Nucleic Acids Research Advance Access published December 7, 2006

Nucleic Acids Research, 2006, Vol. 00, No. 00 ***e1–8***
doi:10.1093/nar/gkl792

Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes

Elhanan Borenstein[1,*], Tomer Shlomi[1,*], Eytan Ruppin[1,2] and Roded Sharan[1]

[1]School of Computer Science and [2]School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel

Received July 23, 2006; Revised September 12, 2006; Accepted October 2, 2006

ABSTRACT

The rate of conservation of a gene in evolution is believed to be correlated with its biological importance. Recent studies have devised various conservation measures for genes and have shown that they are correlated with several biological characteristics of functional importance. Specifically, the state-of-the-art propensity for gene loss (PGL) measure was shown to be strongly correlated with gene essentiality and its number of protein–protein interactions (PPIs). The observed correlation between conservation and functional importance varies however between conservation measures, underscoring the need for accurate and general measures for the rate of gene conservation. Here we develop a novel maximum-likelihood approach to computing the rate in which a gene is lost in evolution, motivated by the same principles as those underlying PGL. However, in difference to PGL which considers only the most parsimonious ancestral states of the internal nodes of the phylogenetic tree relating the species, our approach weighs in a probabilistic manner all possible ancestral states, and includes the branch length information as part of the probabilistic model. In application to data of 16 eukaryotic genomes, our approach shows higher correlations with experimental data than PGL, including data on gene lethality, level of connectivity in a PPI network and coherence within functionally related genes.

INTRODUCTION

Large scale sequencing projects are producing genome data at an ever increasing pace. Interpreting the data to study gene importance and function is a major goal of functional genomics. Recently, several studies have related different measures of evolutionary conservation of a gene to various common measures of gene functional importance, including its essentiality (1–5) and the degree of its encoded protein in a protein–protein interaction (PPI) network (3,6,7). These observations support the long-standing 'knockout rate' hypothesis of Wilson *et al*. (8), which claims that the greater the effect of a gene knockout on fitness the slower is its evolutionary rate. The observed correlation between conservation and importance varies however between various conservation measures, underscoring the need for accurate and general measures for the rate of gene conservation.

Nucleotide substitution rate is a traditional measure of the conservation of a gene in evolution. Hirsh and Fraser (1) have shown that the growth rate of a gene deletion mutant correlates with the gene's evolutionary rate. Further studies, computing more accurate evolutionary rate estimations, extended these findings and showed marked differences between the evolutionary rates of essential and nonessential genes (2,5).

An alternative measure for evolutionary conservation, which measures the propensity for gene loss (PGL), was introduced by Krylov *et al*. (3). PGL is computed based on the pattern of presence and absence of genes across multiple genomes, considering their phylogeny, and was shown to have higher correlations with gene dispensability than sequence evolution rate. Indeed, these two measures of evolutionary conservation capture different characteristics of the genes: Sequence evolution rate represents the selective constraints on protein structure and sequence, whereas PGL captures the essentiality of the gene's function. As pointed by Krylov *et al*., a protein linked to an essential function could potentially have a low propensity to be lost, but still evolve relatively fast due to relaxed functional constraints. However, a moderate correlation between the two measures was found, corroborating the intuitive notion that weakly constrained proteins are lost during evolution significantly more often than strongly constrained ones.

The propensity for gene loss is computed based on three types of biological data sources, characterizing the gene and a group of species: (i) the *phyletic pattern* of a gene, which is the pattern of presence–absence of the gene in the set of species genomes; (ii) a phylogenetic tree topology relating the different species; and (iii) branch length estimates

*To whom correspondence should be addressed. Tel: +972 3 640 5378; Fax: +972 3 640 9357; Email: borens@post.tau.ac.il
*Correspondence may also be addressed to Tomer Shlomi. Tel: +972 3 640 5378; Fax: +972 3 640 9357; Email: shlomito@post.tau.ac.il

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

for the tree. (i) and (ii) are used for constructing the *ancestral phyletic pattern* of the presence–absence of the gene in internal nodes of the tree. This reconstruction is based on the Dollo parsimony principle (9), in which a gene loss is deemed irreversible. Given the ancestral phyletic pattern, each branch is treated as an independent trial where the gene was either preserved or lost. The PGL value of the gene is then defined as the ratio between the total length of branches in which the gene is lost and the total length of branches in which the gene could have been lost (i.e. preserved or lost). This definition captures the idea that the longer the time a gene could have been lost but was not, the lower the propensity of this gene to be lost.

While intuitively reasonable, a careful examination of the definition of PGL reveals two potential weaknesses:

(i) Assuming that a gene loss measure should reflect some evolutionary loss rate, a gene which is lost only once during a long branch, may have a lower rate than that of a gene lost several times during short branches. However, PGL considers only the total lengths of branches in which a gene is lost, ignoring the possible variation in lengths, and may lead to counterintuitive predictions, as shown below.
(ii) PGL is based on a single ancestral phyletic pattern obtained by a parsimony principle. However, the simple parsimony model may not describe the correct ancestral phyletic pattern and ignores alternative possible patterns. Moreover, the parsimonious construction of the ancestral phyletic pattern does not take into account the available branch lengths.

PGL additionally lacks an underlying systematic probabilistic model that properly describes the process of gene loss. In contrast, a conservation measure based on such a model will provide a natural way to incorporate additional sources of biological data (e.g. confidence level of gene presence or absence) and could be easily extended to allow for prediction of various pertaining parameters such as branch lengths.

Here we present a novel approach for computing a conservation measure for genes, the *gene loss rate* (GLR), addressing the above shortcomings. The measure is based on a simple probabilistic model of gene evolution, whose underlying assumption is that each gene has a certain rate in which it is lost during evolution. Given a loss rate estimation, the probability of a gene phyletic pattern can be calculated, considering all possible ancestral phyletic patterns. GLR is defined as the maximum-likelihood estimate of this loss rate, i.e. the loss rate that maximizes the probability of the phyletic pattern associated with the gene. Applying such a maximum-likelihood estimate, our model resembles that of (10–13), who used a stochastic birth and death process model to examine gene family evolution, but assumed a uniform birth and death rate for all genes. Recently, a related measure of gene loss was mentioned as part of an analysis covering numerous characteristics of gene function and evolution (14). However, as the above work did not focus on gene loss measures, no systematic comparison with previous methods was conducted.

In the following, we provide a rigorous definition of GLR and derive an algorithm to efficiently compute it. We show that when considering only a single parsimonious ancestral phyletic pattern and restricting gene losses to branches of equal length, GLR collapses to a simple measure that can be analytically related to PGL. We compare the biological plausibility of GLR and PGL using an extensive data set of 16 Eukaryotic species, showing that GLR better correlates with existing measures of gene functional importance.

METHODS

The GLR measure

We developed a novel measure, GLR, for the loss rate of a gene based on maximum likelihood principles. For a phylogenetic tree T (with estimated branch lengths for the species under study) and a gene's phyletic pattern PP, we define GLR as the maximum-likelihood estimate for the rate μ of gene loss:

$$\text{argmax}_\mu L(\mu \mid \text{PP}) = \text{argmax}_\mu \sum_{\text{APP}} P(\text{PP}, \text{APP} \mid \mu) \qquad \mathbf{1}$$

where APP runs over all possible ancestral phyletic patterns for the gene. In the next sections we provide the details of the probabilistic model underlying GLR and the derivation of the GLR measure. We first consider the simple case of a fixed ancestral phyletic pattern and then present the general case, considering all possible ancestral phyletic patterns.

GLR under a fixed ancestral phyletic pattern

Let T be a phylogenetic tree with branch lengths. Given both a phyletic pattern PP and an ancestral phyletic pattern APP of a gene, GLR is defined as $\text{argmax}_\mu P(\text{PP}, \text{APP} \mid \mu)$. Let $c_1, \ldots, c_{n_1}$ denote the lengths of branches in which a gene is conserved, and let $l_1, \ldots, l_{n_2}$ denote the lengths of branches in which a gene is lost. Then

$$P(\text{PP}, \text{APP} \mid \mu) = \Pi_{i=1}^{n_1} e^{-\mu c_i} \Pi_{i=1}^{n_2} (1 - e^{-\mu l_i})$$

where $e^{-\mu t}$ represents the probability that a gene is conserved along a branch of length t, in accordance with the standard model of nucleotide substitutions (15).

To find the rate, μ, that maximizes the above probability, we take the log of both sides of the equation:

$$\log P(\text{PP}, \text{APP} \mid \mu) = -\mu \sum_{i=1}^{n_1} c_i + \sum_{i=1}^{n_2} \log(1 - e^{-\mu l_i})$$

and get that the probability is maximized when:

$$\sum_{i=1}^{n_1} c_i = \sum_{i=1}^{n_2} \frac{l_i e^{-\mu l_i}}{1 - e^{-\mu l_i}}$$

It is easy to show that the log likelihood function above is concave; thus, the maximizing rate can be obtained via gradient ascent. A naive analytical approximation to the rate μ can be obtained by assuming a uniform length l of all branches in which the gene is lost. Under this assumption, we get that:

$$\sum_{i=1}^{n_1} c_i = \frac{n_2 l e^{-\mu l}}{1 - e^{-\mu l}}$$

And thus, GLR, defined as the rate μ which maximizes the probability function, is given in this simple case by:

$$\mu = -\frac{1}{l}\log\left(1 - \frac{n_2 l}{n_2 l + \sum_{i=1}^{n_1} c_i}\right) \quad \mathbf{2}$$

where $[n_2 l/(n_2 l + \Sigma_{i=1}^{n_1} c_i)]$ is the PGL value.

General ancestral phyletic pattern

In the general case of GLR computation, we again apply gradient ascent to find the rate μ that maximizes the probability of the phyletic pattern, P(PP|μ), but now, weigh all possible ancestral phyletic patterns (Equation 1). To this end, we apply a variant of Felsenstein's tiny maximum-likelihood algorithm (16). For simplicity, we describe it for a binary tree. Let i denote a node in the phylogenetic tree, and let i_1 and i_2 denote its two children. Let T_i denote the subtree rooted at node i. Let $p(i,a), a \in \{0,1\}$ denote the probability that the gene is lost (0) or conserved (1) along the l_i-long branch from the parent of i to i. As in the previous section, we assume that $p(i,0) = 1 - e^{-\mu l_i}$ and $p(i,1) = e^{-\mu l_i}$. Let b_i denote whether the gene is present (1) or absent (0) at i. For any given μ, the probability of the subtree T_i is:

$$P(T_i \mid b_i = 0) = P(T_{i_1} \mid b_{i_1} = 0)P(T_{i_2} \mid b_{i_2} = 0)$$
$$P(T_i \mid b_i = 1) = \sum_{x,y \in \{0,1\}} P(T_{i_1} \mid b_{i_1} = x)p(i_1,x)P(T_{i_2} \mid b_{i_2} = y)p(i_2,y)$$

It should be noted that in this recursion, as in Felsenstein's algorithm (16), the probabilities of the two subtrees, T_{i_1} and T_{i_2}, are calculated given a certain presence-absence assignment and are hence independent. Since the gene is always present in the root node of the tree, the probability of the phyletic pattern, $P(\mathrm{PP} \mid \mu)$, is given by $P(T_{\mathrm{root}} \mid b_{\mathrm{root}} = 1)$. The computation of the derivative in this case is more involved and is done using dynamic programming as follows: Denote by $p'(i,a)$ the derivative of $p(i,a)$ according to μ, i.e. $p'(i,0) = l_i e^{-l_i \mu}$ and $p'(i,1) = -l_i e^{-l_i \mu}$. The derivative of the probability of the subtree T_i is:

$$\begin{aligned} P'(T_i \mid b_i = 1) = \sum_{x,y \in \{0,1\}} & [P'(T_{i_1} \mid b_{i_1} = x)p(i_1,x) \\ & + P(T_{i_1} \mid b_{i_1} = x)p'(i_1,x)] \\ & P(T_{i_2} \mid b_{i_2} = y)p(i_2,y) \\ & + [P'(T_{i_2} \mid b_{i_2} = y)p(i_2,y) \\ & + P(T_{i_2} \mid b_{i_2} = y)p'(i_2,y)] \\ & P(T_{i_1} \mid b_{i_1} = x)p(i_1,x), \end{aligned}$$

where $P'(T_i \mid b_i = 0) = 0$ for any node i, as the probability of a subtree that does not have the gene present in its root is independent of μ. We initialize the recursion by setting $P'(i \mid b_i = x) = 0$ for all leaf nodes i and $x \in \{0,1\}$.

Estimating the expected number of gene losses per branch

Given a GLR value for a phyletic pattern, we compute the probability that the gene is lost in each branch of the tree, while considering all possible ancestral phyletic patterns. Given the phyletic pattern and loss rate, the probability that the gene was lost in the branch leading from node i to node j is:

$$P(b_i = 1, b_j = 0 \mid \mathrm{PP}, \mu) = \frac{\sum_{\mathrm{APP}} P(b_i = 1, b_j = 0, \mathrm{PP}, \mathrm{APP} \mid \mu)}{P(\mathrm{PP} \mid \mu)}$$

where both the numerator and denominator can be calculated as shown in the previous section, using the variant of Felsenstein's algorithm. Specifically, the numerator is calculated by modifying this algorithm to consider only ancestral phyletic patterns that are consistent with a gene loss between nodes i and j, setting $P(T_i \mid b_i = 0) = 0$ and $P(T_j \mid j_i = 1) = 0$ in the recursion. The expected number of gene losses for a branch is the sum of probabilities over all gene phyletic patterns.

Data acquisition and processing

A phylogenetic tree of the following 16 Eukaryotes was obtained from NCBI (17): *Homo sapiens*, *Canis familiaris*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Pan troglodytes*, *Gallus gallus*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, *Neurospora crassa*, *Arabidopsis thaliana*, *Oryza sativa* and *Magnaporthe grisea*. The divergence time estimates were collected from multiple sources (3,18–21).

Clusters of homologous genes for these species were obtained from NCBI's HomoloGene database (17). HomoloGene is a system that automatically arranges genes into putative homology clusters based on DNA and protein-based alignment measures, identifying clear-cut paralogs and orthologs.

The data on the lethality of gene knockouts in yeast was obtained from MIPS database (22). The PPI data along with interactions reliability scores were obtained from (23,24). These reliability scores are essential for controlling errors that stem from noisy large-scale PPI experiments, and were previously calculated based on the type of experiments in which the interaction was observed, and the number of observations in each experimental type. In total we assembled 14 319 and 3926 interactions in yeast and worm, respectively.

Data on protein complexes was obtained from the MIPS catalog (25), considering only manually curated complexes (i.e. removing category 550). We used all leaves of the MIPS complex hierarchy, collapsing nodes of level > 3 to level 3. Overall, 35 complexes were compiled. Large scale data on 21 phenotypic effects of single gene knockouts in yeast was obtained from Dudley *et al.* (26).

RESULTS

In the following we describe the results of applying the GLR measure to analyze a tree of 16 eukaryotic organisms, and a comprehensive comparison to the PGL measure. The performance evaluation was based on comparing the two gene loss measures to an array of experimental data on yeast and worm, including data on gene lethality and level of connectivity in a

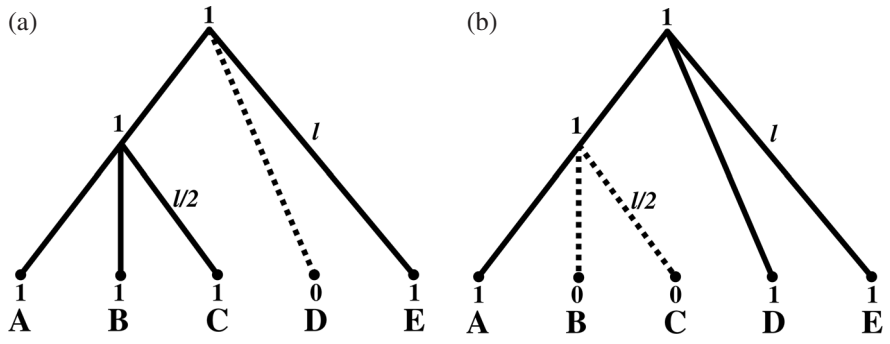

Figure 1. A simple phylogenetic tree, demonstrating the difference between the GLR and PGL measures. Presence or absence of a gene is indicated by 1 or 0, respectively, next to the corresponding node (based on a parsimonious reconstruction of the ancestral phyletic pattern). Dotted lines represent branches in which the gene was lost. (**a**) The gene is lost in one branch of length *l*. (**b**) The gene is lost in two branches, each of length *l*/2.

protein interaction network. To exemplify the difference between the two measures, we start with a simple comparison of GLR and PGL on synthetic data.

An example tree

To demonstrate the difference between PGL and GLR, we first analyze a synthetic tree shown in Figure 1. The tree consists of five species, marked A to E, and two divergence points. The first divergence point, the tree root, occurred *l* years ago, and the second occurred *l*/2 years ago. We consider two phyletic patterns of genes: (i) a gene is present in all species other than D (Figure 1a); and (ii) a gene is present in all species other than B and C (Figure 1b). We reconstruct the corresponding most parsimonious ancestral phyletic patterns for both cases, finding that in phyletic pattern 1, the gene is lost in a branch of length *l*, and in phyletic pattern 2, the gene is lost in two branches of length *l*/2. Given these reconstructions, it would be more reasonable to assign a higher loss rate score to the second phyletic pattern, as it reflects two gene losses during short time periods, compared to a single loss in a longer time period for the first phyletic pattern.

Computing the GLR and PGL scores for both cases reveals that the PGL scores for both phyletic patterns are the same, while GLR correctly captures the above intuitive trend. Specifically, since PGL is defined as the ratio between the total length of the branches in which a gene is lost and the total length of branches in which the gene could have been lost, we get the same PGL value of 1/4 in both cases. In contrast, the GLR score for the second phyletic pattern, calculated based on Equation 2 (see Methods; a fixed ancestral phyletic pattern is assumed), is $\mu = -2/l(log(1-1/4))$ which is twice the size of the GLR for the first phyletic pattern.

Having shown that GLR correctly captures variability in branch lengths in which a gene is lost, while PGL may fail to do so, we turn to compare the two measures using real phylogenetic data and various biological measures of functional importance.

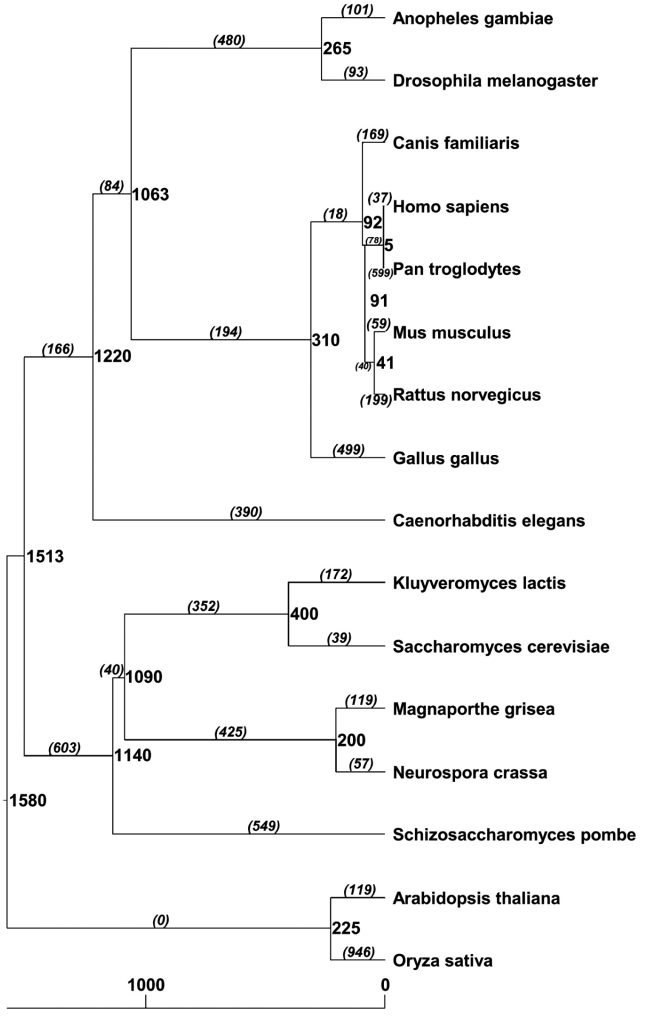

Figure 2. The phylogenetic tree used in our analysis, relating 16 eukaryotes. Estimated divergence times (in millions of years ago) are shown for all internal nodes. The number in parentheses next to each branch indicates the expected number of gene losses (see Methods).

A phylogenetic tree of 16 eukaryotes

We obtained clusters of homologous genes in 16 Eukaryotic species from NCBI's HomoloGene database (17). The Eukaryotic species include nine animals, five fungi, and two plants (Methods). The phyletic pattern associated with each cluster was determined by the species whose genes were present in that cluster. The corresponding phylogenetic tree was taken from NCBI and the divergence time estimates were collected from multiple sources (Figure 2 and Methods). We focused on a subset of the gene clusters in

which a homologous gene was present in either *S.cerevisiae* or *C.elegans*, for which we had additional biological data to validate the conservation measures. Furthermore, as similarly done by Krylov *et al*. (3) on their data, we considered only the clusters for which the gene was present in the common ancestor of the Fungi/Metazoa group and plants under the optimal parsimony reconstruction. These genes are assumed to be present in the root of the phylogenetic tree we study and, hence, form a natural set for gene loss analysis. Furthermore, focusing on simple gene loss dynamics without considering genes duplication, we restrict the analysis to genes whose family size is one, i.e. genes that have only one copy in the genome (see Discussion).

Given a phylogenetic tree, both the GLR and PGL values of a gene are uniquely determined according to the gene phyletic pattern and, hence, cannot distinguish between gene clusters with identical phyletic patterns. Therefore, we group genes clusters according to their phyletic patterns; for each such group and a biological property of interest, we associate with the group the average value of that property over the group's genes. To obtain robust biological data of these phyletic patterns, we included in our analysis only patterns that were associated with at least five gene clusters.

Clearly, the number of different possible values (for either PGL or GLR) is bounded by the number of possible phyletic patterns. PGL was previously calculated based on phyletic patterns consisting of six species (in fact, the phylogenetic tree consisted of seven species, but all phyletic patterns included the species *Arabidopsis thaliana*), markedly limiting the number of possible PGL values. The extended phylogenetic tree used here allows for a significantly higher resolution of gene loss rate values.

The accuracy and robustness of GLR estimations

To estimate the accuracy of the GLR estimations we performed a series of simulations of gene loss dynamics. In each simulation, we chose a 'true' loss rate value, and used it to randomly generate a phyletic pattern according to the gene evolution dynamics assumed in our model (Supplementary Data 1). We computed the GLR value based on the derived phyletic pattern and compared it with the true loss rate used in the simulation. We found a highly significant correlation ($r = 0.802, P < 10^{-300}$, Spearman correlation test; Supplementary Figure 1a) between the true and estimated loss rates. The correlation between PGL and the true loss rates was markedly lower ($r = 0.659, P < 10^{-300}$).

We further examined the robustness of the GLR estimates to modifications of the phylogenetic tree and to phyletic pattern perturbations (Supplementary Data 2). We found that systematically deleting each species in turn from the tree has a relatively minor effect on the estimated GLR values (Supplementary Figure 2a). However, using a substantially smaller tree from (3), which contains only seven species, we got significantly less accurate estimations (Supplementary Data 2). The effect of noise in phyletic patterns was examined by changing the presence–absence values for each species at a time. Following such data modifications, the GLR values obtained were highly correlated with the original estimations, and significantly more robust than the PGL estimates (Supplementary Figure 2b).

GLR, PGL and sequence evolution rate

Next, we examined the correlation between GLR and PGL as well as their correlations with a common sequence evolution rate (SER) measure obtained from (4). We found a statistically significant correlation of 0.833 ($P < 10^{-6}$, Spearman correlation test) between GLR and PGL. Considering that both measures are based on the same data and attempt to capture a similar notion of loss rate, this strong correlation is not surprising. Interestingly, the significant difference between the accuracy of these two measures, demonstrated below, stems from this seemingly low disagreement.

PGL was previously shown to be significantly correlated with SER. The PGL values computed here also show a significant correlation with SER ($r = 0.54$, $P = 7.9 \times 10^{-4}$, Spearman correlation test), higher than the corresponding correlation of GLR with SER ($r = 0.475$, $P = 2.1 \times 10^{-2}$). The correlation between GLR measures and SER is interesting because evolutionary rate reflects selective constraints on protein structure and function whereas GLR captures gene dispensability. However, homology detection by sequence comparison is affected by evolutionary rate, and homology relations for fast evolving genes may remain uncovered. As a consequence of the latter, one would expect that the measured GLR would be larger for fast evolving genes. In other words, the observed correlation is at least in part due to the intrinsic properties of the measurement procedures.

Gene loss correlation with lethality of yeast knockouts

Genes that have low loss rate are assumed to be associated with an essential function for survival and, thus, their knockout is expected to have lethal effects. Such a trend was previously shown for PGL, where the fraction of essential genes was found to be significantly higher for genes with a PGL value of zero, compared to genes with higher PGL values (3). Other studies have also shown similar correlations between SER and lethality (2,5).

The correlation between the GLR measure and the phyletic pattern lethality was -0.655 ($P < 10^{-7}$, Spearman correlation test). The correlation for PGL and lethality was smaller ($r = -0.605, P < 10^{-6}$). The correlation for SER and lethality in these settings was only -0.457 ($P < 10^{-2}$). As the advantages of GLR are expected to be revealed when the gene is lost in several branches, we next restricted the analysis to phyletic patterns for which the most parsimonious reconstruction of the ancestral phyletic pattern included at least two branches in which the gene was lost. Using this restricted set of patterns, the correlation between GLR and lethality was indeed markedly higher and more significant ($r = -0.880$, $P < 10^{-13}$) than that of PGL ($r = -0.709$, $P < 10^{-6}$). The correlation of SER with lethality did not markedly change in this setting ($r = -0.449$, $P = 0.03$). In the rest of the analysis presented in this paper we continue to focus on this subset of phyletic patterns.

To further evaluate the performance of the two measures, we tested their prediction power in classifying the set of yeast genes to lethal and nonlethal. Figure 3 shows the receiver operating characteristic (ROC) curves obtained for each of the measures. Evidently, GLR outperforms PGL in this comparison.

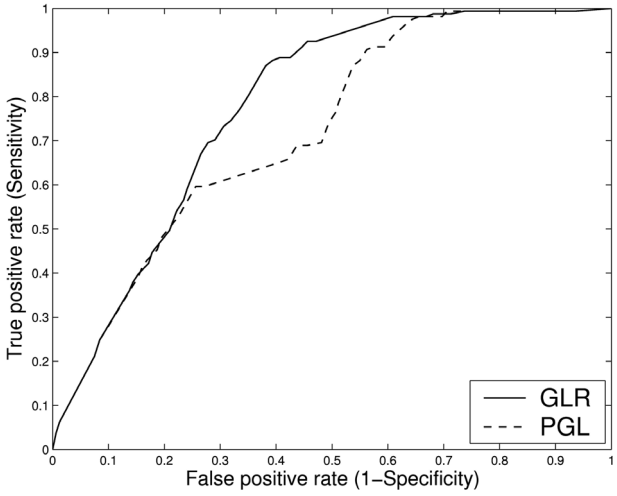

Figure 3. ROC curves for GLR and PGL illustrating the increased specificity and sensitivity of GLR in predicting the lethality of genes. The areas under the ROC curves are 0.78 and 0.72 for GLR and PGL, respectively.

Finally, we examined whether genes that are hardly distinguishable by one of the measures can be correctly classified by the other measure. To this end, we sampled pairs of phyletic patterns for which the PGL scores of the two phyletic patterns were similar (i.e. the difference between the two values falls within the low 10% of differences between all possible pairs), and computed the differences in GLR for those patterns. We found a significant correlation ($r = -0.615$, $P < 10^{-9}$, Spearman correlation test) between these GLR differences and the corresponding differences of lethality values. Conversely, when testing pairs with similar GLR values, no significant correlation could be detected between the differences in PGL values and the differences in lethality values.

Gene loss correlation with protein interaction information

The degree of a protein in a PPI network was shown to be correlated with its dispensability (6,7). Indeed, Krylov *et al.* (3) report a significant correlation between the PGL value of a gene and the degree of its corresponding protein in the yeast PPI network.

To overcome the noise in protein interaction data (27) we assigned reliability estimates to reported interactions and defined the level of connectivity of a protein as the expected number of interactions involving this protein (Methods). Table 1 summarizes the correlations obtained between the GLR or PGL measure of a gene and the connectivity level of its corresponding protein in the PPI networks of yeast and worm. As evident from the table, the correlation of GLR with the experimental network data is significantly higher. Notably, the correlation between SER and connectivity level was not statistically significant.

Gene loss coherence within functionally related genes

Proteins that are part of the same complex (and, hence, share a functional role) are likely to have similar functional importance. Specifically, it has been shown that protein complexes tend to be coherently conserved (28). Motivated by this observation, we examined whether genes that are part of the same complex have coherent GLR (and PGL) values.

Table 1. Correlation between the GLR or PGL measure and the connectivity level in a PPI network

	Yeast	Worm
GLR	-0.429 ($P = 0.004$)	-0.524 ($P = 10^{-4}$)
PGL	-0.316 ($P = 0.04$)	not significant

To this end, we compiled a list of 135 proteins from 59 manually curated complexes (Methods). For each complex, we calculated the standard deviations of the GLR and PGL values of the constituent complex proteins as a measure of coherency. To avoid bias stemming from different distributions of GLR and PGL values, we first normalized both measures to have mean 0 and SD 1. Comparing the coherence of PGL and GLR for all these pairs, we found that genes from the same complex tend to have a smaller GLR standard deviation than PGL ($P < 0.00027$, Wilcoxon signed-rank test for paired data).

Similarly, proteins whose knockout has a certain phenotypic effect are expected to be coherently conserved through evolution. To test whether similarity in phenotypic effect is manifested in coherency of GLR values, we compiled a list of 21 phenotypic effects for 312 single gene knockouts in yeast (26). Comparing the coherency of PGL and GLR values of proteins associated with the same phenotype, as described above, we found higher coherency for GLR than PGL ($P < 0.0034$, Wilcoxon signed-rank test for paired data). This provides further support for the superiority of GLR over PGL.

Expected number of gene losses per branch

We computed the expected number of gene losses in each branch of the phylogenetic tree based on the optimal GLR values (see Methods and Figure 2). Expectedly, we identified a significant correlation of 0.43 ($P < 0.01$, Spearman correlation test) between the expected numbers of gene losses and branch lengths. A few branches in the tree exhibited relatively high gene loss values. Notably, an expected massive loss of 603 genes was identified in the relatively short branch leading from the common ancestor of fungi and metazoa to fungi. As this loss is identified along a branch leading to a relatively large subtree, it is based on the robust evidence concerning the absence of genes from multiple genomes (five, in this case). Other putative massive losses were identified between *O.sativa* and its common ancestor with *A.thaliana* (946 genes), and between *P.troglodytes* and its common ancestor with *H.sapiens* (599 genes). The latter, however, were supported by only a single genome and require further investigation (see Discussion).

DISCUSSION

We have provided a novel maximum-likelihood measure, GLR, for the rate of gene loss. GLR is based on a probabilistic model that takes into account the phylogenetic tree of the species under study, its branch lengths, and a phyletic pattern

representing the presence and absence of the gene in those species. The measure is shown to be highly correlated with experimental data on gene lethality and connectivity levels, and more aligned with biological data compared to the state-of-the-art PGL.

Previously, more complex probabilistic models of gene content evolution that account for events of gene loss, duplication and transfer were proposed (10–13). The estimation of multiple parameters in such complex models requires extensive phylogenetic data and cannot be performed individually per gene based solely on its content data. To obtain reliable rate estimations, Gu *et al*. (11) assumed uniform duplication and loss parameters for all genes under all tree branches, and hence used the entire ensemble of gene content data to estimate these two parameters. The work of Hahn *et al*. (12) assumed uniformity of duplication and loss parameters across all genes but allowed variability within branches. In the recent work of Csuros and Miklos (10), duplication, loss and transfer rates were estimated for large groups of genes that have similar evolutionary dynamics. However, while the above works focus on global parameters characterizing gene evolution, the focus of this work is the variability in evolutionary dynamics between different genes. Estimating evolutionary parameters associated with a single gene based solely on its observed phyletic pattern is a challenging optimization task. Considering a single phyletic pattern at a time prevents the usage of a complex, multi-parameter model. Consequently, we employ a single-parameter model that characterizes simple gene loss dynamics [see also (29,30)] and apply it to those cases that can be plausibly described by such dynamics. Specifically, we focus on Eukaryotic species in which horizontal gene transfer is unlikely, and on genes whose genomic copy number is one. Our analysis suggests that the model accurately predicts gene-specific loss rates for these genes. Moreover, resorting to a simple model of gene loss evolution, our maximum-likelihood algorithm allows for an optimal loss rate estimation. Notably, the works mentioned above do not guarantee the identification of optimal parameters.

Evaluating the correctness of a gene loss measure by correlating it with various estimates of gene importance is a challenging task, owing to the inherent difficulty of defining and measuring gene importance. Gene essentiality is a common and natural measure of importance, and as we have found, it has indeed a markedly high correlation with the calculated loss rates in yeast. However, one cannot expect GLR to be an ideal predictor of gene lethality as gene dispensability is only one out of many factors determining gene evolution dynamics (31). Other measures, previously used to estimate gene importance, are based on various characteristics of the gene functionality. Specifically, the involvement of a protein in various cellular processes is assumed to reflect its functional importance, motivating a heuristic measure based on its degree of connectivity in a PPI network. Membership in the same protein complex also indicates a shared functional role, and, hence, a similar functional importance measure. Such characteristics, however, are only indirect measures of importance. This is also evident by the weaker correlation we have found between gene loss and protein interaction information (Table 1). In the case of PPI networks, the noisy nature of the data (24,32) may further weaken the signal of importance, although this is handled to some extent by taking into account the reliabilities of the interactions.

Although the underlying motivation of GLR is similar to that of PGL, and the two measures are largely in agreement (as exemplified by the high correlation between them), GLR was shown to provide higher correlations with biological measures of gene functional importance and an improved lethal/nonlethal classification (see Figure 3). We believe that the demonstrated superiority of GLR stems from its enhanced probabilistic model, better capturing the gene loss process.

Furthermore, as GLR is based on a probabilistic model it is extensible to utilizing additional data sources in the rate computation. Specifically, additional data on the confidence level of gene presence or absence in homology clusters can be easily incorporated into the model, addressing noise problems in the sequence data. Such data can be obtained for example from the significance scores of BLAST sequence alignment, which is commonly used for gene homology detection (33). Additional data on the confidence level of branch length estimations can also be considered. We believe that such confidence measures can further improve the accuracy of GLR, as branch length estimations are obtained through various computational and experimental methods and are known to be noisy (21).

The GLR model presented here assumes (as does PGL) that the gene was present in the root of the tree and may have been lost multiple times in different branches. To this end, we have focused on genes that are likely to be present in the root of the phylogenetic tree (i.e. present in both the Fungi/Metazoa group and plants, similarly to PGL). In general, for genes whose lowest common ancestor does not lie at the root of the tree, the uncertainty about its location can be incorporated into the probabilistic model underlying GLR.

In our application, GLR was computed based on all possible ancestral phyletic patterns, representing the presence and absence of genes in internal nodes of the tree. Alternatively, GLR can be computed along with a single, most likely ancestral phyletic pattern by employing the ancestral maximum-likelihood approach (34). This can be accomplished by modifying the gradient ascent search to find the loss rate that maximizes the probability of optimal ancestral phyletic pattern using the ancestral maximum-likelihood algorithm.

GLR and its underlying probabilistic model can also be generalized to support different applications. For example, GLR may be utilized to improve current estimations of branch lengths or to estimate branch specific loss rates. This could be done by simultaneously searching for branch specific parameters and GLR values that maximize the probability of a given set of phyletic patterns (for several genes). Another application, providing a maximum-likelihood estimation of the expected number of gene losses per branch (including internal branches) while considering all possible ancestral phyletic patterns was described above.

Determining the importance of genes and, specifically, its effect on genome evolution is a fundamental problem in biology. As genomic data continues to accumulate we expect GLR to capture the biological significance of genes more and more accurately, facilitating an improved analysis of gene functionality and evolution.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Alon Keinan for critical reading of the manuscript. We thank Yuri I. Wolf from the Koonin Evolutionary Genomics Research Group for providing us with the original PGL calculated values. E.B. is supported by the Yeshaya Horowitz Association through the Center for Complexity Science. T.S. is supported by the Tauber Fund. E.R.'s research is supported by the Tauber Fund, the Center for Complexity Science, and the Israeli Science Foundation (ISF). R.S. was supported by an Eshkol Fellowship and by a Sackler Career Development Chair. Funding to pay the Open Access publication charges for this article was provided by School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University.

Conflict of interest statement. None declared.

REFERENCES

1. Hirsh,A.E. and Fraser,H.B. (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.
2. Jordan,K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*., **12**, 962–968.
3. Krylov,D.M., Wolf,Y.I., Rogozin,I.B. and Koonin,E.V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level and interactivity are correlated in eukaryotic evolution. *Genome Res*., **13**, 2229–2235.
4. Wall,D.P., Hirsh,A.E., Fraser,H.B., Kumm,J., Giaever,G., Eisen,M.B. and Feldman,M.W. (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA*, **102**, 5483–5488.
5. Yang,J., Gu,Z. and Li,W.H. (2003) Rate of protein evolution versus fitness effect of gene deletion. *Mol. Biol. Evol*., **20**, 772–774.
6. Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C. and Feldman,M.W. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
7. Jordan,I., Wolf,Y. and Koonin,E. (2003) No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol*., **3**.
8. Wilson,A.C., Carlson,S.S. and White,T.J. (1977) Biochemical evolution. *Annu. Rev. Biochem*., **46**, 573–639.
9. Farris,J.S. (1977) Phylogenetic analysis under dollo's law. *Syst. Zool*., **26**, 77–88.
10. Csuros,M. and Miklos,I. (2006) A probabilistic model for gene content evolution with duplication, loss and horizontal transfer. In *Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. Springer, Berlin, pp. 206–220.
11. Gu,X. and Zhang,H. (2004) Genome phylogeny inference based on gene contents. *Mol. Biol. Evol*., **21**, 1401–1408.
12. Hahn,M., De Bie,T., Stajich,J., Nguyen,C. and Cristianini,N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*., **15**, 1153–1160.
13. Tiuryn,J., Rudnicki,R. and Wojtowicz,D. (2004) A case study of genome evolution: from continuous to discrete time model. *Proc. Math. Found. Comput. Sci*., **3153**, 1–24.
14. Wolf,Y., Carmel,L. and Koonin,E. (2006) Unifying measures of gene function and evolution. *Proc. Biol. Sci*., **273**, 1507–1515.
15. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
16. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*., **17**, 368–376.
17. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **34**, D173–D180.
18. Beltrao,P. and Serrano,L. (2005) Comparative genomics and disorder prediction identify biologically relevant sh3 protein interactions. *PLoS Comput. Biol*., **1**, e26.
19. Gaunt,M.W. and Miles,M.A. (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol. Biol. Evol*., **19**, 748–761.
20. Hamer,L., Pan,H., Adachi,K., Orbach,M.J., Page,A., Ramamurthy,L. and Woessner,J.P. (2001) Regions of microsynteny in *Magnaporthe grisea* and *Neurospora crassa*. *Fungal Genet. Biol*., **33**, 137–143.
21. Hedges,S.B., Blair,J.E., Venturi,M.L. and Shoe,J.L. (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol*., **4**, 2.
22. Guldener,U., Munsterkotter,M., Kastenmuller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al*. (2005) Cygd: the comprehensive yeast genome database. *Nucleic Acids Res*., **33**, D364–D368.
23. Sharan,R., Suthram,S., Kelley,R.M., Kuhn,T., McCuine,S., Uetz,P., Sittler,T., Karp,R.M. and Ideker,T. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.
24. Suthram,S., Shlomi,T., Ruppin,E., Sharan,R. and Idekrer,T. (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**.
25. Mewes,H., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) Mips: a database for genomes and protein sequences. *Nucleic Acids Res*., **27**, 44–48.
26. Dudley,A., Janse,D., Tanay,A., Shamir,R. and Church,G. (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol*. doi:10.1038/msb4100004.
27. Deng,M., Sun,F. and Chen,T. (2003) Assessment of the reliability of protein–protein interactions and protein function prediction. *Pac. Symp. Biocomput*., 140–151.
28. Pellegrini,M., Marcotte,E., Thompson,M., Eisenberg,D. and Yeates,T. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
29. Jordan,I., Makarova,K., Spouge,J., Wolf,Y. and Koonin,E. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res*., **11**, 555–565.
30. Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res*., **10**, 808–818.
31. Pal,C., Papp,B. and Lercher,M. (2006) An integrated view of protein evolution. *Nat. Rev. Genet*., **7**, 337–348.
32. Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res*., **12**, 37–46.
33. Remm,M., Storm,C. and Sonnhammer,E. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*., **314**, 1041–1052.
34. Pupko,T., Pe'er,I., Shamir,R. and Graur,D. (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol*., **17**, 890–896.