

## Major Review

**Cite this article:** Kepp KP (2020). Survival of the cheapest: how proteome cost minimization drives evolution. *Quarterly Reviews of Biophysics* **53**, e7, 1–18. <https://doi.org/10.1017/S0033583520000037>

Received: 13 January 2020

Revised: 27 May 2020

Accepted: 28 May 2020

### Key words:


Amino acid cost; evolution; metabolism; protein misfolding; protein turnover

### Author for correspondence:

Kasper P. Kepp,

E-mail: [kpj@kemi.dtu.dk](mailto:kpj@kemi.dtu.dk)

# Survival of the cheapest: how proteome cost minimization drives evolution

Kasper P. Kepp 

Technical University of Denmark, DTU Chemistry, Kemitorvet 206, DK-2800 Kongens Lyngby, Denmark

## Abstract

Darwin's theory of evolution emphasized that positive selection of functional proficiency provides the fitness that ultimately determines the structure of life, a view that has dominated biochemical thinking of enzymes as perfectly optimized for their specific functions. The 20th-century modern synthesis, structural biology, and the central dogma explained the machinery of evolution, and nearly neutral theory explained how selection competes with random fixation dynamics that produce molecular clocks essential e.g. for dating evolutionary histories. However, quantitative proteomics revealed that selection pressures not relating to optimal function play much larger roles than previously thought, acting perhaps most importantly via protein expression levels. This paper first summarizes recent progress in the 21st century toward recovering this universal selection pressure. Then, the paper argues that proteome cost minimization is the dominant, underlying 'non-function' selection pressure controlling most of the evolution of already functionally adapted living systems. A theory of proteome cost minimization is described and argued to have consequences for understanding evolutionary trade-offs, aging, cancer, and neurodegenerative protein-misfolding diseases.

## Table of contents

<b>Introduction</b>	<b>1</b>
<b>The main determinants of evolution rate</b>	<b>2</b>
<b>The theory of proteome cost minimization</b>	<b>4</b>
<b>Selection dynamics of PCM</b>	<b>5</b>
<b>Typical PCM selection pressures and fixation probabilities for yeast</b>	<b>7</b>
<b>Scaling relations of proteome costs: mass, metabolism, and eukaryote evolution</b>	<b>8</b>
<b>Evidence for PCM during evolution</b>	<b>9</b>
Major evolutionary events mainly represented bioenergetic advantages	9
Energy surplus determines growth of microorganisms	9
Protein turnover is very expensive	9
Life uses cheap amino acids	9
Prokaryote streamlining	10
Highly expressed proteins are more streamlined	11
Unstable proteins reduce cell growth	11
<b>Trading function for cost</b>	<b>11</b>
<b>Time or energy?</b>	<b>12</b>
<b>Temperature, thermostable proteins, and thermophilic organisms</b>	<b>13</b>
<b>PCM, aging, and neurodegenerative diseases</b>	<b>13</b>
<b>Conclusions</b>	<b>14</b>

## Introduction

Protein evolution occurs via mutations that change the composition or expression of the proteome of a population, sometimes by random nearly neutral drift, and sometimes via selection pressures imposed by the habitat (Bajaj and Blundell, 1984; DePristo *et al.*, 2005; Pál *et al.*, 2006; Goldstein, 2008; Hurst, 2009; Worth *et al.*, 2009) After Darwin's theory of natural

selection, Mendel's laws of inheritance, the modern synthesis of the 20th century, and the rise of structural biology and the central dogma, we know that nature selects favorable traits if their impact outweighs the random fixation dynamics, and we know how these changes are actualized via mutations in the DNA that translate to the proteome. Remaining major questions are: (1) how important is selection *versus* random drift and can we predict their relative importance? (Kimura, 1962; Blundell and Wood, 1975; Ohta 1992; Hurst, 2009). (2) What are the molecular properties selected for, and are they universal? (Hurst, 2009; Lobkovsky *et al.*, 2010; Liberles *et al.*, 2012). (3) How do we describe accurately and completely the evolution of populations from the arising mutation in the gene, via the molecular property of the protein, to its fixation and ultimate effect on the population? According to this view, the ultimate goal of biology is to bridge the genome, proteome, phenotype, and population together in one quantitative and predictive theory that explains the history, present, and future of biological structure on this planet.

In the 1960s, the observation of nearly constant evolution of homologous proteins (Margoliash, 1963; Zuckerkandl and Pauling, 1965, 1962) led to the theory of (nearly) neutral evolution implying that most fitness effects are too subtle to dominate over random fixation dynamics of the population, thus producing an almost constant rate of evolution (Kimura, 1962; Ohta, 1992). This resulting, widely applied molecular clock is essential for dating phylogenies and evolutionary histories (Zuckerkandl and Pauling, 1965; Kumar and Subramanian, 2002; Yi *et al.*, 2002; Meredith *et al.*, 2011). When applied to single individuals, variations in the clock specific to the mutated site are used to indicate pathogenicity of a human gene variant (Ng and Henikoff, 2003; Flanagan *et al.*, 2010; Shihab *et al.*, 2013; Tang *et al.*, 2019). The evolution rate varies by many orders of magnitude between sites and proteins (Zuckerkandl and Pauling, 1965; Gillespie, 1984, 1986; Drummond *et al.*, 2005) and can be used to distinguish neutral evolution (Kimura, 1991; Ohta 1992; Fay *et al.*, 2002) from *adaptation or positive selection* toward a new fitness optimum (Hurst, 2009).

Darwin's theory of evolution emphasized that positive selection of optimal function provides the fitness that ultimately determines the structure of life (survival of the fittest). This view has dominated biochemical thinking of enzymes as perfectly optimized catalysts, implying that evolution strives toward optimal function *per se*, e.g. maximal substrate turnover ( $k_{cat}/K_m$ ) of highly optimized and conserved active sites as the main *raison d'être* (Radzicka and Wolfenden, 1995; Cannon *et al.*, 1996; Zhang and Houk, 2005; Hurst, 2009; Soskine and Tawfik, 2010). The connectivity of many proteins (i.e. the extent of their involvement in biochemical pathways) seemed to slow their rate of evolution, consistent with functional constraints on evolution (Fraser *et al.*, 2002; Hahn and Kern, 2004; Wall *et al.*, 2005). However, proteins are also subject to 'non-function' selection pressures directed toward e.g. proteome stability and efficiency of translation (Ehrenberg and Kurland, 1984; Hurst and Smith, 1999; Bloom and Adami, 2003; 2004; Drummond *et al.*, 2005; Lobkovsky *et al.*, 2010; Wylie and Shakhnovich, 2011). During early evolution, fierce competition produced evolutionary innovations in prokaryotes, and the rise of the eukaryotes (Lane and Martin, 2010; Sousa *et al.*, 2013) heralded major biochemical innovations largely relating to advantages of size and metabolism, rather than function *per se* (Lane, 2011). Under these conditions, the ability to efficiently harvest energy and chemical components was critical (Lane and Martin, 2010; Sousa *et al.*, 2013).

The subsequent long periods of relatively stable evolution have seen active sites of proteins highly conserved by purifying selection near their fitness optima (Blundell and Wood, 1975; Casari *et al.*, 1995) and most sequence variation occurs in other sites where nearly neutral substitutions probably dominate most recent evolution (Ohta, 1992). For the same reason, almost all protein evolution involves sequence variations that maintain the already adopted, highly conserved fold structure (Worth *et al.*, 2009). The nearly neutral sites that dominate this evolution are subject to non-function selection pressures, i.e. selection pressures not directly reflecting optimal chemical turnover of the protein. Most importantly, they may contribute to optimal translational efficiency under favorable growth conditions (Ikemura, 1985; Andersson and Kurland, 1990). Selection at the gene level for translational efficiency and precision (Ehrenberg and Kurland, 1984; Andersson and Kurland, 1990; Marais and Duret, 2001; Akashi, 2003; Drummond *et al.*, 2005) is evident e.g. from codon bias and t-RNA isoforms (Robinson *et al.*, 1984; Kanaya *et al.*, 1999; Tuller *et al.*, 2010).

This review concerns the question: What drives protein evolution on most time scales where the function is already nearly optimal? To address this question, we must first discuss the typical properties of proteins. Proteins vary by three orders of magnitude in length (from tens to ten thousands of amino acids), they vary structurally via thousands of folds (Bajaj and Blundell, 1984; Mirny and Shakhnovich, 1999; Qian *et al.*, 2001; Koonin *et al.*, 2002), and by perhaps 5–7 orders of magnitude in abundance in eukaryotic cells (Jansen and Gerstein, 2000; Beck *et al.*, 2011; Milo 2013).

In stark contrast to these enormous variations, proteins across all domains of life are marginally stable in a narrow range of perhaps 30–100 kJ mol<sup>-1</sup>, barely preventing denaturation (DePristo *et al.*, 2005; Goldstein, 2011). There are three possible origins of this phenomenon: marginal stability is a selected beneficial trait, it arises from random mutation-selection dynamics, or it reflects stability-constrained functional optimization. In the first case, marginal stability ensures efficient turnover of aged and damaged proteins and reuse of amino acids; a too stable fold may be hard to degrade. In the second case, because mutations arise randomly and anything random done to an optimized system tends to reduce optimality, protein stability is constantly challenged by mutations that destabilize by perhaps 5 kJ mol<sup>-1</sup> on average (Tokuriki *et al.*, 2007), and responsive selection keeps the protein stable (Taverna and Goldstein, 2002; Goldstein, 2011). If so, marginal stability is not a selected trait but a consequence of the predominance of random drift, with mutation-selection dynamics constantly playing out near the denaturation threshold. Third, optimization of function occurs under the constraint of preventing denaturation. If so, marginal stability is not a selected trait or a consequence of random drift but reflects maximal trading of stability for function by investing protein fold-free energy to minimize transition state barriers of enzymes (Warshel, 1998). Each explanation does not exclude the others, as trade-offs and drift depend greatly on the protein, phenotype, and population, and they can ultimately be linked to the cost of managing the overall proteome, as discussed below.

### The main determinants of evolution rate

To understand the main drivers of evolution we must first understand the protein properties that mostly determine evolutionary rates in proteins on longer time scales. This rate is also used to

**Table 1.** Important correlators of the evolution rate and size of proteins

Features that slow evolution	Effect	Name
Functional active sites	Sites directly involved in e.g. recognition, substrate binding, and catalysis are highly conserved (Blundell and Wood, 1975; Casari <i>et al.</i> , 1995)	Function-rate (F-R) anti-correlation (sequence conservation)
High expression	Highly expressed proteins (measured by mRNA levels) evolve more slowly (Pál <i>et al.</i> , 2001; Drummond <i>et al.</i> , 2005)	Expression-rate (E-R) anti-correlation
Intracellular location	Intracellular proteins evolve more slowly than extracellular proteins (Winter <i>et al.</i> , 2004; Julenius and Pedersen 2006)	Secretion-rate correlation
Buried amino acid sites	Interior sites evolve more slowly than solvent-exposed sites (Overington <i>et al.</i> , 1992; Goldman <i>et al.</i> , 1998; Ramsey <i>et al.</i> , 2011)	Buried-rate (B-R) anti-correlation
Small size	Smaller proteins, all-else being equal, evolve slowly (Bloom <i>et al.</i> , 2006a). Small proteins are less evolvable due to larger functional density (Zuckerandl 1976)	Size-rate (S-R) correlation; functional density
Small contact density/fraction of buried sites	Proteins with smaller fractions of buried sites or contact density evolve slowly (both strongly correlated with size)(Bloom <i>et al.</i> , 2006a, 2006b)	Size-rate (S-R) correlation

classify and predict the functional impact of human variants e.g. in relation to disease (Glaser *et al.*, 2003; Capra and Singh, 2007; Thusberg *et al.*, 2011; Tang *et al.*, 2019). Table 1 provides an overview of the most important relationships between a protein's properties and its evolution rate. As easily verified from sequence alignment, active sites in proteins are highly conserved due to strong purifying selection, because random deleterious mutations impair fitness more in highly optimized parts of the protein. Related to this, solvent-exposed sites in contrast evolve faster than average, consistent with their typically smaller functional and structural effects on the overall protein (Overington *et al.*, 1992; Goldman *et al.*, 1998; Ramsey *et al.*, 2011).

The strongest descriptor of evolutionary rate is protein abundance or equally, mRNA levels, as these correlate (Gygi *et al.*, 1999); it typically spans 5–7 orders of magnitude in eukaryotes (Jansen and Gerstein, 2000; Ghaemmaghami *et al.*, 2003; Beck *et al.*, 2011; Milo 2013). High expression is associated with slower protein evolution in both prokaryotes (Sharp, 1991; Rocha and Danchin, 2004) and eukaryotes (Pál *et al.*, 2001), including mammals (Jordan *et al.*, 2004; Zhang and Li 2004), a phenomenon known as the expression-rate (E-R) anti-correlation (Drummond *et al.*, 2005; Bloom *et al.*, 2006a). Protein expression may explain half of the evolutionary rate variation in yeast (Drummond *et al.*, 2006) indicating a universal driving force of evolution. This remarkable relationship has been studied using many biophysical models focusing on protein stability, misfolding avoidance, and flexibility (Lobkovsky *et al.*, 2010; Geiler-Samerotte *et al.*, 2011; Wylie and Shakhnovich, 2011; Liberles *et al.*, 2012; Serohijos *et al.*, 2012; Yang *et al.*, 2012; Kepp and Dasmeh, 2014; Sikosek and Chan, 2014). All-else-being-equal, a protein's fitness impact should be proportional to its cellular abundance regardless of the specific selection pressure. Thus, any fitness function that scales with protein abundance may seem reasonable. Such models can explain about 60% of site-variations in the evolutionary rate (McInerney, 2006; Echave *et al.*, 2016). Protein stability has mainly been related to fitness via the copy number of misfolded proteins, assuming one-step unfolding (Serohijos *et al.*, 2012; Dasmeh *et al.*, 2014a). These ideas are expanded further below. To summarize the tendencies of Table 1, compared to the average protein, the slowly evolving protein tends to be highly expressed, intracellular, smaller than average, and have a higher functional density, i.e. more important sites *relatively to its size*.

The E-R anti-correlation has been explained (Drummond and Wilke, 2008, 2009) as a selection against inefficient translation leading to toxic misfolded proteins, a theory originally proposed by Kurland and Ehrenberg (Ehrenberg and Kurland, 1984; Kurland and Ehrenberg, 1984, 1987). Protein synthesis is inherently error-prone, and translation operates with typical missense error rates of 1/1000 to 1/10 000 (Kurland and Ehrenberg, 1987). Considering the typical lengths (~100–1000) and total abundance of proteins ( $10^8$ ) in eukaryotic cells, one can expect  $10^{10}$ – $10^{11}$  protein-incorporated amino acids to exist at any time. Without error correction this could imply the constant existence of  $10^6$ – $10^8$  erroneous amino acids in a typical eukaryote cell. This would make translation-error induced proteome variation of similar importance as typical, mostly heterozygote, natural sequence variation in a population. This of course raises the question how much of the actual observed proteome variation is due to genetic inheritance, somatic mutations, and translation errors. To be sure, one needs to sequence each gene and protein many times for several cells. Regardless of this complication, it is clear that the proteome varies much more in composition than implied by genetic variance alone.

Considering this, because the typical non-native residue destabilizes by  $\sim 5$  kJ mol<sup>-1</sup> (Tokuriki *et al.*, 2007), as much as 10% of a proteome could be less stable than commonly assumed purely from wild-type sequence. For a cell with  $10^8$  proteins, this implies that  $10^7$  protein copies are randomly destabilized and subject to higher turnover than expected from their wild-type sequence. Post-translational modifications and specific degrons further diversify the proteome and complicate turnover further. Considering this, the additional destabilization from new arising mutations will aggravate costs only if the affected protein is quite abundant or subject to high turnover.

If the misfolded protein is selected against, regardless of the reason, highly expressed proteins are under stronger selection pressure because the copy number of misfolded proteins  $U_i$  scales with the total abundance of the protein  $A_i$ . Drummond and co-workers suggested a fitness function  $\Phi$  depending exponentially on the total copy number of all misfolded proteins  $U = \sum U_i$ , with an unknown scaling constant  $c$  (Drummond and Wilke, 2008):

$$\Phi \propto \exp(-cU) \quad (1)$$

The constant  $c$  can be derived from fundamental and simple assumptions and related directly to the cost of protein turnover, as discussed below.

## The theory of proteome cost minimization

Darwin's theory of selection and the theory of nearly neutral evolution (Kimura, 1962, 1991; Ohta, 1992) together explain evolution as a process of selection and drift, whereas structural biology explains the molecular language of evolution via the central dogma. However, a complete theory of evolution requires us to also know the properties of the evolving protein that contributes to the organism phenotype, why it contributes, to what extent it contributes, and how this affects the wider evolution of the population in its ecological and historical context. As discussed extensively in the literature, it is increasingly clear that the functional traits selected for in classical positive Darwinian evolution have relatively little importance in many cases relative to other, partly hidden and perhaps universal properties of the proteins (Hurst and Smith, 1999; Bloom and Adami, 2003, 2004; Drummond *et al.*, 2005; Lobkovsky *et al.*, 2010; Wylie and Shakhnovich, 2011).

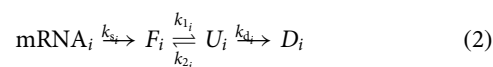
The most obvious universal property subject to selection pressure is arguably the cellular energy state. Before the era of structural biology and proteomics, Boltzmann (1886) and Schrödinger (1944) already speculated that life characteristically represents a well-defined organized (low-entropy) structure that maintains a thermodynamic non-equilibrium state relative to its high-entropy surroundings by constant energy turnover and associated heat dispersion. By this definition, expansion of life (fitness) implies expansion of this energy turnover. Lotka applied these ideas to Darwin's selection theory via his maximum power principle, arguing that evolution occurs by selection of the most energy-efficient organisms (Lotka, 1922). These ideas were then expanded into a much broader ecological view by Odum (1988). Thermodynamically, the system most capable of maintaining its structure by energy dissipation and with the ability to grow and reproduce these structures will prevail over other similar systems, and thus, be most fit.

The theory of proteome cost minimization (PCM) presented below was inspired by these views and further supported by the observations of consistent cost-bias in amino acid use across all kingdoms of life first discovered by Akashi and Gojbori (2002). These findings were confirmed by Swire (2007) and explained in a fitness model by Wagner who showed, among other things, that gene duplications are highly selected against in terms of cellular energy costs (Wagner, 2005). The theory builds substantially on Wagner's seminal quantitative considerations (Wagner, 2005) and the important considerations of Brown *et al.* (1993) who used Lotka's ansatz to explain mass and size optima of biological taxa in terms of evolutionary fitness caused by the different scaling of metabolic rates and reproductive rates with mass. The theory's central ansatz, inspired by these minds, is as follows: 'Fitness is proportional to the energy per time unit available for reproduction after subtracting (proteome) maintenance costs'. Because fitness always has to be measured relative to a wild type after an instant of time, the energy of interest becomes a power (measured in watts or  $\text{J s}^{-1}$ ) as in Lotka's original thinking, and as such directly relates to the respiration rate of the organism, as discussed below.

The mechanistic basis for the theory is that (i) protein degradation increases many-fold with the lack of structure and partial unfolding in protein copies (Gsponer *et al.*, 2008) and (ii) the cost of protein turnover is more than half of total metabolic costs in growing microorganisms (Harold, 1987), and at least 20% in humans (Waterlow, 1995). Accordingly, any increase in these costs reduces the energy available for other energy-demanding

processes, notably reproduction (fitness) of microorganisms (Dasmeh and Kepp, 2017) and cell signaling (cognition) in higher organisms (Kepp, 2019). One of many implications of the theory is that selection against misfolded proteins and toxicity of misfolding proteins measured in cell viability assays is *not due to a specific toxic molecular mode of action as widely assumed, but to the generic adenosine triphosphate (ATP) burden of turning over the misfolded proteins within the cell.*

In its simplest form, which is easily expanded, we assume a life cycle of a protein  $i$  as:



$F_i$  represents the folded proteins,  $U_i$  represents misfolded proteins, and  $D_i$  represent the degradation products, many of which are recycled for use in other proteins; the rate constant of each process is specific to the protein in question. Because the ultimate selection pressure acts only on  $U_i$ , one can easily relax the assumption of one-step unfolding to account for complex situations.

$k_{di}$  is the rate constant (in units of protein molecules per s) for degrading the misfolded protein copies. The *in vivo* rate constants reflect the half-life ( $t_{1/2}$ ) of the fully folded protein, and can thus be written at steady state as:

$$k'_{di} = \frac{k_{di}k_{1i}}{k_{2i}} = \frac{k_{di}}{K_{fi}} = \frac{\ln 2}{t_{1/2}} \quad (3)$$

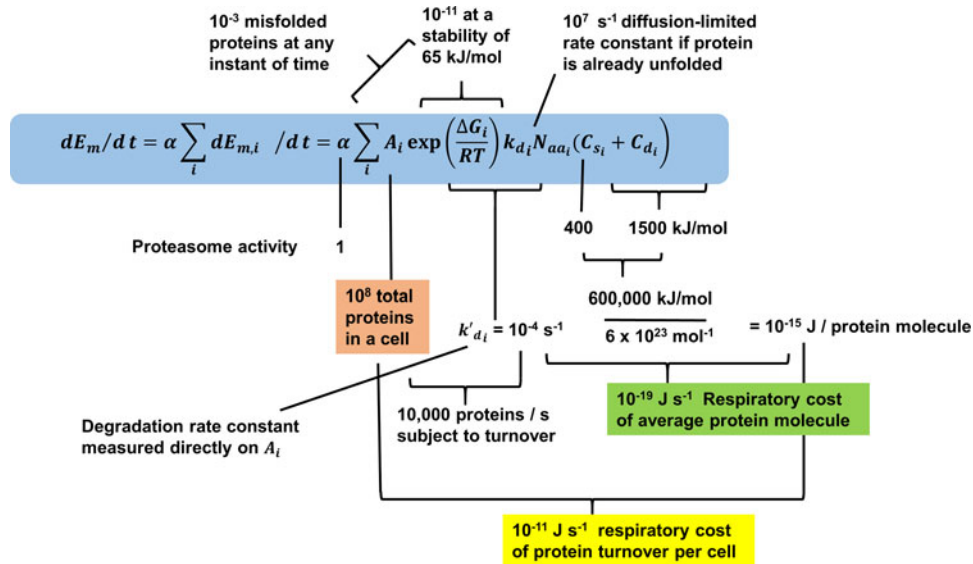
which varies substantially with the protein  $i$ , giving half-lives from minutes to days (Hargrove and Schmidt, 1989). The model assumes that unfolded protein copies are always kept at a very small number in the cell, compared to folded copies, such that  $k_{2i}$  is much larger than  $k_{di}$  and  $k_{1i}$ . This is generally a good approximation, because  $k'_{di}$  is typically of the order of  $10^{-4} \text{ s}^{-1}$  but with order-of-magnitude variations. In contrast,  $k_{di}$  acts directly on already misfolded protein and represents the rate of protein degradation if the chemical activation barrier to unfolding has been removed. Thus,  $k_{di}$  is limited by the number of active proteases, the diffusion and proper orientation of the exposed peptide bond, and the actual  $k_{cat}/K_M$  of the proteases, with an upper limit of perhaps  $10^6$  to  $10^8 \text{ M}^{-1} \text{ s}^{-1}$  per peptide bond hydrolysis (Wolfenden and Snider, 2001; Bar-Even *et al.*, 2011). In terms of steady-state turnover, misfolded proteins are immediately targeted for degradation (Gsponer *et al.*, 2008) and recruited by the ubiquitin-proteasome pathway that takes the protein out of the pool, and thus this process is not rate-limiting the overall protein flux but arguably operates near the diffusion limit.

Assuming one-step misfolding,  $U_i$  is related to the folding free energy of the protein  $\Delta G_i = -RT \ln(K_{fi})$  via the equilibrium constant  $K_{fi} = F_i/U_i$ :

$$U_i = A_i \left( \frac{1}{1 + \exp(-\Delta G_i/RT)} \right) \approx A_i \exp\left(\frac{\Delta G_i}{RT}\right) \quad (4)$$

The last expression follows if there are many more folded than unfolded copies of the protein, which is almost always the case. Because folding equilibrium constants easily reach  $10^{11}$  for a protein of typical stability ( $65 \text{ kJ mol}^{-1}$  at  $37^\circ \text{C}$ ), the number of misfolded proteins at any given time is typically negligible, as they are immediately subject to turnover. Reasonable experimental values





**Fig. 1.** Schematic overview of order-of-magnitude terms of the PCM model. Typical values for yeast used as example. All values are subject to the well-known variations in copy numbers of individual proteins, degradation rate constants, length of proteins, and total number of proteins copies in a cell.

of  $k_{d_i} = 10^7 \text{ s}^{-1}$ ,  $K_{f_i} = 10^{11}$ , and  $k'_{d_i} = 10^{-4} \text{ s}^{-1}$  satisfy the relationship in Eq. (3) and thus justify the use of Eq. (2).

Equation (4) is well established and was first used in a fitness function by Bloom *et al.* (2004) and has been specifically used to explain some of the E-R anticorrelation (Serohijos *et al.*, 2012) and additional variations in evolutionary rates (Dasmeh *et al.*, 2014a). The advantage of this expression is that we can relate the number of misfolded protein copies, which is the property selected upon, directly to the total copy number  $A_i$  of the protein within the cell and to its thermodynamic stability, via the free energy of folding  $\Delta G_i$  (a negative number in  $\text{kJ mol}^{-1}$ ).  $RT$  is the thermal energy of the cell, and thus temperature enters directly as a fundamental physical parameter determining proteome  $U_i$  and ultimately cellular proteome costs and fitness, as discussed further below.

The critical step is now to write the fraction of the total respiration rate (in watts, or  $\text{J s}^{-1}$ ) of the cell due to the maintenance of a single protein  $i$ :

$$\frac{dE_{m,i}}{dt} = A_i \exp\left(\frac{\Delta G_i}{RT}\right) k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i}) \quad (5)$$

In this equation, in addition to the parameters already described above,  $N_{aa_i}$  represents the number of amino acids in the protein  $i$ , and the cost constants  $C_{s_i}$  and  $C_{d_i}$  describe the average synthetic and degradation cost per amino acid in protein  $i$  in units of  $\text{J}$  (Kepp and Dasmeh, 2014).

For the whole proteome of the cell, we can write the *total* cost per time unit as the sum of the costs of maintaining steady-state folded protein copy numbers within the cell:

$$\frac{dE_m}{dt} = \alpha \sum_i \frac{dE_{m,i}}{dt} = \alpha \sum_i A_i \exp\left(\frac{\Delta G_i}{RT}\right) k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i}) \quad (6)$$

Importantly, we see that the total energy costs scale with  $A_i$ . Because  $A_i$  varies substantially for different proteins, e.g. from zero to a million, some proteins are much more important to the cell's energy budget than others. The scaling constant  $\alpha$

represents the activity of the proteasome, which may be controlled with proteasome inhibitors, but a slight expansion of this expression can be done to  $(\alpha + \beta + \dots)$  taking into account the contributions of various degradation pathways (lysosome, proteasome, effects of N-end rule, etc.) to the overall turnover. Figure 1 summarizes some typical values for the parameters of the model applicable to eukaryote cells.

### Selection dynamics of PCM

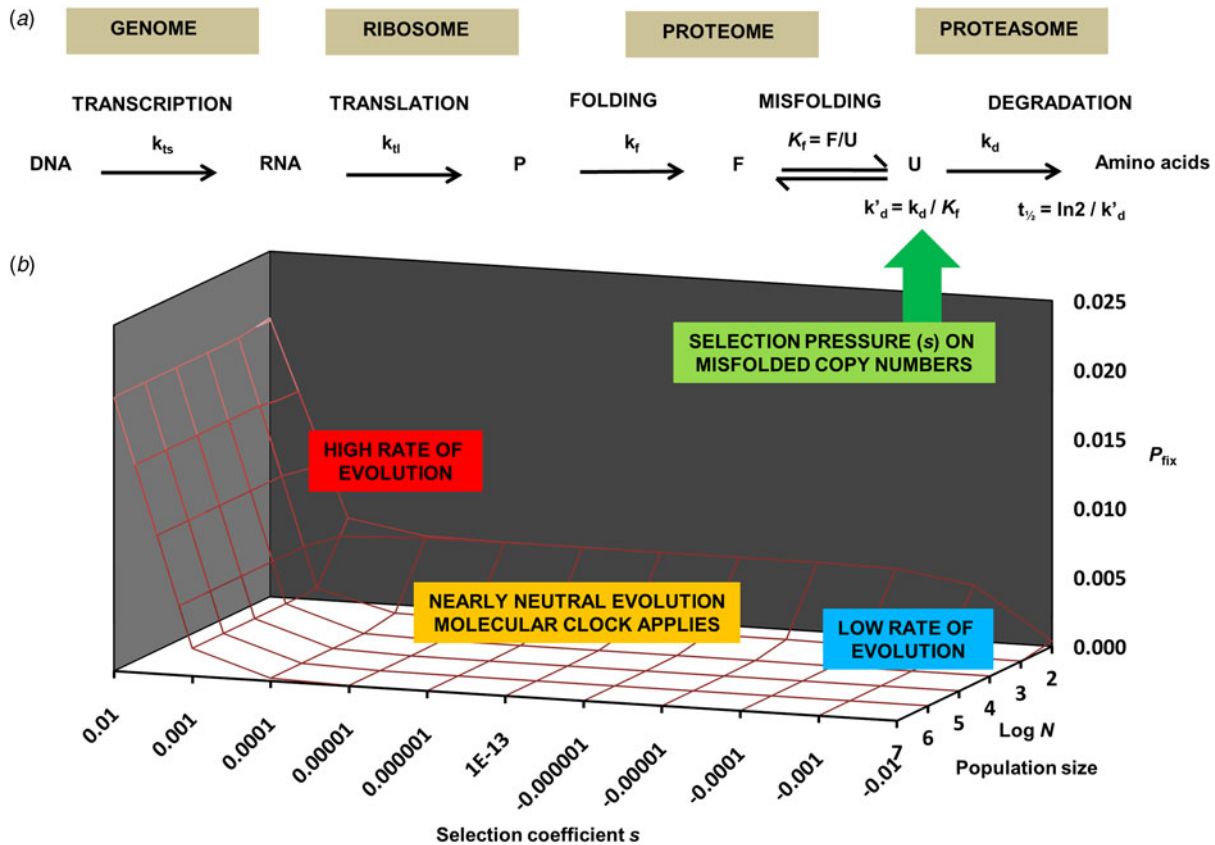
To understand how protein turnover costs affect evolution, we now use the central ansatz that fitness scales with the energy available for reproduction  $dE_r/dt$  after subtracting the proteome costs of Eq. (6) from the total energy available to the cell either by production or supply,  $dE_t/dt$ , divided by the respiration rate needed to run an individual, also taken to  $dE_t/dt$ :

$$\Phi = \frac{dE_r/dt}{dE_t/dt} = \frac{dE_t/dt - dE_m/dt}{dE_t/dt} = 1 - \frac{dE_m/dt}{dE_t/dt} \quad (7)$$

The division by  $dE_t/dt$  formally ensures a dimensionless fitness function. For simplicity, we ignore the non-proteome energy costs because the purpose is to show that the cost of the proteome exerts a major effect on evolution by itself. Assuming that the total energy production is constant for all competing cells, minimization of  $dE_m/dt$  maximizes fitness. When a new mutation arises in protein  $i$ , the selection coefficient is:

$$s_i(M) = \frac{\Phi_i(M)}{\Phi_i(WT)} - 1 = \frac{\Phi_i(M) - \Phi_i(WT)}{\Phi_i(WT)} = \frac{dE_m/dt(WT) - dE_m/dt(M)}{dE_t/dt(WT) - dE_m/dt(WT)} \quad (8)$$

For clarity, we have assumed that the mutation only affects maintenance turnover costs and not energy production, and thus the total energy produced is the same before and after mutation and cancels in Eq. (8). If we further neglect epistasis,



**Fig. 2.** (a) Schematic overview of the processes of protein turnover, with the central dogma to the left and the proteome maintenance, the concern of the present paper, to the right. (b) Probability of fixation ( $P_{fix}$ ) plotted against selection coefficient  $s$  and  $\log N$  (effective population size). Beneficial mutations with  $s > 0.001$  have relevant  $P_{fix}$  of more than 1% for most populations. Only in very small populations ( $< 100$ ) do other mutations get fixated ( $P_{fix} \sim 1\%$ ), and neutral and slightly deleterious mutations become fixated to a similar extent until  $s$  approaches  $-1/N$ , whence  $P_{fix}$  rapidly decreases.

selection only acts on the mutated protein  $i$ :

$$s_i(M) = \frac{A_i \exp(\Delta G_i/RT) k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})(WT) - A_i \exp(\Delta G_i/RT) k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})(M)}{dE_r/dt(WT) - A_i \exp(\Delta G_i/RT) k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})(WT)} \quad (9)$$

This selection coefficient is a function only of protein properties, scaled by the general energy spent for reproduction of the organism,  $dE_r/dt(WT)$ , which can be taken as a constant of the order of  $10^{-11} \text{ J s}^{-1}$  (Harold, 1987). It is perhaps more convenient to write Eq. (9) in terms of copy numbers and half-lives ( $t_{1/2}$ ) which can be measured in live cells:

$$s_i(M) = \frac{A_i N_{aa_i} (C_{s_i} + C_{d_i}) \ln 2 / t_{1/2}(WT) - A_i N_{aa_i} (C_{s_i} + C_{d_i}) \ln 2 / t_{1/2}(M)}{dE_r/dt(WT)} \quad (10)$$

where we have used the relationship:

$$k'_{d_i} = \frac{\ln 2}{t_{1/2}} = \exp\left(\frac{\Delta G_i}{RT}\right) k_{d_i} \quad (11)$$

For a haploid organism, the probability of its fixation  $P_{fix}$  is approximately (Kimura, 1962; Ohta 1992):

$$P_{fix} = \frac{1 - e^{-s_i}}{1 - e^{-s_i N}} \cong \frac{s_i}{1 - e^{-s_i N}} \quad (12)$$

where  $N$  is the effective population size, and the last term comes from expanding the exponential of the small  $s_i$ . For neutral evolution, as  $s_i \rightarrow 0$ ,  $P_{fix} \rightarrow 1/N$ , and does not depend on any properties of the protein. At significant positive selection,  $s_i N$  is large,  $s_i$  is positive, and  $P_{fix} \rightarrow s_i$ . Very similar behavior applies to diploid organisms with slightly different factors of 2 and 4 (Kimura, 1962).

The simple kinetic scheme assumed for the PCM model is highlighted in Fig. 2a. From Eq. (10), considering the variations in the parameters, most of the proteome cost selection occurs by affecting the ratio  $A_i/t_{1/2}$ . Mutations that reduce the half-life of abundant proteins are thus particularly selected against. The typical behavior of  $P_{fix}$  with  $N$  and  $s_i$  is shown in Fig. 2b. The absolute rate of evolution  $\omega$  scales with the mutation rate and the probability of fixing new arising mutations:

$$\omega = u N P_{fix} = u \frac{s_i}{1 - e^{-s_i N}} \quad (13)$$

where  $u$  is the absolute mutation rate; this expression can be expanded by life history variables such as generation time (Martin and Palumbi, 1993), but this is beyond the scope here, as the proportionality of Eq. (13) generally applies, and  $P_{fix}$  thus measures evolution rate. For an optimized evolutionary system, a typical arising mutation has a negative selection coefficient; if small relative to  $1/N$ , it is subject to random fixation drift. From Eq. (13), such mutations will reduce the probability of

**Table 2.** Effect of arising mutants in a haploid organism ( $N_{aa_i} = 400$ ;  $dE_t/dt = 3 \times 10^{-11} \text{ J s}^{-1}$ )

	$A_i$	$\frac{k'_d(WT)}{k'_d(M)}$	$N_{aa_i}(C_{s_i} + C_{d_i})$	$dE_{m}/dt(WT)$	$dE_{m}/dt(M)$	$s_i(M)$	$P_{fix}$ $N = 10^6$	$P_{fix}$ $N = 10^4$
Slightly deleterious mutant that increase $k'_d$ , or $A_i$ , 10-fold (e.g. from 60 to 54 kJ mol <sup>-1</sup> stability at 37 °C)								
Total proteome	$10^8$	$\frac{10^{-4} \text{ s}^{-1}}{10^{-3} \text{ s}^{-1}}$	$10^{-15} \text{ J per protein}$	$10^{-11} \text{ J s}^{-1}$	$10^{-10} \text{ J s}^{-1}$	Cell dies (proteome destabilization corresponds to $T = 72 \text{ °C}$ )		
Typical protein	$10^3$	$\frac{10^{-4} \text{ s}^{-1}}{10^{-3} \text{ s}^{-1}}$	$10^{-15} \text{ J per protein}$	$10^{-16} \text{ J s}^{-1}$	$10^{-15} \text{ J s}^{-1}$	$-4.5 \times 10^{-5}$	$<10^{-20}$	$7.9 \times 10^{-5}$
Abundant protein	$10^5$	$\frac{10^{-4} \text{ s}^{-1}}{10^{-3} \text{ s}^{-1}}$	$10^{-15} \text{ J per protein}$	$10^{-14} \text{ J s}^{-1}$	$10^{-13} \text{ J s}^{-1}$	$-4.5 \times 10^{-3}$	$<10^{-20}$	$<10^{-20}$
Short-lived protein	$10^3$	$\frac{10^{-2} \text{ s}^{-1}}{10^{-1} \text{ s}^{-1}}$	$10^{-15} \text{ J per protein}$	$10^{-14} \text{ J s}^{-1}$	$10^{-13} \text{ J s}^{-1}$	$-4.5 \times 10^{-3}$	$<10^{-20}$	$<10^{-20}$
Positive selection of slightly beneficial mutant that decreases $k'_d$ , 10-fold								
Typical protein	$10^3$	$\frac{10^{-4} \text{ s}^{-1}}{10^{-5} \text{ s}^{-1}}$	$10^{-15} \text{ J per protein}$	$10^{-14} \text{ J s}^{-1}$	$10^{-15} \text{ J s}^{-1}$	$4.5 \times 10^{-6}$	$4.6 \times 10^{-6}$	$1.0 \times 10^{-4}$
Abundant protein	$10^5$	$\frac{10^{-4} \text{ s}^{-1}}{10^{-5} \text{ s}^{-1}}$	$10^{-15} \text{ J per protein}$	$10^{-16} \text{ J s}^{-1}$	$10^{-17} \text{ J s}^{-1}$	$4.5 \times 10^{-4}$	$4.5 \times 10^{-4}$	$4.5 \times 10^{-4}$
Neutral evolution (same for all protein properties, only depends on $N$ )							$10^{-6}$	$10^{-4}$

WT, wild-type value of property; M, Mutant value of property.

fixation (and evolution rate) in proportion to the size of the negative selection coefficient. Figure 2b also illustrates why the molecular clock is generally successful at dating phylogenies, because 90% of randomly occurring mutations in the relevant selection-fixation space are subject to neutral evolution.

To understand the slow evolution of abundant proteins discussed in the literature (Drummond *et al.*, 2005; Bloom *et al.*, 2006a; Drummond and Wilke, 2008), we should identify low values of  $P_{fix}$  in the evolution rate space of Fig. 2b. Most arising mutations (Fig. 2b) remain subject to nearly neutral evolution. However, more extreme selection coefficients will occur for highly abundant proteins, because the selection coefficient of a new arising mutation in a protein scales with the abundance and turnover rate of the affected protein. In contrast, less abundant proteins will typically have numerically smaller selection coefficients at any given effective population size. The next section gives a quantitative estimate of the fixation probabilities.

### Typical PCM selection pressures and fixation probabilities for yeast

Table 2 summarizes some typical selection scenarios in yeast cells. A typical yeast cell respire at  $\sim 1 \text{ J s}^{-1} \text{ g}^{-1}$  and has a mass of  $3 \times 10^{-11} \text{ g}$ , giving  $dE_t/dt \approx 3 \times 10^{-11} \text{ J s}^{-1}$ .  $C_{d_i}$  is perhaps 1 ATP per peptide bond or 30 kJ mol<sup>-1</sup> (Benaroudj *et al.*, 2003). The biosynthetic costs of the amino acids vary from 10 to 80 ATP (Wagner, 2005), the average amino acid composition of the yeast proteome gives  $\sim 25 \text{ ATP}$ , or  $750 \text{ kJ mol}^{-1}$  as typical. If half of the amino acids are recycled, neglecting amino acid transport cost (Waterlow, 1995), this reduces to  $375 \text{ kJ mol}^{-1}$ . Additional costs of the polypeptide chain synthesis, neglecting chaperones, is  $\sim 11\text{--}19 \text{ ATP}$ , or  $330\text{--}660 \text{ kJ mol}^{-1}$  (De Visser *et al.*, 1992). Amino acid transport and chaperones (which need to be synthesized independently) increase costs further. Under growth conditions where most selection probably occurred historically, very few amino acids are recycled, and thus the specific turnover costs per amino acid in a protein molecule ( $C_{s_i} + C_{d_i}$ ) may easily reach  $1500 \text{ kJ mol}^{-1}$ . However, the amino acid-specific

values vary little compared to the protein-specific  $k'_d$ , and thus we use a value of  $1500 \text{ kJ mol}^{-1}$  in Table 2. With a typical protein of 400 amino acids, this implies  $10^{-15} \text{ J s}^{-1}$  of turnover cost per protein molecule, which varies perhaps by 3–4 orders of magnitude, mostly due to  $N_{aa_i}$  (protein length) and  $C_{s_i}$  (the biosynthetic cost of the amino acids) consistent with the empirically known sequence biases (Akashi and Gojobori, 2002; Wagner, 2005; Swire, 2007).

The exponential of Eq. (1) can be expanded as  $1 - cU$  because the values of  $cU$  are much smaller than 1. Accordingly, the empirically proposed (Drummond and Wilke, 2008) fitness cost constant  $c$  can be expressed in terms of fundamental protein turnover parameters, and we argue that  $c$  is protein-specific. The PCM fitness function, Eq. (7), can be written as:

$$\Phi = \frac{dE_t/dt - \sum_i A_i \exp(\Delta G_i/RT) k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})}{dE_t/dt} = 1 - \frac{\sum_i U_i k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})}{dE_t/dt} \quad (14)$$

Comparing the exponential-expanded fitness functions  $1 - cU$  proposed by Drummond and Wilke (2008) and Eq. (14), the dimensionless protein-specific and effective total cost constants are:

$$c_i = \frac{k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})}{dE_t/dt}; \quad c = \frac{\sum_i U_i k_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})}{dE_t/dt U} \quad (15)$$

Separation of  $U_i$  from its cost constant  $c_i$  does not apply in general, as each type of unfolded protein has specific costs, and thus  $c$  represents an average cost of handling all misfolded proteins regardless of type. Using the typical values of  $k_{d_i} = 10^7 \text{ s}^{-1}$  and  $N_{aa_i} (C_{s_i} + C_{d_i}) = 10^{-15} \text{ J s}^{-1}$  (Fig. 1, Table 2) gives  $10^{-8} \text{ J s}^{-1}$  for one molecule of protein  $i$ . When dividing by  $dE_t/dt \sim 10^{-11} \text{ J s}^{-1}$ , this gives a cost constant  $c_i \sim 1000$ . Summing over all misfolded copies ( $U \sim 10^{-3}$ ) gives a correction to the fitness function of the order of unity, in agreement with energy allocated to reproduction and proteome turnover being of the similar magnitudes as total respiration rates of growing cells (Harold, 1987).

A single protein's contribution to fitness is proportional to its relative abundance, all else being equal. If  $A_i = 1000$ , then  $U_i = 10^{-8}$  misfolded copies of this particular protein exist at any time, using the typical parameters given in Fig. 1 and Table 2, giving a total contribution to fitness of  $10^{-5}$ . Typically arising, slightly deleterious mutations in typical proteins will affect evolution rates in small populations of the order of  $N \sim 10^4$ , which probably played a major role in evolution in the wild (Gillespie, 2001; Piganeau and Eyre-Walker, 2009), mainly because historic population bottlenecks dominate the apparent effective population size (Willis and Orr, 1993; Hawks *et al.*, 2000; Bouzat, 2010). The calculation example in Table 2 gives a fixation probability of  $7.9 \times 10^{-5}$  for such typical mutations.

However, some proteins are much more systemically important than such a typical protein. The most important contributor to  $c_i$  is the degradation rate constant  $k_{d_i}$ , which varies by many orders of magnitude for different proteins, and to obtain the fitness we need to multiply this constant by  $A_i$ , or equally, the fold-stability weighted  $U_i$ . Abundance can span 5–7 orders of magnitude (Jansen and Gerstein, 2000; Ghaemmaghami *et al.*, 2003; Beck *et al.*, 2011; Milo 2013), whereas protein length  $N_{aa_i}$  spans about three orders of magnitude, up to  $\sim 30\,000$  amino acids (e.g. titin), with a reasonably small variance of gamma-distributed protein sizes (Zhang, 2000). PCM theory thus suggests that selection acts both on expression level and protein length, as indeed seen experimentally (Bloom *et al.*, 2006a). In small populations ( $N = 10^4$ ), a typical slightly deleterious mutation (less stable by  $5\text{ kJ mol}^{-1}$ , or a 10-fold higher turnover rate) in a highly expressed protein ( $10^5$  copies) will have essentially no probability of fixation ( $< 10^{-20}$ , middle right, Table 2). Cost selection in such moderate-sized populations can thus explain the relatively slower evolution of abundant proteins.

Large effective populations can also contribute to the E-R anticorrelation: random mutation-selection dynamics resulting from purifying or compensatory selection of new residues after accepting slightly deleterious mutations occur more frequently in less abundant proteins that have more neutral selection coefficients. In contrast, these dynamics are less important near the steeper fitness optimum of the more optimized, abundant proteins that pose larger costs to the proteome. The relative importance of these two mechanisms depends on the historic effective population size and the population bottlenecks on long evolutionary timescales. One can model such effects by explicit evolution simulations but this is beyond the scope of the current study.

For comparison to experiment, it is more convenient to use the fitness function:

$$\Phi = 1 - \frac{\sum_i A_i N_{aa_i} (C_{s_i} + C_{d_i}) \ln 2 / t_{1/2i}}{dE_t/dt} \quad (16)$$

where  $t_{1/2i}$  is the experimental *in vivo* half-life of the protein  $i$ , which accounts for real cellular life-times distinct from biophysical protein stability, e.g. effects of the N-end rule (Varshavsky, 1997; Mogk *et al.*, 2007; Gibbs *et al.*, 2014). All the properties in Eq. (16) are either observable or deducible from the protein's sequence.

### Scaling relations of proteome costs: mass, metabolism, and eukaryote evolution

The examples given have centered on yeast as model cell, with  $\sum A_i = 10^8$ . Eukaryote cells vary greatly in size, the total copy number of proteins, and metabolic respiration rates, and

prokaryotes typically feature smaller volumes, protein copy numbers and lower metabolic total respiration rates by 2–3 orders of magnitude (Milo, 2013). The question then emerges how these orders-of-magnitude differences affect the proteome turnover and the associated effects described above. Proteins are degraded differently due to specific degrons of their sequences, but the overall rate of protein turnover typically scales with the general activity of the proteasome (except for those proteins that are not degraded by the proteasome). Accordingly, a scale factor of proteasome activity  $\alpha$  (Eq. (6)), as modulated by proteasome inhibitors, will be an important control parameter in experimental tests of the theory as well as in efforts to understand protein turnover in relation to cellular energy costs, cell viability, and fitness. Although long-term proteasome inhibition is toxic, mild instantaneous proteasome inhibition should prove a useful tool in testing some of the mechanisms described here.

Additional scaling relations are relevant to discuss. Notably, from Eq. (8), any scaling of the metabolic rate by a number  $a$  characteristic of the organism will not affect the selection coefficient, if the fraction of energy devoted to reproduction is constant, commonly between 0.1 and 0.7 of total respiration costs (Harold, 1987; Hawkins, 1991), because the advantage of the mutation with lowered maintenance costs can be considered a perturbation:

$$s_i(\text{scaled})(M) = \frac{adE_m/dt(WT) - adE_m/dt(M)}{adE_r/dt(WT) - adE_m/dt(WT)} = s_i(M) \quad (17)$$

This relation requires comparison of the mutant and wild-type proteins under the same growth conditions.

Based on cell volume and protein copy measurements and associated calculations (Milo, 2013), and using the assumption that a typical protein volume is  $10\,000\text{ \AA}^3$ , proteins take up 1–4% of the cell volume of any cell and more importantly, regardless of the cell type, across prokaryotes and eukaryotes, including human cells. From this, we conclude that the total protein copy number  $A_i$  scales approximately linearly with cell volume. In contrast, the basal specific metabolic rate of both cells and whole organisms tends to scale with  $M^{3/4}$ , rather than  $M$  (Kleiber's law) (Kleiber, 1932, 1947; Savage *et al.*, 2007). Size, all-else-being equal, lowers the specific surface area of the organism and thereby increases metabolic efficiency by reducing the mass-weighted thermodynamic force required to maintain the non-equilibrium boundary (reduced heat dispersion per unit of biomass). Size also potentially minimizes average, mass-specific chemical and electric signaling distances within the organism. Such scaling laws of mass and volume and their implication for bioenergetic costs were discussed by Lynch and Marinov (2015).

For these reasons, the specific resting metabolism decreases with volume or mass, and equally, with total protein copy number of the organism. Accordingly, size carries an evolutionary advantage of the order of the mass-specific metabolic rate, as explained in detail by Brown and co-worker who developed the framework relating mass to fitness (Brown *et al.*, 1993). The advantage is of the order of:

$$\begin{aligned} s(M) &= \frac{\Phi(M)}{\Phi(WT)} - 1 = \frac{dE_r/dt(M)}{dE_r/dt(WT)} - 1 \\ &= \frac{-aM^{3/4}(M)}{-aM^{3/4}(WT)} - 1 \sim \left( \frac{M(M)}{M(WT)} \right)^{3/4} - 1 \end{aligned} \quad (18)$$

However, as pointed out by Brown *et al.* (1993) whereas ecological life-history variables (e.g. foraging efficiency) favor large



organisms, the reproduction rate favors smaller organisms and scales with  $M^{-1/4}$ . Thus, organism size has an evolutionary optimum with respect to both energy and time, which is distinct for different taxa due to the different life-history variables and associated scaling parameters (Brown *et al.*, 1993). A yeast mutant with a larger size of 1%, all-else being equal, would thus be predicted by PCM theory to have a selective advantage of  $(1.01/1)^{3/4} - 1 = 0.007$  if all the saved energy is spent on reproduction. This energy is clearly enough to enforce positive selection at all relevant population sizes from  $10^2$  to  $10^7$ , including early population bottlenecks (Fig. 2).

Combining the ansatz of PCM theory (that fitness scales with the energy left for reproduction per time unit after subtracting maintenance costs) with Kleiber's law leads to several potentially important explanations for size advantage relevant to emergence of life in general and eukaryotes in particular. A central weakness of endosymbiont theory, not mentioned by the otherwise important reviews on this topic (Gray *et al.*, 1999; Lane 2011), is the problem of evolutionary advantage *immediately* after the symbiosis event. The argument goes as follows: at the very beginning, the actual process of symbiosis must have had immediate costs of intrusion and aligning the cellular machineries, and must thus also have provided immediate selective advantages in competition with non-symbiotic cells. According to PCM theory, fitness scales with energy left for reproduction, and thus the immediate total maintenance costs must have reduced.

Imagine a simple doubling of the cell size by a unification event. All else being equal, the new organism would carry the double amount of proteins, the double volume, the double mass, and would require the double amount of energy to reproduce these cell constituents, giving the same fitness as the competing non-symbiotic cells, but then reduced by the costs of the endosymbiosis event itself. However, the immediate advantage offered by reducing the specific surface area of the ancestral eukaryote cell would reduce the basal metabolic maintenance rate. The saved energy could then be immediately converted into a larger fraction of the total energy budget being devoted to the proteome of larger cells and organisms, thus compensating the cost of the actual symbiosis event. If this is correct, endosymbiosis will be successful only when and if the mass-specific metabolic rate saved by mass increase outweighs the energy costs of the symbiosis event itself.

### Evidence for PCM during evolution

Some support for the theory of proteome cost minimization is summarized in Table 3. The following section discusses some of these facts briefly.

#### Major evolutionary events mainly represented bioenergetic advantages

During the longest and earliest timescales where much of the primary cellular biochemistry evolved, unicellular growth conditions provided the context for the evolutionary innovation both in terms of respiration and photosynthesis (Blankenship, 1992; Sousa *et al.*, 2013). Most of the important biochemical pathways being at least qualitatively evolved at the point when eukaryotes had formed (Nisbet and Sleep, 2001; McGuinness, 2010). Early qualitative innovations such as the electron transport chain, fatty acid and amino acid metabolism, and photosynthesis indicate the primary importance of obtaining and maintaining the bioenergy production (Sousa *et al.*, 2013), a tendency further

documented by the rise of eukaryotes whose advantages largely related to energy efficiency by outsourcing and optimizing energy production as argued above and elsewhere (Margulis, 1968, 1975; Gray *et al.*, 1999; Lane 2011).

#### Energy surplus determines growth of microorganisms

For unicellular organisms, the cell cycle determining the decision to grow (and thus contribute to population fitness) is largely based on an assessment of available energy (Cai and Tu 2012): thus, budding yeast grows during the G1 phase until the nutrient level determines whether it commits to reproduction and enters the DNA biosynthesis S phase and subsequent mitosis, or if cell growth is arrested due to low resources (Cai and Tu 2012).

#### Protein turnover is very expensive

Protein turnover is typically the most or second-most expensive process in cells: At one extreme, protein synthesis may account for 3/4 of all energy spent in growing microorganisms (Harold, 1987). In humans, protein synthesis typically requires  $20 \text{ kJ kg}^{-1}$  body mass, or 20% of the basal metabolic rate to produce typically 300 g of protein per day (Reeds *et al.*, 1985; Waterlow, 1995). This number does not include regulation and degradation costs, RNA synthesis, and uncertain costs relating to nitrogen metabolism, reuse, transport, or synthesis of amino acids, which together are substantial (Reeds *et al.*, 1985; Hawkins, 1991). In mammals, protein degradation may cost 10–20% of total energy spent (Hawkins, 1991; Fraser and Rogers, 2007). Ubiquitin requires ATP to bind proteins targeted for degradation, and the lysosome and calcium-dependent proteases require ATP for active calcium and proton transport (Hawkins, 1991). These various features render protein turnover (synthesis and degradation) the most or second-most (next to ion pumping) energy-consuming process even in mammals.

#### Life uses cheap amino acids

The synthetic costs of the 20 amino acids vary roughly from the order of  $\sim 10$  (Glu, Ala, Gly, etc.) to  $\sim 75$  (Trp) phosphate bonds (Akashi and Gojobori, 2002; Heizer *et al.*, 2011). Biosynthetic costs explain some of the amino acid bias in sequences not due to translational efficiency and other effects (Craig and Weber, 1998; Akashi and Gojobori, 2002; Akashi, 2003) and can affect the rate of evolution (Barton *et al.*, 2010). Selection toward cheaper amino acids or smaller proteins can reduce total energy expenditure substantially, by an estimated 0.1% per  $\sim 4$  expensive amino acids (Akashi and Gojobori, 2002). A general evolutionary preference for synthetically cheap amino acids was first suggested (for aromatic residues in *Escherichia coli*) (Lobry and Gautier, 1994) and later demonstrated (Akashi and Gojobori, 2002) and confirmed by others (Wagner, 2005; Heizer *et al.*, 2006) in prokaryotes, where cheaper amino acids tend to be used more in highly expressed proteins across functional classes, with similar observations seen for yeast (Raiford *et al.*, 2008). These findings have been confirmed in many cases (Garat and Musto, 2000; Kahali *et al.*, 2007; Raiford *et al.*, 2008; Heizer *et al.*, 2011) including mammals (Heizer *et al.*, 2011). Biosynthetic cost minimization as an evolutionary driver was identified first in certain bacteria (Akashi and Gojobori, 2002; Schaber *et al.*, 2005) and later in all domains of life (Swire, 2007). Cys is apparently not significantly selected

**Table 3.** Events and facts supporting the PCM theory

Observation	Interpretation
Protein turnover is very expensive, in particular in growing microorganisms	The cost of handling the proteome is the most or second-most costly process in many cells (Reeds <i>et al.</i> , 1985; Waterlow 1995; Fraser and Rogers 2007), and can dominate total energy costs in growing microorganisms (Harold 1987)
Energy surplus determines growth of microorganisms	In the yeast cell cycle, available energy determines whether the cell commits to reproduction or if growth is arrested (Cai and Tu 2012)
All kingdoms of life favor synthetically cheap amino acids (Garat and Musto 2000; Akashi and Gojobori 2002; Schaber <i>et al.</i> , 2005; Kahali <i>et al.</i> , 2007; Swire 2007; Raiford <i>et al.</i> , 2008; Heizer <i>et al.</i> , 2011)	Cheaper amino acids confer a selective advantage by lowering overall protein synthesis costs of the organism
Cheap amino acids are more used in highly expressed proteins (Ikemura 1985; Seligmann 2003; Wagner 2005; Swire 2007)	Abundant proteins contribute more to total fitness, making cheaper amino acids are particularly advantageous, supporting a relation to both abundance and protein-specific costs
Extracellular proteins use cheaper amino acids (Smith and Chapman 2010)	Extracellular proteins are not recycled and thus, their net amino acid costs are larger per protein copy, this seems to have been selected against by favoring cheap extracellular amino acid use
Highly expressed proteins tend to be smaller (Ikemura 1985; Bloom <i>et al.</i> , 2006a)	Seen in 27 of 31 functional categories of yeast, with 12 classes significant (Ikemura 1985; Bloom <i>et al.</i> , 2006a). Length is inversely related to gel-derived protein abundance (Futcher <i>et al.</i> , 1999)
Cheap amino acids are used in large proteins. (Ikemura 1985; Seligmann 2003)	All-else-being-equal, larger proteins constitute larger turnover costs (weighted by their copy numbers) and thus are more relevant for overall PCM.
Large proteins tend to be more stable	Large proteins tend to be more stable (significant but with large variation) (Sawle and Ghosh 2011)
Streamlining theory (the theory that selection favors minimal cell complexity) (Giovannoni <i>et al.</i> , 2014)	The intense streamlining of prokaryote genomes (Lynch 2006; Giovannoni <i>et al.</i> , 2014) reflects selection pressure either via energy, time, or both, and is thus explained by PCM theory
Parasites feature reductive evolution on biosynthesis and metabolism (Loftus <i>et al.</i> , 2005)	Parasites mainly get their energy and nutrients from the host and thus can increase fitness by adaptive loss of biosynthetic and metabolic pathways
Genes with less intronic DNA more highly expressed (Urrutia and Hurst 2003)	Less introns probably reduce the cost of protein translation
Protein synthesis efficiency affects the age-dependent growth of blue mussels (Hawkins <i>et al.</i> , 1986)	Genetic differences in protein turnover efficiency contribute to fitness in some organisms
Misfolded proteins can reduce yeast fitness/growth by 3.2% (Geiler-Samerotte <i>et al.</i> , 2011)	Misfolded proteins impose a cost on the proteome in proportion to the steady state level of misfolded copies and their turnover rate (Eq. (9))
The endosymbiosis leading to eukaryotes was an energy optimization event (Margulis 1975; Lane 2011)	The specialized energy production in mitochondria and the associated genomic asymmetry gave rise to enormous expansions and innovations typical of Eukarya (Lane 2011)
Overflow metabolism (Warburg effect in cancer cells) (Basan <i>et al.</i> , 2015)	The shift in selection pressure from time to energy explains overflow metabolism, because fermentation is faster but respiration is cheaper
Cancer cells use cheaper amino acids (Zhang <i>et al.</i> , 2018)	Cancer cells use ATP-wise cheap amino acids during very fast growth, consistent with an advantage of minimizing proteome energy costs
Synthesis, not toxicity, explains evolution rates of overexpressed proteins (Plata <i>et al.</i> , 2010)	It is widely assumed that misfolded proteins are toxic by a specific mode of action. Plata <i>et al.</i> showed that turnover costs are more important for <i>E. coli</i> cell fate than toxicity at least for the studied proteins
Sickle-cell disease patients display doubling of protein turnover and 20% increase in resting metabolism (Badaloo <i>et al.</i> , 1989)	Mutations in hemoglobin lead to dysfunctional, instable proteins that are compensated by enhanced turnover and synthesis. The numbers suggest that 20% of the normal human metabolic rate is spent on protein turnover, fully consistent with consensus in the field (Hawkins 1991; Waterlow 1995)

for cost (Swire, 2007), perhaps relating to its unique involvement in highly conserved cystine bridges and metal sites.

### Prokaryote streamlining

The fact that prokaryotes have maintained their general morphology until today whereas Eukarya is represented by rich morphological diversity reflects the existence of some selection pressure

that kept prokaryotes simple but afforded major degrees of freedom to Eukarya. The well-known intense streamlining of the small efficient prokaryote genomes has led to the formulation of the so-called streamlining theory of microbial evolution (Lynch, 2006; Giovannoni *et al.*, 2014), which argues that streamlining toward small efficient genomes have been an ongoing selection pressure of prokaryote evolution. Fold structures are the phenotype ultimately selected upon, and structure-based

phylogeny implies that ancestral organisms can have been quite complex, but then later lost some of this complexity (Kurland and Harish, 2015; Harish and Kurland, 2017). This distinction between sequence and phenotype (fold structure) is also central to the debate on two versus three kingdoms of life (Mayr, 1998; Woese, 1998; Kurland and Harish, 2015). Streamlining can result from both selection pressures on time, energy, and space and fits the predictions of PCM theory, as discussed further below.

### Highly expressed proteins are more streamlined

Highly expressed genes tend to code for smaller proteins (Jansen and Gerstein, 2000) with less introns (Urrutia and Hurst, 2003), in support of selection pressure toward minimizing proteome handling costs. Selection against mistranslation can also be understood as selection against biosynthetic cost because translational efficiency is effectively a way to minimize the cost of expensive ‘proofreading’ and other machinery operating on mistranslated gene products (Ikemura, 1985). Additional support for the selection on highly abundant proteins directly relating to turnover costs is the well-known relationship between expression levels and protein half-life (Belle *et al.*, 2006).

### Unstable proteins reduce cell growth

Support for the PCM theory also comes from studies that compare the biophysical properties of overexpressed wild-type and mutant proteins directly. Destabilizing mutants of lacZ in *E. coli* reduce cell growth to a similar extent as wild-type protein expressed at the same level, arguing for quantity (expression levels subject to turnover) as the cause of toxicity rather than qualitative features of the protein variants (Plata *et al.*, 2010). An implication of this is that reduced cell viability in assays of overexpressed misfolding proteins, often used as models of neurodegenerative disease, may in fact reflect energy deficits as described by PCM theory. If so, misfolded proteins are generally not toxic by a specific mode of action (such as membrane pore formation or seeding of misfolding leading to loss of function) but rather because of the ATP costs (Kepp, 2019).

### Trading function for cost

Classical Darwinian evolution considers the struggle and selection for optimal function the primary mode of evolution (Richmond, 1970; Hurst, 2009). This aspect of Darwinism has dominated biochemical views of enzymes as perfectly optimized proficient catalysts that accelerate chemical reactions by orders of magnitude, implying that evolution strives toward optimal function *per se*, including maximal substrate turnover of enzymes (Radzicka and Wolfenden, 1995; Cannon *et al.*, 1996; Zhang and Houk, 2005). However, proteins are also subject to non-function selection pressures that are distinct from, and sometimes in conflict with, optimality of function (Hurst and Smith, 1999; Bloom and Adami, 2003; 2004; Drummond *et al.*, 2005; Lobkovsky *et al.*, 2010; Wylie and Shakhnovich, 2011). Indeed, actual comparison of enzyme kinetic parameters shows that many enzymes are distinctly suboptimal, most likely because of evolutionary and biophysical constraints (Bar-Even *et al.*, 2011).

A standard view is that proteins have evolved to use their excess fold-free energy to optimize the active sites for function, the most notable example being pre-organized active sites with electrostatic fields favoring the free energy of the transition states, to increase

$k_{cat}/K_M$  (Cannon *et al.*, 1996; Warshel, 1998; Adamczyk *et al.*, 2011; Morgenstern *et al.*, 2017; Fuller *et al.*, 2019). Although not directly pointed out by Warshel and co-workers, this mechanism contributes to making proteins marginally stable because, all-else-being equal, any potential excess fold-free energy has been diverted into optimizing the electrostatic field of the folded structure to reduce the transition state’s free energy and thereby increase catalytic proficiency. The mechanism also largely explains the widely observed stability-function trade-offs in protein engineering (Tokuriki *et al.*, 2008). Correspondingly, in the laboratory, without many biological constraints, function of a high-stability starting protein may be optimized beyond the level seen in the wild (Bloom *et al.*, 2006b; Tokuriki and Tawfik, 2009). This is particularly relevant in the context of ‘directed evolution’, i.e. the intended human evolution of new improved protein mutants employing yeast cells with short generation times in static environments where selection pressure can be effectively controlled (Francis and Hansche, 1972; Hall 1981).

PCM theory argues that even functional proficiency often evolved conditionally on cost. To appreciate this, we consider the requirement of a certain total substrate turnover of each enzyme per time unit to maintain homeostasis. The proficiency of function is for enzymes typically defined by  $k_{cat}$ , measuring how many substrate molecules convert into product per time unit per enzyme molecule. At steady-state, both the maximum turnover ( $V_{max}$ ) and the turnover at low substrate concentration are proportional to the total enzyme concentration  $[E]$  and  $k_{cat}$  (Northrop, 1998; English *et al.*, 2005).

Now consider a typical arising mutation in an enzyme  $i$  required to make a product at a certain rate, i.e.  $dP_i/dt$ . Because the protein is evolutionarily optimized (but not necessarily optimal), mutations will tend on average to be hypomorphic and reduce the turnover constant  $k_{cat,i}$  but with a broad scatter and many nearly neutral effects with a random chance of fixation. If the mutation reduces  $k_{cat,i}$  substantially, e.g. by modifying the active site, the substrate turnover will be greatly reduced, and the organism will need to increase the local enzyme concentration  $[E]$  by expressing more enzyme per time unit to maintain a comparable substrate turnover (compensatory expression), thereby increasing  $A_i$ . More specifically, the rate of product formed by enzyme  $i$  under Michaelis–Menten kinetics is (Cannon *et al.*, 1996; Northrop, 1998)

$$\frac{dP_i}{dt} = A_i k_{cat,i} \frac{[S]}{K_{M,i} + [S]} \quad (19)$$

Equation (19) represents the standard equation multiplied on both sides by the cell volume to convert from concentrations to absolute copy numbers. For simplicity, we can ignore the last term and assume zero-order kinetics in  $[S]$ , which represent selection of the enzyme for maximum rate at saturated substrate concentration when  $[S]$  is much larger than the Michaelis constant  $K_{M,i}$ . The cost of maintaining the enzyme is

$$\frac{dE_{m,i}}{dt} = A_i k'_{d_i} N_{aa_i} (C_{s_i} + C_{d_i}) \quad (20)$$

Accordingly, the specific cell-wide cost of maintaining steady state produced concentration of  $P_i$  is

$$\frac{dE_{m,i}}{dP_i} = \frac{A_i k'_{d_i} N_{aa_i} (C_{s_i} + C_{d_i})}{A_i k_{cat,i}} = \frac{k'_{d_i}}{k_{cat,i}} N_{aa_i} (C_{s_i} + C_{d_i}) \quad (21)$$

If measured in concentrations instead, the cost scales with the volume of the cell  $V_{cell}$  to which the steady state applies. We have

ignored the costs associated with producing the substrate and transporting the substrate and products, which can easily be included into the model.

Equation (21) predicts that the ratio of the two time constants for turnover of the enzyme and turnover of the substrate together define the cost of producing  $P_i$  at steady state. The two time constants are in units of  $s^{-1}$ , and  $N_{aa}(C_{S_i} + C_{d_i})$  is of the order of  $10^{-15}$  J for a typical protein. Considering again a typical arising mutation, even if  $k'_{d_i}$  is not increased (which it typically is), a reduction in  $k_{cat,i}$  of a typical hypomorphic mutation will require compensatory expression of the enzyme, increasing  $A_i$  to maintain the rate of production of  $P_i$ , Eq. (19). This increase in  $A_i$  will then increase the total cost of obtaining the product with the same factor (Eq. (20)). Equation (21) summarizes this cost–function relationship because  $k_{cat,i}$  and  $A_i$  are inversely related if homeostasis in  $P_i$  is required. If compensatory expression is 100%, a ten-fold reduction in the enzyme's  $k_{cat,i}$  requires a 10-fold increase in the enzyme's expression, and the specific and total costs of producing  $P_i$  increases 10-fold.

Accordingly, even mutations that only impair function also increase the proteome costs: a 10-fold increase in  $k'_{d_i}$  (loss of kinetic stability, misfolding) or decrease in  $k_{cat,i}$  will have approximately the same 10-fold increase in cellular costs, according to Eq. (21), ignoring the mutation-induced changes in the amino-acid synthesis and degradation costs. If required, the assumption of 100% compensatory expression can easily be modified by a scale factor between 0 and 1 in the above equations. Evidence for compensatory expression is well-known, a dramatic example being homozygous sickle cell disease (Table 3), where dysfunctional, unstable hemoglobin mutants cause a doubling of protein turnover and degradation in patients and a 20% increase in total resting metabolism (Badaloo *et al.*, 1989). Considerations of loss and gain of function mutations associated with other diseases may be viewed in this light (Kepp, 2015, 2019).

Because of the above considerations, we expect a function–cost trade-off acting during evolution of many proteins. We obtain the important possibility that *the main advantage of a mutant may not be a functional improvement of the protein per se, but a reduction of its cost per unit of function, in the simplest case the ratio  $k'_{d_i}/k_{cat,i}$* . Co-optimization of cost versus function is fundamental to many optimization processes and follows the basic principle that if several inputs are available at different functionality and price, the optimal system uses the input whose cost per unit of function is lowest. Such systems will tend to use less functional input if its cheaper price outweighs the loss of function. This suggests that at least some of the widely observed inverse relationships between function and stability (Tokuriki *et al.*, 2008; Bonet *et al.*, 2018; Du *et al.*, 2018) in reality reflect a cost–function trade-off as summarized by Eq. (21). The laboratory can change selection pressures drastically away from those in the wild, notably in the form of ‘directed evolution’ (Francis and Hansche, 1972; Hall 1981). In nature however, the situation is more complicated, because the stability affects the proteome costs and thus fitness. Newly arising mutations may impair both stability and function, but both have a direct negative fitness effect in terms of cost.

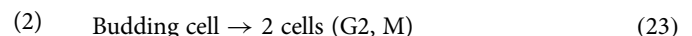
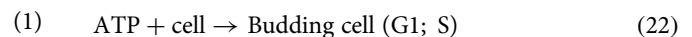
The theory thus predicts that highly abundant proteins, because they are more cost-selected, are more likely to display suboptimal functionality, all else being equal (after adjusting for other correlating variables such as size). The trade-offs will be habitat- and strategy-dependent, and the preferential use of very functional but expensive input may be restricted to high-nutrient habitats and growth media.

## Time or energy?

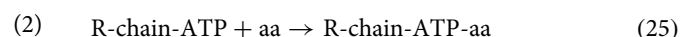
We expect that variations in the habitat's selection pressure should affect the proteome function–cost trade-offs. This should be evident when comparing organisms adapted to different environments. The most obvious biophysical properties of the habitat are time, energy, space, and temperature, which all enter directly in the model, Eq. (12). Selection for time, i.e. ‘survival of the fastest’, can be considered the default mode, and enters via the central ansatz of the theory, that ‘fitness is proportional to the energy per time unit available for reproduction after subtracting maintenance costs’, i.e. Eq. (6):  $\Phi = dE_r/dt = dE_f/dt - dE_m/dt$ . Fitness scales inversely with the time step  $dt$  required for directing a unit of surplus energy sufficient to complete a reproductive event. Temperature enters as a modifier of the protein stability's role in the turnover  $\Delta G_i/RT$ . We also note that a model of protein minimization driven mainly by considering space as a limiting parameter leads to some of the same consequences as protein cost minimization (Brown, 1991). Accordingly, all these biophysical properties may potentially act as selection pressures.

Reasonably, cells have been optimized to maximize growth rates by enabling their proteomes to be produced as fast as possible within the necessary function and stability restrictions. Translational speed and accuracy imply selection for smaller and more streamlined genomes, and accuracy mainly reflects the time–cost trade-off of correcting errors during protein synthesis rather than correcting them later in e.g. a misfolded protein (Kurland and Ehrenberg, 1984; Drummond *et al.*, 2005). We can thus reasonably view time as the ‘default mode’ of selection when energy is plentiful (i.e. survival of the ‘fastest’). If the growth rate is proportional to the synthesis rate of the proteome, then large, highly expressed, and slowly folding proteins will be growth-limiting either at the ribosome or during subsequent folding by chaperones of the rate-limiting proteins.

To account for both time and energy together, for simplicity we only consider two processes, one that is energy-limited and one that is time-limited:



In this simple model, if energy is limited, the cell will enter a dormant state and growth rates are controlled by energy efficiency of the proteome according to PCM theory. If energy is plentiful, growth rates are limited by the rate of producing the new cell, restricted by the speed of synthesizing the proteome rather than its cost. Other models of proteome optimization have emphasized translational speed and accuracy and minimization of protein size (Ehrenberg and Kurland, 1984; Brown, 1991) due to space restrictions on flux control. One can also consider analogous microkinetic models such as:



Here, the ATP needed for the ribosome (R) to catalyze chain elongation by an amino acid (aa) must be available to the ribosome, and if the concentration of ATP is low, then this step is rate-limiting the protein synthesis. If energy is plenty, then step



2, the chain elongation (and subsequent protein folding by e.g. chaperones) is limiting growth and subject to selection pressure.

It should be clear that both the cell-cycle and microkinetic model imply that both energy and time can be relevant selection modes, i.e. survival of the ‘fastest’ (scenario 2) survival of the ‘cheapest’ (scenario 1). One can consider  $r$ - and  $k$ -strategies as resulting from specialization toward these regimes. Experimentally, one may test the two cases via competitive growth assays with variable space and energy restrictions. Importantly, the two selection modes (time and energy) lead to several of the same implications, notably with a selective advantage for streamlining and particular selection on highly expressed proteins as they may limit both time and energy costs of growth (Wang *et al.*, 2011).

One recent study that casts light on this is a study of pathways choices among different sequenced organisms (Du *et al.*, 2018). The study found that different organisms select specific choices of precursor pathways based on both metabolic cost and synthetic efficiency. Cost selection occurs in energy-poor habitats, whereas in energy-rich habitats, the default selection mode is time. There are correlations between time and energy advantages. Notably, the synthesis time of expensive amino acids is all-else-being-equal long as more phosphate bonds must be recruited during synthesis. The cost of handling misfolded proteins can limit growth substantially, as seen in a case of  $\sim 3\%$  growth rate reduction in yeast upon folding-stability-impaired mutants of only one protein (YFP) (Geiler-Samerotte *et al.*, 2011).

The shift in selection pressure from time to energy can also explain the important phenomenon of overflow metabolism, the tendency of using more expensive, but faster fermentation rather than respiration during growth (Basan *et al.*, 2015). PCM theory implies that microorganisms shift to fermentation in rich habitats and growth media, because time is the main selection pressure, whereas in poorer habitats, respiration becomes favored and selected upon because energy is restrictive, although combinations of strategies will probably be common. The choice between these options depending on energy availability could be relevant to many growth assays, but perhaps also to the Warburg effect of cancer cells (Basan *et al.*, 2015). Cancer cells are remarkable by being under selection both for time and space in competition with each other against the selection pressure of the body’s immune system. Cancer cells tend to use cheaper amino acids (Zhang *et al.*, 2018), in accordance with PCM theory, but when energy is widely available, growth-limiting space and time restrictions would favor the Warburg effect over oxidative phosphorylation, although other contributing effects such as mutation impacts and oxygen availability are relevant as well.

### Temperature, thermostable proteins, and thermophilic organisms

As mentioned above, the habitat temperature also imposes a selection pressure on evolution according to the PCM theory, because it directly modifies protein stability  $\Delta G_i/RT$  and thereby, the fitness function, Eq. (11). To appreciate this, we used a sign convention of negative  $\Delta G_i$  for a stable protein, and the  $\Delta G_i$  is the optimal stability of the protein at its temperature of operation (sometimes called  $T^*$ ), typically reflecting to some extent the organism’s experienced extrema temperatures in the relevant habitat (Robertson and Murphy, 1997). The protein has been optimized to display its maximal stability at this  $T^*$ , with  $\Delta G_i$  typically harmonic in the temperature, and increasing or

decreasing the temperature away from  $T^*$  will thus increase the number of misfolded proteins  $U$ ; and increase the associated turnover costs, thereby reducing fitness, Eq. (11) (Robertson and Murphy, 1997).

Using the theory, we can better understand adaptation of proteomes to hot or cold environments (thermophiles and psychrophiles, respectively) (Li *et al.*, 2005; Mozo-Villiarías and Querol, 2006; Luke *et al.*, 2007; Fu *et al.*, 2010). Adaptations to a warmer habitat is largely expected to be a question of optimizing the proteome’s copy-number-weighted median protein  $T^*$  (the most representative  $T^*$  of the proteome of the cell) toward the  $T$  of the habitat, to minimize the average copy number of misfolded protein copies in the cell at any given time, again to minimize proteome costs and maximize energy available for reproduction. Many studies of thermophilic proteins and thermophilic adaptation may be seen in this light, without going into further details, as this is a large and complex topic (Tekaiia *et al.*, 2002; Sawle and Ghosh, 2011; Venev and Zeldovich, 2018), but the essential implications should be clear. In particular, thermophilic organisms are predicted to adjust protein thermostability mainly for the most abundant and quickly turned-over proteins that pose the largest economical cost to the proteome.

### PCM, aging, and neurodegenerative diseases

Proteome cost minimization has been argued to explain a substantial part of the evolution on longer evolutionary timescales, producing clear biases in the use of amino acids and explaining the E-R anti-correlation by slowing the probability of fixating new mutations in abundant, expensive proteins, and giving rise to important cost–function trade-offs. The evolution that shaped these relations mainly occurred in single-cell organisms, and it is thus of interest to consider whether the theory has implications also for evolution of higher organisms and in particular the evolution of aging.

A note is required first on intrinsically disordered proteins (IDPs), which make up a substantial fraction of all proteins in a typical cell. IDPs are disordered as part of their natural function, which can be expected to require structural plasticity or specific conformational changes as the local environment changes, or upon interaction with binding partners (Uversky *et al.*, 2008). The required disorder may lead to particular sensitivity and potential elevated cost of turnover. The common involvement of IDPs in protein misfolding diseases hints to the importance of proteome maintenance, which we argue should be counted in bioenergy units (Kepp, 2019).

All higher organisms use oxidative phosphorylation as the most effective energy-producing process, using the  $O_2$  of the planet’s atmosphere produced by the photosynthetic organisms as primary electron acceptor. The free radical theory of aging argues that aging arises from the incurred damage due to the activity of reducing  $O_2$  to water, as the radical side products of the respiratory chain leads to a consistent mutagenic pressure that needs to be countered by DNA repair and antioxidant defenses (Speakman *et al.*, 2002; Harman, 2003).

Different higher organisms have evolved different trade-offs between life history variables relating mainly to the generation time (Kirkwood and Rose, 1991; Shanley and Kirkwood, 2000; Kirkwood, 2011). Shorter lifespan implies specialization toward shorter generation time, which again implies less energy invested in maintenance of the proteome. Based on the discussion above, this specialization emphasizes time over energy.

Each strategy probably involves an aging program to ‘dispose the soma’ after reproduction to make space for the next generation, although this remains debated (Westendorp and Kirkwood, 1998; Speakman *et al.*, 2002). Aging may thus be a direct consequence of the reproductive strategy. Some organisms specializing toward long lifespan (i.e. *r-* versus *k*-strategists) also diversify toward complex lifestyles with capacity for technology transfer, e.g. cetaceans and apes. Compared to primates, rodents on average have shorter generation times, lifespans, larger litter size, and have traded lifespan for fecundity (Speakman *et al.*, 2002; Wensink *et al.*, 2012). In long-living organisms, proteome misfolding may cause death, perhaps because PCM can no longer be afforded beyond what was evolutionarily beneficial. It is reasonable to argue that the aging program of long-living mammals largely reflect the (active or passive) giving up of the maintenance of the proteostatic machinery to enable the rise of the next generation (Taylor and Dillin, 2011; Hipkiss, 2017).

This discussion is well illustrated by superoxide dismutase 1 (SOD1). SOD1 is one of the most abundant proteins in primates and  $A_i$  can reach 100 000 copies per cell (Dasmeh and Kepp, 2017), it is the central antioxidant defense protein of the mitochondria thus directly linking energy and aging (Perry *et al.*, 2010), it is one of the few proteins known to directly extend lifespan upon induction (Tolmasoff *et al.*, 1980; Landis and Tower, 2005), and one of the few genes of great apes known to have undergone non-synonymous positive selection (Fukuhara *et al.*, 2002; Dasmeh and Kepp, 2017). Deposits of misfolded SOD1 is a hallmark of age-triggered amyotrophic lateral sclerosis (Valentine *et al.*, 2005). The tendency toward aggregation and misfolding of natural human SOD1 variants correlates with their pathogenicity (Lindberg *et al.*, 2005; Wang *et al.*, 2008; Kepp 2015), and wild-type overexpression by itself is enough to trigger disease (Wang *et al.*, 2009). Recent amino acid substitutions in SOD1 of great apes correlate with longer life span and tend to increase the net charge and stability of SOD1, thus increasing the thermodynamic and kinetic stability of the protein ( $k_d$  and  $\Delta G_i$ ) (Dasmeh and Kepp, 2017). Via its abundance and functional importance, any impairment of SOD1 either in terms of function or stability will produce comparatively very large PCM costs. The combination of the features summarized above strongly argues for a relationship between PCM, evolution of aging, and age-triggered neurodegenerative diseases.

According to the PCM theory, neurodegenerative diseases are caused by the increased energy spent on maintaining the proteome of old humans, which leaves less energy available for neuron and motor neuron function. Protein turnover and neuron signaling costs perhaps 20–25% and 50% of the brains energy budget (Hawkins, 1991; Attwell and Laughlin, 2001; Raichle and Gusnard, 2002), respectively, and as age advances, the supply of energy may no longer satisfy the increasing maintenance costs of the proteome (Kepp, 2019). Familial inherited mutations that tend to produce more aggregation-prone protein will increase turnover costs per time units according to PCM theory and will accordingly also accelerate the time at which available energy no longer satisfies the needs of synaptic transmission, leading to earlier clinical age of onset of disease (Kepp, 2019).

## Conclusions

Darwin’s theory of evolution emphasized ‘survival of the fittest’, where the ‘fit’ represented optimal functional proficiency. This

concept has dominated the thinking of the field, including the biochemical view of enzymes as optimally proficient for their catalytic reaction (Radzicka and Wolfenden, 1995; Zhang and Houk, 2005). Proteomic data have shown that most effects on the speed of evolution act via non-functional, universal selection pressures (Pál *et al.*, 2001, 2006; Drummond *et al.*, 2006). The main outstanding challenge in evolution is arguably to provide a predictive quantitative theory that captures these universal selection pressures and predicts real evolutionary histories, including the relative magnitude of drift and selection in specific cases, the nature of the selection pressures, and how it acts upon a population via the individual, the cell, the protein, and the gene.

This paper has reviewed the theory that a universal selection pressure is minimization of the ATP cost of an organism’s proteome (‘survival of the cheapest’). The magnitude and variations of the fundamental parameters show that most of the proteome cost selection acts via the ratio  $A_i/t_{1/2}$ , i.e. the abundance to half-life ratio of the protein. This selection combines with the selection for functional proficiency, typically in a cost–function trade-off between being ‘fit’ and ‘cheap’. The data in Table 2 suggest that cost selection occurred both during the earliest period of prokaryote evolution, during the rise of eukaryotes, particularly explaining the immediate advantages of the larger eukaryote cells due to reduced mass-specific metabolic costs, and during the long periods of relatively uneventful nearly neutral evolution that maintains nearly constant molecular clocks of many phylogenies.

The theory has several implications e.g. for stability–function and time–energy trade-offs, thermophile evolution, and human neurodegenerative diseases. One implication of the theory is that nature has not generally evolved the most proficient enzymes, in terms of turnover numbers ( $k_{cat}/K_M$ ), but the *lowest cost of substrate turnover*, as given by the ratio of Eq. (21). The theory thus predicts that most proteins may be engineered to obtain higher functional proficiency but that this will typically come with an associated increased total cost of the protein pool (e.g. via lower stability), which may however be less of an issue in the laboratory. The breakdown of this cost–function trade-off may be a central reason why directed evolution and protein-engineering strategies that aim to enhance protein performance even for natural functions are successful at all.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Conflict of interest.** The author declares that he has no conflict of interest associated with this study.

## References

- Adamczyk, A. J., Cao, J., Kamerlin, S. C. L. and Warshel, A. (2011). Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 14115–14120.
- Akashi H (2003) Translational selection and yeast proteome evolution. *Genetics* **164**, 1291–1303.
- Akashi H and Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3695–3700.
- Andersson SG and Kurland CG (1990) Codon preferences in free-living microorganisms. *Microbiology and Molecular Biology Reviews* **54**, 198–210.

- Attwell D and Laughlin SB** (2001) An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism* **21**, 1133–1145.
- Badaloo A, Jackson AA and Jahoor F** (1989) Whole body protein turnover and resting metabolic rate in homozygous sickle cell disease. *Clinical Science* **77**, 93–97.
- Bajaj M and Blundell T** (1984) Evolution and the tertiary structure of proteins. *Annual Review of Biophysics and Bioengineering* **13**, 453–492.
- Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS and Milo R** (2011) The moderately efficient enzyme: evolutionary and physico-chemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410.
- Barton MD, Delneri D, Oliver SG, Rattray M and Bergman CM** (2010) Evolutionary systems biology of amino acid biosynthetic cost in yeast. *PLoS ONE* **5**, e11935.
- Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR and Hwa T** (2015) Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature* **528**, 99.
- Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szyborska A, Herzog F, Rinner O, Ellenberg J and Aebersold R** (2011) The quantitative proteome of a human cell line. *Molecular Systems Biology* **7**, 549.
- Belle A, Tanay A, Bitincka L, Shamir R and O'Shea EK** (2006) Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 13004–13009.
- Benaroudj N, Zwickl P, Seemüller E, Baumeister W and Goldberg AL** (2003) ATP hydrolysis by the proteasome regulatory complex PAN serves multiple functions in protein degradation. *Molecular Cell* **11**, 69–78.
- Blankenship RE** (1992) Origin and early evolution of photosynthesis. *Photosynthesis Research* **33**, 91–111.
- Bloom JD and Adami C** (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evolutionary Biology* **3**, 21.
- Bloom JD and Adami C** (2004) Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. *BMC Evolutionary Biology* **4**, 14.
- Bloom JD, Wilke CO, Arnold FH and Adami C** (2004) Stability and the evolvability of function in a model protein. *Biophysical Journal* **86**, 2758–2764.
- Bloom JD, Drummond DA, Arnold FH and Wilke CO** (2006a) Structural determinants of the rate of protein evolution in yeast. *Molecular Biology and Evolution* **23**, 1751–1761.
- Bloom JD, Labthavikul ST, Otey CR and Arnold FH** (2006b) Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5869–5874.
- Blundell TL and Wood SP** (1975) Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature* **257**, 197.
- Boltzmann L** (1886) Der zweite Hauptsatz der mechanischen Warmetheorie. In der feierlichen Sitzung der Kaiserlichen Akademie der Wissenschaften, 29. Mai 1886. Vienna: Gerold, p. 21.
- Bonet J, Wehrle S and Schriever K, Yang C, Billet A, Sesterhenn F, Scheck A, Sverrisson F, Veselkova B, Vollers S, Lourman R, Villard M, Rosset S, Krey T and Correia BE** (2018) Rosetta FunFolDes – a general framework for the computational design of functional proteins. *PLoS Computational Biology* **14**, e1006623.
- Bouzat JL** (2010) Conservation genetics of population bottlenecks: the role of chance, selection, and history. *Conservation Genetics* **11**, 463–478.
- Brown GC** (1991) Total cell protein concentration as an evolutionary constraint on the metabolic control distribution in cells. *Journal of Theoretical Biology* **153**, 195–203.
- Brown JH, Marquet PA and Taper ML** (1993) Evolution of body size: consequences of an energetic definition of fitness. *The American Naturalist* **142**, 573–584.
- Cai L and Tu BP** (2012) Driving the cell cycle through metabolism. *Annual Review of Cell and Developmental Biology* **28**, 59–87.
- Cannon WR, Singleton SF and Benkovic SJ** (1996) A perspective on biological catalysis. *Nature Structural Biology* **3**, 821.
- Capra JA and Singh M** (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics (Oxford, England)* **23**, 1875–1882.
- Casari G, Sander C and Valencia A** (1995) A method to predict functional residues in proteins. *Nature Structural Biology* **2**, 171.
- Craig CL and Weber RS** (1998) Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by *Escherichia coli*. *Molecular Biology and Evolution* **15**, 774–776.
- Dasmeh P and Kepp KP** (2017) Superoxide dismutase 1 is positively selected to minimize protein aggregation in great apes. *Cellular and Molecular Life Sciences* **74**, 3023–3037.
- Dasmeh P, Serohijos AWR, Kepp KP and Shakhnovich EI** (2014) The influence of selection for protein stability on dN/dS estimations. *Genome Biology and Evolution* **6**, 2956–2967.
- DePristo MA, Weinreich DM and Hartl DL** (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics* **6**, 678.
- De Visser R, Spitters CJT and Bouma TJ** (1992) Energy cost of protein turnover: theoretical calculation and experimental estimation from regression of respiration on protein concentration of full-grown leaves. *Molecular, Biochemical and Physiological Aspects of Plant Respiration*. The Hague: Academic Publishing, pp. 493–508.
- Drummond DA and Wilke CO** (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352.
- Drummond DA and Wilke CO** (2009) The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics* **10**, 715–724.
- Drummond DA, Bloom JD, Adami C, Wilke CO and Arnold FH** (2005) Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14338–14343.
- Drummond DA, Raval A and Wilke CO** (2006) A single determinant dominates the rate of protein evolution. *Molecular Biology and Evolution* **23**, 327–337.
- Du B, Zielinski DC, Monk JM and Palsson BO** (2018) Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 11339–11344.
- Echave J, Spielman SJ and Wilke CO** (2016) Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics* **17**, 109–121.
- Ehrenberg M and Kurland CG** (1984) Costs of accuracy determined by a maximal growth rate constraint. *Quarterly Reviews of Biophysics* **17**, 45–82.
- English BP, Min W, van Oijen AM, Lee KT, Luo G, Sun H, Cherayil BJ, Kou SC and Sunney X** (2005) Ever-fluctuating single enzyme molecules: Michaelis–Menten equation revisited. *Nature Chemical Biology* **2**, 87–94.
- Fay JC, Wyckoff GJ and Wu C-I** (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024.
- Flanagan SE, Patch AM and Ellard S** (2010) Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers* **14**, 533–537.
- Francis JC and Hansche P** (1972) Directed evolution of metabolic pathways in microbial populations. I. Modification of the acid phosphatase pH optimum in *S. cerevisiae*. *Genetics* **70**, 59–73.
- Fraser KPP and Rogers AD** (2007) Protein metabolism in marine animals: the underlying mechanism of growth. *Advances in Marine Biology* **52**, 267–362.
- Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C and Feldman MW** (2002) Evolutionary rate in the protein interaction network. *Science (New York, N.Y.)* **296**, 750–752.
- Fu H, Grimsley G, Scholtz JM and Pace CN** (2010) Increasing protein stability: importance of  $\Delta C_p$  and the denatured state. *Protein Science* **19**, 1044–1052.
- Fukuhara R, Tezuka T and Kageyama T** (2002) Structure, molecular evolution, and gene expression of primate superoxide dismutases. *Gene* **296**, 99–109.
- Fuller J, Wilson TR, Eberhart ME and Alexandrova AN** (2019) Charge density in enzyme active site as a descriptor of electrostatic preorganization. *Journal of Chemical Information and Modeling* **59**, 2367–2373.
- Futcher B, Latter GI, Monardo P, McLaughlin CS and Garrels JI** (1999) A sampling of the yeast proteome. *Molecular and Cellular Biology* **19**, 7357–7368.



- Garat B and Musto H** (2000) Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia*. *Biochemical and Biophysical Research Communications* **279**, 996–1000.
- Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL and Drummond DA** (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 680–685.
- Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK and Weissman JS** (2003) Global analysis of protein expression in yeast. *Nature* **425**, 737.
- Gibbs DJ, Bacardit J, Bachmair A and Holdsworth MJ** (2014) The eukaryotic N-end rule pathway: conserved mechanisms and diverse functions. *Trends in Cell Biology* **24**, 603–611.
- Gillespie JH** (1984) The molecular clock may be an episodic clock. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 8009–8013.
- Gillespie JH** (1986) Rates of molecular evolution. *Annual Review of Ecology and Systematics* **17**, 637–665.
- Gillespie JH** (2001) Is the population size of a species relevant to its evolution? *Evolution* **55**, 2161–2169.
- Giovannoni SJ, Thrash JC and Temperton B** (2014) Implications of streamlining theory for microbial ecology. *The ISME Journal* **8**, 1553.
- Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E and Ben-Tal N** (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics (Oxford, England)* **19**, 163–164.
- Goldman N, Thorne JL and Jones DT** (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458.
- Goldstein RA** (2008) The structure of protein evolution and the evolution of protein structure. *Current Opinion in Structural Biology* **18**, 170–177.
- Goldstein RA** (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* **79**, 1396–1407.
- Gray MW, Burger G and Lang BF** (1999) Mitochondrial evolution. *Science (New York, N.Y.)* **283**, 1476–1481.
- Gsponer J, Futschik ME, Teichmann SA and Babu MM** (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science (New York, N.Y.)* **322**, 1365–1368.
- Gygi SP, Rochon Y, Franzosa BR and Aebersold R** (1999) Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology* **19**, 1720–1730.
- Hahn MW and Kern AD** (2004) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution* **22**, 803–806.
- Hall BG** (1981) Changes in the substrate specificities of an enzyme during directed evolution of new functions. *Biochemistry* **20**, 4042–4049.
- Hargrove JL and Schmidt FH** (1989) The role of mRNA and protein stability in gene expression. *The FASEB Journal* **3**, 2360–2370.
- Harish A and Kurland CG** (2017) Empirical genome evolution models root the tree of life. *Biochimie* **138**, 137–155.
- Harman D** (2003) The free radical theory of aging. *Antioxidants and Redox Signaling* **5**, 557–561.
- Harold FM** (1987) *The Vital Force: A Study of Bioenergetics*. New York: W.H. Freeman & Company.
- Hawkins AJS** (1991) Protein turnover: a functional appraisal. *Functional Ecology* **5**, 222–233.
- Hawkins AJS, Bayne BL, Day AJ and Denton EJ** (1986) Protein turnover, physiological energetics and heterozygosity in the blue mussel, *Mytilus edulis*: the basis of variable age-specific growth. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **229**, 161–176.
- Hawks J, Hunley K, Lee S-H and Wolpoff M** (2000) Population bottlenecks and Pleistocene human evolution. *Molecular Biology and Evolution* **17**, 2–22.
- Heizer EM, Raiford DW, Raymer ML, Doom TE, Miller RV and Krane DE** (2006) Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Molecular Biology and Evolution* **23**, 1670–1680.
- Heizer EM, Raymer ML and Krane DE** (2011) Amino acid biosynthetic cost and protein conservation. *Journal of Molecular Evolution* **72**, 466–473.
- Hipkiss AR** (2017) On the relationship between energy metabolism, proteostasis, aging and Parkinson's disease: possible causative role of methylglyoxal and alleviative potential of carnosine. *Aging and Disease* **8**, 334–345.
- Hurst LD** (2009) Genetics and the understanding of selection. *Nature Reviews. Genetics* **10**, 83–93.
- Hurst LD and Smith NGC** (1999) Do essential genes evolve slowly? *Current Biology* **9**, 747–750.
- Ikemura T** (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**, 13–34.
- Jansen R and Gerstein M** (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Research* **28**, 1481–1488.
- Jordan IK, Mariño-Ramírez L, Wolf YI and Koonin EV** (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Molecular Biology and Evolution* **21**, 2058–2070.
- Julenius K and Pedersen AG** (2006) Protein evolution is faster outside the cell. *Molecular Biology and Evolution* **23**, 2039–2048.
- Kahali B, Basak S and Ghosh TC** (2007) Reinvestigating the codon and amino acid usage of *S. cerevisiae* genome: a new insight from protein secondary structure analysis. *Biochemical and Biophysical Research Communications* **354**, 693–699.
- Kanaya S, Yamada Y, Kudo Y and Ikemura T** (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 143–155.
- Kepp KP** (2015) Genotype-property patient-phenotype relations suggest that proteome exhaustion can cause amyotrophic lateral sclerosis. *PLoS ONE* **10**, e0118649.
- Kepp KP** (2019) A quantitative model of human neurodegenerative diseases involving protein aggregation. *Neurobiology of Aging* **80**, 46–55.
- Kepp KP and Dasmeh P** (2014) A model of proteostatic energy cost and its use in analysis of proteome trends and sequence evolution. *PLoS ONE* **9**, e90504.
- Kimura M** (1962) On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713.
- Kimura M** (1991) The neutral theory of molecular evolution: a review of recent evidence. *The Japanese Journal of Genetics* **66**, 367–386.
- Kirkwood TBL** (2011) Systems biology of ageing and longevity. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **366**, 64–70.
- Kirkwood TB and Rose MR** (1991) Evolution of senescence: late survival sacrificed for reproduction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **332**, 15–24.
- Kleiber M** (1932) Body size and metabolism. *Hilgardia* **6**, 315–353.
- Kleiber M** (1947) Body size and metabolic rate. *Physiological Reviews* **27**, 511–541.
- Koonin EV, Wolf YI and Karev GP** (2002) The structure of the protein universe and genome evolution. *Nature* **420**, 218–223.
- Kumar S and Subramanian S** (2002) Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 803–808.
- Kurland CG and Ehrenberg M** (1984) Optimization of translation accuracy. In Cohn WE and Moldave K (eds) *Progress in Nucleic Acid Research and Molecular Biology*, vol. **31**. New York: Elsevier, pp. 191–219.
- Kurland CG and Ehrenberg M** (1987) Growth-optimizing accuracy of gene expression. *Annual Review of Biophysics and Biophysical Chemistry* **16**, 291–317.
- Kurland CG and Harish A** (2015) The phylogenomics of protein structures: the backstory. *Biochimie* **119**, 284–302.
- Landis GN and Tower J** (2005) Superoxide dismutase evolution and life span regulation. *Mechanisms of Ageing and Development* **126**, 365–379.
- Lane N** (2011) Energetics and genetics across the prokaryote-eukaryote divide. *Biology Direct* **6**, 35.
- Lane N and Martin W** (2010) The energetics of genome complexity. *Nature* **467**, 929.
- Li WF, Zhou XX and Lu P** (2005) Structural features of thermozymes. *Biotechnology Advances* **23**, 271–281.
- Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, De Koning AJ, Dokholyan NV, Echave J and Elofsson A**



- (2012) The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science* **21**, 769–785.
- Lindberg MJ, Byström R, Boknäs N, Andersen PM and Oliveberg M** (2005) Systematically perturbed folding patterns of amyotrophic lateral sclerosis (ALS)-associated SOD1 mutants. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9754–9759.
- Lobkovsky AE, Wolf YI and Koonin EV** (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2983–2988.
- Lobry JR and Gautier C** (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research* **22**, 3174–3180.
- Loftus B, Anderson I, Davies R, Davies R, Alsmark UCM, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, Nozaki T, Suh B, Pop M, Duchene M and Hall N** (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865.
- Lotka AJ** (1922) Contribution to the energetics of evolution. *Proceedings of the National Academy of Sciences of the United States of America* **8**, 147–151.
- Luke KA, Higgins CL and Wittung-Stafshede P** (2007) Thermodynamic stability and folding of proteins from hyperthermophilic organisms. *FEBS Journal* **274**, 4023–4033.
- Lynch M** (2006) Streamlining and simplification of microbial genome architecture. *Annual Review of Microbiology* **60**, 327–349.
- Lynch M and Marinov GK** (2015) The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 15690–15695.
- Marais G and Duret L** (2001) Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *Journal of Molecular Evolution* **52**, 275–280.
- Margoliash E** (1963) Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences of the United States of America* **50**, 672.
- Margulis L** (1968) Evolutionary criteria in thallophytes: a radical alternative. *Science (New York, N.Y.)* **161**, 1020–1022.
- Margulis L** (1975) Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symposia of the Society for Experimental Biology*, pp. 21–38.
- Martin AP and Palumbi SR** (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 4087–4091.
- Mayr E** (1998) Two empires or three? *Proceedings of the National Academy of Sciences of the United States of America* **95**, 9720–9723.
- McGuinness ET** (2010) Some molecular moments of the Hadean and Archaean Aeons: a retrospective overview from the interfacing years of the second to third millennia. *Chemical Reviews* **110**, 5191–5215.
- McInerney JO** (2006) The causes of protein evolutionary rate variation. *Trends in Ecology & Evolution* **21**, 230–232.
- Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fischer CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T and Rabosky DL** (2011) Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science (New York, N.Y.)* **334**, 521–524.
- Milo R** (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **35**, 1050–1055.
- Mirny LA and Shakhnovich EI** (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of Molecular Biology* **291**, 177–196.
- Mogk A, Schmidt R and Bukau B** (2007) The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies. *Trends in Cell Biology* **17**, 165–172.
- Morgenstern A, Jaszai M, Eberhart ME and Alexandrova AN** (2017) Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chemical Science* **8**, 5010–5018.
- Mozo-Villiarías A and Querol E** (2006) Theoretical analysis and computational predictions of protein thermostability. *Current Bioinformatics* **1**, 25–32.
- Ng PC and Henikoff S** (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812–3814.
- Nisbet EG and Sleep NH** (2001) The habitat and nature of early life. *Nature* **409**, 1083–1091.
- Northrop DB** (1998) On the meaning of  $K_m$  and  $V/K$  in enzyme kinetics. *Journal of Chemical Education* **75**, 1153–1157.
- Odum HT** (1988) Self-organization, transformity, and information. *Science (New York, N.Y.)* **242**, 1132–1139.
- Ohta T** (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**, 263–286.
- Overington J, Donnelly D, Johnson MS, Šali A and Blundell TL** (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Science* **1**, 216–226.
- Pál C, Papp B and Hurst LD** (2001) Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931.
- Pál C, Papp B and Lercher MJ** (2006) An integrated view of protein evolution. *Nature Reviews. Genetics* **7**, 337.
- Perry J, Shin D, Getzoff E and Tainer J** (2010) The structural biochemistry of the superoxide dismutases. *Biochimica et Biophysica Acta* **1804**, 245–262.
- Piganeau G and Eyre-Walker A** (2009) Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS ONE* **4**, e4396.
- Plata G, Gottesman ME and Vitkup D** (2010) The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biology* **11**, R98.
- Qian J, Luscombe NM and Gerstein M** (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology* **313**, 673–681.
- Radzicka A and Wolfenden R** (1995) A proficient enzyme. *Science (New York, N.Y.)* **267**, 90–93.
- Raichle ME and Gusnard DA** (2002) Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10237–10239.
- Raiford DW, Heizer EM, Miller RV, Akashi H, Raymer ML and Krane DE** (2008) Do amino acid biosynthetic costs constrain protein evolution in *Saccharomyces cerevisiae*? *Journal of Molecular Evolution* **67**, 621–630.
- Ramsey DC, Scherrer MP, Zhou T and Wilke CO** (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* **188**, 479–488.
- Reeds PJ, Fuller MF, Nicholson BA** (1985) Metabolic basis of energy expenditure with particular reference to protein. In Garrow JS and Halliday D (eds), *Substrate and Energy Metabolism in man*, London: Libbey, pp. 46–57.
- Richmond RC** (1970) Non-Darwinian evolution: a critique. *Nature* **225**, 1025–1028.
- Robertson AD and Murphy KP** (1997) Protein structure and the energetics of protein stability. *Chemical Reviews* **97**, 1251–1268.
- Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican M, Eaton G and Humphreys G** (1984) Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Research* **12**, 6663–6671.
- Rocha EPC and Danchin A** (2004) An analysis of determinants of amino acid substitution rates in bacterial proteins. *Molecular Biology and Evolution* **21**, 108–116.
- Savage VM, Allen AP, Brown JH, Gillooly JF, Herman AB, Woodruff WH and West GB** (2007) Scaling of number, size, and metabolic rate of cells with body size in mammals. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 4718–4723.
- Sawle L and Ghosh K** (2011) How do thermophilic proteins and proteomes withstand high temperature? *Biophysical Journal* **101**, 217–227.
- Schaber J, Rispe C, Wernegreen J, Bunes A, Delmotte F, Silva FJ and Moya A** (2005) Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. *Gene* **352**, 109–117.
- Schrödinger E** (1944) *What is life? The physical aspect of the living cell and mind*. Cambridge: Cambridge University Press.
- Seligmann H** (2003) Cost-minimization of amino acid usage. *Journal of Molecular Evolution* **56**, 151–161.
- Serohijos AWR, Rimas Z and Shakhnovich EI** (2012) Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Reports* **2**, 249–256.
- Shanley DP and Kirkwood TB** (2000) Calorie restriction and aging: a life-history analysis. *Evolution* **54**, 740–750.

- Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *Journal of Molecular Evolution* **33**, 23–33.
- Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN and Gaunt TR (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation* **34**, 57–65.
- Sikosek T and Chan HS (2014) Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface* **11**, 20140419.
- Smith DR and Chapman MR (2010) Economical evolution: microbes reduce the synthetic cost of extracellular proteins. *MBio* **1**, e00131–10.
- Soskine M and Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* **11**, 572.
- Sousa FL, Thiergart T, Landan G, Nelson-Sathi S, Pereira IA, Allen JF, Lane N and Martin WF (2013) Early bioenergetic evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20130088.
- Speakman JR, Selman C, McLaren JS and Harper EJ (2002) Living fast, dying when? The link between aging and energetics. *The Journal of Nutrition* **132**(6 Suppl. 2), 1583S–1597S.
- Swire J (2007) Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *Journal of Molecular Evolution* **64**, 558–571.
- Tang N, Dehury B and Kepp KP (2019) Computing the pathogenicity of Alzheimer's disease presenilin 1 mutations. *Journal of Chemical Information and Modeling*, **59**, 858–870.
- Taverna DM and Goldstein RA (2002) Why are proteins marginally stable? *Proteins: Structure, Function, and Bioinformatics* **46**, 105–109.
- Taylor RC and Dillin A (2011) Aging as an event of proteostasis collapse. *Cold Spring Harbor Perspectives in Biology* **3**, a004440.
- Tekaia F, Yeramian E and Dujon B (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* **297**, 51–60.
- Thusberg J, Olatubosun A and Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation* **32**, 358–368.
- Tokuriki N and Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596–604.
- Tokuriki N, Stricher F, Schymkowitz J, Serrano L and Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology* **369**, 1318–1332.
- Tokuriki N, Stricher F, Serrano L and Tawfik DS (2008) How protein stability and new functions trade off. *PLoS Computational Biology* **4**, e1000002.
- Tolmasoff JM, Ono T and Cutler RG (1980) Superoxide dismutase: correlation with life-span and specific metabolic rate in primate species. *Proceedings of the National Academy of Sciences of the United States of America* **77**, 2777–2781.
- Tuller T, Waldman YY, Kupiec M and Ruppin E (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 3645–3650.
- Urrutia AO and Hurst LD (2003) The signature of selection mediated by expression on human genes. *Genome Research* **13**, 2260–2264.
- Uversky VN, Oldfield CJ and Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual Review of Biophysics* **37**, 215–246.
- Valentine J, Doucette P and Potter SZ (2005) Copper-zinc superoxide dismutase and amyotrophic lateral sclerosis. *Annual Review of Biochemistry* **74**, 563–593.
- Varshavsky A (1997) The N-end rule pathway of protein degradation. *Genes to Cells* **2**, 13–28.
- Venev SV and Zeldovich KB (2018) Thermophilic adaptation in prokaryotes is constrained by metabolic costs of proteostasis. *Molecular Biology and Evolution* **35**, 211–224.
- Wagner A (2005) Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution* **22**, 1365–1374.
- Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB and Feldman MW (2005) Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5483–5488.
- Wang Q, Johnson JL, Agar NYR and Agar JN (2008) Protein aggregation and protein instability govern familial amyotrophic lateral sclerosis patient survival. *PLoS Biology* **6**, e170.
- Wang L, Deng H-X, Grisotti G, Zhai H, Siddique T and Roos RP (2009) Wild-type SOD1 overexpression accelerates disease onset of a G85R SOD1 mouse. *Human Molecular Genetics* **18**, 1642–1651.
- Wang M, Kurland CG and Caetano-Anollés G (2011) Reductive evolution of proteomes and protein structures. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11954–11958.
- Warshel A (1998) Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *Journal of Biological Chemistry* **273**, 27035–27038.
- Waterlow JC (1995) Whole-body protein turnover in humans – past, present, and future. *Annual Review of Nutrition* **15**, 57–92.
- Wensink MJ, van Heemst D, Rozing MP and Westendorp RGJ (2012) The maintenance gap: a new theoretical perspective on the evolution of aging. *BioGerontology* **13**, 197–201.
- Westendorp RG and Kirkwood TB (1998) Human longevity at the cost of reproductive success. *Nature* **396**, 743–746.
- Willis JH and Orr HA (1993) Increased heritable variation following population bottlenecks: the role of dominance. *Evolution* **47**, 949–957.
- Winter EE, Goodstadt L and Ponting CP (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research* **14**, 54–61.
- Woese CR (1998) Default taxonomy: Ernst Mayr's view of the microbial world. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 11043–11046.
- Wolfenden R and Snider MJ (2001) The depth of chemical time and the power of enzymes as catalysts. *Accounts of Chemical Research* **34**, 938–945.
- Worth CL, Gong S and Blundell TL (2009) Structural and functional constraints in the evolution of protein families. *Nature Reviews Molecular Cell Biology* **10**, 709.
- Wylie CS and Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 9916–9921.
- Yang J-R, Liao B-Y, Zhuang S-M and Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E831–E840.
- Yi S, Ellsworth DL and Li W-H (2002) Slow molecular clocks in Old World monkeys, apes, and humans. *Molecular Biology and Evolution* **19**, 2191–2198.
- Zhang J (2000) Protein-length distributions for the three domains of life. *Trends in Genetics* **16**, 107–109.
- Zhang X and Houk KN (2005) Why enzymes are proficient catalysts: beyond the Pauling paradigm. *Accounts of Chemical Research* **38**, 379–385.
- Zhang L and Li W-H (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution* **21**, 236–239.
- Zhang H, Wang Y, Li J, Chen H, He X, Zhang H, Liang H and Lu J (2018) Biosynthetic energy cost for amino acids decreases in cancer evolution. *Nature Communications* **9**, 4124.
- Zuckermandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. *Journal of Molecular Evolution* **7**, 269–311.
- Zuckermandl E and Pauling LB (1962) Molecular disease, evolution, and genetic heterogeneity. In Kasha M and Pullman B (eds), *Horizons in Biochemistry*, New York, NY: Academic Press, pp. 189–225.
- Zuckermandl E and Pauling L (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology* **8**, 357–366.