Assessing the Determinants of Evolutionary Rates in the Presence of Noise

Joshua B. Plotkin and Hunter B. Fraser†*

*Department of Biology, The University of Pennsylvania; and †The Broad Institute of Harvard University and MIT

Although protein sequences are known to evolve at vastly different rates, little is known about what determines their rate of evolution. However, a recent study using principal component regression (PCR) has concluded that evolutionary rates in yeast are primarily governed by a single determinant related to translation frequency. Here, we demonstrate that noise in biological data can confound PCRs, leading to spurious conclusions. When equalizing noise levels across 7 predictor variables used in previous studies, we find no evidence that protein evolution is dominated by a single determinant. Our results indicate that a variety of factors—including expression level, gene dispensability, and protein–protein interactions—may independently affect evolutionary rates in yeast. More accurate measurements or more sophisticated statistical techniques will be required to determine which one, if any, of these factors dominates protein evolution.

Introduction

Proteins span more than 3 orders of magnitude in their evolutionary rate, as quantified by the number of nonsynonymous substitutions per site. What determines a protein's rate of evolution has been actively debated over the past several decades. Early hypotheses suggested that the evolutionary rate of a protein is governed by at least 2 factors: the protein's level of functional constraint (i.e., the density of functionally active residues) and its overall importance (or "dispensability") to the organism (Wilson et al. 1977).

Arguments for the role of dispensability are mostly based on theory (Ohta 1973), as very little empirical data have been available until recently (Hurst and Smith 1999; Hirsh and Fraser 2001). A relationship between dispensability and evolutionary rate is hypothesized because mutations in essential proteins are more likely to be deleterious. Such mutations are purged from the population, thereby reducing the evolutionary rate of indispensable proteins (Ohta 1973).

In contrast to dispensability, the relationship between functional constraint and evolutionary rate has been studied empirically for over 40 years. Among the first to address the role of functional constraint was Ingram (1961), who observed that the polypeptide chains of hemoglobin should be differentially constrained depending on the number of other chains with which they physically interact (Ingram 1961). Similarly, when studying cytochrome c, Dickerson (1971) observed that surface residues that interact with other proteins tend to be highly conserved. Although functional constraints are difficult to measure directly, or even define precisely, many studies have used as a proxy the number of physical interactions in which a given protein participates. Authors have recently confirmed early hypotheses of Ingram and Dickerson on a much larger scale using curated sets of interacting protein crystal structures (Mintseris and Weng 2005).

The sudden plethora of sequenced genomes allows us to compare orthologous coding sequences from related species, estimate evolutionary rates, and ask what features of proteins covary with their evolutionary rates. The yeast *Saccharomyces cerevisiae* has emerged as a model system for systematically studying the determinants of evolutionary rates. *Saccharomyces cerevisiae* was the first fully sequenced eukaryote, and its genome remains the most comprehensively annotated. Additionally, *S. cerevisiae* has been the subject of thousands of functional genomic experiments producing diverse information for evolutionary investigation.

Previous studies have reported a variety of functional, biophysical, and fitness-related variables that correlate with the evolutionary rates of proteins (Drummond et al. 2006): proteins evolve more slowly if they have a higher number of mRNA molecules per cell (expression) (Green et al. 1993; Pal et al. 2001), if they have a higher number of protein molecules per cell (abundance) (Drummond et al. 2006), a higher codon adaptation index (CAI) (Pal et al. 2001; Wall et al. 2005), more protein–protein interactions (degree) (Fraser et al. 2002), a larger fitness effect upon gene knockout (dispensability) (Hirsh and Fraser 2001), shorter sequence length (Marais and Duret 2001), or a more central role in the interaction network (centrality) (Hahn and Kern 2005). But these predictor variables are themselves correlated with one another—raising the question of which variables are truly involved in determining evolutionary rates and which variables happen to covary simply because they are influenced by another, causal variable. It has been argued, for example, that the correlation between dispensability and evolutionary rate is simply a side effect of causal relationships between expression level and evolutionary rate and between expression level and dispensability (Pal et al. 2003).

Drummond et al. (2006) undertook a comprehensive analysis of the determinants of protein evolution in yeast. Their work represents a significant advance towards identifying the major, independent correlates of evolutionary rates (McInerney 2006; Pal et al. 2006; Rocha 2006). Prior to the work of Drummond et al. (2006), many authors had used the techniques of multiple regression and partial correlation to assess whether correlates of evolutionary rate are independent of one another (Fraser et al. 2002; Bloom and Adami 2003; Rocha and Danchin 2004). Drummond et al. (2006) demonstrated that colinearity of predictor variables and measurement noise can cause partial correlations to yield spuriously significant results. In lieu of partial correlations, and in order to remove colinearity, Drummond et al. (2006) used a principal component regression (PCR) to analyze evolutionary rates. Surprisingly, they found that a single component—comprised almost entirely of expression level, abundance, and CAI—explained far more variation

Key words: evolutionary rates, noise, pca, PCR, expression levels, dN, dS.

E-mail: jplotkin@sas.upenn.edu.

Mol. Biol. Evol. 24(5):1113–1121. 2007
doi:10.1093/molbev/msm044
Advance Access publication March 7, 2007

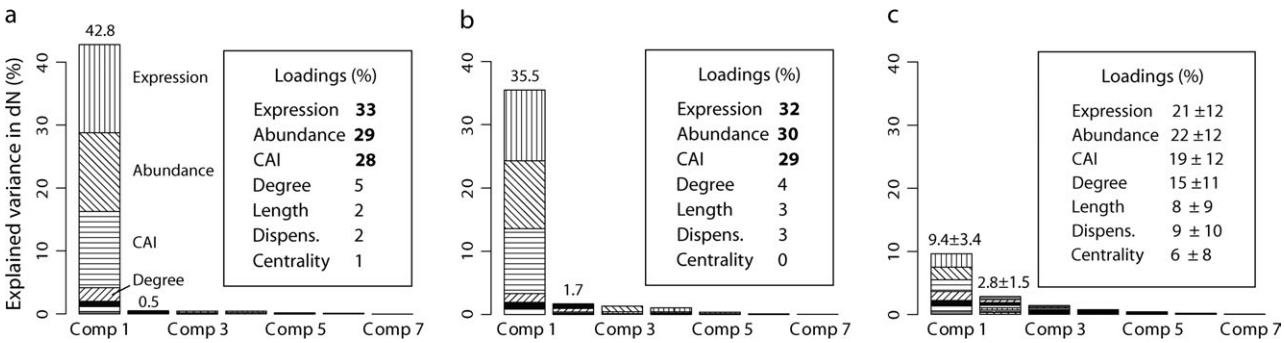


FIG. 1.—PCRs of evolutionary rate against 7 explanatory variables. Components are ordered according to the amount of variation they explain. (*a*) A regression on 568 genes previously reported by Drummond et al. (2006) shows a dominant component—consisting primarily of expression, abundance, and CAI—that explains 42.8% of the variation in dN. (*b*) A regression over 242 genes using a different data set of protein–protein interactions (Ho et al. 2002) instead of Han et al. (2004) also shows a dominant explanatory component. (*c*) A regression on the same data as in (*b*) but with noise levels equalized across predictor variables. Each panel also shows the loadings of predictor variables on the first principal component (±one standard deviation in panel *c*). When noise levels are equalized, there is no evidence of a dominant explanatory component, and all predictor variables explain similar total amounts of variation in dN. Furthermore, when noise levels are equalized, the first component contains roughly equal contributions from each predictor variable.

in evolutionary rates than any other component. On the basis of these results, Drummond et al. proposed that a single determinant, namely, selection against translation-error–induced protein misfolding, dominates protein evolution in yeast (Drummond et al. 2005, 2006). This recent finding has already changed how researchers think about protein evolution (Koonin and Wolf 2006; McInerney 2006; Pal et al. 2006; Rocha 2006).

Here, we reinspect the analyses of evolutionary rates presented by Drummond et al. (2006). We show that the PCR utilized by Drummond et al. can be confounded when predictor variables have been measured with different amounts of noise. We assess the amount of noise associated with each of the 7 predictor variables used in previous studies of yeast protein evolution. We show that after equalizing noise levels across the predictor variables, each predictor has a roughly equal contribution to the evolutionary rate of proteins, and there is no evidence for a single, dominant factor driving protein evolution. Finally, we present a simple mathematical model of evolutionary rates, with parameter values determined by empirical yeast data. Our model demonstrates that the apparent predominance of translational selection as the determinant of yeast protein evolution may be a spurious artifact arising from the variable accuracy of functional genomic measurements.

Results

We have reanalyzed the data sets studied by Drummond et al. (2006), consisting of evolutionary rates (dN) for 3,036 yeast protein sequences (Hirsh et al. 2005) and 7 variables that correlate with evolutionary rates: expression, abundance, CAI, length, dispensability, degree, and centrality.

Given the large number of variables correlated with evolutionary rate, we wish to know which correlations are independent of one another. For example, let R, E, and D denote evolutionary rate, mRNA expression level, and degree of protein–protein interactions, respectively. If these variables are noise free, then the partial correlation coefficient $r(D, R \mid E)$ describes the relationship between degree and rate, controlling for expression. In practice, however, we have only noisy estimates of these variables, denoted E^*, and D^*. Noise in expression data, for example, arises both from variability in expression levels between cells and from inaccuracies in measurement. We refer to both sources of variability as noise.

Drummond et al. (2006) have shown that the method of partial correlations can yield spuriously significant results when applied to noisy variables—that is, $r(D^*, R \mid E^*)$ may show a significant departure from zero even when $r(D, R \mid E)$ equals zero. In other words, as a result of noise, the relationship between protein dispensability and evolutionary rate may appear to be independent of expression levels, even when the underlying, noiseless variables are uncorrelated when controlling for expression. In order to avoid the pitfalls of partial correlations, which are caused by noise and colinearity, Drummond et al. regressed evolutionary rates against the principal components of the 7 predictor variables. The principal component regression corrects for colinearity among predictor variables, but the PCR implicitly assumes that all predictors have been measured with the same amount of noise. In the following sections, we show that the assumption of equal noise is not valid for the data sets analyzed by Drummond et al., and we demonstrate that the apparent dominant determinant of protein evolution may be an artifact of this invalid assumption.

Quantifying Noise in Functional Genomic Data

Some measurable features of genes are virtually free of noise, such as gene length and CAI. But other variables contain significant noise, such as mRNA expression levels, protein abundance, the fitness effects of knockouts, and the number of protein–protein interactions. How can we quantify and estimate the amount of noise associated with each of these genome-wide measurements?

The most straightforward way to quantify noise is to calculate the correlation coefficient between 2 (or more) independent measurements of the same quantity. For example, mRNA expression levels in yeast are quite reproducible: the correlation between 2 independent measurements of mRNA expression levels, using the same oligonucleotide

arrays, has been reported as $r_{expr} = 0.72$ ($n = 5, 555$) (Drummond et al. 2005). We compared expression levels measured by different investigators 3 years apart (Holstege et al. 1998; Causton et al. 2001) and found an even stronger correlation: $r = 0.90$ ($n = 5, 460$). Nevertheless, we will use the value $r_{expr} = 0.72$ so that our analysis is conservative and our estimate of expression noise agrees with that of Drummond et al. (2005).

Protein abundance data are apparently less noisy than mRNA abundance data. In the only systematic study reporting abundances for a large number of yeast proteins, Ghaemmaghami et al. (2003) measured 206 proteins in triplicate. Assuming the noise to be normally distributed, the average correlation in abundance between any 2 sets of 206 replicates is $r = 0.98$. This estimate of noise for protein abundances is not as conservative as for mRNA abundances above—because these replicates were performed using the same method by the same investigators. But this estimate nevertheless suggests that protein abundance measurements in yeast contain relatively little noise. In order to be conservative (see Discussion), we will assume that protein abundances contain approximately the same amount of noise as mRNA levels, $r_{abund} = 0.72$, even though this assumption likely overestimates the noise in protein abundances.

Gene dispensability data in yeast are more noisy than protein abundances or expression levels. The correlation coefficient between 2 independent measurements of the fitness effects of knockouts (Warringer et al. 2003; Deutschbauer et al. 2005) made using the same set of viable single-gene deletion strains is $r_{disp} = 0.56$ ($n = 4, 156$). In order to facilitate comparison with Drummond et al. (2006), we analyze the same data set of gene dispensabilities (Deutschbauer et al. 2005), despite the fact that the other data set correlates more strongly with evolutionary rates (Wall et al. 2005).

The number, or degree, of protein–protein interactions is by far the most noisy variable analyzed in this study and previous related studies. Bloom and Adami (2003) recently summarized the results from 9 protein interaction data sets. Performing all 36 possible pairwise comparisons of these data sets, we find that as many pairs exhibit negative correlations as exhibit positive correlations. Despite the discouraging discordance among the 9 protein interaction data sets, it is well established that some data sets contain less noise than others (Kemmeren et al. 2002; von Mering et al. 2002). Protein interaction data sets assembled from low-throughput studies are difficult to use in this context because of the unknown but certainly extreme bias in which proteins have been studied by individual investigators (Reguly et al. 2006). For this reason, we chose to compare 2 of the highest quality high-throughput data sets that were generated by a single method: mass spectrometry (Gavin et al. 2002; Ho et al. 2002). Any systematic biases inherent in the mass spectrometry method will artificially inflate the correlation between these 2 measurements, leading to a conservative underestimate of noise. The observed agreement between the 2 measurements of protein interaction degree is $r_{degree} = 0.11$ ($n = 524$). Whereas this correlation is much greater than the median correlation among all 36 pairs of interaction data sets ($r = 0.002$), the reproducibility of the protein interaction data is much lower than for all other types of data in this study.

Protein interactions measured by mass spectrometry (Ho et al. 2002) are very similar to the composite data set of interactions analyzed by Drummond et al. (2006). Using the data from Ho et al. in place of the interaction data analyzed by Drummond et al. does not significantly alter the PCR of dN (compare fig. 1*a* and *b*, below). Therefore, because we can estimate the noise in the mass spectrometry data, but not in the composite data set used by Drummond et al., we will use the spectrometry data for all subsequent analyses. (Results are unchanged if we use the composite data set of interactions in our analyses instead of the mass spectrometry data.)

Lastly, we assume that a protein's centrality in the interaction network—a quantity based entirely on protein–protein interaction data—has the same noise level as the underlying interaction data. Relaxing this assumption by raising or lowering the amount of noise in the centrality data does not significantly affect our results.

Equalizing Noise across Predictor Variables

As we have shown above, the 7 correlates of evolutionary rate analyzed by Drummond et al. (2006) contain widely different amounts of noise. What conclusions, then, can we draw from PCRs, given that the PCR method assumes equal noise across all predictors? One way to answer this question is by artificially equalizing the noise levels across the predictor variables and repeating the PCR analysis. (A second way to answer the question is presented in a subsequent section.) We focus on the PCR because this is the technique employed by Drummond et al. (2006) to reach their conclusions about protein evolution in yeast.

The degree of protein interactions is by far the most noisy of the 7 predictor variables in our study. In order to match the level of noise in interaction degree, we can add an appropriate amount of extra noise to each of the other 6 predictors. As described in Appendix, we can solve analytically for the appropriate amount of Gaussian noise to be added to each predictor so that the resulting variables have the same amount of noise as degree, namely, $r_{degree} = 0.11$.

Figure 1 shows PCR analyses of the original predictor variables alongside analyses of modified predictors whose noise levels have been equalized. There are 3 features of each regression that are important to note: 1) the amount of variation in evolutionary rate explained by the dominant component; 2) the amount of variation explained by the secondmost dominant component; and 3) the loadings of predictor variables on components.

As figure 1 shows, variable noise among predictors dramatically affects the PCR analysis of evolutionary rates. Without correcting for variable noise, the PCR identifies a single component that explains at least 20-fold more variance in evolutionary rate than any other component (Drummond et al. 2006); whereas after equalizing noise levels, the dominant component does not explain significantly more variance than the subdominant component (i.e., not more than 2 standard deviations). In other words, after correcting for noise levels, there is no evidence of a single, dominating determinant of evolutionary rates in yeast.

Variable noise levels affect the PCR analysis in several other, important ways. Without correcting for noise, the dominant explanatory component consists almost exclusively of

translation-related variables: mRNA expression, protein abundance, and CAI. The translation-related variables each explain more than 4 times the total variation in evolutionary rate than any of the other predictor variables. These results have been interpreted as conclusive evidence that translational selection governs the rate of protein evolution (Drummond et al. 2006). By contrast, after equalizing noise levels, the dominant explanatory component contains roughly equal loadings from all 7 predictor variables (fig. 1*c*), and each of the predictor variables explains roughly the same amount of total variation in evolutionary rate (within one standard deviation). In other words, after correcting for noise levels, there is no evidence of a dominant, translation-related determinant of evolutionary rates.

A Simple Model of Evolutionary Rates

As with other techniques for multiple regression, the PCR method assumes equal noise levels across predictors, and it is sensitive to violations of this assumption. As seen above, when we equalize noise levels among predictor variables, the resulting PCR paints a very different picture of yeast protein evolution than has been previously reported (Drummond et al. 2006). These results still beg the following question: given the known amount of noise associated with each predictor variable, how much variation in evolutionary rate would be explained by the underlying, noiseless predictors?

In this section, we provide one possible answer to this question by using a simple mathematical model. We present a phenomenological model of evolutionary rates, and we choose parameters consistent with most important features of the observed yeast data. The purpose of this model is not to recapitulate every detail of the empirical data but rather to explore what underlying patterns are consistent with the salient features of the observed, noisy data.

For the sake of simplicity, we focus on the 4 variables that explain the most rate variation: expression (E), abundance (A), CAI (C), and protein–protein interaction degree (D). We demonstrate that the observed, noisy data are consistent with a model in which there are multiple independent determinates of evolutionary rates and in which the underlying (noiseless) protein interactions explain more variance in evolutionary rate than expression level, abundance, or CAI.

We specify our model so as to reflect several important biological features of protein evolution (Drummond et al. 2006): 1) mRNA expression, protein abundance, and CAI all covary because they all reflect, in part, the amount of translation events experienced by a gene; 2) expression, abundance, and CAI also covary with the degree of protein–protein interactions, for reasons unrelated to translation; 3) the amount of translation and the degree of protein interactions both influence the evolutionary rate. These features lead to the following model equations:

$$\begin{aligned} E &= \alpha Z_1 + \beta Z_2 + Z_3, \\ A &= \alpha Z_1 + \beta Z_2 + Z_4, \\ C &= \alpha Z_1 + \beta Z_2 + Z_5, \\ D &= \alpha Z_1 + Z_6, \\ R &= Z_2 + Z_6, \end{aligned}$$

where each Z_i is an independent, normally distributed random variable with mean zero and variance one. The term Z_1 represents a source of covariation shared by expression, abundance, CAI, and degree. The term Z_2 represents the amount of translation, which contributes to covariation in E, A, and C. The terms Z_3 through Z_6 represent sources of variation present in each predictor variable but not shared between them. The parameters α and β quantify the relative importance of translation versus other sources of variation in the predictor variables. In our model, the evolutionary rate (R) is determined by the amount of translation (Z_2) and by the variation in protein interactions unrelated to other variables (Z_6). (Similar results are obtained under related models, such as $R = \alpha Z_1 + \beta Z_2$; see also Supplementary Materials online.)

In addition to the underlying model, we also specify equations that describe noisy versions of the predictor variables representing measurements of expression (E^*), measurements of abundance (A^*), measurements of CAI (C^*), and measurements of interaction degree (D^*):

$$\begin{aligned} E^* &= E + n_E W_1, \\ A^* &= A + n_E W_2, \\ C^* &= C, \\ D^* &= D + n_D W_5, \end{aligned}$$

where each W_i is an independent, normally distributed random variable with mean zero and variance one. The parameter n_E determines the amount of noise in the measured expression data, E^*. We conservatively assume that the amount of noise in protein abundance measurements equals the amount of noise in mRNA abundance measurements (even though our calculations above suggest that abundance has less noise). The parameter n_D determines the amount of noise in measured interaction degree data. Note that CAI is measured without noise (i.e., $C^* = C$).

We choose the 4 parameters of our model so as to match the most important empirical features of the yeast data: 1) the noise in mRNA expression data, $r_{\text{expr}} = 0.72$; 2) the noise in protein interaction data, $r_{\text{deg}} = 0.11$; 3) the correlation between measured expression levels and evolutionary rate, $r(E^*, R) = 0.56$ ($n = 2,840$); and 4) the correlation between measured interaction degree and evolutionary rate, $r(D^*, R) = 0.23$ ($n = 692$). Using a straightforward parameterization technique (see Appendix), we find parameters so as to match all 4 of these observed features in the yeast data.

It is instructive to compare a PCR analysis of the underlying variables in our model against a PCR analysis of the noisy variables, which represent measurable quantities. When applied to the noisy variables (fig. 2*a*), the PCR indicates a dominant explanatory component consisting almost entirely of the translation-related variables—expression (E^*), abundance (A^*), and CAI (C^*). Each of these variables explains significantly more variation in evolutionary rate than protein interactions, which appear in a secondary minor component; this situation is analogous to that seen in the real data (fig. 1*a* and *b*). By contrast, when applied to the underlying noiseless variables, the PCR reveals a dramatically

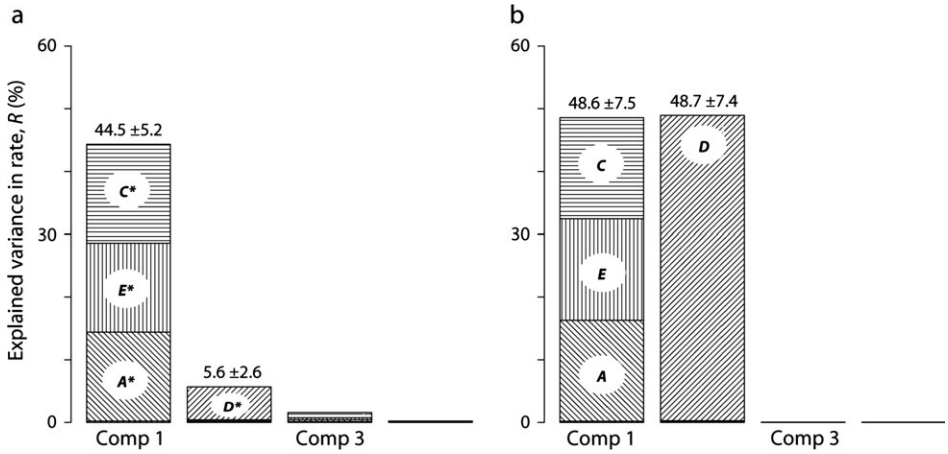


FIG. 2.—A PCR analysis of the noisy (*a*) and underlying (*b*) variables in a simple model of evolutionary rates. When applied to the noisy variables, which represent measurable quantities, the regression indicates a single, dominant determinant of evolutionary rates, and the degree of protein–protein interactions appears to explain significantly less variation in rate than CAI, abundance, or expression. But when applied to the underlying noiseless variables, the PCR analysis reveals the opposite conclusion: no single component dominates evolutionary rates, and protein interactions explain significantly more variation than CAI, abundance, or expression.

different picture (fig. 2*b*): no single component dominates evolutionary rates. Moreover, the true degree of protein interactions explains significantly more variation in rate than expression, abundance, or CAI.

The simple model developed here is consistent with the observed noise levels in yeast genomic data and with the observed correlations with evolutionary rates. Under this model, protein interaction degree explains significantly more variation in evolutionary rates than expression level, abundance, or CAI. Nevertheless, if one were to analyze the noisy versions of predictor variables (disregarding the fact that such a PCR analysis violates the assumption of equal noise levels), one would reach the opposite conclusion. Thus, our model highlights the danger of using a principle component regression to analyze noisy biological data without accounting for known variation in noise levels across predictor variables.

Other Evidence on the Determinants of Evolutionary Rates

The preceding sections demonstrate that the PCR is not robust to violating the assumption of equal noise levels across predictor variables. This weakness is not unique to the PCR, but it is likely shared by most other multiple regression techniques. As a result of this difficulty, however, given the known variation in noise levels across predictors, there is no evidence at present for a single determinant of evolutionary rates in yeast.

In this section, we demonstrate another related line of evidence against our ability to deduce a single determinant of evolutionary rates: namely, the PCR depends strongly on which predictors are included in the regression.

A gene's expression, abundance, and CAI are all related to the total amount of translation events it experiences (Drummond et al. 2006). Therefore, Drummond et al. interpret the dominant explanatory component in their regression—comprised equal parts expression, abundance, and CAI—as the amount of translation, and they conclude that selection against translation-error–induced protein misfolding is the predominant determinant of protein evolution in yeast (Drummond et al. 2005, 2006). If these conclusions were robust, a PCR analysis of the same data excluding CAI, for example, should yield a very similar result—except that the resulting dominant component would be comprised abundance and expression.

As seen in figure 3, a regression excluding CAI paints a very different picture of protein evolution than expected under the translational-selection hypothesis. According to this regression, there is no evidence that translational processes dominate protein evolution. Instead, multiple independent components explain significant variation in evolutionary rates. Moreover, the degree of protein interactions explains more variance in evolutionary rate than protein abundances, and it is more strongly represented in the first component. Regressions excluding mRNA expression, protein abundance, or combinations of these variables yield very similar results. None of these results would be observed if the PCR method were robust and if translational processes dominated protein evolution.

Finally, we note that other techniques for analyzing collinear predictors, such as the sliced inverse regression (Duan and Li 1991), also fail to implicate a single, predominant determinant of evolutionary rates (not shown).

Discussion

What independent factors influence the rate of protein evolution remains an outstanding question. The work of Drummond et al. (2006) is critically important because it demonstrates that partial correlations can yield spurious results, due to noise in a predictor variable. By the same token, we have demonstrated that multiple regressions can yield spurious results due to different levels of noise in predictor variables. At present, a conservative application of the PCR method requires that we equalize noise levels across predictors—in which case there is no evidence that the rate of protein evolution in yeast is dominated by a single determinant.

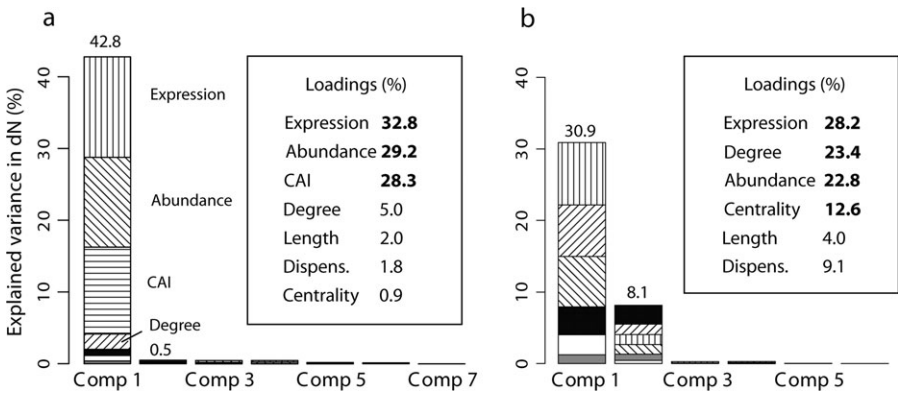


FIG. 3.—PCRs of evolutionary rates. (*a*) A regression of dN against 7 predictor variables, previously reported by Drummond et al. (2006). (*b*) A regression of the same data as in (*a*), omitting CAI as a predictor variable. When CAI is removed, there is no evidence of a dominant explanatory component related to translational selection. In fact, when CAI is omitted, the degree of protein–protein interactions is more strongly weighted on the first component, and it explains more total variance in evolutionary rate than protein abundance. (Black corresponds to centrality; white to dispensability; and grey to length.)

We emphasize that the limitations of the PCR in the face of predictor variables with different degrees of noise are not unique to the PCR method. The same limitations apply to virtually all methods of multiple regression, which typically assume equal noise levels across predictors. We have chosen to focus on the PCR only because this was the method used by Drummond et al. (2006) to reach their conclusions about protein evolution in yeast.

Our analysis has formally demonstrated a concept that is intuitively clear: standard regression techniques cannot meaningfully compare the explanatory power of predictors when the predictors contain different amounts of measurement noise. It is less clear how best to deal with this difficulty, which comes hand in hand with diverse genomic data. Ideally, information about the known noise levels of predictors should be incorporated into statistical methods when partitioning phenotypic variance into independent contributions. Unfortunately, we know of no method that performs such a partitioning while accounting for variable noise levels. The simplistic approach used here for equalizing noise levels may seem Draconian because it throws out some signal contained in the less noisy predictors. Nevertheless, this conservative approach is necessary until more sophisticated statistical methods, which account for variable noise levels, are developed. Indeed, the inadequacy of the PCR in this context is further demonstrated by the fact that its results are not robust to removing predictor variables from the regression.

Our procedure for equalizing noise levels is conservative with respect to our conclusions because we have used one of the lowest estimates of noise in protein interaction data and some of the highest estimates of noise in expression and abundance data. In particular, we have estimated expression noise using the same correlation coefficient reported by Drummond et al. (2005), and our estimate of noise in protein interactions is significantly smaller than the median estimate across 36 pairwise comparisons.

Aside from adding noise to empirical data, we have also presented a simple model of evolutionary rates that is consistent with the observed yeast data. The purpose of this model is not to recapitulate all details of the empirical data but rather to explore what underlying patterns are consistent with the salient features of the observed, noisy data. According to this model, the underlying, noiseless variable for protein interaction degree explains more variance in evolutionary rate than all other variables—despite the fact that a PCR analysis on the noisy, measured variables would yield the opposite conclusion. We emphasize that this model does not establish that protein–protein interactions actually have a stronger influence on yeast evolutionary rates than expression, abundance, or CAI. Rather, the model simply demonstrates that we cannot yet rule out the possibility of an important, independent role for protein–protein interactions (or other noisy variables) in determining evolutionary rates.

We emphasize that our analysis does not rule out the possibility of a single determinant for evolutionary rates. Indeed, in the future, we may conclude definitively that translational processes explain more variance in evolutionary rates than any other feature of yeast proteins. At present, however, given the variable amounts of noise associated with existing genome-wide measurements, PCRs do not provide sufficient evidence to reach this conclusion. It appears that more accurate measurements, or more sophisticated statistical techniques, will be required to tease apart the underlying determinants of protein evolution.

Appendix

Data Sets

All data sets were taken directly from Drummond et al. (2006), with the exception of the protein interaction data. The protein interaction data were measured by mass spectrometry (Ho et al. 2002), and they were compared with another mass spectrometry data set (Gavin et al. 2002) in order to estimate the amount of noise. Our results remain unchanged if we use the interaction data from Drummond et al. instead of the mass spectrometry data. An additional data set (Warringer et al. 2003) was used to estimate the noise in gene dispensability.

Equalizing Noise Levels across Data Sets

Before applying regressions, all variables were log transformed (except dispensability), centered, and variance normalized, as in Drummond et al. (2006). Our results remain essentially unchanged using rank regressions instead of parametric regressions.

Of the data sets used in this study, the protein–protein interaction data are the least precise. In order to equalize noise across predictor variables, we add an appropriate amount of noise to each transformed variable and then rescale each variable by its variance. In order to match the noise level in protein interaction data, $r_{\text{degree}} = 0.11$, we must add enough extra noise to each other variable so that, if we were to add the noise twice independently (so as to imitate 2 independent measurements, each with their own source of noise), the resulting correlation would equal r_{degree}.

To be more explicit, consider a predictor variable E with original noise level, $r_E < r_{\text{degree}}$, given by the correlation between 2 independent measurements: $r_E = r(E_1^*, E_2^*)$. We must find the value of λ such that

$$r\left[\ \text{scale}(E_1^*) + \lambda Z_1,\ \text{scale}(E_2^*) + \lambda Z_2\right] = r_{\text{degree}},$$

where each Z_i is an independent Gaussian random variable with mean zero and variance one and $\text{scale}(E) = (E - \mu_E)/\sigma_E$. Letting $A = \text{scale}(E_1^*) + \lambda Z_1$ and $B = \text{scale}(E_2^*) + \lambda Z_2$ and substituting into the expression for the correlation coefficient, we must solve

$$\frac{\mathbb{E}(AB) - \mathbb{E}(A)\mathbb{E}(B)}{\sigma_A \cdot \sigma_B} = r_{\text{degree}}.$$

Because A and B both have expected value zero, and each Z_i is uncorrelated with E_1^* and E_2^*, our equation reduces to

$$\frac{\mathbb{E}[\text{scale}(E_1^*)\ \text{scale}(E_2^*)]}{1 + \lambda^2} = r_{\text{degree}}.$$

Because the numerator in the equation above equals r_E, we may write

$$\lambda = \sqrt{\frac{r_E}{r_{\text{degree}}} - 1}.$$

This equation gives a simple expression for the amount of noise, λ, that we must add to a variable E so as to equalize its noise level with that of protein interaction degree.

For our variables of interest, we have $r_{\text{expr}} = r_{\text{abund}} = 0.72$, $r_{\text{degree}} = r_{\text{centrality}} = 0.112$, $r_{\text{disp}} = 0.561$, and $r_{\text{CAI}} = r_{\text{length}} = 1$. As a result, $\lambda_{\text{expr}} = \lambda_{\text{abund}} = 2.329$, $\lambda_{\text{degree}} = \lambda_{\text{centrality}} = 0$, $\lambda_{\text{disp}} = 2.003$, and $\lambda_{\text{CAI}} = \lambda_{\text{length}} = 2.815$. Figure 1*c* shows the mean results from PCR analyses of the predictor variables after adding noise. Standard deviations were calculated from >2000 independent random draws.

Principal Components Regressions

All regressions were performed in R (www.r-project.com). In all cases, we have retained all the components in the PCRs. Although there are techniques designed to assess the appropriate number of "nondegenerate" components, such techniques are inherently subjective (Jackson 1993), and so their utility in this context is unclear.

Parameterizing the Model

Our phenomenological model of expression level (E), abundance (A), CAI (C), protein interaction degree (D), and evolutionary rate (R) depends on 4 parameters according to the equations

$$\begin{aligned} E &= \alpha Z_1 + \beta Z_2 + Z_3, \\ A &= \alpha Z_1 + \beta Z_2 + Z_4, \\ C &= \alpha Z_1 + \beta Z_2 + Z_5, \\ D &= \alpha Z_1 + Z_6, \\ R &= Z_2 + Z_6, \end{aligned}$$

where each Z_i is an independent, normally distributed random variable with mean zero and variance one. In addition to the underlying model, we also specify equations that describe noisy versions of the predictor variables, representing 2 independent measurements of expression (E_1^*, E_2^*), a measurement of abundance (A^*), a measurement of CAI (C^*), and 2 independent measurements of interaction degree (D_1^*, D_2^*):

$$\begin{aligned} E_1^* &= E + n_E W_1, \\ E_2^* &= E + n_E W_2, \\ A^* &= A + n_E W_3, \\ C^* &= C, \\ D_1^* &= D + n_D W_4, \\ D_2^* &= D + n_D W_5. \end{aligned}$$

In these equations, each W_i is an independent Gaussian variable, and we have conservatively assumed that abundance data are as noisy as expression data ($n_A = n_E$). We wish to choose the 4 parameters α, β, n_E, and n_D so as to match important features of the yeast data: 1) the noise in mRNA expression data, $r_{\text{expr}} = 0.72$; 2) the noise in protein interaction data, $r_{\text{deg}} = 0.11$; 3) the correlation between measured expression levels and evolutionary rate, $r(E_1^*, R) = 0.56$; and 4) the correlation between measured interaction degree and evolutionary rate, $r(D_1^*, R) = 0.23$. In other words, our parameters should satisfy the following 4 conditions as precisely as possible:

$$\begin{aligned} r(E_1^*, E_2^*) &= 0.72, \\ r(D_1^*, D_2^*) &= 0.11, \\ r(E_1^*, R) &= 0.56, \\ r(D_1^*, R) &= 0.23. \end{aligned}$$

Under the assumptions of our model, we can write analytic expressions for left hand sides of these equations, in terms of our 4 parameters:

$$\begin{aligned} r(E_1^*, E_2^*) &= (\alpha^2 + \beta^2 + 1)/(\sigma_{E^*})^2, \\ r(D_1^*, D_2^*) &= (\alpha^2)/(\sigma_{D^*})^2, \\ r(E_1^*, R) &= \beta/(\sigma_{E^*} \cdot \sigma_R), \\ r(D_1^*, R) &= 1/(\sigma_{D^*} \cdot \sigma_R), \end{aligned}$$

where

$$\sigma_{E^*} = \sqrt{\alpha^2 + \beta^2 + 1 + n_E^2},$$
$$\sigma_{D^*} = \sqrt{\alpha^2 + 1 + n_D^2},$$
$$\sigma_R = \sqrt{2}.$$

Given these analytic expressions, we numerically minimize the square Euclidean distance

$$[r(E_1^*, R) - 0.56]^2 + [r(D_1^*, R) - 0.23]^2 + [r(E_1^*, E_2^*) - 0.72]^2 + [r(D_1^*, D_2^*) - 0.11]^2$$

and thereby find parameters that satisfy our desired conditions to 2 significant digits:

$$\alpha = 0.1992,$$
$$\beta = 2.651,$$
$$n_E = 1.771,$$
$$n_D = 2.900.$$

Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank D. A. Drummond, C. O. Wilke, and A. Raval for productive conversations and contributions to the manuscript. J.B.P. acknowledges support from the Burroughs Wellcome Fund. Note added in proof: recently Kim and Yi (Genetica 2007, DOI 10.1007/s10709-006-9125-2) have independently demonstrated that principle component regression can yield spurious results when predictor variables contain different amounts of noise.

Literature Cited

Bloom JD, Adami C. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol. 3.

Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA. 2001. Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell. 12:323–337.

Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics. 169:1915–1925.

Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. J Mol Evol. 1:26–45.

Drummond D, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Duan N, Li KC. 1991. Slicing regression: a link-free regression method. Ann Stat. 19:505–530.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science. 296:750–752.

Gavin AC, Bosche M, Krause R, et al. (38 co-authors). 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 415:141–147.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature. 425:737–741.

Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM. 1993. Ancient conserved regions in new gene sequences and the protein databases. Science. 259:1711–1716.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803–806.

Han JD, Bertin N, Hao T, et al. (11 co-authors). 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 430:88–93.

Hirsh AE, Fraser HB. 2001. Protein dispensability and the rate of evolution. Nature. 411:1046–1049.

Hirsh AE, Fraser HB, Wall DP. 2005. Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol Biol Evol. 22:174–177.

Ho Y, Gruhler A, Heilbut A, et al. (46 co-authors). 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature. 415:180–183.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell. 95:717–728.

Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? Curr Biol. 9:747–750.

Ingram VM. 1961. Gene evolution and the haemoglobins. Nature. 189:704–708.

Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology. 74:2204–2214.

Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol Cell. 9:1133–1143.

Koonin EV, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 17:481–487.

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in caenorhabditis elegans. J Mol Evol. 52:275–280.

McInerney JO. 2006. The causes of protein evolutionary rate variation. Trends Ecol Evol. 21:230–232.

Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA. 102:10930–10935.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246:96–98.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927–931.

Pal C, Papp B, Hurst LD. 2003. Genomic function: rate of evolution and gene dispensability. Nature. 421:496–497.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.

Reguly T, Breitkreutz A, Boucher L, et al. (20 co-authors). 2006. Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. J Biol. 5:11.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21:108–116.
Rocha EPC. Forthcoming 2006. The quest for the universals of protein evolution. Trends Genet. 22:412–416.
von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 417:399–403.
Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA. 12:5483–5488.
Warringer J, Ericson E, Fernandez L, Nerman O, Blomberg A. 2003. High-resolution yeast phenomics resolves different physiological features in the saline response. Proc Natl Acad Sci USA. 100:15724–15729.
Wilson A, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem. 46:573–639.

Michele Vendruscolo, Associate Editor

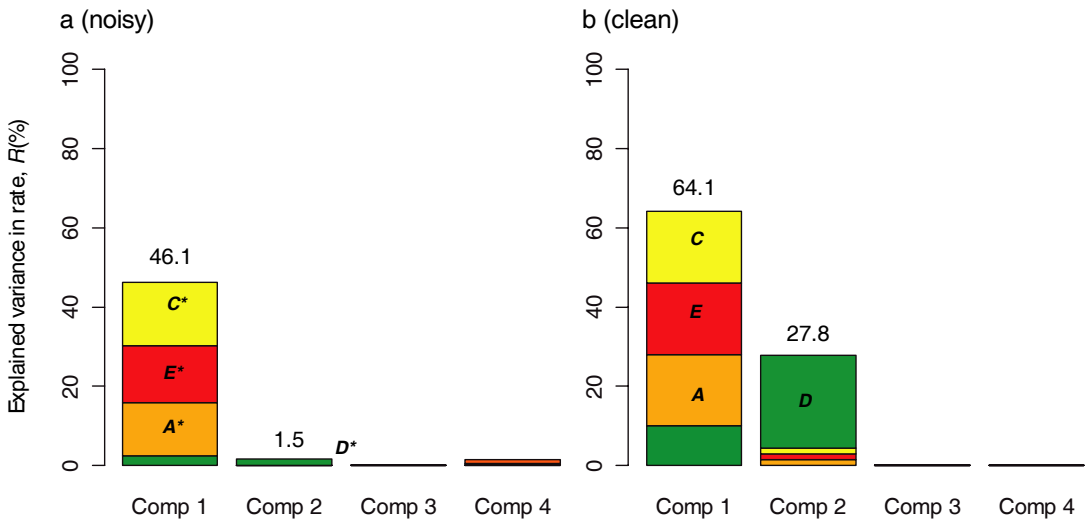
Accepted February 8, 2007

Assessing the determinants of evolutionary rates in the presence of noise

Joshua B. Plotkin and Hunter B. Fraser

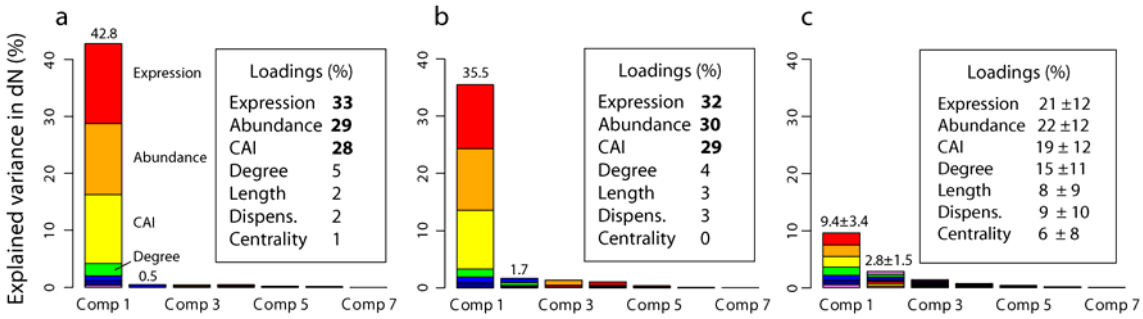
Supplementary Information

In this supplement we present the results of an alternative model of evolutionary rates that is slightly more detailed that the model presented in the paper:

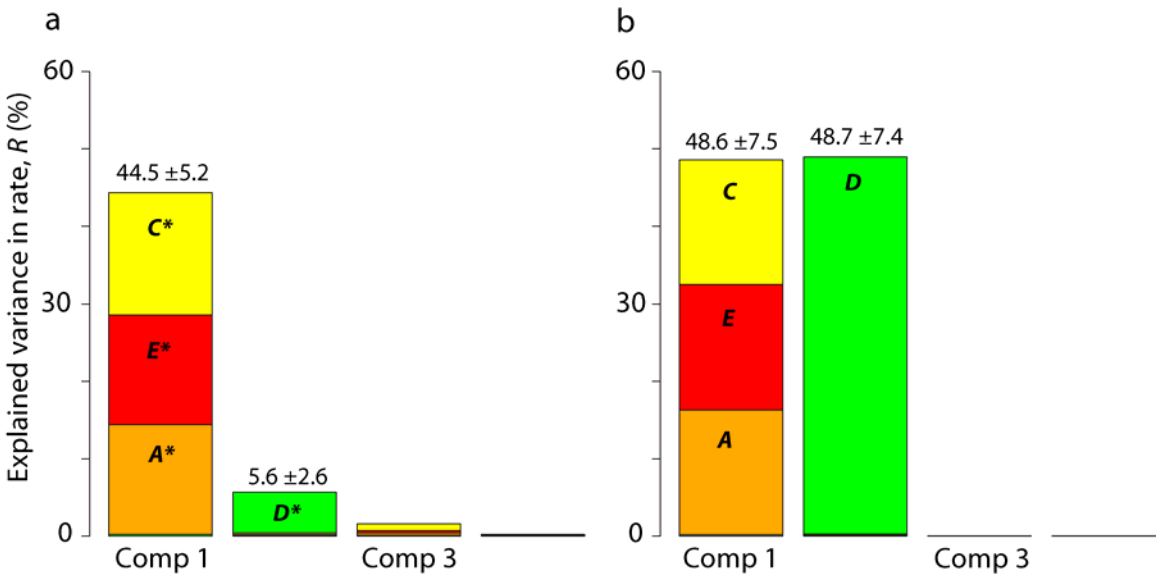


Supplementary Figure A. PCR analysis of the noisy (a) and underlying (b) variables in a simple model of evolutionary rates. When applied to the noisy variables, which represent measurable quantities, the regression indicates a single, dominant determinant of evolutionary rates; and the degree of protein-protein interactions appears to explain significantly less variation in rate than CAI, abundance, or expression. But when applied to the underlying noise-less variables, the PCR analysis reveals the opposite conclusion: no single component dominates evolutionary rates; and the degree of interactions explains significantly more variation than CAI, abundance, or expression.

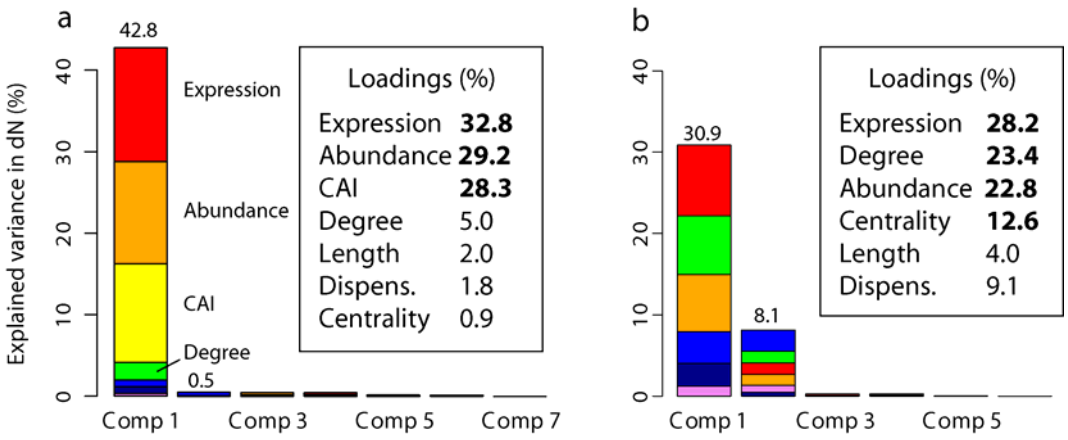
The model shown above is identical to the one presented in the paper, except for an additional term in the degree equation: $D = \alpha Z_1 + \beta Z_2 + \gamma Z_6$. As a result of this minor modification, the first component of the noisy regression contains a small loading by degree, similar to that observed in the empirical data (Fig. 1a). This modified model demonstrates the same conclusion as the model presented in the paper: the noisy data implicate a single determinant, whereas the underlying variables reveal multiple determinants. Parameters are given by α=0.199, β=2.651, γ=0.5, n_E=1.77, n_D =2.0.



Color version of Figure 1



Color version of Figure 2



Color version of Figure 3