Supplementary Materials for

The Origins and Impact of Constraints in Evolution of Gene Families

Boris E. Shakhnovich and Eugene V. Koonin

Construction of the Diffusion and Divergence Graph (DDG) for model organisms

Yeast

The reference sequences for the genome and the annotated proteins of *Saccharomyces cerevisiae* were extracted from GenBank (NIH, Bethesda), and each open reading frame (ORF) was assigned a node on the graph. Amino acid sequences of these ORFs were compared to each other using all-against-all BLAST (Altschul, Madden et al. 1997). The results of this comparison, i.e., scores and amino acid sequence identities represent edges between the nodes. To calculate the weight of the edges we use amino-acid alignment strings. First, we translate the amino-acid to nucleotide alignments. We then use PAML(Yang 1997; Yang and Nielsen 2000) to calculate the Ks (synonymous substitutions per synonymous site) and Ka (nonsynonymous substitutions) values between the ORFs in the alignments. This procedure produces three weighted graph representations of the yeast genome where the nodes are the genes and the edges are weighed by BLAST scores, amino-acid sequence identities and Ka,Ks values.

Paralogous families were identified by finding all strongly connected components as described in detail in the manuscript and elsewhere(Cormen 2001). This procedure requires that we translate the weighted graph into an unweighted one using a cutoff (Table 1). After finding all strongly connected components, we used essentiality data to divide the families into two classes: *E* and *N-families*. If a family included at least one essential gene, it was annotated as an *E-family*, else the family was designated is an *N-family*. Since we are interested in dynamics of paralogs, we only considered families that consisted of two or more genes. For yeast, we used the essentiality data obtained from high-throughput knockout experiments described in (Giaever, Chu et al. 2002).

The number of genes and families shows a near-linear dependency on the cutoff (Table S1). Most results described in the paper and in this Supplementary Material use the E-value cutoff of 1e-15 and Ks cutoff of 5. However, all trends described in the paper, e.g., slower divergence rate of non-essential genes in *E* families and larger average separation between paralogs in *E* families, were found to be independent of the cutoff (Table S1).

Table S1. *Dependence of the number of genes in E and N-families with respect to cutoff Ks used in building the DDG for S. cerevisiae. Average divergence in nonsynonymous mutations Ka was calculated over all pairs of vertices in the graph. All differences in <Ka> are significant above cutoff S<1.*

Ks Cutoff	Num Genes in E families	Num Genes in N Families	<Ka> in E-families	<Ka> in N-families
5	275	658	0.509344	0.147632
4.7	271	660	0.510581	0.147968
4.4	270	659	0.509814	0.147657
4.1	266	655	0.508682	0.14724
3.8	264	642	0.507339	0.144439
3.5	258	629	0.507252	0.142173
3.2	246	625	0.500473	0.140768
2.9	235	609	0.498432	0.136404
2.6	204	585	0.490572	0.128384
2.3	159	568	0.445489	0.124553
2	114	491	0.404539	0.104033
1.7	69	434	0.307063	0.091604
1.4	34	369	0.20307	0.070387
1.1	29	322	0.128399	0.054944

Escherichia coli K12 and C. elegans:

The procedure for building the DDG graph for *E. coli* and *C. elegans* was the same as described above for yeast. Genomes and open reading frame annotations were from GenBank. Essentiality data for *Escherichia coli K12* was obtained from (genbank; Gerdes, Scholle et al. 2003). We consider a gene essential if both PEC and Gerdes et al. assigned essentiality. However, using either one of these data sets alone dis not qualitatively change the results. For determining essential genes in *C. elegans*, we used the RNAi knockdown experiments described in (Fraser, Kamath et al. 2000; Kamath, Fraser et al. 2003; Simmer, Moorman et al. 2003). We considered all genes whose RNAi knockdown imparts sterility, lethality or other major phenotypic deficiencies as lethal. As in other cases, the exact definition of lethality did not qualitatively affect the results (data not shown).

Estimating the Speed of Divergence using SFP data and sequence comparison of orthologs in Yeast using Ka/Ks

In the paper, we describe evidence that essential genes evolve slower using SFP data from (Winzeler, Castillo-Davis et al. 2003). That data set provides us with the most direct evaluation of the strength of purifying selection as it calculates mutations on a very short time scale by comparing genomes of different strains of *S. cerevisiae*. While the SFP dataset does not separate the mutations into synonymous and non-synonymous ones, we assume that approximately 25% of the observed substitutions are silent (Lynch and Conery 2003). Using that assumption, we found that the SFP density in essential genes θ_{ess}=.01567 was less than in non-essential genes θ_{ne}=.02158. (*Table 1*). We then performed the same comparison for genes that have not been annotated as essential (Winzeler, Shoemaker et al. 1999; Giaever, Chu et al. 2002), but are members of *E* families. We find that for those genes SFP density is .012. Finally, we calculated the SFP density for genes in *N* families and found a value of .027. From these results, we conclude that, although approximately 2/3 of the genes in *E* families were not essential according to the knockout data, they show evidence for similar strength of purifying selection as essential genes.

Calculating strength of selection using Ka/Ks for S. cerevisiae-S. paradoxus orthologs

We further tested whether SFP density calculations show the same qualitative trends as the more common calculations of purifying selection using Ka/Ks ratios. First, we performed an all-against-all genome comparison between *S. cerevisiae* ORFs and *S. paradoxus* ORFs. The sequence of *S. paradoxus* genome from (Kellis, Patterson et al. 2003). We identified 4706 pairs of orthologs if we required alignments over 80% of the sequence length and BLAST E-value < 1e-15. After finding orthologs, we used PAML (Yang 1997; Yang and Nielsen 2000) to calculate Ka and Ks values between orthologs using the amino-acid alignments as the guide for nucleotide alignment (*Table S2*).

We found that the results were qualitatively similar to ones calculated with SFP data and reported in *Table 1*. We first compared the Ka/Ks means between all essential and non-essential genes. We found, as reported previously (Hurst and Smith 1999; Hirsh and Fraser 2001; Jordan, Rogozin et al. 2002; Kondrashov, Rogozin et al. 2002; Wall, Hirsh et al. 2005), that essential genes evolve slower than nonessential ones. This difference was more pronounced in genes that have paralogs. In fact, essential genes in *E* families had a mean Ka/Ks ratio of .064, i.e., almost twofold lower than the value for non-essential genes and also significantly lower than the value for essential genes without paralogs. Similar results have been reported previously (Yang, Gu et al. 2003)

Table S2. *Calculation of Ka/Ks for S. cerevisiae and S. paradoxus orthologs. All genes identified as essential are compared to the rest of the ORFs in the genome that had calculable Ka and Ks values. Ka/Ks for all non-essential genes that are in E families are compared to genes in N families.*

	Essential genes	*NonEssential genes*	P-Val
All Genes	0.10	0.13	2e-8
	E- families	*N- families*	
Only Non-Essential Genes in Families	0.08	0.12	4e-11

Calculation of the strength of purifying selection using Ka/Ks for Escherichia coli K12 and Escherichia coli CFT073 orthologs

The genome sequences and annotated ORFs for *Escherichia coli K12* and *Escherichia coli CFT073* were extracted from GenBank, and an all-against-all ORF sequence comparison was performed as described above. We identified 4996 pairs of orthologs when we required alignments over 80% of the sequence length and BLAST E-value < 1e-15. Very similar results were from *K12* comparison with *O157H7*(data not shown). After finding orthologs, we used PAML(Yang 1997; Yang and Nielsen 2000) to calculate Ka and Ks values between the orthologs using the amino-acid alignments as guide for nucleotide alignments (*Table S3*). As in yeast, we founnd that essential genes with duplicates evolved slower (Ka/Ks = .04) than all annotated non-essential genes (Ka/Ks = .099) and essential genes without duplicates (Ka/Ks = .054). However, the difference between two groups of essential genes was not statistically significant due to the small number of essential genes with duplicates (P-Val=.19). Importantly, as in yeast, we confirmed that, in *Escherichia coli K12,* non-essential genes in *E* families evolve slower (Ka/Ks=.056) than other non-essential genes with paralogs. *(Table S3)*

Table S3. *Calculation of Ka/Ks for Escherichia coli K12 and Escherichia coli CFT073 orthologs. In the first row all genes identified as essential are compared to the rest of the ORFs in the genome that had calculable Ka and Ks values. In the second row Ka/Ks is compared between non-essential genes in E-families and non-essential genes in N-families.*

	Essential genes	*Non-essential genes*	P-Val
Genes	0.054	0.099	1.6e-6
	E-families	*N-families*	
Only Non-Essential Genes in Families	0.056	0.1	1e-8

Controls for CAI and Abundance:

Previous research has suggested that abundance, codon adaptation index (CAI), and expression level also correlate with evolutionary rate (A. Drummond personal Communication, (Pal, Papp et al. 2003; Drummond, Bloom et al. 2005; Drummond, Raval et al. 2005). A nagging problem has been to evaluate the relative contributions of essentiality and each of the other variables (CAI, expression, and abundance) to the selective pressure experienced by a gene. In a recent study, Wall and coworkers imply that there is no way to assess the relative importance of each characteristic (Wall, Hirsh et al. 2005),. This is in contrast to the conclusions of Drummond et. al.(Drummond, Raval et al. 2005) who claim that CAI, expression and abundance are the only statistically significant determinants of evolutionary rate. These authors justify their conclusion by presenting a model where the constraint is imposed because of higher cost of protein instability in highly expressed proteins. While we observe that both CAI and, to a lesser extent, abundance correlate with essentiality, that correlation disappears completely in our analysis because we only compared genes with paralogs in *E* and *N* families. Neither the essential genes alone in *E-families* nor their non-essential paralogs differed in CAI or abundance from non-essential genes in *N* families. *(Table S4, S5)*

Calculation of CAI

We used the codonw program (Sharp and Li 1987) to calculate CAI with the *S. cerevisiae* background distribution. While we observed the previously reported difference between essential genes(Winzeler, Shoemaker et al. 1999; Giaever, Chu et al. 2002) and non-essential genes, this difference disappeared when we confined our analysis to essential genes with paralogs in *E* families. Non-essential genes in *E* families showed no significant difference in codon usage from other non-essential genes with paralogs (genes in *N* families) either. Interestingly, there was a large (almost twofold) difference in CAI between genes with paralogs and genes without paralogs. Codon adaptation index is known to correlate well with expression (Coghlan and Wolfe 2000), which is consistent with our observations where the average CAI and expression increase twofold between all-genes and genes with paralogs. The relationship between CAI and expression has a corollary in two-fold difference in protein abundance (see below).

Table S4. *Codon adaptation index compared between essential and nonessential genes and genes with paralogs. While Essential genes have slightly larger CAI when compared to all non-essential genes, members of E and N families do not differ significantly in CAI.*

	Essential Genes	*NonEssential Genes*	P-Val
All Genes	0.197	0.176	<1e-05
	E-families	*N-families*	
Only genes in families	0.28337	0.26238	.31
Only non-essential genes in families	0.26826	0.26238	.76

Abundance

Finally, using protein abundance data from (Ghaemmaghami, Huh et al. 2003), we tested whether the observed difference in the rate of evolution can be attributed to previously observed correlation between evolutionary rate and protein abundance (Pal, Papp et al. 2001; Drummond, Bloom et al. 2005; Drummond, Raval et al. 2005). We found that abundance does not vary between *E* and *N*-families in a statistically significant way. There was a slight difference between all essential genes and all non-essential genes. The difference in protein abundance between essential and non-essential is consistent with CAI (see above). We found that abundance varied twofold when comparing all genes and only genes with paralogs *(Table S5)* This is similar to the variance observed in CAI *(Table S4).* Like with CAI, when we consider all genes, there is a significant, albeit relatively weak (P-val = .02) correlation between essentiality and abundance. Essential genes show slightly greater abundance. However, when we compared only genes in families of paralogs, we observed no significant difference between genes in E-families and N-families. We did not observe a statistically significant difference when we compared essential genes in E-families vs non-essential in N-families, non-essential genes in E families vs non-essential genes in N-families or all genes in E and N-families.

Table S5: *Comparison of protein abundance between essential and non-essential genes. First, we divide all genes into 772 essential and 3096 non-essential ones that have observable abundance levels. Here, the difference in abundance level is significant at P =.02. The means for non-essential genes in paralogous families represent 150 genes in E-families and 412 genes in N-families.*

	Essential Genes	*NonEssential Genes*	P-Val
All Genes	1.7e4	1e4	0.025837
	E-families	*N-families*	
Non-Essential Genes in Paralogous Families	2.9e4	2.7e4	0.81076

Average sequence separation of paralogous families in E. coli and C. elegans

To confirm that the results reported in (Fig 2a,b) were not specific to yeast or a particular cutoff value, we constructed DDGs for *E. coli K12 and C. elegans.* We picked these organisms because they both have high-throughput essentiality data. Each graph was partitioned into *E* and *N* families using essentiality data from (Gerdes, Scholle et al. 2003) for *Escherichia Coli K12* and (Fraser, Kamath et al. 2000; Kamath, Fraser et al. 2003; Simmer, Moorman et al. 2003) for *C. elegans*. If the family had at least one essential gene, it was classified as an E-family, else as an N-family. We then calculated Ka values using PAML(Yang 1997; Yang and Nielsen 2000) and sequence identity using BLAST between all pairs of paralogs. Analogous to the results in Fig 2a,b and Table S1 in *S. cerevisiae*, we found that paralogs were, on average, farther separated in E-families in both organisms. This is consistent whether sequence separation between paralogs is calculated using Ka with PAML or sequence identity with BLAST. (Table S6a,S6b) Furthermore, we showed that the results were independent of the choice of cutoff used to define paralogous families.

Table S6a. *Average separation of paralogs in E. coli. At 20% cutoff, E families have 353 pairwise distances while N-families have 521. Sequence similarity between paralogs is smaller in E-families while divergence in Ka is significantly larger. Independence of the results on the cutoff is shown. As the cutoff is increased, the difference between the families increases slightly. Even though at the highest cutoff of 40% E-families have only 40 paralog distances while N-families have 265, the difference is still highly significant. Calculations of distance using Ka behave exactly the same with respect to cutoff. (Data not shown)The table below shows comparisons between E and N families in both <Ka> and sequence identity.*

	E-families	*N-families*	P-Val
<Ka>	0.63	0.43	1.1e-16
Sequence Identity @20% cutoff	33%	51%	<1e-20
Sequence Identity @30% cutoff	36%	55%	<1e-20
Sequence Identity @35% cutoff	40%	60%	<1e-20
Sequence Identity @40% cutoff	44%	68%	<1e-20

Table S6b. *Average separation of paralogs in C. elegans. E- families represent 277pairwise paralog distances while N families represent 1300. Sequence similarity between paralogs is smaller in E- families while divergence calculated using non-synonymous substitutions (Ka) is significantly larger. Independence of the results on the cutoff is shown. As the cutoff is increased, the difference between the families increases slightly. Even though at the highest cutoff of 40% E- families have only 165 paralog distances while N families have 997, the difference is still highly significant. Calculations of distance using <Ka> behave exactly the same with respect to cutoff.*

	E-families	*N-families*	P-Val
<Ka>	0.38	0.30	1e-10
Sequence Identity @20% cutoff	54%	67%	1e-12
Sequence Identity @30% cutoff	60%	72%	8e-12
Sequence Identity @35% cutoff	65%	75%	1e-10
Sequence Identity @40% cutoff	69%	78%	2e-7

Control for family size distribution in S. cerevisiae
To address the question of whether the results are simply a function of the distribution of family sizes, we recalculated the Ka (as in Fig 2a) and sequence identity (as in Fig 2b) using only families of specific size. Table S7a shows that, even when the families were all of the same size, paralogs in E-families were farther diverged as calculated by sequence identity. We also recalculated the difference using PAML (Ka) with the same results (Table S7b)

Table S7a *Average sequence separation of paralogs in S. cerevisiae controlling for family size. Here, we consider only families of specific sizes of 2,3,4,5 members. Independently of sequence family size, divergence between paralogs (calculated using BLAST Sequence Identity) is significantly larger in E-families. Its worthwhile to note that average sequence identity falls for both types of families as we consider families of larger sizes. This is expected because the larger family sizes have farther diverged members.*

Family Size	Total Num Paralogs in E-families	Total Num Paralogs in N-families	E-families Average Seq Id	N-families Average Seq Identity
2	80	444	53	69
3	15	114	51	66
4	28	40	47	67
5	10	35	38	61

Table S7b *Average sequence separation of paralogs in S Cerevisiae controlling for family size. We build DDG as described above. Independently of sequence family size, divergence calculated using non-synonymous substitutions (Ka) is significantly larger for paralogs in E-families.*

Family Size	Total Num Paralogs in E-families	Total Num Paralogs in N-families	Average Ka E-families	Average Ka N-families
2	80	444	.45	.21
3	15	114	.75	.27
4	28	40	.59	.26
5	10	35	1.07	.26

The annotated gene families for yeast *S. cerevisiae* can be found at:
romi.bu.edu/td/nonlethal_ssc_family.dat – N families
romi.bu.edu/td/lethal_ssc_family.dat – E families
an alternative set of files annotated using SGD TermFinder is at
romi.bu.edu/td/nonlethal_functions.xls –N families
romi.bu.edu/td/lethal_functions.xls – E-families

References:

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

codonw "http://codonw.sourceforge.net/."

Coghlan, A. and K. H. Wolfe (2000). "Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae." Yeast **16**(12): 1131-45.

Cormen, T. H. (2001). "Introduction to algorithms, second edition." 2nd. from http://aeryn.mit.edu/emetrics/count.php?http://libproxy.mit.edu/login?url=http://library.books24x7.com/library.asp? B&isbn=0262032937 Click here for the electronic version.

Drummond, A. D., J. D. Bloom, et al. (2005). "Why highly expressed proteins evolve slowly." arxiv.org(q-bio.PE/0506002).

Drummond, A. D., A. Raval, et al. (2005). "A single determinant for the rate of yeast protein evolution." arxiv.org(q-bio.PE/0506011).

Fraser, A. G., R. S. Kamath, et al. (2000). "Functional genomic analysis of C. elegans chromosome I by systematic RNA interference." Nature **408**(6810): 325-30.

genbank "ftp://ftp.ncbi.nih.gov."

Gerdes, S. Y., M. D. Scholle, et al. (2003). "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655." J Bacteriol **185**(19): 5673-84.

Ghaemmaghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-41.

Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.

Hirsh, A. E. and H. B. Fraser (2001). "Protein dispensability and rate of evolution." Nature **411**(6841): 1046-9.

Hurst, L. D. and N. G. Smith (1999). "Do essential genes evolve slowly?" Curr Biol **9**(14): 747-50.

Jordan, I. K., I. B. Rogozin, et al. (2002). "Essential genes are more evolutionarily conserved than are nonessential genes in bacteria." Genome Res **12**(6): 962-8.

Kamath, R. S., A. G. Fraser, et al. (2003). "Systematic functional analysis of the Caenorhabditis elegans genome using RNAi." Nature **421**(6920): 231-7.

Kellis, M., N. Patterson, et al. (2003). "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature **423**(6937): 241-54.

Kondrashov, F. A., I. B. Rogozin, et al. (2002). "Selection in the evolution of gene duplications." Genome Biol **3**(2): RESEARCH0008.

Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science **302**(5649): 1401-4.

Pal, C., B. Papp, et al. (2001). "Highly expressed genes in yeast evolve slowly." Genetics **158**(2): 927-31.

Pal, C., B. Papp, et al. (2003). "Genomic function: Rate of evolution and gene dispensability." Nature **421**(6922): 496-7; discussion 497-8.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-95.

Simmer, F., C. Moorman, et al. (2003). "Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions." PLoS Biol **1**(1): E12.

Wall, D. P., A. E. Hirsh, et al. (2005). "Functional genomic analysis of the rates of protein evolution." Proc Natl Acad Sci U S A **102**(15): 5483-8.

Winzeler, E. A., C. I. Castillo-Davis, et al. (2003). "Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays." Genetics **163**(1): 79-89.

Winzeler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." Science **285**(5429): 901-6.

Yang, J., Z. Gu, et al. (2003). "Rate of protein evolution versus fitness effect of gene deletion." Mol Biol Evol **20**(5): 772-4.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-6.

Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Mol Biol Evol **17**(1): 32-43.