BMC Evolutionary Biology

Research article

Upstream plasticity and downstream robustness in evolution of molecular networks

Sergei Maslov*[1], Kim Sneppen[2], Kasper Astrup Eriksen[3,1] and Koon-Kiu Yan[1,4]

Address: [1]Department of Physics, Brookhaven National Laboratory, Upton, New York 11973, USA, [2]Nordita, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark, [3]Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden and [4]Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA

Email: Sergei Maslov* - maslov@bnl.gov; Kim Sneppen - sneppen@nbi.dk; Kasper Astrup Eriksen - kasper@thep.lu.se; Koon-Kiu Yan - kyan@grad.physics.sunysb.edu

* Corresponding author

Published: 08 March 2004

BMC Evolutionary Biology 2004, **4**:9

Received: 10 October 2003
Accepted: 08 March 2004

This article is available from: http://www.biomedcentral.com/1471-2148/4/9

Abstract

Background: Gene duplication followed by the functional divergence of the resulting pair of paralogous proteins is a major force shaping molecular networks in living organisms. Recent species-wide data for protein-protein interactions and transcriptional regulations allow us to assess the effect of gene duplication on robustness and plasticity of these molecular networks.

Results: We demonstrate that the transcriptional regulation of duplicated genes in baker's yeast *Saccharomyces cerevisiae* diverges fast so that on average they lose 3% of common transcription factors for every 1% divergence of their amino acid sequences. The set of protein-protein interaction partners of their protein products changes at a slower rate exhibiting a broad plateau for amino acid sequence similarity above 70%. The stability of functional roles of duplicated genes at such relatively low sequence similarity is further corroborated by their ability to substitute for each other in single gene knockout experiments in yeast and RNAi experiments in a nematode worm *Caenorhabditis elegans*. We also quantified the divergence rate of physical interaction neighborhoods of paralogous proteins in a bacterium *Helicobacter pylori* and a fly *Drosophila melanogaster*. However, in the absence of system-wide data on transcription factors' binding in these organisms we could not compare this rate to that of transcriptional regulation of duplicated genes.

Conclusions: For all molecular networks studied in this work we found that even the most distantly related paralogous proteins with amino acid sequence identities around 20% on average have more similar positions within a network than a randomly selected pair of proteins. For yeast we also found that the upstream regulation of genes evolves more rapidly than downstream functions of their protein products. This is in accordance with a view which puts regulatory changes as one of the main driving forces of the evolution. In this context a very important open question is to what extent our results obtained for homologous genes within a single species (paralogs) carries over to homologous proteins in different species (orthologs).

Background

Biological processes are rarely performed by single isolated molecules. Instead, they typically involve a coordinated activity of many molecules forming a neighborhood in biomolecular networks. Changes in these networks are thus coupled to the evolution of new functions and functional relationships in the organism. Gene duplication is an important source of raw material for the molecular evolution [1]. Immediately after a duplication event the pair of freshly duplicated genes is thought to be identical in both sequences and functional roles in the cell. However, with time their properties including their positions within molecular networks diverge. Here we quantify this divergence in the baker's yeast *Saccharomyces cerevisiae* using several recent system-wide data sets. To this end we measure: 1) The similarity of positions of duplicated genes in the transcription regulatory network [2] given by the number of transcription regulators that regulate both of them; 2) The similarity of the set of binding partners [3,4] of their protein products, and their ability to substitute for each other in knock-out experiments [5]. These measures reflect, correspondingly, the upstream and downstream properties of molecular networks around duplicated genes. We then repeat this analysis using species-wide data on protein interaction networks in a bacterium *Helicobacter pylori* [6] and a fruit fly *Drosophila melanogaster* [7], as well as a systematic RNAi gene inactivation assay [8] in a nematode worm *Caenorhabditis elegans*.

Results and discussion

Divergence of the upstream transcriptional regulation of duplicated genes in S. cerevisiae

The first measure of the divergence of duplicated genes compares sets of their transcriptional regulators. Such a set contains information about different conditions under which a given gene is expressed, and thus reflects the spectrum of its functional roles in the cell. To quantify the similarity of transcriptional regulation of a pair of genes we use "regulatory overlap" Ω_{reg} given by the number of transcription factors that bind to upstream regulatory regions of *both* these genes (see Fig. 1 for a general illustration). The information about gene duplications used in this study was extracted from the list of all pairs of paralogous (evolutionary related) proteins found in the yeast genome by the blastp program [9] with a conservative 10^{-10} E-value cutoff (see Methods for more details). The system-wide data for the transcription regulatory network in yeast was taken from the chip-on-chip experiment by Lee *et al*. [2] which investigated in-vivo binding patterns between 106 yeast transcription factors and upstream regulatory regions of all 6270 yeast genes. Fig. 2A shows the distribution of the regulatory overlap for different values of the percent identity (PID) of amino acid sequences of paralogous proteins. From this figure one can see that the regulatory overlap has a tendency to decrease as a function of PID. While multiple overlaps dominate the distribution for PID ≥ 80%, they gradually disappear at lower PIDs.

Fig 2B shows the average value of the regulatory overlap as a function of PID. The regulatory overlap in this plot is normalized by a proxy to the ancestral connectivity of a gene, estimated as the total number of distinct transcription factors that are involved in regulation of at least one of the pair of proteins (see Fig 1). The correlation between the normalized regulatory overlap Ω_{reg} and the PID is highly statistically significant: the Pearson correlation is 0.34 (P-value around 10^{-70} for 2275 data points). Even for the lowest value of PID = 20% the average Ω_{reg} significantly exceeds its value in non-paralogous proteins. One interesting feature of the graph in Fig. 2B is that even pairs of proteins whose amino acid sequences are 100% identical to each other on average have only about 30% overlap in their upstream regulation. Such low regulatory overlap of recently duplicated genes can be partially attributed to false positives and false negatives present in the dataset of Ref. [2] (see Methods for extended discussion.) It might also be sometimes caused by an incomplete duplication of the upstream regulatory region of a gene, or by a burst of very rapid evolution of the regulatory region immediately following the duplication event. The second feature of the Fig. 2B is a gradual decline of the average regulatory overlap over the whole range of sequence similarities. The data in Fig. 2B can be fitted with an exponential decay with a rate corresponding to an average 3% loss of common regulators of a paralogous pair for every 1% decrease in their amino acid sequence identity. Thus already at PID = 80% about half of the common regulations present at PID = 100% are lost. The decline in the regulatory overlap at lower PIDs clearly visible in Fig. 2A,2B is in accord with a recently published analysis [11] of similarity between microarray profiles of paralogs. In fact, due to a more direct information about transcriptional regulation contained in the chip-on-chip dataset of Ref. [2] compared to microarray experiments, our analysis extends the gradual decline to much lower PID than was detected in Ref. [11]. After we submitted this manuscript another group of authors [12] has reported a rapid decline in the number of shared regulatory motifs of duplicated genes. This study, carried out as a function of a much faster silent substitution rate K_s, nicely complements our own findings. Indeed, in their analysis Papp *et al*. [12] logarithmically binned the K_s into four broad bins: below 0.01, 0.01–0.1, 0.1–1, and above 1. Since the reliability of the measured silent substitution rate dramatically decreases at high values of K_s, the whole long-time behavior (i.e. that for PID < 75% which in yeast roughly corresponds to $K_s > 1$) of the regulatory overlap remained inaccessible to the analysis of Ref. [12].

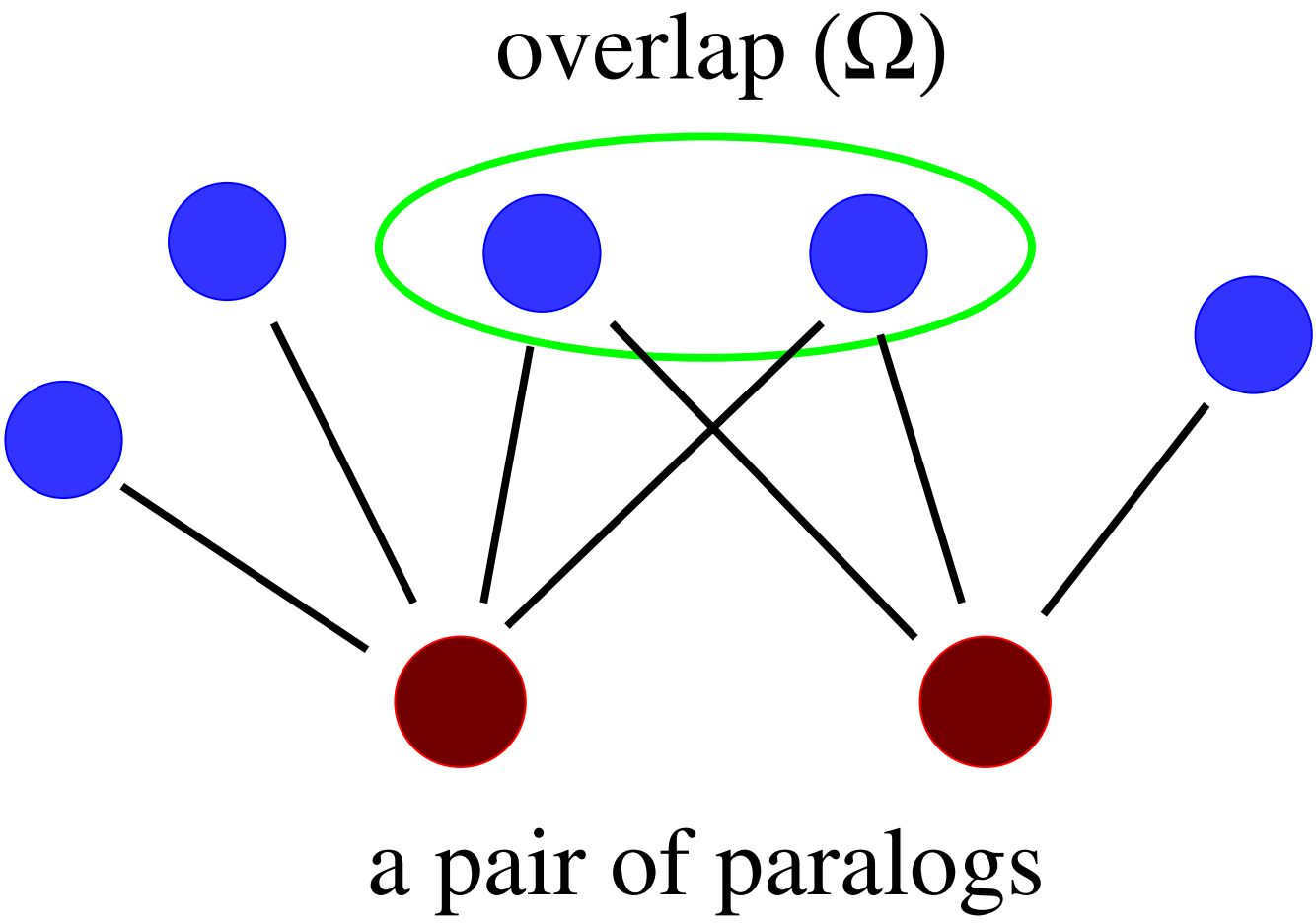

Figure 1
The illustration of a concept of overlap in a molecular network. For a pair of paralogs the overlap Ω is defined as the number of their common neighbors in the network. In the case of a transcription network the (upstream) regulatory overlap Ω_{reg} is given by the number of transcription factors regulating both paralogs, while for the physical interaction network the interaction overlap Ω_{int} counts their common binding partners. The pair of paralogs used in this illustration has the overlap Ω = 2 out of the total of 5 distinct neighbors of the pair. That corresponds to a normalized overlap of 2/5 = 0.40.

***Divergence in downstream functional roles of duplicated genes in* S. cerevisiae**

The rate of divergence between sets of *upstream* transcriptional regulators of paralogous proteins has an obvious *downstream* counterpart: it is the rate at which paralogous transcription factors loose their downstream targets. Unfortunately, an attempt to quantify this rate using the same dataset that we used above for the rate of upstream divergence would be limited to only 4 paralogous pairs formed by 106 transcriptional regulators studied in Ref. [2]. In general, relatively small number of paralogous transcription factors in any given species makes it difficult to go beyond just describing anecdotal cases in such an analysis. Thus in the remaining part of this study we concentrate on another measure of the downstream divergence, systematically comparing functional roles of duplicated (paralogous) proteins. The functional similarity of a pair of proteins is in part reflected in the "interaction overlap" Ω_{int} given by the number of other proteins that physically interact with both of them (See Fig. 1). In our study we use the system-wide information about protein-protein physical interactions obtained by combining two high-throughput two-hybrid experiments [3,4]. Fig 3A shows the average value of the interaction overlap Ω_{int} between pairs of paralogous proteins as a function of PID – their amino-acid similarity. Again Ω_{int} typically decreases with decreasing PID, reflecting the gradual loss/change of binding partners of proteins in the course of evolution. A similar analysis, but as a function of the silent substitution rate (K_s) was previously reported by

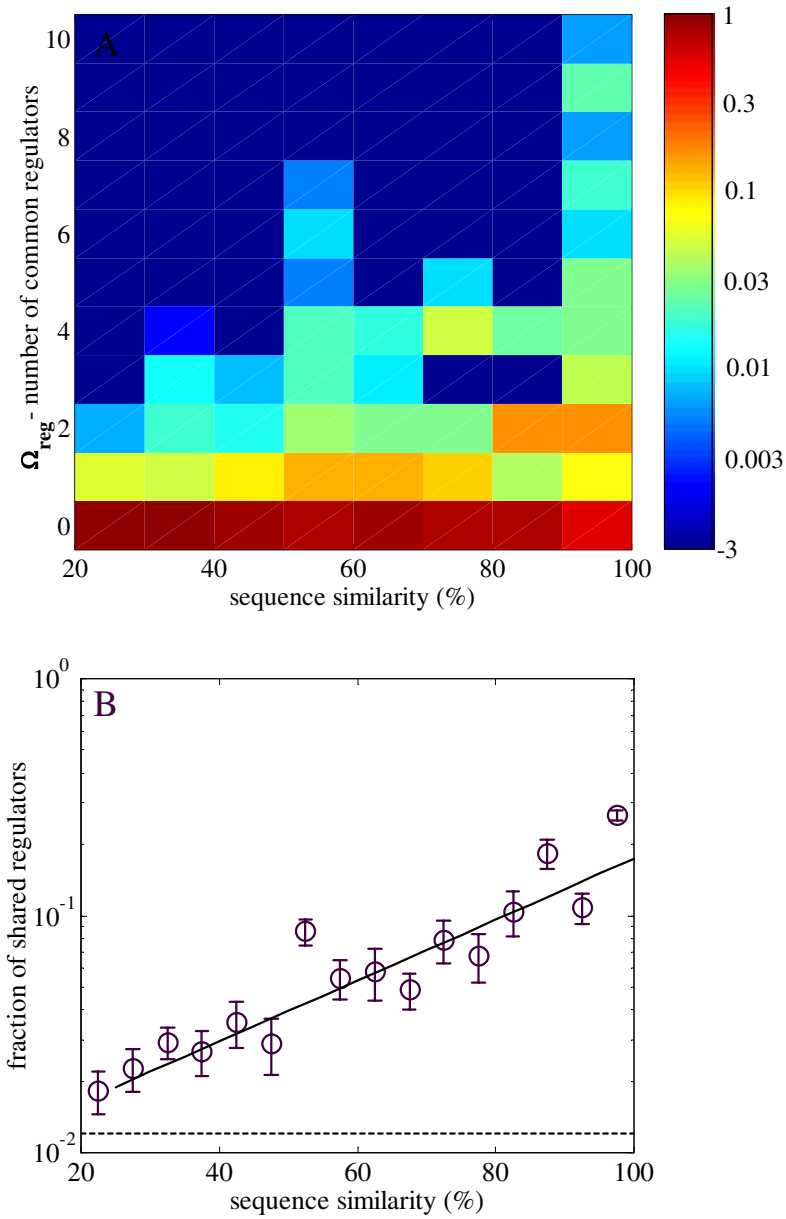

Figure 2
Divergence of the upstream transcriptional regulation of duplicated genes in yeast. A) The distribution of the regulatory overlap Ω_{reg} of paralo-gous proteins. The y-axis – Ω_{reg} – is the number of transcription factors that cis-regulate *both* genes encoding a given pair of paralogous proteins. The x-axis is the percent identity (PID) of amino acid sequences of these two proteins. The colorbar shows the likelihood of finding a given Ω_{reg} in a given 10% PID bin (note the logarithmic scale). The data describing the yeast regulatory network were taken from the whole genome chip-on-chip binding assay of 106 transcription factors [2], while the list of pairs of paralogous proteins was obtained by the whole genome blastp search (see Methods for more details.). B) The PID dependence of the average regulatory overlap Ω_{reg} normalized by the total number of regulators of either one or the other paralog. Relative error bars are estimated by the inverse square root of the total number of shared regulators in a given PID bin. The solid line is the best fit to the exponential form: $\Omega_{reg} \sim \exp(\gamma \text{ PID})$ with $\gamma = 0.03$ (3% change in Ω_{reg} for every 1% change in PID.) The dashed horizontal line at 0.015 is a null-model expectation of the normalized overlap of two randomly selected proteins (not necessarily paralogs).

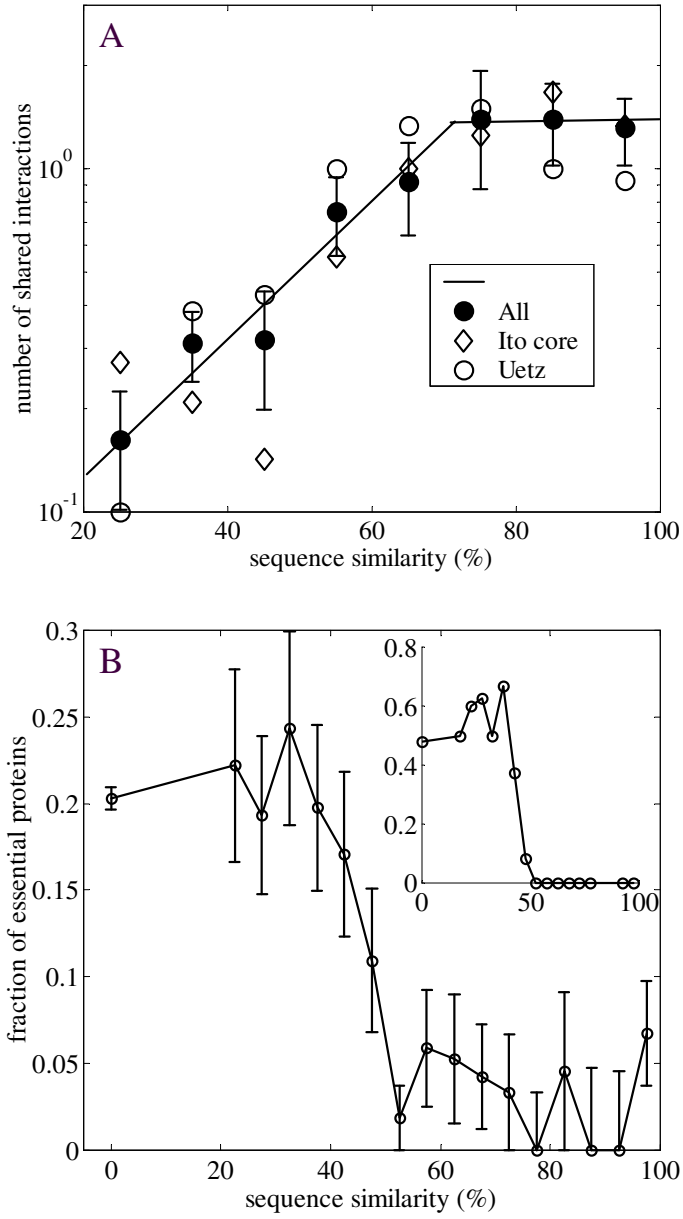

Figure 3
Divergence of downstream functions of duplicated genes in the baker's yeast *S. cerevisiae*. A. The average value of the interaction overlap Ω_{int} – the number of physical interaction partners shared by a pair of paralo-gous proteins – as a function of the similarity of their amino acid sequences. The physical interaction data are taken from the set of Uetz *et al.* [3] (open circles), the core dataset of Ito *et al.* [4] (diamonds), and the non-redundant combination of the two (filled circles). Note the apparent plateau for PID's between 70% and 100% in all three datasets. Solid lines are guides for the eye. A randomly selected (usually non-paralogous) pair of proteins in the combined dataset on average has Ω_{int} around 8×10^{-3} (off-limits in this figure). All data points at all PIDs are significantly above this null-model value. B. The fraction of essential (lethal null-mutant) proteins among all proteins tested in Ref. [5] as a function of PID to their most similar paralog in the yeast genome. Proteins with no paralogs (singletons) are binned at 0% PID. Note the apparent plateau between 50% and 100% PID. The inset (note the change of scale on the y-axis) shows the fraction of essential proteins in the subset of all proteins known to be localized in the yeast nucleus [17]. Here the effect becomes even more pronounced so that all 18 nuclear proteins protected by a paralog with at least 50% similarity were found to be non-essential.

Wagner [13]. In agreement with that study, we find that paralogous proteins are more likely to share interaction partners than one expects by pure chance alone (see the caption to the Fig. 3). Our set of yeast paralogs contains 189 paralogous pairs such that both paralogs physically interact with at least one other protein in the combined dataset of Refs. [3,4]. Out of these pairs 60 (30%) share at least one interaction partner. The correlation between the Ω_{int} and the PID in the combined two-hybrid dataset is highly statistically significant: the Pearson correlation is 0.36 (P-value around 5×10^{-6} for 189 data points). We also find that in yeast the divergence in the set of binding partners becomes systematic only for PID < 70%, while above 70% it remains roughly constant in both Uetz [3], Ito [4], and combined datasets (Fig. 3A).

An alternative way to quantify the extent of divergence/redundancy of duplicated genes is to examine phenotypes of of null-mutants lacking one of them. A systematic gene-deletion study in yeast [10] was recently used [14] to compare the fraction of essential genes (so that their null-mutants have lethal phenotype) between genes with and without paralogs in the genome. It was found that the fraction of essential genes is approximately 4 times higher among singleton genes than among ones protected by a highly similar paralog. It was also demonstrated that such protection by a paralog persists down to rather low levels of its amino-acid sequence similarity (PID) with the deleted protein. In Fig. 3B we confirm these findings using a more recent and larger systematic study [5] of viability of null-mutants in yeast as well as demonstrate that the magnitude of this protective effect is the strongest in the nucleus, where the largest fraction of essential proteins resides. Notice that the fraction of essential proteins (especially that of nuclear proteins) shows a dramatic increase as the PID to their closest paralog falls below 50%. Thus paralogous proteins with sequence similarity above 50% can typically substitute for each other.

Having presented different measures of upstream and downstream divergence of duplicated genes in yeast *S. cerevisiae* we are now in a position to discuss them in a wider context. Comparing Fig. 2B to Figs 3A,3B one concludes that changes in the upstream regulation of duplicated genes happen more readily than changes in their downstream function. The overlap in the set of binding partners (Fig. 3A) and the ability of duplicates to substitute for each other (Fig. 3B) remain virtually constant down to PID of 70%, at which point their average regulatory overlap has dropped to about 40% of its maximum (Fig. 2B). To summarize: our results indicate that duplicated genes would still have the ability to partially substitute for downstream functions of each other even at the time when the repertoire of their regulatory connections has already substantially changed from its ancestral state before the duplication. Such genes would be less constrained in evolving new functions [15], and thus would contribute to a greater evolutionary plasticity of the network.

Functional redundancy of paralogous proteins from RNAi experiments on C. elegans

One might expect the protective effect of paralogs to be unique to single-celled organisms such as yeast. Indeed, in multicellular organisms duplicated proteins are often expressed only in specific tissues and therefore unable to substitute for each other. However, using a systematic study of RNAi (RNA Interference) phenotypes in a nematode worm *C. elegans* [8] we found such protection [16] to be equally strong in this multicellular organism (See Fig. 4). As in Fig. 3B, the x-axis in Fig. 4 is PID – the similarity of amino acid sequences between a given protein and its closest related paralog (all singleton proteins without paralogs are clumped into the 0% PID bin). The y-axis is the fraction of tested proteins whose elimination by the RNAi technique was found [8] to give rise to a nonviable phenotype (embryonic or larval lethality or sterility). In worm the protection of having a paralog starts to gradually weaken for PID < 70%. In both worm and yeast there seems to be a four-fold drop in the fraction of essential proteins between PID = 0% and 100%.

In the inset to Fig. 4 we kept all successfully cloned genepairs, while in the main panel we dropped those genepairs whose product was predicted [8] to target mRNA product of more than one gene in the genome (see Methods for more details). It is instructive that the fraction of essential genes as a function of PID shown in the inset to Fig. 4 has a well pronounced minimum around PID = 70% and then subsequently starts to rise for higher values of PID. The tentative explanation for this behavior is that unlike single-gene deletion technique used in yeast, the RNAi technique is based on RNA complementarity and can eliminate several different mRNAs with similar sequences. Therefore, paralogous genes with nearly identical DNA sequences prove to be useless from the point of view of protection against RNAi since their mRNA products would be eliminated at nearly the same rate as the intended targets. This neatly explains why in the inset to Fig. 4 the fraction of nonviable phenotypes for genes with a 100% identical paralog in the genome approaches that of unprotected genes without paralogous partners (keep in mind that in this plot we use amino acid sequence identity of proteins and not of their mRNA precursors.) This observation also reinforces the point of view that the decline in the fraction of essential genes vs PID shown in Figs 3B,4 is indeed caused by protective effects of paralogs and cannot be explained by a possible tendency of nonessential genes to duplicate more frequently.

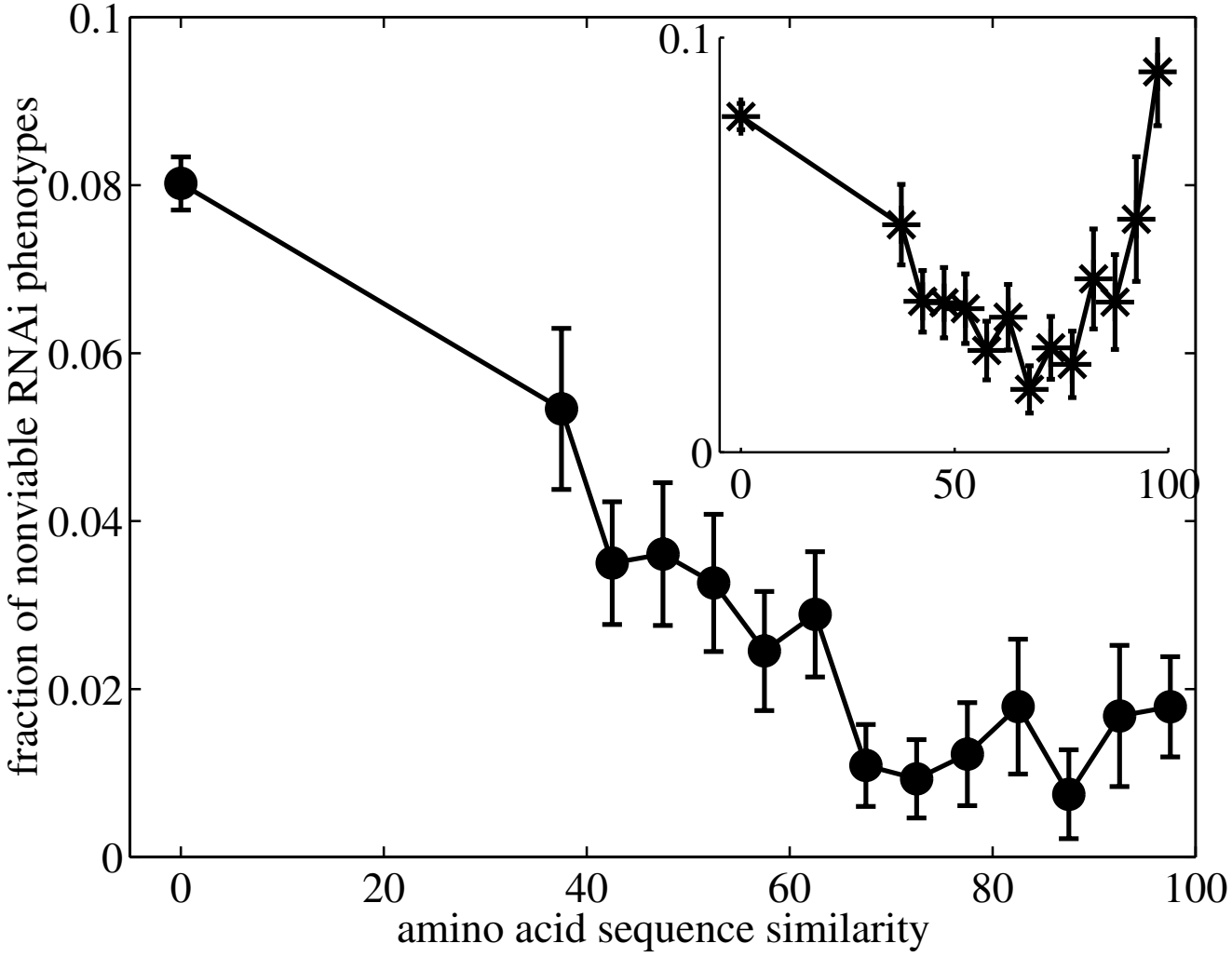

Figure 4
Protective effect of paralogs in a nematode worm *C. elegans*. The fraction of essential (non-viable RNAi phenotype [8]) proteins among all tested worm proteins as a function of PID to their most similar paralog in the worm genome. Note the apparent plateau between 70% and 100% PID. The plot in the inset shows the fraction of essential proteins among all RNAis tested in Ref. [8], while that in the main panel drops RNAis that are predicted [8] to target mRNA products of more than one gene. Note that while the graph in the main panel is qualitatively similar to that in Fig. 3B, in the inset the fraction of essential proteins at PID = 100% rises to its level for singleton proteins. Thus when mRNAs of highly similar paralogs are eliminated along with the targeted mRNA, the protective effect of paralogs totally disappears.

Divergence of physical interactions of paralogous genes in H. pylori *and* D. melanogaster

The analysis of evolution of molecular networks advocated in this paper requires a *large* (preferably genome-wide) and *unbiased* (i.e. no anthropogenic selection present in databases) dataset describing a molecular network in a given species. Apart from yeast, which is arguably the best studied model organism, system-wide two-hybrid physical interaction assays were published for a simple bacterium *Helicobacter pylori* [6], and a fly *Drosophila melanogaster* [7]. In Fig. 5 we used these two datasets to quantify the decay of the average interaction overlap as a function of amino-acid sequence similarly (see Fig. 3A for the same analysis in yeast.) The correlation between Ω_{int} and PID is highly statistically significant in both cases: the Pearson correlation of 0.43 (P-value around 3×10^{-4} for 65 data points) for *H. pylori*, and 0.19 (P-value around 10^{-26} for 2843 data points) in *D.*

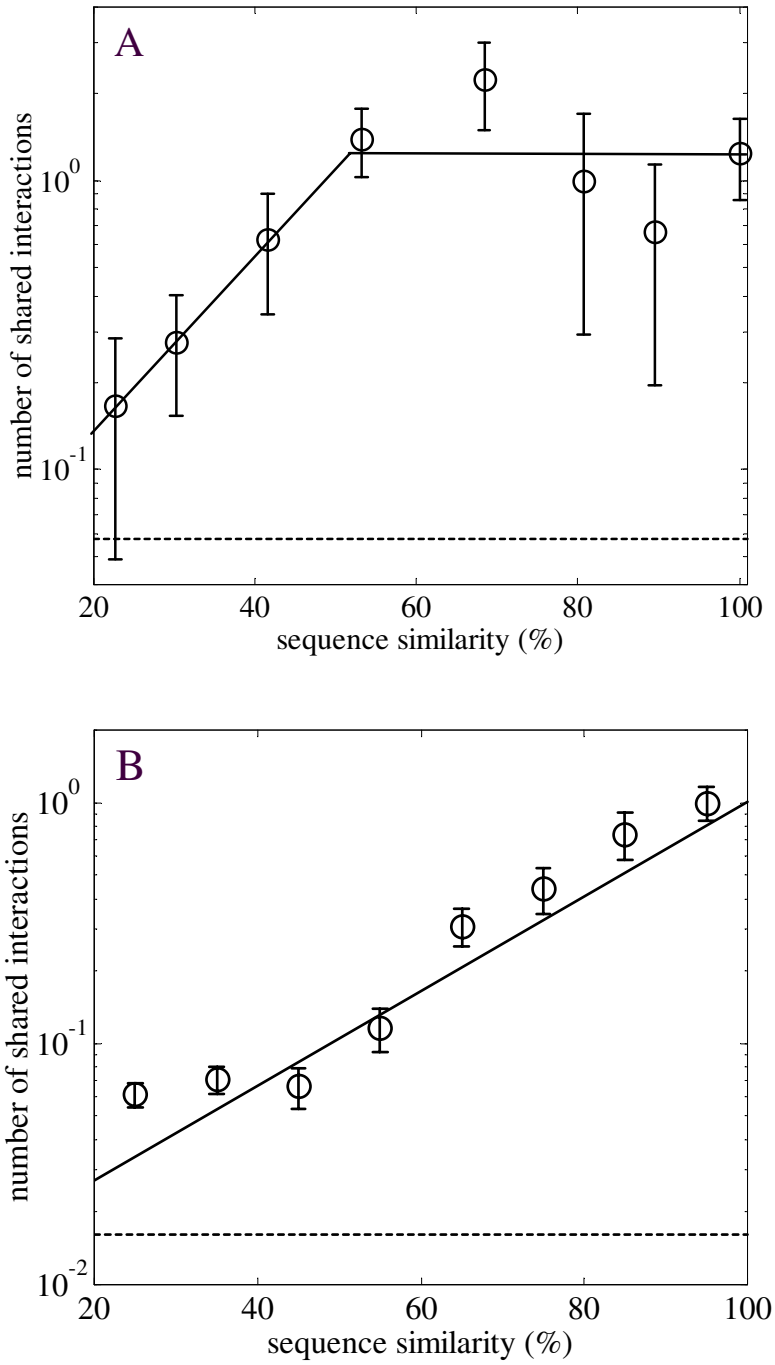

Figure 5
Divergence of physical interaction neighborhoods of duplicated genes in a bacterium *H. pylori* and a fly *D. melanogaster*. The average value of the interaction overlap Ω_{int} of paralogous proteins in H. pylori (A) and D. melanogaster (B) as a function of the amino acid sequence similarity. The physical interaction data are taken from Ref. [6] for *H. pylori* (A) and from Ref. [7] for *D. melanogaster*. Note the apparent plateau for PID's between 50% and 100% in panel A and its absence in panel B. Dashed horizontal lines show the average interaction overlap of a random (usually non-paralogous) pair of proteins. The solid line in B is the best fit to the exponential form: $\Omega_{int} \sim \exp(\gamma_{FLY} PID)$ with $\gamma_{fly} = 0.045$.

melanogaster. Our basic conclusions agree for all quite diverse organisms used in this study: paralogous proteins are much more likely to share binding partners than expected by pure chance alone. Furthermore, the number of common interaction partners goes down as PID of their amino acid sequences decreases. In the yeast and *H. pylori* we see the evidence of an initial plateau at which the average overlap appears to be independent of PID. On the other hand in the fly there is no evidence of such plateau, which makes the average rate of loss of common binding partners (about 4.5% for every 1% of change in PID) quite high in this organism. However, in the absence of system-wide data on transcription factors' binding in the fly and *H. pylori* we could not quantify rates of upstream changes in these two organisms, and consequently cannot compare them to the corresponding downstream rates.

Conclusions

The evolution of a biological organism modifies it on multiple levels ranging from sequences of individual molecules, to their coordinated activity in the cell (molecular networks), all the way up to the phenotype of the organism itself. While its manifestations both on the level of protein sequences and phenotypes are reasonably well documented, the data needed to quantify evolutionary changes taking place on the level of molecular networks have appeared only very recently. Systematic experiments such as high-throughput two hybrid assays of protein-protein interactions [3,4,6,7], chip-on-chip studies of whole-genome binding of a large number of transcription factors [2], and whole-genome assays of inactivations of single genes [5] or proteins [8] allowed us to go beyond describing particular cases of evolution of molecular networks and look at its large scale dynamics.

For all molecular networks studied in this work we found that even the most distantly related paralogous proteins with amino acid sequence identities around 20% on average have more similar positions within a network than a randomly selected pair of proteins. That means that some pairs of paralogous proteins at least partially retain their functional redundancy for extremely long time after the duplication event.

Our results also indicate that the genetic regulation of paralogous proteins changes faster than both their amino acid sequences and the set of their protein interactions partners. It is tempting to extend this observation to pairs of homologous proteins in different species (orthologs) that diverged from each other as a result of a speciation (as opposed to a gene duplication) event. This would help to explain how species with very similar gene contents can evolve novel properties on a relatively short timescale. However, such an inter-species comparison of molecular networks has to wait for the appearance of whole-genome data on molecular networks in closely related model organisms.

Methods

As a source of information about yeast duplicated genes we use the dataset consisting of 3909 pairs of paralogous yeast proteins. This set was obtained by blasting all yeast proteins against each other with a conservative E-value cutoff of 10^{-10} and leaving only pairs in which the aligned region constituted at least 80% of the length of a longer protein. This prevented the appearance of pairs of multidomain proteins paralogous over only one of their domains. We further curated this dataset by removing 72 known [17] transposable elements and all their paralogs (108 proteins all together). That left us with 2299 paralogous pairs formed by 1596 yeast proteins (about 25% of the genome). These pairs are characterized by a broad and relatively uniform distribution of the percent identity (PID) of amino acid sequences ranging from 20% to 100% (See Fig. 6) The histogram in the Fig. 6 is binned at 5% PID (as the data used to plot the Fig. 2B), and one can see that even in the least represented bins there are over 40 paralogous pairs providing sufficient statistics for our analysis. Our set of all possible pairs of paralogous proteins contains some redundant information especially for large protein families. Indeed, a family of, say, 4 proteins would contribute $(4 \cdot 3)/2 = 6$ paralogous pairs to our analysis, while it contains at most 3 true duplicated pairs. However, in the situation where the data describing molecular networks are incomplete and noisy such redundancy is rather beneficial by providing better statistics. We have verified that apart from somewhat larger errorbars all our quantitative findings remained virtually unchanged when we repeated our analysis of upstream regulations in yeast using only 938 pairs of putative duplicated proteins. These pairs were obtained from the full set of 2299 paralogous pairs by the detailed phylogenetic analysis of individual families. It is also worthwhile to note that while the average number K_s of silent substitutions per substitution site in a pair of duplicated genes is commonly used as a proxy of the time elapsed since the duplication event [1], the PID (or K_a – the number of non-silent substitutions per site – related to PID via $\text{PID} = 100 \exp(-2K_a)$) is rather a crude estimate of the extent of their functional similarity. Hence, our analysis emphasizes function-dependent rather than time-dependent divergence between paralogous proteins.

The system-wide data describing the transcription regulatory network of yeast was taken from the Ref. [2], which reports the so-called "chip-on-chip" study of in-vivo binding of 106 transcription factors to upstream regulatory regions of genes encoding all 6270 of yeast proteins. Since the number of transcriptional regulators in this dataset is quite large, the probability that by pure chance the same

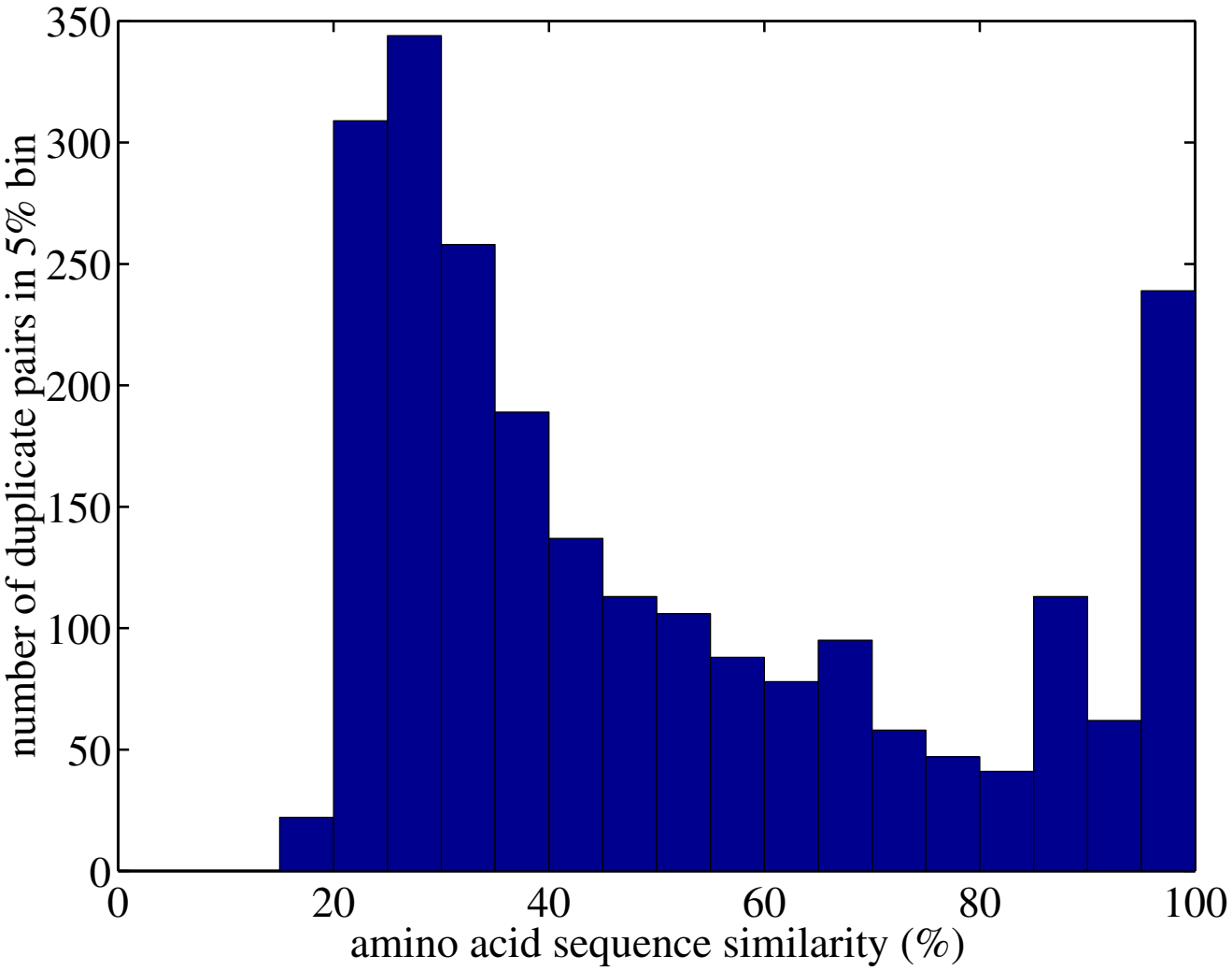

Figure 6
The histogram of amino acid sequence identities (PID) of 2299 pairs of paralogous yeast proteins used in our study.

transcription factor would be incorrectly detected among upstream regulators of *both* duplicated genes is small (of order of 1%). Thus the contribution of false positives of the dataset of Ref. [2] to the regulatory overlap Ω_{reg} is quite insignificant. This allowed us to use a P-value cutoff equal to 10^{-2} (12854 regulations) less conservative than the 10^{-3} cutoff (4418 regulations) of Lee *et al*. [2]. On the other hand, false positives (if present in the data) could significantly affect the average number of regulatory inputs of individual proteins used to normalize the regulatory overlap in Fig. 2B. However, we found that both the initial drop and the rate of exponential decay of the *normalized* regulatory remains virtually unchanged when Fig. 2B is repeated for different values of the P-value cutoff ranging from 10^{-2} to 10^{-4} (data not shown). In the same range of P-values the average number of regulations per gene changes six-fold (from 2 to 0.33)! This suggests that false positives are not a significant part of the experimental dataset of Ref. [2] at least up to 10^{-2}, and validates the robust nature of parameters extracted from the Fig. 2B. In the analysis shown in Fig. 2 we have dropped 3 paralogous pairs sharing the same intergenic sequence since by design of the chip-on-chip experiment [2] such pairs would have 100% regulatory overlap. We also checked that Fig. 2A does not change significantly if one limits the analysis to genes without diverging promoters ensuring that a given intergenic could possibly regulate only one gene.

As a source of information about binding partners of yeast proteins we combined the data from two independent high-throughput two-hybrid experiments: the core dataset of Ito *et al*. [4] (806 interactions among 797 proteins) and the extended Uetz *et al*. dataset [3], downloaded from the website of this group (1446 interactions among 1340 proteins). The resulting network consists of 1734 proteins joined by 2111 non-redundant interactions. Using this combined dataset we found that even 100% identical proteins share on average only 30% of their binding partners. However, unlike for upstream regulation, the set of interaction partners of a protein is fully determined by its amino acid sequence. Therefore, an imperfect overlap in the set of binding partners of identical proteins has to be attributed to false positives/negatives inevitably present in high-throughput two-hybrid experiments. The relatively high rate of false negatives in genome-wide two-hybrid experiments is further corroborated by the fact that datasets used in our study coming from two independent experiments [3,4] have only 141 interactions in common. The abundance of missing interactions makes the normalization of the interaction overlap impractical. That was the reason why unlike in Fig. 2B in Fig. 3A we used the raw (unnormalized) interaction overlap. To make sure that differences between Figs. 2B and 3A are not caused by differences in normalization we repeated them using various normalization schemes as well as altogether unnormalized (data not shown). We found that apart from the overall scale of the y-axis, changes in normalization do not affect exponential decay parameters of Figs 2B,3A.

The system-wide data on viability of *S. cerevisiae* null-mutants used in our study was obtained from Ref. [5] in which 1103 essential (non-viable null-mutants) and 4678 non-essential (viable null-mutants) yeast proteins were reported. The lists of viable and non-viable null-mutants as discovered in Ref. [5] were downloaded from the Saccharomyces Genome Database [17].

Our analysis of protective effects of paralogs in *C. elegans* is based on the set of 15587 viable and 1170 non-viable (embryonic or larval lethality or sterility) RNAi phenotypes reported in [8]. The information about worm paralogs is obtained from the EuGenes database [18] and consists of 30036 paralogous pairs involving 10071 worm proteins (blastp with 10^{-30} cutoff and no requirements on the length of aligned region). In Fig. 4 we used 13884 RNAi phenotypes for which we were able to uniquely map the genepair name to the worm protein name used in EuGenes.

The two-hybrid assay of protein-protein inetractions in *H. pylori* [6] used in Fig. 5A contains 1465 interactions between 732 proteins, while there are only 260 paralogous pairs involving 140 proteins. As in yeast this set was obtained by blasting all protein sequences found in the fully sequenced genome against each other with a conservative E-value cutoff of 10^{-10} and leaving only pairs in which the aligned region constituted at least 80% of the length of a longer protein.

Finally, our analysis of the interaction overlap between paralogous proteins in *D. melanogaster* is based on the full dataset of the high-throughput two-hybrid experiment [7]. It consists of 20671 protein-protein physical interactions involving 7002 of fly proteins obtained in. To generate Fig. 5B we also used the set of 16713 paralogous pairs involving 2827 fly proteins.

Authors Contributions

SM, KS, and KAE contributed to both the ideas and writing of the manuscript in close collaboration. Koon-Kiu Yan has generated blastp datasets for *H. pylori*, yeast, and fly, as well as performed the analysis of RNAi experiment shown in Fig. 4. All authors read and approved the manuscript.

Acknowledgments

Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy. Two of us (K.E and K.S.) thank the Institute for Strongly Correlated and Complex Systems at Brookhaven National Laboratory for hospitality and financial support during visits when part of this work was completed. S.M. and K.S. acknowledge the support of the NSF grant PHY99-07949 (work at the KITP, University of California at Santa Barbara). We thank John Little for critically reviewing the manuscript.

References

1. Ohno S: *Evolution by gene duplication Berlin: Springer-Werlag*; 1970.
2. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcription Regulatory Networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
3. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403:**623-627.
4. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori Ma, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
5. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B *et al.*: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418:**387-391.
6. Rain JC, Selig L, Reuse HD, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V *et al.*: **The protein-protein interaction map of Helicobacter pylori.** *Nature* 2001, **409:**211-215.
7. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al.*: **A Protein Interaction Map of Drosophila melanogaster.** *Science* 2003, **302:**1727-1736.
8. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapink A, Le Bot N, Moreno S *et al.*: **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.** *Nature* 2003, **421:**231-237.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
10. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H *et al.*: **Systematic screen for human disease genes in yeast.** *Nature Genetics* 2002, **31:**400-404.

11. Gu Z, Nicolae D, Lu HH-S, Li W: **Rapid divergence in expression between duplicate genes inferred from microarray data.** *Trends in Genetics* 2002, **18:**609-613.
12. Papp B, Pál C, Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast.** *Trends in Genetics* 2003, **19:**417-422.
13. Wagner A: **The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.** *Mol Biol Evol* 2001, **18:**1283-1292.
14. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate genes in genetic robustness against null mutations.** *Nature* 2003, **421:**63-66.
15. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biology* 2002, **3(2):**RESEARCH0008.10008.9.
16. After our study was completed we learned of a similar analysis submitted for publication, Conant GC, Wagner A: **Duplicate genes and robustness to transient gene knockouts in Caenorhabditis elegans.** *Proc R Soc Lond B* in press.
17. **Saccharomyces Genome Database,** [http://genome-www.stanford.edu/Saccharomyces]
18. Gilbert DG: **euGenes: a eucaryote genome information system.** *Nucleic Acids Res* 2002, **30:**145-148.