J Mol Evol (2003) 56:151–161
DOI: 10.1007/s00239-002-2388-z

JOURNAL OF MOLECULAR EVOLUTION

Cost-Minimization of Amino Acid Usage

Hervé Seligmann

Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

Received: 30 April 2002 / Accepted: 13 August 2002

Abstract. The negative correlation between the frequencies of usage of amino acids and their biosynthetic cost suggests that organisms minimize costs of protein biosynthesis. Empirical results support that: (1) free-living organisms (Archaea, Bacteria, and Eucaryota) minimize the usage of heavy amino acids more than intracellular organisms (viruses, chloroplasts, and mitochondria), a result confirmed by comparing intracellular Bacteria with other Bacteria; (2) avoidance of amino acids with low impact on protein structure (Chou-Fasman indices) is greater than for those with equal molecular weight but greater structural impact: constraints on protein function limit cost-minimization; (3) amino acid weight minimization (WM) for a protein correlates positively with the protein's expression level and with its size; (4) preliminary results suggest that for different proteins, the evolutionary rate of amino acid replacements correlates negatively with WM in these proteins; (5) results suggest that WM decreases with genome-size; and (6) developmental rates correlate positively with WM (within primates and rodents), even after confounding factors were accounted for. Effects of biosynthetic cost-minimization at whole-organism levels vary with metabolic and ecological strategies. Biosynthetic cost-minimization is an adaptive hypothesis that yields a semi-mechanistic explanation for small differences in allele fitness.

Key words: Metabolic strategy — Sequence evolution — Codon bias — Developmental rate — Genome-size — Replacement rate — Endocellularity

Introduction

Abundances of amino acids in protein composition correlate negatively with their molecular weights (Barrai et al. 1995). This observation suggests that biases in amino acid usage in proteins minimize costs of protein synthesis (Dufton 1997). The idea that economical constraints, in addition to structural and functional ones, shape the evolution of proteins, is not new. Sulfur-depleted versions of some abundant proteins are specifically expressed under sulfur-starvation in a cyanobacterium (Mazel and Marlière 1989). Using synonymous-codon bias as an index proportional to translation rate, Akashi and Gojobori (2002) show in most functional groups of proteins in *Bacillus subtilis* (3055 genes) and *Escherichia coli* (3397 genes) that the abundance of energetically less costly amino acids increases proportionally to the level of expression of the genes. The aims of this study are to: (1) compare levels of cost-minimization among different groups of organisms; (2) develop a method that quantifies amino acid cost-minimization that does not make assumptions on the particular metabolism of the organism, in contrast to the method used by Akashi and Gojobori (2002); and (3) test and explore for evolutionary and life-history correlates and consequences of the amino acid usage cost-minimization principle at whole-organism levels.

Current address: Dept. of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, LA 70803, USA; *email:* Herve@uchicago.edu

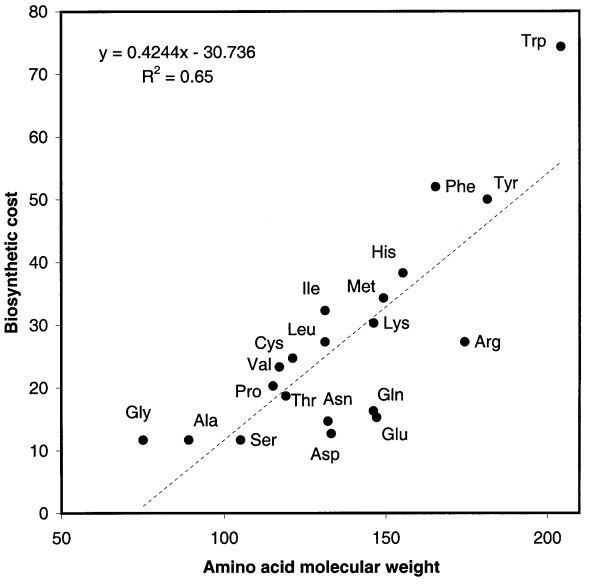

Fig. 1. Metabolic costs of amino acid biosynthesis in *E. coli* as a function of amino acid molecular weight. The y axis is the total energetic cost of an amino acid, the sum of the energy invested in the phosphate bonds in ATP and GTP molecules and the number of available hydrogen atoms carried in NADH, NADPH, and FADH2 molecules to metabolize an amino acid, according to Table 1 in Akashi and Gojobori (2002).

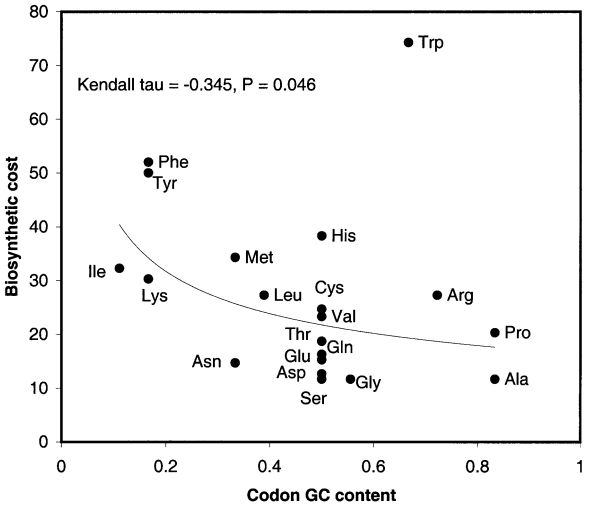

Fig. 2. Metabolic costs of amino acid biosynthesis in *E. coli* as a function of the mean GC content of codons coding for a particular amino acid. The y axis is the same as in Fig. 1.

Quantifying Biosynthetic Costs: Molecular Weight Versus Chemical Energy

Amino acid molecular weights and their structural complexity might reflect biosynthetic costs, resource investment, and perhaps also cytoplasmic toxicity (Dufton 1997). Indeed, Dufton indicated that large amino acids tend to have more complex and diverse chemical properties, which makes their chemical behavior less predictable. It is also possible that the biosynthesis of large and structurally complex amino acids is less accurate (Fresht 1986). These factors would increase different types of costs associated with usage of amino acids proportionally to their size. The estimates of metabolic costs of Akashi and Gojobori (2002) quantify the amount of chemical energy invested in the metabolization of amino acids, but this approach does not estimate other forms of costs associated with the use of amino acids. Figure 1 shows the positive correlation ($r = 0.80$, $df = 18$, $p < 0.01$) of amino acid weight with the metabolic cost of amino acid biosynthesis (Akashi and Gojobori 2002). The correlation in Fig. 1 shows that about 35% of the variation in amino acid molecular weight is not explained by the amount of chemical energy. Tests of the cost-minimization hypothesis using amino acid molecular weight as measure of costs might complement the approach of Akashi and Gojobori (2002), because molecular weight probably estimates additional components of costs, besides the investment of chemical energy.

In this study I use amino acid molecular weight as an estimate of costs rather than the energy invested in each chemical bond in the amino acid because molecular weight is (a) probably proportional to additional types of cost, beyond the chemical energy invested in the amino acid's biosynthesis; (b) the method does not require precise knowledge of the metabolic pathways used by the organism, (c) as a result of point (b), organisms with different metabolic pathways are compared on a common scale, and (d) using molecular weight as an estimate of costs prevents the potential confounding effects of codon GC content with amino acid biosynthetic cost (next section). However, I wish to stress here that for the aim of estimating amino acid costs, amino acid molecular weight and chemical energy should be considered as complementary rather than competing methods, because each method estimates a different aspect of cost.

Quantifying Biosynthetic Costs: GC Content as Confounding Factor

Akashi and Gojobori (2002) noted that A- or T-rich codons tend to encode more costly amino acids. Indeed, a significant correlation exists between the mean GC content of the synonymous-codons coding for a specific amino acid and the biosynthetic costs of that amino acid as calculated by Akashi and Gojobori (2002) (Fig. 2, Kendall rank correlation coefficient $tau = -0.345$, $df = 18$, $p < 0.05$). A similar analysis did not detect any significant association of mean codon GC content with amino acid molecular weight. Hence the association of gene expression levels with the measure of chemical energy invested into the metabolism of amino acids, reported by Akashi and Gojobori

(2002), might be confounded by GC-content effects. Functional, rather than economic, adaptations affect GC content: for example, increases in genome-wide GC content increase the thermic stability of double-stranded DNA under high temperatures, and lead to amino acid replacements that increase the thermal stability of proteins (Bernardi and Bernardi 1986).

Predictions of the Cost-Minimization Hypothesis

The major assumption underlying the idea of cost-minimization of amino acid use is that natural selection would tend to restrict the usage of larger amino acids to those relatively few sites in proteins where the specific properties of the larger amino acids are crucial to the function of the protein. The power of Dufton's hypothesis is in the fact that it generates many testable predictions. Until now, population geneticists have described genetic variability at given loci, and relative fitnesses of each allele, but hypotheses making explicit predictions on why an allele is selectively advantaged as compared to others are few, and often very specific to the gene under study. The cost-minimization hypothesis could be a simple tool of general use in this respect.

I present empirical tests for a set of explicit, falsifiable predictions developed from Dufton's original hypothesis.

Prediction 1: The extent of amino acid weight minimization (WM, the negative of the correlation coefficient of the relative frequency of amino acids with their molecular weight) varies among different groups of organisms. WM is less marked in organisms that absorb amino acids from their environment than in organisms that mostly synthesize amino acids. In general, WM increases with the overall metabolic cost of amino acid synthesis in different organisms.

Prediction 2: For amino acids with the same molecular weight, cost-minimization decreases the usage of amino acids with low structural impact more than the usage of amino acids with the same molecular weight, but greater structural impact. This is because cost-minimization should happen more freely for amino acids with low structural impact.

Prediction 3: Constraints to decrease costs of a protein are proportional to the costs associated with that protein. Therefore, WM of different proteins is positively related to the protein's abundance, and increases with the protein's size.

Prediction 4: Constraints of cost-minimization on protein structure decrease the evolutionary rate of amino acid replacements. The alternative hypothesis is that structural and functional constraints, rather than economic ones, should limit non-synonymous rates of evolution of protein sequences.

Prediction 5: The same ultimate (adaptive) factors that cause reduction in genome-size increase WM, because reduction of genome-size decreases costs of replication. Hence if the reduction of replication and metabolic costs have common ecological causes, negative correlations of genome-size and WM should frequently be statistically significant.

Prediction 6: Developmental rates at the whole-organism level increase with WM. This should have similar ultimate, but no common proximate causes as the negative correlations described between genome-size and developmental rates (Sessions and Larsson 1987).

Note that predictions involve different levels of organismic organization: from properties of single amino acids (prediction 2); to properties of different proteins within an organism (predictions 3 and 4); to variation among organisms (predictions 1, 5, 6); to association with another, independent trait (genome-size) that is presumably affected by the same causal factors (prediction 5); to associations with whole-organism life-history traits (predictions 1 and 6). Prediction 1 implies that WM results from the life-history of the organism, and prediction 6 implies that WM determines developmental rates. Predictions 5 and 6 seem cumbersome as they imply a 'black box' as to the mechanisms that link WM and differentiation rates. Yet, such associations would hint at links between molecular and whole-organism levels of organization, and indicate that natural selection affects WM via life-history traits. Confirming or infirming predictions 5 and 6 would indicate the limits of the consequences of cost-minimization at the molecular level on whole-organism biology. They should not be considered as predictions in the strictest sense, but rather as justifications for exploring the correlations. Prediction 3 is identical to the one tested by Akashi and Gojobori (2002); however, tests in the present study include non-microbial organisms; use an alternative measure of cost-minimization that does not correlate with GC content, a potential confounding factor; and explore the association of protein size with cost-minimization.

Results

Prediction 1: Cost-Minimization in Different Groups of Organisms

The analysis of Dufton (1997) used a pool of proteins to calculate amino acid frequencies in proteins, without indicating possible differences among taxa in levels of cost-minimization of amino acid composition. Table 1 shows significant variation in mean WM in a pool of different proteins in different groups of organisms. Amino acid usages are from Nakamura et al. (2000). Only organisms for which amino acid usage distributions were derived from more than

Table 1. Mean of correlations coefficients of amino acid frequencies in proteins with the number of synonymous-codons for an amino acid according to the standard genetic code (redundancy), with amino acid molecular weight (WM), and alpha and beta Chou-Fasman indices[a]

Organisms	N	Redundancy	WM	Alpha	Beta
Archaea	22	0.33 (0.14)	0.44 (0.12)	0.33 (0.07)	0.05 (0.17)
Bacteria	256	0.46 (0.13)	0.52 (0.09)	0.38 (0.09)	0.08 (0.14)
Free living Firmicutes	88	0.43 (0.13)	0.55 (0.07)	0.35 (0.09)	0.09 (0.16)
Intracellular Firmicutes	11	**0.23 (0.16)**	0.38 (0.11)	0.32 (0.09)	0.16 (0.16)
Free living Proteobacteria	104	0.51 (0.08)	0.57 (0.06)	0.42 (0.07)	0.09 (0.13)
Intracellular Proteobacteria	17	0.45 (0.16)	**0.52 (0.09)**	0.39 (0.10)	0.09 (0.11)
Eucaryota	310	0.51 (0.12)	0.57 (0.11)	0.27 (0.10)	0.03 (0.15)
Chloroplasts	14	0.50 (0.14)	0.42 (0.14)	0.24 (0.06)	0.21 (0.07)
Mitochondria	30	0.40 (0.16)	0.32 (0.14)	0.52 (0.17)	0.40 (0.09)
ds DNA viruses	96	0.51 (0.17)	0.46 (0.12)	0.26 (0.11)	0.02 (0.13)
ds RNA viruses	14	0.53 (0.13)	0.41 (0.11)	0.21 (0.09)	0.02 (0.08)
ss DNA viruses	9	0.56 (0.16)	0.50 (0.10)	0.06 (0.14)	0.05 (0.14)
(−) ss RNA viruses	26	0.54 (0.11)	0.50 (0.08)	0.21 (0.08)	0.07 (0.18)
(+) ss RNA viruses	38	0.63 (0.14)	0.54 (0.11)	0.18 (0.13)	0.02 (0.13)
Retroid viruses	14	0.55 (0.16)	0.45 (0.10)	0.15 (0.13)	0.02 (0.09)
Total	841	0.49 (0.14)	0.52 (0.12)	0.30 (0.13)	0.06 (0.16)
F		11.48	23.86	48.34	24.26

[a] WM is the negative of the correlation coefficient. Alpha and beta indicate the mean of correlations of the residual abundance of amino acid (after regressing out the effect of amino acid molecular weight) with the alpha and beta Chou-Fasman conformational indices, respectively. Numbers between parentheses are standard deviations. Results are means for N species, for each of which codon frequencies were derived from more than 10000 codons, at http://www.kasuza.or.jp/codon/ (Nakamura et al. 2000). F statistics are for ANOVA among means of z-transformed correlation coefficients from the major groups of organisms. Categories used for ANOVA tests did not include subdivisions within bacteria. Boldface for intracellular bacteria indicates significant differences (one-tailed *t*-tests) with free-living species in the same taxonomic group.

10000 codons were used. WM is compared with Redundancy, the mean correlation coefficient of the number of synonymous-codons for an amino acid (according to the standard genetic code) with the amino acid's respective usage in proteins. This comparison is done because previous studies show that amino acid abundances in proteins increase with the number of synonymous-codons for that amino acid (see for example King and Jukes 1969, and Arquès and Michel 1997). The correlation with codon redundancy was greater than WM in 72% of the 253 intracellular organisms in Table 1, while in free-living organisms, this correlation was greater than WM only in 27% of the organisms. Amino acid molecular weight predicts their abundances to a greater extent than does codon redundancy in Archaea, Bacteria, and Eucaryota. However, in most intracellular organisms (viruses, chloroplasts, and mitochondria), the effect of cost-minimization on amino acid usage is generally weaker than the effect of redundancy. This observation suggests that intracellular organisms minimize costs to a lesser extent than free-living organisms, perhaps because they absorb amino acids from their host rather than synthesize them. Comparisons within bacterial taxonomic groups confirm this result. WM was significantly weaker (*t*-tests, $p < 0.05$) in two independent groups of intracellular bacteria from different lineages (Firmicutes: *Mycoplasma* (Bacillus/Clostridium complex, Entomoplasmataceae), *Rhodococcus* (Actinobacteria), and *Spiroplasma* (Bacillus/Clostridium complex, Spiroplasmataceae); Proteobacteria: *Brucella* (alpha division), *Buchnera* (gamma division), *Burkholderia* (beta division), *Legionella* (gamma division), *Photorhabdus* (gamma division), and *Rickettsia* (alpha division)) than in free-living species from the same taxonomic groups (Table 1). Frequent horizontal gene transfers in bacteria and the associated uncertainties of bacterial phylogenetic relationships make corrections accounting for phylogenetic dependence among taxa difficult and the results not obligatorily more rigorous than the "simple" analysis of the raw data. However, patterns are similar in different lineages, suggesting that one can give some confidence in the conclusion, even without corrections for phylogenetic dependence among taxa. It seems that some characteristics of intracellularity cause WM to be lower in intracellular organisms than in others. This is despite the fact that most intracellular organisms, independently of their phylogenetic relationships, tend to have a low GC content (Heddi et al. 1998). Indeed, Fig. 2 shows that high GC content associates with low biosynthetic costs, so that the confounding effect of low GC content in intracellular bacteria does not explain the low WM observed in these organisms. In intracellular bacteria, low WM is rather in spite of low GC content.

Further testable predictions can be developed on the base of these results: WM in nitrogen-fixing bacteria should be greater than in other, comparable organisms, because nitrogen fixation is a biochemically very costly process. An explorative study fo-

cusing on this point might use an approach similar to the one described here. In this case, one should consider using correlations of amino acid usage with the number or proportion of nitrogen atoms within the amino acid. This approach is similar to the one presented by Baudouin-Cornu et al. (2001), who showed that the atomic composition of nitrogen- and carbon-assimilatory enzymes is depleted in nitrogen and carbon, respectively.

Prediction 2: Protein Function and Amino Acid Weight Minimization

It is likely that economical considerations shaped the amino acid composition of proteins at a later evolutionary stage than constraints governing folding, activity, specificity, and stability of proteins. Amino acid weight minimization probably works within the limits allowed by functional constraints: cost-minimization should affect more strongly those amino acids that are less crucial to protein structure than are others. Amino acids vary in their impact on protein structure, as quantified by Chou-Fasman conformational indices (Chou and Fasman 1978) that quantify the probability that a specific amino acid is at the physical origin of abrupt differences in the protein's three-dimensional structure (alpha helices or beta sheets). Hence the cost-minimization hypothesis predicts that the usage of amino acids with high Chou-Fasman indices should be less cost-minimized than that of amino acids with the same molecular weight, but lower conformational impact. Indeed, correlations of residual frequencies of amino acids (regressing out the effect of molecular weight) and their Chou-Fasman conformational index are positive for the alpha index in 98% of the species used in Table 1 (mean r = 0.30, sd = 0.13) and in 60% of all species for the beta index. For alpha indices, 80% of the exceptions were viruses. This pattern was also found for the beta index. These results confirm that the higher an amino acid's conformational impact, the less likely its usage is to be avoided because of cost-minimization. The effect seems weaker for beta sheets, perhaps because of different abundances of alpha helices and beta sheets in the majority of proteins.

While levels of cost-minimization seem low in all intracellular organisms, the analyses of the correlations of amino acid frequencies with Chou-Fasman indices reveal differences among groups of intracellular organisms: some (viruses) have the least conformational constraint on amino acid use (low correlations with Chou-Fasman indices in Table 1), while amino acid usage in mitochondria and to some extent chloroplasts show among the highest conformational impacts. This difference probably reflects the necessity for highly efficient proteins in these major organelles, possibly the cause of low levels of cost-minimization.

Prediction 3: Cost-Minimization in Different Proteins

Variability in WM might exist between different proteins in the same organism, depending on the protein. I tested the prediction that WM correlates positively with the amount a protein is produced by the cell, and the protein's size. I used as proxy of expression level the optimization of the usage of synonymous-codons in a gene.

Presumably, synonymous-codon optimization increases translational rates and correlates positively with gene expression (in *E. coli*, Gouy and Gautier 1982, Ikemura 1982; in *S. cerevisiae*, Bennetzen and Hall 1982). I defined as optimal the synonymous-codon that is most used over the whole proteome, and estimated codon-usage optimization by the proportion of codons in a protein-coding sequence that are consistent with the synonymous-codons most used genome-wide. Results in six organisms, with the notable and unexplained exception of *Arabidopsis*, confirm the results of Akashi and Gojobori (2002) that cost-minimization increases with synonymous-codon optimization (Pearson correlation coefficients of WM with synonymous-codon optimization, number of proteins, statistical significance, sequence data from GenBank): *E. coli*, r = 0.174, n = 4288, $p = 10^{-30}$; *Halobacterium* sp., r = 0.093, n = 2058, $p = 10^{-5}$; man, r = 0.177, n = 27357, $p = 10^{-192}$; *Saccharomyces cerevisiae*, r = 0.021, n = 6357, $p = 0.0502$; *Drosophila melanogaster*, r = 0.026, n = 13969, $p = 0.001$; and *Arabidopsis thaliana*, r = −0.126, n = 24229, $p = 4 \times 10^{-87}$. One should bear in mind that positive evidence for synonymous-codon optimization as proxy of expression levels exists only for *E. coli*, *Drosophila*, and *Saccharomyces cerevisiae*.

Figure 3 shows that in all organisms, the mean WM of proteins (grouped according to the length of their coding sequence) increases with the protein's size, confirming the prediction that WM increases for large, hence, costly proteins. Associations of protein size and WM might be confounded by associations of synonymous-codon optimization with gene length, because in *Drosophila*, the level of synonymous-codon optimization decreases with the length of the protein's coding sequence (mainly the region proximal to the 5′ end of the mRNA is optimized; Comeron et al. 1999). Figure 4 shows that in *Homo sapiens*, as in *Drosophila*, codon optimization decreases with protein size, but for the four remaining organisms, synonymous-codon-bias optimization increases with size. The causes for that variation are not within the scope of this study, but associations of coding-sequence length and synonymous-codon op-

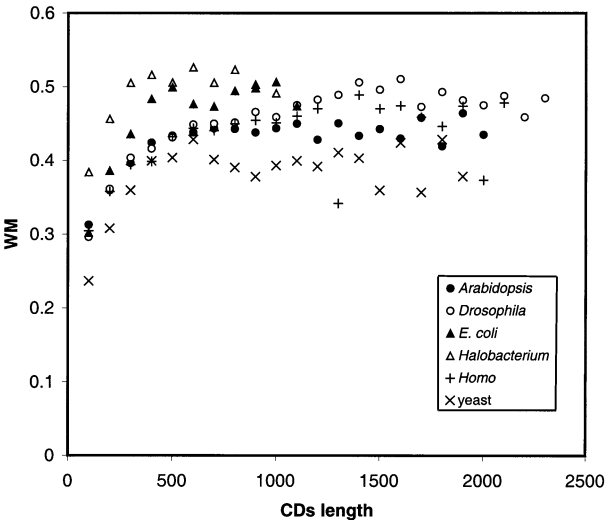

Fig. 3. Mean cost-minimization (WM) in amino acid composition of proteins as a function of the length of their coding sequence in six organisms. Genes were grouped according to their length. Correlations of gene length with mean cost minimization (WM) are: *Arabidopsis thaliana*, r = 0.83; *Drosophila melanogaster*, r = 0.92; *Escherichia coli*, r = 0.90; *Halobacterium*, r = 0.81; *Homo sapiens*, r = 0.68; *Saccharomyces cerevisiae*, r = 0.71.

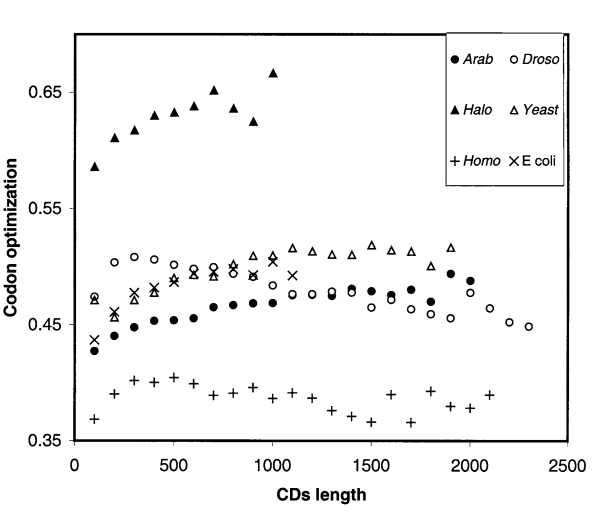

Fig. 4. Mean synonymous-codon optimization of proteins as a function of the length of their coding sequence in six organisms, same data as in Fig. 3. Correlations of gene length with mean synonymous-codon optimization are: *Arabidopsis thaliana*, r = 0.91; *Drosophila melanogaster*, r = −0.86; *Escherichia coli*, r = 0.83; *Halobacterium* sp., r = 0.82; *Homo sapiens*, r = −0.41; *Saccharomyces cerevisiae*, r = 0.84.

timization do not confound the increase of WM with size, because protein size and WM correlate positively in all organisms (Fig. 3), independent of the direction of the association of synonymous-codon optimization and protein size (Fig. 4).

There is an alternative to the hypothesis that the pressure for cost-minimization is proportional to the cost of synthesis of the protein. "Active" sites where the specific properties of the costly amino acids are required represent a smaller proportion of the amino acid's sequence in large than in small proteins. Hence the number of opportunities for decreasing the abundance of costly amino acids without altering the protein's function is proportional to the protein's size, which would also explain the patterns in Figure 3: the same level of selective pressure could achieve greater WM in a larger than a smaller protein. Further tests are required to explore these two possibilities.

The results in this section confirm for estimates of cost minimization that are not confounded by GC content that cost minimization increases with the level of expression of genes and show that, independent of the latter observation, cost-minimization increases with the size of the protein.

Prediction 4: WM Decreases the Rate of Amino Acid Replacements

Constraints limit the number of potential states of a system. In proteins, for example, functional requirements define a limited number of amino acids that can be at a given position in the protein, because other amino acids would alter its structure and enzymatic activity. Indeed, the average alpha Chou-Fasman index of the amino acids in proteins correlates negatively with the rate of amino acid replacement estimated from comparisons between *Drosophila melanogaster* and *D. obscura* (Li 1997) in 31 proteins ($r = -0.59$, $p < 0.01$, one-tailed test, not shown). This result is expected because numbers of potentially neutral replacements decrease for amino acids with high impact on the protein's structure. Cost-minimization of amino acid usage is an adaptive constraint of a different nature. Presumably, cost-minimization should be a secondary constraint that limits the number of replacements, which makes its detection more difficult. In the case of the 31 proteins in *Drosophila*, the correlation of replacement rates with the mean Chou-Fasman index indicates that the latter is a major factor in protein evolution. However, economic constraints may also affect replacement rates. This is suggested by a multiple regression analysis of replacement rates (estimated from the replacements in 40 proteins, occurring between *Mus musculus* and *Homo sapiens* (Li 1997)): the multiple regression of replacement rates (dependent) on the mean alpha Chou-Fasman index of the proteins and their WM is significant ($R^2 = 0.43$, $p = 0.025$). The two independent variables are correlated ($r = -0.38$, $p = 0.018$). In order to assess the relative contribution of each independent, I used partial correlation analysis, a method that calculates the correlation between two variables, adjusting for each variable's covariation with a third (or more) variable(s) (Sokal and Rohlf 1995). Adjusting for the effect of the mean Chou-Fasman index, I found a significant decrease of

Table 2. Correlation coefficients of amino acid weight minimization (WM) in two mitochondrial proteins, cytochrome B and NADH 4, with genome-size (WM-GS), and correlation coefficients of genomic GC contents with genome-size (GC-GS) in five groups of vertebrates[a]

Organisms	WM-GS	GC-GS	P-WM-Gs	P-GC-GS
Fish (CYT B)	−0.45 (14, 0.11)	0.02 (40)	−0.47	0.01
Salamanders (CYT B)	−0.51 (23, 0.013)	0.09 (17)	−0.48	0.02
Salamanders (NADH4)	−0.64 (17, 0.005)	0.09 (17)	−0.28	0.02
Frogs (CYT B)	0.49 (15, 0.06)	0.52 (37)		
Reptiles (CYT B)	−0.95 (13, 0.000)	0.44 (25)	−0.73	0.43
Birds (CYT B)	−0.58 (7, 0.171)	0.41 (8)		
Mammals (CYT B)	0.25 (22, 0.27)	−0.10 (27)	0.04	−0.01

[a] Numbers in parentheses indicate sample sizes followed by p values of correlation coefficients (two-tailed tests). P-WM-GS and P-GC-GS indicate correlation coefficients for phylogenetic independent contrasts. These were never significant at $p < 0.05$, and were not calculated for birds (data for too few species for such an analysis) and frogs (the phylogeny of Ranidae, which contains most of the species used here, is not yet resolved).

replacement rates with WM (partial correlation coefficient $r = -0.27$, $p = 0.026$, one-tailed test). This partial correlation is comparable with the partial correlation between the mean Chou-Fasman index and replacement rates, neutralizing effects of WM (partial correlation coefficient $r = -0.24$, $p = 0.034$). According to these results, in this set of 40 mammal proteins, economic and functional constraints have similar impacts on decreasing replacement rates. These preliminary results in the small sample of proteins above do not take into account the positive association of cost-minimization with gene-expression levels (Akashi and Gojobori 2002), and protein size (results of prediction 3, above). A more detailed study on larger numbers of genes should explore the relative importance of functional and cost-minimization constraints in proteins grouped according to expression levels, size, and function.

Prediction 5: Genome Size and Protein Cost-Minimization

Genome-size correlates with various whole-organism traits in amphibians: it decreases developmental rates (embryonic, Horner and MacGregor 1983; and during limb regeneration, Sessions and Larsson 1987), and with brain histology (Roth et al. 1994). Presumably, the time and the energetic cost of replicating large genomes decrease differentiation rates and rates of development. Replication costs and costs of "normal" cell metabolism are part of the costs of growth and development, hence the ultimate adaptive causes for genome-size reduction should be similar to those increasing WM. However, the proximal processes that cause evolutionary changes in genome-size and in WM differ (for example, polyploidy for genome-size, and point mutations for amino acid replacements). Hence positive associations of genome-size and WM should indicate that changes in genome-size and WM have a common adaptive cause, and strengthen confidence in the interpretation of WM as a measure of protein cost-minimization. In viruses, all groups combined, WM decreases with genome-size ($r = -0.35$, $n = 71$, $p < 0.01$, results not shown): amino acid weight is less minimized in viruses with large genomes than viruses with shorter genomes, suggesting a common cause to decreases in genome-size and increases in WM. In bacteria, the correlation is also significant, but in the opposite direction ($r = 0.52$, $n = 38$, $p < 0.01$), an effect that further analyses showed to be confounded by a GC-content effect, but not by phylogenetic constraints. (I found that GC content and bacterial genome-size correlate positively, also for phylogenetically independent contrasts). Data on genome-sizes from different vertebrate groups are available (Vinogradov 1998; additional data for salamanders from Sessions and Larsson 1987; Pagel and Johnston 1992; for frogs from Fritz et al. 1994; for salamanders and frogs from Roth et al. 1994). Only sequences of two mitochondrial genes (Cytochrome B and NADH 4) were available at GenBank for a sufficient number of vertebrates with known genome-sizes to make correlation analyses possible. Table 2 shows that in 5/6 groups (results were statistically significant at $p < 0.05$ in salamanders and reptiles), genome-size correlates negatively with WM. In salamanders, results were qualitatively similar for cytochrome B and NADH 4. In one group, frogs, the association was positive and close to significant. Considering all groups of organisms, WM and genome-size correlated negatively in 5/8 cases, and correlations were statistically significant at $p < 0.05$ in 50% of the cases (two tailed tests), in 3/5 negative correlations and 1/3 positive correlations. These results suggest that genome-size reduction and WM may have in some cases common ultimate causes (half of the tests are statistically significant, which is more than the 5% significant cases expected by pure chance from the multiplicity of tests). Correlations based on phylogenetic contrasts were not significant. Hence the association of genome-size and cost-minimization seems to result in

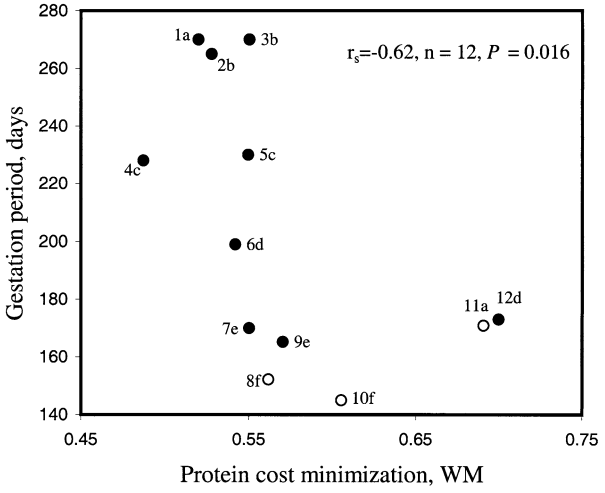

Fig. 5. Primate gestation time as a function of cost-minimization of amino acid protein composition. Gestation times are from Brizzee and Dunlap (1986). Only species for which amino acid protein composition was derived from more than 5000 codons were included (Nakamura et al. 2000). Cost-minimization is quantified by WM, the negative of the correlation coefficient of amino acid molecular weight and its usage. Species are: 1, *Pongo pygmaeus*; 2, *Gorilla gorilla*; 3, *Homo sapiens*; 4, *Pan paniscus*; 5, *Pan troglodytes*; 6, *Cercopithecus aethiops*; 7, *Macaca mulatta*; 8, *Callithrix jacchus*; 9, *Macaca fasciata*; 10, *Saguinus oedipus*; 11, *Saimiri sciureus*;12, *Papio hamadryas*. Circles indicate South American species. Species numbers followed by the same letter form a pair used in the correlation analysis of phylogenetic-contrasts ($r = -0.82$, $n = 6$, $p < 0.01$).

large part from causes that covary with phylogeny. This observation deserves further scrutiny, because overall, the same trends were observed independently in different phylogenetic groups, and that is unlikely to happen because of circumstantial historical constraints that would lead to similar overall patterns in independent lineages. Indeed, Cheverud et al. (1985) remarked that phylogenetic-contrast methods adjust data for variation among species due to common history, and the interaction between common history and species-specific evolution, in a way that only purely species-specific components of variance are analysed. The interaction component, if non-zero, reflects ortho-evolutionary constraints, which imply regulatory mechanisms where the evolution of a trait is in part a function of former evolution in the trait. Such orthogenetic mechanisms seem likely for the regulation of genome-size. Hence standard methods to account for phylogenetic constraints are not adequate to test the hypothesis, and no clear conclusion can be drawn from the reported analyses. Overall, results suggest that the same evolutionary factor(s) might cause the associations of genome-size and WM. Hence selective pressures on replication rates, for example, would cause organisms to 'cut' costs of genome replication, and the same ecological constraints are likely to select for cost-minimization of metabolism, including protein synthesis, resulting in weak negative correlations between genome-size and WM.

GC Content as a Confounding Factor?

Genome-size correlates positively with the genomic GC content, and genomic GC content associates with biases in codon-usage and amino acid composition (Bernardi and Bernardi 1986). Hence the correlation of WM and genome-size might be indirect, due to triangular relations with GC content. This hypothesis does not account for the cases where I observed negative correlations of genome-size with WM, because results in Table 2 show that in 5/6 groups, correlation coefficients of GC with genome-size were lower than those reported for WM and genome-size. Genomic GC contents are an unlikely confounding factor also because Fig. 2 shows that amino acids coded by codons with high GC content are relatively 'cheap'. Accordingly, high GC content is confounded with high levels of cost-minimization. Yet in most correlations in the previous section, especially the significant ones, WM is greater in organisms with small genomes than in those with large ones: the opposite would have been expected if GC contents were a confounding factor, because GC content correlates positively with genome-size. This situation where the correlations described are obtained in spite of the known trends existing between genome-size and GC content, and not because of these, is similar to the one described for comparisons of free-living and intracellular organisms (prediction 1).

Does Protein-Cost-Minimization Affect Higher Levels of Organization of Whole-Organisms?

Genome-size associates with decrease in the rate of development in amphibians (Sessions and Larsson 1987; Pagel and Rufus 1992) and with increase in body size in copepods and flatworms (Gregory et al. 2000). The associations of genome-size with WM reported in the previous section suggest that developmental rates might correlate with WM. This would be because WM decreases the costs of both components of development, growth and differentiation, by minimizing costs of protein synthesis. Indeed, preliminary results show that WM in cytochrome B and in NADH 4 correlates negatively with rates of differentiation (days until hatching, and regeneration rates of hindlimbs in salamanders, Jockusch 1997). However, genome-size is likely to be a confounding variable in amphibians, because a negative correlation exists between differentiation rate and genome-size (Sessions and Larsson, 1987). I hence focused on groups where genome-size varies much less than within Amphibia. Gestation periods of primates (here considered as inversely proportional to developmen-

tal rates) correlate negatively with WM (Spearman rank correlation coefficient, $r_s = -0.62$, $n = 12$, $p = 0.016$; Fig. 5). The correlation for the phylogenetic contrasts was also significant, so that phyletic constraints do not account for the trend shown in Fig. 5. The variation among primate species in WM might result from biases in gene sampling for the different species, because a pool of different genes was used to derive amino acid abundances in different species. Figure 5 includes only species for which amino acid usage was determined from more than 5000 codons. Restricting this choice to those species for which at least 10000 codons were available (decreasing effects of sampling biases), excludes species numbered 4, 10, 11, and 12 from the analysis, most of which seem to be outliers in Fig. 5. This procedure contracts by more than half the range of variation of WM. However, the qualitative result that the gestation period correlates negatively with WM remains similar ($r_s = -0.66$, $n = 8$, $p = 0.038$), and does not result from outliers or gene-sample bias. Although such a correlation fits the bioenergetic cost-minimization hypothesis, this positive result seems unlikely to be due to the presumed mechanisms, because of the long chain of unknown factors between protein synthesis and organogenesis. For example, brain size, even more than neonate body size, correlates with gestation time (Sacher and Staffeldt 1974). However, even after controlling for the effects of brain size on gestation period, a statistical treatment that accounts for 80% of the variation in gestation periods, the correlation of residual gestation period with WM remains negative and is almost significant (Spearman rank correlation coefficient, $r_s = -0.44$, $p = 0.066$, one-tailed test). In rodents, I found that cost-minimization of cytochrome B and length of gestation correlate negatively ($r_s = -0.25$, $n = 125$, $p = 0.003$). Accounting for covariation of gestation period and neonate size in those rodents for which estimates of neonate weight was available, the correlation remains significant ($r_s = -0.30$, $n = 40$, $p = 0.03$). For time until hatching in *Drosophila* spp. (data from Ashburner 1989), results were qualitatively similar but not statistically significant ($r_s = -0.28$, $n = 25$, $p = 0.097$). These results are unconclusive, especially that I did not test for phyletic constraints in rodents, *Drosophila* and the reduced primate dataset after removing outliers from the analyses, but the overall trend suggests that cost-minimization of biosynthesis is part of the r- versus K-strategy syndrome. It is likely that the processes that link molecular cost-minimization and estimates of metabolic rates are more complex than assumed by the amino acid usage cost-minimization hypothesis. However, the results also indicate that this ultra-reductionist approach might bear some insights at whole-organism levels.

Random Mutations Do Not Account for Cost-Minimization

Previous studies show that the properties of the genetic code are such that nucleotide substitution rates tend, on average, to minimize the distance between the chemical properties of replaced and replacing amino acids (Grantham 1974). Gojobori et al. (1982) showed that this effect is stronger for nucleotide-substitution patterns observed in functional genes than in non-functional ones. Some studies even suggest that the genetic code's codon-amino acid assignments are optimized in this respect, because the existing genetic code minimizes some measures of physico-chemical distances between replaced and replacing amino acids more than random sets of genetic codes (Freeland et al. 2000). Cost-minimization of amino acid usage might result from a similar phenomenon: amino acid replacements, as a result of the existing substitution patterns, could lead, in average, to the replacement of heavier amino acids by lighter ones. My analyses, using the substitution patterns described for mammalian pseudogenes (Li et al. 1984) and those described for non-coding regions of *Drosophila* genomes (Bergman and Kreitman 2001) suggest the opposite: under these constraints of substitution frequencies, there is a weak, non-significant tendency for the molecular weight of replacing amino acids to be heavier than the replaced amino acids. This suggests that amino acid cost-minimization occurs in spite of spontaneous substitution rates, rather than because of these, and strengthens the view that natural selection is involved in cost-minimization.

Conclusions

I present positive evidence that supports the hypothesis that organisms minimize costs of protein synthesis by avoiding the usage of heavy amino acids. Levels of cost-minimization are lower (1) in intracellular organisms than in free-living organisms; (2) for amino acids with high structural impact on the protein than for those with the same weight but lower structural impact; and (3) in genes with low expression levels than in those with higher ones, and in short genes as compared to long ones. (4) Among different proteins, cost-minimization of amino acid usage decreases evolutionary rates of amino acid replacements. (5) Genome-size frequently correlates negatively with levels of protein cost-minimization, suggesting a common adaptive (ultimate) cause to both independent processes. Presumably, cost-minimization of protein synthesis is an important component of whole-organism metabolic strategies. Results show that cost-minimization as observed in the amino acid composition of single proteins (such as cytochrome B) may reflect the metabolic-ecologi-

cal strategy of the whole-organism. Although I consider that the semi-mechanistic hypothesis of protein cost-minimization is incorrect, or at least too simplistic, I cannot dismiss the power of the hypothesis in predicting observed trends.

Acknowledgments. Discussions of results with Leigh Van Valen and Marty Kreitman contributed to the elaboration of the ideas presented and of the manuscript itself. Prediction 4 was the original idea of Marcos Antezana, who helped me to get the extensive data analyzed in Fig. 3 and 4.

References

Aguilera M (1985) Growth and reproduction in *Zygodontomys microtinus* (Rodentia, Cricetidae) from Venezuela in a laboratory colony. Mammalia 49:75–83

Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. PNAS 99:3695–3700

Arquès DG, Michel CJ (1997) A code in the protein-coding genes. Biosystems 44:107–134

Ashburner M (1989) *Drosophila*. Cold Spring Harbor, New York

Barrai I, Volinia S, Scapoli C (1995) The usage of oligopeptides in proteins correlates negatively with molecular weight. Int J Peptide Protein Res 45:326–331

Baudouin-Cornu P, Surdin-Kerjan Y, Marlière P, Thomas D (2001) Molecular evolution of protein atomic composition. Science 293:297

Begall S, Burda H, Gallardo MH (1999) Reproduction, postnatal development, and growth of social coruros, *Spalacopus cyanus* (Rodentia: Octodontidae), from Chile. J Mammal 80:210–217

Bennett NC, Jarvis JUM (1988a) The social structure and reproductive biology of colonies of the mole-rat, *Cryptomys damarensis* (Rodentia, Bathyergidae). J Mammal 69:293–302

Bennett NC, Jarvis JUM (1988b) The reproductive biology of the Cape mole-rat, *Georychus capensis* (Rodentia, Bathyergidae). J Zool Lond 214:95–106

Bennett NC, Jarvis JUM, Cotterill FPD (1994) The colony structure and reproductive biology of the afrotropical Mashona mole-rat, *Cryptomys darlingi*. J Zool Lond 234:477–487

Bennetzen JL, Hall BD (1982) Codon selection in S*accharomyces cerevisiae*. J Biol Chem 257:3026–3031

Bergmann CM, Kreitman M (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res 11:1335–1345

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Blumstein DT, Armitage KB (1998) Life-history consequences of social complexity: a comparative study of ground-dwelling sciurids. Behav Ecol 9:8–19

Brizzee KE, Dunlap WP (1986) Growth. In: Dukelow WR, Erwin J (eds) Comparative primate biology. Reproduction and development, Vol III. AR Liss, New York, pp 363–413

Bryant SL, Rose RW (1989) Growth and role of the corpus luteum throughout delayed gestation in the potoroo, *Potorous tridactylus*. J Reprod Fert 76:409–414

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: Sexual dimorphism in body weight among primates. Evolution 39:1335–1351

Chou PY, Fasman GD (1978) Empirical predictions of protein conformation. Ann Rev Biochem 47:251–276

Comeron JM, Kreitman M, Aguadè M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151:239–249

Dufton MJ (1997) Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? J Theor Biol 187:165–173

Duplantier JM, Granjon L, Bouganaly H (1996) Reproductive characteristics of three sympatric species of *Mastomys* in Senegal, as observed in the field and in captivity. Mammalia 60:629–638

Fersht AR (1986) The charging of tRNA. In: Kirkwood TBL, Rosenberger RF, Galas DJ (eds) Accuracy in molecular processes. Chapman & Hall, New York, pp 159–189

Freeland SJ, Hurst LD (1998) Load minimization of the genetic code: history does not explain the pattern. Proc R Soc Lond B 265:2111–2119

Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000) Early fixation of an optimal genetic code. Mol Biol Evol 17:511–518

Fritz B, Vences M, Glaw F (1994) Comparative DNA content in *Discoglossus* (Amphibia, Anura, Discoglossidae). Zool Anzeiger 233:135–145

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Gouy M, Gautier C (1982) Codon-usage in bacteria: correlation with gene expressivity. Nucl Acids Res 10:7055–7074

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Gregory TR, Hebert PDN, Kolasa J (2000) Evolutionary implications of the relationship between genome-size and body size in flatworms and copepods. Heredity 84:201–208

Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P (1998) Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: A peculiar G + C content of an endocytobiotic DNA. J Mol Evol 47:52–61

Hoogland JL (1997) Duration of gestation and lactation for Gunnison's prairie dogs. J Mammal 78:173–180

Horner HA, MacGregor HC (1983) *C* value and cell volume: their significance in the evolution and development of amphibians. J Cell Sci 63:135–146

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

Innes DGL, Millar JS (1944) Life histories of *Clethrionomys* and *Microtus* (Microtinae). Mammal Rev 24:179–207

Jockusch E (1997) An evolutionary correlate of genome-size change in plethodontid salamanders. Proc R Soc Lond B 264:597–604

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Mazel D, Marlière P (1989) Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature 341:245–248

Moscarella RA, Aguilera M (1999) Growth and reproduction of *Oryzomys albigularis* (Rodentia: Sigmodontinae) under laboratory studies. Mammalia 63:349–362

Nakamura Y, Gojobori T, Ikemura, T (2000) Codon-usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Res 28:292

Neal BR (1990) Observations on the early post-natal growth and development of *Tatera leucogaster*, *Aethomys chrysophilus* and *A. namaquensis* from Zimbabwe, with a review of the pre- and

post-natal growth and development of African muroid rodents. Mammalia 54:245–270

Orr RT (1970) Development: Prenatal and postnatal. In: Wimsatt WM (ed) Biology of bats, vol I. Academic Press, New York, p 404

Pagel M, Johnston RA (1992) Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. Proc R Soc Lond B 249:119–124

Roth G, Blanke J, Wake DB (1994) Cell size predicts morphological complexity in the brains of frogs and salamanders. PNAS 91:4796–4800

Sacher GA, Staffeldt EF (1974) Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth. Amer Nat 108:593–615

Scharff A, Begall S, Gruetjen O, Burda H (1999) Reproductive characteristics and growth of Zambian giant mole-rat, *Cryptomys mechowi* (Rodentia: Bathyergidae). Mammalia 63:217–230

Schmid M (1980) Chromosome banding in amphibia. 5. Highly differentiated ZW-ZZ sex-chromosomes and exceptional genome-size in *Pyxicephalus adspersus* (Anura, Ranidae). Chromosoma 80:69–96

Sessions SK, Larsson A (1987) Developmental correlates of genome-size in plethodontid salamanders and their implications for genome evolution. Evolution 41:1239–1251

Sokal RR, Rohlf FJ (1995) Biometry. Third ed. Freeman & Co. New York

Thompson SD (1987) Body size, duration of parental care, and the intrinsic rate of natural increase in eutherian and methaterian mammals. Oecologia 71:201–209

Vinogradov AE (1998) Genome-size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. Cytometry 31:100–109

Waterman JN (1996) Reproductive biology of a tropical, non-hibernating ground squirrel. J Mammal 77:134–146