MODELLING

Causes of evolutionary rate variation among protein sites

Julian Echave[1], Stephanie J. Spielman[2] and Claus O. Wilke[2]

Abstract | It has long been recognized that certain sites within a protein, such as sites in the protein core or catalytic residues in enzymes, are evolutionarily more conserved than other sites. However, our understanding of rate variation among sites remains surprisingly limited. Recent progress to address this includes the development of a wide array of reliable methods to estimate site-specific substitution rates from sequence alignments. In addition, several molecular traits have been identified that correlate with site-specific mutation rates, and novel mechanistic biophysical models have been proposed to explain the observed correlations. Nonetheless, current models explain, at best, approximately 60% of the observed variance, highlighting the limitations of current methods and models and the need for new research directions.

Evolutionary rates
Number of substitutions (fixed mutations) per unit of evolutionary time.

Structural constraints
Structural features that correlate with sequence conservation (for example, solvent accessibility).

Functional constraints
Functional features that correlate with sequence conservation (for example, involvement in the active site).

[1]*Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, 1650 San Martín, Buenos Aires, Argentina.*
[2]*Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712, USA.*

Correspondence to J.E. and C.O.W.
jechave@unsam.edu.ar; wilke@austin.utexas.edu

doi:10.1038/nrg.2015.18
Published online 19 Jan 2016

Different protein-coding genes within the same species have widely varying rates of evolution. For example, genes that encode proteins that are highly expressed or that carry out critical functions tend to evolve more slowly than do genes that encode other proteins[1]. In addition to this gene-wide variation, and perhaps more interestingly, evolutionary rates vary among residues within a given protein. Although some of this variation is attributable to positive diversifying selection, for example, selection pressure triggering adaptation to environmental or other changes, substantial rate heterogeneity exists even at sites that are not subject to such selection pressure. This heterogeneity probably emerges from the differing functional and/or biophysical constraints that affect different protein sites. Accurately modelling this among-site heterogeneity is critically important in evolutionary studies, particularly in phylogenetic inference[2–8]. Phylogenetic models that allow for among-site rate heterogeneity universally provide better fits to data than models that assume constant rates across sites[3,9–13]. However, such models are generally phenomenological in nature and contain no information about the mechanistic source of among-site rate heterogeneity[14]. Although it is clear that substantial rate variation exists, the underlying mechanisms that generate the observed rate heterogeneity remain elusive.

Over the years, it has become apparent that site-specific evolutionary rates are influenced by a dynamic interplay between structural constraints and functional constraints (FIG. 1). In the 1960s, Perutz *et al.*[15] investigated site-specific sequence variability in globin proteins and found that internal sites were generally more conserved than superficial sites (that is, sites on the protein surface). They reasoned that special functions had to be influencing the sites that did not conform to this pattern[15]. Later, Kimura and Ohta built upon these observations by proposing the governing principle that "[f]unctionally less important molecules or parts of a molecule evolve (in terms of mutant substitutions) faster than more important ones" (REF. 16). Kimura and Ohta additionally recognized that surface protein residues "are usually not very critical to maintaining the function or tertiary structure, and the evolutionary rates in these parts are expected to be much higher" (REF. 16).

Following these early studies, most work on the sequence–structure–function relationship has been carried out from the perspective of structural biology. In general, such studies have not considered evolutionary rates, but have considered conservation only qualitatively or through conservation scores that do not take into account the nature of the evolutionary processes or the phylogenetic relationships. Therefore, our current understanding of how functional and structural constraints interact to shape evolutionary rate heterogeneity remains limited. To develop a complete picture of protein evolution, we need to identify the precise structural and functional properties that ultimately govern protein evolutionary rates and develop mechanistic explanatory models of these.

In recent years, there has been substantial progress on this front. Advances in computational evolutionary modelling have provided a variety of robust methods for estimating site-specific rates both from amino acid sequences and from protein-coding DNA sequences. Furthermore, numerous studies have discovered functional, structural and dynamic molecular features that correlate with rates[17–23], and biophysical models have been proposed that predict site-specific rates from

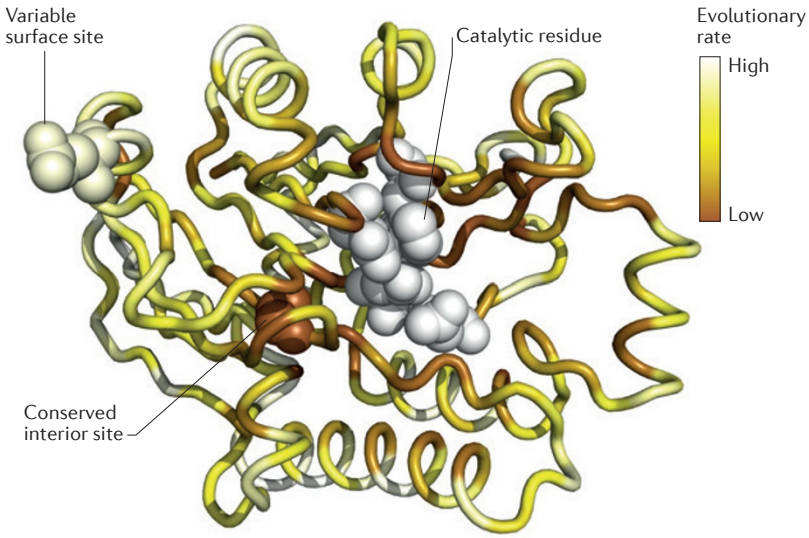


Figure 1 | **Structural and functional constraints shape site-specific evolutionary divergence.** Substitution rates for Exonuclease III of *Escherichia coli* are mapped onto its structure (Protein Data Bank (PDB) ID: 1AKO) using a divergence scale from darker low rates to lighter high rates. Owing to structural constraints, substitution rates are low in the protein interior and high on the protein surface. Residues close to the catalytic sites (shown in grey) also evolve slowly, indicating functional constraints. Evolutionary rate data are taken from REF. 75.

protein thermodynamics[24,25]. Studies on the relationship between rate and structure have yielded consistent findings, suggesting that the current data and methods provide a solid foundation upon which further knowledge can be built. Therefore, the time is ripe to review and synthesize our current understanding of site-specific evolutionary rates and the identified structural and functional aspects that influence them.

In this Review, we focus as much as possible on work that has studied site-specific rates estimated using state-of-the-art molecular evolution methods. To provide context, or in cases for which we lack rate-based studies, we also discuss relevant work based on other measures of sequence conservation. We first describe current methods to estimate rates from sequence data. We then consider molecular traits related to site-specific conservation. Next, we discuss two recent mechanistic biophysical models that have been used to explain site-specific rates. We conclude with a discussion of current challenges and future directions.

Estimation of site-specific rates

Computing site-specific evolutionary rates requires two pieces of data: a multiple sequence alignment, with either codon or amino acid data, and a corresponding phylogeny. Estimating the substitution rate at each individual site in a protein can be a computationally burdensome endeavour, much more so than estimating a mean rate for an entire protein sequence. Indeed, individual sites contain far less information than entire protein alignments, and thus large and diverse data sets are needed for reliable inference. In particular, most alignment sites need to have experienced several substitution events for a reasonably accurate rate estimate to be made[26].

Broadly speaking, site-specific inference methods follow one of two paradigms[27–31]: first, directly counting observed substitutions along a phylogeny; and second, using a Markov model of sequence evolution to infer evolutionary rate parameters, typically in a maximum-likelihood (ML) framework. Other methods, which compute sequence entropy or conservation scores[32–34], are useful for assessing the tolerance of a given residue to mutation. However, they cannot be substituted for true measures of evolutionary rate, as they do not typically account for phylogeny, which represents the evolutionary relationships among sequences. Indeed, it is possible for a given site to have high entropy and a low evolutionary rate or vice versa.

Inferring rates from codon data. In the context of protein-coding sequences, evolutionary rates are typically estimated with the ratio $\omega = dN/dS$, where dN is the evolutionary rate of non-synonymous substitutions and dS is the evolutionary rate of synonymous substitutions. To make dN and dS directly comparable, they are normalized to account for the approximately threefold higher likelihood that a random mutation is non-synonymous than synonymous[35]. The ratio ω has been primarily developed to detect sites under adaptive evolution (for which $\omega > 1$), but it can also be used to estimate site-specific rates[30,36].

Counting-based methods, the oldest class of dN/dS inference methods, calculate dN/dS simply by enumerating the observed changes either between pairs of sequences or along a phylogenetic tree[5,29,37–39]. Although they are relatively fast, these methods do not adequately account for multiple substitutions, variation in branch lengths or other biases, and therefore they tend to produce biased dN/dS estimates[5,29,35].

Conversely, most current inference approaches estimate rates in a ML framework with an explicit Markov model of sequence evolution. By implicitly accounting for any hidden substitutions along branches, ML-based methods are more robust and less biased than counting methods. Site-specific rates are obtained either by fitting a rate parameter individually to each site in the coding sequence (known as the fixed-effects likelihood (FEL) approach)[28,29,40] or by considering the rate to be a random variable drawn from a distribution governing the entire protein (known as the random-effects likelihood (REL) approach)[9,28,29,41]. In the REL approach, site-specific rates are calculated using a Bayes Empirical Bayes framework[42]. FEL lacks power for small data sets, and thus it is most appropriate for use on large data sets (that is, at least 200 sequences in which most sites have experienced recurrent changes)[29]. Conversely, REL is best suited for data sets of intermediate size (50–200 sequences), as its estimates on either very small or very large data sets are usually quite biased[29]. Importantly, although smaller data sets (for example, 16–50 sequences) may suffice when detecting episodic and/or diversifying selection[43], obtaining reliable site-specific dN/dS point estimates requires larger data sets. We note that FEL methods are primarily implemented in HyPhy[44] and its corresponding web-server DataMonkey[45], whereas REL methods are implemented in both HyPhy[44] and PAML[46].

Substitutions
Mutations that have spread to all members of the population (that is, have fixed), substituting the ancestral variant.

dN/dS
Ratio of non-synonymous to synonymous evolutionary rates.

dN
Non-synonymous evolutionary rate: that is, the rate at which non-synonymous substitutions (fixed mutations) occur per unit of evolutionary time.

Non-synonymous substitutions
DNA substitutions that change from a codon that codes for one amino acid to a codon that codes for a different amino acid.

dS
Synonymous evolutionary rate: that is, the rate at which synonymous substitutions (fixed mutations) occur per unit of evolutionary time.

Synonymous substitutions
DNA substitutions that change from a codon that codes for one amino acid to a codon that codes for the same amino acid.

There are two possible strategies for parameterizing *dN*/*dS* during rate inference. One can either fix *dS* across sites and simply estimate site-specific *dN* values, or one can estimate a separate *dN* and *dS* parameter for each site. Importantly, in the context of protein evolution, *dN* is the primary parameter of interest, and *dS* serves only as a normalization factor to determine the selection regime (for example, purifying selection, neutral or positive selection) in which a given residue falls. Given that site-specific rate estimates are inherently noisy, normalizing the inferred *dN* of each site with a corresponding site-specific *dS* is likely to introduce substantial, and potentially confounding, error. Therefore, models that either fix *dS* to 1 (REFS 9,41,47), or similarly infer gene-wide *dS* estimates for normalization may represent a more robust strategy for obtaining reliable rates of protein sequence evolution.

Although the inference approaches described above are generally implemented in a ML framework, several evolutionary rate inference approaches have recently emerged that use Bayesian, rather than frequentist, statistics. For example, a novel method that is known as renaissance counting[30], used in the BEAST software package[48], combines a counting-based approach with empirical Bayes regularization, thus leveraging the power of large data sets to produce site-specific estimates of accuracy that are comparable to FEL and REL. In addition, the inference method FUBAR (fast, unconstrained Bayesian approximation for inferring selection) adapts the REL framework to rapidly fit a large, pre-specified grid of evolutionary rates to the data in a hierarchical Bayesian framework[49]. This approach is exceptionally fast but yields reliable rate estimates for data with sufficient divergence. Finally, an approach for estimating gene-wide *dN*/*dS* values using a Bayesian statistical framework has recently been described, and future development may see this method extended to site-specific estimation[50].

Inferring rates from amino acid data. The primary approach for inferring rates from amino acid data is implemented in the Rate4Site program[51]. Rate4Site estimates a per-site rate-scaling factor that indicates how rapidly each residue evolves relative to the mean protein rate. It is implemented in both ML-based and Bayesian frameworks, with Bayesian frameworks being the default[52]. Under the Bayesian framework, Rate4Site uses a random-effects approach, specifying either a single gamma distribution[52], or a mixture of gamma distributions[11], as the prior rate distribution. A Bayes Empirical Bayes approach is then used to calculate site-specific rates. Importantly, Rate4Site can only accommodate data sets with fewer than approximately 300 sequences[51], and thus future research endeavours may seek to extend this method for use on larger data sets.

Alternatively, Fernandes and Atchley proposed a fixed-effects framework for estimating site-specific evolutionary rates from proteins[53]. Unlike the Bayes Empirical Bayes approach in Rate4Site, this method provides an independent rate estimate at each site, thus avoiding the confounding influences of mis-specified prior distributions on the rate. Finally, a relatively new method known as GP4Rate uses a Gaussian process to infer site-specific evolutionary rates while taking protein tertiary structure into account[54,55]. This approach effectively accounts for non-independence among site-specific rates.

Structural and environmental rate constraints

The among-sites rate variation observed in natural sequence alignments is, to a large extent, driven by the requirement that proteins fold properly and stably into their required, active conformation. In evolutionary terms, this requirement corresponds to purifying selection, such that sites at which mutations would disrupt folding or stability the most will be the most conserved. In addition, proteins experience selection pressure to avoid disrupting their native environment, for example, by forming nonspecific protein–protein interactions[56]. We refer below to these kinds of selective forces as structural and environmental constraints on sequence evolution.

Early studies of structural constraints established a basic paradigm dividing the protein into two general regions: the interior, which evolves slowly, and the surface, which evolves more rapidly[15,16] (FIG. 1). This paradigm poses two questions: first, why is the protein interior more conserved than the surface? And second, what are the salient structural differences between these two regions? Several biophysical measures have been proposed to explain the observed rate differences in structurally distinct regions. These measures include the solvent accessibility of a residue, and its packing density and flexibility. Although these measures are distinct, they all quantify the position of a given residue relative either to other nearby residues or to the protein as a whole.

Solvent accessibility. The most obvious difference between the surface and the interior is that the surface is accessible to the external environment, for example, water, but the interior is not. This observation defines solvent accessibility (accessible surface area (ASA) or solvent accessible surface area (SASA)), which indicates the surface area of a given residue that is accessible to water. ASA values are commonly normalized by the largest possible ASA for a given amino acid[57], resulting in the relative measure relative solvent accessibility (RSA). RSA ranges from 0 for completely buried residues to 1 for completely exposed residues.

Because early studies showed a relatively high conservation of protein cores, solvent accessibility formed the basis of research investigating the relationship between protein evolution and structure in the mid-1990s[58–63]. A broad consensus emerged that amino acid substitution rates and properties differ between buried and exposed sites, with buried sites being more conserved and tending more towards hydrophobic residues, owing to the local environment in the protein core. Over time, it was generally assumed that solvent accessibility represented the dominant structural constraint on evolutionary rate. For example, one study partitioned residues according

Purifying selection
Loss of mutations that decrease fitness (deleterious).

Positive selection
Fixation of mutations that increase fitness (adaptive).

Rate4Site
Popular software to estimate relative site-specific rates from amino acid sequence data.

Accessible surface area
(ASA). Same as solvent accessible surface area.

Solvent accessible surface area
(SASA). Surface area of a given residue that is accessible to water.

Relative solvent accessibility
(RSA). Measures the proportion of the surface of an amino acid that is accessible to solvent (that is, water) in the folded protein structure, from 0 (completely inaccessible) to 1 (completely accessible). Calculated as the ratio of the solvent accessible surface area (SASA) of a given residue in the protein structure and the maximum SASA of that residue in a fully solvent-accessible conformation.

to RSA and secondary structure (for example, helix, sheet, coil, turn, and so on) and found that exposed sites evolved more rapidly than buried ones, regardless of secondary structure[64]. A study of site-specific substitution rates for 25-α/β barrel enzymes found that a minimal model that considered RSA as the main factor could not be improved upon by adding other properties such as secondary structure or H-bonding information[17]. A third study showed that amino acid properties, such as hydrophobicity or size, had little influence on site rates beyond the strong effect of solvent exposure[65].

More recently, studies have leveraged the power of genomic data sets and sophisticated rate inference methods to carry out more comprehensive analyses. For example, Franzosa and Xia[18] examined the correlation of site-specific *dN*/*dS* with several structural properties across nearly 1,000 *Saccharomyces cerevisiae* proteins. They found that evolutionary rate increased linearly with RSA. Since then, several additional studies have reinforced the strong, positive relationship between RSA and site-specific rate[66–68].

Packing density. As we have discussed above, solvent accessibility has become the *de facto* structural measurement to use in protein evolution studies. However, more recent work has called the central role of solvent accessibility into question, suggesting instead that different structural measures more strongly correlate with evolutionary rate.

In particular, instead of quantifying the extent to which a given residue comes into contact with solvent, as RSA and ASA do, we can quantify the extent to which a residue comes into contact with other residues in the protein. This alternative concept, known as packing (or contact) density, indicates how densely packed a residue is within the protein tertiary structure. The two packing measures most commonly used in evolutionary studies are the contact number (CN) and the weighted contact number (WCN). For a given amino acid, CN simply counts the number of other residues within a local, structural neighbourhood. By contrast, WCN considers all residues in the protein and weighs them by the square of their inverse distance to the focal amino acid[19,69].

Packing density was initially introduced into the protein evolutionary rate literature because theoretical calculations predicted that more densely packed proteins would be easier to design[70] and, as a consequence, would evolve more rapidly[71]. The first studies relating contact density to evolutionary rate focused on whole-protein rates rather than site-specific rates, and broadly found that average protein rates are higher for proteins with residues that have, on average, higher packing densities[71–73]. Once packing density was established as a veritable predictor of evolutionary rate, several groups began investigating the packing–rate relationship on a site-specific basis. Franzosa and Xia found a modest but significant partial correlation of CN with site-specific rates while controlling for RSA, prompting them to conclude that CN influences rate independently of RSA[18]: more densely packed sites evolve more slowly. Importantly, Franzosa and Xia concluded that, although CN does predict evolutionary rate, RSA is a much stronger predictor[18]. However, subsequent studies have challenged this finding, pointing out that using WCN instead of CN to estimate packing density results in stronger predictive power, and moreover that the independent contribution of RSA becomes relatively small when WCN is controlled for[19,23,74].

In all the above-mentioned studies, WCN was calculated using C_α carbons to represent residues. Recent work has proposed that using side chains rather than C_α carbons may provide a more robust determinant[75]. Indeed, side-chain-based WCN has consistently outperformed both C_α-based WCN and RSA as a rate determinant (FIG. 2). Therefore, it now seems that WCN is the main determinant of site-specific substitution rates, and RSA provides a comparatively minor independent contribution.

Even so, as can be seen in FIG. 2, both the relative performance of WCN and RSA and the overall performance of either predictor vary widely among protein structures. Whereas for some proteins we can explain more than 60% of the observed rate variation with simple structural measures such as RSA and WCN, for other proteins we can explain less than 10% using the same measures. Similarly, for some structures WCN outperforms RSA by over 15 percentage points, but for others the two measures perform comparably, and in certain cases RSA can even outperform WCN. The underlying causes of these discrepancies are not well understood, but there are several, not mutually exclusive possibilities: first, other predictors, such as those that are related to protein function (see below), may be more important in some structures than in others. For example, consider the case of structure 1AKO shown in FIG. 2b. Second, in some cases alignments may be poor or may contain insufficient or excessive divergence. In general, alignments need to be sufficiently diverged for accurate rate inference at individual sites[26] but must not be saturated with mutations. Finally, standing polymorphisms, slightly deleterious mutations or mutations hitchhiking on recent selective sweeps may cause biased rate estimates[76], and as a consequence structural predictors may not work well on alignments in which any of these factors are highly prevalent.

Flexibility. Proteins are not static structures; they are dynamic polymers that undergo constant conformational fluctuations. Such movements are frequently crucial for protein function. For example, enzymes must shift their structural conformation to expose the active site before a substrate can be accommodated. Similarly, conformational changes could control the mutational tolerance of a site, such that a site in a highly flexible region of a protein structure would probably be more tolerant to mutations than a site in a less flexible region[20–22,77].

At the site level, conformational dynamics can be quantified using measures of local flexibility, such as mean square fluctuations (MSFs) or B-factors. These quantities measure the extent to which a given residue changes its position over time. Using these and similar quantities, several studies have found that site-specific sequence variation correlates with local flexibility[20,22,24,77], such that flexible sites evolve more rapidly than rigid sites.

Contact number
(CN). Number of neighbouring residues present in a protein structure within a given distance (for example, 10 Å) from a focal residue.

Weighted contact number
(WCN). Similar to the contact number, but the neighbouring residues are weighted by their inverse square distance to the focal residue, and all residues in a structure are considered to be neighbouring residues.

Mean square fluctuations
(MSFs). Time-average of the square norm of the vector that connects the instantaneous coordinates of a site to its equilibrium coordinates; measures the amount of movement a residue undergoes over time.

B-factors
(Also known as temperature factors). Quantity that measures the amount of thermal motion of an atom in a protein crystal structure.

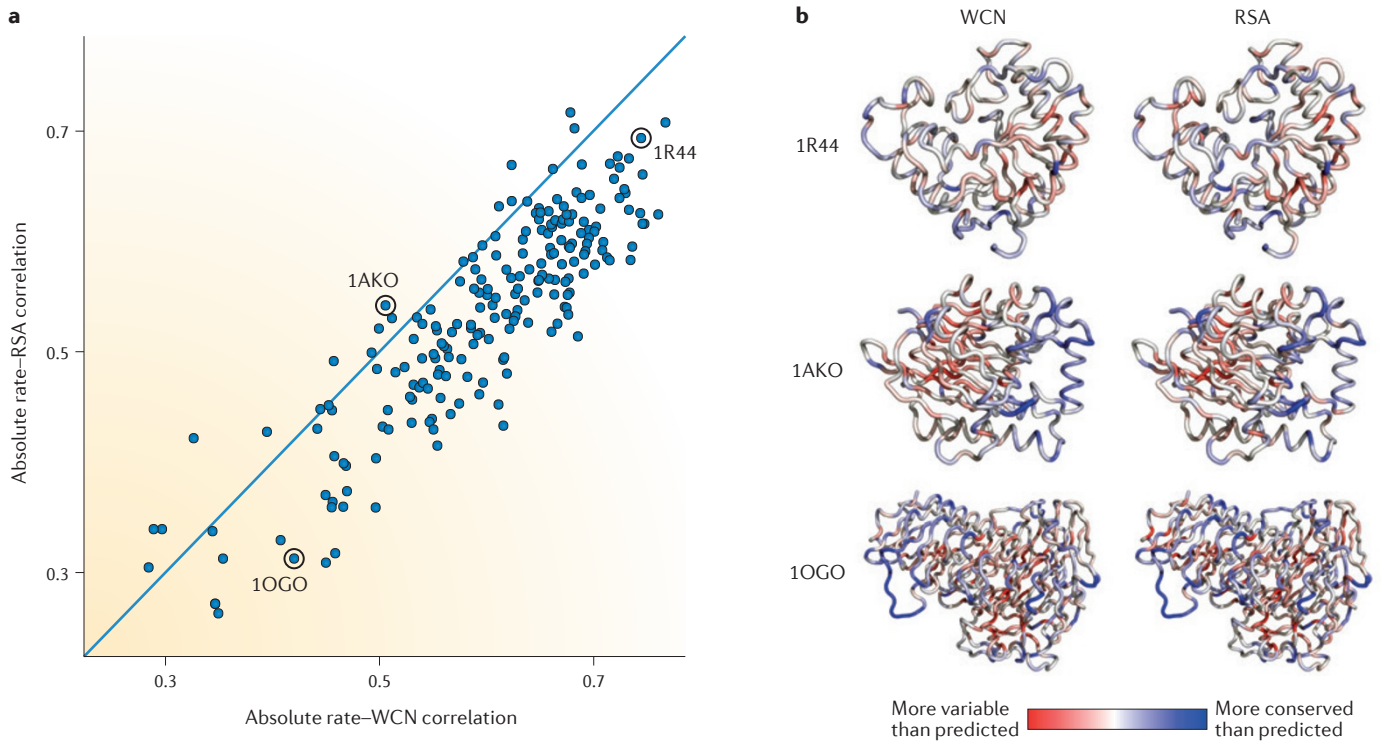


Figure 2 | **WCN correlates more strongly than RSA with site-specific rate. a** | Absolute rate–relative solvent accessibility (RSA) correlation versus absolute rate–weighted contact number (WCN) correlation for 209 enzyme structures is shown[75]. The solid line represents the $x = y$ line. The rate–WCN correlations are systematically stronger than the rate–RSA correlations. However, for some proteins RSA performs better than WCN, as can be seen for the highlighted structure 1AKO. **b** | Observed versus predicted rate, mapped onto the backbone of three structures is shown. Rate predictions were obtained from either WCN or RSA. The structures were chosen to represent low (1OGO), moderate (1AKO) and strong (1R44) structure–rate correlations, as highlighted in part **a**. Colours represent the differences between the observed and predicted rates at each site, with white representing a perfectly accurate prediction. As can be seen for structures 1AKO and 1OGO, poor predictions often coincide with surface loops that are more conserved than predicted from structure alone. These surface loops probably experience additional purifying selection due to function; compare, for example, the location of the conserved surface loops in 1AKO with the location of the active site of the protein, as shown in FIG. 1.

The fact that flexibility and evolutionary divergence correlate has been interpreted as evidence that protein dynamics imposes substantial constraints on sequence evolution[20,77]. However, whether flexibility is the underlying causal factor in the observed flexibility–rate correlation is unclear. Local flexibility directly relates to packing density[78], which, as discussed above, strongly correlates with rate. Therefore, it is possible that either flexibility or packing density represents the underlying causal factor that affects rate.

One of the first studies of the flexibility–evolution relationship used the inverse of the contact number as a proxy of flexibility, making the implicit assumption that flexibility was the causal factor[79]. Several later studies have made similar arguments[20,77]. However, if flexibility were the actual determinant and packing only a proxy, site-specific rates would have higher correlations with flexibility than with packing, which is not the case. Instead, the reverse is true: site-specific substitution rates correlate more strongly with measures of packing density (such as WCN) than with measures of flexibility (such as MSF)[22,24]. Moreover, when packing density is controlled for, no residual correlation remains between rate and MSF[24]. Therefore, it seems that flexibility correlates with rate simply because both quantities are determined by local packing density, and not because of a direct, causal relationship between flexibility and rate.

Other structural constraints: folding kinetics, protein expression and cellular environment. There are other quantities and constraints that broadly relate to the requirement for proper and stable protein expression. For example, structural factors that may constrain evolution at the site level include secondary structure, side-chain hydrogen bonds, unusual side-chain rotamers, nonplanar peptide bonds, strained main-chain conformations and buried hydrophilic-charged residues[80]. However, most of these factors have little explanatory power for site-specific rates once solvent accessibility is controlled for[17,64,81]. One factor that does matter is structural disorder: sites in disordered regions tend to evolve more rapidly, and with fewer conservative amino acid substitutions, than ordered regions[82–84].

In addition to being stable, proteins need to fold sufficiently rapidly. It is reasonable to expect this requirement for rapid folding to further constrain evolutionary divergence. Folding occurs via a transition state involving a small number of sites that assume their native

conformation: a folding nucleus[85]. A classic study of the cytochrome *c* family speculated that sites that were very conserved but not involved in activity could be the sites that form the folding nucleus[86]. In another study, the experimentally determined folding nucleus sites of nine proteins were found to be more conserved than average[87]. By contrast, a more exhaustive and systematic study found no significant evidence for the extra conservation of folding nucleus residues and, moreover, argued that the previously discovered special conservation was due to biases of the experimental data[88]. Note that the cited studies quantified conservation using entropy-based measures. A more recent study that was based on substitution rates found no significant differential conservation of folding nucleus sites[89].

The level at which a protein is expressed and the cellular location where it functions also influence site variation. For example, it is well known that more highly expressed proteins evolve more slowly[90]. Analyses at the site level have shown that this evolutionary constraint is RSA dependent[67,68]. In yeast, the difference in mean evolutionary rates for lowly and highly expressed genes increases linearly with RSA. For the most buried residues (RSA = 0), sites in highly expressed genes evolve approximately twofold slower than sites in lowly expressed genes, and for the most solvent-exposed genes (RSA = 1), the relative ratio in mean rates grows to above three[67]. An example of environment dependence can be seen in membrane proteins. In these proteins, the transmembrane regions are more evolutionarily conserved than the extramembrane regions, and this effect seems to be separate to conservation due to solvent accessibility[91,92].

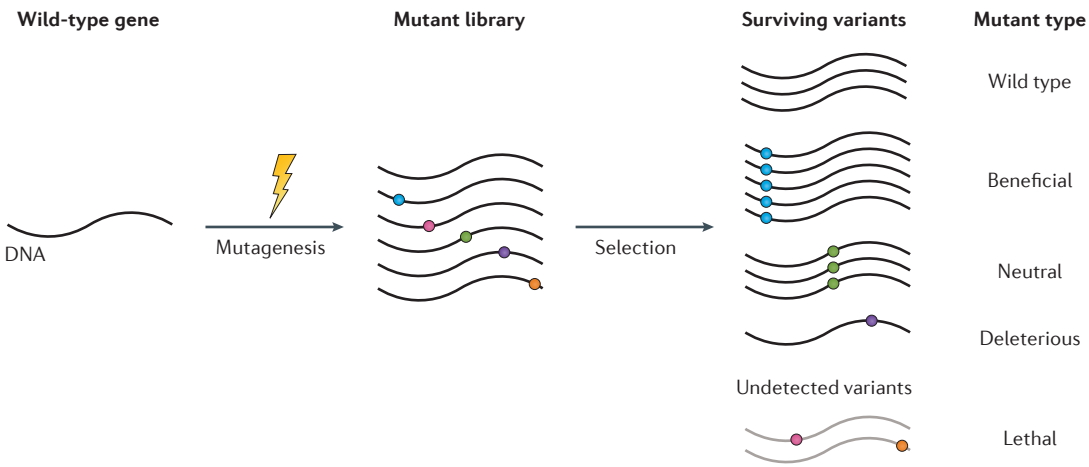
Rate variation caused by protein function

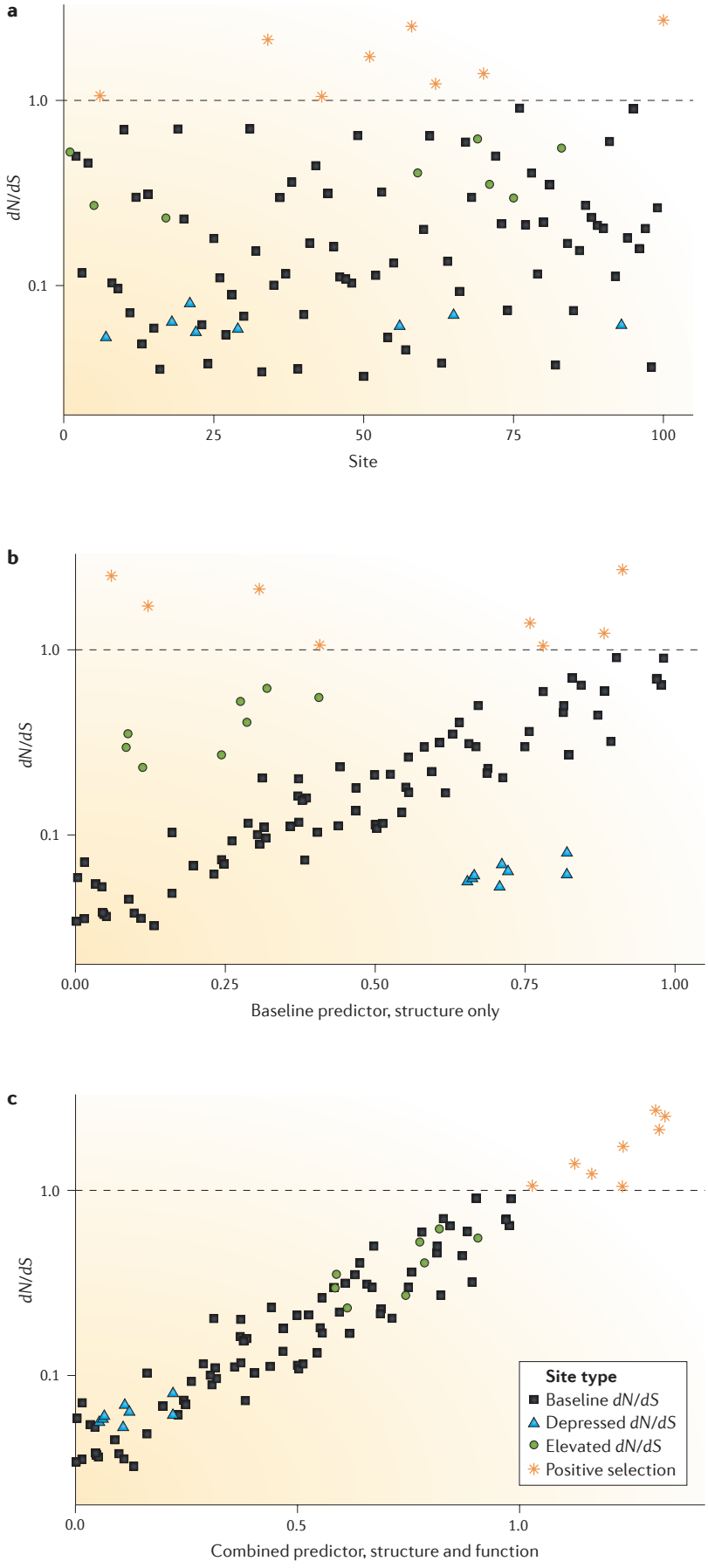
None of the structural and environmental constraints discussed above is directly related to protein function. Indeed, although proteins need to fold stably into their active conformation, this requirement alone does not guarantee that they will function properly, except perhaps for proteins the sole purpose of which is structural, as building blocks of organs and tissues. All other proteins, including enzymes, transcription factors, molecular motors and antibodies, have specific functional or active sites at which they experience additional selection

Box 1 | Experimental approaches to measure site-specific variation

Most of the work quantifying site-specific rate variation has been conducted computationally; however, a growing body of literature has emerged that takes an experimental approach. Specifically, several experimental lines of research have sought to determine site-specific amino acid preferences and/or tolerance to mutations. These quantities are intimately tied to evolutionary rate: sites that are more tolerant to mutation, or sites at which more amino acids are selectively tolerated, will generally evolve more rapidly. Conversely, sites with low mutational tolerance will evolve more slowly[66,119]. Results from experimental work sampling all possible mutations across all residues in a given protein have supported these theoretical predictions. For example, McLaughlin Jr *et al.*[95] have shown that functionally important residues are generally less mutationally tolerant than residues with less stringent functional constraints. Leferink *et al.*[141] demonstrated that mutations that increase solvent accessibility at an active site have a strong influence on the catalytic efficiency of an enzyme, demonstrating a tight relationship between evolutionary rate and function.

One new and powerful approach to addressing site-specific properties in proteins is deep mutational scanning, an experimental approach that samples as many as 1 million protein variants at a time[142–144]. Under deep mutational scanning, many different variants of a given gene are subjected to selection in a high-throughput procedure. By measuring the relative enrichment or depletion of variants after selection, this procedure allows for the precise quantification of the relative fitness of the gene variants (see the figure). Deep mutational scanning studies conducted on proteins from bacteria[145,146] and viruses[147,148] have revealed extensive heterogeneity in mutational tolerance within a given protein, and that tolerance generally correlates with solvent accessibility. Moreover, deep mutational scanning on TEM lactamase has shown that residues near active sites can sustain very few substitutions[146]. Finally, Bloom linked experimentally measured mutational tolerances to evolutionary models with site heterogeneity and showed that these experimentally informed models better account for observed variation in natural sequences than standard phylogenetic models[145,147].





pressures. These pressures may act in the form of purifying selection, causing increased evolutionary conservation, or in the form of positive, diversifying or balancing selection, causing increased evolutionary variability.

Purifying selection. In many cases, selection for function adds additional evolutionary constraints to the specific amino acids (such as catalytic sites) involved in the function of a protein[93,94]. These sites — as well as their neighbours — are often particularly conserved[93] (for example, consider the pattern of conservation near the active site in FIG. 1). Curiously, selection for function seems to extend beyond the active site and its immediate neighbours. Dean *et al.* found that distance to the active site correlated with site-specific variation in several enzymes[17]. Similarly, several experimental studies have observed that mutations far away from the active site can disable protein function by inducing protein-wide structural alterations (see also BOX 1). These findings provide evidence for the presence of long-range, indirect interactions in protein structures, probably mediated by steric interactions among neighbouring amino acids[95].

Besides catalytically active sites, residues that are involved in protein–protein or protein–nucleic acid interactions also experience added functional constraints and are generally more conserved than other surface sites[18,96–98]. The extent to which protein–protein interactions constrain site evolution seems to depend on the exact nature of the interaction. For example, obligate interactions, which often persist for the lifetime of the protein, are associated with lower rates compared with transient interactions that occur only occasionally[96]. As residues involved in protein–protein interactions experience reduced solvent accessibility when the interacting protein partner is present, Franzosa and Xia asked whether this reduction could explain the added evolutionary constraint on interface residues[18]: their answer was "not entirely". They found that although the evolutionary constraint increases linearly with increasing amount of solvent-accessible

Figure 3 | **Predictors of evolutionary variation can help to identify important sites in a protein. a** | When plotted against the linear position (site) in the protein, site-specific evolutionary rates appear to be random. We can identify sites that are under positive selection ($dN/dS > 1$; indicated in yellow) but we cannot easily identify other important sites (indicated in green and blue). **b** | If we can identify a baseline predictor that captures the effect of protein structure on site-specific evolutionary rate, then sites that deviate from this baseline expectation clearly stand out (sites that evolve more rapidly than expected are shown in green; sites that are more conserved than expected are shown in blue). Such sites are likely to be functionally important. **c** | If we can develop a predictor that can capture both the effects of structure and the effects of functional importance on evolutionary rate, then previously outlying sites now appear to follow the overall trend. This result indicates that we have identified the proper underlying reasons for why blue sites evolve slower and green or yellow sites evolve faster than expected under the model shown in part **b**.

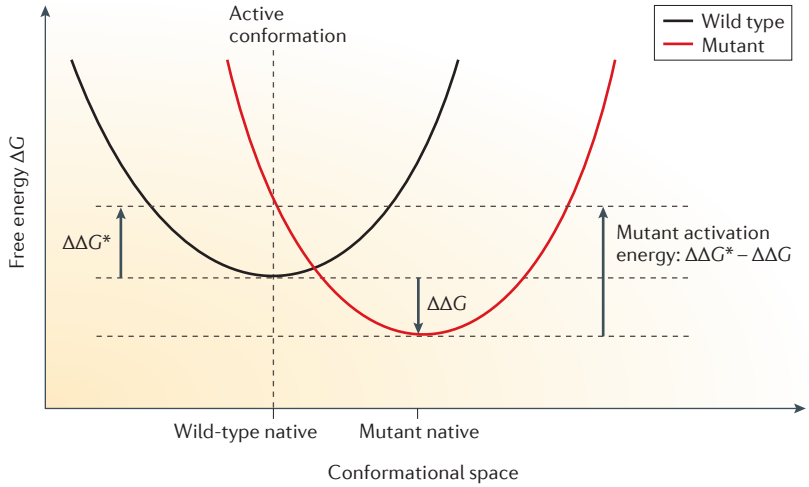


Figure 4 | **The trade-off between native stability and active stability.** A mutation shifts the free energy landscape from that of the wild-type protein (black curve) to that of the mutant protein (red curve). The mutant has a different equilibrium conformation and its stability differs by an amount of $\Delta\Delta G$ from that of the wild type (difference between red and black minima). The stability of the active conformation changes by an amount of $\Delta\Delta G^*$ due to the mutation (difference between the intersections of the vertical active conformation line with the red and black curves). To function, the mutant protein must deform from its equilibrium conformation to the active conformation, which requires an activation energy $\Delta\Delta G^{\ddagger} = \Delta\Delta G^* - \Delta\Delta G$. We assume here that the wild-type native structure is the active conformation. In this scenario, a mutation may stabilize the native state ($\Delta\Delta G < 0$) but destabilize the active state ($\Delta\Delta G^* > 0$). Thus, even a stabilizing mutation can create an energy barrier that may reduce or eliminate proper protein function.

surface area lost due to the interaction, there is also an additional, albeit low, fixed cost that can be attributed to the mere fact that a residue is participating in the interface interaction[18]. This fixed cost is independent of RSA. Its low magnitude is consistent with recent experiments and computer simulations showing that protein–protein interfaces can maintain function despite extensive divergence of one partner[99].

Finally, ligand-binding sites tend to be more conserved than other sites, and this conservation is exploited in ligand-binding site prediction methods[100,101]. However, the degree of conservation varies. For example, although catalytic sites are highly conserved, allosteric sites — although more conserved than average — vary more than catalytic sites, either because of weaker constraints or because of positive selection on the regulatory mechanisms[102].

Positive selection. When organisms are faced with novel or changing environments, their protein-coding genes may experience positive selection, that is, a selection pressure to change rather than to remain the same. This selection pressure will primarily act on the sites that are directly involved in the specific function under selection. That individual sites can experience positive selection for specific function was first recognized more than 30 years ago. Early examples include positive diversifying selection in the active sites of three related rodent protease inhibitors[103], over-dominance in the antigen recognition site of human and mouse class I MHC (major histocompatibility complex) genes[104], and positive selection in the V3 region of the HIV-1 envelope gene[41], probably reflecting immune escape or adaptation to cell tropism.

Positive selection is probably the most prevalent in viral surface proteins, which experience intense selection pressure to adapt to their host or to escape their host's immune response. For example, positive selection strongly shapes the evolution of the influenza haemagglutinin protein[105–108], which initiates the fusion of the viral envelope with the cellular membrane. Indeed, in this protein, positive selection explains nearly as much rate variation as RSA[109].

As positively selected sites can provide meaningful insights into the functioning of and selective constraints on a gene, the molecular evolution community is broadly interested in identifying such sites under many different scenarios. This interest has spurred the development of numerous tests for positive selection, based on the approaches and inference frameworks discussed above. Importantly, these methods rarely consider the baseline structural constraints that act on most sites in a protein. Consequently, they tend to be overly conservative and are likely to miss many important sites[36].

For example, the widely used test for $dN/dS > 1$ makes the implicit assumption that sites evolve at $dN/dS = 1$ in the absence of selection. However, for protein-coding sequences, selection is almost never absent; structural constraints will induce purifying selection that will push dN/dS to values much lower than 1 in nearly all cases. Thus, we can expect that positive selection at a site that would otherwise have been highly conserved may yield elevated dN/dS values that nevertheless remain below 1 (FIG. 3). Purely statistical methods that consider only the value of dN/dS at individual sites will not be able to identify such sites (FIG. 3a). However, methods that incorporate appropriate baseline expectations derived from protein structure may be able to so identify these sites (FIG. 3b). In general, functional constraints may be the reason why sites evolve faster or slower than expected from structural constraints (FIGS 2b,3b). Finally, methods that incorporate both structural and functional information may accurately predict the rate of evolution at functionally selected sites, thus providing mechanistic insights into why a given site evolves at the rate that it does (FIG. 3c).

Predicting rates from first principles

Correlations between rates and predictor variables allow us to identify factors that influence rate variation, but they do not ultimately provide explicit mechanistic insights into why a given site is variable or conserved. To gain mechanistic insight, we need to develop biophysical models that are grounded in first principles. Several biophysical models have been used to study issues such as marginal protein stability and site–site co-evolution[110,111]. We focus below on the models that have been used to study site-specific rates.

Although phenomenological models depend directly on predictor variables such as RSA and WCN, biophysical models are essentially based on protein stability. This choice is reasonable because, from a physicochemical

Table 1 | **Quantities observed to influence site-specific rate***

Quantity and physical effect	Effect on rate	Refs
Structural constraints		
Contact number (CN and WCN)	Decreases with increasing CN/WCN	18,19,22–24, 74,75,79
Relative solvent accessibility	Increases with increasing RSA	17,18,23,24,64, 66–68,74,75,81
Structural flexibility	Increases with increasing flexibility	20–22,24
Structural disorder	Increased in disordered regions	82–84
Functional constraints		
Protein–protein interfaces	Depressed in interface regions	18,96,97
Protein–nucleic acid interfaces	Depressed in interface regions	98
Catalytic sites	Depressed at and near catalytic sites	17
Environmental constraints		
Gene expression level	Decreases with increasing expression level, in particular at surface sites	67,68

CN, contact number; RSA, relative solvent accessibility; WCN, weighted contact number. *Rates in the cited studies have been estimated either from codon or from amino acid data.

perspective, the function of a protein is determined by its thermodynamics and kinetics, which are related to stability. In addition, stability is related to all the molecular features that correlate with evolutionary rates. For example, solvent accessibility (RSA) is related to stability via the energetic cost of burying a side chain into the core[112,113]. Similarly, packing density (CN and WCN) and flexibility (MSF) are related to the mean interaction energy of a site with the rest of the protein[78]. Therefore, molecular features such as RSA, WCN and MSF could be mere proxies of stability, which would be the true determinant of protein fitness and, therefore, site-specific evolutionary rates.

Two distinct biophysical models have been proposed in the literature. The stability-threshold model, referred to here as the native-stability model, links site-specific substitution rates to mutational changes in protein thermodynamic stability[114]. Specifically, it assumes that all proteins with sufficient stability in the native state function equally well and have identical fitness, and proteins that are not sufficiently stable have zero fitness. Therefore, this model imposes a stability-threshold condition that the protein needs to meet at all times during its evolution. We note that variations of this model may use a sigmoidal function instead of a hard threshold, but that these variations show similar behaviour and make the similar assumption that native stability is the crucial factor in the function of a protein[115,116].

In the threshold model, the probability of fixation of a given mutation is equivalent to the probability that the mutation will push the stability below the threshold. This probability can be calculated under the assumption that the free-energy changes $\Delta\Delta G_{ij,k}$ for mutations from amino acid i to amino acid j at site k are known[25] ($\Delta\Delta G$ measures the change in free energy ΔG between two protein variants, and ΔG is the measure of the stability of the protein fold). These free-energy changes can be estimated from atomic force fields such as FoldX[117] and can be subsequently converted into rate estimates[25].

In contrast to the native-stability model, the active-stability or stress model assumes that a mutation will affect not only the native conformation of a protein but also its whole energy landscape. For example, if a protein needs to adopt a certain active conformation to function, the stability of this active state probably affects fitness. The active-stability model postulates that the fixation probability is proportional to the probability of finding the mutant in the active conformation, which in turn is a function of the stability change of the active conformation, $\Delta\Delta G^*$ (FIG. 4). This stability change can be calculated analytically via perturbed elastic network models (ENMs)[24,75]. In particular, using the parameter-free anisotropic network model (pfANM)[118], one can show that the substitution rate should be proportional to the weighted contact number WCN. Thus, the active-stability model provides a mechanism for the observed rate–WCN correlation[24,75].

The site-specific rates predicted by both the active-stability and the native-stability models are in good agreement with empirical rates[24,25,75]. However, the active-stability model tends to perform better, and there is little independent contribution from the native-stability model once the active-stability contribution is accounted for. Therefore, the empirical evidence so far favours the active-stability model. Mechanistically, the active-stability model can explain why native-stability predictions correlate with empirical rates: as shown in FIG. 4, $\Delta\Delta G^* = \Delta\Delta G + \Delta\Delta G^{\ddagger}$, where the first term is the change in native-state stability and the second term is the change in activation free energy. From the perspective of the active-stability model, $\Delta\Delta G$ affects evolutionary constraints via its effect on the stability of the active conformation. Importantly, a given mutation may destabilize the active state even if it increases native stability (FIG. 4), and thus the native-state model predictions for such mutations would be incorrect.

Challenges and future directions

Our current understanding of site-specific rate variation broadly agrees with the initial picture developed more than 40 years ago. However, whereas early work suggested definite structural regions (interior and surface), the emerging view is more nuanced (see also TABLE 1). Structural constraints continuously decrease from the solvent-inaccessible, tightly packed and rigid protein interior towards the solvent-exposed, loosely packed and flexible protein surface. Moreover, active sites and protein–protein interfaces exert additional evolutionary constraints, in the form of either positive or purifying selection, and these constraints seem to extend beyond the immediate residues involved in the function of a protein.

However, a complete and accurate predictive model of rate variation remains elusive. Our best current models can explain only ~60% of the observed rate variation, and only in some structures. Model performance varies widely among different proteins, for unknown reasons. Therefore, although the field has made considerable progress, many important questions remain unanswered (BOX 2). To make further progress, we will have to pursue

$\Delta\Delta G$
Mutational change of stability; the folding free energy difference between mutant and wild type when each is in its own native conformation.

$\Delta\Delta G^*$
Mutational change of stability of the active conformation; free energy difference between the active conformation of the mutant and the active conformation of the wild type.

$\Delta\Delta G^{\ddagger}$
Mutational change of the activation free energy; difference between mutant and wild type of the free energy needed to deform the protein from the native into the active conformation.

Box 2 | **Open questions**

How accurate are current rate estimates, and how can we improve them?

- Can we quantify the expected errors and biases of rate estimates?
- How do rate estimates depend on the number of sequences and the degree of sequence divergence?
- How robust are rate estimates with respect to violation of model assumptions, such as prior rate distributions and site independence?

Do we know all the molecular determinants of site rates?

- How much of the variation of rates among sites do the factors we currently know actually explain?
- Is the unexplained variation due to unidentified factors, errors of rate estimates or unexplainable variation (that is, noise and biases of estimates)?
- What other molecular features affect evolutionary rates? How can we incorporate the effect of specific functional features, such as an active site, into quantitative predictors of rates?

What are the mechanisms that produce site-specific evolutionary rate variation?

- Does natural selection favour more stable proteins?
- Is there a stability optimum due to stability–activity trade-offs?
- Is there a stability threshold above which all mutants are neutral?
- What is more important: stability or the correct active-site conformation?
- How can we explicitly incorporate protein function into mechanistic models?

three distinct research areas: first, we need to improve rate estimates by developing better inference methods and by quantifying the errors of these estimates for realistic data sets. Second, we must try to improve predictions by finding as yet undiscovered relevant molecular traits. Third, to advance mechanistic understanding, we need further research on theoretical models. All of these efforts are likely to benefit from stronger integration with experimental work, as discussed in BOX 1. Some questions that could help to orient future research in these three areas are listed in BOX 2.

The three main challenges. Although the field of rate estimation is mature, we still see ample room for improvement. In particular, rate estimation is subject to both stochastic and systematic errors. Estimation methods are typically based on a particular model and assessed using data simulated using the same type of model. This practice serves to assess stochastic errors and their convergence with, for example, number of sequences and divergence[29,51]. It also serves to assess whether Bayesian or ML methods result in better estimates[49,52]. However, differences between the process that generated the actual data and the model used for analysis will lead to systematic errors[119]. Using the most rigorous statistical approaches and increasing the amount of data cannot compensate for the use of incorrect models; on the contrary, it may lead to even more biased estimates[120]. Among the most important assumptions that may affect rate estimates are the codon or amino acid replacement model, prior rate distributions, and the assumptions that rates are constant over time and lineages and that sites evolve independently of each other. Investigating the effects of violating such model assumptions is, we believe, the most important current challenge for improving rate inference methods.

Phenomenological models that combine predictors such as RSA and WCN are not in perfect agreement with observed rates, and the origin of the mismatch, especially of the wide variation of explanatory power among proteins, is unknown. RSA and WCN are the best currently known predictors, but they are not the only predictors available. Even though other constraints, such as local flexibility, secondary structure and side-chain hydrogen bonding, do not seem to have a large effect on determining the overall pattern of site-specific rates, these properties have been found to affect the evolutionary process of some sites, and further work in this area may be worthwhile, in particular comparing and contrasting these quantities with RSA and WCN. More importantly, we need to develop useful predictors that quantify functional constraints. Beyond the high conservation of a few sites that are directly involved in function, there is some evidence that functional constraints induce longer-range patterns. In some cases, measures such as the distance to the active site improve site-specific rate predictions, and these and similar functional predictors of site-specific rate variation are the obvious next direction for the field. Finally, phenomenological models can also be directly integrated into the rate-inference framework[36,121], and such integrated models could provide both better rate estimates and novel insights into structural and functional evolutionary constraints.

Ultimately, we aim at a mechanistic understanding of protein evolution, derived as much as possible from first principles. We have described the two biophysical models that have been applied to the study of site-specific rates. These models are based on the idea that fitness depends on protein activity which, in turn, depends on stability changes. One of the models depends on changes in the stability of the native conformation, and the other depends on the (de)stabilization of an active conformation. Even if the active-stability model results in better predictions of site-specific rates, whether active-state stability or native-state stability or both are the primary drivers of site-specific evolution is not currently known. Moreover, no framework currently exists to incorporate functional constraints into biophysical mechanistic models. For example, what is the biophysical origin of the increase in site-specific rates with distance to the active site? Including function explicitly in mechanistic biophysical models is one of the main challenges for the further development of mechanistic models.

Other limitations of current work. All the structural predictors of rate that we discuss above suffer from one important shortcoming: they ignore pairwise interactions between amino acids. Although quantities such as solvent accessibility or packing density implicitly take into account the extent to which other amino acids are nearby, they cannot explicitly model the increased or decreased substitution rate at one site in response to a substitution at another site. However, such co-evolution among sites is well documented[122–126] and has been used to infer protein tertiary structure[127–129] and protein–protein interactions[130,131] from sequence alignments. How pairwise

interactions among sites affect rate heterogeneity, however, is poorly understood. Indeed, studies have primarily focused on inferring structure from sequence alignments, and few attempts have been made to solve the inverse problem of predicting site co-variation from structure.

As a generic approach towards developing more accurate models of structural constraints, one could move away from simple summary statistics such as RSA or WCN and could instead use mechanistic, all-atom models of protein folding. In principle, we should be able to recover any evolutionary constraints that are attributable to protein-folding stability from the detailed energetic models used in protein design algorithms, which naturally consider interactions among residues in the structure[132]. However, attempts have so far fallen short of expectations[22,133,134]; the simple quantities RSA and WCN perform much better than sophisticated all-atom protein-design calculations in predicting site-specific rates. In particular, protein design underestimates the amount of co-variation among sites observed in natural protein sequences[133]. It also tends to overestimate the variability of buried sites and underestimate that of exposed sites[134].

Importantly, we are limited in the extent to which we can compare findings from distinct existing studies, because they frequently use widely diverging inference methods, data sets or models. For example, although rates estimated with Rate4Site[51] should correlate with *dN*/*dS* estimates, the extent to which this relationship holds true has never been tested. Similarly, results found for highly diverged globular enzymes[23,75] may not be comparable to the results found for viral proteins that have experienced little divergence[22], and which may inherently possess distinct structural features relative to cellular proteins[135]. Also, although some studies are based on highly diverged data sets that include all major taxa[23,75]others are based on only a few closely related species[18]. Even though the mutational response of proteins of different major taxa seems to be universally distributed[136], the extent to which patterns of rate variation among sites can be compared among different taxa has not been assessed. Thus, future work will need to explicitly test to what extent the results obtained under one method, for one taxonomic group or for one type of data carry over to very different methods, taxonomic groups and data types.

Finally, we have focused in this Review on the variation of evolutionary rates among sites while implicitly assuming that rates are constant. However, rates vary also over evolutionary time, a phenomenon known as heterotachy[137]. If, as we have seen, site-specific rates depend on structural and functional constraints, the rate of a site should change due to evolutionary divergence of protein structure and function. Indeed, heterotachy has been observed in relationship with functional[26,138,139] and non-functional divergence[137]. In addition, even under constant structural or functional constraints, changes in the amino acids of the local environment of a site also affect its substitution rate[140]. Addressing the challenges proposed above, especially the development of mechanistic biophysical models of evolution, coupled with studies of structural and functional divergence, should advance our understanding of the variation of evolutionary rates over time.

1. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420 (2015).
2. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
3. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372 (1996).
4. Lartillot, N. & Phillipe, H. A. Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
5. Yang, Z. *Computational Molecular Evolution* (Oxford Univ. Press, 2006).
6. Holder, M. T., Zwickl, D. J. & Dessimoz, C. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Phil. Trans. R. Soc. B* **363**, 4013–4021 (2008).
7. Wang, H. C., Li, K., Susko, E. & Roger, A. J. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.* **8**, 331 (2008).
8. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).
9. Yang, Z. H., Nielsen, R., Goldman, N. & Pedersen, A. M. K. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
10. Buckley, T. R., Simon, C. & Chambers, G. K. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.* **50**, 67–86 (2001).
11. Mayrose, I., Friedman, N. & Pupko, T. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* **21**, ii151–ii158 (2005).
12. Delport, W., Scheffler, K., Gravenor, M. B., Muse, S. V. & Kosakovsky Pond, S. L. Benchmarking multi-rate codon models. *PLoS ONE* **5**, e11587 (2010).
13. Lartillot, N. Probabilistic models of eukaryotic evolution: time for integration. *Phil. Trans. R. Soc. B* **370**, 20140338 (2015).
14. Liberles, D. A., Teufel, A. I., Liu, L. & Stadler, T. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biol. Evol.* **5**, 2008–2018 (2013).
15. Perutz, M. F., Kendrew, J. C. & Watson, H. C. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669–678 (1965).
16. Kimura, M. & Ota, T. On some principles governing molecular evolution. *Proc. Natl Acad. Sci. USA* **71**, 2848–2852 (1974).
17. 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).
One of the first studies to consider both structural and functional determinants of site-specific amino acid substitution rates.
18. Franzosa, E. A. & Xia, Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26**, 2387–2395 (2009).
This landmark study found that that site-specific rate (*dN*/*dS*) increases linearly with solvent accessibility in yeast.
19. Shih, C.-H. & Hwang, J.-K. Evolutionary information hidden in a single protein structure. *Proteins* **80**, 1647–1657 (2012).
20. Nevin Gerek, Z., Kumar, S. & Banu Ozkan, S. Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol. Appl.* **6**, 423–433 (2013).
21. Marsh, J. A. & Teichmann, S. A. Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays* **36**, 209–218 (2014).
22. Shahmoradi, A. *et al.* Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *J. Mol. Evol.* **79**, 130–142 (2014).
23. Yeh, S.-W. *et al.* Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol. Biol. Evol.* **31**, 135–139 (2014).
First study showing that site-specific rates correlate more strongly with WCN than with RSA.
24. Huang, T.-T., Del Valle Marcos, M. L., Hwang, J.-K. & Echave, J. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evol. Biol.* **14**, 78 (2014).
This paper introduces the stress model of protein evolution, a biophysical model based on mutational changes of active-state stability.
25. Echave, J., Jackson, E. L. & Wilke, C. O. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys. Biol.* **12**, 025002 (2015).
Study of rate variation among sites using the native-stability threshold biophysical model.
26. Meyer, A. G., Spielman, S. J., Bedford, T. & Wilke, C. O. Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. *Virus Evol.* **1**, vev006–vev010 (2015).
27. Nielsen, R. Mapping mutations on phylogenies. *Syst. Biol.* **51**, 729–739 (2002).
28. Kosakovsky Pond, S. L. & Frost, S. D. W. A simple hierarchical approach to modeling distributions of substitution rates. *Mol. Biol. Evol.* **22**, 223–234 (2004).

29. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
Landmark paper benchmarking different methods of site-specific rate inference.
30. Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L. & Suchard, M. A. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics* **28**, 3248–3256 (2012).
31. Rodrigue, N. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* **193**, 557–564 (2013).
32. Valdar, W. S. Scoring residue conservation. *Proteins* **48**, 227–241 (2002).
33. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).
34. Johansson, F. & Toh, H. A comparative study of conservation and variation scores. *BMC Bioinformatics* **11**, 311–388 (2010).
35. Muse, S. V. Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13**, 105–114 (1996).
36. Meyer, A. G. & Wilke, C. O. Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.* **30**, 36–44 (2013).
37. Li, W.-H., Wu, C.-I. & Luo, C.-C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution consider the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150–174 (1985).
38. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
39. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–42 (2000).
40. Meyer, S. & von Haeseler, A. Identifying site-specific substitution rates. *Mol. Biol. Evol.* **20**, 182–189 (2003).
41. Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope. *Genetics* **148**, 929–936.
42. Yang, Z., Wong, W. S. W. & Nielsen, R. Bayes Empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118 (2005).
43. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
44. Kosakovsky Pond, S. L., Frost, S. D. W. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005).
45. Delport, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455–2457 (2010).
46. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
47. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
48. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
49. Murrell, B. *et al.* FUBAR: a fast, unconstrained Bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.* **30**, 1196–1205 (2013).
This paper introduces an extremely rapid but accurate method to infer *dN*/*dS*.
50. Angelis, K., dos Reis, M. & Yang, Z. Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. *Mol. Biol. Evol.* **31**, 1902–1913 (2014).
51. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**, S71–S77 (2002).
This paper introduced the Rate4Site method that is now widely used to calculate site-specific rates from amino acid sequence data.
52. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791 (2004).
53. Fernandes, A. D. & Atchley, W. R. Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. *Bioinformatics* **24**, 2177–2183 (2008).
54. Huang, Y. F. & Golding, G. B. Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Comput. Biol.* **10**, e1003429–e1003412 (2014).
55. Huang, Y.-F. & Golding, G. B. FuncPatch: a web server for the fast bayesian inference of conserved functional patches in protein 3D structures. *Bioinformatics* **31**, 523–531 (2015).
56. Yang, J.-R., Liao, B.-Y., Zhuang, S.-M. & Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl Acad. Sci. USA* **109**, E831–E840 (2012).
57. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* **8**, e80635 (2013).
This paper provides accurate normalization constants required for the calculation of relative solvent accessibility.
58. Hubbard, T. J. & Blundell, T. L. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171 (1987).
59. Lim, W. A. & Sauer, R. T. Alternative packing arrangements in the hydrophobic core of λrepressor. *Nature* **339**, 31–36 (1989).
60. Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.* **241**, 132–145 (1990).
61. Topham, C. M. *et al.* Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* **229**, 194–220 (1993).
62. Wako, H. & Blundell, T. L. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. solvent accessibility classes. *J. Mol. Biol.* **238**, 682–692 (1994).
63. Koshi, J. M. & Goldstein, R. A. Context-dependent optimal substitution matrices. *Protein Eng.* **8**, 641–645 (1995).
64. 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).
65. Conant, G. C. & Stadler, P. F. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol. Biol. Evol.* **26**, 1155–1161 (2009).
66. Ramsey, D. C., Scherrer, M. P., Zhou, T. & Wilke, C. O. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* **188**, 479–488 (2011).
67. Scherrer, M. P., Meyer, A. G. & Wilke, C. O. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol. Biol.* **12**, 179 (2012).
68. Franzosa, E. A. & Xia, Y. Independent effects of protein core size and expression on residue-level structure-evolution relationships. *PLoS ONE* **7**, e46602 (2012).
69. Lin, C.-P. *et al.* Deriving protein dynamical properties from weighted protein contact number. *Proteins* **72**, 929–935 (2008).
70. England, J. L. & Shakhnovich, E. Structural determinant of protein designability. *Phys. Rev. Lett.* **90**, 218101 (2003).
71. Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761 (2006).
72. Shakhnovich, B., Deeds, E., Delisi, C. & Shakhnovich, E. Protein structure and evolutionary history determine sequence space topology. *Genome Res.* **15**, 385–392 (2005).
73. Zhou, T., Drummond, D. A. & Wilke, C. O. Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Evol.* **66**, 395–404 (2008).
74. Yeh, S.-W. *et al.* Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *BioMed Res. Int.* **2014**, 572409 (2014).
75. Marcos, M. L. & Echave, J. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ* **3**, e911 (2015).
76. Mugal, C. F., Wolf, J. B. W. & Kaj, I. Why time matters: codon evolution and the temporal dynamics of d*N*/d*S*. *Mol. Biol. Evol.* **31**, 212–231 (2014).
77. Liu, Y. & Bahar, I. Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.* **29**, 2253–2263 (2012).
Study of the correlation between flexibility and site-specific sequence entropy.
78. Halle, B. Flexibility and packing in proteins. *Proc. Natl Acad. Sci. USA* **99**, 1274–1279 (2002).
79. Liao, H., Yeh, W., Chiang, D., Jernigan, R. L. & Lustig, B. Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Eng. Des. Sel.* **18**, 59–64 (2005).
80. Worth, C. L., Gong, S. & Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **10**, 709–720 (2009).
81. 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).
82. Brown, C. J. *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110 (2002).
83. Brown, C. J., Johnson, A. K. & Daughdrill, G. W. Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.* **27**, 609–621 (2010).
84. Tóth-Petróczy, A. & Tawfik, D. S. Slow protein evolutionary rates are dictated by surface-core association. *Proc. Natl Acad. Sci. USA* **108**, 11151–11156 (2011).
Systematic study of the distributions of site-specific rates for yeast proteins.
85. Finkelstein, A. V., Ivankov, D. N., Garbuzynskiy, S. O. & Galzitskaya, O. V. Understanding the folding rates and folding nuclei of globular proteins. *Curr. Protein Pept. Sci.* **8**, 521–536 (2007).
86. Ptitsyn, O. B. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278**, 655–666 (1998).
87. Mirny, L. & Shakhnovich, E. Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129 (2001).
88. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxco, K. W. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* **316**, 225–233 (2002).
Study that shows that sites involved in the folding nucleus are not particularly conserved.
89. Tseng, Y. Y. & Liang, J. Are residues in a protein folding nucleus evolutionarily conserved? *J. Mol. Biol.* **335**, 869–880 (2004).
90. 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).
91. Franzosa, E. A., Xue, R. & Xia, Y. Quantitative residue-level structure–evolution relationships in the yeast membrane proteome. *Genome Biol. Evol.* **5**, 734–744 (2013).
92. Spielman, S. J. & Wilke, C. O. Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. *J. Mol. Evol.* **76**, 172–182 (2013).
93. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121 (2002).
94. Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504 (2004).
95. McLaughlin R. N. Jr, Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **490**, 138–142 (2012).
96. Mintseris, J. & Weng, Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci. USA* **102**, 10930–10935 (2005).
This paper shows that sites that participate in obligate protein–protein interactions are more conserved than those involved in transient interactions.
97. Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938–1941 (2006).

98. Huang, Y.-W., Chang, C.-M., Lee, C.-W. & Hwang, J.-K. The conservation profile of a protein bears the imprint of the molecule that is evolutionarily coupled to the protein. *Proteins* **83**, 1407–1413 (2015).
99. Kachroo, A. H. *et al.* Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921–925 (2015).
100. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins Struct. Funct. Genet.* **62**, 479–488 (2006).
101. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **5**, e1000585 (2009).
102. Yang, J. S., Seo, S. W., Jang, S., Jung, G. Y. & Kim, S. Rational engineering of enzyme allosteric regulation through sequence evolution analysis. *PLoS Comput. Biol.* **8**, e1002612–e1002610 (2012).
103. Hill, R. E. & Hastie, N. D. Accelerated evolution in the reactive centre regions of serine protease inhibitors. *Nature* **326**, 96–99 (1987).
104. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
105. Bush, R. M., Fitch, W. M., Bender, C. A. & Cox, N. J. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16**, 1457–1465 (1999).
106. Shih, A. C., Hsiao, T., Ho, M. & Li, W. Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. *Proc. Natl Acad. Sci. USA* **104**, 6283–6288 (2007).
107. Pan, K. & Deem, M. W. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. *J. R. Soc. Interface* **8**, 1644–1653 (2011).
108. Tusche, C., Steinbrück, L. & McHardy, A. C. Detecting patches of protein sites of influenza A viruses under positive selection. *Mol. Biol. Evol.* **29**, 2063–2071 (2012).
109. Meyer, A. G. & Wilke, C. O. Geometric constraints dominate the antigenic evolution of influenza H3N2 hemagglutinin. *PLoS Pathog.* **11**, e1004940 (2015).
110. Liberles, D. A. *et al.* The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* **21**, 769–785 (2012).
111. Harms, M. J. & Thornton, J. W. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).
112. Zhou, H. & Zhou, Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* **322**, 315–322 (2004).
113. Shaytan, A. K., Shaitan, K. V. & Khokhlov, A. R. Solvent accessible surface area of amino acid residues in globular proteins: correlation of apparent transfer free energies with experimental hydrophobicity scales. *Biomacromolecules* **10**, 1224–1237 (2009).
114. Bloom, J. D. & Glassman, M. J. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput. Biol.* **5**, e1000349 (2009).
115. Wylie, S. C. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl Acad. Sci. USA* **108**, 9916–9921 (2011).
116. Wylie, S. C. & Shakhnovich, E. I. Mutation induced extinction in finite populations: lethal mutagenesis and lethal isolation. *PLoS Comput. Biol.* **8**, e1002609 (2012).
117. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
118. Yang, L., Song, G. & Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl Acad. Sci. USA* **106**, 12347–12352 (2009).
119. Spielman, S. J. & Wilke, C. O. The relationship between *dN*/*dS* and scaled selection coefficients. *Mol. Biol. Evol.* **32**, 1097–1108 (2015).
This paper establishes a mathematical relationship between mutation–selection models and *dN*/*dS* ratios.
120. Kolaczkowski, B. & Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984 (2004).
121. Kleinman, C. L., Rodrigue, N., Lartillot, N. & Philippe, H. Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.* **27**, 1546–1560 (2010).
122. Pagel, M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. B Biol. Sci.* **255**, 37–45 (1994).
123. Muse, S. V. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* **139**, 1429–1439 (1995).
124. Poon, A. F. Y., Lewis, F. I., Kosakovsky Pond, S. L. & Frost, S. D. W. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput. Biol.* **3**, 2279–2290 (2007).
125. Carlson, J. M. *et al.* Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.* **4**, e1000225 (2008).
126. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* **7**, e1001301 (2011).
127. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **6**, e1000633 (2010).
128. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
129. Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
130. Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043–1054 (2008).
131. Cheng, R. R., Morcos, F., Levine, H. & Onuchic, J. N. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl Acad. Sci. USA* **111**, E563–E571 (2014).
132. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
133. Ollikainen, N. & Kortemme, T. Computational protein design quantifies structural constraints on amino acid covariation. *PLoS Comput. Biol.* **9**, e1003313 (2013).
134. Jackson, E. L., Ollikainen, N., Covert, A. W., Kortemme, T. & Wilke, C. O. Amino-acid site variability among natural and designed proteins. *PeerJ* **1**, e211 (2013).
135. Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N. & Tawfik, D. S. Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* **34**, 53–59 (2009).
136. Faure, G. & Koonin, E. V. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys. Biol.* **12**, 035001 (2015).
137. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).
138. Gu, X. Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* **16**, 1664–1674 (1999).
139. Gu, X. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol. Biol. Evol.* **23**, 1937–1945 (2006).
140. Pollock, D. D., Thiltgen, G. & Goldstein, R. A. Amino acid coevolution induces an evolutionary stokes shift. *Proc. Natl Acad. Sci. USA* **109**, E1352–E1359 (2012).
This paper introduces the concept of evolutionary Stokes shift: when an amino acid substitution occurs at a site, its neighbours evolve more rapidly to accommodate the substitution.
141. Leferink, N. G. H. *et al.* Impact of residues remote from the catalytic centre on enzyme catalysis of copper nitrite reductase. *Nat. Commun.* **5**, 4395 (2014).
142. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
143. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
144. Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl Acad. Sci. USA* **112**, 7159–7164 (2015).
145. Bloom, J. D. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homolog. *Mol. Biol. Evol.* **31**, 2753–2769 (2014).
146. Abriata, L. A., Palzkill, T. & Dal Peraro, M. How structural and physicochemical determinants shape sequence constraints in a functional enzyme. *PLoS ONE* **10**, e0118684 (2015).
This paper shows one example (TEM lactamase) for which functional constraints relax slowly with distance to the active site.
147. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **31**, 1956–1978 (2014).
One of the first studies to parameterize a phylogenetic model with experimentally measured, site-specific parameters.
148. Doud, M. B., Ashenberg, O. & Bloom, J. D. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.* **32**, 2944–2960 (2015).

Acknowledgements
J.E. is Principal Investigator of CONICET. This work was also supported in part by US National Institutes of Health (NIH) grant F31 GM113622-01 to S.J.S. and by NIH grant R01 GM088344, NSF Cooperative agreement DBI-0939454 (BEACON Center), and ARO grant W911NF-12-1-0390 to C.O.W.

Competing interests statement
The authors declare no competing interests.