

# **Mutation rate and the cost of complexity**

Ralph Haygood

Biology Department, Duke University, Durham, NC 27708

rhaygood@duke.edu

*Molecular Biology and Evolution* **23**:957–963 (2006)

## **Abstract**

Two recent theoretical studies of adaptation suggest that more complex organisms tend to adapt more slowly. Specifically, in Fisher's "geometric" model of a finite population where multiple traits are under optimizing selection, the average progress ensuing from a single mutation decreases as the number of traits increases—the "cost of complexity." Here, I draw on molecular and histological data to assess the extent to which on a large phylogenetic scale, this predicted decrease in the rate of adaptation per mutation is mitigated by an increase in the number of mutations per generation as complexity increases. As an index of complexity for multicellular organisms, I use the number of visibly distinct types of cell in the body. Mutation rate is the product of mutational target size and population mutation rate per unit target. Despite much scatter, genome size appears to be positively correlated with complexity (as indexed by cell-type number), which along with other considerations suggests that mutational target size tends to increase with complexity. In contrast, effective population mutation rate per unit target appears to be negatively correlated with complexity. The net result is that mutation rate probably does tend to increase with complexity, although probably not fast enough to eliminate the cost of complexity.

## Introduction

Students of evolution have an abiding interest in the tempo of evolution, including rates at which traits respond when environmental changes impose directional selection on them. In this regard, two recent theoretical studies of adaptation offer an intriguing suggestion. Using the “geometric” model introduced by Fisher (1930, ch. 2), Orr (2000) and Welch and Waxman (2003) studied the rate of adaptation in a finite population where multiple traits are under optimizing selection. When a mutation occurs, the population may make progress toward the optimal trait values. Orr showed that as the number of traits increases, the average progress ensuing from a single mutation decreases, a phenomenon he dubbed the “cost of complexity.” Welch and Waxman showed that Orr’s finding is robust with respect to several variations on his assumptions. Fisher’s model is highly abstract, and relating it to empirical observations is challenging. Still, the suggestion that more complex organisms tend to adapt more slowly is intriguing (Reznick and Ghalambor, 2001).

However, it is important to recognize that the cost of complexity studied by Orr, Welch, and Waxman is in the rate of adaptation *per mutation*. In some contexts, the rate of adaptation *per generation* is more pertinent. The two adaptation rates are connected through the number of mutations per generation,

$$\text{adaptation per generation} = \text{adaptation per mutation} \times \text{mutations per generation.} \quad (1)$$

Mutation rate can usefully be decomposed into two factors,

$$\begin{aligned} \text{mutations per generation} = & \text{mutational target size in nucleotide sites} \times \\ & \text{mutations per generation per nucleotide site.} \end{aligned} \quad (2)$$

Mutational target size is the size of the functional portion of the genome in which mutations affect organismal traits, for better or worse. Population mutation rate per site is the number of mutations per generation per nucleotide site, in the entire population. These factors might vary with complexity and mitigate or aggravate the cost of complexity. For example, mutational target size might be positively correlated with complexity, say, because more complex organisms tend to have larger

genomes, whereas population mutation rate per site might be negatively correlated with complexity, say, because more complex organisms tend to have smaller populations. Thus, the net effect of mutation rate might depend on the relative magnitudes of opposing tendencies.

That mutation rate might vary with complexity is fairly obvious. Interesting questions include (1) do available data suggest that this possibility is realized, (2) if so, does mutation rate tend to increase with complexity, and (3) if so, is the increase fast enough to eliminate the cost of complexity? Here, I draw on molecular and histological data to address these questions on a large phylogenetic scale, using the number of visibly distinct types of cell in the body as an index of complexity for multicellular organisms. First, I sketch the findings of Orr, Welch, and Waxman and explain more fully what it means to say that mutation rate varies with complexity. I then discuss cell-type number as an index of complexity and argue that it is worth considering despite its limitations. Next, I present evidence that despite much scatter, genome size is positively correlated with complexity (as indexed by cell-type number), and I argue that mutational target size probably tends to increase with complexity. I then present evidence that population mutation rate per site is negatively correlated with complexity, and I assess the net variation of mutation rate with complexity. Finally, I discuss my findings, their limitations, and directions for future research. In brief, on the scale of all multicellular eukaryotes, the answers to the questions above appear to be yes, yes, and no—(1) mutation rate probably does vary with complexity, and (2) it probably tends to increase, but (3) the increase is probably too slow to eliminate the cost of complexity.

## **Fisher's model and the cost of complexity**

The following is merely a sketch; see Orr (2000) and Welch and Waxman (2003) for details. In Fisher's model, the state of a finite population of a haploid organism is represented by a point in  $n$ -dimensional space, whose coordinates are the values of  $n$  traits under optimizing selection; standing variation, whether genetic or environmental, is neglected. The optimal state is the origin  $\vec{0}$ , with fitness  $w[0] = 1$ , and the initial state is  $\vec{z}$ , with fitness  $w[z] = \exp[-z^2/2]$ , where  $z$  is the Euclidean length of  $\vec{z}$ . A mutation changes the phenotype of an individual from  $\vec{z}$  to  $\vec{z}' = \vec{z} + \Delta\vec{z}$ .

The direction of  $\Delta\vec{z}$  is random. In Orr’s analysis, the size of  $\Delta\vec{z}$  is a constant, whereas in Welch and Waxman’s, it is drawn from a probability distribution with mean  $\rho$ . If  $z' > z$ , the mutation is disfavored and assumed to be lost. Alternatively, if  $z' < z$ , the mutation is favored and assumed to be fixed with probability  $2sN_e/N$ , where  $s = (w[z']/w[z]) - 1$ ,  $N_e$  is effective population size, and  $N$  is census population size (Kimura, 1962; the possibility that  $N_e \neq N$  is usually neglected). Either way, the mutation is assumed to meet its fate before another mutation occurs. The change in  $\ln[w]$  is a convenient measure of the progress of the population. Averaging over the probability distributions of mutation direction and size yields

$$E\{\Delta_{\text{mut}} \ln[w]\} \approx \frac{2z^2\rho^2}{n} F\left[\frac{2z^2}{n\rho^2}\right] \frac{N_e}{N}. \quad (3)$$

The subscript “mut” signifies that the change is per mutation.  $F$  is a monotonically increasing function depending on the probability distribution of mutation size but not  $n$ ,  $z$ , or  $\rho$  except through its argument (Welch and Waxman, 2003). This expected value corresponds to “adaptation per mutation” in (1). As  $n$  increases, the expression  $2z^2\rho^2/n$  decreases, and the argument  $2z^2/(n\rho^2)$  and hence the value of  $F$  decreases, albeit slowly unless the argument is small (Welch and Waxman, 2003). Thus, the rate of adaptation per mutation decreases as  $n$  increases, the minimum cost of complexity being given by the initial factor of  $1/n$ .

The cost of complexity transcends some limitations of Fisher’s model. For example, Welch and Waxman investigated the implications of grouping traits into modules, such that each mutation affects a single module. They showed that modularity can mitigate the cost to a modest extent, but the minimum cost  $1/n$  persists. However, Fisher’s model has other limitations (Barton, 1998; Barton and Partridge, 2000). Perhaps the most significant are its neglect of standing genetic variation, although adaptation does sometimes involve such variation, and its assumption of a simple peak on the fitness landscape, although selection for coadaptation of traits must sometimes lead to ridges. Although we will not consider them further, these limitations call for investigation.

As in (1), the rate of adaptation per generation is the product of the rate of adaptation per mutation and the number of mutations per generation, and as in (2), the latter is the product of

mutational target size and population mutation rate per site, yielding

$$E\{\Delta_{\text{gen}} \ln[w]\} = E\{\Delta_{\text{mut}} \ln[w]\}(\phi G)(Nu) \approx \frac{2z^2 \rho^2}{n} F\left[\frac{2z^2}{n\rho^2}\right] \phi G N_e u. \quad (4)$$

The subscript “gen” signifies that the change is per generation.  $\phi$  is the functional fraction of the genome, as explained below, and  $G$  is genome size in nucleotide sites, so  $\phi G$  is mutational target size.  $N$  is census population size, and  $u$  is the number of mutations per generation per nucleotide site per individual, so  $Nu$  is population mutation rate per site. (Readers of Orr’s analysis should note that his  $\mu$  is my  $\phi Gu$ .) The multiplication by  $N_e/N$  from (3) replaces  $Nu$  with  $N_e u$ , effective population mutation rate per site.  $\phi$  and  $G$  themselves are presumably mutable, but we will assume the rates are negligible.

Available data furnish estimates of  $G$ ,  $N_e u$ , and, very roughly,  $\phi$  for modest numbers of organisms. Here, we ask whether these estimates suggest that  $\phi G$  and  $N_e u$  vary with  $n$ , if so, whether their product tends to increase, and if so, whether the increase is linear or faster, canceling out the minimum cost of complexity  $1/n$ . Presumably, many factors influence  $\phi G$  and  $N_e u$ , and their relationships to  $n$  are not determinate. We seek trends, around which there may be scatter. Moreover, as I will discuss, it is not obvious how to measure  $n$ , so we must consider a practical index of complexity rather than  $n$  itself.  $n$  almost certainly tends to increase linearly or faster with the index  $X$  we will consider. Thus, we will assess the relationships of  $\phi G$  and  $N_e u$  to  $X$ , with particular attention to whether  $\phi G N_e u$  tends to increase linearly or faster with  $X$ , which is almost certainly necessary to cancel out even the minimum cost of complexity.

$\phi$ , the functional fraction of the genome, is essential to this formulation. Eukaryotic genomes contain introns, transposons, and other sequences thought to have little effect on organismal traits (Filipski and Kumar, 2005). The fraction of the genome consisting of such putatively nonfunctional sequences might be positively correlated with complexity, say, because more complex organisms tend to have smaller populations, in which selection against the subtle burden of nonfunctional sequences is less efficient (Lynch, 2006). Estimating  $\phi$  is difficult, as I will discuss, but representing the functional fraction is essential for conceptual clarity. If  $\phi$  were omitted, mean mutation

size  $\rho$  would include mutations in nonfunctional sequences, which have size 0, so an increase in the nonfunctional fraction of the genome would entail a decrease in  $\rho$ .  $\rho$  might vary with  $n$  for other reasons we will not consider (Welch and Waxman, 2003), but variation of the nonfunctional fraction is apportioned to  $\phi$ .

## **A practical index of complexity**

It is not obvious how to measure  $n$ . An organism may be described in terms of an indefinitely large number of traits, any of which may be under optimizing selection. However, for two traits of an organism to correspond to two traits in Fisher's model, they must have some degree of genetic autonomy. The number of traits in such a description of an organism is not indefinitely large, if only because the genome is finite. Still, obtaining such a description seems impractical even for a relatively simple organism. As a thought experiment, Orr (2000) sketched a procedure for estimating  $n$  by analyzing the fitness effects of the same large set of mutations in both an optimally adapted strain and one or more suboptimally adapted strains. As Orr acknowledged, this is probably impractical even for a microbe. Thus, we must consider not  $n$  itself but something that can be (and has been) measured and is strongly correlated with  $n$ —a practical index of complexity.

Unfortunately, there is probably no fully satisfactory index. Organismal complexity is notoriously difficult to define and compare across the breadth of organismal diversity (McShea, 1996). For our purposes, phylogenetic breadth is appropriate, because estimates of genome size and, more so, effective population mutation rate per site are available only for modest numbers of organisms with wide phylogenetic distributions, so we should expect to discern trends only on large phylogenetic scales. Breadth is also desirable in order to consider, say, the evolutionary response of an entire ecological guild to a climate change. However, it is problematic, because defining and comparing complexity requires comparable parts or processes, but on large phylogenetic scales, none are evident above the organizational level of the cell (McShea, 1996).

Accordingly, one practical index of complexity, arguably the most practical index available on a large phylogenetic scale, is the number of visibly distinct types of cell in the body, which I will

denote  $X$ . Estimates of  $X$  have been culled from the histological literature and used to explore topics including the relationship between body size and complexity (Bell and Mooers, 1997; Bonner, 2004; McCarthy and Enquist, 2005) and evolutionary trends in complexity (Valentine et al., 1994; McShea, 1996; McCarthy and Enquist, 2005). The best compilations known to me are those of Bell and Mooers (1997) and McCarthy and Enquist (2005), which include 134 and 64 species, respectively, with overlap. In what follows, all estimates of  $X$  are from these compilations, which standardize estimates in the same way.

Obviously, cell-type number  $X$  is not a fully satisfactory index for several reasons. It does not apply to unicellular organisms, and it ignores differences in complexity between cell types within multicellular organisms. Exactly how the complexity of a tissue, organ, or organism relates to the number of cell types it contains is unclear. Different authors recognize different cell types, and some distinctions depend on staining procedures and other technicalities. Bell and Mooers' and McCarthy and Enquist's compilations are impressive, but few of their species are model organisms, and several major model organisms are missing. Nonetheless,  $X$  is worth considering. One motivation is that although the relationship of  $n$  to  $X$  is unclear, it seems very likely that each cell type gives rise to at least one organismal trait and that the genetic substrate for the differentiation of the type imparts some degree of genetic autonomy to the trait or traits. It follows that  $n$  almost certainly tends to increase linearly or faster with  $X$ , so mutation rate must do likewise to cancel out the minimum cost of complexity  $1/n$ . Another motivation for considering  $X$  is that, as I will show, available data suggest a surprisingly strong correlation of genome size with  $X$ , which warrants further investigation.

In what follows,  $X$  is the independent variable in several (logarithmic) regressions, but this is not meant to impute causation. We are interested in the relationship between mutation rate and complexity, whatever the causes of the relationship may be.  $X$  is presumably correlated with many organismal traits through many causal mechanisms. However, one trait is probably central to this network of relationships, namely, body size, which I will denote  $B$ . Bell and Mooers (1997) and

McCarthy and Enquist (2005) compiled estimates of  $B$  as well as  $X$  and showed that on large phylogenetic scales,  $B$  and  $X$  are positively correlated—more complex organisms tend to be larger (cf. Bonner, 2004). Moreover, Lynch (2006) argued that effective population mutation rate per site  $N_e u$  and  $B$  are negatively correlated—larger organisms tend to have smaller effective populations (cf. Lynch and Conery, 2003); they may also tend to have higher  $u$ , but not enough to overcome their lower  $N_e$ . We might therefore expect that  $N_e u$  tends to decrease as  $X$  increases, and as I will show, this appears to be the case. Less is known about the relationship of mutational target size  $\phi G$  to  $B$ . We will use estimates of  $B$  from Bell and Mooers and McCarthy and Enquist to assess the relationships of genome size  $G$ ,  $N_e u$ , and  $X$  to  $B$ . We will also assess how far the relationships of  $G$  and  $N_e u$  to  $X$  are determined by the relationships of these variables to  $B$ , that is, we will assess the residual relationships of  $G$  and  $N_e u$  to  $X$  after statistically eliminating  $B$ .

Following standard practice (Felsenstein, 1985; Garland et al., 1992; Sokal and Rohlf, 1995), we will compute statistics using base-10 log-transformed estimates of  $X$ ,  $B$ , and other variables. If the regression coefficient of, say,  $\log[X]$  on  $\log[B]$  is  $b$ , it follows that  $X$  tends to vary with  $B$  like  $B^b$ . One of the statistics we will consider, the rank-order (Spearman) correlation coefficient, is unaffected by log transformation.

## **Mutational target size and complexity**

We will first assess the relationship of genome size  $G$  to cell-type number  $X$  and then consider functional fraction  $\phi$  and mutational target size  $\phi G$ , which are difficult to estimate. For diploid and polyploid organisms, we will take  $G$  to be haploid genome size.

Seeking a trend in  $G$  with  $X$  might seem unpromising, because  $G$  is well known to vary widely among organisms of similar apparent complexity. For example, a recent review states that “genome size varies enormously among species, and... except at a very basic level (e.g., prokaryotes versus eukaryotes) there is no correlation between genome size and notions of organismal complexity.” (Filipski and Kumar, 2005, p. 531) This is known as the “C-value paradox,” the C-value of an organism being the amount of DNA in a gamete nucleus (Gregory, 2005).

Surprisingly, despite much scatter, available data suggest an appreciable positive correlation of  $G$  with  $X$  on the scale of all multicellular eukaryotes. The data are presented in the supplement and plotted in Figure 1, and statistics are presented in the first three rows of Table 1. There are data for 37 species, including 22 animals, 7 algae, and 8 plants, representing 14 phyla or, for plants, divisions; these are all of Bell and Mooers' (1997) and McCarthy and Enquist's (2005) species for which estimates of  $G$  are available through the Animal Genome Size Database (Gregory, 2001) or the Plant DNA C-values Database (Bennett and Leitch, 2003), which includes algae. Over all species, the standard (Pearson) and rank-order (Spearman) correlation coefficients are moderately large,  $r = 0.47$  and  $r_s = 0.56$ , respectively. Inspection of Figure 1 suggests that for a given  $X$ ,  $G$  tends to be smaller for animals than for algae and plants. The coefficients within these groups are larger.

The standard  $p$ -values of all the statistics in the left half of Table 1 are less than 0.05, and most are less than 0.01. However, these  $p$ -values may not be meaningful (hence they are omitted from Table 1) due to phylogenetic dependencies among the data. The standard way of addressing this issue is the method of phylogenetically independent contrasts (PIC) (Felsenstein, 1985; Garland et al., 2005). As implemented in the PDAP package of Mesquite (Maddison and Maddison, 2004; Midford et al., 2005), using the tree topology presented in the supplement, and setting all branch lengths to unity (suggestive of "punctuational" evolution, Garland et al., 1992), PIC yields the statistics in the right half of Table 1. Over all species, the PIC correlation coefficient is moderately large,  $r = 0.44$ , and significant,  $p = 0.011$ . Within animals, the coefficient is larger and more significant, but within algae and plants, it is small and nonsignificant. The latter may indicate a weak relationship between  $G$  and  $X$  within this group.

Caution is appropriate in interpreting these results for several reasons. The phylogenetic scale is large, and the trend may not prevail on smaller scales. The data are sparse, omitting many phyla of animals, algae, and plants and the entire kingdom of fungi. There are measurement errors, and multiple estimates of  $X$  and  $G$  for several species suggest that they are substantial. For PIC, the tree topology may be wrong, and the branch lengths and evolutionary model (Brownian motion) are

certainly wrong. (However, apart from one case mentioned below, the regressions of the absolute values of the standardized contrasts on the standard deviations of the standardized contrasts are nonsignificant, so according to this standard diagnostic, the tree and model are not incoherent with the data, Garland et al., 1992.)

Despite these concerns, the trend may well be genuine. The PIC results are fairly insensitive to branch lengths. Over all species, Grafen's (1989) arbitrary branch lengths with Grafen's parameter  $\rho = 0.5$  yield  $r = 0.37$  with  $p = 0.035$ , Pagel's (1992) arbitrary branch lengths yield  $r = 0.30$  with  $p = 0.087$ , and Nee's (Purvis, 1989) arbitrary branch lengths yield  $r = 0.38$  with  $p = 0.032$ . There have presumably been evolutionary trends in  $X$  and  $G$ , which the evolutionary model does not accommodate (Garland et al., 2005). For example, the PIC estimated root node has  $X \approx 12$ , but the most recent common ancestor of all multicellular eukaryotes was almost certainly unicellular. However, PIC sometimes estimates correlations well even when it estimates ancestors poorly (Oakley and Cunningham, 2000). Moreover, when the root node is forced (using a "ghost node," Garland et al., 2005) to have  $X = 1$  and  $G = 39$  Mbp, the smallest value of  $G$  in the data set, PIC yields  $r = 0.66$  with  $p < 0.001$ . (In this case, however, the regressions of the absolute values of the standardized contrasts on the standard deviations of the standardized contrasts are significant.) Thus, although the relationship should be reassessed using additional data, and although there is much scatter around the trend, the present analysis suggests that  $G$  tends to increase with  $X$ , at least within animals. The PIC regression coefficients in Table 1 suggest that the increase is roughly linear or somewhat faster.

The relationship of  $G$  to  $X$  is mostly determined by the relationships of these variables to body size  $B$ , as indicated in the first three rows of Table 2. Over all species, within animals, and within algae and plants,  $G$  and  $X$  are moderately to strongly correlated with  $B$ , and their residuals after regression on  $B$  are weakly correlated with each other. The moderate to strong correlations of  $G$  and  $X$  with  $B$  are broadly consistent with previous studies (Bell and Mooers, 1997; Bonner, 2004; Gregory, 2005; McCarthy and Enquist, 2005). Over all species, the PIC regression coefficients of  $G$  and  $X$  on  $B$  are 0.15 and 0.07, respectively.

Functional fraction  $\phi$  almost certainly tends to decrease as  $G$  increases. Larger eukaryotic genomes certainly tend to contain higher fractions of non-protein-coding sequences (Lynch and Conery, 2003; Lynch, 2006). However, some of these sequences function in gene regulation (Wray et al., 2003), and functional transposons and even pseudogenes are known (Hirotsume et al., 2003; Kidwell, 2005). The prevalence of such phenomena is unknown, so it seems premature to estimate how fast  $\phi$  tends to decrease as  $G$  increases. It seems very unlikely, however, that the decrease is as fast as  $1/G$ . For example, Lynch and Conery (2003, Fig. 4) showed that for sequenced genomes with  $G \gtrsim 10$  Mbp, the fraction of the genome *not* consisting of transposons tends to decrease as  $G$  increases roughly like  $1/G^{0.05}$ . Given  $G$  probably tends to increase with  $X$ , mutational target size  $\phi G$  probably does likewise but somewhat slower.

## **Effective population mutation rate per site and complexity**

Estimates of effective population mutation rate per site  $N_e u$  are as scarce as estimates of cell-type number  $X$ . The best compilation known to me is that of Lynch (2006), which includes 130 eukaryotic species, many of them unicellular. Most of these estimates derive from measurements of nucleotide diversity at intronic or silent exonic sites. Estimates of both  $N_e u$  and  $X$  are available for very few species. Lynch's species include only six of Bell and Mooers' (1997) and McCarthy and Enquist's (2005), three of which (dog, human, and mouse) are close relatives. Thus, we must consider higher taxa.

For several phyla or, for plants, divisions, Bell and Mooers and/or McCarthy and Enquist give an estimate of  $X$  for at least one species, and Lynch gives an estimate of  $N_e u$  for at least one species. A data point representing the higher taxon can be constructed by averaging over the species. This procedure yields 12 data points, involving 160 species of animals, algae, plants, fungi, and protists. The supplement presents the species data and higher-taxon averages, along with the coefficients of variation within the higher taxa. Most of the coefficients exceed 0.1, indicating substantial heterogeneity within the taxa. Grouping species at a lower taxonomic level might lead to more homogeneous groups, but it would take in fewer species.

These data suggest an appreciable negative correlation of  $N_e u$  with  $X$  on the scale of all multi-cellular eukaryotes. The data are plotted in Figure 2, and statistics are presented in the last row of Table 1. The standard and rank-order correlation coefficients are large,  $r = -0.76$  and  $r_S = -0.71$ , respectively. Using the tree topology presented in the supplement and setting all branch lengths to unity, the PIC correlation coefficient is large,  $r = -0.65$ , and significant,  $p = 0.023$ . Interpretive caution is again appropriate, not only for the reasons mentioned before but also due to the use of higher taxa rather than species and the possibility that estimates of  $N_e u$  derived from intronic or silent exonic sites are distorted by selection. However, the PIC results are again fairly insensitive to branch lengths, and the PIC estimated root node has  $X \approx 4$ , a modest improvement. Thus, the present analysis suggests that  $N_e u$  tends to decrease as  $X$  increases. The PIC regression coefficient,  $b = -0.54$ , suggests that the decrease is slower than  $1/X$ .

As indicated in the last row of Table 2, the relationship of  $N_e u$  to  $X$  is partly determined by the relationships of these variables to body size  $B$ , but the residual relationship is appreciable. In this analysis, the PIC regression coefficient of  $X$  on  $B$  is 0.10.

The PIC regression coefficients in Table 1 suggest that  $GN_e u$  tends to increase with  $X$  roughly like  $X^{0.97-0.54} = X^{0.43}$ . Given the uncertainties in the coefficients, the exponent could be less than 0 or greater than 1, although neither is very probable. Assuming functional fraction  $\phi$  tends to decrease no faster than roughly like  $1/G^{1-(0.54/0.97)} = 1/G^{0.44}$  as  $G$  increases, it follows that  $\phi GN_e u$  probably tends to increase with  $X$ . However, the increase is probably slower than linear, which is too slow to cancel out even the minimum cost of complexity  $1/n$ . This conclusion is strengthened if  $n$  tends to increase faster than linearly with  $X$ , which is plausible in view of the combinatorial character of tissues, organs, and organisms.

## Discussion

The cost of complexity is a predicted decrease in the average progress toward the optimal trait values ensuing from a single mutation as the number of selected traits increases. Per generation, it might be mitigated by an increase in mutation rate. The present study suggests that on the

scale of all multicellular eukaryotes, mutation rate probably does tend to increase with complexity, although probably not fast enough to eliminate the cost of complexity.

The latter conclusion is not wholly secure, not only because the data are sparse and the assumptions of PIC are not perfectly fulfilled but also due to measurement errors in the data. In particular, measurement errors in cell-type number  $X$  are difficult to estimate but may be substantial. Some researchers would therefore advocate using reduced major axis (RMA) regression coefficients on  $X$  rather than standard ones; if  $b$  is the standard coefficient and  $r$  is the correlation coefficient, the RMA coefficient is  $b/|r|$  (Sokal and Rohlf, 1995). This would suggest that  $GN_{eu}$  tends to increase with  $X$  roughly like  $X^{(0.97/0.44)-(0.54/0.65)} = X^{1.37}$  instead of  $X^{0.43}$ . Assuming functional fraction  $\phi$  tends to decrease slowly as  $G$  increases,  $\phi GN_{eu}$  would tend to increase somewhat faster than linearly with  $X$ . This would cancel out the minimum cost of complexity  $1/n$ , provided  $n$  did not increase much faster than linearly with  $X$ . Whether RMA regression is indeed more appropriate in this context is a complicated issue (Garland et al., 1992; Sokal and Rohlf, 1995). For future studies, uniform data collection procedures would be helpful, so that measurement errors could be treated systematically.

This study also suggests that mutational target size probably tends to increase with complexity, which is intuitively appealing. However, this conclusion builds on the more surprising conclusion that genome size itself probably tends to increase with complexity, at least within animals, contrary to the strongest statements of the C-value paradox. It should be noted that even if this trend is genuine, there is much scatter around it, and it may not prevail on smaller phylogenetic scales. Further investigation is warranted, using cell-type number and other measurable, quantitative indices of complexity.

Although their  $p$ -values are persuasive, the analyses presented here involve small sets of noisy data. Eventually, they should be repeated with high-quality estimates of  $X$ , genome size  $G$ , effective population mutation rate per site  $N_{eu}$ , and, if possible, functional fraction  $\phi$  for all of a sizable number of organisms with a wide phylogenetic distribution. Presumably, estimates of  $G$ ,  $N_{eu}$ , and

even  $\phi$  will accumulate in the normal course of genomic research, but estimates of  $X$  will require a special effort. The strong correlations reported here provide impetus for such an effort.

Even if these trends are genuine, they may not prevail on smaller phylogenetic scales—magnitudes and even signs of correlation and regression coefficients may vary. For example, McCarthy and Enquist (2005) argued that the correlation of body size with cell-type number is absent within several animal phyla. However, better indices of complexity may be available on smaller scales. For example, appendage specialization has been used as an index of complexity for aquatic arthropods (Cisne, 1974).

So do more complex organisms tend to adapt more slowly? This study offers some support for this suggestion, but there are at least three additional qualifications. First, not only the relationship between mutation rate and complexity but also the cost of complexity itself are trends, around which there is scatter. Presumably, the scatter varies from one group of organisms to another, because Fisher's model applies better to some organisms (and adaptive challenges) than others, and mutation rate is more influenced by factors uncorrelated with complexity within some groups than others. Undoubtedly, more complex organisms sometimes adapt more quickly. Second, as I have mentioned, Fisher's model neglects standing genetic variation and ridged fitness landscapes, significant limitations that call for investigation. Third, greater complexity may lead to slower but more complete adaptation. For example, if two traits are completely mutationally correlated, they evolve as one trait rather than two, and if the initial relationship of one to the other differs from the optimal relationship, the population cannot reach the optimal state.

## **Acknowledgments**

Thanks to Mike Lynch for sending me his most recent compilation of estimates of  $N_e u$  prior to its publication, Megan McCarthy for sending me the data underlying Fig. 3b of McCarthy and Enquist, 2005, and Jim Balhoff for introducing me to Mesquite and PDAP. Thanks to Allen Orr, Marta Wayne, Greg Wray, and two anonymous reviewers for comments on drafts of this paper and the participants in the 2005 SMBE Tri-National Young Investigators' Workshop for comments

on a talk about this research. This research was supported by a National Science Foundation Postdoctoral Fellowship in Biological Informatics (Grant No. 0434655).

## References

- Barton, N., and Partridge, L., 2000. Limits to natural selection. *BioEssays* **22**:1075–1084.
- Barton, N. H., 1998. The geometry of natural selection. *Nature* **395**:751–752.
- Bell, G., and Mooers, A. O., 1997. Size and complexity among multicellular organisms. *Biological Journal of the Linnean Society* **60**:345–363.
- Bennett, M. D., and Leitch, I. J., 2003. Plant DNA C-values database. <http://www.rbgekew.org/cval/homepage.html>.
- Bonner, J. T., 2004. The size–complexity rule. *Evolution* **58**:1883–1890.
- Cisne, J. L., 1974. Evolution of the world fauna of aquatic free-living arthropods. *Evolution* **28**:337–366.
- Felsenstein, J., 1985. Phylogenies and the comparative method. *American Naturalist* **125**:1–15.
- Filipski, A., and Kumar, S., 2005. Comparative genomics in eukaryotes. Gregory, T. R. (Ed.), *The evolution of the genome*, pp. 521–583. Elsevier Academic Press, Burlington, MA.
- Fisher, R. A., 1930. *The genetical theory of natural selection*. Oxford University Press, Oxford, UK.
- Garland, Jr., T., Bennett, A. F., and Rezende, E. L., 2005. Phylogenetic approaches in comparative physiology. *Journal of Experimental Biology* **208**:3015–3035.
- Garland, Jr., T., Harvey, P. H., and Ives, A. R., 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology* **41**:18–32.
- Grafen, A., 1989. The phlogenetic regression. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **326**:119–157.

- Gregory, T. R., 2001. Animal genome size database. <http://www.genomesize.com>.
- Gregory, T. R., 2005. Genome size evolution in animals. Gregory, T. R. (Ed.), *The evolution of the genome*, pp. 3–87. Elsevier Academic Press, Burlington, MA.
- Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A., 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**:91–100.
- Kidwell, M. G., 2005. Transposable elements. Gregory, T. R. (Ed.), *The evolution of the genome*, pp. 165–221. Elsevier Academic Press, Burlington, MA.
- Kimura, M., 1962. On the probability of fixation of mutant genes in populations. *Genetics* **47**:713–719.
- Lynch, M., 2006. The origins of eukaryotic gene structure. *Molecular Biology and Evolution* **23**:450–468.
- Lynch, M., and Conery, J. S., 2003. The origins of genome complexity. *Science* **302**:1401–1404.
- Maddison, W. P., and Maddison, D. R., 2004. Mesquite: A modular system for evolutionary analysis, version 1.05. <http://mesquiteproject.org>.
- McCarthy, M. C., and Enquist, B. J., 2005. Organismal size, metabolism and the evolution of complexity in metazoans. *Evolutionary Ecology Research* **7**:681–696.
- McShea, D. W., 1996. Metazoan complexity and evolution: Is there a trend? *Evolution* **50**:477–492.
- Midford, P. E., Garland, Jr., T., and Maddison, W. P., 2005. PDAP package of Mesquite, version 1.06. [http://mesquiteproject.org/pdap\\_mesquite](http://mesquiteproject.org/pdap_mesquite).
- Oakley, T. H., and Cunningham, C. W., 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution* **54**:397–405.
- Orr, H. A., 2000. Adaptation and the cost of complexity. *Evolution* **54**:13–20.

- Pagel, M. D., 1992. A method for the analysis of comparative data. *Journal of Theoretical Biology* **156**:431–442.
- Purvis, A., 1989. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **348**:405–421.
- Reznick, D. N., and Ghalambor, C. K., 2001. The population ecology of contemporary adaptations: What empirical studies reveal about the conditions that promote adaptive evolution. *Genetica* **112–113**:183–198.
- Sokal, R. R., and Rohlf, F. J., 1995. *Biometry*, 3rd ed. W. H. Freeman and Company, New York, NY.
- Valentine, J. W., Collins, A. G., and Meyer, C. P., 1994. Morphological complexity increase in metazoans. *Paleobiology* **20**:131–142.
- Welch, J. J., and Waxman, D., 2003. Modularity and the cost of complexity. *Evolution* **57**:1723–1734.
- Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., and Romano, L. A., 2003. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20**:1377–1419.

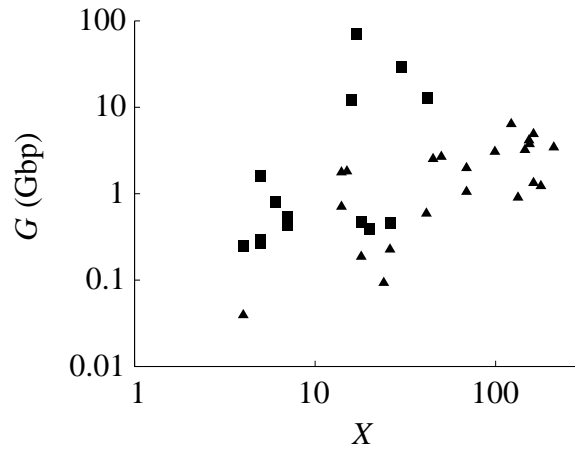


Figure 1: Log–log plot of genome size  $G$  (in billions of base pairs) versus cell-type number  $X$ . Triangles represent animals, and squares represent algae and plants. (Estimates of  $G$  are from Gregory, 2001 and Bennett and Leitch, 2003, and estimates of  $X$  are from Bell and Mooers, 1997 and McCarthy and Enquist, 2005. See supplement for details.)

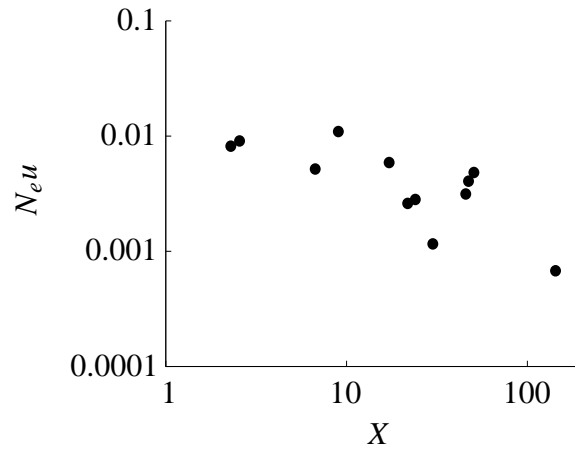


Figure 2: Log–log plot of effective population mutation rate per site  $N_e u$  versus cell-type number  $X$ . (Estimates of  $N_e u$  are from Lynch, 2006, and estimates of  $X$  are from Bell and Mooers, 1997 and McCarthy and Enquist, 2005. See supplement for details.)

variables	taxa	one point per taxon			PIC <sup>b</sup>		
		$r^c$	$r_S^d$	$b^e$	$r^c$	$b^e \pm SE^f$	$p^g$
	all	0.47	0.56	0.58	0.44	$0.97 \pm 0.36$	0.011
$G$ vs. $X$	animals	0.72	0.63	0.88	0.65	$1.30 \pm 0.36$	0.003
	algae and plants	0.62	0.64	1.48	0.16	$0.41 \pm 0.72$	0.581
$N_{eu}$ vs. $X$	all	-0.76	-0.71	-0.50	-0.65	$-0.54 \pm 0.20$	0.023

Table 1: Statistics on genome size  $G$  and effective population mutation rate per site  $N_{eu}$  versus cell-type number  $X$ .<sup>a</sup>

<sup>a</sup>See text for methods and supplement for data. All statistics are computed using base-10 log-transformed data.

<sup>b</sup>Phylogenetically independent contrasts.

<sup>c</sup>Standard (Pearson) correlation coefficient.

<sup>d</sup>Rank-order (Spearman) correlation coefficient.

<sup>e</sup>Regression coefficient.

<sup>f</sup>Standard error.

<sup>g</sup>Two-tailed  $p$ -value, rounded up.

variables	taxa	one point per taxon			PIC <sup>b</sup>		
		$r_{G,B}^c$	$r_{X,B}^c$	$\tilde{r}_{G,X}^d$	$r_{G,B}^c$	$r_{X,B}^c$	$\tilde{r}_{G,X}^d$
$G$ and $X$ vs. $B$	all	0.58	0.59	0.25	0.59***	0.65***	0.05
	animals	0.74	0.85	0.25	0.63**	0.74***	0.31
	algae and plants	0.37	0.27	0.50	0.55*	0.55*	0.29
$N_{eu}$ and $X$ vs. $B$		$r_{N_{eu},B}^c$	$r_{X,B}^c$	$\tilde{r}_{N_{eu},X}^d$	$r_{N_{eu},B}^c$	$r_{X,B}^c$	$\tilde{r}_{N_{eu},X}^d$
	all	-0.59	0.85	-0.69	-0.21	0.79**	-0.80**

Table 2: Statistics on genome size  $G$ , effective population mutation rate per site  $N_{eu}$ , and cell-type number  $X$  versus body size  $B$ .<sup>a</sup>

<sup>a</sup>See text for methods and supplement for data. All statistics are computed using base-10 log-transformed data.

<sup>b</sup>Phylogenetically independent contrasts.

<sup>c</sup>Standard (Pearson) correlation coefficient.

<sup>d</sup>Standard (Pearson) correlation coefficient of residuals after regression on  $B$ .

\*\*\*  $p < 0.001$ , \*\*  $0.001 < p < 0.01$ , \*  $0.01 < p < 0.05$ , where  $p$  is two-tailed  $p$ -value (for PIC only).