Letter

Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution

Dmitri M. Krylov, Yuri I. Wolf, Igor B. Rogozin, and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Lineage-specific gene loss, to a large extent, accounts for the differences in gene repertoires between genomes, particularly among eukaryotes. We derived a parsimonious scenario of gene losses for eukaryotic orthologous groups (KOGs) from seven complete eukaryotic genomes. The scenario involves substantial gene loss in fungi, nematodes, and insects. Based on this evolutionary scenario and estimates of the divergence times between major eukaryotic phyla, we introduce a numerical measure, the propensity for gene loss (PGL). We explore the connection among the propensity of a gene to be lost in evolution (PGL value), protein sequence divergence, the effect of gene knockouts on fitness, the number of protein–protein interactions, and expression level for the genes in KOGs. Significant correlations between PGL and each of these variables were detected. Genes that have a lower propensity to be lost in eukaryotic evolution accumulate fewer substitutions in their protein sequences and tend to be essential for the organism viability, tend to be highly expressed, and have many interaction partners. The dependence between PGL and gene dispensability and interactivity is much stronger than that for sequence evolution rate. Thus, propensity of a gene to be lost during evolution seems to be a direct reflection of its biological importance.
We sought to investigate the connection between the two distinct measures of the evolutionary conservation of a gene: (1) the newly introduced propensity for gene loss (PGL) and the rate of sequence evolution and (2) the major variables that determine the functional importance of a gene, namely, the effect of gene knockout on fitness, interactivity, and expression level. For this analysis, we used the recently developed collection of clusters of eukaryotic orthologous groups (KOGs) of proteins from seven (nearly) completely sequenced eukaryotic genomes (Tatusov et al. 2003), which allowed us to construct a parsimonious scenario of gene losses along the branches of the eukaryotic phylogenetic tree. We introduce here a numerical measure for gene loss, PGL, and show a statistically significant positive correlation between PGL and evolutionary rate of a KOG. Furthermore, both PGL and sequence divergence strongly and negatively correlate with the fitness effect of knockout, interactivity, and expression level of the respective gene. The protein sequences of genes that are rarely lost during evolution change relatively slowly; these genes tend to be essential for the survival of an organism and are highly expressed.

RESULTS

The Data Set of Conserved KOGs and Distribution of Gene Losses Over the Eukaryotic Phylogenetic Tree

The KOG database contains 5873 KOGs represented in two to seven eukaryotic genomes: the plant Arabidopsis thaliana; animals C. elegans, Drosophila melanogaster, and Homo sapiens; fungi S. cerevisiae and Schizosaccharomyces pombe; and the microsporidian E. cuniculi (Tatusov et al. 2003; http://www.ncbi.nlm.nih.gov/COG/new/shokog.cgi). According to the phylogeny of the eukaryotic crown group that is currently considered most likely (Hedges 2002), plants branched off first, followed by the divergence of the fungi-microsporidian and metazoan (animal) clades (Fig. 1). For the purposes of the present analysis, we chose a subset of KOGs that are represented in at least three species and could be traced back to the last common ancestor of plants, animals, and fungi. If the amount of HGT between complex eukaryotes is considered to be negligible, reconstruction of the ancestral gene set becomes straightforward: All 3140 KOGs shared by Arabidopsis and any two of the other species should be considered ancestral (KOGs consisting of only two species were not analyzed).

Given a tree topology, the most parsimonious evolutionary scenario resulting in the observed distribution of the phyletic patterns of KOGs can be reconstructed by using the evolutionary parsimony principle. For the purpose of this reconstruction, the phyletic pattern of each KOG was treated as a string of binary characters (one, the presence of the given species; zero, its absence in the given KOG). Given the implausibility of HGT between eukaryotes, the Dollo parsimony principle, under which gene loss is treated as irreversible (a gene can be lost independently in several evolutionary lineages but cannot be regained), was adopted (Farris 1977).

In the resulting parsimonious scenario, each branch was associated with the number of gene losses such that the sum total of losses was minimal, with the exception of the plant branch and the branch leading to the common ancestor of fungi and animals: Gene losses could not be assigned to these branches with the current set of genomes (Fig. 1). The evolutionary scenario includes a massive gene loss in the fungal clade, with additional loss in the microsporidian, and subsequent substantial gene loss in each of the animal lineages, particularly in the nematodes and arthropods (Fig. 1).

Figure 1  The phylogeny of eukaryotes and PGL calculations. (A) Estimated divergence times in millions of years ago (MYA) are shown for all internal nodes of the tree; the estimates are from Hedges et al. (2001). The number of lost genes according to the reconstructed parsimonious scenario is shown next to each branch. (B, C) Examples of PGL calculation. The presence and absence of a gene in each of the extant species is indicated by “+” and “−”, respectively. Red branches are those that retained the gene; blue branches are those to which a loss was mapped. (D) The loss of gene in the branch leading to the common ancestor of yeasts and microsporidian is shown by a blue dot because this branch formally has zero length.

Propensity for Gene Loss

The simplest numerical measure for gene loss in a group of orthologs is the fraction of lineages in which a given gene has been lost. However, the one/zero scoring scheme for gene loss and preservation in different lineages does not reflect the time during which a particular gene was lost or preserved. This time can be different for different lineages, which renders the binary measure inaccurate. In our reconstruction of the parsimonious evolutionary scenario, we mapped gene losses onto the widely accepted phylogenetic tree for the analyzed lineages. The PGL for each gene (KOG) was then calculated by taking into account the tree topology and the available time estimates for each divergence point (Hedges et al. 2001; Hedges 2002; Hedges and Kumar 2003). The logic behind this calculation was as follows. Each branch of the phylogenetic tree was treated as an independent trial during which the given gene was either preserved or lost. The longer the time during which a gene could have been lost, but was not, compared with the total time available, the lower the propensity of this gene to be lost (Fig. 1; for details, see Methods).

A PGL value of zero corresponds to KOGs that are represented in all seven species. A PGL value of one, in theory, would be assigned to a gene present in the last common ancestor of the analyzed species but lost in all lineages. Such genes, for obvious reasons, cannot be detected, and in practice, PGL values can
range from zero to some maximum value less than one. In the data set analyzed here, the PGL values varied from zero to 0.49, the upper limit of PGL being a function of the number of lineages included and the times since their divergence. Genes with PGL value that was estimated as zero using the current data set of seven species (i.e., that were not lost in any of these seven species) might, in reality, have some propensity to be lost in other species. Nevertheless, the PGL values remain meaningful and internally consistent for this data set inasmuch as they are used to estimate the relative propensity for gene loss among all analyzed genes over the time elapsed since the last common ancestor of the compared species. The highest PGL value obtained here, 0.49, is the maximum only for the genes and species considered in this analysis; as additional genomes are included, greater PGL values will result.

The Dependence Between Gene Loss and Sequence Evolution Rate

The tendency of a gene to be lost and sequence evolution rate are two variables that characterize the evolutionary conservation of the gene. A priori, these variables could be considered independent. For example, a protein potentially could evolve relatively fast due to relaxed functional constraints but have a low propensity for loss linked to an essential function. For the purposes of the present analysis, we used the mean evolutionary distance between the KOG member from Arabidopsis (the outgroup with respect to the other analyzed species; Fig. 1) and the rest of the KOG members as the measure of the sequence evolution rate characteristic of the KOG (gene) as a whole. When the PGL values for the analyzed sample of 3140 KOGs were plotted against the evolutionary rates (determined with several methods, see Methods), clear positive correlation was observed (Table 1). The correlation coefficient (R) ranged from 0.3–0.4; depending on the distance measure used, whereas all correlations were statistically highly significant (p < 10^–6). Thus, the assumption of independence of the two variables could be rejected with a high degree of confidence. There is a definite connection between the two facets of evolutionary conservation: The more often a gene is lost, the more substitutions it typically accumulates. However, it is equally notable that the interdependence of the two values is not overwhelmingly strong as only 10%–15% of the variation in the sequence evolution rate can be explained by variation in PGL (and vice versa).

Viability of Knockouts of Yeast Genes With Different Propensities for Loss

Intuitively, it appears that the propensity of a gene to be lost should strongly correlate with the effect of gene knockouts on the viability of the organism. Indeed, one would surmise that if a gene is never lost during a long span of evolution, this is because its function is essential for survival. The PGL values for those KOGs that are represented in S. cerevisiae were superimposed over the available data on the effect of gene knockout on yeast viability (Giaever et al. 2002). More than half of the genes with PGL equal to zero, that is, those that have not been lost in any of the seven lineages considered here, are essential; that is, the respective knockouts are lethal (Fig. 2). The fraction of essential genes was dramatically lower in all other PGL classes (P < 10^–4 by the x^2 criterion). Thus, genes with the lowest propensity for loss during evolution seem to be involved in indispensable functions to a much greater extent than are those genes that have been lost in some lineages. Although one might expect that the fraction of essential genes among those with PGL = 0 could be somewhat lower in more complex organisms due to functional redundancy among paralogs, the conservation pattern of a gene expressed numerically through PGL still could be a reasonable predictor of essential gene functions.

In contrast to the strong connection between the PGL and (in)dispensability of a gene, and in agreement with the previous report (Hirsh and Fraser 2001), we found no appreciable correlation between the sequence evolution rate and dispensability. Among the genes with PGL = 0, the sequence evolution rate was slightly lower for essential genes, but the difference in rates between essential and nonessential genes was statistically significant (p < 0.05) for only one method of evolutionary rate calculation, the PAM distances (Table 2). Thus, although PGL positively, and strongly, correlates with both sequence evolution rate and dispensability, the latter two variables are not significantly correlated; that is, they appear to be (nearly) independently linked to PGL.

Propensity for Gene Loss, Substitution Rates, and Expression Levels

A highly significant negative correlation between the evolutionary rate of yeast genes has been reported. Highly expressed genes appear to evolve slowly (Pal et al. 2001). We examined the correlation between the gene expression levels in various organisms, PGL, and the sequence evolution rate. A significant negative correlation was detected between the expression level and both measures of evolutionary conservation; that is, highly expressed genes tend to evolve more slowly and to be less prone to loss in various lineages than are genes expressed at lower levels. Although the correlation coefficient varied for different measures of evolutionary distance, it was consistently greater for sequence evolution rate than for PGL (Table 1).

Number of Protein–Protein Interactions, PGL, and Substitution Rates

Genes with products that are involved in numerous protein–protein interactions tend to evolve more slowly than do those that have few interaction partners, although the magnitude of the difference varied in different studies and was not dramatic in any of them (Fraser et al. 2002; Jordan et al. 2003). We examined the correlation between PGL and sequence evolution rate, on the one hand, and the number of protein–protein interactions for the KOG members from yeast on the other hand. To this end, the data set collected in the General Repository for Interaction Datasets (GRID) database (Breitkreutz et al. 2003) was used as the

| Table 1. Correlation (R) Between the Propensity for Gene Loss, Substitution Rates, Gene Expression Level, and the Number of Protein–Protein Interactions |
|---------------------------------|-----------|-----------|-----------|---|
|                                | PGL      | Expression in | Interactions |
|                                |          | Yeast       | Worm       | Human     | in yeast |
| PGL                            |          | –0.179      | –0.120     | –0.202    | –0.341   |
| P-distance                      | 0.336    | 0.312       | 0.260      | 0.359     | 0.188    |
| PAM (average from A.t.)b        | 0.368    | 0.164       | 0.133      | 0.204     | 0.169    |
| JTT (average from A.t.)         | 0.317    | 0.271       | 0.226      | 0.299     | 0.178    |
| JTT (three-kingdom average)    | 0.403    | 0.286       | 0.227      | 0.311     | 0.213    |
| JTT (average from A.t., γ-corrected) | 0.300    | 0.230       | 0.198      | 0.263     | 0.165    |

*Different method for evolutionary distance (a surrogate for substitution rate) calculation are introduced in Methods.

b A.t., Arabidopsis thaliana.
source of protein–protein interaction data. We found a strong
negative correlation between the number of protein–protein in-
teractions per protein and PGL, and a weaker correlation with
various measures of sequence evolution rate (Table 1). Both cor-
relations were highly statistically significant ($P < 10^{-16}$). Further-
more, when the KOGs were binned according to their PGL val-
ues, the difference in the mean number of interactions of yeast
proteins between the bins appeared dramatic (Fig. 3). Thus, pro-
teins that have many interaction partners seem to be substan-
tially less prone to loss during evolution than those with
fewer partners, and this connection is much stronger than that
between the interactivity and sequence evolution rate. This is
compatible with the observation that highly connected proteins
in the yeast interaction network include a higher proportion of
essential gene products than do proteins with fewer interactions
(Jeong et al. 2001).

**DISCUSSION**

Sequence evolution rate is a traditional measure of the conserva-
tion during evolution of a gene. Early molecular evolutionary
studies have unequivocally shown that different genes evolve at
substantially different rates (Kimura 1983). However, only with
the advent of genomics and other kinds of “omics”, such as ge-
nome-wide analysis of gene expression and protein–protein in-
teractions, has the opportunity presented itself to systematically
explore the connections between the evolution rate and various
other characteristics of genes (Wolfe and Li 2003). The results of
these studies so far have been somewhat disappointing, in that a
truly strong correlate of the evolution rate had not been identi-
fied. It has been shown that slow-evolving genes tend to be
highly expressed (Pal et al. 2001) and encode longer proteins
(Lipman et al. 2002) that tend to be involved in a somewhat
greater number of protein–protein interactions than are fast-
evolving gene products (Fraser et al. 2002; Jordan et al. 2003).
However, establishing the significance of each of these correla-
tions required careful examination of statistical evidence. In
other words, none of these correlations is particularly strong, and
none can explain much of the variation in evolution rate, al-
though they are statistically significant thanks to the massive
amounts of genomics data. Notably, the results of direct tests of
Wilson’s knockout rate hypothesis are in the same category: Knockout of slow-evolving genes tends to have a greater effect on
fitness than does knockout of fast-evolving genes, but the con-
nection is relatively weak, to the point that some studies have
failed to support its significance (Hurst and Smith 1999; Hirsh
and Fraser 2001; Jordan et al. 2002; Pal et al. 2003).

These observations incite the iconoclastic idea that sequence
evolution rate might not be the most biologically relevant
measure of the evolutionary conservation of a gene. Here
we explored an alternative, the propensity of a gene to be lost
during evolution, a characteristic that obviously can be measured
only through comparison of multiple complete gene sets. PGL is
a much more intuitive correlate of the dispensability of a gene
than is sequence evolution rate; indeed, if a gene is never lost
during evolution, that is probably because it is essential for vi-
ability. However, the connection is not as trivial as it seems to be
at first glance because it is based on a strong assumption, namely,
the transfer of the information on the essentiality of a gene in
one organism (e.g., yeast) to its ortholog in another, vastly dif-
ferent organism (e.g., worm). Actually, the conservation of essen-
tiality is not guaranteed because a gene might be rendered non-
essential by the evolution of redundancy, in the form of paralogs
or unrelated but functionally analogous genes. This might be
followed by the loss of a formerly essential gene, resulting in
nonorthologous gene displacement (Koonin and Mushegian
1996).

Empirically, we observed a strong connection, but definitely
not a one-to-one correspondence, between PGL and knockout
viability, and a highly significant positive correlation between
PGL and sequence evolution rate. In contrast, sequence evolu-
tion rate and viability are linked weakly at best. This suggests that
PGL carries with it a strong biological signal, which is directly
linked to the dispensability of a gene and less directly, even if
indisputably, to the sequence evolution rate. By transitivity, it
should be expected that the latter two variables are also corre-
lated but that connection is nearly lost in the statistical noise.
Thus, a gene shown to be essential in a particular organism has a
strong tendency to be retained and, by implication, to be essen-
tial even in phylogenetically remote lineages; the protein se-
quences encoded by such genes also might tend to evolve slightly
slower than do those of nonessential genes.

These conclusions are supported by the detected strong cor-
relation between PGL and the interactivity of a protein: Hubs of
the protein interaction network are lost during evolution much
less readily than are proteins with few interaction partners, and
this connection is much stronger than that between interactivity

![Figure 2](image_url) Distribution of essential and nonessential yeast genes among PGL classes. Yeast proteins were binned into four classes according to the PGL values for the corresponding KOGs. The number of essential (E) and nonessential (N) genes in each class is indicated. If there were multiple yeast paralogs in a KOG, the KOG was counted as essential if at least one of the paralogs was essential.

![Figure 3](image_url) PGL and number of protein–protein interactions for yeast proteins. Yeast proteins were binned into four classes according to the PGL values for the corresponding KOGs. The average number of interactions for all paralogs was used, with the rationale that the latter two variables are also correlated but that connection is nearly lost in the statistical noise. PGL is a much more intuitive correlate of the dispensability of a gene than is sequence evolution rate; indeed, if a gene is never lost during evolution, that is probably because it is essential for viability. However, the connection is not as trivial as it seems to be at first glance because it is based on a strong assumption, namely, the transfer of the information on the essentiality of a gene in one organism (e.g., yeast) to its ortholog in another, vastly different organism (e.g., worm). Actually, the conservation of essentiality is not guaranteed because a gene might be rendered nonessential by the evolution of redundancy, in the form of paralogs or unrelated but functionally analogous genes. This might be followed by the loss of a formerly essential gene, resulting in nonorthologous gene displacement (Koonin and Mushegian 1996).
and sequence evolution rate. This is compatible with the previous reports on the connection between interactivity and dispensability (Jeong et al. 2001) and with the general notion that scale-free networks, such as the network of protein–protein interactions, are tolerant to error (random elimination of weakly connected nodes) but are highly vulnerable to attack (directed elimination of the hub; Albert et al. 2000; Barabasi 2002). Because protein–protein interaction domains generally show limited sequence conservation (whereas structure conservation is crucial), it is perhaps not unexpected that the connection between interactivity and sequence evolution rate could be detected (at best) only as a relatively weak statistical trend. Surprisingly, however, the observations reported here indicate that gene expression level more strongly correlated with sequence evolution rate than with PGL. Generally, one would expect the same trends to be seen with dispensability, interactivity, and expression level. If validated by further analysis of more robust and extensive expression data, this inversion could suggest a non-trivial connection between expression level and sequence conservation, the nature of which remains to be explored.

PGL and sequence evolution rate are measures of evolutionary conservation that seem to capture substantially different aspects of evolution. PGL is a much more direct reflection of the biological dispensability of a gene, whereas sequence evolution rate is a much more refined analysis will become feasible as the collection of sequenced eukaryotic genomes grows. The ratio of the previously estimated times since divergence to expected species was chosen. The sequences from the respective KOG were compared to each other by using the BLASTP program (Altschul et al. 1997), and for each species, the sequence that had the best cumulative score with the sequences from the other species was selected.

Other Data
The data on gene knockout effects in yeast were primarily from Giaever et al. (2002). The SGD database (http://genome-www. stanford.edu/Saccharomyces/) was used to collect the knockout viability data for each individual gene. For KOGs with multiple yeast paralogs, only the most conserved paralog identified as described above was considered.

The GRID database was used as the source of data on protein–protein interactions (Breitkreutz et al. 2003). All duplicate interactions were collapsed into one entry. Absence of interactions in GRID for a given gene was interpreted as zero interactions.

Average expression levels of C. elegans genes were from Hill et al. (2000). Expression levels of human genes were estimated in the following fashion: The human CDSS were used as queries in a BLASTN search against the dbEST database. Hits with >98% identity for alignment length >400 nt or with >95% identity for alignment between 100 to 400 nt were tallied, and the number of ESTs was taken as the expression level for the respective gene. Expression levels of yeast genes were obtained from the published microarray analysis by averaging the control (no diauxic shift) data (DeRisi et al. 1997). For all three organisms, gene expression data were mapped to KOGs, and if more than one paralog was present in a KOG, the maximum expression level for the given organism was assigned to the KOG.

Divergence Times of E. cuniculi, S. cerevisiae, and S. pombe
Phylogenetic trees for CDC28 kinase, glyceraldehyde-3-phosphate dehydrogenase (GPDH), small chain of ribonucleoside-di-phosphate reductase (RDR), and triosephosphate isomerase (TIM) families were constructed by using the Mega and ProtML packages (Adachi and Hasegawa 1992; Kumar et al. 1994). The lengths of the branches connecting E. cuniculi, S. cerevisiae, and S. pombe were taken to be proportional to the divergence times for these lineages. The divergence times were calculated by using the estimates for the other eukaryotic lineages (Wang et al. 1999). The ratio of the previously estimated times since divergence to branch lengths for A. thaliana, H. sapiens, C. elegans, and D. melanogaster was used to calibrate the branches of the tree in years. An average estimate over the CDC28, GPDH, RDR, and TIM families was used as the estimate of the time of divergence of E. cuniculi, S. cerevisiae, and S. pombe.

PGL Calculations
By using the published estimates (Wang et al. 1999) and our own estimates for the divergence times of E. cuniculi, S. cerevisiae, and S. pombe, specific divergence times were assigned to each internal node (ancestral form) in the phylogenetic tree of the eukaryotic crown group (Fig. 1A). Given a phyletic distribution pattern,
branches of the tree associated with gene loss (B_L) can be identified (Fig. 1B,C). Designating those branches of the tree, in which the given gene was preserved B_P, we have

\[ PGL = \frac{\Sigma B_L}{(\Sigma B_P + \Sigma B_L)} \]

In terms of Fig. 1, B and C, this is the ratio of the sum of the lengths of blue branches to the sum of the lengths of all colored branches. Thus, for a gene present in Arabidopsis, human, and C. elegans but lost in the Drosophila branch and the Fungi-Microsporidia branch (Fig. 1B),

\[ PGL = \frac{(1063 + 0)}{(1642 + 100 + 322 + 1220 + 157 + 1063 + 1063 + 0)} = 0.19 \]

Similarly, for a gene found in Arabidopsis and the two yeast species (lost in the Metazoa branch and in the E. caniculi branch, Fig. 1C),

\[ PGL = \frac{(322 + 1542)}{(1642 + 100 + 705 + 837 + 837 + 322 + 1542)} = 0.31 \]

Calculation of Evolutionary Distance Between Protein Sequences

Evolutionary distances between proteins in a KOG were calculated from multiple alignments. To obtain the P-distance multiple alignments of protein sequences were constructed, and distances between orthologs were calculated as the proportion of different amino acids. All positions in the alignment containing a deletion or insertion in at least one of the sequences were removed prior to calculating P-distance. P-distances were measured relative to A. thaliana orthologs for all KOGs; their mean value was used as the distance characteristic for the given KOG. Similarly, evolutionary distances between proteins were calculated by using the PAM (Dayhoff et al. 1983) or JTT (Jones et al. 1992) substitution matrices and the mean distance from using the PAM (Dayhoff et al. 1983) or JTT (Jones et al. 1992) substitution matrices and the mean distance from A. thaliana to other species was used for further analysis. The three kingdom mean distance was calculated as the unweighted average of the mean distances among plants, animals, and fungi. JTT matrix distances were also calculated with \( \gamma \)-correction by using the ProtDist program with the \( \gamma \)-parameter of 1.0 (Felsenstein 1996).

ACKNOWLEDGMENTS

We thank Fyodor Kondrashov for help in the expression data analysis and I. King Jordan and Roland Walker for useful discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


WEB SITE REFERENCES
http://biodata.mshri.on.ca/grid/servlet/Index; the General Repository for Interaction Datasets (GRID).

Received April 27, 2003; accepted in revised form August 4, 2003.