test the null-hypothesis by numerically generating an expected distribution of divergence values (Figs 2b, 3c). This distribution can then be compared with the observed distribution.

A second null-hypothesis is that of symmetric divergence through *gain-of-interactions.* It assumes that immediately after duplication, the two proteins P and P* had only b interactions, and that P gained d_1-b interactions, whereas P* gained d_2-b interactions. Symmetric divergence means that each protein is equally likely to gain an interaction, i.e., that the total number of gained interactions is partitioned between the proteins with a binomial distribution B($d_{1+}d_2$-$2b$, ½).

Environmental stress and gene expression. To assay the differential expression response of yeast paralogues to environmental stresses, I used data provided by Gasch and collaborators (Gasch *et al*., 2000; http://www-genome-stanford.edu/yeast_stress) for the following conditions: heat shock (25°C to 37°C, after 30 minutes), reverse heat shock (37°C to 25°C, 30'), H_2O_2 and Menadione exposure, both of which generate reactive oxygen species (60' and 80', respectively), dithiothreitol, a reducing agent interfering with protein folding (90'), diamide, an agent oxidizing sulfhydryl groups, (40'), hyperosmotic shock mediated by 1M sorbitol (60'), hypo-osmotic shock mediated by transfer of cells from 1M sorbitol to medium lacking sorbitol (30'), amino acid starvation (2 hours), nitrogen depletion (1d), and stationary phase (7d). I considered genes whose expression level was changed at least three-fold relative in response to a stressor to be affected significantly. Because the expression response to most environmental stresses is transient, I chose a time point (indicated above in parentheses) approximately halfway through the measured response time series for each environmental stress to assess significant change. I then counted the number of stressors to which each member of a paralogous gene pair responded, and did so for all duplicate pairs with $0.5<K_s<3$. Due to cross-hybridization, very closely related duplicates can not be distinguished through microarray analysis. However, the analysis of Gasch (Gasch *et al*. 2000, Fig. 5) suggests that gene pairs with K_s>0.5 are readily distinguishable. I excluded paralogues with K_s<0.5 from the analysis. For 2.6% of all paralogous gene pairs, there was at least one stress condition where the expression of one gene was induced and that of the other was repressed. I excluded all such gene pairs from the analysis.

Gene perturbations and gene expression. Data summarizing the effects of 271 gene deletions (and other treatments) on gene expression was made available as supplemental material to (Hughes *et al*. 2000), file 'data_expts_1-300_ratios.txt'. From this data set, which contains $\log_{10}$-transformed expression ratios of 6312 genes for each mutation, I eliminated all data derived from haploid and aneuploid deletion strains, as well as data on non-genetic treatments. The remaining data contains information on null-mutation effects for a total of 21 paralogous gene pairs, the most closely related 11 of which (K_s<3.2) are discussed in here. For each member gene of each paralogue, I determined what other genes were affected in their expression level by a synthetic-null mutation in the gene. I also determined the number of genes that were affected by a null-mutation in each paralogue. I considered a gene as affected by a null-mutation if its level of mRNA expression had changed by more then 3-fold in response to the mutation.

The tests of two null-hypothesis of divergence proceed analogously to those above, except that because of the larger number of genes affected, a meaningful test is possible for each gene pair individually. Let d_1 and d_2 be the number of genes affected by a synthetic null-mutation in gene 1 and 2, respectively, of a paralogous pair. Let b be the number of genes affected by both mutations. The first null-hypothesis is that all divergence is due to equiprobable loss-of-effects on other genes. An exact test is most expedient in this case. It assumes that immediately after the duplication, a null mutation in either duplicate would have affected d_1+d_2-b other genes. $l_1=d_2-b$ and $l_2=d_1-b$ of these effects were subsequently lost in gene 1 and 2, respectively, adding to a total of d_1+d_2-2b lost effects. A disparity between l_1 and l_2 indicates asymmetry in divergence. The probability **P** of a disparity as big or bigger as that actually observed, by chance alone, is calculated by summing over the tails of a binomial distribution B(d_1+d_2-2b, ½), so

$$\mathbf{P} = 2\sum_{0}^{\min(l_1,l_2)} \binom{d_1+d_2-2b}{1/2}\left(\frac{1}{2}\right)^{d_1+d_2-b} \qquad l_1 \neq l_2,$$

where **P**=1 for $l_1=l_2$. The factor 2 in front of the summation sign indicates that this is a two-tailed test. The second null-hypothesis, that all divergence is due to equiprobable gain-of-effects by the two duplicates, is tested in the same way. The only difference is that *min(l_1,l_2)* above is replaced with *min(g_1,g_2),* where $g_1=d_1-b$ and $g_2=d_2-b$ is the number of effects gained.

Results and Discussion

Asymmetric divergence in protein-protein interactions. Genome-scale screens of protein interactions using the yeast two-hybrid assay have been carried out in several organisms (Bartel *et al*. 1996; Fromont-Racine *et al*. 1997; Ito *et al*. 2000; Uetz *et al*. 2000). Their results are comprehensive maps of protein-protein interactions comprising most proteins encoded by a genome. Interpreting these maps is still difficult, because they may contain significant numbers of false positive and false negative interactions (Ito *et al*. 2001), and because they collapse the spatial and temporal dimensions of gene expression into a still-life image of protein interactions. However, these maps have also demonstrated usefulness in predicting the spatial expression domain and functional annotation of many proteins from their interaction partners (Schwikowski *et al*. 2000). They can also answer questions about global patterns of interactions, questions whose answer does not depend on the veracity of each individual interaction, but only on statistical interaction patterns.

Gene duplications are important in shaping protein interaction networks. For instance, more than 50% of yeast genes whose products interact with proteins are part of gene families (Wagner 2001). How do gene duplications affect protein interactions? Fig. 1 shows a hypothetical protein P that interacts with four other proteins. Immediately after duplication of the gene encoding P, P and its duplicate P^* share all four interactions. As the duplicates diverge in sequence, they also diverge in their protein interactions. Each protein may occasionally gain new interactions. But if mutations are more likely to cause loss of an interaction, as suggested by the prevalence of degenerative mutations in general (Li 1997), then most divergence will be due to loss of originally common protein interactions. Here, I use the number of interaction partners a protein has as a crude one-dimensional indicator of protein function. The number of common and different interactions between two duplicates then indicates their functional divergence.

Fig. 2a shows the number of interaction partners for 2185 pairs of paralogous genes in the network described by Uetz and collaborators (Uetz *et al*. 2000). These comprise all paralogous pairs with K_s<3 synonymous substitutions per synonymous site, corresponding - albeit with a large error margin - to genes duplicated within the last 300Myrs (Wagner 2001). The abscissa and ordinate axes show the number of protein interactions for the first and second protein member of each pair. The number of common interactions in these pairs is small: even among the most recent paralogues (K_s<0.5) less than 60 percent share any interactions at all, and this number dwindles to less than 15 percent for more distant paralogues (K_s>1) (Wagner 2001).

Fig. 2a shows a distinct L-shape, indicating that in many protein pairs, where one partner has many interactions, the other one has disproportionately few. This negative correlation in the number of interaction partners between duplicates is statistically highly significant (Spearman r_s=-0.62, **P**<<10-3; Pearson r=-0.185, **P**<<10-3, df=2183). Could it have occurred by chance alone, that is, through random equiprobable loss of interactions in either member of a pair? Fig. 2b shows the results of testing this hypothesis, a hypothesis based on three assumptions. First, all observed differences in interactions among paralogues are due to loss of originally common interactions. Second, each protein interaction of the preduplication state is lost with equal probability in each duplicate (symmetric divergence). Third, the total number of lost interaction per gene pair is the same as that observed empirically. The plot shown in Fig. 2b stems from a stochastic simulation of the divergence of 2185 genes using these assumptions. The L-shape of the plot in Fig. 2a disappears in this scenario of symmetric divergence, as does the highly negative statistical association (Spearman r_s=-0.1, **P**<<10-3; Pearson r=0.44, **P**<<10-3, df=2183).

Another possible, although less likely null-hypothesis of divergence, is the independent and equiprobable (symmetric) *gain* of interactions in the two paralogues. Simulating the divergence of paralogues under this null-hypothesis also does not reproduce the characteristic L-shape and the highly negative statistical association of the empirical data (Fig. 2c; Spearman r_s=-0.1, **P**<<10-3; Pearson r=0.42, **P**<<10-3, df=2183).

Independent genome-scale two-hybrid experiments using different experimental designs (Ito *et al*. 2001; Uetz *et al*. 2000) show limited overlap in the interactions they detect. It is thus advisable to assure that the observed patterns of divergence are not artefacts of a particular experimental technique. I have repeated the above analysis with yeast protein interaction data taken from the MIPS database (Mewes *et al*. 1999), from which I eliminated all protein interaction information generated by two-hybrid experiments. The remaining 899 interactions among 680 yeast proteins have been experimentally confirmed using techniques ranging from Western blotting to coimmunoprecipitation. The global pattern of interactions among paralogues follows closely that of the two-hybrid data, an L-shaped distribution indicating asymmetry (Fig. 3a) and a highly negative statistical association (Spearman r_s= - 0.57, **P**<<10-3; Pearson r=-0.2, **P**<<10-3, df=1803). This pattern is not explicable through either symmetric loss of interactions (Fig. 3b; Spearman r_s=0.11, **P**<<10-3; Pearson r=0.5, **P**<<10-3, df=1803) or symmetric gain of interactions (Fig. 3c; Spearman r_s=0.06, **P**<<10-3; Pearson r=0.48, **P**<<10-3, df=1803).

In sum, protein interactions among products of duplicate genes diverge asymmetrically, i.e., one paralogue has more protein interactions than the other. This asymmetry is statistically highly significant and is neither explicable through independent (equiprobable) loss of function in the duplicates, nor through independent gain of function.

Asymmetric response to environmental stresses. Unicellular organisms like yeast have evolved elaborate cellular responses allowing them to adapt to drastic environmental changes. They can not only withstand fluctuations in temperature, osmolarity, environmental acidity, and types and quantity of nutrients. They can also survive the influence of radiation and toxic chemicals. During environmental change, many genes alter their transcriptional activity. Such changes in mRNA expression profile provide valuable insights into gene functions (Chu *et al*. 1998; Eisen *et al*. 1998; Gasch *et al*. 2000; Spellman *et al*. 1998). A recent study examined the genomic mRNA expression response of most yeast genes to a variety of environmental stressors (Gasch *et al*. 2000). To assess the differential response of duplicate genes to these stressors, I analyzed data from 11 different stress-responses, including heat shock, hyperosmotic shock, amino acid, and nitrogen starvation (Gasch *et al*. 2000). I excluded the most closely related paralogues (K_s<0.5) from the analysis, because cross-hybridization does not allow them to be distinguished by microarray analysis. I also excluded paralogues where neither member gene responded to any environmental stressor. For the remaining 3815 paralogous gene pair with 0.5<K_s<3, I identified the number of stressors to which each member of the pair responds.

There is again a pronounced asymmetry in the response of gene duplicates to these stresses, as indicated by a significantly negative statistical association between the number of stresses the first and second gene respond to (Spearman r_s=-0.33, **P**<<10-3; Pearson r = -0.12, **P**<<10-3, df=3813). Completely analogous to the tests for symmetric divergence in protein interactions, I analyzed whether this association is consistent with the null-hypothesis that the paralogues originally responded identically to these 11 stresses, but that some responses got lost and did so with equal probability between the duplicates. This null-hypothesis of symmetric divergence must be rejected (Spearman r_s=-0.0094, **P**>0.5; Pearson r=0.18, **P**<<10-3, df=3813). Nor is the observed pattern explicable if the paralogous genes gained stress responses independently and equiprobably since the duplication (Spearman r_s=-0.0087, **P**>0.5; Pearson r=0.18, **P**<<10-3, df=3812).

In sum, the distinct asymmetry in divergence observed for protein interactions also holds for another aspect of gene function, the response to environmental stress.

Asymmetric response to genetic perturbations. The results of a large-scale gene perturbation experiment in yeast, involving several hundred gene-knockout mutations in combination with microarray measurements of changes in the expression of 6312 yeast genes, have been reported (Hughes *et al*. 2000). Measuring the effect of a null mutation in a gene on the expression of all other genes does not distinguish between direct and indirect effects of the mutation. Its advantage, however, is that it is a very comprehensive means to assay genetic interactions.

For the purpose of this paper it is relevant that the available data (Hughes *et al*. 2000) contains information on the knockout effect of 11 moderately distant paralogous gene pairs with K_s<3.5. For these 11 gene pairs, I compared the number of genes whose expression is affected by a null-mutation in each member of the pair (Table 1). Interpreting differences between paralogues in the number of affected genes is complicated because these differences are not only the result of divergence between the paralogues. They also include effects from the divergence of genes interacting with each paralogue. However, the advantage of a perturbation approach is that it provides a more comprehensive assessment of functional differences between paralogues than a mere analysis of direct physical protein interactions. It exposes how the effects of a mutation ripple through a transcriptional regulation network.

Analogous to the analysis above, one can ask whether the observed differences between paralogues can be attributed to independent and equiprobable loss of genetic interactions, or to independent gain of interactions since the duplication. For 7 out of 11 gene pairs in Table 1, both these null-hypotheses must be rejected. That is, these seven gene pairs show statistically significant asymmetries in divergence. Eliminating one of two paralogous genes affects a substantially greater number of other genes than eliminating the other.

Asymmetric divergence and mutational robustness. The number of genes whose expression is affected by gene perturbations is moderately large. This makes it possible to derive statistical evidence for asymmetric divergence of paralogous genes from individual gene pairs. Seven out of 11 perturbed gene pairs show such evidence. The number of environmental stresses to which a gene responds is typically smaller. And so is the number of protein interactions of gene products. These smaller numbers make it more difficult to derive solid evidence for asymmetric divergence from individual gene pairs. However, such evidence emerges when analyzing multiple gene pairs.

It can not be excluded that some divergence among duplicates arises through mutations causing gain of interactions, gain of stress responses etc. However, shortly after a duplication, such mutations are probably rare compared to degenerative loss-of-function mutations (Force *et al*. 1999; Li 1997). Consider the example of a gene's transcriptional response to a stressor. Given the small size of yeast transcriptional regulatory regions, mutations are much more likely to eliminate an existing transcription factor binding site than to generate a new site mediating a stress response. Divergence of functions through gain of transcription factor binding sites—while probably occurring occasionally—must be much less frequent than loss of binding sites. I will thus focus on a simple model of divergence through loss of common functions. Figure 4 explains the basic idea behind this model. It applies to the divergence of genes that have several suitably defined functions (represented by white boxes in Fig. 4a), as indicated by observed molecular interactions or patterns of gene expression. Immediately after a duplication, two duplicates are identical in all these functions. The model makes only two assumptions about the process of divergence, both of them very simple. First, every function must be exercised by at least one of the two genes. Organisms where this does not hold will suffer reduced fitness. Second, a loss-of-function mutation (i) affects each of the duplicates with equal probability (1/2), and (ii) eliminates one of the affected gene's functions. In this context, what is the probability P_Δ of suffering a deleterious mutation if the two duplicates have diverged symmetrically vs. asymmetrically? Asymmetric divergence means that one duplicate has lost more functions than the other. Assume that since the duplication, duplicates 1 and 2 have lost a fraction l_1 and l_2 of their functions, respectively ($0<l_1, l_2\leq 1$). Let $l=l_1+l_2$ be the total fraction of functions lost ($0\leq l\leq 1$). If no function is allowed to have been lost in both genes, the probability that a mutational loss of one further function has a deleterious effect is equal to

$$P_\Delta = \frac{1}{2}\left(\frac{l_2}{1-l_1}\right)+\frac{1}{2}\left(\frac{l_1}{1-l_2}\right).$$

Upon expressing l_1 and l_2 in terms of the total fraction of functions lost, l, using x:=l_1/l ($1-x=l_2/l$) this expression becomes

$$P_\Delta(x,l) = \frac{1}{2}\left(\frac{1-x}{(1/l)-x}\right)+\frac{1}{2}\left(\frac{x}{(1/l)-1+x}\right).$$

In this expression, a value of $x=0.5$ indicates symmetric divergence. The pertinent feature of $P_\Delta(x,l)$ is that it is unimodal: Regardless of l, it has a maximum at $x=0.5$ (Fig. 4b). This means that the probability of a deleterious mutation is greatest if two genes have diverged symmetrically. Thus, asymmetric divergence minimizes the risk of deleterious mutations.

If this model is correct, we see asymmerically diverged gene pairs because organisms harboring them have survived preferentially in the past. Importantly, natural selection would act in an indirect, “second order” manner on such gene pairs. In a population polymorphic for gene duplicates at different stages of divergence, different individuals would not necessarily have different fitness. Rather, the propensity of such individuals to suffer deleterious mutations would be different. Individuals with symmetrically diverged duplicates would be preferentially eliminated from the population through deleterious mutations.

One might assume that the selective advantage of having asymmetrically diverged gene duplicates must be minute. After all, differences in fitness do not manifest themselves until new loss-of-function mutation arise. For any organism, the expected waiting time for such a new loss-of-function mutation is proportional to the inverse of the mutation rate μ (Hartl and Clark 1988). During this time, symmetrically diverged gene duplicates are free to go to fixation via random drift. Formal population genetic analysis (Wagner 2000b) shows that for sufficiently large population sizes ($N>1/\mu$), the lens of natural selection has sufficient resolving power to perceive differences in mutational robustness and act on them. For microorganisms like yeast, attainable population sizes may well be in the required range. In addition, this minimally required population size is based on the evolution of only one diverging gene pair (Wagner 2000b). It may be much smaller for multiple gene pairs, and their cumulative effects on mutational robustness.

The requirement for large effective population sizes suggests a test for the model. In organisms with small effective population sizes, such as many higher vertebrates, we would not expect asymmetric divergence of gene duplicates. (The necessary data is not yet available.) A requirement for persistently large population sizes may also be one of the reasons why the asymmetry observed is not perfect and does not hold for all genes. Depending on a gene and its functions, a loss-of-function mutation may have very subtle fitness effects. In conjunction with fluctuating effective population sizes, the selection pressures for asymmetrically divergence may fluctuate as well. Some genes thus diverge symmetrically whereas others do not.

Mutational robustness – the insensitivity of a biological process to mutations – is a prominent feature of genetic networks. It occurs in all organisms, from single-celled to mammalian. It occurs in processes as different as metabolism and development. And it has been detected using every approach from classical genetics to functional genomics (Hartl *et al*. 1985; Hartman *et al*. 2001; Rendel 1979; Rutherford and Lindquist 1998; Smith *et al*. 1996; vonDassow *et al*. 2000).

There are two possible evolutionary origins for observed robustness. First, whenever gene products interact to perform a biological function, the resulting gene network and its function might show some resilience to mutations. Second, mutational robustness may be a result of stabilizing natural selection, sorting more robust genetic networks from less robust ones on enormous time scales. The second hypothesis, if true, has profound consequences on our understanding of biological evolution. It implies that natural selection itself can change the amount of phenotypic variation produced by a genetic network through mutations. It could thus also affect any further response of a population to natural selection – the population's evolvability. Despite these intriguing consequences, empirical evidence for evolved robustness is hard to come by. First, robustness is most likely a global property of genetic networks (Edwards and Palsson 2000; Wagner 2000a), and thus not easily subject to experimental manipulation. Second, if evolution of robustness occurs, it probably occurs on very large time scales. The third reason has to do with the detection of robustness itself. It is impossible to prove by direct experimentation that phenotypic robustness to mutations translates into robust fitness. The reasons are twofold. First, experimentally undetectably small fitness differences in organisms may be perceived by natural selection. Second, it is usually impossible to reconstitute in the laboratory an environment representing all environments in which an organism may need to survive and reproduce.

For all these reasons, our best chance of assessing whether genuine robustness exists and evolves is to detect patterns of molecular evolution consistent with it (Hirsh and Fraser 2001). Patterns like the one detected here may thus be ideally suited to answer two difficult but important questions in one sweep. First, do differences in mutational robustness exist? And second, are they relevant for evolution, are they seen by natural selection? The data presented here suggest the answer to both questions is yes.

Acknowledgments. Financial support through NIH grant GM63882 is gratefully acknowledged.

References

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 Research **25:** 3389-3402.

Bartel, P. L., J. A. Roecklein, D. SenGupta and S. Fields, 1996 A protein linkage map of Escherichia coli bacteriophage T7. Nature Genetics **12:** 72-77.

Bender, W., M. Akam, F. Karch, P. A. Beachy, M. Peifer, P. Spierer, E.B. Lewis and D.S. Hogness, 1983 Molecular-Genetics of the Bithorax Complex in Drosophila-Melanogaster. Science **221:** 23-29.

Chu, S., J. Derisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz, 1998 The transcriptional program of sporulation in budding yeast. Science **282:** 699-705.

Coen, E. S., and E. M. Meyerowitz, 1991 The War of the Whorls: Genetic Interactions Controlling Flower Development. Nature **353:** 31-37.

Edwards, J. S., and B. O. Palsson, 2000 The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proceedings of the National Academy of Sciences of the United States of America **97:** 5528-5533.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein, 1998 Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America **95:** 14863-14868.

Ekker, M., M. A. Akimenko, M. L. Allende, R. Smith, G. Drouin, R.M., Langille, E.S. Weinberg and M. Westerfield, 1997 Relationships among msx gene structure and function in zebrafish and other vertebrates. Molecular Biology and Evolution **14:** 1008-1022.

Ekker, S. C., A. R. Ungar, P. Greenstein, D. P. Vonkessler, J. A. Porter, R.T. Moon and P.A. Beachy1995 Patterning Activities of Vertebrate Hedgehog Proteins in the Developing Eye and Brain. Current Biology **5:** 944-955.

Force, A., M. Lynch and J. Postlethwait, 1999 Preservation of duplicate genes by subfunctionalization. American Zoologist **39:** 460-460.

Fromont-Racine, M., J. C. Rain and P. Legrain, 1997 Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nature Genetics **16:** 277-282.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P.O. Brown, 2000 Genomic expression programs in the response of yeast cells to environmental change. Molecular Biology of the Cell **11:** 4241-4257.

Gerhart, J., and M. Kirschner, 1998 *Cells, embryos, and evolution*. Blackwell, Boston.

Hartl, D. L., and A. G. Clark, 1988 *Principles of Population Genetics*. Sinauer associates, Sunderland MA.

Hartl, D. L., D. E. Dykhuizen and A. M. Dean, 1985 Limits of Adaptation : the Evolution of Selective Neutrality. Genetics **111:** 655-674.

Hartman, J. L., B. Garvik and L. Hartwell, 2001 Cell biology : Principles for the buffering of genetic variation. Science **291:** 1001-1004.

Hirsh, A. E., and H. B. Fraser, 2001 Protein dispensability and rate of evolution. Nature **411:** 1046-1049.

Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, Armour, C. D. H.A. Bennett, E. Coffey, H.Y. Dai, Y.D.D. He,M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend, 2000 Functional discovery via a compendium of expression profiles. Cell **102:** 109-126.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of the United States of America **98:** 4569-4574.

Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara and Y. Sakaki 2000 Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proceedings of the National Academy of Sciences of the United States of America **97:** 1143-1147.

Jack, J., and Y. Delotto, 1995 Structure and Regulation of a Complex Locus : the Cut Gene of Drosophila. Genetics **139:** 1689-1700.

Kirchhamer, C. V., C. H. Yuh and E. H. Davidson, 1996 Modular Cis-Regulatory Organization of Developmentally Expressed Genes: 2 Genes Transcribed Territorially in the Sea-Urchin Embryo and Additional Examples. Proceedings of the National Academy of Sciences of the United States of America **93:** 9322-9328.

Lee, K. H., Q. H. Xu and R. E. Breitbart, 1996 A new tinman-related gene, nkx2.7, anticipates the expression of nkx2.5 and nkx2.3 in zebrafish heart and pharyngeal endoderm. Developmental Biology **180:** 722-731.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer, Massachusetts.

Li, W. H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Journal of Molecular Evolution **36:** 96-99.

Li, X. L., and M. Noll, 1994 Evolution of distinct developmental functions of 3 Drosophila genes by acquisition of different cis-regulatory regions. Nature **367:** 83-87.

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

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459-473.

Mehlhorn, K., and S. Naher, 1999 *LEDA: A platform for combinatorial and geometric computing.* Cambridge University Press, Cambridge, UK.

Mena, M., B. A. Ambrose, R. B. Meeley, S. P. Briggs, M. F. Yanofsky and R.J. Schmidt, 1996 Diversification of C-function activity in maize flower development. Science **274:** 1537-1540.

Mewes, H. W., K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker and D. Frishman, 1999 MIPS: a database for genomes and protein sequences. Nucleic Acids Research **27:** 44-48.

Rendel, J. M., 1979 Canalisation and Selection, pp. 139-156 in *Quantitative Genetic Variation*, edited by J. N. Thompson and J. M. Thoday. Academic Press.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L.G. Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M. Gelbart, R.A. George, L.S.B. Goldstein, F.C. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A.

Hoskins, J.Y. Li, Z.Y. Li, R.O. Hynes, S.J.M. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, O.F. PH, O.K. Pickeral, C. Shue, L.B. Vosshall, J. Zhang, Q. Zhao, X.Q.H. Zheng, F. Zhong, W.Y. Zhong, R. Gibbs, J.C. Venter, M.D. Adams, and S. Lewis. 2000. Comparative genomics of the eukaryotes. *Science* **287**: 2204-2215.

Rutherford, S. L., and S. Lindquist, 1998 Hsp90 Buffers Development Against Genetic-Variation and Could Link Capacity For Morphogenic Change With Environmental-Stress. Molecular Biology of the Cell **9:** 2511-2511.

Schmidt, R. J., B. Veit, M. A. Mandel, M. Mena, S. Hake and M.F. Yanofsky, 1993 Identification and Molecular Characterization of Zag1, the Maize Homolog of the Arabidopsis Floral Homeotic Gene Agamous. Plant Cell **5:** 729-737.

Schwikowski, B., P. Uetz and S. Fields, 2000 A network of protein-protein interactions in yeast. Nature Genetics **18:** 1257-1261.

Slusarski, D. C., C. K. Motzny and R. Holmgren, 1995 Mutations That Alter the Timing and Pattern of Cubitus Interruptus Gene-Expression in Drosophila-Melanogaster. Genetics **139:** 229-240.

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.

Spellman, P. T., G. Sherlock, B. Futcher, P. O. Brown and D. Botstein, 1998 Identification of cell-cycle regulated genes in yeast by DNA microarray hybridization. Molecular Biology of the Cell **9:** 2155-2155.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. QureshiEmili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M.J. Yang, M. Johnston, S. Fields, and J.M. Rothberg, 2000 A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature **403:** 623-627.

vonDassow, G., E. Meir, E. M. Munro and G. M. Odell, 2000 The segment polarity network is a robust development module. Nature **406:** 188-192.

Wagner, A., 2000a Mutational robustness in genetic networks of yeast. Nature Genetics **24:** 355-361.

Wagner, A., 2000b The role of pleiotropy, population size fluctuations, and fitness effects of mutations in the evolution of redundant gene functions. Genetics **154:** 1389-1401.

Wagner, A., 2001 The yeast protein interaction network evolves rapidly and contains few duplicate genes. Molecular Biology and Evolution. **18:** 1283-1292.

Figure Captions

Fig. 1.: Asymmetric divergence in protein-protein interactions. (a) Circles stand for proteins, lines for interactions among proteins. Shortly after a gene duplication, the products P and P^* of a duplicate gene that are part of a protein interaction network interact with the same proteins. Eventually, some or all of the common interactions may be lost, and new interactions may be gained by either protein.

Fig. 2.: Asymmetric divergence in protein-protein interactions: Two-hybrid data (a) The number of interaction partners of protein 1 versus protein 2, plotted for the two protein products of 2185 paralogous yeast gene pairs with $K_s<3$. Protein interaction data is taken from (Uetz *et al*. 2000). Values of zero on either axis indicate that one member of a paralogous pair is not part of the protein interaction network. It may have lost all protein interactions (but may have retained biological functions not mediated through protein interactions.) One protein pair with 0 and 24 interactions is outside the scale shown here but was included in the statistical analysis. **(b)** and **(c)** The expected distribution of interactions for the same 2185 paralogues if divergence after duplication had occurred through independent equiprobable interaction loss or gain, respectively, in either duplicate.

Fig. 3.: Asymmetric divergence in protein-protein interactions: Non two-hybrid data (a) The number of interaction partners of protein 1 versus protein 2, plotted for the two protein products of 1805 paralogous yeast gene pairs with $K_s<3$. Interaction data supported by experiments not using the two-hybrid assay was obtained from the MIPS data base (Mewes *et al*. 1999). Values of zero on either axis indicate that one member of a paralogous pair is not part of the protein interaction network. It may have lost all protein interactions (but may have retained biological functions not mediated through protein interactions.) **(b)** and **(c)** The expected distribution of interactions for the same 1805 paralogues if divergence after duplication had occurred through independent equiprobable interaction loss or gain, respectively, in either duplicate.

Fig. 4.: Asymmetric divergence and mutational robustness. **(a)** Schematic depiction of symmetric vs. asymmetric divergence of two duplicate genes with 20 (sub)functions, represented by white boxes. Black boxes indicate that a gene has suffered a mutational loss of the respective

function. In asymmetric divergence, this loss occurs preferentially in one gene. **(b)** The probability P_{Δ} that a loss of function mutation has a deleterious effect, that is, that it eliminates a function not "covered" by the other gene, as a function of x, the degree of asymmetry in divergence. P_{Δ} is smallest for maximal asymmetry in divergence, i.e., for $x=0$ and $x=1$.

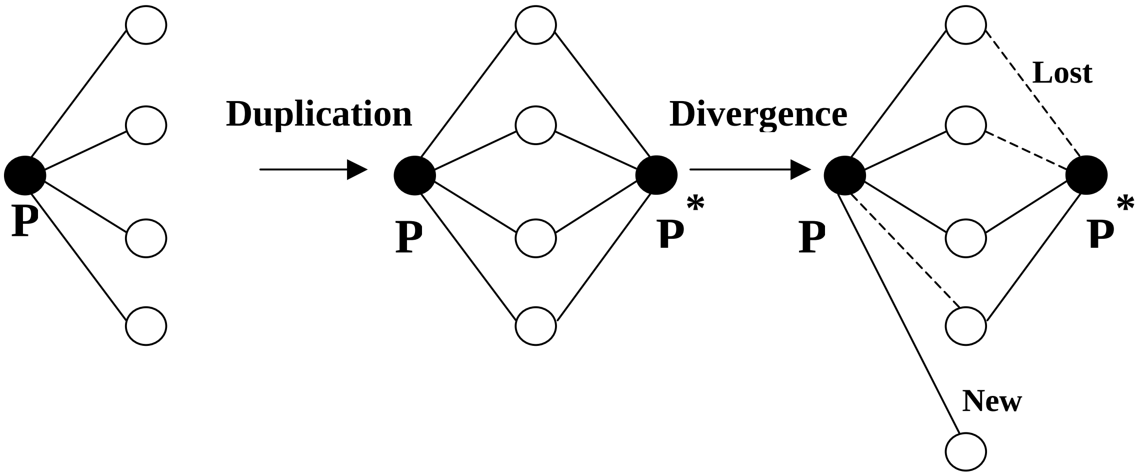

Fig. 1

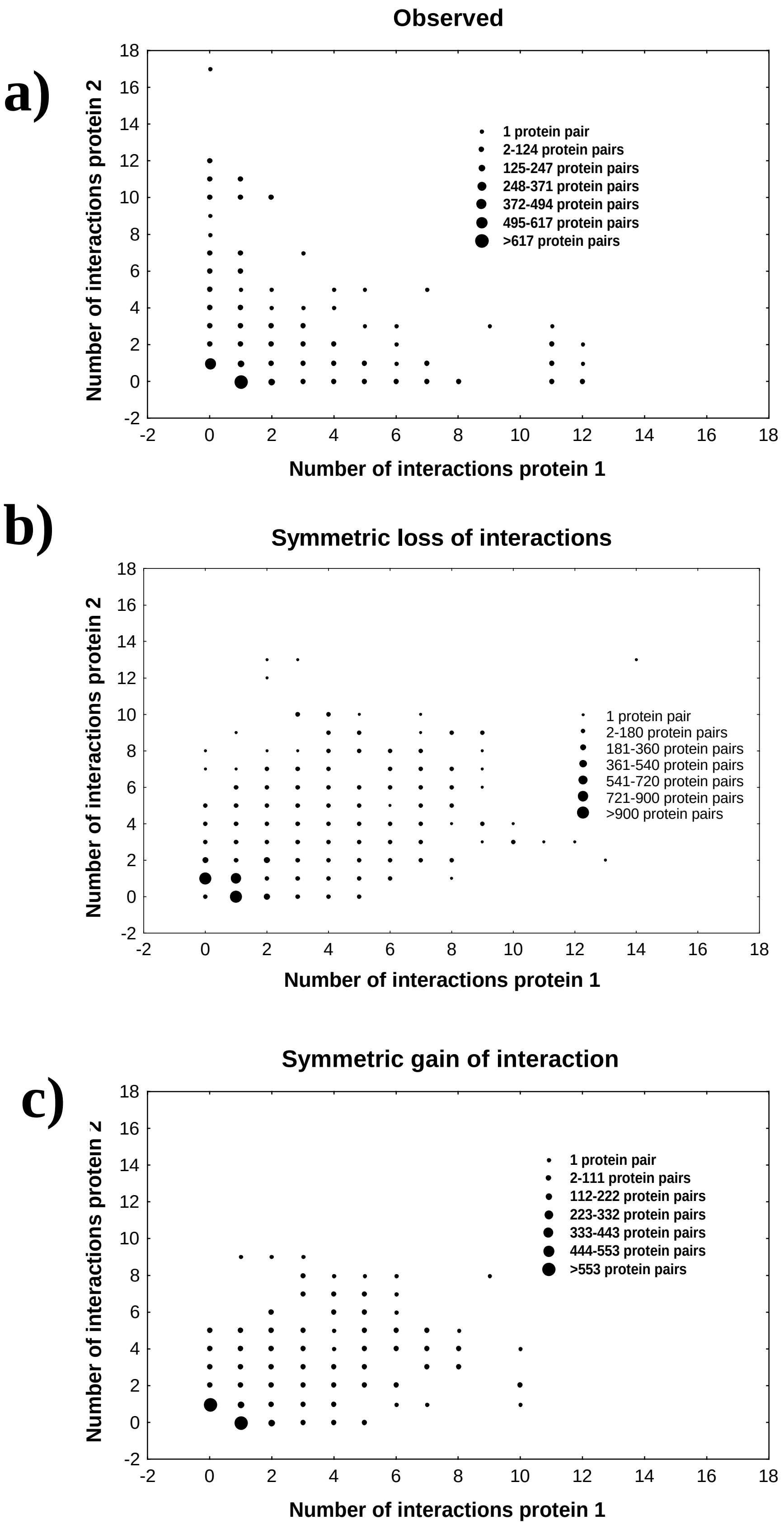
a)
Observed
Number of interactions protein 2
Number of interactions protein 1
-2
0
2
4
6
8
10
12
14
16
18
1 protein pair
2-124 protein pairs
125-247 protein pairs
248-371 protein pairs
372-494 protein pairs
495-617 protein pairs
>617 protein pairs
b)
Symmetric loss of interactions
Number of interactions protein 2
Number of interactions protein 1
1 protein pair
2-180 protein pairs
181-360 protein pairs
361-540 protein pairs
541-720 protein pairs
721-900 protein pairs
>900 protein pairs
c)
Symmetric gain of interaction
Number of interactions protein 2
Number of interactions protein 1
1 protein pair
2-111 protein pairs
112-222 protein pairs
223-332 protein pairs
333-443 protein pairs
444-553 protein pairs
>553 protein pairs

Fig. 2

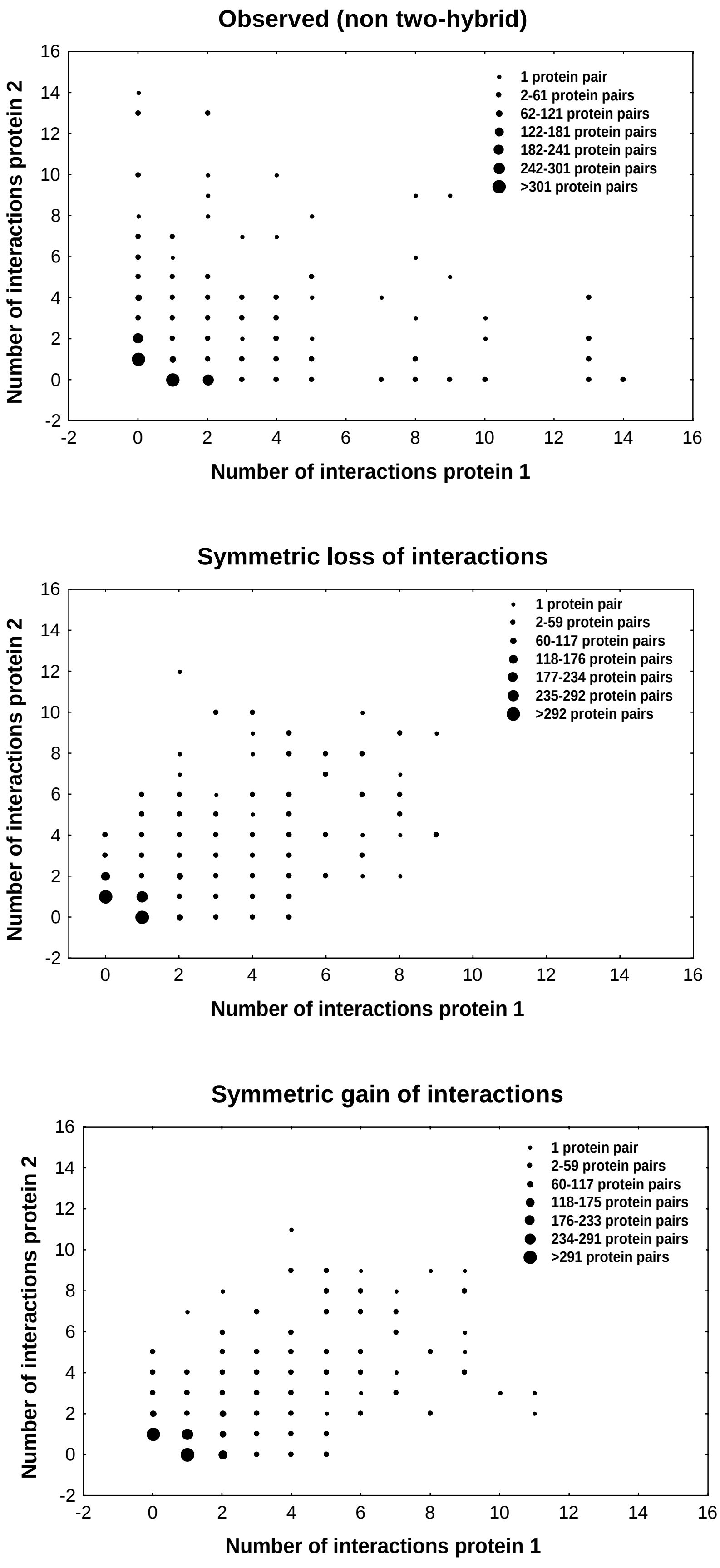
a)
Observed (non two-hybrid)
● 1 protein pair
● 2-61 protein pairs
● 62-121 protein pairs
● 122-181 protein pairs
● 182-241 protein pairs
● 242-301 protein pairs
● >301 protein pairs
Number of interactions protein 2
Number of interactions protein 1
-2 0 2 4 6 8 10 12 14 16
b)
Symmetric loss of interactions
● 1 protein pair
● 2-59 protein pairs
● 60-117 protein pairs
● 118-176 protein pairs
● 177-234 protein pairs
● 235-292 protein pairs
● >292 protein pairs
Number of interactions protein 2
Number of interactions protein 1
-2 0 2 4 6 8 10 12 14 16
c)
Symmetric gain of interactions
● 1 protein pair
● 2-59 protein pairs
● 60-117 protein pairs
● 118-175 protein pairs
● 176-233 protein pairs
● 234-291 protein pairs
● >291 protein pairs
Number of interactions protein 2
Number of interactions protein 1
-2 0 2 4 6 8 10 12 14 16

Fig. 3

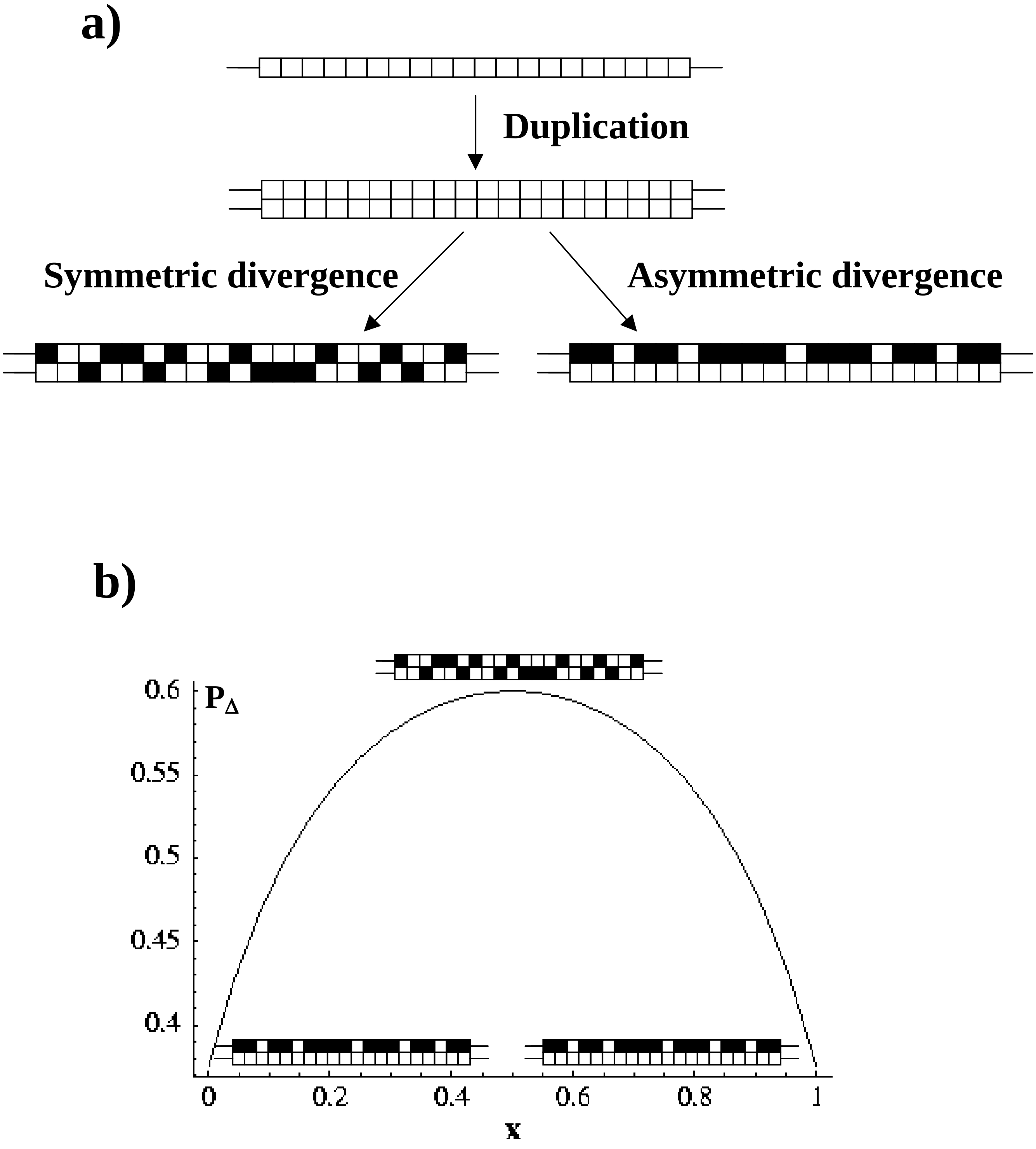
a)
Duplication
Symmetric divergence
Asymmetric divergence
b)
0.6
0.55
0.5
0.45
0.4
P_Δ
0
0.2
0.4
0.6
0.8
1
x

Fig. 4

Table 1: Differential effects of paralogous gene null-mutations on yeast gene expression. The column labeled ‘1/2’ indicates the number of genes whose expression is affected by a null-mutation in gene 1 and 2, respectively. The column labeled ‘common’ indicates the number of genes affected by both. **P-**values indicate the probability that a difference in the number of affected genes equal or greater than the observed difference is due to equiprobable loss or gain of effects in the duplicates. Rows in bold type have **P**<0.05.

		Number of effects		
Gene 1 / Gene 2	K_s	1 / 2	Common	Symmetric Divergence
YDL056W / YER111C	**1.29**	**5 / 149**	**1**	**7.69×10^{-39}**
YAL007C / YOR016C	1.58	6 / 1	0	0.13
YGR109C / YPR119W	1.83	17 /17	0	1
YBR245C /YOR304W	1.90	20 / 13	3	0.1
YER041W / YKL113C	**2.01**	**0 / 14**	**0**	**1.2×10^{-4}**
YHR022C / YOR089C	**2.08**	**12 / 3**	**0**	**0.035**
YLR014C / YMR280C	**2.17**	**14 / 0**	**0**	**1.2×10^{-4}**
YDL042C / YOR025W	**2.36**	**16 / 97**	**2**	**1.9×10^{-16}**
YEL049W / YOR009W	**2.36**	**1 / 10**	**0**	**0.011**
YER073W / YHR039C	2.64	6 / 4	0	0.75
YDR480W / YPL049C	**3.21**	**9 / 23**	**2**	**5.2×10^{-3}**