An integrated view of protein evolution

Csaba Pál[*‡], *Balázs Papp*[§] *& Martin J. Lercher*[*||]

Abstract | Why do proteins evolve at different rates? Advances in systems biology and genomics have facilitated a move from studying individual proteins to characterizing global cellular factors. Systematic surveys indicate that protein evolution is not determined exclusively by selection on protein structure and function, but is also affected by the genomic position of the encoding genes, their expression patterns, their position in biological networks and possibly their robustness to mistranslation. Recent work has allowed insights into the relative importance of these factors. We discuss the status of a much-needed coherent view that integrates studies on protein evolution with biochemistry and functional and structural genomics.

Genetic drift
The stochastic changes in allele frequencies in a population that occur owing to random sampling effects in the formation of successive generations.

Purifying selection
The removal of a deleterious genetic variant from the population owing to the reduced reproductive success of its carriers.

*European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany.
‡Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.
§Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK.
||Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK.
Correspondence to M.J.L.
e-mail:
M.J.Lercher@bath.ac.uk
doi:10.1038/nrg1838

Understanding the causes of variation in protein evolutionary rates is central for many fields, including molecular evolution, comparative genomics and structural biology. Determining the rate of protein evolution (BOX 1) is arguably the most powerful general tool to quantify the relative importance of selection and genetic drift, and to identify selective forces from genomic data. Analyses of protein evolution also provide a unique tool for investigating issues such as the evolution of speciation[1], senescence[2] and social lifestyle[3]; and they facilitate the identification of functionally important sites (to be used in protein design, for example)[4], peptides that are involved in human genetic diseases[5], drug targets[6] or protein interaction partners[7]. Observed rates of evolution can also be used to predict how different mutations might contribute to disease (BOX 2). The power of such analyses can be increased substantially if confounding factors that affect protein evolution are recognized and accounted for.

There has been a large increase in the amount of available genome-scale data in the past few years, prompting a re-examination of some classical assumptions about protein evolution. It is no longer tenable to suppose that protein evolution is only affected by selection on protein structure and function (for a discussion of this issue ahead of its time see REF. 8). There is now an increasing need to form a new integrated theory of protein evolution. We have both progressively sophisticated methods to test the neutral theory of evolution[9] and several, largely isolated ideas on how genomic, cellular and physiological properties affect the ratio of neutral and selected sites. An integrated view would combine these ideas and consider the global properties of proteins under a single conceptual framework. We anticipate that such a coherent theory will have far-reaching consequences on crucial problems in evolutionary biology (BOX 3), but, as discussed below, this description will require the integration of many elements.

Our review concentrates on the causes of rate variation across proteins that are encoded within the same genome; a brief discussion of the variation across species is given in BOX 4. After reviewing the factors that are linked to genomic position, and therefore are independent of individual protein properties, we discuss the causes of protein-specific differences in the strength of purifying selection and summarize the general patterns that have emerged from studying positive selection.

These analyses have revealed some unexpected findings. Genomic variations in mutation and recombination rates seem to have only a small (albeit measurable) influence on protein evolution; the same seems to be true for protein interactions. Surprisingly, the overall importance of proteins (or conversely their dispensability) also seems to be a relatively poor predictor of evolutionary rate. Studies on the yeast *Saccharomyces cerevisiae* indicate instead that the strongest predictor of evolutionary rate is the expression level of a protein. Furthermore, and in contrast to expectations from the neutral theory, many amino-acid changes seem to be due to positive selection, often reflecting arms races or compensatory mutations rather than adaptation to changed environments.

Box 1 | **Measuring protein evolution**

The rate of evolution is obtained by calculating the evolutionary distance between sequences that are derived from a common ancestor, and then dividing this distance by the evolutionary time elapsed since the divergence of the species concerned. If we are interested in relative rates among proteins within the same genome, we can equally use the distances themselves. Evolutionary distances are measured as the number of substitutions per site. However, multiple substitutions can occur at the same site. To correct for such hidden substitutions, we require a model, given in the form of a transition matrix between different states (amino acids, codons or nucleotides). See REF. 110 for an excellent discussion of the different methods to estimate evolutionary rates. Although some simpler models can be solved analytically, calculations are usually carried out numerically, for example, in a maximum-likelihood framework (for an overview of computer packages see REF. 111).

To estimate distances, it is first necessary to identify orthologous proteins, and to align those amino-acid positions that are derived from the same ancestral site. Different amino-acid sites might be under different selective pressures, and so are likely to change in different ways. This variation can be modelled by assuming that each site has a rate that is drawn from a distribution, with the distribution shape constituting another model parameter.

Estimation from amino-acid sequences

The simplest substitution model assumes that all types of amino-acid substitution are equally likely. In this case, the probability of *k* substitutions at any given site follows a Poisson distribution. However, substitutions occur more often between amino acids that are similar in their biochemical properties. Empirical transition matrices are estimated from large sets of protein sequence alignments; statistical criteria can be used to select the best fit for the sequences under study[112].

Estimation from nucleotide sequences

Amino-acid changes are based on nucleotide mutations of the encoding genes. In contrast to amino-acid-based models, it is feasible to estimate the transition matrix for each analysed pair of sequences individually; again, the best-fitting model can be selected on the basis of statistical criteria[113]. The rate of protein evolution is then measured as non-synonymous distances (termed d_N or K_A), which are obtained by dividing the number of non-synonymous changes by the number of non-synonymous sites (that is, sites at which substitutions lead to amino-acid changes)[110].

Estimation from codon sequences

Nucleotide-based models can account for nucleotide mutational biases, whereas amino-acid-based methods can account for different substitution probabilities between amino-acid pairs. To combine the strengths of these two approaches, it is necessary to consider substitutions between codons. This method is feasible for reasonably long alignments[114], and results in estimates of d_N and $\omega = d_N/d_S$, the ratio of non-synonymous to synonymous rates, which reflects selection pressures on the protein[110].

Positive selection
The accelerated spread of a beneficial genetic variant in the population owing to the increased reproductive success of its carriers.

Dispensability
A measure that is inversely related to the overall importance of a gene. It is usually approximated by the fitness (or growth rate) of the corresponding gene knockout strain under various laboratory conditions.

Transition matrix
A matrix that contains the probabilities of each type of amino-acid substitution for a given period of evolution.

The evolution of amino-acid sequences is inextricably linked to the evolution of associated regulatory sequences and synonymous coding sites (which affect translational efficiency and mRNA stability). With the exception of considerations that relate to protein structure and interactions, the topics outlined below might be equally relevant to the evolution of such sites.

Influence of regional genomic properties

Protein evolution requires two steps: the mutation of nucleotides that code for amino acids, and the fixation of new variants in the population. The probability of fixation depends on the fitness effect of mutations[10–12]; the new variant can be neutral or nearly neutral (and so governed purely or largely by genetic drift, respectively), deleterious (and consequently opposed by purifying selection), or advantageous (and therefore supported by positive selection). Although most analyses of protein evolution aim to identify selective forces, both mutation rate and efficiency of selection (and so fixation probability) vary systematically across genomic regions, as discussed in this section. This variation results in a component of evolutionary rate that is independent of the properties of individual proteins.

Variation in mutation rate. Although genomic variation in mutation patterns has been observed in diverse bacterial and eukaryotic species, the best evidence for such variation comes from mammals, in which it is found on different genomic scales[13]: between nucleotides[14], and within and between chromosomes[15].

Variation in nucleotide mutation rate causes an associated variation in the rate of protein evolution. This relationship is supported by a recent analysis[16], which concluded that the amino-acid changing (non-synonymous)[15,17] and amino-acid conserving (synonymous)[15,18] nucleotide sites of neighbouring mammalian genes evolve at similar rates. What fraction of the variation in protein evolution is attributable to these mutation-rate differences? The answer is unclear, partly because some of the factors that are associated with variation in mutation rate also influence the fixation probability of new mutants. For example, genomic regions with high mutation rates also have high expression levels[19] and recombination rates[20], where both high expression and high recombination are associated with a reduced fixation rate of deleterious mutation.

Recombination gives a particularly clear example of this problem. The association between recombination and mutation rates is probably caused by error-prone polymerases, which are involved in the repair of double-strand breaks that initiate recombination[21]. However, the mutagenic effects of recombination have remained neglected by theoreticians, which instead focus exclusively on the role of recombination in decoupling the evolution of physically linked loci[22]. This omission is especially worrying for models that examine the spread of favourable mutations, as expectations from such models are similar to those from a neutral null model that is based on mutational effects alone: both predict that proteins encoded by genes under high recombination rates should evolve quickly[23].

Another potential problem with assessing the effect of mutation-rate variation on protein evolution lies in the fact that nucleotide mutation rates are often measured by the substitution rate at synonymous sites. Based on this method, it has recently been suggested that as much as 40% of the variation in protein evolution might be attributable to differences in the underlying mutation rate[24]. However, there is evidence for selection on synonymous sites in many organisms, including mammals[25], and therefore rate of evolution at synonymous sites is only a crude approximate measure of mutation rate. To solve this question conclusively, variation in nucleotide mutation rates across the genome would have to be estimated experimentally. Obtaining such an estimate is beyond the scope of current studies, although progress in sequencing technology might make advances possible in the near future.

Maximum-likelihood framework
A method that takes a model (for example, of sequence evolution) and searches for the combination of parameter values that best describes the observed data (for example, the aligned sequences).

Synonymous (change)
A nucleotide change in the protein-coding region of a gene that leaves the encoded amino acid unchanged.

Nearly neutral (mutation)
A mutation is nearly neutral when its fitness effect is too small to be governed only by selection, and so its fate is determined largely by genetic drift.

Non-synonymous (change)
A nucleotide change in the protein-coding region of a gene that alters the encoded amino acid.

Interference (Hill–Robertson effects)
A phenomenon that describes a reduction in the efficiency at which selection functions simultaneously at genetically linked sites, especially in regions of low recombination.

Variation in recombination rate. Protein evolution requires that a mutant spreads until it is fixed in the population (that is, it reaches a frequency of 100%). The fixation probability depends on selection on the mutants. However, efficient selection on individual proteins can be hindered by genetic linkage: proteins that are encoded by neighbouring genes do not evolve independently[23,26]. Harmful mutations can impede the spread of linked advantageous alleles; conversely, if the latter make it to fixation, then the harmful mutations might hitchhike with them. By breaking up the linkage between neighbouring loci in meiosis, recombination increases the efficiency of selection. Consequently, in regions of high recombination, harmful mutations have a lower chance of spreading, whereas advantageous mutations can accumulate more easily. Depending on the relative occurrence of these types of mutation, higher recombination rates can therefore either decrease or increase the rate of protein evolution.

Consistent with the idea that recombination boosts the action of selection, the mean and variance of protein evolution rates are reduced in regions of low recombination in *Drosophila melanogaster* and *Drosophila simulans*[27]. This observation cannot be attributed to regional differences in mutation rate, and is consistent with the claim that positive selection drives a substantial fraction of amino-acid substitutions in these *Drosophila* species[28]. Further support comes from the recent observation that proteins that are encoded in genomic regions with reduced recombination rates suffer more segregating, mildly deleterious mutations, and fix fewer beneficial mutations than genes that experience more recombination[29]. However, it is unclear how many of these correlations remain after controlling for confounding variables.

One such potential confounding factor is transcription: highly expressed proteins tend to evolve slowly[30–33], whereas transcriptionally active regions also seem to be more prone to meiotic recombination[34,35]. Accordingly, although a lower rate of evolution in regions of high recombination in yeast seems to be superficially consistent with increased efficiency of purifying selection against mostly deleterious mutations, this association disappears when controlling for expression level[35]. If this observation can be generalized to other species, previous attempts to study variation in the intensity of selection by examining the covariance of protein evolution with recombination must be treated with caution: the covariance of both variables with germ-cell expression levels has to be eliminated as causative.

From the existing analyses, it seems that fine-scale variation in recombination rate has had little effect on protein evolution. Why might this be so? Interference across sites can be largely alleviated with a very low rate of recombination[22]. Therefore, it is conceivable that only asexual chromosomes experience enough interference to accumulate harmful mutations. For example, on the neo-Y chromosome of *Drosophila miranda* (which became non-recombining ~1 million years ago), 24% of amino-acid substitutions evolved as if they were neutral, compared with 6% on the recombining homologous neo-X chromosome[36]. Many of the changes on this chromosome are deleterious, and seem to have hitchhiked to fixation with a few strongly advantageous substitutions[37].

In summary, although the role of recombination-rate variation in protein evolution seems to be limited, to fully understand it we have to first understand the genomic distribution of fitness effects and control for confounding variables such as expression levels.

Purifying selection

Most variation in the rate of protein evolution is due to protein-specific properties rather than to genomic influences. Several properties have long been suggested to be important factors, including overall protein importance, structural constraints and pleiotropy. More recently, a strong influence of protein-expression patterns has been discovered. These factors are discussed in turn below.

Fitness density. Zuckerkandl proposed 30 years ago that the evolutionary rate of a protein is primarily determined by its functional density, that is, by the proportion of its sites that are involved in specific functions[38]. However, the term functional density might be misleading, as selection that is unrelated to protein function can also affect protein sequences, for example, to increase translational efficiency or accuracy. For this reason, the term fitness density has been suggested instead[39].

To predict the effect of selection on the protein as a whole, we must consider not only the proportion of sites that are affected by selection, but also the distribution of selection strength at these sites (FIG. 1). Fitness (or functional) density measures how much the fitness (or function) of a mutant protein is changed relative to

Box 2 | Predicting disease-causing mutations from evolutionary patterns

The debilitating effect of disease-causing mutations creates selective pressures, the same pressures that have probably resulted in sequence conservation over evolutionary time. Therefore, variation in evolutionary rate across amino-acid sites can be used to predict the severity of medically relevant phenotypes. As expected, disease-associated amino-acid changes occur more often at evolutionarily conserved residues. Moreover, the types of amino-acid change that are associated with disease are those that are not commonly found among closely related species: physico-chemical differences are much more radical for disease-causing mutations than for substitutions between related species[115].

Sequence conservation across species has been used to classify the human SNPs that affect amino-acid sequences[116,117] (for a more detailed account see REF. 118); web servers are available that attempt to predict which substitutions are most likely to pathologically disrupt function for any protein of interest (for example, the Sorting Intolerant from Tolerant (SIFT) database, which predicts, from sequence comparisons, whether a mutation results in a deleterious phenotype[116], and PolyPhen — Prediction of Functional Effect of Human nsSNPs, which predicts the functional effects of mutations from sequence comparisons and biophysical variables[117]). The underlying algorithms (some of which also use structural or functional protein annotation) rely on the assumption that evolutionary rates are determined mostly by stabilizing selection; a comprehensive understanding of the other factors that contribute to rate variation would greatly improve the predictive power of these algorithms.

Although mutations at the most conserved sites of disease-associated genes are those most likely to be involved in pathology, it is currently unclear whether disease genes as a class evolve slower[103] or faster[5] than the rest of the genome.

Box 3 | **Some unsolved problems in protein evolution**

What is the distribution of fitness effects of mutations?
Despite its obvious relevance, only a few pioneering studies have estimated the distribution of fitness effects of mutations (for example, see REF. 119). Even the relative occurrence of neutral, beneficial and deleterious mutations is still a matter of intensive debate[11,12,28,90]. Obtaining precise population-diversity data on the genomic scale, together with new methodological developments[120], will help to resolve these issues.

How does adaptation proceed at the molecular level?
This seemingly simple question incorporates many of the fundamental problems in evolutionary biology. First, it remains unclear how often adaptation at the molecular level proceeds through maladaptive states[40,41]. Theoretical and empirical studies indicate that adaptation frequently involves few mutations with large contributions to fitness that are later adjusted by numerous mutations with small effects[67,121]. How potential trade-offs between different functional and structural requirements affect the fitness distribution is largely unclear (but see REF. 122).

Is neutrality an evolving trait?
It has been claimed that selection might favour organisms that are especially robust to genetic and environmental perturbations, which leads to an increase in selectively neutral variants. Therefore, rather than being a mere result of constraints on protein structure and function, the fraction of neutrally evolving sites might itself reflect an evolved property[91]. Consistent with this proposal, conserved genes are especially likely to undergo diversification by either gene duplication[123,124] or alternative splicing[125]; both these processes relax purifying selection pressure, which leads to an immediate increase in evolutionary rates[87,126].

Metabolic efficiency and protein evolution
If two amino acids at a given position on the protein can do the same job, then selection might favour the retention of the one for which synthesis requires less energy. The amino-acid compositions in the proteomes of *Escherichia coli* and *Bacillus subtilis* reflect the action of such selection pressure[88]. The effect of this force on protein evolution is largely unknown (but see REF. 33).

Is haploid selection a significant force on protein evolution?
In diploid organisms, mutations are exposed to more efficient selection pressure when expressed in genomic regions that have haploid expression patterns (for example, imprinted genes, or sex chromosomes in the heterogametic sex). Therefore, harmful mutations are expected to have a lower chance of spreading in genomic regions with haploid expression, whereas advantageous mutations can accumulate there more easily. Consistent with this idea, genes on the human X chromosome show an elevated tendency for positive selection[92].

that of the wild type. To estimate the selection strength on the mutant organism, this fitness change has to be scaled by the overall importance of the protein (which is inversely related to its dispensability). These considerations predict that fitness density and dispensability are the two most important factors in protein evolution.

The concept of fitness (or functional) density assumes that fitness effects are additive across sites. Whereas most sites might be approximately decoupled in this way, there are strong epistatic interactions between pairs or groups of individual sites. After a mutation at one such site, a compensatory mutation at a paired site can be selectively advantageous even if it was deleterious in the wild-type background[40–42]. Because their fitness effects depend on a genetic background that might vary within a population, such compensatory mutations blur the distinction between purifying and positive selection (see below).

Protein dispensability. Any reduction in protein performance will produce stronger fitness effects in a protein that makes a higher overall contribution to fitness — a protein that is less dispensable when knocked out experimentally. As a consequence, the fitness effects of mutations in less dispensable proteins are less likely to fall below the threshold $1/2N_e$ (where N_e denotes effective population size). This threshold separates nearly neutral from deleterious mutations[11]: the former can go to fixation by genetic drift, whereas the latter are efficiently opposed by purifying selection. If the spread of nearly neutral mutations is the dominant source of evolutionary change (an assumption that is at the heart of the 'nearly neutral theory'[11]), then the higher fraction of such mutations in more dispensable proteins would cause them to evolve faster[8]. Population genetic simulations indicate that this effect should be especially strong when only genes with relatively small fitness effects (around $1/2N_e$) are compared[43], as proteins with much stronger fitness effects are under selective constraints that are comparable to essential proteins[43].

There is indeed a significant correlation between protein dispensability and evolutionary rate in *S. cerevisiae*[43] (FIG. 2), in bacterial species[44] and in *Caenorhabditis elegans*[45]. Although most of these studies lack appropriate controls for confounding variables such as expression level[46], a recent detailed statistical analysis showed that protein dispensability and expression rate make independent contributions to protein evolution[47] (but see REF. 48).

However, there are two substantial problems with linking protein evolution and dispensability. First, contrary to theoretical expectations and original claims[43], this relationship only holds when essential and non-essential proteins are contrasted[49]: quantitative growth-rate data from yeast provide no strong support for the theory when essential proteins are excluded[46]. Second, dispensability explains only a relatively low fraction of the rate variation[33,46,48] (but see REF. 47). There are several potential explanations for these two observations.

Fitness that is measured under nutritionally generous laboratory conditions provides only a crude approximation to the importance of the protein in the wild. Indeed, most yeast enzymes that are marked as dispensable seem to make important fitness contributions in specific environments[50]. As long as organisms regularly encounter these environments, selection might act efficiently on these proteins. Furthermore, dispensability can usually only be measured in extant species, and values might not be representative for the past evolution of the protein[49]. Indeed, the correlation between dispensability and rate of evolution is stronger for closely related species[49].

Moreover, fitness is measured for complete gene knockouts, whereas evolution proceeds largely through point mutations. How protein performance reductions correspond to fitness reductions is probably very different for different classes of protein. Consistent with this reasoning, protein dispensability is indeed much more strongly correlated with the propensity for complete loss of genes than with sequence evolution rates[51].

More fundamentally, and contrary to the nearly neutral theory of evolution, positive selection might not be a negligible force[46] (see below). Indeed, after controlling for positively selected genes, there is no correlation between dispensability and rate of evolution in rodents[52]. Similarly, a recent analysis of mammalian genes found that essential

Fitness density
The proportion of residues in a protein that are under natural selection, with the contribution of each site weighted by the fitness effects of mutations. Besides functional requirements, selection can favour many fitness components, including stability and robustness against errors. Therefore, fitness density is expected to be higher than functional density.

Imprinted gene
A gene in which expression is determined by the parent from which it is inherited.

fertility proteins that affect only one sex (mostly the male) evolve faster than essential viability genes[53]. This difference indicates that the intensity of positive selection rather than dispensability determines the evolutionary rates of these proteins[53].

Overall, functional importance (or dispensability) seems to be a statistically significant, but overrated, determinant of protein evolution.

Protein structure and stability. Irrespective of their dispensability, most proteins require a suitable three-dimensional structure to function. This native structure must be sufficiently thermodynamically stable to ensure that enough active proteins are available in the cell. However, mutation experiments indicate that a large fraction of amino-acid substitutions have a biologically significant effect on protein stability[54] and activity[55]. Most mutations will destabilize rather than stabilize protein structure[41]. Although the most obvious effect of changes in stability and structure might be loss of function, even slight reductions in the efficiency and accuracy of protein folding can lead to protein aggregation and toxicity[56]. Because individual mutations can both decrease and increase structural stability, many amino-acid substitutions might in fact be compensating for deleterious substitutions at other sites of the protein[41] (see below).

The best evidence for selection on protein stability *per se* (with little or no change in protein function) comes from studies of temperature adaptation[57]. Proteins of thermophilic organisms are thermodynamically more stable than orthologous proteins of relatives that live at moderate temperatures, thereby ensuring resilience to high temperatures in thermophilic organisms[58]. A fundamental role of selection for thermodynamic stability in shaping molecular evolution has been demonstrated by studies that simulated sequence evolution under structural constraints. Sequences that are required to be thermodynamically compatible with a given structure showed amino-acid conservation patterns that are similar to those observed in natural proteins[59,60]. That structural constraints enhance the strength of selection on specific amino-acid sites is also supported by studies on site-to-site variation in evolutionary rates[61,62]. For example, rates of non-synonymous changes are about twice as high on the surface of globular proteins as on sites that are less accessible to solvent[62]. Polymorphism data from *Escherichia coli* and *Salmonella enterica* indicate that this difference reflects a decrease in the strength of purifying selection with increasing solvent accessibility[63].

Although structural constraints seem to be associated with a large fraction of among-site variation in protein evolution rates, it remains less clear what fraction of the variation across proteins they explain. Computational methods have revealed that protein designability — the number of possible sequences that are compatible with a given protein structure — is highly variable across structures[64] (where structures with high designability are represented by more diverse sequences in nature[64]). It is therefore tempting to suggest that designability might partly determine the overall substitution rate of proteins.

To further understand the role of structure and stability in protein evolution, researchers now need to link the structural effect of individual amino-acid substitutions to the associated selection pressures. One promising strategy might be the combination of mutagenesis experiments and structural genomics with competitive selection experiments in microbes.

Effective population size
The number of individuals in a population that contribute to the next generation. It is generally much smaller than the number of individuals in the population, and is influenced by factors that include population structure, sex ratio, mating system and age distribution.

Essential protein
One for which deletion of the encoding gene results in a lethal phenotype, which is usually measured under laboratory conditions.

Orthologous
Proteins that are encoded by genes that evolved from a common ancestral gene through speciation.

Protein designability
The number of possible amino-acid sequences that are compatible with a given protein structure.

Overdispersion
When the variance in the substitution rate across lineages exceeds its mean. This indicates that the substitution process does not follow a Poisson distribution.

Box 4 | **Rate variation in protein evolution across species**

The molecular clock hypothesis, proposed in the early 1960s, suggests that protein evolution proceeds at an approximately constant rate over time (for details and further references see REF. 127). Recent systematic studies have convincingly shown that, overall, the evolutionary rates of the proteome vary considerably across species[127], and current research focuses on the extent and causes of such deviations from a universal molecular clock.

One simple explanation is that variation in proteome evolution across species might reflect differences in the underlying mutation rates[127], which are possibly caused by differences in DNA methylation, fidelity of DNA-repair mechanisms or production of DNA-damaging agents. As expected from the relationship between metabolic rate and mutation rate, there is an inverse relationship between body size (which determines metabolic rate) and protein evolution in mammals[128]. A confounding variable in this analysis is, however, generation time[127,128].

A second possibility is that the efficiency of selection against deleterious mutants varies across species, owing to variations in effective population size and/or mode of reproduction. Enhanced rates of protein evolution in intracellular endosymbiotic bacteria is a well-studied example[129]. These organisms not only suffer increased mutation rates that are due to loss of repair enzymes, but also have markedly reduced population sizes and are generally asexual. Recent theoretical analyses claim that the rate of molecular evolution is less sensitive to population size differences than to the extent of genomic linkage[130]. However, comparisons of species with similar ecological niches indicate that substitution rates are indeed affected independently by both population size[131] and mode of reproduction (for example, conversion to asexuality[132] or inbreeding that is due to partial self-fertilization[133]).

Finally, rate variation across lineages could be caused by species-specific differences in the timing and frequency of adaptive evolution. Theoretical studies by Gillespie showed that the frequency of favourable variants sharply reduces as adaptation proceeds[12]. Moreover, adaptation at the molecular level frequently involves only a few adaptive steps, which indicates that protein evolution at a gene occurs in small bursts of adaptive substitutions. Although overdispersion of the molecular clock is generally considered to support these ideas, overdispersion could also arise from purely neutral evolution occurring under structural constraints[134]. So far, few empirical studies have directly investigated the role of adaptive evolution in shaping the molecular clock[67].

Module
A discrete entity that is isolated through spatial localization, gene-expression pattern, chemical specificity or position in biological network (for example, protein complex, metabolic or signal-transduction pathways). Ideally, the biological function of a module is separable from that of other modules.

Overlapping reading frames
Adjacent protein-coding genes that share one or more nucleotides.

Position in biological networks. Protein structure might be further constrained by selection for interactions among proteins[65], leading to a further increase in fitness density and a corresponding reduction in evolutionary rate[38]. This effect is similar to Fisher's classical finding that pleiotropy reduces the likelihood of advantageous mutations and so limits the rate of adaptive evolution[66,67]. The first broad survey on yeast argued that central proteins in protein interaction networks evolve slowly[68]. However, the original claim was probably affected by biases that are inherent in some protein interaction data sets[69], and later studies found that the correlation between the number of protein interactions and the rate of protein evolution is weak or non-existent in yeast[70,71] and *Helicobacter pylori*[71]. Furthermore, expression rate might again be a confounding factor[72].

The aforementioned studies did not distinguish between different types of interaction[73]. This omission might be an important shortcoming, as residues at the interfaces of obligate complexes — but not of transient interactions — tend to evolve slowly[74]. Moreover, the properties of the interacting partner also have an influence[75]. The difference between obligate and transient interactions might also explain the claim that evolutionary innovations tend to occur by altering the interactions between rather than within modules, which is reflected in the accelerated evolution of the connecting proteins (defined as those with transient interactions)[76]. Overall, whereas on theoretical grounds we expect that further interactions will increase the fitness density of a protein, current evidence does not support protein interaction as a strong evolutionary force.

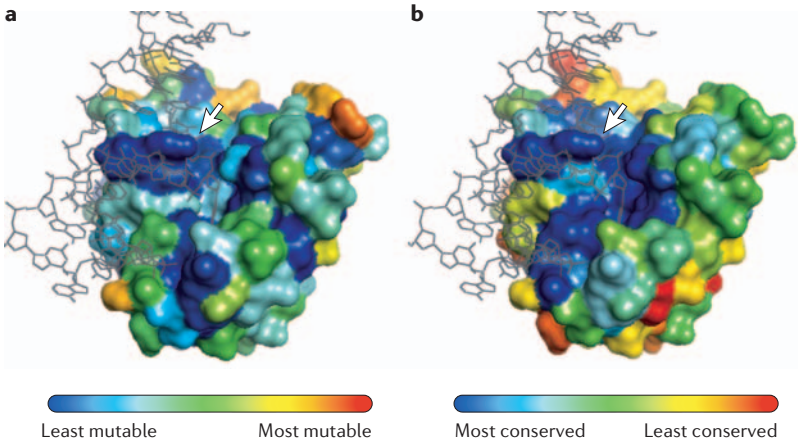

Figure 1 | **Distribution of mutation effects and evolutionary conservation across a DNA-repair enzyme.** The three-dimensional structure of human 3-methyladenine DNA glycosylase (Protein Data Bank: 1F4R). Colours in panel **a** quantify the fraction of mutations that abolish function at each amino-acid site of the enzyme[55]; this value provides a rough estimate of the contribution of each site to fitness density. Colours in panel **b** quantify variability among 159 homologous proteins contained in the UniProtKB knowledgebase, calculated using the ConSurf web server (a server that calculates site-specific amino-acid conservation scores using phylogenetic methods)[135]. The least mutable sites are usually those that are least variable, that is, most conserved in evolution. The interacting DNA molecule is depicted as grey lines; interaction sites, especially the nose pushing the DNA into the active site (see arrow), are among the least mutable and most conserved sites. Panel **a** is re-drawn from the data in REF. 55 using PyMOL.

How does the network position for other types of interaction affect protein evolution? Mammalian proteins with many co-expressed partners (the 'hubs' of co-expression networks) evolve slowly[77]. Although the rate reduction might be due to increased pleiotropy, expression breadth (see below) was not excluded as a confounding factor. Conversely, neither the position in regulatory networks[78] nor the position in metabolic networks[70] seems to have measurable effects on the rate of protein evolution. However, for these networks it is not obvious how network topology and protein pleiotropy are linked, and so the absence of a correlation is not necessarily surprising.

A different measure of pleiotropy is provided by the number of biological processes in which a protein is involved. Although there is a highly significant negative correlation between this number and the rate of protein evolution in *S. cerevisiae*, multifunctionality explains less than 1% of the variation[79]. There is also a positive correlation between environmental specificity and the rate of protein evolution, but again the effect seems to be weak[79]. A more convincing example of the importance of pleiotropic constraints comes from the analysis of neighbouring viral genes with overlapping reading frames[80]. As expected, the overlapping nucleotide regions evolve at very low rates compared with those of non-overlapping regions of the genes[80].

Developmental timing. Two easily accessible biological measures are probably informative about the level of pleiotropy: the timing of gene expression in development and its tissue specificity (discussed below). Taxonomically diverse studies have revealed that morphological and molecular conservation of animal development follows an 'hourglass' model[81], where both early and late stages of development are variable, with an intermediate, conserved 'phylotypic' stage[81]. One potential explanation for this pattern is that pleiotropy is highest in the middle stages of development, whereas mutations in genes that function either in earlier stages (within blocks of undifferentiated cells) or in later stages (within differentiated tissues) do not influence the overall course of development.

Although the original observations were based on gene content and expression conservation across phyla, studies in *D. melanogaster*[82] and the mouse[83] indicate that the hourglass model also applies to protein sequence conservation. By contrast, there is no such trend in *C. elegans*[84]. It is difficult to discern whether this discrepancy simply reflects differences in the available data sets, or whether it is due to the peculiarly low flexibility of worm development.

Caenorhabditis elegans proteins that are expressed predominantly after reproductive maturity evolve more rapidly than larval proteins[2], a pattern that is also found in *D. melanogaster*[82]. This observation seems to be consistent with theories of senescence that propose that there is relaxed selection late in life[2].

Although there is evidence that the timing of expression affects the rate of protein evolution, this relationship could reflect not only differences in the strength of purifying selection, but also varying levels of positive

selection that are, for example, due to sexual selection or host–pathogen interactions in adults (see below). In support of this interpretation, the intensity of positive selection that is associated with the evolution of male gametes varies considerably through development and functions primarily on phenotypes that develop late in spermatogenesis[85].

Expression breadth and expression level. An obvious correlate of pleiotropy in multicellular organisms is breadth of expression: proteins that are expressed in many tissues have to operate under diverse cellular conditions and might interact with diverse proteins. As expected, broadly expressed proteins in mammals[30,86], insects[30] and plants[31] evolve more slowly than tissue-specific proteins. A similar relationship can be seen when exons of a given gene are considered: alternatively spliced exons (which are generally also tissue specific) evolve at higher rates than constitutively spliced exons[87].

However, expression breadth is strongly correlated with expression level, that is, the number of transcripts per cell[30]. Therefore, expression level rather than high pleiotropy might affect evolutionary rate. Studies on both multicellular[30,31] and unicellular[32,33] organisms have confirmed a negative correlation between expression level and rate of protein evolution. However, at least in mammals, it seems that expression breadth explains more of the variation than expression level: the negative correlation between divergence and expression rate disappears when controlling for expression breadth (M.J.L., unpublished observations). In the unicellular yeast (where, by definition, there is no variation in expression breadth), expression level explains 30–50% of the variation in the rate of protein evolution[32,39,48] (FIG. 3), much more than any other known variable.

What causes this strong correlation between evolutionary rate and expression level, at least in yeast? It is unlikely to be a secondary effect that is caused by correlations of both factors with one of the usual suspects, such as dispensability, pleiotropy that is due to protein interactions, functional classification, gene-conversion rate or mutation rate[32,39]. Moreover, it is unlikely that selection on mRNA secondary structure is responsible: such selection would affect synonymous sites more than non-synonymous sites, the opposite to what is seen[39]. One might argue that highly expressed genes should use energetically cheap amino acids to reduce the total cellular costs of amino-acid synthesis[88]. However, although the amino-acid composition in the proteomes of *E. coli* and *Bacillus subtilis* reflects the action of such selection pressure[88], this pressure does not seem to influence the rate of protein evolution[33].

Recently, it has been suggested that selection on the speed and accuracy of translation might influence not only codon usage bias, but also protein evolution. The latter would be expected if translation into certain amino acids were more efficient than into others[89]. However, translational accuracy is unlikely to explain the correlation between expression level and the rate of protein evolution, as this correlation remains after controlling for adaptive codon usage[39].

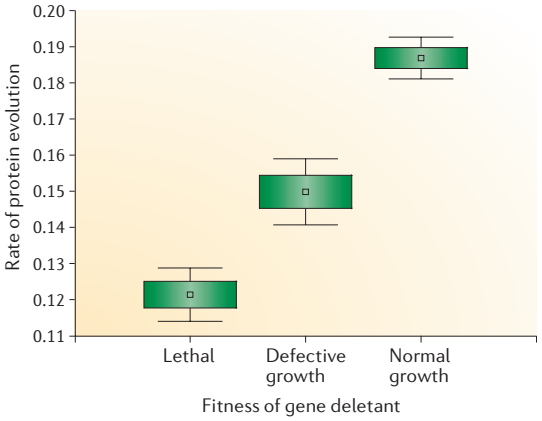

Figure 2 | **Gene dispensability and rate of protein evolution.** The rate of protein evolution is weakly associated with the severity of the fitness effect of gene deletions in yeast (ANOVA: $R^2 = 0.073$, $P < 10^{-9}$, $N = 2,979$, $df = 2$). However, most of this effect is due to the difference between essential (lethal) and non-essential (defective or normal growth) genes (ANOVA: $R^2 = 0.061$, $P < 10^{-9}$, $N = 2,979$, $df = 1$). Among the significantly slow growing strains (defective growth), relative growth rate (fitness) does not correlate with rate of evolution (Pearson: $R^2 = 0.001$, $P = 0.47$, $N = 478$). Gene dispensability data are from REF. 136, where the authors measured the growth rate of each gene deletant strain in a rich medium and identified genes with significantly slow-growing phenotypes (defective growth). Evolutionary rate (non-synonymous divergence) was calculated by Wall *et al.*[47] using sequences from four yeast species of the *Saccharomyces* genus. Boxes show mean ± standard error.

However, the reduction in protein evolutionary rate seems to be coupled to the production rate of proteins rather than their abundance[39]. This observation is consistent with a recent suggestion that robustness against mistranslations might be responsible for the observed effects[39]. Missense translation errors might affect nearly 20% of all produced proteins[39]. By increasing the probability of misfolding, they result in a higher risk of protein aggregation and toxicity[56]. This aggregation is unproblematic for proteins with low abundance, but might result in a substantial cellular burden for the most highly expressed proteins[39]. Therefore, selection might favour amino-acid sequences that reduce the risk of incorrect folding even in the presence of incorrectly translated amino acids. By directly comparing stability and aggregation risk in highly and lowly expressed genes, mutagenesis experiments could test this exciting hypothesis.

Overall, four factors that are associated with purifying selection seem to influence the rate of protein evolution: gene dispensability (albeit to a lesser extent than commonly assumed); protein structure and stability; pleiotropy (reflected in the role of diversity in protein interactions, biological processes and expression); and expression level. Expression level seems to be the strongest predictor of evolutionary rate, at least in unicellular species; selection for structural

Sexual selection
Competition among members of one sex for mating opportunities with the other sex.

Gene conversion
Non-reciprocal transfer between a pair of non-allelic or allelic DNA sequences during meiosis and mitosis, such that the receiving sequence becomes more similar to the donating sequence.

Codon usage bias
The non-random usage of synonymous codons for the same amino acid.

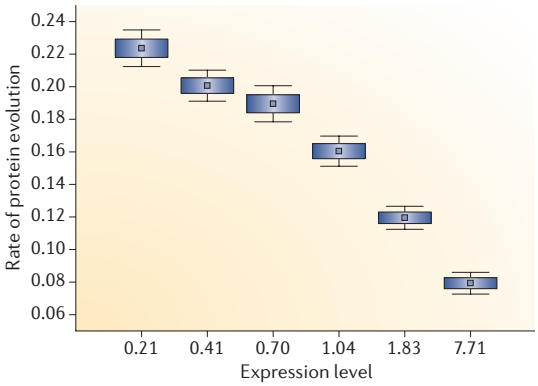

Figure 3 | **Gene-expression level and rate of protein evolution.** Gene-expression level (measured as mRNA abundance on a rich medium[137]) correlates strongly and negatively with the rate of protein evolution in yeast (R^2 = 0.29 for individual genes). Evolutionary rate (non-synonymous divergence) was calculated by Wall *et al.*[47] using sequences from four species of the *Saccharomyces* genus. The same number of genes was assigned to each bin. Boxes show mean ± standard error.

robustness against mistranslations has been put forward as a possible explanation[39], but further experiments are necessary before we can consider this issue resolved.

Positive selection

So far our discussion has concentrated on the strength of selection against deleterious mutations. However, recent surveys indicate that positive selection on advantageous protein changes can drive at least 20–45% of all amino-acid substitutions[28,90].

The proportion of amino-acid substitutions that are fixed by adaptive evolution seems to be remarkably constant across the genome: fast-evolving genes have higher numbers of both adaptive and neutral substitutions[28]. This correlation indicates that neutral evolution might have a constructive role during adaptation by enabling local exploration of sequence space[91]. By contrast, genes with high fitness density (due to high pleiotropy for example) might have relatively few unconstrained sites and are less likely to contribute to adaptation[66,67]. Factors that influence pleiotropy, which were discussed above in the context of purifying selection, might therefore be equally relevant to positive selection.

Although similar factors seem to influence the strength of purifying selection in all lineages (see above), the weight of positive selection is placed on different biological processes across lineages[84]. As outlined below, some general patterns still emerge.

Arms races. Positive selection is not necessarily a sign of adaptation at the organism level[67]: arms races might enforce change just to maintain the *status quo*. Arms races occur between competitors either within or across species. An obvious example are host–parasite interactions, which probably explain the fast evolution of human immunity genes[92].

Arms races within species might reflect intraspecies competition for limited resources. A striking example is the observation of continuing positive selection over thousands of phage generations in a non-selective setting[93]. In many sexually reproducing species, the ultimate limiting resource is access to reproduction. The ensuing competition leads to an arms race among members of the same sex, usually most pronounced in the male (where variation in reproductive success is higher). Sexual selection might be behind the fast evolution of male-biased genes in *D. melanogaster*[94] and humans[92].

Compensatory substitutions. A very different reason for positive selection without adaptation is the compensation of a deleterious mutation by a mutation at another, epistatically interacting, site in the genome; the latter mutation would then be positively selected, even if it was deleterious in the wild-type background[40,41]. The commonness of compensatory mutations is indicated by the observation that for each deleterious amino-acid substitution, there might be as many as 10–12 potential mutations at other sites that can compensate for the loss in fitness[42]. Unsurprisingly, organisms with high mutation rates (for example, viruses[95] and endosymbionts[96]) provide the clearest examples of compensatory evolution. But even for humans, a remarkably high fraction of pathogenic mutations are fixed in other closely related species, probably reflecting compensatory substitutions[40].

Because individual mutations can either decrease or increase the stability of protein structures depending on other sites, it has been suggested that compensatory mutations are particularly important in maintaining protein structure in evolution[41]. Consistent with this view, a recent study[97] found that after an amino-acid substitution has occurred, the probability that a second substitution occurs at a structurally interacting site is substantially increased; for ionic interactions it is amplified almost 5 times above chance expectations.

The importance of compensatory mutations for fixed deleterious amino-acid substitutions was recently studied experimentally in a DNA bacteriophage[95]. Fixation of a compensatory mutation was found to be twice as likely as reversion to the wild-type sequence. However, even in this simple organism, approximately half the compensatory mutations were observed outside the originally affected coding sequence, and therefore were not related to maintaining protein structure. As expected, the probability of compensatory substitutions increased with the severity of the deleterious fitness effect[42,95].

The compensatory mutations discussed above function locally and compensate for mutations at one or a few sites. Conversely, global compensatory mutations (also termed global mutation suppressors) are able to mitigate the effect of mutations at numerous other sites within the same protein by increasing its stability[39]. An even higher level of global suppression is exemplified by a chaperone protein in the endosymbiotic *Buchnera aphidicola*, which evolved to compensate for the accumulation of harmful mutations in protein-coding genes across the genome[96].

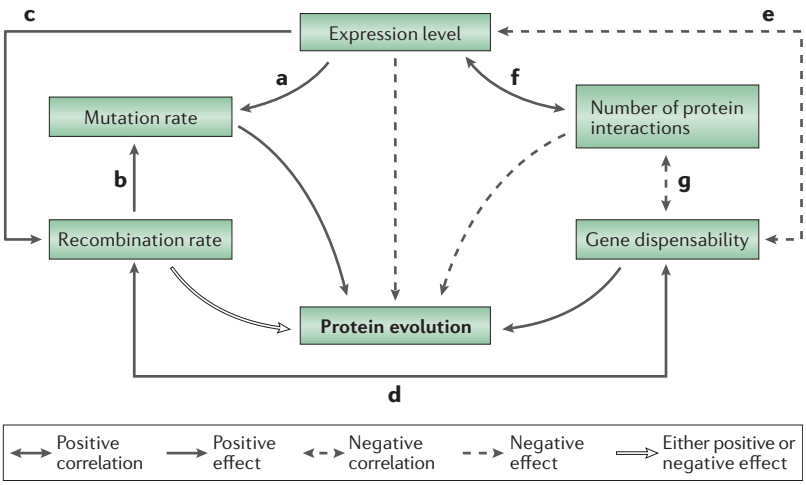

Figure 4 | **Interdependence between the factors that affect protein evolution.** **a** | Transcription causes increased spontaneous mutation rates in *Saccharomyces cerevisiae*[19] and *Escherichia coli*[138], probably by exposing the non-transcribed ssDNA to mutagenic chemicals. **b** | Recombinational repair of double-stranded breaks in *S. cerevisiae* increases the frequency of nearby point mutations[21]. **c** | Genes that are close to recombination hotspots in *S. cerevisiae* are expressed at higher levels during vegetative growth than most other genes[34]. **d** | Essential genes are clustered in regions of low recombination in *S. cerevisiae* and *Caenorhabditis elegans*[139]. **e** | Proteins that are more dispensable tend to be expressed at lower levels than less dispensable ones[46]. **f** | More protein–protein interactions have been reported for highly expressed proteins than for low-abundance proteins in *S. cerevisiae*[140]. However, this correlation is not supported by all interaction-detection methods[140], and might reflect a detection bias towards high-abundance proteins. **g** | It has been reported that essential genes have more protein–protein interactions than non-essential genes[141]. However, this correlation might be an artefact of biases in certain interaction data sets[142].

Adaptation. Despite the above emphasis on non-adaptive evolution, there are of course many cases in which positive selection reflects adjustment to new or changing environments. There are at least several such examples per ecological niche and type of species; we therefore restrict ourselves to mentioning some of the most exciting case studies on human evolution. In our lineage, recent positive selection has affected many proteins that are involved in sensory perception, as well as proteins that are involved in the determination of brain size and in language processing[98]. Genes that are expressed in the CNS of primates have generally evolved at accelerated rates compared with rodents[98]. Remarkably, selection on some cognitive traits has persisted in anatomically modern humans[99].

It has been suggested that adaptation often progresses through changes in protein expression rather than protein sequence. Consistent with this view, the most significant enrichment in positively selected proteins is found for those that are associated with transcriptional regulation[100].

Overall, and in contrast to the nearly neutral theory, positive selection seems to be an important force in protein evolution. Despite obvious examples of adaptation, many positively selected mutations occurred to uphold the *status quo*, endangered by substitutions either in the same genome (compensatory substitutions) or in competitors (arms races).

Conclusions and outlook

Systematic analyses of genomic data have demonstrated the influence of a range of factors on the evolution of proteins[101], encompassing positive selection (due to adaptation, arms races or compensatory interactions), purifying selection (due to selection on protein function, structure and folding), and regional genomic influences (due to variation in mutation and recombination rates). However, we are not yet in the position to conclusively judge the relative importance of all these factors. Research needs to move from discussing whether each factor has any significant effect to what proportion of the variation it can explain. This step might sound easier than it actually is. Most of the important factors are correlated with each other (FIG. 4), and are measured at different accuracies. Currently, we have relatively few integrated analyses. One of the rare examples shows that in both *E. coli* and *B. subtilis*, expression rate affects the rate of evolution more than functional category, essentiality or biosynthetic cost of amino acids[33]. In a similar vein, expression level was identified as the key variable in yeast among a large number of investigated factors, explaining 40-fold more variation in evolutionary rate than any other variable[48] (but see REF. 47). New techniques, some borrowed from artificial intelligence[102], trained neural nets[103] or structural equation models[47] will be needed to find the fundamental determinants of protein evolution.

It is also unclear to what extent the different factors influence the number of amino-acid sites that are under selection or the intensity of selection on particular sites. Systematic mutagenesis experiments have revealed that mutations affecting structural stability and other biophysical properties are distributed across the whole protein, unlike mutations that drastically affect function only[54,55]. In fact, these experiments show that most amino-acid replacements have biologically significant effects on protein stability[41], which indicates that structural constraints are among the strongest contributors to fitness density.

More fundamentally, it is even unclear to what extent the factors that contribute to evolutionary rate variation derive their effect from influencing the strength of purifying selection (protein 'importance') versus the probability of positive selection (protein 'adaptability')[104]. This topic requires further study through, for example, microbial selection experiments, or by combining comparative analyses with large-scale polymorphism data.

Further elucidation of the factors discussed in this review will be crucial to, and benefit from, the development of a new integrated theory of protein evolution. Although population genetics provides a solid base for such a theory[10–12], there have been only limited attempts so far to include the diverse factors discussed above in a single conceptual framework.

Several important hypotheses on protein evolution have so far remained largely untested. This is partly attributable to the dependence of protein evolution on the fine details of many unknown population

genetic parameters, such as population size, selective conditions, and rates of mutation and recombination. Long-term microbial selection experiments[105], combined with genomic analyses[106], are a promising tool to resolve this problem in the laboratory. It is relatively straightforward to manipulate these variables experimentally, and recent cost-efficient technologies facilitate the sequencing of complete genomes of ancestral and derived microbial strains[106].

Clearly our discussion on protein evolution, which focused on point mutations, is incomplete. There are at least three other principal genetic mechanisms that contribute to the evolution of new functions: duplication and functional divergence of genes, protein domain shuffling[107], and horizontal gene transfer across species. Many of the factors discussed in this paper will similarly influence the fixation of these events in populations. For example, the evolution of both gene duplicates[108] and horizontally transferred genes[109] is influenced by protein interactions.

The recent advances in our understanding of protein evolution can improve the predictive power of methods that rely on estimates of evolutionary rates. For example, when attempting to validate protein interaction data by comparing evolutionary rates[7], or when identifying potential drug targets (for example, essential proteins) in microbes, under the assumption that they are slowly evolving[6], it is paramount to control for gene expression as a confounding variable. The development of an integrated theory would take such corrections from being *ad hoc* and approximate to being a fundamental aid in understanding the processes under study.

1. Webster, A. J., Payne, R. J. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478 (2003).
2. Cutter, A. D. & Ward, S. Sexual and temporal dynamics of molecular evolution in *C. elegans* development. *Mol. Biol. Evol.* **22**, 178–188 (2005).
3. Bromham, L. & Leys, R. Sociality and the rate of molecular evolution. *Mol. Biol. Evol.* **22**, 1393–1402 (2005).
4. Brakmann, S. & Schwienhorst, A. (eds) *Evolutionary Methods in Biotechnology: Clever Tricks for Directed Evolution* (Wiley, Weinheim, 2004).
5. Smith, N. G. & Eyre-Walker, A. Human disease genes: patterns and predictions. *Gene* **318**, 169–175 (2003).
6. Searls, D. B. Pharmacophylogenomics: genes, evolution and drug targets. *Nature Rev. Drug Discov.* **2**, 613–623 (2003).
A summary of the potential links between evolutionary genomics and pharmacology.
7. Ramani, A. K. & Marcotte, E. M. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284 (2003).
8. Wilson, A. C., Carlson, S. S. & White, T. J. Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639 (1977).
A classical early study that recognized several potential determinants of protein evolution.
9. Fay, J. C. & Wu, C. I. The neutral theory in the genomic era. *Curr. Opin. Genet. Dev.* **11**, 642–646 (2001).
10. Kimura, M. *The Neutral Theory of Evolution* (Cambridge Univ. Press, Cambridge, 1983).
11. Ohta, T. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).
12. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
References 10–12 are landmark reviews (frequently with opposing views) on the neutral and nearly neutral theories.
13. Ellegren, H., Smith, N. G. C. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13**, 562–568 (2003).
14. Smith, N. G. C. & Hurst, L. D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395–1402 (1999).
15. Lercher, M. J., Williams, E. J. B. & Hurst, L. D. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**, 2032–2039 (2001).
An analysis of mutation-rate variation across mammalian genomes and its effect on protein evolution.
16. Lercher, M. J., Chamary, J. V. & Hurst, L. D. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* **14**, 1002–1013 (2004).
17. Williams, E. J. & Hurst, L. D. The proteins of linked genes evolve at similar rates. *Nature* **407**, 900–903 (2000).
18. Matassi, G., Sharp, P. M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791 (1999).
19. Datta, A. & Jinks-Robertson, S. Association of increased spontaneous mutation-rates with high levels of transcription in yeast. *Science* **268**, 1616–1619 (1995).
20. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**, 337–340 (2002).
21. Rattray, A. J. & Strathern, J. N. Error-prone DNA polymerases: when making a mistake is the only way to get ahead. *Annu. Rev. Genet.* **37**, 31–66 (2003).
22. Hurst, L. D. & Peck, J. R. Recent advances in understanding the evolution and maintenance of sex. *Trends Ecol. Evol.* **11**, 46–52 (1996).
23. Birky, C. W. Jr & Walsh, J. B. Effects of linkage on rates of molecular evolution. *Proc. Natl Acad. Sci. USA* **85**, 6414–6418 (1988).
24. Wyckoff, G. J., Malcom, C. M., Vallender, E. J. & Lahn, B. T. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet.* **21**, 381–385 (2005).
A remarkable study that suggests that up to 40% of the variation in protein evolutionary rates might be attributable to variation in the underlying mutation rate.
25. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Rev. Genet.* **7**, 98–108 (2006).
26. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
27. Betancourt, A. J. & Presgraves, D. C. Linkage limits the power of natural selection in *Drosophila*. *Proc. Natl Acad. Sci. USA* **99**, 13616–13620 (2002).
This paper claims that regional recombinational differences have a strong influence on the fixation of positively selected mutations.
28. Bierne, N. & Eyre-Walker, A. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* **21**, 1350–1360 (2004).
29. Presgraves, D. C. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr. Biol.* **15**, 1651–1656 (2005).
30. Subramanian, S. & Kumar, S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168**, 373–381 (2004).
31. Wright, S. I., Yau, C. B., Looseley, M. & Meyers, B. C. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol. Biol. Evol.* **21**, 1719–1726 (2004).
32. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
The first identification of protein-expression level as a strong predictor of evolutionary rate in yeast.
33. Rocha, E. P. C. & Danchin, A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21**, 108–116 (2004).
This work (like reference 48) compares the relative importance of several factors that are implicated in protein evolution, identifying expression level as the most important variable.
34. Gerton, J. L. *et al.* Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 11383–11390 (2000).
35. Pal, C., Papp, B. & Hurst, L. D. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol. Biol. Evol.* **18**, 2323–2326 (2001).
36. Bachtrog, D. Protein evolution and codon usage bias on the neo-sex chromosomes of *Drosophila miranda*. *Genetics* **165**, 1221–1232 (2003).
37. Bachtrog, D. Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*. *Nature Genet.* **36**, 518–522 (2004).
38. Zuckerkandl, E. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J. Mol. Evol.* **7**, 167–183 (1976).
39. 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).
Might highly expressed proteins be under strong selection to avoid protein misfolding? Several tests in this remarkable study indicate that this is the case.
40. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99**, 14878–13883 (2002).
An original study on the frequency and importance of compensatory substitutions.
41. 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).
An original and thought-provoking review that links protein stability and compensatory evolution.
42. Poon, A., Davis, B. H. & Chao, L. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. *Genetics* **170**, 1323–1332 (2005).
43. Hirsh, A. E. & Fraser, H. B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
A classical study on the effect of gene 'importance' on protein evolution.
44. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).
45. Cutter, A. D. *et al.* Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. *Genome Res.* **13**, 2651–2657 (2003).
46. Pal, C., Papp, B. & Hurst, L. D. Rate of evolution and gene dispensability. *Nature* **421**, 496–497 (2003).
47. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102**, 5483–5488 (2005).
A sophisticted analysis that aims to disentangle the influences of expression level and dispensability.

48. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.*, 327–337 (2005).
This work (like reference 33) compares the relative importance of several factors that are implicated in protein evolution, and identifies expression level as the most important variable.
49. Zhang, J. Z. & He, X. L. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol. Biol. Evol.* **22**, 1147–1155 (2005).
50. 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).
The 'importance' of a gene is highly environment-specific: about half of all 'dispensable' enzymes in the laboratory are essential in specific environments.
51. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**, 2229–2235 (2003).
52. Hurst, L. D. & Smith, N. G. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
53. Torgerson, D. G., Whitty, B. R. & Singh, R. S. Sex-specific functional specialization and the evolutionary rates of essential fertility genes. *J. Mol. Evol.* **61**, 650–658 (2005).
Shows that function-specific positive selection, rather than essentiality, seems to explain the evolution of fertility genes.
54. Pakula, A. A. & Sauer, R. T. Genetic analysis of protein stability and function. *Annu. Rev. Genet.* **23**, 289–310 (1989).
55. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**, 9205–9210 (2004).
56. Dobson, C. M. Principles of protein folding, misfolding and aggregation. *Semin. Cell Dev. Biol.* **15**, 3–16 (2004).
57. Haney, P. J. *et al.* Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc. Natl Acad. Sci. USA* **96**, 3578–3583 (1999).
58. Sterner, R. & Liebl, W. Thermophilic adaptation of proteins. *Crit. Rev. Biochem. Mol. Biol.* **36**, 39–106 (2001).
59. Dokholyan, N. V. & Shakhnovich, E. I. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312**, 289–307 (2001).
60. Parisi, G. & Echave, J. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. *Gene* **345**, 45–53 (2005).
61. Dean, A. M., Neuhauser, C., Grenier, E. & Golding, G. B. The pattern of amino acid replacements in α/β-barrels. *Mol. Biol. Evol.* **19**, 1846–1864 (2002).
62. Goldman, N., Thorne, J. L. & Jones, D. T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458 (1998).
63. Bustamante, C. D., Townsend, J. P. & Hartl, D. L. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol. Biol. Evol.* **17**, 301–308 (2000).
64. Koehl, P. & Levitt, M. Protein topology and stability define the space of allowed sequences. *Proc. Natl Acad. Sci. USA* **99**, 1280–1285 (2002).
65. Aris-Brosou, S. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol. Biol. Evol.* **22**, 200–209 (2005).
66. Fisher, R. *The Genetical Theory of Natural Selection* (Dover, New York, 1958).
67. Orr, H. A. The genetic theory of adaptation: a brief history. *Nature Rev. Genet.* **6**, 119–127 (2005).
An excellent review on molecular adaptation.
68. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
An influential, but controversial study on the effect of protein interactions on evolution.
69. Bloom, J. D. & Adami, C. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evol. Biol.* **3**, 21 (2003).
70. Hahn, M. W., Conant, G. C. & Wagner, A. Molecular evolution in large genetic networks: does connectivity equal constraint? *J. Mol. Evol.* **58**, 203–211 (2004).
71. Jordan, I. K., Wolf, Y. I. & Koonin, E. V. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1 (2003).
72. Agrafioti, I. *et al.* Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. *BMC Evol. Biol.* **5**, 23 (2005).
73. Teichmann, S. A. The constraints protein–protein interactions place on sequence divergence. *J. Mol. Biol.* **324**, 399–407 (2002).
74. Mintseris, J. & Weng, Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci. USA* **102**, 10930–10935 (2005).
75. Makino, T. & Gojobori, T. The evolutionary rate of a protein is influenced by features of the interacting partners. *Mol. Biol. Evol.* **23**, 784–789 (2006).
76. Fraser, H. B. Modularity and evolutionary constraint on proteins. *Nature Genet.* **37**, 351–352 (2005).
77. Jordan, I. K., Marino-Ramirez, L., Wolf, Y. I. & Koonin, E. V. Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* **21**, 2058–2070 (2004).
78. Evangelisti, A. M. & Wagner, A. Molecular evolution in the yeast transcriptional regulation network. *J. Exp. Zool. B* **302**, 392–411 (2004).
79. Salathe, M., Ackermann, M. & Bonhoeffer, S. The effect of multi-functionality on the rate of evolution in yeast. *Mol. Biol. Evol.* **23**, 721–722 (2006).
80. Mizokami, M. *et al.* Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.* **44** (Suppl. 1), 83–90 (1997).
81. Raff, R. *The Shape of Life* (Univ. Chicago Press, Chicago, 1996).
82. Davis, J. C., Brandman, O. & Petrov, D. A. Protein evolution in the context of *Drosophila* development. *J. Mol. Evol.* **60**, 774–785 (2005).
83. Hazkani-Covo, E., Wool, D. & Graur, D. In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J. Exp. Zool. B* **304**, 150–158 (2005).
In agreement with the 'hourglass' model of animal development, genes that are expressed during the phylotypic stage are under strong stabilizing selection.
84. Castillo-Davis, C. I., Kondrashov, F. A., Hartl, D. L. & Kulathinal, R. J. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res.* **14**, 802–811 (2004).
85. Good, J. M. & Nachman, M. W. Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. *Mol. Biol. Evol.* **22**, 1044–1052 (2005).
86. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68–74 (2000).
The first demonstration of faster evolution of tissue-specific proteins.
87. Xing, Y. & Lee, C. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl Acad. Sci. USA* **102**, 13526–13531 (2005).
Shows that exons that are used in minor isoform proteins evolve at higher rates than constitutive exons.
88. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **99**, 3695–3700 (2002).
89. Akashi, H. Translational selection and yeast proteome evolution. *Genetics* **164**, 1291–1303 (2003).
90. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
91. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett.* **579**, 1772–1778 (2005).
92. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
A comprehensive overview of the gene classes that were shaped by positive selection in human evolutionary history (see also reference 100).
93. Wichman, H. A., Millstein, J. & Bull, J. J. Adaptive molecular evolution for 13,000 phage generations: a possible arms race. *Genetics* **170**, 19–31 (2005).
This work indicates that intraspecies competition might lead to selection for perpetual change.
94. Zhang, Z., Hambuch, T. M. & Parsch, J. Molecular evolution of sex-biased genes in *Drosophila*. *Mol. Biol. Evol.* **21**, 2130–2139 (2004).
95. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage ϕX174. *Genetics* **170**, 989–999 (2005).
96. Fares, M. A., Moya, A. & Barrio, E. Adaptive evolution in GroEL from distantly related endosymbiotic bacteria of insects. *J. Evol. Biol.* **18**, 651–660 (2005).
This paper (along with others from the same group) indicates that a heat-shock protein might have evolved to mitigate the effect of deleterious substitutions in endosymbionts.
97. Shim Choi, S., Li, W. & Lahn, B. T. Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. *Nature Genet.* **37**, 1367–1371 (2005).
98. Fisher, S. E. & Marcus, G. F. The eloquent ape: genes, brains and the evolution of language. *Nature Rev. Genet.* **7**, 9–20 (2006).
99. Mekel-Bobrov, N. *et al.* Ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*. *Science* **309**, 1720–1722 (2005).
100. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
A comprehensive overview of the gene classes that were shaped by positive selection in human evolutionary history (see also reference 92).
101. Koonin, E. V. Systemic determinants of gene evolution and function. *Mol. Syst. Biol.* 13 Sep 2005 (doi:10.1038/msb4100029).
102. Chen, Y. & Xu, D. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* **21**, 575–581 (2005).
103. Kondrashov, F. A., Ogurtsov, A. Y. & Kondrashov, A. S. Bioinformatical assay of human gene morbidity. *Nucleic Acids Res.* **32**, 1731–1737 (2004).
104. Wolf, Y. I., Carmel, L. & Koonin, E. V. Unifying measures of gene function and evolution. *Proc. R. Soc. B* (in the press).
105. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Rev. Genet.* **4**, 457–469 (2003).
106. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
107. Patthy, L. *Protein Evolution* (Blackwell Science, Oxford, 1999).
108. Papp, B., Pal, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–197. (2003).
109. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806 (1999).
110. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, Oxford, 2000).
111. Whelan, S., Lio, P. & Goldman, N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**, 262–272 (2001).
112. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
113. Posada, D. & Crandall, K. A. *MODELTEST*: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818 (1998).
114. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
115. Miller, M. P. & Kumar, S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10**, 2319–2328 (2001).
116. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
117. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
118. Rebbeck, T. R., Spitz, M. & Wu, X. F. Assessing the function of genetic variants in candidate gene association studies. *Nature Rev. Genet.* **5**, 589–597 (2004).
119. Piganeau, G. & Eyre-Walker, A. Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. *Proc. Natl Acad. Sci. USA* **100**, 10335–10340 (2003).
120. Loewe, L., Charlesworth, B., Bartolome, C. & Noel, V. Estimating selection on non-synonymous mutations. *Genetics* **172**, 1079–1092 (2006).

121. Rokyta, D. R., Joyce, P., Caudle, S. B. & Wichman, H. A. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nature Genet.* **37**, 441–444 (2005).
References 119–121 attempt to estimate the fitness distribution of mutations; these values are highly relevant to understanding the relative influence of deleterious and advantageous mutations on protein evolution.
122. Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nature Genet.* **37**, 73–76 (2005).
123. Davis, J. C. & Petrov, D. A. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* **2**, 318–326 (2004).
124. Jordan, I. K., Wolf, Y. I. & Koonin, E. V. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* **4**, 22 (2004).
125. Cusack, B. P. & Wolfe, K. H. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol. Biol. Evol.* **22**, 2198–2208 (2005).
126. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, RESEARCH0008 (2002).
Shows that selection pressure is relaxed for a short period after gene duplication.
127. Kumar, S. Molecular clocks: four decades of evolution. *Nature Rev. Genet.* **6**, 654–662 (2005).
A comprehensive overview of the reasons for evolutionary rate variation across species.
128. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102**, 140–145 (2005).
129. Wernegreen, J. J. Genome evolution in bacterial endosymbionts of insects. *Nature Rev. Genet.* **3**, 850–861 (2002).
130. Gillespie, J. H. The role of population size in molecular evolution. *Theor. Popul. Biol.* **55**, 145–156 (1999).
131. Eyre-Walker, A., Keightley, P. D., Smith, N. G. & Gaffney, D. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* **19**, 2142–2149 (2002).
132. Paland, S. & Lynch, M. Transitions to asexuality result in excess amino acid substitutions. *Science* **311**, 990–992 (2006).
133. Bustamante, C. D. *et al.* The cost of inbreeding in *Arabidopsis*. *Nature* **416**, 531–534 (2002).
References 132 and 133 show the effect of sex and breeding system on the accumulation of deleterious substitutions.
134. Bastolla, U., Porto, M., Eduardo Roman, M. H. & Vendruscolo, M. H. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J. Mol. Evol.* **56**, 243–254 (2003).
135. Glaser, F. *et al.* ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**, 163–164 (2003).
136. Deutschbauer, A. M. *et al.* Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–1925 (2005).
137. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
138. Wright, B. E., Longacre, A. & Reimers, J. M. Hypermutation in derepressed operons of *Escherichia coli* K12. *Proc. Natl Acad. Sci. USA* **96**, 5089–5094 (1999).
139. Pal, C. & Hurst, L. D. Evidence for co-evolution of gene order and recombination rate. *Nature Genet.* **33**, 392–395 (2003).
140. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
141. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
142. Coulomb, S., Bauer, M., Bernard, D. & Marsolier-Kergoat, M. C. Gene essentiality and the topology of protein interaction networks. *Proc. Biol. Sci.* **272**, 1721–1725 (2005).

Acknowledgements
The authors wish to thank L. Hurst and L. Loewe for their insightful comments. C.P. and B.P. are supported by the Hungarian Scientific Research Fund (OTKA). C.P. is also supported by an EMBO (European Molecular Biology Organization) Long-term Fellowship. B.P. is a fellow of the Human Frontier Science Program. M.J.L. acknowledges financial support from the DFG (Deutsche Forschungsgemeinschaft).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Genome-wide Analysis papers: http://www.yeastgenome.org/cache/genome-wide-analysis.html
Joe Felsenstein's web page of Phylogeny Programs: http://evolution.genetics.washington.edu/phylip/software.html
Martin Lercher's web page: http://www.bath.ac.uk/bio-sci/lercher.htm
MEGA — Molecular Evolutionary Genetics Analysis: http://www.megasoftware.net
PolyPhen — Prediction of Functional Effect of Human nsSNPs: http://www.bork.embl-heidelberg.de/PolyPhen
Protein Data Bank: http://www.rcsb.org/pdb/Welcome.do 1F4R
PyMOL homepage: http://pymol.sourceforge.net
Sorting Intolerant From Tolerant (SIFT) database: http://blocks.fhcrc.org/sift/SIFT.html
The ConSurf server: http://consurf.tau.ac.il
UniProtKB: http://ca.expasy.org/sprot
Access to this links box is available online.