æ

16. THE COSTS OF CELLULAR FEATURES

15 December 2018

Although it is common to view all cellular features as products of adaptive processes, it should now be clear that this starting point is generally no more than an assumption grown out of a tradition in which natural selection is viewed as the only evolutionary force of significance. Studies of the adaptive value of alternative manifestations of a trait can be pursued by measuring the fitness of individuals with different phenotypic values (Walsh and Lynch 2018). However, it is one thing to identify optimum phenotypic values, but quite another to determine the overall consequences of developing and exhibiting the trait at all. Virtually all structural features and functions of cells require energetic expenditures for construction and maintenance, and the total benefit of a trait needs to be considered in the light of this baseline cost.

A cellular feature that appears to be adaptive in the context of its current lineage of occupancy need not have endowed any net benefits relative to the ancestor from which it first emerged. If the overt benefit of trait expression on growth, reproduction, and survival is insufficiently greater than the baseline price, then it is plausible that the establishment of some cellular features has proceeded on a nearly neutral basis, possibly guided to a degree by mutation pressure. To understand whether such scenarios are possible, and further whether mildly deleterious cellular features can emerge in certain population-genetic contexts, it is essential to have a quantifiable measure of the cost of cellular attributes. In principle, this cost would be measured as the decline in cell fitness were the trait to be expressed while conveying no benefits. In reality, such a measure is nearly impossible for many well-established traits for the simple reason that once integrated into critical cellular pathways, a trait will have indirect side effects accommodated over evolutionary time.

Thus, we require an indirect way of summarizing the baseline costs of simply expressing and maintaining a trait that does not require invasive manipulation of the cell. At the heart of this subject are two central concepts in biology: 1) the costs of building and maintaining cellular features, and any resultant benefits in terms of biomass production; and 2) the degree to which such costs and benefits translate into fitness differences influencing long-term evolutionary trajectories. Here, an attempt is made to provide a quantitative framework for addressing these issues.

The starting point is an outline of the conceptual issues involving this fundamental link between cell biology and evolution. This will then be followed by applications to several key features exhibiting a substantial gradient in complexity across the Tree of Life, most notably the expansion of gene number and gene-architectural complexity and of the investment in membrane-bound organelles in the eukaryotic

domain. These analyses will show that such elaborations need not have been driven by positive selection for cellular complexity, but rather may be inevitable passive consequences of the diminished efficiency of selection resulting from a reduction in effective population size. Although such a scenario may leave the impression of the cumulative development of a long-term fitness drag on a population, this need not be the case. Embellishments incorporated into the genomic/cellular real estate by nonadaptive mechanisms can also serve to open up alternative pathways to future adaptive evolution by descent with modification.

The Bioenergetic Cost of a Cellular Feature

We start with the cost of a simple cellular feature, e.g., a noncoding RNA, a protein molecule or complex, or the membrane of a cellular inclusion. The basic principle is that, regardless of their fitness benefits, all such features entail some baseline costs of construction and maintenance. To this end, we desire a universal currency by which costs can be measured for a wide variety of traits in ways that generalize across all phylogenetic lineages, and there seems to be little alternative than to rely on a measure of energy utilization. In the fields of ecology and traditional evolutionary biology, the maximization of energy flow through biomass production has long been viewed as a target of natural selection (e.g., Lotka 1922; Van Valen 1976, 1980). The simple reason for this is that selection favors successful processes of survival and offspring production, all of which of which require energy.

Although analyses might be carried out with alternative limiting factors (e.g., carbon or some other key nutrient), most elemental measures are of restricted value as the nature of a limiting nutrient can vary among phylogenetic lineages and even within species growing in different environments. For example, silicon availability can be critical to diatom growth, but nearly irrelevant to most other organisms. Likewise, nitrogen will rarely be limiting to a species capable of nitrogen fixation. In contrast, regardless of the substrates and elemental building blocks being consumed, their entry into all aspects of cellular maintenance and biomass production requires energy.

There remains the question as to the logical energetic currency to be used, but here there seems to be little alternative than the consumption of ATP molecules. Hydrolysis of ATP to ADP is universally deployed across the Tree of Life in the vast majority of energy-requiring processes, generally even involving orthologous pathways such as the TCA cycle, modes of amino-acid biosynthesis, and mechanisms of assembly of cellular polymers (e.g., DNA, RNA, and protein). Some processes involve the hydrolysis of other nucleotides such as CTP and GTP, but the energy release here can still be counted in terms of ATP equivalents, as can the use of coenzymes such as NADH and NADPH in electron-transport chains to produce ATP. Thus, in keeping with previous efforts in microbiology (Bauchop and Elsden 1960; Atkinson 1970; Stouthamer 1973; Tempest and Neijssel 1984; Russell and Cook 1995), all energetic costs described below will be defined in units equivalent to numbers of ATP hydrolyses. Additional justification for this approach is that although the yields of microbes grown on alternative substrates (per unit carbon consumed) can vary substantially with the nature of the substrate, once the appropriate en-

ergetic conversions are made, the number of ATP hydrolyses necessary to build an offspring cell are found to be relatively constant (Chapter 3).

The direct cost of a trait can be subdivided into three component costs of: 1) biosynthesis of the basic building blocks; 2) assembling of the components into the full structure; and 3) maintenance (Figure 16.1). First, with respect to biosynthesis, nearly all cellular features are assembled from four types of monomeric building blocks: amino acids, nucleotides, lipid molecules, and carbohydrates. If not provided by the outside environment, such molecules must be synthesized within the cell by processes requiring carbon skeletons and energetic expenditures (derived, for example, from transformations of glucose or acetate to precursor metabolites such as pyruvate and acetyl CoA). When monomeric building blocks are available externally (as is the case for predators), the reliance on *de novo* biosynthesis will be diminished. However, there will still be costs of resource acquisition and sometimes of transformation of the precursors (the branch-point metabolites) in various ways. Second, the assembly cost of a cellular feature is comprised of the summed requirements for construction from its monomeric building blocks. For example, protein assembly requires polymerization of the constituent amino acids, addition of post-translational modifications, and folding the amino-acid chain into the appropriate globular form. Finally, there will often be maintenance costs, e.g., accommodation of molecular turnover, and identification and elimination of cumulative errors.

The sum of costs noted above represents the baseline investment that must be made in a cellular feature regardless of its benefit to the host cell. Given the near universality of many biosynthetic pathways and enzyme-reaction mechanisms, all three sets of costs can often be calculated from information in cell biological literature. Such an approach is a highly desirable alternative to laborious *in vivo* cell biological methods, such as modifications of gene-expression levels, as these can have additional side effects (e.g., promiscuous binding or aggregation) that are irrelevant to construction/maintenance costs.

These direct-cost components of a trait do not fully describe the consequences of a trait's presence for the cell. Even if the trait under consideration pays for itself by endowing the cell with increased fitness (in excess of the direct baseline costs), the construction and maintenance of the trait will still impose a drain on resources that could have been allocated to other essential cellular functions. When metabolic precursors that are generally processed for ATP production and/or utilized as carbon skeletons are instead allocated to the production/maintenance of a focal trait, the loss of availability for other purposes represents an opportunity cost (Figure 16.1). This follows from the fact that the production of an additional building block (or entire trait) adds to a cell's lifetime energy budget requirements, while subtracting from the prior pool of resources available for other functions, which must be replaced (generally by extending the cell cycle). Atkinson (1970) defined opportunity costs as the "prices of metabolites," and seemingly independent of him, Craig and Weber (1998) and Akashi and Gojobori (2002) used this approach to partition the costs of amino acids into components associated with the investment in synthesis and the utilization of metabolites.

Summing up, the total cost of synthesizing and maintaining a trait and diverting components from alternative usage to do so is the sum of the direct and opportunity

costs,

$$c_T = c_D + c_O, \tag{16.1}$$

where all costs represent the cumulative expenditures over the entire lifespan of the cell. Whereas c_D is expected to reflect actual ATP hydrolysis reactions resulting in heat dissipation in the cell, c_O will not be manifested in heat production, given that no ATP is produced or consumed. Nonetheless, c_O should be included in evolutionary analyses except in cases where energy is a nonlimiting resource. If the energy extracted from the food source is in excess supply relative to some other key elemental resource (e.g., nitrogen, phosphorus, or iron), c_O is irrelevant to the fitness of the cell, reducing the energetic cost from a fitness perspective to $c_T \simeq c_D$.

A relevant observation with respect to this latter point exists for microbes growing on a defined medium with all carbon and energy being provided by a single compound and all other nutrients being in excess supply. In this case, the growth yield per carbon consumed increases linearly with the substrate heat of combustion (which is inversely related to the degree of oxidation) up until a threshold value, thereafter leveling off (Figure 3.11). This suggests that below a critical value of $\simeq 10$ kcal/g carbon, growth is limited by energy, whereas above this threshold the food supply contains excess energy relative to carbon content required for growth. Notably, the most common substrate used in growth experiments with microbes, glucose, has a heat of combustion of 9.3 kcal/g carbon, close to the threshold at which growth is equally limited by carbon and energy. Very few commonly used substrates have heats of combustion much beyond the apparent threshold (values being 11.0, 13.6, and 14.8 for glycerol, ethanol, and methanol, respectively).

The Evolutionary Cost of a Cellular Feature

Given a measure of the cost of a cellular feature, the overall implications depend on the cell's lifetime energy requirements. From a cell physiological perspective, the absolute energetic cost of the feature must be scaled relative to the total cost of building and maintaining the cell (Chapter 4), and evolutionary considerations further require that this relative measure be converted to an appropriately scaled metric of the impact on fitness.

Supposing the cell has a baseline total energy budget per cell cycle of C_T (which includes the costs of both growth and maintenance, with a capital C denoting a whole-cell cost), the presence of the trait under consideration alters the lifetime energy budget such $C_T = C_T' + c_T$, C_T' is the total energy budget in the absence of the trait. This additional investment in the trait is expected to increase the cell-division time (ignoring for the time being any direct advantages conferred by the trait), owing to the additional amount of resource acquisition necessary to complete the cell cycle.

Understanding the fitness consequences of the energetic investment in a trait requires a definition of the cell's reproductive rate relative to that for a cell without such an excess investment, and this can be shown to be nearly equal to the ratio c_T/C_T (Foundations 16.1). More specifically, the baseline fitness effects of building and maintaining a trait leads to a negative selection coefficient $\simeq -\ln(2)(c_T/C_T)$, with $\ln(2) \simeq 0.69$ simply being a scaling factor relating to the continuous-growth

scale.

One caveat with respect to this definition is that it assumes that the addition of the trait does not somehow alter the cell's basic metabolic makeup in ways that would modify the total baseline energy budget C_T. However, even if this does occur, the result given in Foundations 16.1 will be only slightly modified if the fractional alteration to C_T is small, which seems likely for cellular modifications involving just one or two genes. In some instances, there may be additional costly side effects.For example, a novel protein might promiscuously interact with inappropriate substrates, aggregate with other cellular components, and/or excessively occupy cellular volume or membrane real estate.

These additional issues notwithstanding, the general point remains – there will always be some baseline cost of expressing a trait, and the net benefit of a trait needs to be derived by subtracting the construction/maintenance costs from the direct benefits accrued from increased survival and/or reproduction (Figure 16.2). If a trait is to be maintained by natural selection, the baseline costs must be sufficiently small that the net benefit s_n is greater than the power of random genetic drift ($1/N_e$ for a haploid, and $1/2N_e$ for a diploid population). When this is not the case, an existing trait will generally be vulnerable to loss by degenerative mutations. It also follows that in the domain of effective neutrality, mutation pressure alone can lead to the slow accumulation of modifications that impose a net drain on a cell's energetic capacity, as will be discussed below in the context of alterations to genome architecture and gene structure in various eukaryotic lineages.

Foundations 16.1. The relationship of bioenergetic costs to the strength of selection. Understanding the fitness consequences of the total energetic investment in a trait requires a definition of its baseline (construction plus maintenance) effects on the cell's reproductive rate. The selective disadvantage is defined as

$$s = r - r', \tag{16.1.1a}$$

where $r = \ln(2)/\tau$ denotes an exponential rate of growth, τ is the mean cell division time (Chapter 5), and r' and r denote the growth rates in the presence and absence of the attribute under consideration. Denoting the increase in cell-division time as $\Delta_\tau = (\tau' - \tau)$,

$$s \simeq \frac{\ln(2) \cdot \Delta_\tau}{\tau}. \tag{16.1.1b}$$

Further assuming that the cost of the trait is much less than the lifetime energetic expenditure of a cell, $c_T \ll C_T$, so that the increment in cell-division time scales linearly with the proportional increase in investment,

$$\tau' \simeq \tau \left(1 + \frac{c_T}{C_T}\right), \tag{16.1.2}$$

leads to

$$s \simeq \frac{\ln(2) \cdot c_T}{C_T} \tag{16.1.3}$$

showing that the intrinsic selective disadvantage associated with the bioenergetic cost of a trait scales directly with the proportional increase in the total energy demand per cell cycle.

The Energetic Cost of a Gene

As a first application of the preceding ideas, consider the total cost of a maintaining and operating a gene, which involves up to three levels of investment. Even for an unexpressed gene, there are DNA-level costs in terms of replication and chromosome maintenance (which in eukaryotes, include the cost of nucleosomes around which the DNA is wrapped). Transcription and transcript processing imposes additional costs, and for protein-coding genes, there are still more costs in terms of amino-acid biosynthesis and polypeptide processing. The sum of costs at these one to three levels constitute the total cost of any genomic sequence.

Individual genes face a "use it or lose it" challenge. If the net fitness advantage of a gene is smaller than the power of random genetic drift, it will be vulnerable to eventual, passive loss by the accumulation of degenerative mutations in an effectively neutral fashion, and if there is a strong enough net selective disadvantage, removal will be accelerated by directional selection. On the other hand, excess and even nonfunctional DNA can often accumulate in genomes by insertion mechanisms, as in the case of the expansion of mobile-genetic elements, and mutation pressure alone can facilitate such genomic expansion provided the cost of the excess genomic material is smaller than the power of random genetic drift.

One of the many mysteries of genome evolution concerns the number of genes contained within genomes and the mechanisms responsible for the lineage-specific expansions of such numbers in eukaryotes, especially in multicellular species (Lynch 2007a; Lynch et al. 2011). The genomes of most prokaryotes contain 5000 or fewer protein-coding genes, whereas those in most eukaryotes exceed 10,000 genes, with the genomes of most multicellular eukaryotes (metazoans and land plants) containing 15,000 to 30,000 genes. Most significantly, the expansion of total genome sizes from prokaryotes (generally 1 to 10 million base pairs) to unicellular eukaryotes (generally of order 10 to 100 million base pairs) to multicellular eukaryotes (hundreds to thousands of million base pairs) is much less a consequence of an increase in gene number than of the proliferation of introns, mobile-genetic elements, and other sorts of genomic insertions. Most prokaryotic genomes are highly streamlined, typically containing $< 5\%$ intergenic DNA and devoid of introns and mobile-genetic elements, whereas the genomes of multicellular eukaryotes tend to be highly bloated, often containing $< 5\%$ coding DNA and harboring massive numbers of large introns, mobile elements, and other forms of DNA insertions.

This raises the question as to whether expansions in genome size and gene-architectural complexity are driven by adaptive processes or simply inevitable consequences of the increased power of random genetic drift in organisms of larger size. A common view is that that there is an intrinsic advantage to both cellular complexity and multicellularity, but such a stance is nothing more than an assumption (Lynch 2007b; Booth and Doolittle 2015), presumably fostered by the fact that one multicellular species has come to dominate the biological and intellectual world.

There is no direct evidence that what we regard as complexity has been directly promoted by natural selection. If the advancement of complexity is a goal of selection, given that all extant organisms are temporally equidistant from the last universal common ancestor,the more astounding observation is the extreme phylogenetic rarity of multicellularity involving large numbers of cell types (only being represented by two eukaryotic lineages, metazoans and land plants).

To help explain the gradient from extreme streamlining of genomes in prokaryotes to the extraordinary bloating of genomes in metazoans and land plants, we now draw from a wide variety of observations from cell biology and biochemistry to evaluate how expensive a gene (or segment of DNA) is from an energetic perspective. Through its phenotypic manifestations, a gene may have a multiplicity of advantages, but the energetic costs associated with replication, maintenance, and expression represent a minimum burden that must be overcome to achieve a net selective advantage large enough to ensure gene survival (Figure 16.2). The following subsections present first-order approximations of the cumulative cost of a gene, subdividing the total into expenses at the genomic, transcriptional, and protein levels, and evaluating how the selective consequences of such costs vary among organisms. The consideration is the cost of production of the basic monomeric building blocks that comprise DNA, RNA, and protein molecules.

Biosynthetic costs of nucleotides and amino acids. Although there are many energetic costs associated with assembly, processing, and maintenance of DNA, RNA, and protein, the primary costs of a gene are associated with the synthesis of the monomeric building blocks of the polymeric chains. Because the basic biosynthetic pathways for nucleotides and amino acids are highly conserved across the Tree of Life (Chapter 18), it is relatively straight-forward to compute the costs of these basic units. As discussed above, such costs will be quantified in units of ATP usage, specifically relying on the number of phosphorus atoms released via ATP hydrolyses, the primary source of energy in most endergonic cellular reactions. CTP and GTP, which are utilized in a few reaction steps in lipid biosynthesis, will be treated as equivalent to ATP, and electron transfers resulting from conversions of conenzyme NADH to $NADH^+$, NADPH to $NADPH^+$, and $FADH_2$ to FAD can be expressed in ATP equivalents using conventions in biochemistry based on energetic analyses.

As an entrée into the logic of estimating the cost of a basic building block, consider the situation for the adenine nucleotide. All nucleotide synthesis involves the use of the precursor PRPP (phosphoribosyl pyrophosphate, an early derivative from the pentose phosphate pathway) and amino-acid molecules, which we assume must first also be synthesized, and there are additional investments of ATP in various steps in the conversion of the precursors to nucleotides. Assuming a starting point of glucose, the diversion of PRPP deprives the cell of an additional potential nine ATP hydrolyses and ten hydrogen ions (from the electron carriers NADH and NADPH, which we assume to be equivalent to ~ 2 ATP hydrolyses each, as the latter would result from the utilization of the former in the electron-transport chain). This leads to a total opportunity loss equivalent to ~ 29 ATPs per utilized PRPP molecule. One glycine molecule is also consumed, the production of which results in an opportunity loss of ~ 12 ATPs (below). There are ten steps in the production of the final

precursor (IMP) to the purine nucleotides (A and G) from PRPP and glycine, and these in total consume the equivalent of seven ATPs. The final conversion of IMP to dATP consumes an additional three ATPs, so the total opportunity and direct costs for ATP production are 41 and 10 ATPs, respectively. Conversion to dGTP consumes just a single additional ATP.

Pyrimidine (C and T) production also starts with PRPP, but consumes an aspartate molecule with an opportunity cost of 13 ATPs, yielding an opportunity cost of 42 ATPs. The route to the final pyrimidine precursor UMP consumes an additional three ATPs. Conversion to dTTP requires the equivalent of an additional four ATPs, which yields a total direct cost of seven ATPs. dGTP production consumes just one additional ATP. Ribonucleotides have total costs that are just three ATPs less than those for the corresponding dNTPs. Thus, taken together, these results indicate that the direct and opportunity costs of building each ribonucleotide and deoxyribonucleotide can be taken to be 40 and 10 ATPs as a first approximation. Very similar estimates were made by Akashi and Gojobori (2002) and Wagner (2005).

Similar logic can be used to estimate the costs of amino acids, the details of which can be found elsewhere (Atkinson 1970; McDermitt and Loomis 1981; Craig and Weber 1998; van Milgern 2002; Arnold et al. 2015), with slight variances in the estimates among studies resulting from differences in assumptions regarding the conversion of coenzyme reactions to ATP (Barton et al. 2010). The central point again is that amino-acid biosynthetic pathways initiate with various metabolic derivatives that could otherwise be broken down to generate energy. The average opportunity and direct costs per amino acid are ~ 25 and 2 ATPs, so the synthesis of an average amino acid increases a cell's energy budget by roughly half that induced by average nucleotide production.

Contrary to the situation with the four nucleotides, all of which have nearly identical biosynthetic costs, the total costs for the twenty different amino acids have a nearly eight-fold range, from 10 ATPs for aspartic acid to 76 for tryptophan (Table 11.1), and these are strongly correlated with the molecular weights of the individual residues (Seligmann et al. 2003). It has been suggested that the differences in bioenergetic costs among amino acids may be sufficiently high to be perceived by natural selection (Akashi and Gojobori 2002), and averaged across the Tree of Life, there is indeed a nearly 30-fold reduction in the use of the most expensive relative to the cheapest amino acids (Krick et al. 2014). However, such differences might be due to mutation bias or simple functional constraints, as opposed to selection based on energy demands. For the latter to be sufficiently strong to drive differences in amino-acid utilization, the cost differential between alternative amino acids at single codon sites must be of sufficient magnitude for selection to overcome background random forces due to drift.

Why the focus on amino acids at single sites rather than the full genome content? In principle, selection might operate on differences in the amino-acid contents of stretches of sequence longer than single codons, but for this to be effective, multiple codons would have to be segregating amino-acid variants simultaneously within a population, and recombination among sites would have to be sufficiently rare to retain the composite properties of aggregates of selected mutations prior to fixation. Given that most populations have levels of heterozygosity at silent sites < 0.02,

that such levels are typically on the order of 25% smaller for amino-acid replacement sites, and that $\sim 75\%$ of nucleotide sites within a gene are replacement sites (Chapters 8 and 11), then for a gene of average length (~ 1000 bp), the number of amino-acid polymorphisms segregating simultaneously will generally be fewer than 5, many of which by chance will involve amino acids with small cost differences. Moreover, because the rate of recombination per nucleotide site is roughly equal to that of mutation (Chapter 8), substantial recombination may decouple the majority of segregating amino-acid substitutions. Thus, although there is need for theoretical work in this area, it is prudent to first consider whether natural selection is capable of perceiving energetic-cost differences among amino acids on a single-codon basis. As just noted, the maximum cost differential at a particular amino-acid site is $(76 - 10) = 66$ ATP.

Consider first the situation for an *E. coli*-sized cell with volume 1 μm^3, which requires $\sim 3 \times 10^{10}$ ATP hydrolyses to build (Chapter 4). The mean number of protein molecules associated with an average gene in a cell of this size is ~ 1700, with almost all genes falling in the range of 10 to 10^4 protein copies/cell (Chapter 3). Thus, relative to the total cell budget, the maximum energetic impact of the substitution of a single amino-acid residue (for a highly expressed gene) is on the order of $(66 \times 10^4)/(3 \times 10^{10}) \simeq 2.2 \times 10^{-5}$. As this quantity times $\ln(2)$ translates into a selection coefficient (Foundations 16.1), then recalling that selection is effective in a haploid population if $2N_e s > 1$ (Chapter 8), the effective population size N_e need only exceed $30,000$ for selection to promote this extreme an amino-acid change in a highly expressed gene. For a gene with just 10 protein copies per cell, the critical N_e to 3×10^7, which is near the typical N_e for microbial populations.

This rough analysis suggests that the selective promotion of amino acids with low biosynthetic costs (assuming they do not compromise protein function) can be quite effective in prokaryotes, although with diminishing strength in lowly expressed genes. Consistent with these arguments, Akashi and Gojobori (2002) find that the average cost per amino-acid residue is inversely correlated with gene-expression levels in *E. coli* and *B. subtilis,* with a decline of nearly 10% average residue costs over the full range of gene-expression levels. A similar $\sim 10\%$ reduction of average amino-acid cost over the range of expression level has been observed for proteins in other bacterial species (Heizer et al. 2006; Raiford et al. 2012). Thus, at least in bacteria, amino-acid substitutions that may be neutral with respect to function may nonetheless be advanced via their relative metabolic demands. Notably, amino-acid sites containing expensive residues in *E. coli* and yeast tend to evolve more slowly than those harbored by the cheapest residues, suggesting that in these species energetically costly residues are primarily relied upon for key structural or functional reasons (Seligmann et al. 2003; Barton et al. 2010).

Now consider a yeast-sized cell, ~ 100 μm^3 in volume, with a range of 100 to 10^6 protein copies per gene (Chapter 3). Given the near linear scaling of lifetime cellular energy budgets with cell volume (Chapter 4), the latter is $\sim 100\times$ greater than that for a typical bacterium, but so is the upper limit for protein number, so the upper limit to the relative cellular expense is the same as for *E. coli.* The lower limit is $\sim 10\times$ smaller. Thus, the critical N_e values become 3×10^4 to 3×10^8, with the latter being near the upper bound of what is seen in unicellular eukaryotes (Chapter 8). Similar calculations for a metazoan cell size of ~ 1000 μm^3 in volume yield a range

of critical N_e values of 3×10^4 to 3×10^9 for the most highly to most lowly expressed genes. The latter critical point is orders of magnitude greater than what is observed in multicellular species.

Consistent with these disparities, indirect analyses suggest that selection may promote energetically cheap amino acids in highly expressed genes in unicellular eukaryotes, and perhaps even in some genes in multicellular eukaryotes (Swire 2007; Heizer et al. 2011). However, the overall effect is substantially lower than that in bacteria – only a 1% reduction in the average amino-acid cost in highly expressed genes in *S. cerevisiae* (Raiford et al. 2008), and 3% in the flour beetle *Tribolium castaneum* (Williford and Demuth 2012).

These observations suggest that in eukaryotes, the efficiency of selection for usage of energetically cheap amino acids can still be quite high in the most highly expressed genes, but approaches effective neutrality for a large fraction of genes with lower expression, and increasingly so in organisms that are larger in size. However, all of these computations have been carried out with the most extreme bioenergetic cost difference between amino acids. It is worth noting that all but three amino acids have ATP costs in the ranges of either 10 to 16 or 22 to 33 ATPs (Table 11.1). Amino-acid cost differences on the order of 7 ATPs (as opposed to 66) require effective population sizes to be 10-fold higher than those noted above for selection to be efficient, so for species with large cell sizes, this greatly reduces the likelihood of selective promotion of amino acids based on their energetic demands.

One final issue of concern here involves the nature of amino-acid acquisition. Although most (and in some cases all) amino acids are synthesized within the cell starting from other organic compounds, many cases exist in which one or more amino acids can not be synthesized and must be acquired directly from food sources. Yet the same pattern of declining usage of expensive amino acids even occurs in microbes that are auxotrophic (unable to synthesize) such residues (Swire 2007; Raiford et al. 2012). Although there is no cost of direct biosynthesis of directly utilizable amino acids, Swire (2007) makes the argument that there is still an opportunity cost – an amino acid taken up by a heterotroph can either be directly incorporated into a protein or degraded to produce ATP that can be utilized in other cellular processes. Thus, given that the amount of energy extracted from the breakdown of an amino acid is about the same as the energy for building one, that most of the total cost of amino-acid biosynthesis involves opportunity loss, and that energy must be expended for amino-acid uptake, the differences in cost for directly acquired vs. internally synthesized amino acids may be relatively minor.

Chromosome-associated costs. We now consider the cost of harboring a segment of DNA, regardless of its expression level. Genome replication requires the synthesis of two new DNA strands from each parental double-helix DNA molecule. Taking the total cost per nucleotide to be 50 ATPs, the cost of replicating a gene L_n nucleotides in length is then $2 \cdot 50 \cdot L_n = 100 L_n$ P. There are numerous additional costs of a gene at the DNA level, a major one being the unwinding of the parental double helix, which requires ~ 1 ATP per nucleotide. All other replication-related costs – associated with opening of origins of replication, clamp loading, proofreading, production of the RNA primers used for replicate-strand extension, ligation of Okazaki fragments, and DNA repair – are an order of magnitude or so smaller than those just noted,

and can be ignored for purposes herein.

There is, however, one additional major chromosome-level cost specific to eukaryotes – the highly ordered, dense coverage of nucleosomes, each of which contains two heterotetrameric histone complexes followed by a linker histone. Throughout eukaryotes, each nucleosome wraps ~ 147 bp, and with an average linker length between nucleosomes of 33 bp, there is on average one nucleosome per 180-bp interval. Weighting by the cost of synthesizing the amino acids that comprise histone proteins and the cost of translating such proteins (below), the total nucleosome-associated cost is $\sim 160L_n$ ATPs (Lynch and Marinov 2015).

Taking all of the above issues into consideration, the chromosome-level cost of a bacterial DNA segment (in units of ATP hydrolyses) is

$$c_{\mathrm{DNA},b} \simeq 100L_n, \tag{16.2a}$$

whereas for a haploid eukaryote,

$$c_{\mathrm{DNA},h} \simeq 260L_n, \tag{16.2b}$$

and doubling the preceding cost for a diploid eukaryote yields

$$c_{\mathrm{DNA},d} \simeq 520L_n. \tag{16.2c}$$

These results provide a quantitative basis for understanding the evolutionary mechanisms underlying the dramatic differences in gene structure and genomic architecture between prokaryotes and eukaryotes. Because replication is essentially a one-time investment in the life of a cell, whereas a longer cell-division time increases the total basal metabolic requirement and hence the total lifetime energy requirements of a cell, the maximum fractional contribution of the DNA-level cost of a gene to a cell's total energy budget, c_{DNA}/C_T, occurs at minimum cell-division times.

Ignoring time-dependent costs of maintaining a cell (as we are attempting to determine maximum relative chromosome-level costs, and hence desire the minimum possible denominator), a prokaryotic cell with a representative volume of 1 μm^3 has a maximum replication-associated cost of DNA $\simeq (4 \times 10^{-9})L_n$, so the fractional drain on the total cellular energy budget can be as high as (4×10^{-8}) for a small 10-bp insertion, and (4×10^{-6}) for a gene-sized insertion of 1000 bp. Thus, because free-living prokaryotes typically have effective population sizes $\sim 10^8$ (Chapter 8), when growing at maximum rates, prokaryotes experience efficient enough selection to remove insertions as small as 10 bp (and even 1-bp when N_e approaches the apparent upper bound of 10^9).

In contrast, for a unicellular eukaryote with a moderate-sized 100 μm^3 cell containing a haploid genome (e.g., a yeast), the fractional cost of DNA is $\simeq 10^{-10}L_n$, yielding maximum relative chromosome-level costs of 10^{-9} and 10^{-7} for 10- and 1000-bp segments of DNA, respectively. Because unicellular eukaryotes often have $N_e < 10^8$, sometimes ranging down to 10^6, these results imply that insertions of small to moderate size will frequently be unmovable by natural selection in such species. For a larger cell size of 2500 μm^3, more typical of a multicellular eukaryote, and a diploid genome, the relative cost of DNA declines to $\simeq 10^{-11}L_n$, so even a 10^5-bp segment of DNA has a relative cost of just 10^{-6}. The effective population sizes of

invertebrates tend to be in the neighborhood of 10^6, with that of some vertebrates (including humans) ranging down to 10^4.

Thus, even though the DNA-level cost of a DNA insertion in a diploid multicellular eukaryote is $\sim 5\times$ that in a prokaryote, the disparity in total cellular energy budgets is so much greater that the power of random genetic drift in eukaryotes is sufficient to overwhelm the ability of selection to prevent the accumulation of quite large insertions on the basis of DNA-level costs. These results provide a clear mechanistic explanation for the highly streamlined genomes of prokaryotes relative to eukaryotes.

As outlined in Foundations 16.2, however, there is an another cost of excess DNA, unassociated with bioenergetics – all excess DNA, even when nonfunctional, is mutationally dangerous in that it increases the substrate for mutations to gene malfunction. On a per-nucleotide basis, this too is typically a weak evolutionary cost, strong enough to be perceived in many microbes but often effectively neutral in eukaryotes (multicellular species in particular).

Foundations 16.2. The mutational hazard of excess DNA. All genes have a mutational target size equal to the number of nucleotide sites (at the DNA level) for which the nucleotide identity has the potential to influence fitness (Lynch 2007a). This will include most amino-acid replacement sites in the coding region, and to a much lesser (but not always insignificant) extent silent sites (for which the nucleotide identity has no impact on the encoded amino acid, but may influence the rate of translation). Here, we consider the impact of slightly larger segments of gene-associated noncoding DNA that are relevant to successful gene expression. These include introns, which commonly populate eukaryotic genes, a variety of transcription-factor binding sites, and numerous other regulatory sequences in the $5'$ untranslated regions (UTRs) of transcripts.

What is the contribution of such elements for the vulnerability of a gene to mutational inactivation? Introns are transcribed into pre-mRNAs, and must be properly spliced out to yield a productive mature mRNA, with accurate recognition by the splicing machinery depending on motifs at both ends comprising a total $\sim$ 20 to 30 nucleotides. Transcription-factor binding sites typically consists of motifs 8 to 16 bases in length, and other elements in UTRs are usually of this same size. Thus, to simplify discussion, we will consider an embellishment to a gene that magnifies the mutational target size by 10 nucleotides. Notably, even entirely nonfunctional DNA can magnify the vulnerability to deleterious mutations, as such material can acquire detrimental gain-of-function mutations, although the magnitude of such effects is difficult to quantify.

The mutation rate per nucleotide site per generation ranges from $\sim 10^{-10}$ in prokaryotes and many unicellular eukaryotes to 10^{-8} in vertebrates, so the addition of a gene-structural embellishment that introduces 10 nucleotides whose identity is critical to gene function is equivalent to increasing the mutation rate to null alleles by 10^{-9} to 10^{-7}. This type of mutation-rate inflation for an embellished gene operates in a manner effectively identical to selection, as it is a measure of the excess rate of removal of such alleles from the population by conversion to null alleles.

How do these mutational hazards compare with the energetic costs of nucleotides? Returning to the results in the text, a 10-bp segment of DNA imposes an energetic penalty (relative to the total cost of building a cell) of $\sim 10^{-8}$ in a typical bacterium, $\sim 10^{-9}$ in a unicellular eukaryote, and $\sim 10^{-10}$ in a multicellular eukaryote. These

rough calculations suggest that for bacteria, the primary selective disadvantage of excess DNA is associated with energetic costs (this, at most, being of the same order of magnitude as the mutational hazard). In contrast, the mutational cost starts to exceed the energetic penalty in unicellular eukaryotes, and greatly exceeds it in multicellular species. In the latter case, however, even the mutational cost of a gene-structural embellishment is insufficient to overcome the power of random genetic drift. Hence, from both the perspectives of energetics and genetics, we expect a gradient in genome size and gene-structural complexity from prokaryotes to multicellular eukaryotes, provided there is a mutational bias towards insertions.

Transcription-associated costs. The costs of transcription are numerous, and although not all of them can be quantified with certainty, the major contributors are well understood. Thus, it is possible to achieve order-of-magnitude estimates of the investments required to produce individual transcripts, and again we will adhere to the costs summarized in Lynch and Marinov (2015).

Before proceeding, it must be understood that because transcripts are typically degraded (and must be replaced) on time scales much shorter than cell-division intervals, the total cost of transcription depends on the lifespan of a cell. If we consider a cell containing an average number of transcripts N_r at birth and a degradation rate of δ_r per transcript, during its lifetime, a cell must produce N_r additional surviving transcripts to create a daughter cell at the same steady-state level, and during the cell-division time T an additional $\delta_r N_r T/\ln(2)$ replacement molecules must be produced to offset molecular degradation (Foundations 16.3). Here, it will be assumed that the set of N_r transcripts necessary to provision the equivalent of a daughter cell require *de novo* synthesis of nucleotides, and that the remaining $\delta_r N_r T/\ln(2)$ transcripts are produced via recycled nucleotides.

Several forms of transcription-associated costs are general across prokaryotes and eukaryotes, but only two of these are quantitatively significant enough to be of concern here. The primary investment is the synthesis of ribonucleotides, which as noted above entails an average cost of ∼50 ATP per ribonucleotide. The total cost of *de novo* ribonucleotide synthesis associated with a gene with transcript length L_r is then $50N_rL_r$. The second major cost involves the replacement of degraded transcripts within the lifespan of the cell, and here the total expenditurre is simply taken to be the two ATPs that must be expended per nucleotide for each chain-elongation step, which leads to a cost of $2\delta_r N_r L_r T/\ln(2)$.

A third cost associated with helix unwinding is $< 5\%$ of that associated with ribonucleotide recycling; and a fourth cost is associated with aborted transcripts (as not all transcription-initiation events lead to completed transcripts), but because such events generally occur within the first ten nucleotides, this cost is even smaller than that for helix unwinding. Still smaller is the cost of activating and initiating transcription. Thus, to a close approximation, the sum of the two predominant costs of transcription, expenditure on ribonucleotide synthesis and chain elongation, closely approximate the total cost of transcribing a gene in the lifetime of a bacterial cell (in units of ATP),

$$c_{\mathrm{RNA},b} \simeq 2N_rL_r(25 + 1.4\delta_r T). \tag{16.3a}$$

Several additional energy-consuming features of transcription are incurred by

eukaryotes, but only two of them are quantitatively relevant here. First, eukaryotic mRNAs are terminated by poly(A) tails, which are added co-transcriptionally with initial lengths of ~ 250 nucleotides. Taking into consideration the full costs of As necessary for the standing pool of mRNAs and the two ATPs per bp necessary for chain elongation associated with excess degraded transcripts, the total cost of poly(A) tails is $\sim 250 N_r[50 + 2(\delta_r T/\ln(2))]$. Second, as noted above, eukaryotic DNA is populated by regularly spaced nucleosomes, and in order for RNA polymerases to proceed, the DNA must be unwrapped from nucleosomes, and this and other related processing entails a total energetic cost of $\sim 0.17 N_r L_r \delta_r T/\ln(2) \simeq 0.25 N_r L_r \delta_r T$. For intron-containing genes, there is a small additional cost of splicing, and there are also costs associated with transcript termination, $5'$ mRNA capping, phosphorylation cycles associated with the RNA polymerase II, and nuclear export, but all of these are quite small relative to the two costs noted above. Summing the eukaryotic-specific components with Equation (1a), the total cost associated with transcription for a eukaryotic gene is

$$c_{\mathrm{RNA},e} \simeq N_r[(11,500 + 50L_r) + (720 + 3L_r)\delta_r T], \tag{16.3b}$$

where L_r is the length of the primary transcript (before splicing). Note that in Equations (16.3a,b) the total cost associated with transcription is subdivided into two components, the first defining the baseline requirement for building a cell, and the second being a linear function of the cell-division time.

Observations from single-cell methodologies provide quantitative insight into some of the key parameters in these formulations, demonstrating in particular that average standing numbers of transcripts per gene are generally quite small (Chapter 3). For example, for *E. coli,* the average number of transcripts/gene/cell (N_r) is 5.0 (with a range of 0 to 100 among genes) (Lu et al. 2007; Li and Xie 2011). The mean N_r is 10 per gene in *S. cerevisiae* (Lu et al. 2007; Zenklusen et al. 2008), and the median is ~ 20 in mammalian cells (Islam et al. 2011; Schwanhäusser et al. 2011; Marinov et al. 2014). In all cases, there is a broad distribution around the mean, so that genes with numbers of transcripts deviating 10-fold from the mean are not uncommon (Golding et al. 2005; Raj et al. 2006; Taniguchi et al. 2010; Csárdi et al. 2015).

Estimates of transcript-decay rates available suggest that the half-lives of mRNAs are typically much shorter than cell-division times. In *E. coli,* $\sim 80\%$ of mRNAs have decay rates (δ_r) in the range of 7 to 20/hour, with a median of 12/hour (Bernstein et al. 2002; Taniguchi et al. 2010). The vast majority of mRNAs in *Bacillus subtilis* have decay rates > 6/hour, with a median of 8/hour (Hambraeus et al. 2002); and in *Lactococcus lactis,* mean and median mRNA decay rates are in the range of 3 to 7/hour, decreasing with decreasing cellular growth rates (Dressaire et al. 2013). For eukaryotes, median mRNA decay rates range from 3 to 6/hour in *S. cerevisiae* (Wang et al. 2002; Neymotin et al. 2014), and average 0.1/hour in mouse fibroblast cells (Schwanhäusser et al. 2011). These results suggest to a first-order approximation that $\delta_r T$ is generally within the range of 10 to 100, which from Equations (16.3a,b) further implies that rapid mRNA decay typically inflates the total cost of transcription by a factor 2 to $6\times$ that expected on the basis of the steady-state number of transcripts per cell, $50 N_r L_r$ and $N_r[(11,500 + 50L_r)]$, for bacteria and eukaryotes, respectively.

By comparison with the preceding results for chromosome-level costs, it can be seen that the costs at the level of transcription will often be several fold higher. Considering microbes for example, and noting that the length of a transcript is very close to the length of a gene ($L_n \simeq L_r$), the ratio of Equations (16.31) and (16.2a), $N_r(0.5 + 0.028\delta_r T)$, shows that the ratio of these two costs always exceeds 1.0, provided the steady-state number of transcripts is 2.0, and can be several-fold higher for genes with higher expression levels and/or multiple transcript half-lives per cell cycle.

For the yeast *S. cerevisiae,* $> 95\%$ of genes are intron-free, and the remaining 5% contain only a single small intron, so as a first approximation, it can again be assumed that genes and transcripts have essentially the same lengths $L_n \simeq L_r = 2000$ bp. As noted above, the mean number of mRNAs per gene per cell is $\overline{N}_r \sim 10$, and with an average decay rate of 4.5/hour and a doubling time of ~ 1.5 hours under optimal growth conditions $\delta_r T \simeq 7$. From Equation (16.3b), the cost of transcription for a typical yeast gene is then on the order of $10 \cdot \{(11,500+100,000)+[(720+6,000)\cdot 7]\} \simeq 2 \times 10^6$ ATPs. By contrast, from Equation (16.2b), the chromosome-level cost of a gene in this species is 0.5×10^6 ATPs.

Finally, we consider the situation for a typical human gene, where the median of number of mRNAs per gene is ~ 20, and the average mRNA decay rate is $\sim$ 1.4/day. Assuming a cell- division time of one day and an average primary transcript length of 47.0 kb, and ignoring the small contribution from the cost of splicing ~ 7 introns per gene, the cost of transcription per gene is then on the order of $20 \cdot \{(11,500 + 2,350,000) + [(720 + 141,000) \cdot 1.4]\} \simeq 5 \times 10^7$ ATPs. Total gene lengths are difficult to define in metazoans, but a 50% inflation relative to the pre-mRNA ($\sim 70,000$ bases) is not unreasonable. Equation (16.2c) then implies a typical cost at the chromosome level of 4×10^7 ATPs. Taken together, all of these results suggest that transcription-associated costs are typically of the same order of magnitude as those at the chromosome level, although the former can greatly exceed the latter for highly expressed genes.

Foundations 16.3. Numbers of molecules required in a cellular lifespan. From the time of birth to the time of cell division, for any particular cellular feature, a cell must accumulate new constituent molecules to a level consistent with the birth of a new cell, as well as acquire replacement molecules to balance any decay processes. Here, we consider the situation in which a newborn cell contains N_0 molecules of the type being considered (e.g., the number of transcripts or protein molecules associated with a particular gene, or the number of lipid molecules in a cell membrane), which then must double to $2N_0$ molecules to allow for binary fission. Letting β be the rate of production of the molecule and δ be the rate of decay, so that $r = \beta - \delta$ is the net growth rate in cell size, then assuming exponential growth

$$\frac{dN}{dt} = rN \tag{16.3.1}$$

denotes the rate of increase in cell size. This expression integrates to

$$N_t = N_0 e^{rt}, \tag{16.3.2}$$

so with the cell dividing when $N_t = 2N_0$, the cell-division time is $T = \ln(2)/r$.

The average number of molecules in the parental cell over its entire lifespan is

$$\overline{N} = \frac{N_0}{T}\int_0^T e^{rt}\cdot dt = \frac{N_0(e^{rT}-1)}{rT} = \frac{N_0}{\ln(2)}, \tag{16.3.3}$$

with the final simplification following from the fact that $rT = \ln(2)$. The total number of molecules produced per cellular lifespan (N_p) is then the product of the average number of molecules ($\overline{N}$), the production rate per molecule (β), and the cell-division time (T),

$$N_p = \frac{\beta N_0 T}{\ln(2)} \tag{16.3.4}$$

which by using $\beta = r + \delta$ and $T = \ln(2)/r$ becomes

$$N_p = N_0\left(1 + \frac{\delta T}{\ln(2)}\right). \tag{16.3.5}$$

A simple interpretation of this expression is that during its lifespan, a cell must produce N_0 new molecules (to yield an offspring cell) and $\delta N_0 T/\ln(2)$ replacement molecules to offset molecular degradation/loss. Note that this second (maintenance) term increases linearly with the cell-division time.

Translation-associated costs. The conceptual approach employed in the preceding section can be extended to the protein level by again assuming that the full cost of production of the steady state number of proteins must be covered by *de novo* synthesis of amino acids, with the excess number of molecules needed to compensate for protein decay being acquired from salvaged amino acids.

Although several subcategories of costs underlie protein production and subsequent management, the overwhelming contributions are associated with just three functions, the biochemical details of which follow the logic noted in the preceding section (Lynch and Marinov 2015). First, the cost associated with replacement of the standing level of protein derived from a particular gene is $N_p L_p \overline{c}_{\rm AA}$, where N_p is the number of protein molecules per newborn cell, L_p is the number of amino acids per protein, and $\overline{c}_{\rm AA}$ is the average cost of synthesis per amino-acid residue (assumed to be equivalent to 27 ATP hydrolyses, based on results given above). Second, the total cost associated with chain elongation of all proteins produced during the cell cycle is $4N_pL_p[1 + (\delta_p T/\ln(2))]$, where δ_p is the rate of protein decay (two ATPs are required for activating the cognate tRNA, one for transferring the tRNA to the ribosome, and one for the movement of the ribosome to an adjacent mRNA triplet). Third, degradation of proteins imposes an approximate cost of $N_pL_p\delta_pT/\ln(2)$ ATP hydrolyses. Additional costs small enough be ignored are associated with translation initiation and termination, post-translational modification, and protein folding. Summing up the three primary expenses, the total protein-level cost of a gene in both bacteria and eukaryotes is

$$c_{\rm PRO} \simeq N_pL_p(31 + 7\delta_pT), \tag{16.4}$$

where again the first term represents a one-time cost incurred regardless of the length of the cell cycle, and the second term grows linearly with the cell-division time owing to the cumulative costs of protein turnover and replacement.

Some insight into the relative magnitudes of the two terms in Equation 16.4 can be obtained from results from high-throughput proteomics. Most notably, the decay rates of proteins are typically much lower than those of their cognate mRNAs. In the bacterium *Lactococcus lactis,* the vast majority of protein decay rates are in the range of 0.04 to 6.0/hour, with the median being 0.1 to 0.9/hour depending on the growth rate (Lahtvee et al. 2014), and those for other bacteria are commonly in the range of 0.05 to 0.20/hour (Trötschel et al. 2013). In *S. cerevisiae,* the median and mean decay rate is $\sim 1.4/$ hour, with most values for individual proteins falling in the range of 0.2 to 5.5/hour under optimal growth conditions (Belle et al. 2006), and the median declining to 0.1/hour in nutrient limiting conditions (Shahrezaei and Swain 2008; Helbig et al. 2011). In mouse fibroblast cells, the median decay rate of a protein is ~ 0.02/hour (with a range of 0.002 to 0.3/hour) (Schwanhäusser et al. 2011), and in a human cancer cell line, the range is from 0.04 to 1.3/hour (Eden et al. 2011). Given the known division times for the cell types noted above, $\delta_p T$ will generally be on the order of 1.0 or smaller, and seldom greater than 10. This implies that the second term in Equation (16.4), the cost of protein degradation, will generally be of the same order of magnitude of the *de novo* amino-acid and protein synthesis or smaller.

Cellular abundances of proteins (N_p) are much higher than those for their cognate mRNAs, with the average ratio of the two per gene being 450 in *E. coli,* 5100 in *S. cerevisiae,* and 2800 in mammalian fibroblasts (Ghaemmaghami et al. 2003; Lu et al. 2007; Schwanhäusser et al. 2011). How do these numbers translate into the total protein-level cost of a gene? To keep the computations simple but still accurate to a first-order approximation, it will be assumed here that $7\delta_p T \simeq 9$, so that $c_{\rm PRO} \simeq 40 N_p L_p$. In *E. coli,* the average number of proteins per gene is ~ 2250, and the average protein contains ~ 300 residues, implying an average $c_{\rm PRO}$ on the order of 3×10^7 ATP/protein-coding gene. By comparison, the average chromosome-level cost is $\sim 10^5$, and as noted above, the average transcription-associated cost is only a few fold greater. Thus, the vast majority of the energetic cost of a protein-coding gene in bacteria is associated with translation. In the case of *S. cerevisiae,* there is an average of $\sim 50,000$ proteins per genetic locus per cell, and the average protein length is $\sim 50\%$ greater than in *E. coli,* yielding an average total cost of translation of $\sim 8 \times 10^8$ ATP per protein-coding gene, or approximately two orders-of-magnitude greater than the summed costs at the chromosome and transcription levels.

Evolutionary implications. The full slate of data necessary to estimate the total cost of a gene are only available for a few species (Lynch and Marinov 2015), but these are fully in accord with the hypothesis that baseline selective consequences (s_c) of such costs tend to exceed the power of random genetic drift in microbes and then progressively become smaller than the power of genetic drift in larger eukaryotic species (Figure 16.3).

For the bacterium *E. coli,* s_c for almost all genes falls in the range of 10^{-6} to 10^{-3}, far above the likely minimum values that can be perceived by selection in this large-N_e species (Figure 16.3). If such genes were to find themselves in an

environment where their functions were no longer useful, inactivating mutations would be strongly selected for. Within eukaryotes, small peaks of lowly expressed genes exist with roughly the same absolute costs of *E. coli* genes. However, owing to the increased total cellular energy budgets, s_c for many eukaryotic genes falls below 10^{-6} and in some cases to as low as 10^{-9}. This is significant because of the reduction in the effective population sizes of such species, which increases the power of random genetic drift. For the majority of genes in the eukaryotic species *S. cerevisiae, C. elegans* and *A. thaliana,* the costs at the chromosomal and transcriptional levels are below or near the power of random genetic drift, implying that without translation most gene sized insertions will be essentially invisible to the eyes of selection. The major contribution that pushes s_c of some genes past the drift barrier in eukaryotes is the cost of translation.

A more general analysis over a large number of species indicates that the total cost of a gene declines with increasing cell size across the Tree of Life (Figure 16.4). Average estimates of all three cost measures in bacteria are generally substantially greater than those in eukaryotes, although there is continuity in the scaling between groups, and in most cases are likely large enough to be opposed by selection. In addition, as noted above, there is a consistent ranking of $s_{\mathrm{DNA}} < s_{\mathrm{RNA}} < s_{\mathrm{PRO}}$, with a one to two order of magnitude increase from the former to the latter.

Taken together, these results suggest that by reducing the contribution of single genes to a cell's total energy budget, evolutionary increases in cell size promote a shift in the selective environment such that gene-sized insertions in eukaryotes, particularly in multicellular species, will typically be effectively neutral from a bioenergetic perspective. This effect works in concert with the increased power of random genetic drift experienced with increases in organismal size. The energetic cost of a DNA segment of even just a few nucleotides (even if nontranscribed) can be sufficient to be perceived by selection in a typical bacterial population with N_e on the order of 10^8. In contrast, insertions of even thousands of kb often impose a small enough energetic burden relative to the overall requirements of eukaryotic cells to be immune to selection.

Although costs at the RNA level are frequently greater than those of at the DNA level, these are often still not large enough to overcome the power of random genetic drift in eukaryotic cells. This means that many nonfunctional DNAs that are inadvertently transcribed in eukaryotes (especially in multicellular species) cannot be opposed by selection. On the other hand, with the cost at the protein level generally being much greater than that at the RNA level, segments of DNA that are translated can sometimes impose a large enough energetic cost to be susceptible to selection, even in multicellular species. This may explain why redundant duplicate genes commonly experience high rates of nonfunctionalization (Chapter 10).

These observations are quite relevant to Lane and Martin's (2010) hypothesis that an enhanced ability to generate energy made possible by the origin of the mitochondrion was a prerequisite for the evolution of increased gene numbers, protein lengths, protein folds, protein-protein interactions, and regulatory elements in eukaryotic cells. As already noted in Chapter 4, there is no dichotomous break in the size-scaling of the metabolic properties in prokaryotic vs. eukaryotic cells, and here we see that increased cell size does not induce a condition in which gene addition becomes an increasing energetic burden, but quite the contrary. From an evolution-

ary perspective, the evolution of increased cell size has the opposite effect. Although the absolute cost of a gene does increase with cell size, in terms of the fractional contribution to a cell's energy budget, which ultimately determines whether selection can oppose genome expansion, the cost of an average gene decreases at the DNA, RNA, and protein levels. Thus, population-genetic arguments based on both the mutational-hazard hypothesis (Foundations 16.2) and on the observed features of cellular energetics lead to the conclusion that passive increases in genome size are expected to naturally arise in organisms with increased cell sizes (which, by correlation, have reduced effective population sizes). This supports the view that variation in the power of random genetic drift has played a central role in the historical diversification of genome and possibly cellular architecture across the Tree of Life.

Finally, it should be noted that genes may have costs beyond those noted above. For example, there may be opportunity costs with respect to transcription and translation, as RNA polymerases, tRNAs, and ribosomes must be deployed in gene expression, reducing their availability in servicing other genes. Aggregation of proteins, associated with misfolding, can reduce the operation of key cellular functions, etc. Experimental work suggests, however, that the predominant cost of genes is associated with the biosynthesis of the elemental building blocks rather than with toxic problems associated with harmful misfolding and protein-protein misinteractions (Stoebel et al. 2008; Plata et al. 2010), with the quantitative effects being in reasonable accord with the numbers cited above (Tomala and Korona 2013; Adler et al. 2014). Nonetheless, high gratuitous expression of proteins can leading to significant misfolding problems that induce secondary biosynthesis costs associated with the up-regulation of chaperones (Geiler-Samerotte et al. 2011; Frumkin et al. 2017).

Before proceeding, a brief return to the issue of element- vs. energy-based currencies in evaluating constraints on evolutionary processes is warranted. A number of authors have suggested that the elemental composition of nucleotides and amino acids in different phylogenetic lineages is driven by environmental conditions, most notably nitrogen availability (Acquisti et al. 2009; Elser et al. 2011; Grzymski and Dussaq 2012). The arguments for such repatterning are based on the fact that at the nucleotide level, purines contain five nitrogen atoms and pyrimidines three, whereas seven of the amino acids (arginine, aspargine, glutamine, histidine, lysine, proline, and tryptophan) contain nitrogens in their side chains, whereas the remaining 13 do not. Bragg and Hyder (2004) further note that the genetic code is structured in such a way that the nitrogen content of codons is correlated with that of the encoded amino acids, so there might e a reinforcing effect at the level of nucleotides and amino acids.

There are two key uncertainties here. First, given the limitations on the ability of natural selection to perceive the energetic differences of single nucleotides / amino-acid residues, it appears quite unlikely that discrimination based on the presence of just one or two atoms is possible. A more likely explanation for any environmental association with the nitrogen contents of transcriptomes and proteomes is that other evolutionary forces drive genome-wide nucleotide composition, with species assorting into environments that are most compatible with elemental demands. One such force is mutational bias, which plays a strong role in guiding nucleotide content of genomes across the Tree of Life, often operating in a direction in conflict with that

of natural selection (Long et al. 2017). Notably absent from studies that invoke elemental limitation as a selective force operating on genome/proteome content are formal evolutionary analyses. In one of the only attempts to do so, Günther et al. (2012) found that high-frequency derived alleles in *Arabidopsis* populations have elevated nitrogen content, contrary to expectations for a plant expected to be under nitrogen limitation.

The Cost of Lipids and Membranes

The major disparities in cellular structure between prokaryotes and eukaryotes involve internal membrane-bound organelles in the latter, with functions including sequestration and gated access to the genomic material, vesicle transport of a multiplicity of cargoes, power production, and platforms for molecular assembly. As outlined in Chapter 14, numerous features of these complex membrane systems appear to have evolved by repeated rounds of gene duplication and divergence. Given that the rate of *de novo* gene duplication in prokaryotes is comparable to that in eukaryotes (Lynch 2007), and that some prokaryotes with organelles and internal membranes do exist, why is the typical internal layout of almost all prokaryotes devoid of membranes? One possibility is that the evolution of internal membranes is the null state, driven by mutational bias and basic biophysical forces (Ramadas and Thattai 2013; Mani and Thattai 2016), but that the construction and maintenance of such embellishments is energetically expensive enough that incremental changes are thwarted by the efficiency of natural selection in prokaryotes, but impervious to selection in larger eukaryotic cells. This could then lead to the emergence of cellular complexity as a simple consequence of nearly neutral, drift-like processes rather than by direct promotion by positive selection.

Because eukaryotic cells are typically larger than those of prokaryotes, often substantially so, there is an increase in the absolute investment of lipids based on the cell membrane alone. However, given a constant shape, the surface area of a cell increases with the square of cell length, whereas the volume increases with the cube of length, so the relative investment in the external membrane declines with cell volume. Nonetheless, with their additional investment in membrane-bound organelles, a larger fraction of cellular biomass is allocated to lipids in eukaryotes than prokaryotes, although not enormously so – the mean fractional contribution of lipids to total dry weight is ~ 0.06 for bacteria, ~ 0.08 for yeast species, and ~ 0.15 for unicellular photosynthetic algae (Chapter 3).

To understand the considerable bioenergetic costs of membrane production, we require information on the numbers of lipid molecules required for membrane production over the life of the cell as well as the cost of biosynthesis of such molecules. Again, the total number of molecules of a particular type required in cell's lifetime can be subdivided into a fixed quantity, equivalent to the number of molecules that comprise a newborn cell, N_l, and a time-dependent maintenance quantity associated with molecular turnover, $\delta_l N_l T/\ln(2)$, where δ_l is the molecular decay rate per lipid molecule, and T is the cell-division time.

Given the unfortunate absence of information on the rate of lipid-molecule turnover, it will be necessary here to rely on N_l as an estimate of the minimum

lifetime requirement for lipids, but as there is no evidence that membrane lipids are rapidly degraded, such estimates are likely to closely approximate the total investment in lipids per cell cycle. This quantity can be determined from the membrane areas of the cell if the occupancy of individual lipid molecules is known. Estimates of the head-group areas of common membrane lipids are mostly within 10% of an average of $a_l = 0.65$ nm^2 (Nagle and Tristram-Nagle 2000; Petrache et al. 2000; Kucerka et al. 2011). It is also essential to know the thickness of the bilayer, as this determines the areas of the inner and outer layers. The thickness of a bilayer is approximately twice the radius of the head-group area, which is ~ 0.5 nm in all cases, plus the total length of the internal hydrophobic tail domain (Lewis and Engelman 1983; Mitra et al. 2004), which is ~ 3.0 nm, and so sums to $h \simeq 4.0$ nm. There are slight increases in bilayer thickness with the length of the fatty-acid chain deployed (Rand and Parsegian 1989; Wieslander et al. 1995; Rawicz et al. 2000), as each single and double carbon-carbon bond adds ~ 0.15 and 0.13 nm, respectively.

To gain some appreciation for the number of lipid molecules per cell, now consider a spherical cell with radius r, which implies a surface area for the outer side of the bilayer of $4\pi r^2$, and $4\pi(r-h)^2$ for the inner layer. The summed area is then $4\pi[r^2 + (r-h)^2]$, or $8\pi r^2$ for $h \ll r$. Dividing by a_l gives the total number of lipid molecules in the bilayer. With $a_l = 0.65$ nm and $h = 4.5$ nm, for a bacterial-sized cell with radius $r = 1000$ nm (1 μm), there are then $\sim 4 \times 10^7$ total lipid molecules in the cell membrane, and this increases to $\sim 4 \times 10^9$ for a 10 μm sphere and $\sim 4 \times 10^{11}$ for a 100 μm sphere. For Gram negative bacteria (which include *Escherichia, Pseudomonas, Helicobacter,* and *Salmonella,* there are two cell membranes, which roughly doubles these numbers. Thus, the number of lipids in the cell membrane alone is almost always greater than the number of nucleotides in a cell's genome. These values need to be discounted somewhat to account for the occupancy of membranes by proteins; so, for example, if 50% of the membrane surface is occupied by proteins (probably an upper limit), the preceding values would be diminished by 50%.

We now consider the total (direct plus opportunity) biosynthetic costs of individual lipid molecules, the details of which are presented in Foundations 16.4. Most cellular membranes are predominantly comprised of glycerophospholipids (Table 16.1), which despite containing a variety of head groups (e.g., glycerol, choline, serine, glycerol, and inositol), all have biosynthetic costs per molecule (in units of ATP hydrolyses) of

$$c_L \simeq 301 + [36 \cdot (n_L - 16)] + (8 \cdot n_U), \tag{16.5a}$$

$$c_L \simeq 317 + [38 \cdot (n_L - 16)] + (8 \cdot n_U), \tag{16.5b}$$

in bacteria and eukaryotes respectively, where n_L is the mean fatty-acid chain length (number of backbone carbons), and n_U is the mean number of unsaturated carbons per fatty-acid chain. Although variants on glycerophospholipids are utilized in a variety of species (Guschina and Harwood 2006; Geiger et al. 2010), these are structurally similar enough that the preceding expressions should still provide excellent first-order approximations. The direct costs, which ignore the opportunity loss of ATP-generating potential from the diversion of metabolic precursors, are

$$c'_L \simeq 130 + [9 \cdot (n_L - 16)] + (8 \cdot n_U), \tag{16.6a}$$

$$c'_L \simeq 146 + [10 \cdot (n_L - 16)] + (8 \cdot n_U), \tag{16.6b}$$

in bacteria and eukaryotes.

For most lipids in biological membranes, $14 \leq n_L \leq 22$, and $0 \leq n_U \leq 6$, so the total cost per lipid molecule is generally in the range of $c_L \simeq 200$ to 600 ATP, although the average over the pooled population of lipids deployed in species-specific membranes is much narrower (Table 16.2). Cardiolipin, which rarely comprises more than 20% of membrane lipids is exceptional, having an evolutionary cost of ~ 700 ATP/molecule (and a reduced cost of 290 ATP). Thus, on an individual molecule basis, lipids are an order of magnitude more expensive than the other two major monomeric building blocks of cells (nucleotides and amino acids).

Application of the preceding expressions to known membrane compositions indicates that the biosynthetic costs of eukaryotic lipids are somewhat higher than those in bacteria (Table 16.1). For example, for a diversity of bacterial species, the average cost per lipid molecule in the plasma membrane is $c'_L = 153\,(\mathrm{SE} = 4)$ ATP, whereas that for eukaryotes is 173 (2). The latter estimate is very similar to the mean obtained for whole eukaryotic cells, 172 (3), but the average cost of mitochondrial lipids is especially high, 188 (6). These elevated expenses in eukaryotes are joint effects of two factors: 1) the added cost of mitochondrial export of oxaloacetate to generate acetyl-CoA necessary for lipid biosynthesis in the cytoplasm; and 2) the tendency for eukaryotic lipids to have longer chains containing more desaturated carbons.

We are now in position to estimate the total bioenergetic cost associated with membrane lipids. Recalling the surface occupancy of individual lipid molecules and the bilayer nature of membranes, the biosynthetic cost of a membrane with surface area A (in units of $\mu\mathrm{m}^2$, and ignoring the unknown contribution from lipid turnover) is

$$C_L \simeq 2A \cdot \bar{c}'_L/(0.65 \times 10^{-6}), \tag{16.7}$$

where $\bar{c}'_L$ is the average direct cost of a lipid (e.g., as given in Table 16.2). Dividing C_L by the total cost of building a progeny cell (Chapter 4) yields the proportion of a cell's total growth budget allocated to a membrane.

There are only a few cell types for which the internal anatomy has been scrutinized well enough to estimate the allocation to a cell's energy budget for the full set of membrane types. However, the data uniformly suggest that a substantial fraction (~ 10 to 30%) of a cell's growth budget is allocated to lipid biosynthesis (Table 16.2). As expected, based on the surface area:volume relationship, the plasma membrane constitutes a diminishing cost with increasing cell size, from $> 18\%$ in bacterial-sized cells with volumes < 2 $\mu\mathrm{m}^2$ to $< 4\%$ in large mammalian cells. Thus, the addition of internal membranes to eukaryotic cells imposes a very large energetic burden (up to 30%) not incurred by prokaryotes.

Enough information is available on the total investment in mitochondrial lipid membranes that a general statement can be made for this particular organelle. Over the eukaryotic domain, the total surface area of mitochondria (inner plus outer membranes, summed over all mitochondria, in $\mu\mathrm{m}^2$) scales with cell volume (V, in units of $\mu\mathrm{m}^3$) as $2.98V^{0.99} \simeq 3V$ (Figure 3.5b). Applying this to Equation (16.7), with the average total cost of mitochondrial lipids (427 ATP/ molecule; Table 16.1), and letting the total growth requirements of a cell be $(27 \times 10^9)V$ from Chapter 3, yields a measure of the relative cost of mitochondrial lipids of 0.15, essentially independent of eukaryotic cell size.

These results indicate that the construction of mitochondrial membranes, critical to ATP synthase operation in eukaryotes, amounts to a 15% drain on the cellular energy budget beyond what would be necessary had ATP synthase remained in operation on the plasma membrane (with mitochondria being absent). Not including the opportunity cost of building lipids reduces this measure to $\sim 44\%$ of the total evolutionary cost. These calculations are first-order approximations for rapidly growing cells, for which the contributions of cell maintenance and lipid-molecule turnover to the total cellular energy budget will be minor. For slowly growing cells, the costs will be higher or lower depending on whether the cost of mitochondrial-membrane maintenance is above or below that for total cellular maintenance. However, the central point remains – the costs of mitochondrial membranes represent a baseline price, not incurred by prokaryotes, associated with relocating bioenergetics to the interior of eukaryotic cells.

Although the data are not as extensive, this approach can be extended to show that quite different cost scalings exist for other types of internal membranes. Across the Tree of Life, the outer surface area of the nuclear envelope scales with the 0.56 power of cell volume, whereas the area of the golgi scales with the 1.45 power of cell volume (Figure 16.5). Assuming the average cost of a lipid in these organelles to be about the same as for the entire cell ($\sim$ 400 ATPs, from Table 16.2), and recalling that the nuclear envelope has a double membrane, the total cost of the nuclear envelope is $0.2V^{-0.44}$, whereas that for the golgi is $0.0008V^{0}.45$. Thus, the cost of the nuclear envelope (relative to a cell's total energy budget declines with increasing cell volume, whereas that for the golgi increases, with the two have equal costs of ~ 0.013 at a cell volume of 500 μm^3)

There are a number of other costs associated with membranes and their processing, but as in the case of nucleic acids and proteins, the major costs appear to be associated with the biosynthesis of the basic building blocks noted above. Consider, for example, the cost of molding membranes into specific shapes. The problem is most simply evaluated for spherical vesicles, for which the bending energy is ~ 400 K_BT independent of vesicle size (Phillips et al. 2012) – there is more surface area to bend in larger vesicles, but the curvature is reduced, and the two effects cancel. Knowing the rate of membrane flux for a cell then allows a rough estimate of the total bending energy required per unit time. For example, the entire cell membranes of a mammalian fibroblast and macrophage cells are interiorized by pinocytosis in about 0.5 to 2.0 hours (Steinman et al. 1976). The same is true for *Dictyostelium* (slime mold) cells (Thilo and Vogel 1980). Knowing the surface area of the cell and the average surface area of a vesicle, one can then estimate that about 1000 vesicles must be produced per minute. Assuming that the bending energy is acquired via ATP hydrolyzing reactions in the production of protein-coating cages (Chapter 14), and an energy of ~ 16 K_BT associated with each ATP hydrolysis, the total energy demand for bending would be $\sim 25,000$ ATP/minute. The cost of membrane fusion is even smaller, being equivalent to ~ 20 K_BT, or about one ATP hydrolysis, per fusion event (François-Martin et al. 2017). By contrast, as noted above, the plasma membrane of eukaryotic cells generally contains at least 10^9 lipid molecules, each of which costs ~ 400 ATPs for biosynthesis.

Foundations 16.4. The biosynthetic costs of lipid molecules. Estimation of the total cellular expenditure on the synthesis of a lipid molecule requires a separate consideration of the investments in the three molecular subcomponents: the fatty-acid tails, head groups, and linkers (Lynch and Marinov 2017). As in the preceding applications, we will quantify such costs in units of the number of phosphorus atoms released via ATP hydrolyses. CTP, which is utilized in a few reaction steps in lipid biosynthesis, will be treated as equivalent to ATP, and electron transfers resulting from conversions of NADH to $NADH^+$, NADPH to $NADPH^+$, and $FADH_2$ to FAD will be taken to be equivalent to 3, 4, and 2 ATP hydrolyses (all conventions in biochemistry based on energetic analyses). As noted in Chapter 14, a broad diversity of lipid molecules exist, and just the general approach is sketched out here.

The starting point for the synthesis of most fatty acids is the production of one particular linear chain, palmitate, containing 16 carbon atoms. Production of this molecule takes place within a large complex, known as fatty-acid synthase, and in bacteria biosynthesis of each molecule requires the consumption of 8 acetyl-CoA molecules, 7 ATPs, and reductions involving 14 NADPH. Each molecule of acetyl-CoA is generally derived from a pyruvate molecule, a process that produces 1 NADH, but each acetyl-CoA molecule diverted to lipid production is one less rotation of the energy producing citric-acid cycle, which would otherwise yield 3 NADH, 1 $FADH_2$, and 1 ATP per rotation; this leads to a net loss of the equivalent of 9 ATPs per acetyl-CoA molecule. Thus, the total cost of production of one molecule of palmitate is $(8 \times 9) + (7 \times 1) + (14 \times 4) = 135$ ATP in bacteria. Fatty-acid production is slightly more expensive in nonphotosynthetic eukaryotes, where acetyl-CoA is produced in the mitochondrion and reacts with oxaloacetate to produce citrate, which must then be exported. Cleavage of oxalacetate in the cytosol yields acetyl-CoA at the expense of 1 ATP, and a series of reactions serve to return oxaloacetate to the citric-acid cycle in an effectively ATP neutral way. Thus, the cost of palmitate increases to 143 ATP. Each additional pair of carbons added to this fatty-acid chain requires one additional acetyl-CoA and ATP, and two additional NADPHs, or an equivalent of 18 ATPs, and each subsequent desaturation of a bond consumes one NADPH, or 4 ATP equivalents.

To evaluate the total cost of a lipid molecule, we first consider the situation for glycerophospholipids, for which glycerol-3-phosphate (G3P) serves as the linker between the fatty-acid chains and the headgroup. G3P emerges from one of the last steps in glycolysis, and its diversion to lipid production deprives the cell of one further step of ATP production as well as a subsequent pyruvate molecule. As pyruvate normally can yield the equivalent of 3 ATPs in the conversion to acetyl-CoA, which then would generate a net 12 ATPs following entry into the citric-acid cycle, the use of G3P as a linker in a lipid molecule has a cost of $1 + 3 + 12 = 16$ ATP. Linking each fatty-acid tail requires 1 ATP, and linking the head group involves two CTP hydrolyses.

All that remains is the cost of synthesis of the head group. In the case of phosphatidylglycerol, the head group is G3P, the cost of which is 16 ATP, as just noted, so the total cost of this molecule in a bacterium is $\simeq (2 \cdot 135) + 20 + 16 = 306$ ATP. From Table 11.1, the cost of a serine is 14 ATP, so the total cost of a phosphatidylserine is 304 ATP, and because ethanolamine and choline are simple derivatives of serine, this closely approximates the costs of both phosphatidylethanolamine and phosphatidylcholine. The headgroup of phosphatidylinositol is inosital, which is derived from glucose-6-phosphate, diverting the latter from glycolysis and depriving the cell of the equivalent of 9 ATPs, so the total cost of production of this molecule is 299 ATP. Finally, cardiolipin is synthesized by the fusion of two phosphatidylglycerols and the release of one glycerol, so taking the return from the latter to be 15 ATP, the total cost per molecule produced is 703 ATP.

Estimation of the cost of biosynthesis of sphingolipids follows many of the steps

just outlined. Construction of the linker molecule requires a palmitate molecule and the expenditure of one NADPH, for a total of 139 and 147 ATP in bacteria and eukaryotes, respectively. Then, a single fatty-chain is added, so assuming this is palmitate, this requires the expenditure of another 136 or 144 ATP. Finally, there are the costs of synthesizing and adding the head group, both of which are outlined in the preceding paragraphs.

Table 16.1. Bioenergetic costs for the synthesis of lipid molecules. Data are provided for species with lipid composition measurements of the parameters needed to solve Equations (16.1a-16.2b), along with additional information on the contribution from cardiolipin. PL denotes glycerophospholipid, and C cardiolipin, with the cost for the remaining small fraction being taken to be the average of the preceding two per molecule. Total cost denotes the opportunity plus direct cost per molecule incorporated into the membrane. Mean costs are obtained by weighting the PL and cardiolipin costs by their fractional contributions. Standard deviations among species are given in parentheses. Summarized from Lynch and Marinov (2017).

	PL Cost		Composition		Mean Cost	
Source	Total	Direct	PL	C	Total	Direct
Bacteria, whole cell	342 (27)	110 (7)	0.89 (0.09)	0.09 (0.06)	372 (17)	123 (7)
Euks., whole cell	393 (27)	139 (7)	0.95 (0.03)	0.05 (0.03)	404 (23)	143 (6)
Euks., plasma memb.	396 (21)	139 (5)	0.95 (0.05)	0.03 (0.03)	406 (22)	143 (6)
Euks., mitochondrion	404 (52)	142 (15)	0.85 (0.08)	0.11 (0.05)	436 (45)	155 (14)

Table 16.2. Contributions of membranes to total cellular growth costs (in units of ATP equivalents). Cell volumes (CV) and total membrane (TM) surface areas are in units of μm^3 and μm^2, respectively. Fractional contributions to the total cell growth requirements are given for the plasma membrane (Pm), mitochondrial membranes (inner + outer, Mt), nuclear envelope (Nu), endoplasmic reticulum and golgi (ER/G), vesicles and vacuoles (V), and total (Tot). The fraction of the total cell growth budget allocated to membranes is obtained by applying Equation 8.3, using the species-specific lipid biosynthesis costs (Table 16.2) where possible (and otherwise applying the averages for eukaryotic species), and normalizing by the allometric equation for ATP growth requirements in Chapter 4. The results for the two algae, *O. tauri* and *D. salina,* do not include the investment in plastid membranes. From Lynch and Marinov (2017).

			Fractional contributions to total cell growth:					
Organism	CV	TM	Pm	Mt	Nu	ER/G	V	Tot
Bacteria:								
Staphylococcus aureus	0.29	4.00	0.217					0.217
Escherichia coli	0.98	15.85	0.284					0.284
Bacillus subtilis	1.41	10.69	0.138					0.138
Eukaryotes:								
Ostreococcus tauri	1.0	15	0.180	0.079	0.037	0.011	0.018	0.324
Saccharomyces cerevisiae	44	204	0.031	0.034	0.008	0.011	0.011	0.096

Dunaliella salina	591	2299	0.013	0.019	0.003	0.030	0.030	0.094
Sus scrofa, pancreas	1060	12952	0.013	0.067	0.002	0.213	0.006	0.302

Literature Cited

Akashi, H., and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA 99: 3695-3700.

Acquisti, C., J. J. Elser, and S. Kumar. 2009. Ecological nitrogen limitation shapes the DNA composition of plant genomes. Mol. Biol. Evol. 26: 953-956.

Adler, M., M. Anjum, O. G. Berg, D. I. Andersson, and L. Sandegren. 2014. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. Mol. Biol. Evol. 31: 1526-1535.

Arnold, A., M. Sajitz-Hermstein, and Z. Nikoloski. 2015. Effects of varying nitrogen sources on amino acid synthesis costs in *Arabidopsis thaliana* under different light and carbon-source conditions. PLoS One 10: e0116536.

Atkinson, D. E. 1970. Adenine nucleotides as universal stoichiometric metabolic coupling agents. Adv. Enzyme Regul. 9: 207-219.

Barton, M. D., D. Delneri, S. G. Oliver, M. Rattray, and C. M. Bergman. 2010. Evolutionary systems biology of amino acid biosynthetic cost in yeast. PLoS One 5: e11935.

Bauchop, T., and S. R. Elsden. 1960. The growth of micro-organisms in relation to their energy supply. J. Gen. Microbiol. 23: 457-469.

Belle, A., A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea. 2006. Quantification of protein half-lives in the budding yeast proteome. Proc. Natl. Acad. Sci. USA 103: 13004-13009.

Bernstein, J. A., A. B. Khodursky, P. H. Lin, S. Lin-Chao, and S. N. Cohen. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. USA 99: 9697-9702.

Booth, A., and W. F. Doolittle. 2015. Eukaryogenesis, how special really? Proc. Natl. Acad. Sci. USA 112: 10278-10285.

Bragg, J. G., and C. L. Hyder. 2004 Nitrogen versus carbon use in prokaryotic genomes and proteomes. Proc. Biol. Sci. 271 Suppl 5: S374-7.

Craig, C. L., and R. S. Weber. 1998. Selection costs of amino acid substitutions in ColE 1 and colIa gene clusters harbored by *Escherichia coli* . Mol. Biol. Evol. 15: 774-776.

Csárdi, G., A. Franks, D. S. Choi, E. M. Airoldi, and D. A. Drummond. 2015. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. PLoS Genet. 11: e1005206.

Dressaire, C., F. Picard, E. Redon, P. Loubière, I. Queinnec, L. Girbal, and M. Cocaign-Bousquet. 2013. Role of mRNA stability during bacterial adaptation. PLoS One 8: e59059.

Eden, E., N. Geva-Zatorsky, I. Issaeva, A. Cohen, E. Dekel, T. Danon, L. Cohen, A. Mayo, and U. Alon. 2011. Proteome half-life dynamics in living human cells. Science 331: 764-768.

Elser, J. J., C. Acquisti, and S. Kumar. 2011. Stoichiogenomics: the evolutionary ecology of macromolecular elemental composition. Trends Ecol. Evol. 26: 38-44.

François-Martin, C., J. E. Rothman, and F. Pincet. 2017. Low energy cost for optimal speed and control of membrane fusion. Proc. Natl. Acad. Sci. USA 114: 1238-1241.

Frumkin, I., D. Schirman, A. Rotman, F. Li, L. Zahavi, E. Mordret, O. Asraf, S. Wu, S. F. Levy, and Y. Pilpel. 2017. Gene architectures that minimize cost of gene expression. Mol. Cell 65: 142-153.

Geiger, O., N. González-Silva, I. M. López-Lara, and C. Sohlenkamp. 2010. Amino acid-containing membrane lipids in bacteria. Prog. Lipid Res. 49: 46-60.

Geiler-Samerotte, K. A., M. F. Dion, B. A. Budnik, S. M. Wang, D. L. Hartl, and D. A. Drummond. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded

protein response in yeast. Proc. Natl. Acad. Sci. USA 108: 680-685.

Ghaemmaghami, S., W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. 2003. Global analysis of protein expression in yeast. Nature 425: 737-741.

Golding, I., J. Paulsson, S. M. Zawilski, and E. C. Cox. 2005. Real-time kinetics of gene activity in individual bacteria. Cell 123: 1025-1036.

Grzymski, J. J., and A. M. Dussaq. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 6: 71-80.

Günther, T., C. Lampei, and K. J. Schmid. 2013. Mutational bias and gene conversion affect the intraspecific nitrogen stoichiometry of the *Arabidopsis thaliana* transcriptome. Mol. Biol. Evol. 30: 561-568.

Guschina, I. A., and J. L. Harwood. 2006. Lipids and lipid metabolism in eukaryotic algae. Prog. Lipid Res. 45: 160-186.

Helbig, A. O., P. Daran-Lapujade, A. J. van Maris, E. A. de Hulster, D. de Ridder, J. T. Pronk, A. J. Heck, and M. Slijper. 2011. The diversity of protein turnover and abundance under nitrogen-limited steady state conditions in *Saccharomyces cerevisiae.* Mol. Biosyst 7: 3316-3326.

Heizer, E. M., Jr., D. W. Raiford, M. L. Raymer, T. E. Doom, R. V. Miller, and D. E. Krane. 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol. Biol. Evol. 23: 1670-1680.

Heizer, E. M., Jr., M. L. Raymer, and D. E. Krane. 2011. Amino acid biosynthetic cost and protein conservation. J. Mol. Evol. 72: 466-473.

Hambraeus, G., C. von Wachenfeldt, and L. Hederstedt. 2003. Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. Mol. Genet. Genomics 269: 706-714.

Islam, S., U. Kjällquist, A. Moliner, P. Zajac, J. B. Fan, P. Lönnerberg, and S. Linnarsson. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21: 1160-1167.

Krick, T., N. Verstraete, L. G. Alonso, D. A. Shub, D. U. Ferreiro, M. Shub, and I. E. Sánchez. 2014. Amino acid metabolism conflicts with protein diversity. Mol. Biol. Evol. 31: 2905-2912.

Kučerka, N., M. P. Nieh, and J. Katsaras. 2011. Fluid phase lipid areas and bilayer thicknesses of commonly used phosphatidylcholines as a function of temperature. Biochim. Biophys. Acta 1808: 2761-2771.

Lahtvee, P. J., A. Seiman, L. Arike, K. Adamberg, and R. Vilu. 2014. Protein turnover forms one of the highest maintenance costs in *Lactococcus lactis.* Microbiology 160: 1501-1512.

Lane, N., and W. Martin. 2010. The energetics of genome complexity. Nature 467: 929-934.

Lewis, B. A., and D. M. Engelman. 1983. Lipid bilayer thickness varies linearly with acyl chain length in fluid phosphatidylcholine vesicles. J. Mol. Biol. 166: 211-217.

Li, G. W., and X. S. Xie. 2011. Central dogma at the single-molecule level in living cells. Nature 475: 308-315.

Long, H., et al. 2017. Evolutionary determinants of genome-wide nucleotide composition. Nature Ecol. Evol. 2: 237-240.

Lotka, A. J. 1922. Contribution to the energetics of evolution. Proc. Natl. Acad. Sci. USA 8: 147-151.

Lu, P., C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25: 117-124.

Lynch, M. 2007a. The Origins of Genome Architecture. Sinauer Assocs., Inc., Sunderland, MA.

Lynch, M. 2007b. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. USA 104 (Suppl.): 8597-8604.

Lynch, M., and G. K. Marinov. 2015. The bioenergetic costs of a gene. Proc. Natl. Acad. Sci. USA 112: 15690-15695.

Lynch, M., and G. K. Marinov. 2017. Membranes, energetics, and evolution across the prokaryote-eukaryote divide. eLife 6: e20437.

Mani, S., and M. Thattai. 2016. Stacking the odds for Golgi cisternal maturation. elife 5: e16231.

Marinov, G. K., B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 24: 496-510.

McDermitt, D. K., and R. S. Loomis. 1981. Elemental composition of biomass and its relation to energy content, growth efficiency, and growth yield. Ann. Bot. 48: 275290.

Mitra, K., I. Ubarretxena-Belandia, T. Taguchi, G. Warren, and D. M. Engelman. 2004. Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. Proc. Natl. Acad. Sci. USA 101: 4083-4088.

Nagle, J. F., and S. Tristram-Nagle. 2000. Structure of lipid bilayers. Biochim. Biophys. Acta 1469: 159-195.

Neymotin, B., R. Athanasiadou, and D. Gresham. 2014. Determination of *in vivo* RNA kinetics using RATE-seq. RNA 20: 1645-1652.

Petrache, H. I., S. W. Dodd, and M. F. Brown. 2000. Area per lipid and acyl length distributions in fluid phosphatidylcholines determined by ^{2}H NMR spectroscopy. Biophys. J. 79: 3172-3192.

Phillips, M. J., and G. K. Voeltz. 2016. Structure and function of ER membrane contact sites with other organelles. Nat. Rev. Mol. Cell. Biol. 17: 69-82.

Phillips, R., J. Kondev, J. Theriot, and H. Garcia. 2012. Physical Biology of the Cell, 2nd Ed. Garland Science, New York, NY.

Plata, G., M. E. Gottesman, and D. Vitkup. 2010. The rate of the molecular clock and the cost of gratuitous protein synthesis. Genome Biol. 11: R98.

Raiford, D. W., E. M. Heizer, Jr., R. V. Miller, H. Akashi, M. L. Raymer, and D. E. Krane. 2008. Do amino acid biosynthetic costs constrain protein evolution in *Saccharomyces cerevisiae*? J. Mol. Evol. 67: 621-630.

Raj, A., C. S. Peskin, D. Tranchina, and D. Y. Vargas, and S. Tyagi. 2006. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 4: e309.

Raiford, D. W., E. M. Heizer, Jr., R. V. Miller, T. E. Doom, M. L. Raymer, and D. E. Krane. 2012. Metabolic and translational efficiency in microbial organisms. J. Mol. Evol. 74: 206-216.

Ramadas, R., and M. Thattai. 2013. New organelles by gene duplication in a biophysical model of eukaryote endomembrane evolution. Biophys. J. 104: 2553-2563.

Rand, R. P., and V. A. Parsegian. 1989. Hydration forces between phospholipid bilayers. Biochim. Biophys. Acta 988: 351-376.

Rawicz, W., K. C. Olbrich, T. McIntosh, D. Needham, and E. Evans. 2000. Effect of chain length and unsaturation on elasticity of lipid bilayers. Biophys. J. 79: 328-339.

Russell, J. B., and G. M. Cook. 1995. Energetics of bacterial growth: balance of anabolic and catabolic reactions. Microbiol. Rev. 59: 48-62.

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. 2011. Global quantification of mammalian gene expression control. Nature 473: 337-342.

Seligmann, H. 2003. Cost-minimization of amino acid usage. J. Mol. Evol. 56: 151-161.

Shahrezaei, V., and P. S. Swain. 2008. Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. USA 105: 17256-17261.

Steinman, R. M., S. E. Brodie, and Z. A. Cohn. 1976. Membrane flow during pinocytosis: a stereologic analysis. J. Cell Biol. 68: 665-687.

Stoebel, D. M., A. M. Dean, and D. E. Dykhuizen. 2008. The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. Genetics 178: 1653-1660.

Stouthamer, A. H. 1973. A theoretical study on the amount of ATP required for synthesis of microbial cell material. Antonie Van Leeuwenhoek 39: 545-65.

Swire, J. 2007. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. J. Mol. Evol. 64: 558-571.

Taniguchi, Y., P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329: 533-538.

Tempest, D. W., and O. M. Neijssel. 1984. The status of YATP and maintenance energy as biologically interpretable phenomena. Annu. Rev. Microbiol. 38: 459-486.

Thilo, L., and G. Vogel. 1980. Kinetics of membrane internalization and recycling during pinocytosis in *Dictyostelium discoideum.* Proc. Natl. Acad. Sci. USA 77: 1015-1019.

Tomala, K., and R. Korona. 2013. Evaluating the fitness cost of protein expression in *Saccharomyces cerevisiae.* Genome Biol. Evol. 5: 2051-2060.

Trotschel, C., S. P. Albaum, and A. Poetsch. 2013. Proteome turnover in bacteria: current status for *Corynebacterium glutamicum* and related bacteria. Microb. Biotechnol. 6: 708-719.

van Milgen, J. 2002. Modeling biochemical aspects of energy metabolism in mammals. J. Nutr. 132: 3195-3202.

Van Valen, L. 1976. Energy and evolution. Evol. Theory 1: 179-229.

Van Valen, L. 1980. Evolution as a zero-sum game for energy. Evol. Theory 4: 289-300.

Walsh, J. B., and M. Lynch. 2018. Selection and Evolution of Quantitative Traits. Oxford Univ. Press, Oxford, UK.

Wang, Y., C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown. 2002. Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. USA 99: 5860-5865.

Wieslander, A., S. Nordström, A. Dahlqvist, L. Rilfors, and G. Lindblom. 1995. Membrane lipid composition and cell size of *Acholeplasma laidlawii* strain A are strongly influenced by lipid acyl chain length. Eur. J. Biochem. 227: 734-744.

Zenklusen, D., D. R. Larson, and R. H. Singer. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. Nat. Struct. Mol. Biol. 15: 1263-1271.

Figure 16.1. The distinction between direct and opportunity costs associated with synthesizing molecular building blocks (e.g., an amino acid). As the energy resource (e.g., glucose) is partially metabolized into precursor metabolites necessary for synthesis of the building block, the additional energy that could have been captured from the complete metabolism of the resource is the opportunity cost. The conversion of precursor metabolites to some molecular building blocks can also consume electron carrier molecules such as NADH, which if not used in building-block synthesis would have generated ATP, and thus represent additional source of opportunity cost. The direct consumption of ATP in the biosynthetic process defines the direct cost of building-block synthesis. The assembly of macromolecules such as proteins from building blocks requires additional post-synthesis costs such as the cost of polymerization and maintenance. Finally, for proteins and nucleic acids, the polymerization of building blocks incurs an assembly cost, and there may be subsequent maintenance costs as well (e.g., associated with molecular turnover, or DNA repair).

Figure 16.2. The evolutionary distinction between the construction/maintenance cost of a trait and the direct benefits. s_c represents the reduction in fitness that would be expected from the presence of the trait in the absence of any direct benefits; it may be viewed as the selective advantage of a mutant relieved of the trait in an environment in which no advantages of the trait are experienced. s_d is a measure of the increase in fitness that would accrue in the absence of any assembly/maintenance costs. The difference $s_n = s_d - s_c$ denotes the net fitness advantage of the trait; if this value is negative the trait is selectively disadvantageous despite any ecological benefits accrued. Clearly, a gene will be selected against if $s_d < s_c$, but $s_d > s_c$ is not a sufficient condition for gene preservation by natural selection, as the absolute value of s_n must exceed the power of random genetic drift, $1/N_e$ in a haploid species and $1/(2N_e)$ in a diploid, to be readily perceived by natural selection (Chapter 8).

Figure 16.3. The distribution of the total costs for all genes (and their three components, associated with replication, transcription, and translation) in four species for which the detailed transcriptomic and proteomic data necessary for full analyses are possible (from Lynch and Marinov 2015). The data are presented as frequency histograms, so that the peaks represent the mode of the costs over the full set of genes. The lower axis ($\log_{10}$) denotes the total (direct plus opportunity) costs, whereas the upper axis divides these numbers by the total cost of building a cell (from Chapter 4), and is a measure of the selective cost of the maintenance and operation of a gene relative to the cell's total energy budget. The vertical dashed lines denote the drift barrier, with the middle line approximately demarcating $1/N_e$ for the species, and the two flanking lines simply providing order-of-magnitude margins for error. The fact that the relative costs of virtually all genes in *E. coli* are far to the right of the drift barrier implies that their baseline costs can easily be perceived by selection – if such genes do not pay their way by endowing the cell with benefits in excess of such costs, they would be rapidly purged by degenerative mutation and negative selection. The effective population size declines with organism size, moving the drift barrier to the right, whereas the relative costs of genes move to the left owing to the predominant effect of an increase in the total cost of the cell. As a consequence, with increasing organism size, the baseline costs of genes tend to move to the left of the drift barrier, implying that gene-specific costs are too small, particularly at the DNA and RNA levels, to be perceived by selection.

Figure 16.4. Estimated costs (relative to the estimated total cellular energy budgets) for average genes in 44 species of prokaryotes and eukaryotes. For all components of cost, there is a clear negative scaling with increasing cell volume, both within and among groups, with no discontinuity in the scaling behavior between prokaryotes and eukaryotes. From Lynch and Marinov (2015).

Figure 16.5. Total surface area of the nucleus and the stacked golgi scales with cell volume across phylogenetic groups as $A = 2.27V^{0.56}$ and $A = 0.018V^{1.45}$, respectively, with A and V having units of μm^2 and μm^3.

Figure 16.1

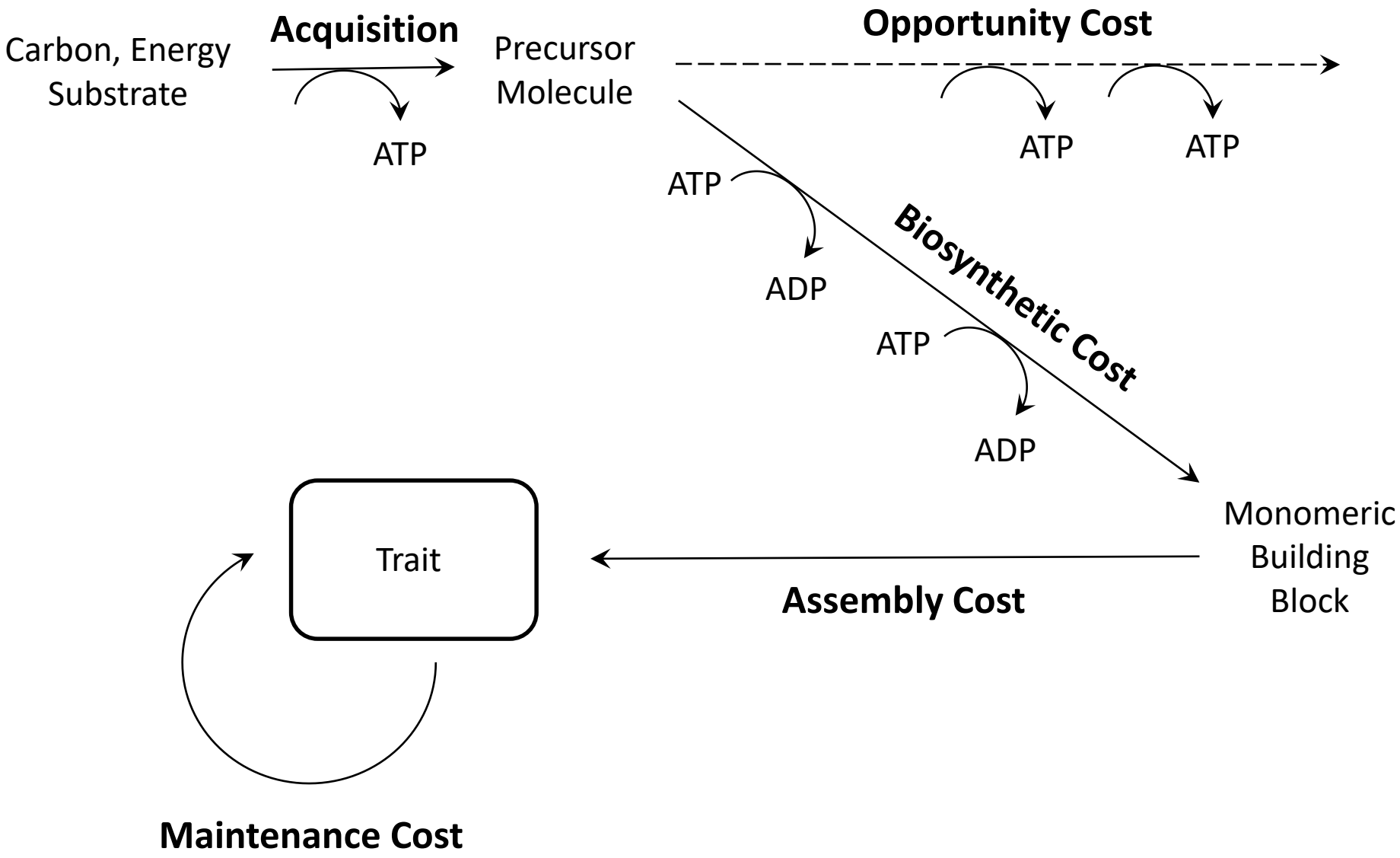

Carbon, Energy Substrate
Acquisition
ATP
Precursor Molecule
Opportunity Cost
ATP
ATP
ATP
ADP
Biosynthetic Cost
ATP
ADP
Monomeric Building Block
Trait
Assembly Cost
Maintenance Cost

Figure 16.2

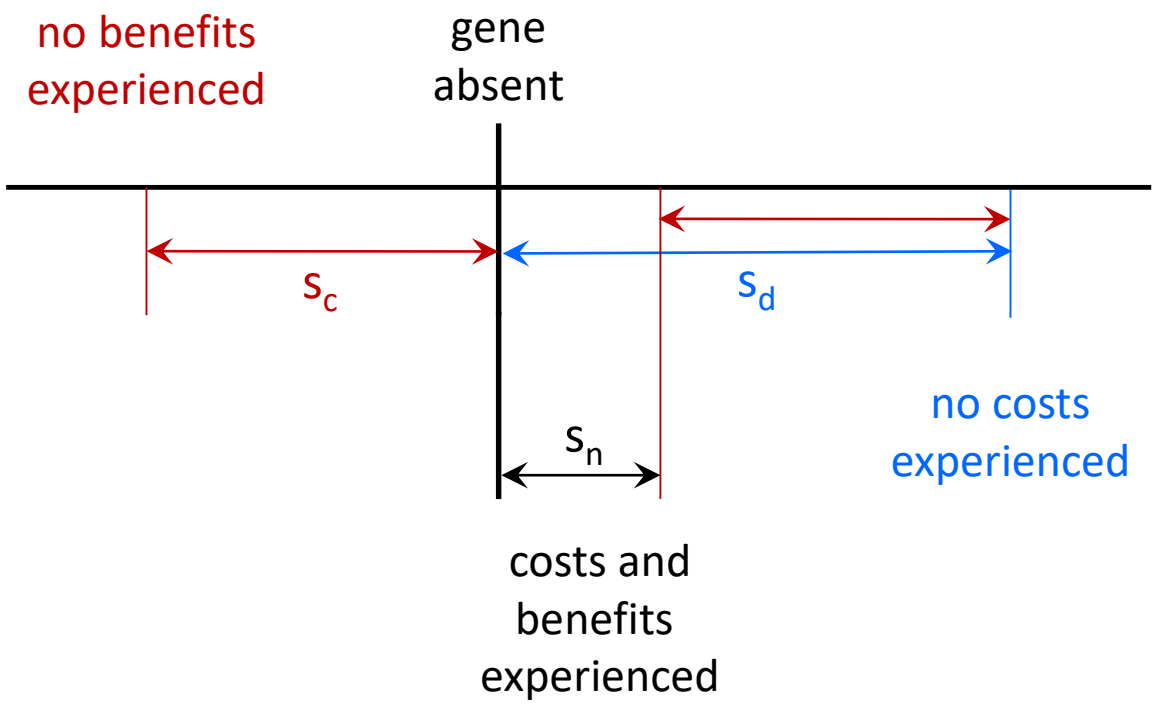

no benefits
experienced
gene
absent
s_c
s_d
s_n
no costs
experienced
costs and
benefits
experienced

Figure 16.3

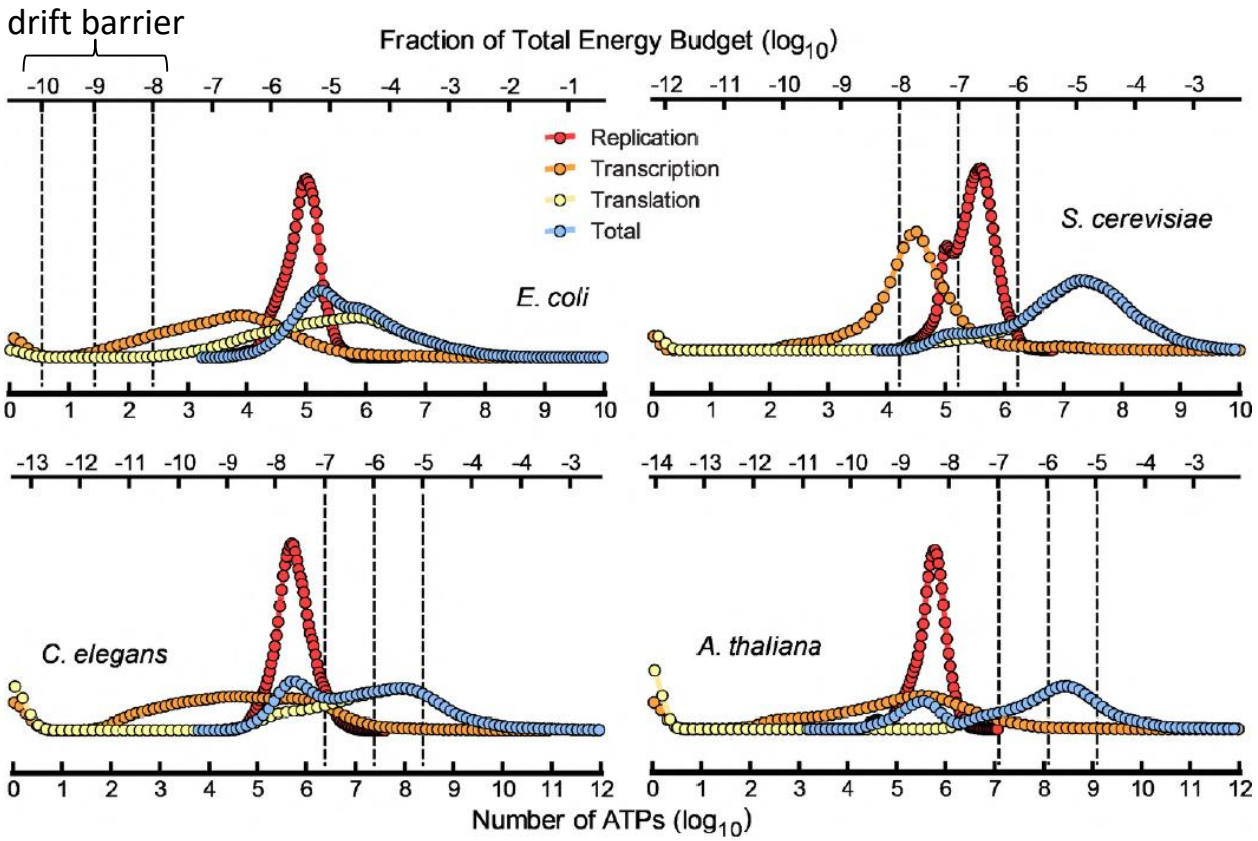

drift barrier
Fraction of Total Energy Budget (log_{10})
Replication
Transcription
Translation
Total
E. coli
S. cerevisiae
C. elegans
A. thaliana
Number of ATPs (log_{10})

Figure 16.4

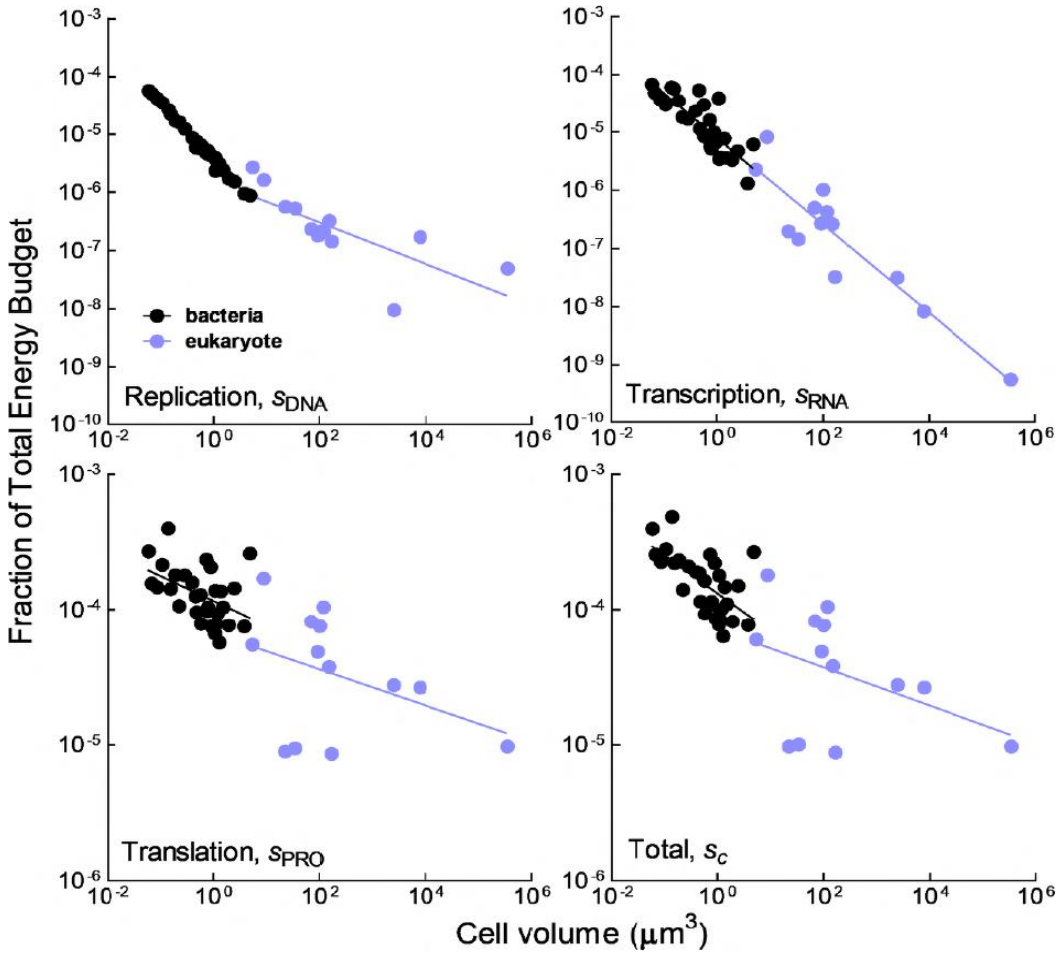
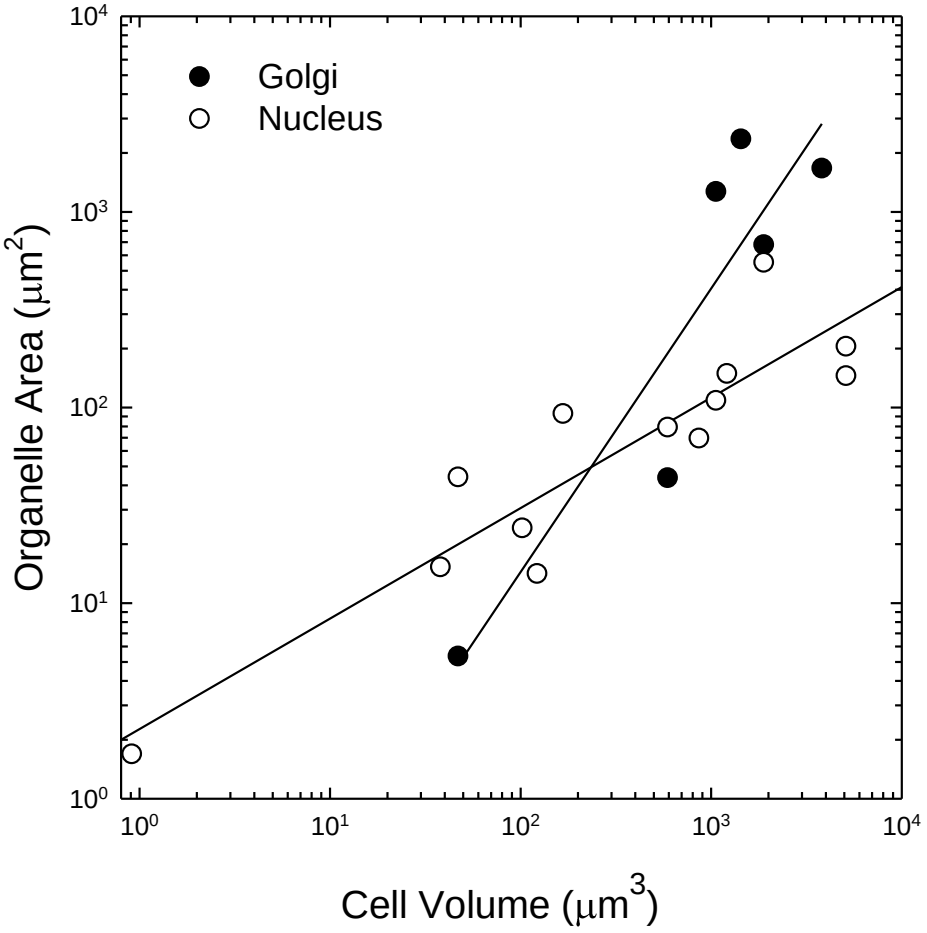

Fraction of Total Energy Budget
Cell volume (μm3)
bacteria
eukaryote
Replication, sDNA
Transcription, sRNA
Translation, sPRO
Total, sc

Figure 16.5


Golgi
Nucleus
Organelle Area (μm2)
Cell Volume (μm3)
10^0
10^1
10^2
10^3
10^4