

# Mutational effects and the evolution of new protein functions

Misha Soskine and Dan S. Tawfik

**Abstract** | The divergence of new genes and proteins occurs through mutations that modulate protein function. However, mutations are pleiotropic and can have different effects on organismal fitness depending on the environment, as well as opposite effects on protein function and dosage. We review the pleiotropic effects of mutations. We discuss how they affect the evolution of gene and protein function, and how these complex mutational effects dictate the likelihood and mechanism of gene duplication and divergence. We propose several factors that can affect the divergence of new protein functions, including mutational trade-offs and hidden, or apparently neutral, variation.

## Protein mutations

Missense mutations that occur in encoded open reading frames.

## Trade-offs

Gains of a new activity or property at the expense of other activities or properties.

## Protein stability

The capacity of a protein to adopt its native, functional structure. Stability also correlates with cellular protein levels.

## Sub-functionalization

Degenerate mutations that result in a gene and its duplicated copy sharing the burden of one function.

To understand the evolution of gene and protein function necessitates a mechanistic model that describes how one gene diverges to give two paralogous genes that encode two proteins with related sequences, structures and functions. Different mechanisms and models have been proposed for such divergence processes, and these mechanisms differ by assuming that different timings and different selection forces act on the starting gene versus its duplicated copy. However, the relevance and feasibility of these various models is still unclear. The influence of several key factors, including population genetics parameters, has been addressed<sup>1,2</sup>. However, the divergence of new gene and protein functions should also be considered in light of the effects of mutations, and specifically of protein mutations. This Review describes our current knowledge of the effects of mutations on the structural integrity and activity of proteins. It provides insights into the mechanisms by which new protein functions diverge from existing ones through gene duplication and through mutations that modulate protein function. As shown below, these elements are inseparable. This discussion is part of a broader, ongoing effort to integrate molecular evolution, population genetics and protein science (the study of protein structure, function and biophysics), and that aims to provide a comprehensive understanding of protein evolution<sup>3-6</sup>.

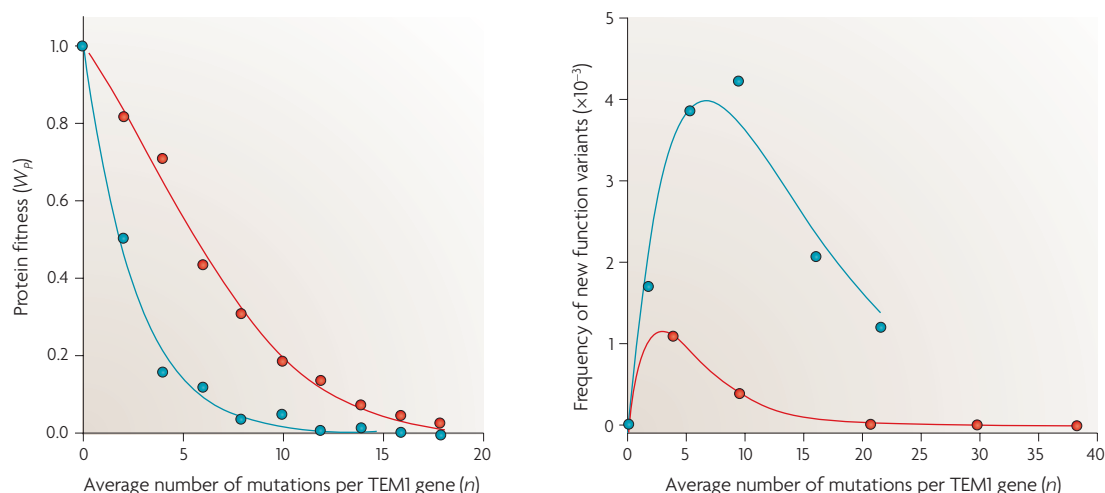
In this article, we adopt a protein perspective by considering the effects of protein mutations. Similar to the analysis of mutational fitness effects that has been undertaken at the organismal level<sup>7</sup>, we examine the frequency of mutations with varying fitness effects — such as neutral, deleterious or advantageous — on protein structure and function. Curiously, the comparison

of the fitness effects of mutations on proteins versus intact organisms shows some unexpected trends; for example, organisms could be more sensitive to mutations than their individual proteins. We describe the pleiotropic effects of protein mutations and the various trade-offs that arise from this pleiotropy — for example, a mutation that positively affects the activity of a protein may negatively affect protein stability and thereby reduce the level of soluble, functional protein. Similarly, a mutation that is beneficial for an alternative, future function can be deleterious or neutral for a protein's existing function. We also discuss the various mechanisms of buffering and compensating for mutational effects (and thus alleviating their trade-offs) in addition to the related notions of hidden, or neutral, variation and neutral networks. We subsequently argue that mutational effects, their trade-offs and the corresponding buffering mechanisms influence not only whether a given protein function can evolve but also the mechanism by which this process is likely to occur. To address this issue, we show how the likelihood of occurrence of three representative models of divergence (Ohno's model, the 'divergence prior to duplication' (DPD) model and the 'sub-functionalization' model) is influenced by the effects of protein mutations and their trade-offs.

## The effects of protein mutations

The effects of mutations on protein structure, stability and function have been extensively examined<sup>8</sup>. However, few studies have provided systematic data that can be used in evolutionary analyses. Here, we primarily present data that are derived from experiments using TEM1  $\beta$ -lactamase as a model protein. TEM1 confers

Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel.  
Correspondence to D.S.T.  
e-mail: [tawfik@weizmann.ac.il](mailto:tawfik@weizmann.ac.il)  
doi:10.1038/nrg2808



**Figure 1 | Rapid fitness declines result in increased likelihood of a new function emerging under selection for the existing one.** **a** | Protein fitness ( $W_p$ ) rapidly declines when random mutations accumulate under no selection.  $W_p$  corresponds to the fraction of TEM1 genes that are able to confer growth to *Escherichia coli* under a given concentration of the antibiotic ampicillin.  $W_p$  was measured for populations evolved by drift, each carrying a different average number of mutations per gene ( $n$ ). The measurement was performed under two conditions: wild-type-like levels of antibiotic resistance (2,500 mg/L ampicillin; in blue), and a 200-fold lower resistance level (12.5 mg/L ampicillin; in red). At wild-type levels, fitness decline is very rapid, and accumulation of >10 mutations per gene results in non-functionalization of  $\geq 99\%$  of the mutated genes. However, a substantially higher fraction of functional genes is maintained at a lower level of fitness (12.5 mg/L ampicillin). Data were fitted to exponential decays:  $W_p = e^{-\alpha n}$ , in which  $\alpha$  is the fraction of deleterious mutations<sup>10</sup>. At the low fitness level (12.5 mg/L ampicillin; in red), the decay is steeper than exponential owing to negative epistasis mediated by a margin of protein stability that buffers the effect of the first mutations<sup>6,10</sup>. **b** | The likelihood of acquiring a new function under no selection versus under selection for the existing function. Plotted are the measured frequencies of TEM1 variants exhibiting a new function in gene populations with increasing average numbers of mutations ( $n$ )<sup>11</sup>. Libraries that were drifted under no selection are shown in red (as in part a; corresponding to Ohno's model, FIG. 4) and TEM1 genes drifted under selection for the existing function are in blue (the 'divergence prior to duplication' model, FIG. 4). Selection on ampicillin served as the purifying selective regime for the original penicillinase function, and cefotaxime resistance modelled the acquisition of a new enzymatic specificity. The lines represent a fit to a model indicating that the frequency of new-function mutations is similar under both regimes ( $\sim 1.3 \times 10^{-3}$ ). However, the fraction of deleterious and non-functionalization mutations under purifying selection ( $\alpha = 0.14$ ) is much smaller than under no selection ( $\alpha = 0.36$ )<sup>11</sup>.

#### Negative epistasis

The combined effect of mutations being more deleterious than expected from their individual effects.

#### Protein fitness

Levels of physiological function exerted by a given protein variant under a certain selection pressure.

#### Non-functionalization

The complete inactivation of a gene or protein by highly deleterious mutations.

#### Neo-functionalization

The divergence of a duplicated gene or protein to execute a new function.

#### $\Delta\Delta G$

The stability difference for a protein variant versus its wild-type reference ( $\Delta\Delta G > 0$  indicates lower stability).

antibiotic resistance to gram-negative bacteria. Protein fitness ( $W_p$ ) therefore corresponds to the concentration of an antibiotic that can be tolerated by *Escherichia coli* cells that carry a given enzyme variant<sup>9–11</sup>. TEM1 was subjected to random mutagenesis *in vitro* and the levels of antibiotic resistance were measured for many variants. The average fitness of a population of TEM1 genes could thereby be determined as a function of the average number of mutations<sup>10</sup> (FIG. 1a).

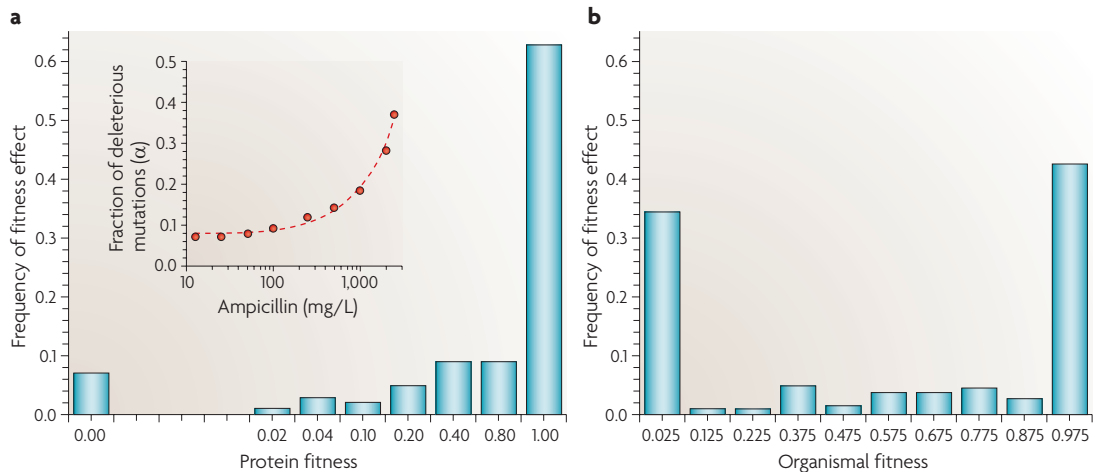
#### The distribution of fitness effects of protein mutations.

The TEM1 measurements can be used to derive a distribution of fitness effects, even though the measurements were not processed in this way in the original publications<sup>10,11</sup>. This distribution can be compared to the distribution of fitness effects of mutations in the organismal genomes (BOX 1). This distribution and other data<sup>8,9</sup> indicate that, in the absence of selection, non-functionalization is inevitable. As mutations accumulate, the likelihood of a gene or protein losing its function increases exponentially — or even more steeply than exponentially (FIG. 1a). A substantial subgroup of deleterious mutations ( $\sim 8\%$ ; BOX 1) lead to the loss of all functions in a

way that makes neo-functionalization impossible. These non-functionalization mutations arise primarily from a sizeable fraction of mutations that severely undermine protein stability ( $\Delta\Delta G \geq 3$  kcal per mol). The overall stability ( $\Delta G$ ; the free energy difference between the folded, native state of a protein and its unfolded state) of most proteins is in the range of 10 kcal per mol (for example, 7.3 kcal per mol for TEM1)<sup>12</sup>. Thus, even a single destabilizing mutation can cause a substantial reduction, or even complete loss, of protein levels owing to misfolding, aggregation or proteolytic clearance<sup>6,13</sup>. Although less frequent than destabilizing mutations, mutations that alter protein residues that are absolutely essential for function also lead to non-functionalization.

Approximately 30% of TEM1 mutations cause a partial reduction in fitness, primarily owing to mildly destabilizing mutations that reduce the levels of soluble, folded protein. The remaining fraction ( $\sim 62\%$ ) has no immediate measurable effects on fitness (BOX 1). Although this detailed distribution is available for only one protein, experiments with other proteins show similar trends; approximately 40% of mutations reduce or completely abolish the activity of the mutated protein<sup>8</sup>. However, the

## Box 1 | The distribution of fitness effects of protein mutations

**Protein fitness**

Fitness is an organismal feature that relates to population growth rates. What then does the term ‘protein fitness’ mean? Proteins are a key component of organismal fitness. Under some conditions, variation in only one protein contributes to fitness, and the effects of mutations in this protein can be directly connected to organismal fitness. In the presence of antibiotics, for example, the survival of a microorganism depends on the function of a single protein that mediates resistance, such as an enzyme that degrades the antibiotic. In such cases, the level of physiological function exerted by this protein can be easily measured and is denoted as protein fitness ( $W_p$ ).  $W_p$  would be proportional to the concentration of folded, active enzyme in a living cell (protein level or dosage) and to the activity per protein molecule (for example, the catalytic efficiency of the degrading enzyme,  $k_{cat}/K_M$ ). Such simple scenarios enable the description of the distribution of fitness effects for an individual protein and allow it to be compared to the distribution of fitness effects of mutations in an intact organism.

The figure shows the distribution of fitness effects of mutations for an individual protein (part **a**) and an organism (part **b**). In part **a**, the fraction of mutations that result in a loss of fitness was derived from measurements of fitness declines upon the accumulation of mutations, as exemplified in FIG. 1a. Below a certain level (<50 mg/L ampicillin),  $\alpha$  remains constant (inset in part **a**), indicating that ~8% of the mutations cause complete non-functionalization ( $W = 0$ ). The intermediate levels of fitness ( $0 < W < 1$ ) were derived from the  $\alpha$  values for 50–2,000 mg/L ampicillin<sup>10</sup>. The fraction of mutations with no effect on fitness ( $W = 1$ ) is 0.63 ( $1 - 0.37$ , in which 0.37 is the  $\alpha$  value obtained for the maximal, wild-type-like levels of fitness, 2,500 mg/L ampicillin). Part **b** shows the distribution of fitness effects for an intact organism. The effects were averaged for spontaneous and mutagen-induced mutations in yeast and re-binned similarly to the protein distribution. The fraction of non-functionalization mutations for yeast can only be estimated from the lowest fitness threshold ( $W = 0.025$ ), as complete fitness loss ( $W = 0$ ) in an organism cannot be measured.

**Is an organism a sum of its proteins?**

The graphs for a single protein (part **a**) and whole organism (part **b**) only provide a preliminary basis for discussion, as they differ in several key parameters, including the fitness thresholds. These caveats aside, the overall shapes of these distributions are strikingly similar. A notable difference seems to be that the fraction of non-functionalization mutations ( $W \leq 0.025$ ) is much higher for organisms than for an individual protein (~0.35 in organisms versus <0.1 for an individual protein). This difference suggests that complex, multi-component organisms could be more easily perturbed than their individual protein components. This is somewhat unexpected given the various redundancy and backup mechanisms that confer organisms with robustness to mutations. Indeed, mutations in metabolic enzymes show no loss of organismal fitness despite partial loss of enzyme function, and a large fraction of genes can be knocked out completely with no apparent fitness effect<sup>90</sup>.

The similarities and differences between the protein and organismal distributions therefore raise fundamental questions that need to be explored further. One possible explanation is that most proteins exhibit more deleterious distributions than that of TEM1 (BOX 2), although similar distributions were obtained for other enzymes<sup>8</sup>. Alternatively, partial loss of function of certain proteins may result in near-complete loss of fitness for the whole organism and, conversely, many mutations may have more deleterious effects than the complete removal of the gene (as in dominant-negative mutants).

The distributions in panel **b** were extracted from REFS 7,94.

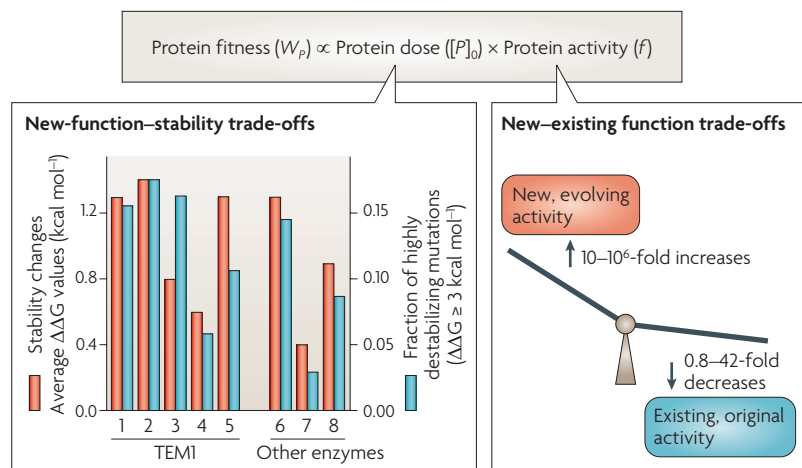
**Disordered domains**

Protein domains with a high degree of random coil and loop regions and a low degree of highly ordered secondary structure.

available data are largely limited to one class of proteins — single-domain, soluble enzymes. Other classes of proteins, such as membrane proteins, remain unexplored and their distributions may differ substantially. Similarly, disordered domains may not behave in the same way as completely folded enzymes and viral proteins that show partial disorder may also show a different fitness distribution<sup>14</sup>.

**Mutational trade-offs**

The above scenario is simplistic in that mutations are portrayed as showing uni-dimensional effects — a mutation is deleterious, neutral or advantageous. In reality, the effects of mutations are context-dependent or pleiotropic. Fitness effects vary depending on the evolutionary context or growth environment — a mutation



**Figure 2 | Protein fitness and mutational trade-offs.** Protein fitness ( $W_p$ ) is proportional to the protein's specific activity (activity per protein molecule,  $f$ ) and to the level of folded and functional protein (protein dose,  $[P]_0$ ), which in turn relates to protein stability. The fitness of an evolving protein increases via the accumulation of mutations that increase the new function. However, these mutations may also affect the protein's existing function and its stability. New-existing function trade-offs are defined by the fold-increase in the new function (in terms of affinity or catalytic efficiency) induced by the mutation versus the decrease in the existing function. Most mutations selected *in vitro* for improvements in an evolving promiscuous function show  $\geq 10$  times higher increases in the new function relative to the decrease in the existing one, although the existing function was not maintained under selection<sup>21</sup>. Strong trade-offs are also seen, in which the decrease in the primary function is larger than the gain in the evolving new function<sup>19-21</sup>. New-function-stability trade-offs refer to mutations that improve a new, evolving function but also reduce protein stability and may thereby reduce protein dose<sup>15,16</sup>. The red bars indicate the destabilizing effect of each group of mutations (average  $\Delta\Delta G$  values, computationally predicted). Blue bars indicate the fraction of highly destabilizing mutations ( $\Delta\Delta G \geq 3$  kcal mol<sup>-1</sup>) within these groups. The left set of bars (bars 1-5) correspond to mutations in TEM1  $\beta$ -lactamase<sup>10,15,51</sup>, and the right set of bars (bars 6-8) correspond to mutations identified in laboratory evolution experiments in a large set of enzymes<sup>16</sup>. The destabilizing effects of mutations found in genes drifted under no selection, both in TEM1 (bar 1) and in other enzymes (bar 6), resemble those predicted for all possible mutations in TEM1 (bar 2). Selection purges destabilizing mutations, and strongly destabilizing mutations in particular, as seen in the analysis of mutations in TEM1 genes that drifted under purifying selection (bars 3 and 4, which represent low- and high-selection stringencies, respectively) and in other enzymes (bar 7). Mutations that confer new functions in TEM1 (bar 5) and in other enzymes (bar 8) show significantly higher destabilizing effects than those that accumulated under purifying selection to maintain the existing function.

can be neutral or deleterious in one environment but can become beneficial if the context changes. Similarly, a mutation may improve one protein activity at the expense of another. This pleiotropy gives rise to mutational trade-offs that make the fitness landscape multi-dimensional. We describe here two main types of trade-off that are well characterized and allude to others that have yet to be examined.

**New-function-stability trade-offs.** The configurational stability of a protein dictates the levels of soluble, functional protein<sup>6</sup>. Most mutations are destabilizing (FIG. 2), and once protein stability has dropped below a certain threshold the levels of soluble and functional protein decrease, thereby resulting in reduced protein and organismal fitness<sup>3,6,8,13</sup>. Purifying selection therefore

purges mutations with strong or even mildly destabilizing effects, depending on the threshold of selection, and the remaining neutral or apparently neutral mutations exhibit low destabilizing effects. However, on average, mutations that confer new functions (new-function mutations) exhibit stronger destabilizing effects than these neutral mutations. This trade-off was first described for TEM1<sup>15</sup> and was later confirmed for a large number of laboratory-evolved enzymes<sup>16</sup>, as well as for protein-protein interactions<sup>17</sup>. Owing to their destabilizing effect, new-function mutations often result in lower protein levels, and such mutations are less likely to cause the divergence of new function. Destabilizing mutations may also lead to misfolding and the resulting aggregates may reduce fitness<sup>18</sup>.

**New-existing function trade-offs.** New-existing function trade-offs relate to the exquisite specificity of proteins that is reflected in the classic description of a 'lock-and-key' complementarity between the shape and charge of the ligand or substrate and the protein's active site. Implicit in this view is the notion that a mutation that changes the shape of an active site to accommodate a new ligand and is bound to disturb the interactions of the active site with the original ligand. Indeed, certain mutations exhibit strong new-existing function trade-offs, such that improvements in the evolving function are accompanied by large drops in the original function<sup>19,20</sup>. In general, strong trade-offs often relate to marked differences in the size, or charge, of the new versus the original ligand or substrate, and to the location of the mutation (these aspects are discussed elsewhere<sup>21</sup>). As discussed below, in such cases functional divergence depends on duplication or on compensatory mechanisms, such as the upregulation of protein expression.

Surprisingly, however, most mutations exhibit weak trade-offs with respect to latent and promiscuous protein functions<sup>22</sup>. This is highlighted by the effects of mutations that have been identified in many different cases of laboratory and natural evolution in a range of proteins<sup>21</sup>: many of these mutations show almost no trade-off and the vast majority result in a  $\geq 10$ -fold gain in new function versus loss of the existing one (FIG. 2). Weak new-existing function trade-offs also underlie the evolution of protein-protein interactions<sup>17</sup>. Weak trade-offs are connected with the conformational flexibility of proteins because alternative conformations that mediate new functions can gain higher representation without severely compromising the conformation that mediates the existing function<sup>21,23</sup>. If the trade-off is sufficiently weak, new-function mutations can accumulate — or can even reach fixation — under purifying selection as part of 'neutral drift'.

**Regulatory trade-offs.** A new gene function usually means not only a new protein activity, but also a new regulatory regime. The original and the evolving functions may contribute to organismal fitness under conflicting regulatory regimes — that is, they might be needed at conflicting times and locations and may therefore trade off.

#### Apparently neutral mutations

Mutations that have no significant or observable fitness effect under a given environment.

#### New-function mutations

Mutations that mediate changes in protein activity, typically by increasing a weak, latent promiscuous function.

### The advantage of purging deleterious mutations

The high frequency of deleterious mutations, and of non-functionalization mutations in particular, greatly decreases the likelihood of divergence. Indeed, a laboratory evolution experiment using TEM1  $\beta$ -lactamase that was aimed at testing this hypothesis indicated that when deleterious mutations are purged under selection, the emergence of variants that exhibit the new function becomes far more likely<sup>11</sup> (FIG. 1b). The much lower frequency of new-function variants and the narrower window for their emergence are the outcome of a larger fraction of deleterious and non-functionalization mutations that accumulate under no selection ( $\alpha = 0.36$ , in which  $\alpha$  represents the fraction of deleterious and non-functionalization mutations) in comparison to the population drifting under purifying selection ( $\alpha = 0.14$ ). However, divergence under selection for the existing function is feasible only when the new–existing function trade-offs are weak enough, and when the level of purifying selection that acts on the drifting gene is sufficiently low to enable new-function mutations to accumulate. In the TEM1 model, both of these conditions are easily met<sup>11</sup>.

The divergence of a new function under no selection versus the divergence of a new function under selection for the existing function are represented in different models — Ohno's model and the DPD model, respectively. These models are discussed in a later section.

### Buffering and compensatory mechanisms

The numbers and types of mutations that accumulate are influenced not only by the level of selection that acts on the drifting genes, as shown in FIG. 1a, but also by the availability of buffering and compensatory mechanisms, as discussed below.

**Gene duplication.** Duplication is not only a vehicle of functional divergence but also a means to buffer the deleterious effects of mutations. Duplication increases gene and protein doses and thereby reduces the level of purifying selection that acts on individual protein molecules. Under these reduced selection levels, a larger number of mildly deleterious mutations can accumulate (FIG. 1a; for example, a fitness level of  $W_p = 0.5$  would enable an average of  $\sim 2$  mutations per gene under strong selection pressure (2,500 mg/L ampicillin) versus  $\sim 5$  under weak selection (12.5 mg/L ampicillin)). These mildly deleterious mutations include destabilizing but potentially adaptive mutations. Indeed, a common outcome of duplication is an increase in the accumulation of activity-reducing mutations (ARMs)<sup>24</sup>. Most of these mutations decrease protein levels and are therefore readily compensated for by duplication.

Duplication may not only alleviate stability–new-function trade-offs and new–existing function trade-offs but may also buffer regulatory trade-offs. By virtue of the duplicate being placed in a different genomic location, it is also likely to be expressed under a new regulatory regime. Indeed, gene duplications may give rise to new functions merely by having the same protein expressed in a different context (for example, a retrogene duplication of fibroblast

growth factor 4 (*FGF4*) resulted in different regulatory control from the endogenous *FGF4* and thereby gave rise to the short-legged phenotype in certain dog breeds)<sup>25,26</sup>.

In bacteria<sup>27</sup> and eukaryotes<sup>28</sup>, gene duplication is a relatively fast process that provides an immediate advantage by increasing gene and protein doses. The presence of duplicated genes that encode the same, or similar, function is often due to selection for higher enzymatic fluxes<sup>29,30</sup>. Retrogenes, which originate from the reverse transcription of mRNAs<sup>31</sup>, comprise another mode of duplication<sup>25,26</sup> that is sometimes described as 'Lamarckian'<sup>32,33</sup>. In this mode, regulatory, physiological responses lead to higher protein doses through higher transcription levels, and higher mRNA levels increase the likelihood of retrogene formation. Thus, the higher protein doses are made heritable.

The compensatory potential of duplication is also apparent in its transient nature. As selection pressure is relaxed or as a protein's specific activity (activity per protein molecule) is increased through an adaptive mutation, duplicates are rapidly removed<sup>34</sup>. Therefore, we propose that duplication may act as an intermediate, bridging step. The absence of an overall correlation between gene dosage and gene duplicability<sup>35</sup> also supports a transient role for duplication.

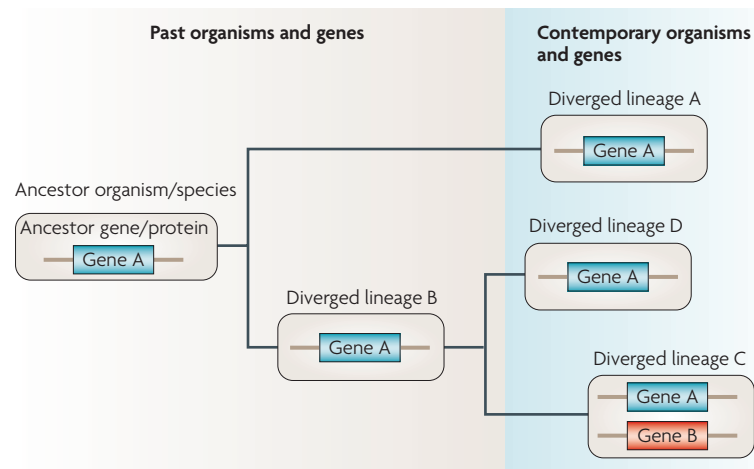
**Upregulation of protein expression.** Protein doses can also be increased by mutations in regulatory elements. For example, mutations in the promoter region and in the signal peptide that augment protein levels by up to 10-fold are commonly observed in TEM1 under selection for higher antibiotic resistance<sup>36,37</sup>. Similarly, in Hall's classical experiment, the evolution of a new  $\beta$ -galactosidase function in *E. coli* resulted in the recruitment of an enzyme (evolved  $\beta$ -galactosidase (*ebg*)) that exhibits weak, promiscuous  $\beta$ -galactosidase activity. The first mutation inactivated the suppressor of *ebg* and thereby dramatically increased the expression levels of *ebg*<sup>38,39</sup>. This is not a case of buffering trade-offs — *ebg* is a non-essential gene — but a demonstration of the buffering potential of upregulation.

A marked example of trade-offs that are alleviated by upregulation is that of glutamylphosphate reductase (*proA*). *ProA* reduces glutamylphosphate to yield glutamate 5-semialdehyde, an intermediate in proline biosynthesis. However, it also exhibits weak promiscuous activity that yields *N*-acetylglutamate 5-semialdehyde — an intermediate in arginine biosynthesis that is normally produced by *argC*. Selection of *argC* knockout cells for growth with no supplemented amino acids led to the fixation of a mutation in *proA*. However, the *proA* function is essential for proline synthesis, and the mutation exhibited a strong trade-off (12-fold increase in the evolving *argC* function versus 2,800-fold decrease in the original *proA* function). Upregulation of *proA* levels enabled growth despite this trade-off<sup>19</sup>.

However, elevated expression levels only provide temporary relief. Expression is costly<sup>40,41</sup>. Moreover, for a significant fraction of proteins, increased dosages result in reduced fitness owing to undesirable promiscuous interactions driven by high protein concentrations<sup>42</sup>

#### New–existing function trade-offs

The acquisition of a new function through mutations that undermine the existing function.



**Figure 3 | Divergence of protein-coding genes.** Protein-coding genes diverge along the lineage of the organisms in which they reside. The process is manifested in the structure, sequence and functional homology between genes and proteins found in various organisms and within the same organism. Genes annotated as Gene A represent orthologues — genes that diverged along the divergence of species. Orthologues usually differ in sequence but exhibit the same structure and function. Gene A and Gene B represent paralogues that resulted from gene duplication in one lineage (lineage C). Paralogues differ in sequence (typically to a larger degree than the Gene A orthologues) as well as in function.

or disturbed balance of protein complexes<sup>29,43</sup>. Thus, although increased protein doses can make a weak, promiscuous activity come into action and thereby provide an evolutionary starting point, these increased doses may also become deleterious owing to the very same effect. Higher expression levels might also increase the protein's sensitivity to mutations as misfolding and aggregation are concentration-dependent<sup>44</sup>. Thus, placement of an evolving protein under a relevant regulatory regime necessitates optimal, long-term solutions, the most common of which is the divergence of a duplicated gene that possesses higher activity as well as a new regulation scheme.

**Chaperones.** Chaperones also comprise a system for buffering genetic perturbations and can thus promote the acquisition of higher genetic diversity<sup>45,46</sup>. In addition, chaperones mediate adaptive evolution by buffering the deleterious effects of mutated genes that mediate new functions<sup>47</sup>. The capacity of the bacterial GroEL to buffer destabilizing mutations has been demonstrated<sup>48,49</sup>. Mutational drift (*in vitro* mutagenesis and purifying selection for the maintenance of protein function) in the presence of GroEL overexpression doubled the number of accumulating mutations by enabling the correct folding of enzyme variants that carry mutations with much higher destabilizing effects (>3.5 kcal per mol  $\Delta\Delta G$  values on average, versus ~1 kcal per mol in the absence of GroEL)<sup>49</sup>. The chaperones also enabled an evolving protein to circumvent the stability trade-off of a new-function mutation, and this facilitated the acquisition of a new enzymatic specificity — in terms of the number of newly evolved variants — and in their higher specificity and activity ( $\geq 10$ -fold)<sup>49</sup>.

**Chaperones**  
Proteins that mediate the correct folding and assembly of other proteins.

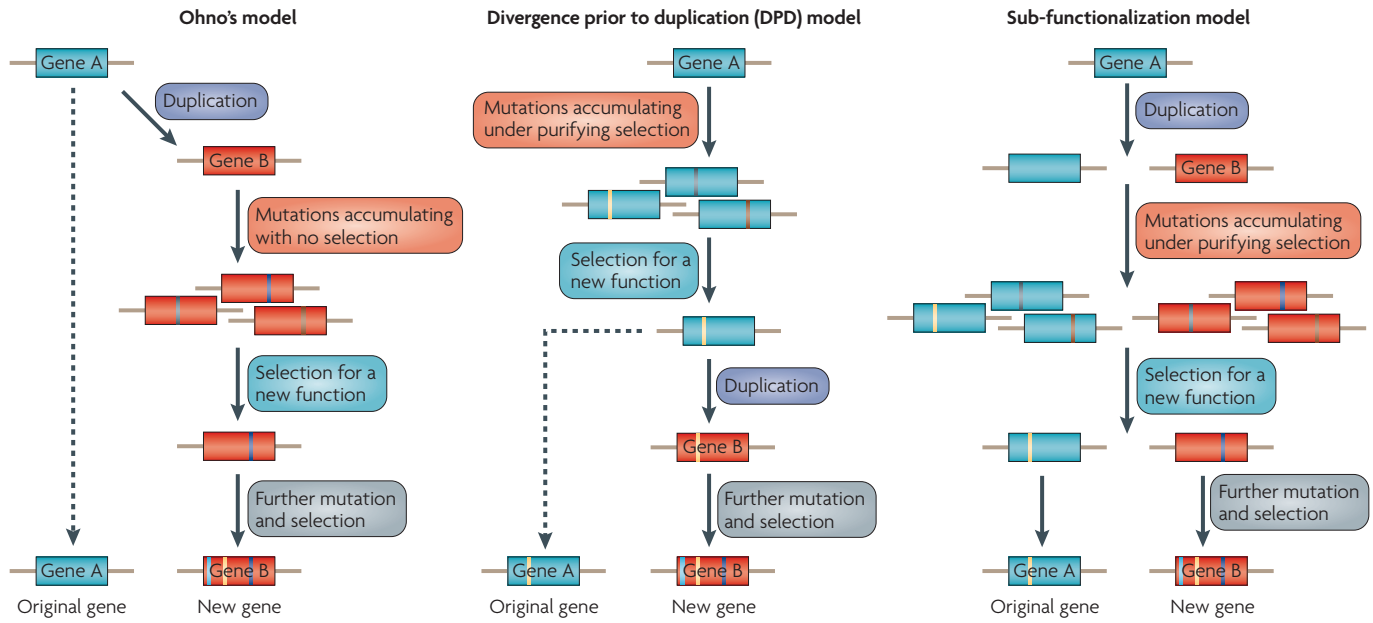
**Compensatory, stabilizing mutations.** Because new-function mutations are destabilizing, they are usually followed by stabilizing, compensatory mutations in the same protein<sup>15</sup>. Indeed, compensatory mutations are crucial in evolution and in particular in small populations<sup>50</sup>. The frequency of such mutations can be high. In TEM1, for example, more than 15 different stabilizing mutations have been identified<sup>51–54</sup>. At least ten of these were shown to act as global suppressors and to compensate for the destabilizing effects of a wide range of mutations, including new-function mutations<sup>51</sup>. Compensatory mutations are largely neutral (or even slightly deleterious) on a wild-type background, but may become highly advantageous on other backgrounds<sup>55</sup> and particularly in the background of new-function mutations. The accumulation of new-function mutations under purifying selection is therefore possible<sup>11,51</sup>. Once present, such stabilizing mutations dramatically increase the likelihood of adaptation<sup>51,56</sup>. Other compensatory mutations can be deleterious on their own and advantageous only in combination with a new-function mutation. They are therefore likely to appear in genes that drift under no — or very weak — selection. At present, however, the relative occurrence of this epistatic mode of compensation is unknown.

#### Divergent evolution of protein-coding genes

The divergence of new genes and their encoded proteins is a key evolutionary process that occurred multiple times throughout evolution. It is manifested in the existence of families and super-families of proteins that have each diverged from one common ancestor. Family members are characterized by similarities in structure, sequence and function, but also by the diversification of these features — in particular, of sequence and function. For example, members of a receptor family might all bind a ligand and subsequently activate a downstream response. However, each family member recognizes a distinctly different ligand. The variability of ligands in one family can be very high, for example, members of the Per-Arnt-Sim (PAS) receptor family recognize signals as diverse as photons and polyaromatic hydrocarbons<sup>57</sup>.

As a fundamental evolutionary process, the divergence of new genes and proteins has been examined from many angles. Our discussion focuses on the effects and trade-offs of protein mutations. For clarity, and owing to space constraints, we are forced to generalize some of the arguments and to omit important aspects of evolutionary theory and population genetics that are addressed elsewhere<sup>1,2</sup>.

**Gene duplication.** Divergence can occur through duplication of the ancestral gene to give two related genes, or 'paralogues' (*genA* and *genB*; FIG. 3). Duplication is a very frequent event. This is apparent in a variety of eukaryotic genomes in which 1–10% of genes seem to coexist with their nearly identical duplicates<sup>58</sup> and by the high variability of gene copy number between individual genomes<sup>59,60</sup>. The duplicated segments span from several bases up to whole genomes. However, our discussion revolves around the schematic case of duplication of one gene.



**Figure 4 | Divergent evolution via gene duplication.** Three basic models are described that differ in several ways. In both Ohno's model and the sub-functionalization model, duplication is a neutral event that is neither selected for nor against. By contrast, the 'divergence before duplication' (DPD) model assumes that duplication is positively selected because higher protein doses provide an immediate advantage. In Ohno's model, the adaptive mutation (or mutations) accumulates in the duplicate under no selection. However, in the two other models, mutations occur under selection: in sub-functionalization, purifying selection acts on both copies to maintain the original function, and in the DPD model, mutations with adaptive potential are selected for before duplication. Finally, Ohno's model is asymmetric — by default, the new function arises in the duplicated copy. By contrast, the DPD and sub-functionalization models are in principle symmetric (the new function may appear in either the original or the duplicated copy), although in effect the new function is more likely to be mediated by the duplicated copy. For a detailed discussion of these models, see REFS 1,2.

**Divergence in the absence of duplication.** Divergence of sequence and function also occurs without duplication. Sequence divergence occurs constantly and gives rise to orthologues, the functions of which are considered to be identical (FIG. 3). Many of the sequence differences between orthologues are therefore considered neutral — that is, the sequence differences are not associated with adaptation towards new functions. The degree to which these assumptions are true is unclear. Many genes that are assigned as orthologues encode proteins with different activities in different organisms, in particular when these functions are not part of the well-defined core metabolism<sup>61</sup>. Features other than activity per se also evolve — for example, the adaptation of a protein to different cellular compartments or tissues and to the partners with which they interact (compartmental adaptation). Although our discussion focuses on the functional divergence of paralogues through gene duplication, the mutational effects that are discussed here also apply to the divergence of sequence and function in orthologues.

**Mechanisms of divergence: prevailing models**

What are the driving forces for gene duplication, and what are the selection pressures that shape the evolving gene pair? To address these questions, we outline three basic models of divergence and examine them in light of the mutational effects and trade-offs that are discussed above.

**Ohno's model.** According to this classical model<sup>62,63</sup>, duplication is a neutral event and the redundant duplicated copy freely accumulates mutations under no selection. Only when a new need arises do mutations that endow a new function come under selection (FIG. 4). Over the years, Ohno's model has been questioned, primarily because it predicts that duplicated copies (close or nearly identical paralogues) would drift under no selection, whereas the analysis of genomes indicates the contrary<sup>24,64</sup>. Several other lines of evidence run counter to the expectations of Ohno's model. The expression of a redundant copy carries an energetic cost<sup>40,41</sup> that may lead to a selective pressure to inactivate it. Genes that evolve under no selection also accumulate commonly occurring destabilizing mutations that lead to protein aggregation and that may reduce organismal fitness<sup>18</sup>. Indeed, empirical evidence frequently indicates that both members of a gene pair are expressed and maintained under purifying selection<sup>24,65,66</sup>. Furthermore, as discussed above, gene duplication is often not a neutral event, but is positively selected under demands for higher protein doses<sup>67,68</sup>.

From the point of view of mutational effects, the feasibility of Ohno's model is affected by two factors. On the one hand, because the frequency of mutations that completely inactivate proteins is high, non-functionalization is the most likely fate of genes that accumulate mutations under no selection (BOX 1; FIG. 1 a).

On the other hand, the assumption of a fitness trade-off between the existing and the new function underlines the need to duplicate and generate a redundant copy that is free from the burden of selection<sup>69</sup>. Inherent to Ohno's model is therefore the view of exquisite, absolute specificity; that genes and proteins are specialists — one sequence equals one structure and one function. This view of protein function has become untenable<sup>23,70</sup>. Therefore, although certain cases of strong trade-offs (as exemplified earlier) may render Ohno's model the most feasible, alternative modes of divergence are based on weak trade-offs and on the notion of 'gene sharing'<sup>71,72</sup> — that is, one gene or protein performing more than one function.

**Genes and proteins as generalists.** Gene sharing was inspired by the discovery that several structural eye-lens proteins (crystallins) are identical, or nearly identical, in sequence to certain metabolic enzymes<sup>71,72</sup>. Since then, numerous examples of proteins with multiple functions have accumulated, including proteins that are renowned for their exquisite specificity. For example, in mast cells, lysyl-tRNA-synthetase catalyses the synthesis of a signalling molecule, diadenosine tetraphosphate (Ap4A)<sup>73</sup>. Under specific pathogenic circumstances, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), a common cytosolic metabolic enzyme, is transferred to the nucleus and signals a death cascade<sup>74</sup>. RNase A is another example of a protein that has evolved host-defence functions that are unrelated to its original enzymatic activity<sup>75</sup>. These secondary functions are all organism and/or tissue specific and have therefore evolved well after the primary function. Thus, although specialization may ultimately depend on duplication, generalist intermediates are a feasible option.

The feasibility that new functions may develop in an existing protein without necessarily compromising its original function is also supported by the notion of promiscuous and/or moonlighting functions. Many proteins perform functions that they did not evolve for. This functional pleiotropy is related to the conformational pleiotropy of proteins. That is, the existence of multiple structures in the same sequence also enables the existence of multiple functions<sup>23,70</sup>. These coincidental, promiscuous activities often serve as starting points for the evolution of new functions if, or when, the need arises<sup>76–79</sup>. For example, enzymes that evolved millions of years ago can exhibit promiscuous activities towards chemicals that were introduced only several decades ago. This provides ample starting points for the divergence of new enzymes that specialize in degrading man-made chemicals<sup>80,81</sup>. The structural, mechanistic and evolutionary aspects of protein promiscuity are discussed in recent reviews<sup>21,82</sup>.

**Divergence before duplication model.** Several divergence models that are based on generalist intermediates have been proposed<sup>22,65,69</sup>, and they are grouped here under the title of DPD. Models such as the 'innovation–amplification–divergence' model<sup>68</sup> and the 'escape from adaptive conflict' model<sup>83–85</sup> also belong to this category.

According to the DPD model, an initial level of the new, evolving function is acquired while the original function is maintained (FIG. 4). Duplication occurs after the new function also becomes under positive selection. Duplication may provide an immediate advantage by increasing protein levels, thereby compensating for the low efficiency of the new, evolving function — it also eventually enables two specialists to emerge from a generalist intermediate.

**Sub-functionalization model.** The third model — the sub-functionalization model, also known as the 'duplication–degeneration–complementation' (DDC) model<sup>86</sup> — combines elements from both Ohno's model and the DPD model (FIG. 4). It is based on the hypothesis that deleterious mutations can accumulate in either the original or the copy owing to the relief in selection pressure that is afforded by duplication<sup>64,87</sup>. The two copies may therefore acquire complementary loss-of-function mutations such that both genes are now required to maintain the function of a single ancestral gene. This model was originally devised for regulatory elements<sup>88</sup>, but it can readily be extended to proteins<sup>11,24</sup>.

**Mutational effects and divergence models.** With respect to mutational fitness effects, the DPD and sub-functionalization models minimize the threat of highly deleterious and non-functionalization mutations because the diverging gene remains constantly under selection (FIG. 1 b). Furthermore, following duplication, mildly deleterious ARMs can occur with higher frequency<sup>24</sup> (BOX 1). Thus, duplication allows the accumulation of ARMs, whereas ARMs cause the duplicated copy to come under selection, and this results in a mechanism that maintains duplicated genes and ARMs because they are interdependent. The feasibility of the DPD and sub-functionalization models depends, however, on the new-function mutations having weak effects on the existing function, namely on having no (or weak) trade-offs. Therefore, from the point of view of mutational effects, the forces that favour the DPD and the sub-functionalization models also work against them. By contrast, new–existing function trade-offs do not interfere with Ohno's model, in which new function arises from redundant gene copies that drift under selection.

The magnitude of mutational trade-offs therefore affects the feasibility of these divergence models, and the mutation in *genA* that becomes selectable for the new function is the crucial 'decision point'. The relative likelihood of the different divergence modes (FIG. 4) would be determined by the interplay among the trade-offs that are associated with this mutation, the respective buffering mechanisms, and the magnitude of selection that acts on the existing and the new functions. If trade-offs are sufficiently weak or masked by buffering mechanisms other than duplication, the need for immediate duplication is alleviated and the DPD model becomes most feasible. If the trade-offs (the new-function–stability trade-offs in particular) are stronger, so that the new-function mutation reduces protein

#### Specialists

Genes or proteins that exert one specific function with high proficiency.

#### Generalist

A gene or protein that exerts multiple functions, typically one primary function and additional secondary or promiscuous functions.

#### New-function–stability trade-offs

Mutations that increase the new, evolving function but reduce protein stability and protein dosage.

## Box 2 | Open questions and future directions

**The distribution of mutational effects — beyond single proteins**

The fitness effects of mutations in a cellular, let alone organismal, context are much more complex than in isolated proteins such as TEM1  $\beta$ -lactamase. Future aims include the examination of fitness effects under native conditions (such as the expression of a chromosomal gene from its endogenous promoter) and the investigation of how the fate of mutated variants is affected by various buffering mechanisms and by mechanisms of protein trafficking and clearance.

**The distribution of mutational effects — is an organism a sum of its proteins?**

Surprisingly, on preliminary examination, organisms seem to be more sensitive to mutations than their component proteins (BOX 1). The overall similarity in the distributions of fitness effects of mutations for a single protein and an intact organism, in addition to the differences, demand further exploration.

**Studies of natural protein divergence**

Our knowledge of protein mutations and their effects comes primarily from the study of model cases. In only a few cases has the natural divergence of a protein (or its clinical divergence, as in TEM1) been subsequently studied in the laboratory. Studies of actual adaptations will provide an interesting and more conclusive picture of how new genes and proteins diverge<sup>5</sup>. Potential study targets include adaptations towards man-made chemicals, such as pesticides and herbicides in bacteria<sup>81,95,96</sup>, insects<sup>97</sup> or plants<sup>98</sup>. Secondary metabolism in plants is another rich source of functional diversification<sup>99</sup>.

**Reconstructions of ancestors and their divergence paths**

Reconstructing ancestral genes and proteins and the divergence paths that led to contemporary proteins is a powerful approach that can also be applied to examine the mutational paths and the effects of individual mutations on the divergence process<sup>5</sup>.

**Compartmental adaptations**

'Function' refers not only to what a protein does in a living organism but also to where it does it. The same activity (for example, catalysing a given enzymatic transformation) can take place in different compartments or organisms. This means that proteins can be placed under different regulation schemes and be processed differently with regard to synthesis, transport, pH optimum, stability, and so on. It will be interesting to explore the sequence changes that drive compartmental adaptation, as these could be more intense than the sequence alterations that drive changes in activity itself.

**Divergence by horizontal gene transfer**

Horizontal (lateral) gene transfer is a common source of evolutionary innovation. A related gene imported from another organism partly resembles a newly formed duplicate, so it will be worth investigating whether the above-discussed mechanisms, mutational effects and their trade-offs apply.

**Evolutionary rates**

The rates of evolution of proteins (the average number of amino acid exchanges per position, per generation) are widely distributed between organisms (viral proteins provide a clear example)<sup>14</sup>, as well as within the same organism. Many factors might be involved<sup>4,44</sup>, but the effects of the protein's structure and its response to mutations (the distribution of mutational effects) remain unclear.

**Epistasis and protein evolution**

Although it is beyond the scope of this Review, epistasis is an important factor in protein evolution<sup>100–102</sup>. If the effect of a given mutation depends on whether another mutation (or mutations) is present, the likelihood and mechanism of divergence will be affected. For example, new-function mutations may not be fixed unless a stabilizing, compensatory mutation is already present. Conversely, a compensatory mutation can be neutral or even deleterious on its own but beneficial in combination with a destabilizing mutation<sup>17,49,56,103</sup>. Exploring the mutations that underline various divergence paths may provide new insights regarding the role of epistasis in directing the mechanism of divergence.

dosage, duplication via either the sub-functionalization or the DPD model is the most plausible route. Finally, if the above-mentioned mutation exhibits prohibitively high trade-offs, Ohno's model would be the most likely mode of divergence.

The DPD and sub-functionalization models also change the fitness landscape such that previously neutral — or even slightly deleterious — mutations that possess some adaptive potential may become advantageous under a different context or environment. This change is possibly the most fascinating facet of mutational pleiotropy and is also related to the notions of hidden variation and neutral networks.

**Hidden variation and neutral networks.** The frequency of new-function mutations is of the order of  $10^{-3}$  (FIG. 1b). If evolution of a new function depends on the simultaneous acquisition of two or more mutations, the frequency of adaptive events becomes impossibly low. However, the frequency dramatically increases if adaptive mutations accumulate as 'apparently neutral'. Thus, neutrality could be a facilitator rather than an opposing force to evolutionary innovation<sup>89,90</sup>. Neutrality, or hidden genetic variation (variation that has no apparent effect on fitness), and mechanisms such as those mediated by chaperones that increase its prevalence may therefore promote adaptation<sup>46,49</sup>.

## Productive variation

Genetic variation that does not compromise fitness in the dwelling environment but holds the potential for adaptation to new environments.

The role of hidden variation is also generalized under the description of neutral networks. These are sets of genotypes that exhibit the same phenotype and that are connected by single mutations. However, phenotype and fitness are defined by the traits that are under selection at a given time. A change in genotype may have no apparent effect on phenotype at present, but can change the potential for future adaptations: “by moving neutrally something does vary: the potential for change”<sup>91</sup>. The notion of neutral networks is tightly linked with the prevalence of weak new-existing function trade-offs and with proteins exhibiting alternative structures and functions alongside their primary one<sup>21,23,70,92</sup>. This latent pleiotropy provides the basis for the expansion of neutral networks and for the manner in which they change the adaptive potential of proteins.

## Summary and future research directions

We have proposed that, among other factors<sup>1,2</sup>, the likelihood of a particular mechanism of divergence occurring is determined by mutational effects on protein function and stability, by trade-offs between mutational effects and by alleviation of these trade-offs. The key to adaptation is genetic variation or, more precisely, productive variation — namely the sequence variations that do not compromise organismal fitness under the current state but maintain the potential to adapt to new states. Duplication is a key mechanism for the acquisition of productive variation. However, laboratory evolution

experiments indicate that purging of non-functionalization mutations provides a major advantage. Thus, although duplication is a primary means of increasing variation and, in particular, of circumventing trade-offs, it increases variation not necessarily through the complete removal of the selection pressure (as in Ohno's model) but by reducing selection pressures under modes of divergence such as sub-functionalization. Indeed, the stringency of selection that acts on the diverging gene is of key importance<sup>93</sup> — total relaxation of selection results in rapid non-functionalization (FIG. 1), whereas a highly stringent selection purges nearly all new-function mutations due to trade-offs<sup>11</sup>. A relaxed level of purifying selection owing to the higher gene and protein doses that are afforded by duplication<sup>24</sup> and to the other buffering and compensatory mechanisms that are noted above, enables the accommodation of a larger fraction of mutations that includes adaptive mutations.

Although our knowledge of how mutations affect protein structure and function, and subsequently protein evolution, has increased dramatically, many aspects remain unexplored — some of these are outlined in BOX 2. The overall, long-term goal is to provide an integrated yet detailed description of protein evolution that includes the selection forces that acted on the evolving organism, as well as the role and effects of mutations in shaping the function and structure of its individual proteins, and of the accompanying changes in non-coding, regulatory regions.

- Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nature Rev. Genet.* **9**, 938–950 (2008).
- Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Rev. Genet.* **11**, 97–108 (2010).
- DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6**, 678–687 (2005).
- Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7**, 337–348 (2006).
- Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* **8**, 675–688 (2007).
- Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
- Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* **8**, 610–618 (2007).
- Camps, M., Herman, A., Loh, E. & Loeb, L. A. Genetic constraints on protein evolution. *Crit. Rev. Biochem. Mol. Biol.* **42**, 313–326 (2007).
- Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA* **102**, 606–611 (2005).
- Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
- Bershtein, S. & Tawfik, D. S. Ohno's model revisited: measuring the frequency of potentially adaptive mutations under various mutational drifts. *Mol. Biol. Evol.* **25**, 2311–2318 (2008).
- Hecky, J. & Muller, K. M. Structural perturbation and compensation by directed evolution at physiological temperature leads to thermostabilization of  $\beta$ -lactamase. *Biochemistry* **44**, 12640–12654 (2005).
- Yue, P. & Moul, J. Identification and analysis of deleterious human SNPs. *J. Mol. Biol.* **356**, 1263–1274 (2006).
- Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezhovsky, I. N. & Tawfik, D. S. Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* **34**, 53–59 (2009).
- Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320**, 85–95 (2002).
- Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput. Biol.* **4**, e1000002 (2008).
- Levin, K. B. *et al.* Following evolutionary paths to protein–protein interactions with high affinity and selectivity. *Nature Struct. Mol. Biol.* **16**, 1049–1055 (2009).
- Lindner, A. B., Madden, R., Demarez, A., Stewart, E. J. & Taddei, F. Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proc. Natl Acad. Sci. USA* **105**, 3076–3081 (2008).
- McLoughlin, S. Y. & Copley, S. D. A compromise required by gene sharing enables survival: implications for evolution of new enzyme activities. *Proc. Natl Acad. Sci. USA* **105**, 13497–13502 (2008).
- Vick, J. E., Schmidt, D. M. & Gerlt, J. A. Evolutionary potential of  $(\beta/\alpha)_8$ -barrels: *in vitro* enhancement of a 'new' reaction in the enolase superfamily. *Biochemistry* **44**, 11722–11729 (2005).
- Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Ann. Rev. Biochem.* **79**, 471–505 (2010).
- Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nature Genet.* **37**, 73–76 (2005).
- Tokuriki, N. & Tawfik, D. S. Protein dynamism and evolvability. *Science* **324**, 203–207 (2009).
- Scannell, D. R. & Wolfe, K. H. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res.* **18**, 137–147 (2008).
- Kaessmann, H. Genetics. More than just a copy. *Science* **325**, 958–959 (2009).
- Parker, H. G. *et al.* An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**, 995–998 (2009).
- Andersson, D. I. & Hughes, D. Gene amplification and adaptive evolution in bacteria. *Annu. Rev. Genet.* **43**, 167–195 (2009).
- Schimke, R. T. Gene amplification in cultured cells. *J. Biol. Chem.* **263**, 5989–5992 (1988).
- Papp, B., Pal, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661–664 (2004).
- Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genet.* **39**, 1256–1260 (2007).
- Fablet, M., Bueno, M., Potrzebowski, L. & Kaessmann, H. Evolutionary origin and functions of retrogene introns. *Mol. Biol. Evol.* **26**, 2147–2156 (2009).
- Jablonka, E. & Lamb, M. J. *Epigenetic Inheritance and Evolution: The Lamarckian Dimension* (Oxford Univ. Press, Oxford, UK, 1995).
- Steele, E. J., Lindley, R. A. & Blanden, R. V. *Lamarck's Signature: How Retrogenes Are Changing Darwin's Natural Selection Paradigm*, (Allen & Unwin; Perseus Books, Australia, 1988).
- Chen, G. K. *et al.* Preferential expression of a mutant allele of the amplified *MDR1* (*ABCB1*) gene in drug-resistant variants of a human sarcoma. *Genes Chromosomes Cancer* **34**, 372–383 (2002).
- Qian, W. & Zhang, J. Gene dosage and gene duplicability. *Genetics* **179**, 2319–2324 (2008).
- Goldsmith, M. & Tawfik, D. S. Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc. Natl Acad. Sci. USA* **106**, 6197–6202 (2009).
- Siu, L. K., Ho, P. L., Yuen, K. Y., Wong, S. S. & Chau, P. Y. Transferable hyperproduction of TEM-1  $\beta$ -lactamase in *Shigella flexneri* due to a point mutation in the *prb* box. *Antimicrob. Agents Chemother.* **41**, 468–470 (1997).
- Hall, B. G. Evolution of a regulated operon in the laboratory. *Genetics* **101**, 335–344 (1982).
- Hall, B. G. The EBG system of *E. coli*: origin and evolution of a novel  $\beta$ -galactosidase for the metabolism of lactose. *Genetica* **118**, 143–156 (2003).
- Stoebel, D. M., Dean, A. M. & Dykhuizen, D. E. The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. *Genetics* **178**, 1653–1660 (2008).

41. Wagner, A. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**, 1365–1374 (2005).
42. Vavouri, T., Sempke, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198–208 (2009).
43. Veitia, R. A. Gene dosage balance: deletions, duplications and dominance. *Trends Genet.* **21**, 33–35 (2005).
44. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14338–14343 (2005).
45. Fares, M. A., Ruiz-González, M. X., Moya, A., Elena, S. F. & Barrio, E. Endosymbiotic bacteria: GroEL buffers against deleterious mutations. *Nature* **417**, 398 (2002).
46. Rutherford, S., Hirate, Y. & Swalla, B. J. The Hsp90 capacitor, developmental remodeling, and evolution: the robustness of gene networks and the curious evolvability of metamorphosis. *Crit. Rev. Biochem. Mol. Biol.* **42**, 355–372 (2007).
47. Cowen, L. E. & Lindquist, S. Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. *Science* **309**, 2185–2189 (2005).
48. Parent, K. N., Ranaghan, M. J. & Teschke, C. M. A second-site suppressor of a folding defect functions via interactions with a chaperone network to improve folding and assembly *in vivo*. *Mol. Microbiol.* **54**, 1036–1050 (2004).
49. Tokuriki, N. & Tawfik, D. S. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* **459**, 668–673 (2009).
50. Zhang, L. & Watson, L. T. Analysis of the fitness effect of compensatory mutations. *Hfsp J.* **3**, 47–54 (2009).
51. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).
52. Hecky, J., Mason, J. M., Arndt, K. M. & Muller, K. M. A general method of terminal truncation, evolution, and re-elongation to generate enzymes of enhanced stability. *Methods Mol. Biol.* **352**, 275–304 (2007).
53. Kather, I., Jakob, R. P., Dobbek, H. & Schmid, F. X. Increased folding stability of TEM-1  $\beta$ -lactamase by *in vitro* selection. *J. Mol. Biol.* **383**, 238–251 (2008).
54. Marciano, D. C. *et al.* Genetic and structural characterization of an L201P global suppressor substitution in TEM-1  $\beta$ -lactamase. *J. Mol. Biol.* **384**, 151–164 (2008).
55. Kimura, M. The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **64**, 7–19 (1985).
56. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874 (2006).
57. McIntosh, B. E., Hogenesch, J. B. & Bradfield, C. A. Mammalian Per-Arnt-Sim proteins in environmental adaptation. *Annu. Rev. Physiol.* **72**, 625–645 (2010).
58. Lynch, M. Genomics. Gene duplication and evolution. *Science* **297**, 945–947 (2002).
59. Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Rev. Genet.* **8**, 639–646 (2007).
60. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Rev. Genet.* **10**, 551–564 (2009).
61. Liao, B. Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl Acad. Sci. USA* **105**, 6987–6992 (2008).
62. Ohno, S. *Evolution by Gene Duplication* (Allen & Unwin; Springer, New York, 1970).
63. Kimura, M. & Ota, T. On some principles governing molecular evolution. *Proc. Natl Acad. Sci. USA* **71**, 2848–2852 (1974).
64. Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
65. Hughes, A. L. Adaptive evolution after gene duplication. *Trends Genet.* **18**, 433–434 (2002).
66. Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *Trends Genet.* **20**, 544–549 (2004).
67. Kondrashov, F. A. & Koonin, E. V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* **20**, 287–290 (2004).
68. Bergthorsson, U., Andersson, D. I. & Roth, J. R. Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl Acad. Sci. USA* **104**, 17004–17009 (2007).
69. Kondrashov, F. A. In search of the limits of evolution. *Nature Genet.* **37**, 9–10 (2005).
70. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem. Biol.* **5**, 789–796 (2009).
71. Piatigorsky, J. *et al.* Gene sharing by  $\Delta$ -crystallin and argininosuccinate lyase. *Proc. Natl Acad. Sci. USA* **85**, 3479–3483 (1988).
72. Piatigorsky, J. *Gene Sharing and Evolution: The Diversity of Protein Functions*, (Harvard Univ. Press, Cambridge, Massachusetts, USA; London, UK, 2007).
73. Lee, Y. N., Nechushtan, H., Figov, N. & Razin, E. The function of lysyl-tRNA synthetase and Ap4A as signaling regulators of MITF activity in Fc $\epsilon$ R1-activated mast cells. *Immunity* **20**, 145–151 (2004).
74. Sedlak, T. W. & Snyder, S. H. Messenger molecules and cell death: therapeutic implications. *JAMA* **295**, 81–89 (2006).
75. Rosenberg, H. F. RNase A ribonucleases and host defense: an evolving story. *J. Leukoc. Biol.* **83**, 1079–1087 (2008).
76. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1974).
77. O'Brien, P. J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).
78. Palmer, D. R. *et al.* Unexpected divergence of enzyme function and sequence: 'N-acetylamino acid racemase' is *o*-succinylbenzoate synthase. *Biochemistry* **38**, 4252–4258 (1999).
79. James, L. C. & Tawfik, D. S. Catalytic and binding poly-reactivities shared by two unrelated proteins: the potential role of promiscuity in enzyme evolution. *Protein Sci.* **10**, 2600–2607 (2001).
80. Afriat, L., Roodveldt, C., Manco, G. & Tawfik, D. S. The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry* **45**, 13677–13686 (2006).
81. Copley, S. D. Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nature Chem. Biol.* **5**, 559–566 (2009).
82. Copley, S. D. *Comprehensive Natural Products II: Chemistry and Biology* (eds Mander, L. & Liu, H.-W.) (Elsevier, Oxford, 2010).
83. Hughes, A. L. The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256**, 119–124 (1994).
84. Barkman, T. & Zhang, J. Evidence for escape from adaptive conflict? *Nature* **462**, e1; discussion e2–e3 (2009).
85. Des Marais, D. L. & Rausher, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**, 762–765 (2008).
86. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).
87. Dykhuizen, D. & Hartl, D. L. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* **96**, 801–817 (1980).
88. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
89. Nei, M. The new mutation theory of phenotypic evolution. *Proc. Natl Acad. Sci. USA* **104**, 12235–12242 (2007).
90. Wagner, A. *Robustness and Evolvability in Living Systems* (Princeton Univ. Press, Princeton, USA, 2005).
91. Schuster, P. & Fontana, W. Chance and necessity in evolution: lessons from RNA. *Physica D* **133**, 427–452 (1999).
92. Wroe, R., Chan, H. S. & Bornberg-Bauer, E. A structural model of latent evolutionary potentials underlying neural networks in proteins. *Hfsp J.* **1**, 79–87 (2007).
93. Klassen, J. L. Pathway evolution by horizontal transfer and positive selection is accommodated by relaxed negative selection upon upstream pathway genes in purple bacterial carotenoid biosynthesis. *J. Bacteriol.* **191**, 7500–7508 (2009).
94. Wloch, D. M., Szafraniec, K., Borts, R. H. & Korona, R. Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. *Genetics* **159**, 441–452 (2001).
95. Kivisaar, M. Degradation of nitroaromatic compounds: a model to study evolution of metabolic pathways. *Mol. Microbiol.* **74**, 777–781 (2009).
96. Wackett, L. P. Questioning our perceptions about evolution of biodegradative enzymes. *Curr. Opin. Microbiol.* **12**, 244–251 (2009).
97. Newcomb, R. D., Gleeson, D. M., Yong, C. G., Russell, R. J. & Oakeshott, J. G. Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the Rop-1 locus of the sheep blowfly, *Lucilia cuprina*. *J. Mol. Evol.* **60**, 207–220 (2005).
98. Patzoldt, W. L., Hager, A. G., McCormick, J. S. & Tranel, P. J. A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. *Proc. Natl Acad. Sci. USA* **103**, 12329–12334 (2006).
99. O'Maille, P. E. *et al.* Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chem. Biol.* **4**, 617–623 (2008).
100. Lozovsky, E. R. *et al.* Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc. Natl Acad. Sci. USA* **106**, 12025–12030 (2009).
101. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).
102. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99**, 14878–14883 (2002).
103. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).

### Acknowledgements

D.S.T. is the incumbent of the Nella and Leon Benozziyo Professorial Chair. Financial support from the Meil de Botton Aynsley and the EU network BioModularH2 are gratefully acknowledged. We are very grateful to S. Bershtein, N. Tokuriki, F. Kondrashov and J. G. Zhang for their insightful comments regarding this manuscript and to A. Eyre-Walker for providing the data for the figure in Box 1.

### Competing interests statement

The authors declare no competing financial interests.