JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Headline review

Cite this article: Sikosek T, Chan HS. 2014 Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface* **11**: 20140419.
http://dx.doi.org/10.1098/rsif.2014.0419

Received: 22 April 2014
Accepted: 28 July 2014

Subject Areas:
biophysics, bioinformatics, computational biology

Keywords:
adaptation, promiscuous functions, conformational dynamics, hidden states, protein folding, protein–protein interactions

Authors for correspondence:
Tobias Sikosek
e-mail: t.sikosek@utoronto.ca
Hue Sun Chan
e-mail: chan@arrhenius.med.toronto.edu

Royal Society Publishing

Biophysics of protein evolution and evolutionary protein biophysics

Tobias Sikosek[1,2,3] and Hue Sun Chan[1,2,3]

[1]Department of Biochemistry, [2]Department of Molecular Genetics, and [3]Department of Physics, University of Toronto, Toronto, Ontario, Canada M5S 1A8

TS, 0000-0001-9929-3525

The study of molecular evolution at the level of protein-coding genes often entails comparing large datasets of sequences to infer their evolutionary relationships. Despite the importance of a protein's structure and conformational dynamics to its function and thus its fitness, common phylogenetic methods embody minimal biophysical knowledge of proteins. To underscore the biophysical constraints on natural selection, we survey effects of protein mutations, highlighting the physical basis for marginal stability of natural globular proteins and how requirement for kinetic stability and avoidance of misfolding and misinteractions might have affected protein evolution. The biophysical underpinnings of these effects have been addressed by models with an explicit coarse-grained spatial representation of the polypeptide chain. Sequence–structure mappings based on such models are powerful conceptual tools that rationalize mutational robustness, evolvability, epistasis, promiscuous function performed by 'hidden' conformational states, resolution of adaptive conflicts and conformational switches in the evolution from one protein fold to another. Recently, protein biophysics has been applied to derive more accurate evolutionary accounts of sequence data. Methods have also been developed to exploit sequence-based evolutionary information to predict biophysical behaviours of proteins. The success of these approaches demonstrates a deep synergy between the fields of protein biophysics and protein evolution.

1. Introduction

Biological evolution uses mutations as its basic working material. Mutations occur in DNA molecules through various mechanisms. Some mutations are relatively 'silent' in that their effects are less appreciable, whereas others have a more prominent impact on the biological function. The most immediate effect of a mutation is the alteration of the DNA molecule itself and thus, possibly, its affinities to bind certain proteins or RNA. Given the vastness of many genomes, it was once believed that many mutations in DNA fall in regions that have no biological function. However, with increasing knowledge of the functional roles of non-coding DNA sequences, the proportion of genomes that is considered non-functional has decreased significantly [1]. Regions of the genome that do encode for a functional RNA or protein can undergo several different kinds of mutations, such as insertions, deletions and duplications of entire segments of DNA. The present review focuses primarily on the effect of point mutations (change of a single nucleotide) and will consider only proteins but not RNA, although many general principles of evolution are applicable to both classes of biomolecules. We refer to other authors for the evolution of protein structures via sequence re-arrangements such as domain-wise evolution [2–4], the fusion of small peptide fragments [5] or the 'chimeric' recombination of fragments that is also exploited in protein engineering [6–9].

Current study of molecular evolution can benefit from a huge amount of sequence data, but only a relatively small body of structural data. Consequently, many approaches in evolutionary studies are predominantly sequence-based. A prime example is phylogenetic inference methods based upon multiple sequence alignments. Mostly, the biophysical foundation of

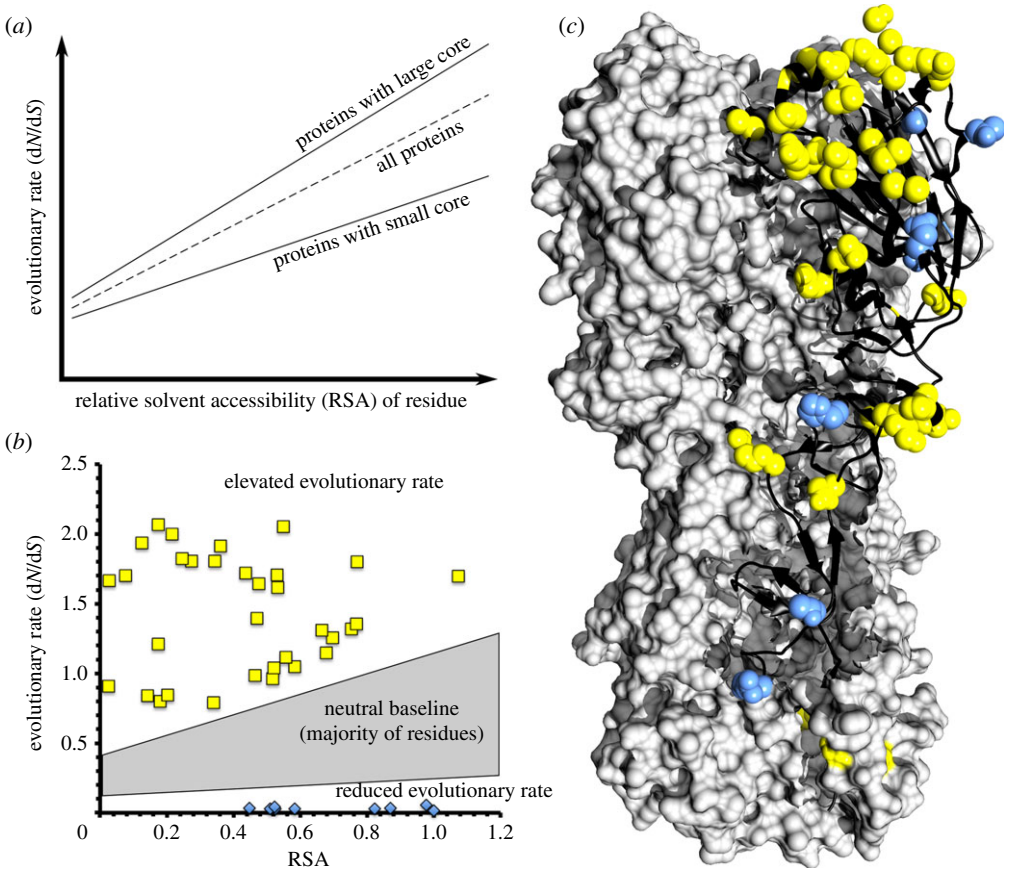

Figure 1. Enhancing evolutionary methods with biophysical information. (*a*) Relative solvent accessibility (RSA), the exposure of an amino acid to solvent in the folded structure, is strongly correlated with the evolutionary rate $\omega \equiv dN/dS$ [22], where dN/dS is the ratio of non-synonymous over synonymous substitution rates. (*b*) RSA was used to improve the neutral baseline for the detection of positive and negative selection in influenza proteins [23,24]. Data points in yellow have elevated evolutionary rates (larger ω) but can be either below or above the conventional $\omega = 1$ divide that typically distinguishes positive ($\omega > 1$) and negative/purifying ($\omega < 1$) selection. Blue data points show sites evolving at reduced evolutionary rates compared with the neutral baseline. (*c*) The homo-trimeric haemagglutinin [25] is an influenza surface glycoprotein. The structure of one of the three monomeric units is shown as black ribbons with the residues under positive or negative selection highlighted as spheres using the same colour code as that in (*b*); the other two monomeric units are depicted by the grey surface. (Adapted from [24]). A majority of the positively selected sites are found around the region that is most frequently targeted by antibodies (top right of the structure in (*c*)) and are thus under strong selection pressure to diversify.

these mathematical methods is provided only rudimentarily by the BLOSUM [10] or PAM [11] substitution matrices that are empirical summaries of the posterior probabilities of various amino acid substitutions. These models can roughly capture the tendency to conserve the physico-chemical properties of amino acids when they undergo mutations, like polar amino acids that are mostly substituted by other polar amino acids but less frequently by hydrophobic ones. However, such trends capture only a tiny aspect of the many biophysical implications of mutations that can be important for the biological function of proteins. For instance, they often do not even consider the local structural environment of a given amino acid residue position such as backbone conformation and hydrogen bonding pattern that might constrain evolutionary choices [12].

In this context, a number of authors from within the biophysics community have recently called for a stronger collaboration between the fields of molecular evolution and protein biophysics in order to achieve new and deeper insights into protein evolution [13–17]. At the same time, within the phylogenetics community there is a growing realization of the need for including structure information into evolutionary models [18–20]. As a first step in pursuing this direction of investigation, the effect of mutations on protein stability or binding affinities is probably the most promising example of how biophysics can contribute to a better understanding of evolution [21].

The two fields can clearly benefit from each other. For example, a common evolutionary method to identify gene positions that have undergone significant mutational changes and to quantify the degree of selection is to compute the ratio of non-synonymous to synonymous substitutions. However, a correlation between this ratio and the solvent exposure of the site in the folded protein structure has been noted recently [22] (figure 1), suggesting that this ratio may not be purely a measure of adaptive selection but may also reflect the site's contribution to protein stability. Based on this finding, solvent exposure of residues has been used in establishing a new neutral baseline that reflects this biophysical constraint under which natural selection must operate. Notably, this procedure has led to recognition of new amino acid positions in the influenza protein haemagglutinin that have undergone adaptation (figure 1), highlighting how biophysical/structural knowledge can improve evolutionary analysis [23,24].

Conversely, evolutionary information can also provide novel biophysical understanding of proteins. One earlier example in using such an approach that may be termed *evolutionary protein biophysics* is the utilization of evolutionary

data on the PDZ domain family to predict energetically coupled positions on the protein, some of which are spatially far apart [26]. Another example is the inference of structural information from protein sectors, which are co-evolving clusters of spatially proximate and physically interacting amino acids within a protein structure. A protein such as rat trypsin [27], for example, can have several such clusters that have distinct functions and evolve independently (figure 2). The existence of protein sectors raises fundamental concerns over phylogenetic methods that assume no such biophysical interactions, because those methods led to inconsistent phylogenetic trees depending on whether they are deduced from all mutations of the protein or from considering only mutations within a sector. However, with appropriate analysis, biophysical studies of proteins can use this type of evolutionary information to predict the correct fold of a protein, deduce interactions between protein monomeric units in a multiple-chain protein complex and identify hitherto unknown functional conformations [28–32].

In the following, we first discuss the basic constraints of biophysics on evolution by surveying salient biophysical consequences of protein mutations. We then outline recent advances in using biophysical concepts to shed light on experimentally observed evolutionary behaviours.

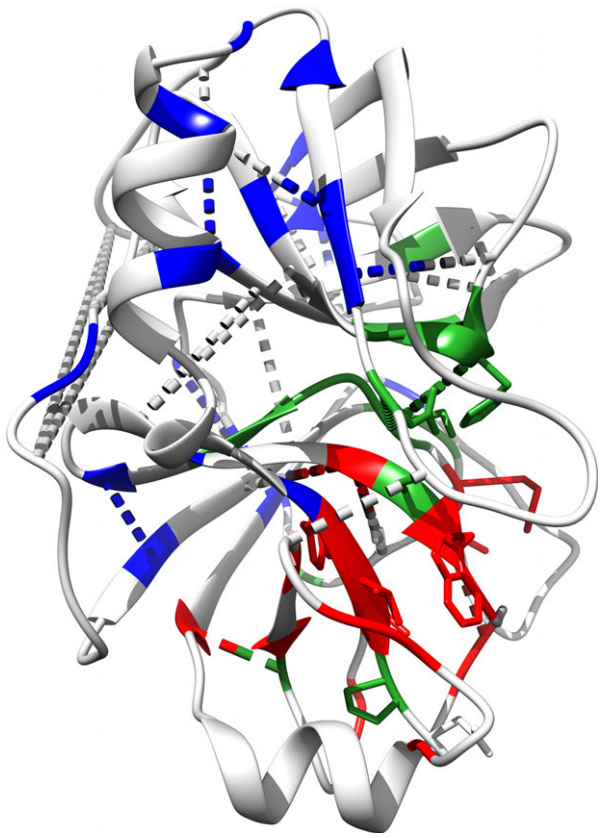

Figure 2. Co-evolving residues in rat trypsin (PDB code 3TGI), a serine protease. Protein sectors are networks of co-evolving residues with independent functions [27]. Here, the three sectors of serine proteases are shown in red (substrate specificity), blue (thermal stability) and green (catalysis). Known functional residues are shown as sticks. The existence of protein sectors has important consequences for phylogenetic analyses, since each sector evolves independently. Protein sectors were identified using the statistical coupling analysis (SCA) approach, whereas a different approach, direct coupling analysis (DCA), yielded a partially different set of co-evolving residues (dashed lines). Residue pairs from DCA have successfully been used in combination with structure-based models to predict native structure, protein–protein interactions and conformational changes [28,29]. These examples illustrate how the fields of biophysics and molecular evolution can benefit from each other. (Adapted from [27,29].)

2. Biophysical consequences of protein mutations

2.1. Mutational effects on the thermodynamic stability of protein folded states

For proteins that have a globular folded native structure, the thermodynamic stability of the folded structure relative to the ensemble of unfolded conformations is determined by the balance between the interactions that favour the folded state and the conformational entropy that favours the unfolded state. The more stable a protein, the more difficult it is to unfold (denature) under high temperatures or high concentrations of denaturing chemicals. To illustrate the energetic balance governing protein stability and its kinetic implications, the conformational diversity of the unfolded state and the essentially unique native structure of a globular protein is often depicted by a funnel-like representation of the free energy landscape of the protein conformations. The folded state is situated at the bottom of the funnel whereas the unfolded state populates the top of the funnel [33–36] (see for example the top-left drawing in figure 3).

In protein evolution studies, stability is often used as a proxy for the fidelity of a protein function, because a sufficient stability of the native state is often required for function [21]. Although a protein's function is not equivalent to its stability, experimental support exists for a positive correlation between protein functionality and native stability (e.g. [39–41]). This relationship can be seen very clearly in a recent experiment demonstrating how the evolutionary trajectory of influenza nucleoprotein is probably constrained to avoid low-stability sequences [42] (see further discussion in §3.8). In general, a mutation that decreases the stability of a protein is more probable than a mutation that does not decrease the protein's stability to lead to the formation of other non-functional structures that would be detrimental to the protein's original (wild-type) biological function, and in the worst case can cause serious harm to the organism.

The qualitative impact of a mutation on the folded state of a protein can often be anticipated. In globular proteins, surface residues are mostly polar and charged, while core residues have a higher tendency to be hydrophobic [43,44]. Mutations that conserve these properties are less likely to result in a large change in stability. In addition, the statistical propensities for certain amino acids to occur in a particular type of secondary structure have also been compiled and can be used to predict probable mutational effects on secondary structure (e.g. [45]). A recent comprehensive review of numerous studies of mutants occurring in natural protein families and superfamilies shows clearly that amino acid substitutions are constrained differently—i.e. their viabilities vary—in different local environments as defined by the main-chain secondary structure, solvent accessibility and hydrogen bonding [12].

Using stability as proxy for function, quantitative stability prediction is widely used to address the effect of mutations on protein function. Many tools exist to calculate an estimated $\Delta\Delta G$, or change in free energy, after one or more mutations

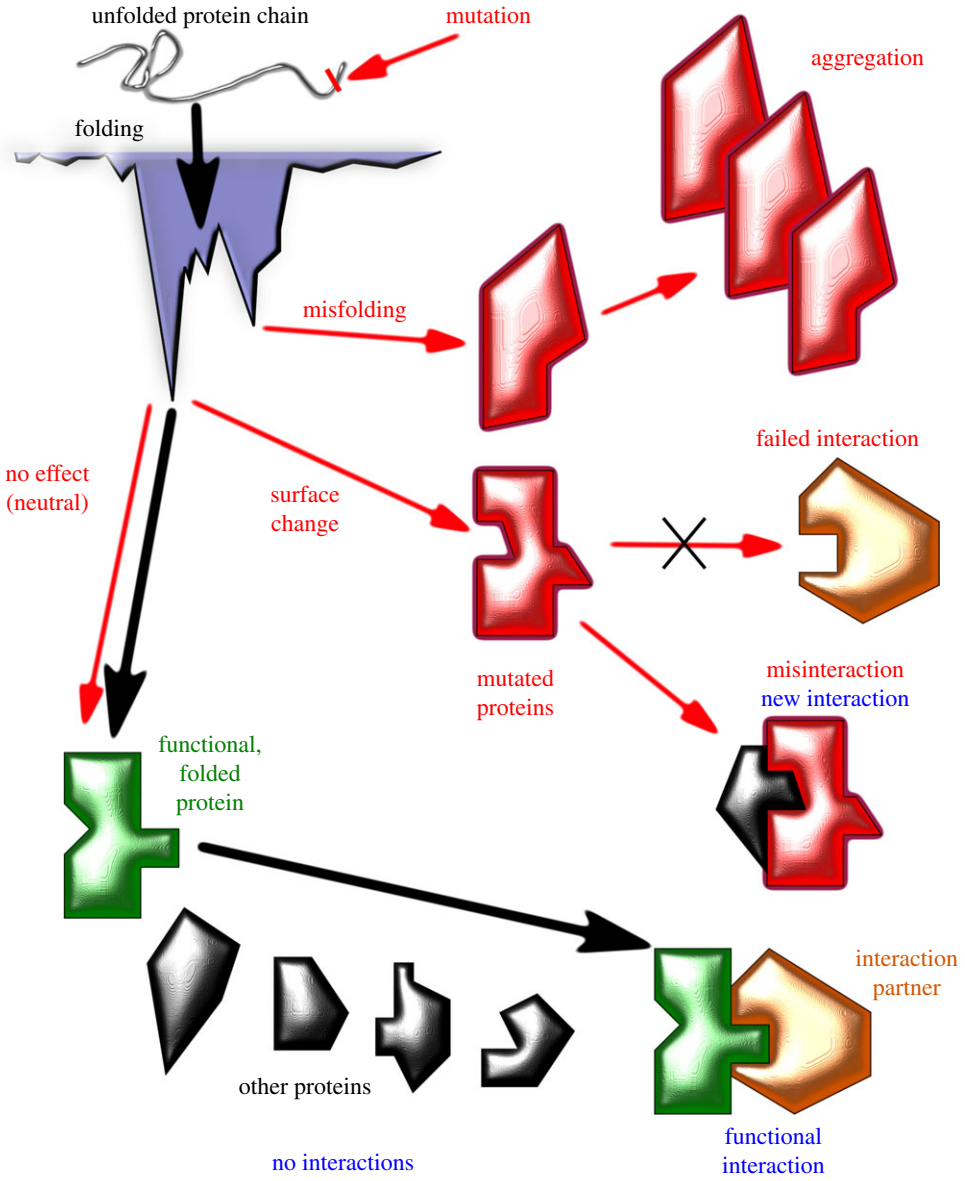

Figure 3. Schematics of some of the possible effects of mutations on protein folding and interaction. The top-left cartoon of the folding landscape of a globular protein shows the correctly folded structure as the global free energy minimum, whereas a shallower minimum corresponds to a misfolded structure. Interactions of the original protein are indicated by black arrows; those of the mutant are indicated in red. Mutations can lead to misfolding and/or aggregation and/or mis-interactions. Mutations can also lead to no apparent changes (neutral mutation). Some non-neutral mutations, however, can lead to new functional interactions that can then be subject to evolutionary selection. Note that the depiction of interactions between folded proteins as a 'lock and key' fit between specific shapes is adopted here merely to simplify the schematic representation. The perspective conveyed by the present figure does not preclude more dynamic binding mechanisms such as induced fit [37] and conformational selection [38].

[46–52]. Most of these methods focus on a static reference structure for which an energy or a score is calculated according to an empirical forcefield. To implement the mutation, the structure is computationally modified; energy is then re-calculated and compared against the pre-mutation wild-type value. $\Delta\Delta G$ prediction is widely used to screen large numbers of mutations, often in combination with laboratory experiments [53–57]. The approach has also served as fitness estimators in simulation studies of protein evolution [58,59].

One obvious limitation of these $\Delta\Delta G$ prediction methods is that, with few exceptions [60–62], they consider only a single 'native' protein conformation. In essence, these methods disregard mutational effects on the unfolded state and often ignore the possibility of structural adjustment of the folded state in response to the mutation. The accuracy of these methods is limited because in reality the mutational effects on protein stability are determined by the balance between the impact of the mutation on the folded and the unfolded states. Moreover, these methods do not address possible change from one folded structure to another, nor the possibility of misfolding; but conformational transition is crucial for exploring new protein functions during evolution, with polar-to-hydrophobic substitutions having a higher potential to lead to alternative folded structures [63–65]. In fact, sometimes a mutation may seem harmless in the native structure but can have dramatic effects during the folding process so that the native state might not even be formed (see §2.2).

In principle, with improved algorithms and appropriate atomistic forcefields, extensive molecular dynamics simulations that sample both the folded and unfolded conformations may provide more accurate stability predictions [66], even predictions of conformation transition [67–69]. But currently the computational cost for such simulations is very high; thus molecular dynamics cannot yet be used for large-scale mutation screenings.

2.2. Effects of mutation on folding kinetics and intermediate states

The impact of mutations on a globular protein is not limited to its folded structure. The folding process itself is altered by mutations, even when the end-point of the folding kinetics of the mutant is essentially the same folded structure as that of the original sequence. Kinetics of folding is often two-state-like for small, single-domain proteins [70] but transiently populated intermediate states are observed in many other proteins [71]. Mutations can affect folding speeds of both two-state-like and non-two-state proteins by modulating the interactions that favour the native state [72–75] or through strengthening certain non-native interactions not present in the folded structure [76,77].

Folding kinetics can be subject to natural selection. A recent estimate pointed to an overall increase in folding speed during evolution. Specifically, the folding speeds of α-proteins (folded structures consisting mostly of α-helices) have increased throughout evolution whereas those of β-proteins (folded structures consisting mostly of β-sheets) appear to have been decreasing in the last 1.5 billion years [78]. In an earlier study of conserved amino acid positions across protein families, it was concluded that conserved sites are important for function or stability, and that there has been 'evolutionary pressure towards fast (not necessarily the fastest) folding of several proteins' [79]. By contrast, a subsequent investigation of 48 natural mutants with single-site substitutions in the hydrophobic core of the SH3 domain (a β-protein; not considered in [79]) indicated that conservation correlates well with unfolding rates but not the folding rates of the mutants. In other words, mutants with slower unfolding rates occur more frequently than mutants with faster unfolding rates, but a positive or negative correlation between folding rate with occurrence frequency was not observed. This finding suggests that evolution selects more strongly for a slower unfolding rate than faster folding rate, at least for the SH3 family [80].

In this regard, a recent survey argued that protein kinetic stability, i.e. a slow unfolding rate, is often more strongly selected by evolution than thermodynamic stability, most probably because kinetic instability (a faster unfolding rate) facilitates irreversible alteration processes such as amyloid formation and other forms of detrimental protein aggregation even if overall thermodynamic stability is maintained by a higher folding rate [81]. Echoing the aforementioned study of SH3 domains, an investigation of 27 single-substitution variants of thioredoxin—the fold of which is apparently extremely ancient in evolutionary history [82]—indicates that viable mutants can at most be 2 kcal mol^{-1} less stable than the wild-type, but a significant correlation exists between slower unfolding rate and the occurrence frequency of a given residue in sequence alignments, again suggesting a significant natural selection for slower unfolding rates [83].

For proteins that undergo folding with significantly populated transient intermediates, a mutation may stabilize or destabilize the intermediate conformations, or even abrogate the intermediates encountered in the folding of the original sequence, or create new intermediates. In fact, in some experiments, mutations were intentionally introduced to stabilize various folding intermediates to facilitate their characterization [84,85]. In one case, swapping certain hydrophobic core residues between two related proteins could also swap the associated folding intermediates [86]. In more extreme cases, a mutation could lead to the formation of different folding intermediates or even different folded structures with potentially severe implications for protein function and aggregation [87]. In particular, highly abundant proteins with relatively low solubilities are prone to aggregate [88]. An increasing number of neurodegenerative and other varieties of prion and amyloid diseases are now known to be caused by misfolded structures (different 'native' structures) or by aggregation/oligomerization of intermediate conformational states, with propensity for misfolding increased by certain mutations [87,89] (figure 3). Cataracts in the human eye are also found to be caused by accumulation of misfolded proteins [90] and associated with mutations that led to abnormal folding behaviour [91,92]. As exemplified by the mouse prion protein and consistent with the general observation of evolutionary selection for kinetic stability [81], the folding and maintenance of the non-disease folded form of some of the pertinent proteins (the misfolded forms of which are implicated in diseases) is under kinetic rather than thermodynamic control [93]. Consistent with these observations, the experimentally observed distribution of protein evolution rates may be rationalized by an evolutionary process that selects against misfolding [94].

In the cellular environment, mutations can affect not only the folding kinetics of a protein in isolation but also how it interacts with the complex cellular machinery while it is folding. Inasmuch as folding kinetics is concerned, the *in vivo* translational rate can affect co-translational folding [95,96] because, for example, fast-translating codons can be useful for avoiding misfolding. In this regard, even synonymous mutations that do not change the amino acid sequence of a protein can lead to altered folding pathways in the cell [97].

2.3. Interactions and misinteractions

The biological functions of most proteins require them to interact with other proteins and/or other biomolecules [98]. Mutations affect these interactions and can lead to misinteractions [99]. A classic example is the glutamic acid to valine mutation in haemoglobin [100] that causes aggregation of haemoglobin and consequently sickle-cell anaemia [101]. More recent examples include mutations implicated in prion, amyloid and other misfolding diseases mentioned above [102] as well as disease-causing mutations that disrupt or weaken the proper binding between two proteins [103,104].

The cellular environment is crowded [105,106]. This crowdedness is probably dictated by biophysical constraints imposed by a living cell's need for efficient rates of biochemical reactions [107]. Within the cellular confine, a given protein can potentially come into contact with a large number of other proteins [108,109]. Although the possibility of non-specific binding probably constitutes a biophysical constraint that might have restricted the number of proteins in a cell [110], natural proteins can function by being remarkably specific binders. This interaction specificity entails not only favourable binding with a protein's target molecule(s) but also extremely unfavourable—essentially absence of—binding with many other molecules. This requirement is conceptually similar to the well-known principle for protein design, i.e. that an optimized sequence has to 'design in' the target structure as well as 'design out' alternative structures [111]. Many natural proteins have evolved not only to fold to the functional native state but also to strongly destabilize non-native intermediate states [112] by increasing the energetic

separation between the folded and unfolded states [113,114] such that the folding–unfolding transition is switch-like [36,115]. Therefore, in line with both the folding and interaction requirements, functional proteins have to disfavour non-native intra-protein interactions as well as discriminate against detrimental inter-protein misinteractions (figure 3).

There is a biophysical limit to evolutionary optimization of protein binding specificity, however. Because proteins are made up of a finite alphabet of amino acid residues [116], the heterogeneity, or designability, of their interactions are constrained by the physico-chemical properties of the alphabet. It is not physically possible to eliminate all favourable interactions between a protein and all other proteins except its presumed functional partner(s). In other words, misinteractions cannot be eliminated completely by optimization. In the living cell, there can be more misinteractions because some evolving proteins have not had time to minimize them [117]. In fact, even the folded form of a globular protein is probably a metastable state, whereas amyloid [118] or prion-like [119] aggregates are expected to be thermodynamically more stable configurations at longer timescales. Therefore, binding should not be understood as an all-or-none proposition; instead it is a question of binding affinities that can vary over a wide range. Although proteins bind their evolved interaction partners particularly strongly, they probably also interact transiently with many other proteins, albeit with low affinities. Currently it is not feasible to identify the effects of a given mutation on the many possible interactions a protein can engage in, especially when the mutation has no detectable effect on the main function. Nonetheless, computational prediction methods are being developed to perform efficient tests for potential binding between large numbers of proteins [120].

Any mutation on a protein can potentially increase the binding strength with some molecular partners. If this change alters the cellular biochemistry, the mutation may be subject to either positive or negative natural selection (figure 3). A misinteraction is created by mutation if an originally negligible protein–protein interaction is strengthened to an appreciable level. If the misinteraction is beneficial, it can underpin a new oligomeric state or promiscuous function of the protein which can then be positively selected [121,122] (see further discussion in §3.5). In those cases, computational modelling suggests that positive selection of an interacting region can also facilitate evolution of globally well-packed globular structures in the interacting proteins [123,124].

Protein–protein interactions require geometric coupling of the protein interfaces. Mutations within the interfaces naturally have a direct impact on binding; mutations outside the interface can affect binding allosterically as well [125] (see further discussion in §2.7.1). Biophysically, new protein–protein interactions are not unlikely to emerge. A recent survey of heterodimers found that functional binding interfaces bury a surface area between 380 and 3400 $Å^2$ [126]. Another recent study indicated that only two amino acid substitutions are needed to shift the average amino acid composition of a 1000 $Å^2$, approximately 28-residue non-interacting protein surface to that of a protein–protein interface [127]. In this light, transient binding may be possible with even smaller interfaces. One can imagine a 'grey area' of interface sizes where a single surface mutation may significantly increase the binding affinity to a new substrate. There are also overlapping binding interfaces that bind different substrates [128,129], which can be created easily via mutations from an original interface that binds only one substrate. This perspective is consistent with a recent directed evolution study on the bacterial immunity protein Im9. The wild-type Im9 primarily inhibits deoxyribonuclease ColE9 but also inhibits ColE7 promiscuously, i.e. to a much lesser extent. The experiment shows that it can evolve readily into a primary ColE7-inhibitor with an approximately 10^5-fold increase in affinity and 10^8-fold increase in selectivity via a 'generalist' intermediate that allows for rapid evolutionary divergence [130].

2.4. Marginal native stability

Since native stability is required for globular proteins to perform their biological functions (§2.1) and to avoid misfolding and aggregation (§2.2), it might seem that a higher native stability should always be desirable and therefore favoured by evolution. However, natural globular proteins are not extremely stable. An early survey of the thermal stability of 12 proteins at 25°C showed considerable variation of native stability among them, with average stabilizing free energies of 0.05–0.12 kcal per mole of amino acid residues [131]. This and other experimental data indicate an approximate native stability of 5–15 kcal mol^{-1} for a natural globular protein with about 100 amino acids. These findings have since been rationalized theoretically by considering the strength of intra-protein interactions and conformational entropy [44,132]. This experimental level of stability of natural globular proteins is often characterized as 'marginally stable'. 'Marginal' here points to the relatively small free energies of folding. Sometimes the term also refers to the fact that the net balance of 5–15 kcal mol^{-1} for native stability is the result of a partial cancellation of two much larger free energies on the order of 100–200 kcal mol^{-1} contributed by favourable intra-protein interactions on one hand and conformational entropy on the other [44].

If evolutionary selection for stability is expected, why are natural proteins only marginally stable? One possible reason is that native stability is not the only requirement on a functional globular protein. Conformational flexibility is crucial for certain biological functions. Therefore, adaptation towards increased conformational flexibility might have acted as a check against proteins evolving to become extremely stable [21,133,134], suggesting that marginal stability can be an adaptive trait.

2.4.1. Marginal stability may not be an adaptive property

Is a strong selection pressure for marginal stability necessary to account for the experimentally observed marginal stability of natural proteins? Biophysics-based models have suggested otherwise by showing that marginal stability could be a non-adaptive property [135,136]. The number of sequences encoding for a given structure generally decreases with native stability. Hence, even in the absence of any evolutionary selection, there are more sequences encoding for a given native structure with low stabilities than sequences encoding for the same structure with high stabilities. This phenomenon is a basic property of protein sequence space and is consistent with the 'superfunnel' perspective [137] (§3.2.3). Therefore, as long as a certain minimum stability requirement for folding and function is met, random mutational drift will lead an evolving population to a region of sequence space that

encodes with marginal stabilities (close to the minimum required stability) simply because there are more sequences with that property [135]. In a more recent model, an evolved population is seen to prefer marginal stability even when the model fitness function increases exponentially with native stability [136]. In this view, if marginal stability of a protein is functionally beneficial, it may represent a 'spandrel' [135], i.e. a tendency occurring originally for non-adaptive reasons that is exploited subsequently by biology [138].

This population consideration argues convincingly that there might not have been extensive positive evolutionary selection to *decrease* the stabilities of globular proteins. A fundamental issue that remains to be addressed, however, is the extent of evolutionary selection to *increase* stability. This question asks whether the stabilities of natural proteins are close to their biophysical maximum, as envisioned in the superfunnel picture (§3.2.3) or are far from a biophysically possible maximum that was not selected evolutionarily. Notably, both of the models discussed above [135,136] posit that there are amino acid sequences that can fold to a given structure uniquely with native stabilities far exceeding the experimentally observed stabilities of natural proteins. Results of the random mutation model of Taverna & Goldstein [135] show a significant population of sequences encoding with higher native stabilities than the sequences around the peak of the steady-state population. Therefore, if the sequences near the peak of the population distribution are taken as models for natural proteins, their results suggest that a significant fraction of mutations of natural proteins would lead to higher native stabilities (although that fraction is smaller than the fraction of mutations leading to lower native stabilities). In a more recent model of Goldstein [136], it is stated specifically that the 300-residue protein used in the study can potentially reach an extremely high stability of 118 kcal mol^{-1} but the evolved population has a stability of only about 9 kcal mol^{-1}.

2.4.2. How stable can real proteins be?

Is it physically possible for some amino acid sequences to fold with exceedingly high stability? The perspective from experiments is different from that suggested by Goldstein [136]. Among 290 single-residue substitutions of staphylococcal nuclease created artificially by Shortle and co-workers [139–141], 257 are destabilizing, five lead to stabilities essentially the same as that of the wild-type (approx. 5.5 kcal mol^{-1}), only 28 are stabilizing. Moreover, each destabilizing artificial mutation destabilizes by more than 2.08 kcal mol^{-1} on average (maximum = 7.5 kcal mol^{-1}), whereas each stabilizing artificial mutation stabilizes by only 0.36 kcal mol^{-1} on average (maximum = 1.0 kcal mol^{-1}). A similar trend is exhibited by the 98 artificial mutants of chymotrypsin inhibitor 2 studied by Fersht and co-workers [142] (77 with a single substitution, 17 with two substitutions, and four with three substitutions): 90 artificial mutants are less stable than the wild-type (7.6 kcal mol^{-1}), only eight artificial mutants are more stable than the wild-type. On average, a destabilizing mutation destabilizes by 1.67 kcal mol^{-1} (maximum = 4.93 kcal mol^{-1} among single-substitution mutants), whereas a stabilizing mutation stabilizes by only 0.18 kcal mol^{-1} (maximum = 0.42 kcal mol^{-1}). These data suggest that the stabilities of natural proteins are close to, albeit not exactly at, the maximum achievable by sequences in the immediate sequence-space neighbourhood of the wild-type sequence. However, when larger numbers of amino acid substitutions are applied to a wild-type, an increase in thermodynamic and/or kinetic stability of 3–4 kcal mol^{-1} has been observed in several proteins (e.g. [143,144]).

There is no experimental evidence to date indicating the existence of polypeptides that encode for an essentially unique folded structure with native stability as high as approximately 0.4 kcal per mole of amino acid residues as posited by Goldstein [136]. A case in point is the 93-residue designed protein Top7, which is already characterized as extremely stable. Its stability is approximately 13 kcal mol^{-1} at 25°C [145]. Although this level of native stability is significantly higher than several single-domain proteins [146] including the 97-residue S6 with similar secondary structure (native stability = 8.5 kcal mol^{-1}) [147], the stability of the artificially designed Top7 is still within the 5–15 kcal mol^{-1} native stability range long recognized for natural proteins [44,131]. The highest stability achieved by more recent attempts to design stable proteins is 14.9 kcal mol^{-1}, or 0.14 kcal per mole of amino acid residues for a 110-residue construct [148]. In this light, the 118 kcal mol^{-1} stability estimated in [136] is physically unrealistic. This exceedingly high estimate is probably an artefact of the non-explicit-chain approach used in the study (for a discussion of explicit- versus non-explicit-chain protein models, see [36]), which tends to underestimate mutational effects on the unfolded states. From a protein biophysics standpoint, however, any given mutation not only impacts the free energy of the native state but can also have a significant effect on the denatured (unfolded) state, and the effects on the two states often partially cancel, such that extremely high native stability is physically not possible [149].

2.4.3. Reconciling evolutionary selection for stability with marginal stability

Taken together, the above discussion indicates that fundamentally, natural globular proteins without disulfide and other cross-links are marginally stable because of the physical constraints on native stability itself. Exceedingly high native stability is physically impossible. Because there are more sequences encoding for lower stabilities than higher stabilities [135,137], extensive evolutionary selection to decrease native stability is not necessary, though selection for local flexibility may sometimes result in functional globular proteins that are not the most stable possible for the given folds [21,133,134]. Experimental evidence abounds, however, for evolutionary selection for higher native stability [40,83] (see §2.1), though not necessarily the highest once a certain threshold for function is achieved [150], as illustrated by the data on the artificial mutants of staphylococcal nuclease and chymotrypsin inhibitor 2 discussed in §2.4.2. Therefore, natural globular proteins are marginally stable (because of biophysical constraints) but they are nonetheless nearly maximally stable (by evolution) for the structures they fold to. This conclusion is supported by theory: neutral net topology in protein sequence space tends to concentrate large evolving populations toward sequences that are mutationally most robust [137,151]. These sequences are often also thermodynamically most stable [137]. But random mutations alone—in the absence of a fitness drive towards higher native stability—are not sufficient to produce a highly concentrated population at the most stable 'prototype' sequence at the

bottom of the sequence-space superfunnel because of the large number of sequences that are less stable [40,137,152,153]. Therefore, the experimental observation that natural proteins are often a nearly most stable sequence that behaves like a prototype sequence suggests strongly that they are results of positive selection for higher native stability (see further discussion in §3.2.3).

2.5. Geometric/topological constraints imposed by the native structure

The Top7 example mentioned in §2.4.2 also offers insights into other aspects of the interplay between biophysical constraints and evolution. It shows that the tendency to misfold does not necessarily diminish with increasing native stability: despite the high native stability of Top7, its folding kinetics is complex, probably involving multiple kinetic traps [154,155]. Theoretical considerations indicate that the lack of two-state-like behaviour of Top7 is probably caused more fundamentally by its peculiar native structure, more so than the fact that it is an artificially designed protein that did not undergo natural selection [156]. Thus, native geometry or topology (the pattern of residue–residue contacts in the native structure) probably impose a physical constraint on the level of stability and folding cooperativity that natural or artificial selection can achieve [156,157]. In this connection, it has been shown using simple lattice protein models that not all protein structures are equally encodable [158] or designable [159–161]. Some structures may not be encodable at all [158,162]. This represents another set of biophysical constraints under which protein evolution must operate.

2.6. Chaperones and *in vivo* folding

In molecular biology, chaperones are a class of proteins that assist the folding and assembly of other proteins, or even reverse misfolding [163]. Many mutant proteins fail to fold or be expressed in the cell because of reduced native stability, increased probability of misinteractions during folding, or other changes in folding kinetics that are detrimental to productive folding. These biophysical constraints hinder evolution because they limit the number of mutants that can be explored. Mutations that decrease native stability below a certain threshold cannot participate in the evolutionary process even if they possess superior functionality—provided they are properly folded—because relative native instability compromises protein folding and expression. In the cellular environment, chaperones offer a degree of relief from these constraints. Molecular chaperones enhance evolvability—i.e. a genome's ability to produce adaptive variants [164] (see §3.5)—because they help mutants that are less stable to fold to functional structures and to avoid non-functional aggregation, thus allowing more mutants with potentially beneficial new functions to be explored *in vivo* [165,166].

This principle was borne out in experiments involving the *Escherichia coli* GroEL/GroES chaperonin complex. In a set of laboratory evolution experiments on four enzymes, the divergence of modified enzymatic specificity was found to be much more speedy when GroEL/GroES is overexpressed, most probably because GroEL/GroES assist folding of enzyme variants, allowing mutants that lose as much as 3.5 kcal mol^{-1} in native stability to be viable whereas only approximately 1 kcal mol^{-1} loss in stability is permitted in the absence of GroEL/GroES [165]. In a more recent experiment to evolve a phosphotriesterase into an arylesterase *in vivo*, GroEL/GroES is again seen to increase the ability to adapt to new functions by allowing for more genetic variation. Moreover, it was found that mutational tolerance is not determined by *in vitro* native stability per se, but rather by the level of soluble expression of the mutant protein in the cell. In this case, the GroEL/GroES chaperone enhances soluble expression by apparently stabilizing a folding intermediate against detrimental aggregation and thus indirectly promotes productive folding, underscoring the critical importance of mutants' *in vivo* folding kinetics on the course of protein evolution [166].

Consistent with this trend, there is also strong evidence at the genome level that proteins that use GroEL/GroES obligately for folding evolve faster [167] and are less dependent on optimal codon usage to avoid translation-induced misfolding [168] than proteins that do not require these chaperones for folding but rather rely more on optimal codons [169]. The link between translation errors and evolutionary rates will be discussed further in §3.10.

On the theory front, a recent simulation of protein evolution considered a model cell containing a few interacting protein species that can adopt either 'folded' or 'molten-globule' structures [170]. Consistent with the trend seen in experiments [165,166], the simulation indicated that chaperones that actively catalyse folding also accelerate evolutionary adaptation because the increased chaperone-assisted folding rates allow for deeper searches of the sequence space [170].

2.7. Multi-basin folding landscapes, allostery and conformational dynamics

Protein structures are dynamic; and conformational dynamics is crucial in many biomolecular interactions [171,172]. Even for globular proteins that fold to an essentially unique native structure under physiological conditions, other less favourable 'excited-state' conformations are always populated, albeit to a much lesser extent than the dominant native conformation that is commonly identified as the ground-state structure. The balance between the dominant ground-state and excited-state populations can be altered by mutations. For instance, a recent NMR experiment demonstrated that a mutant T4 lysozyme populates an excited state to about 3% at 25°C [173] (figure 4*a*).

Besides uniquely folding proteins, there are globular proteins that have more than one dominant folded conformation. For these proteins, the same amino acid sequence adopts more than one structure with similarly high probabilities. Thus, instead of a single funnel, the energy landscape of such a protein has multiple basins of attraction [177,178]. In some cases, these alternative structures freely interconvert during the lifetime of the protein as for the cytokine lymphotactin [179] and the cell cycle control protein Mad2 [180]. Sometimes it takes an additional factor to stabilize an alternative structure, such as a change in the solvent conditions or a binding event (e.g. [181–184]).

2.7.1. Conformational diversity is often needed for function

Multi-basin energy landscapes are widely used by Nature to regulate protein function. A prime example is allostery, by which the function of a protein is regulated through binding a ligand (effector) at a site (the allosteric site) on the protein

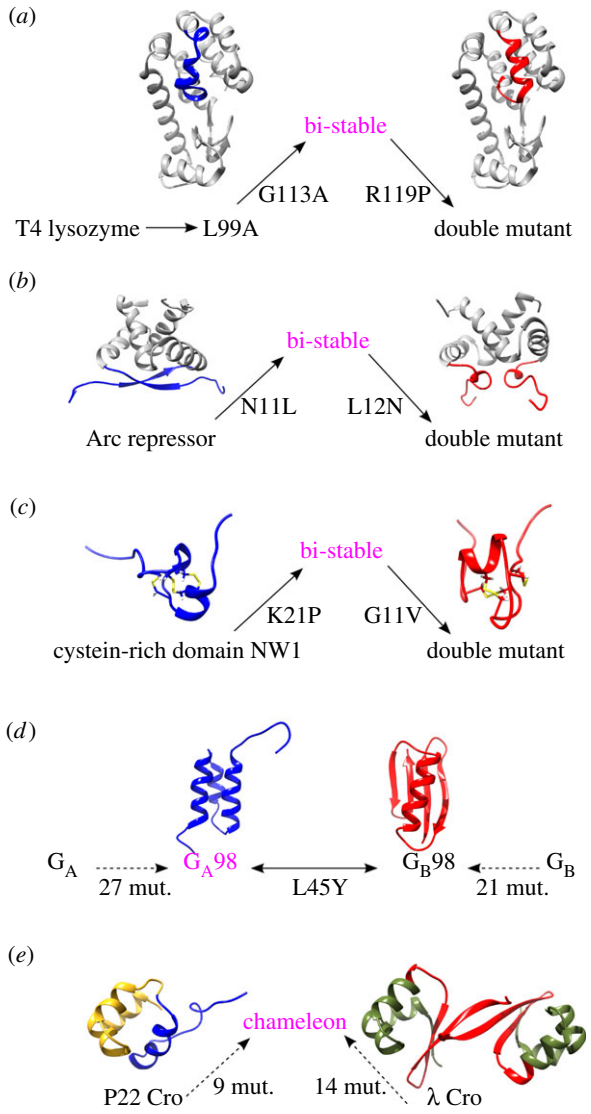

Figure 4. (*Caption opposite.*)

Figure 4. (*Opposite.*) Examples of experimentally designed bi-stable proteins and mutation-induced structural switches. (*a*) Wild-type T4 lysozyme was mutated (L99A) to create an internal cavity that allows for the population of an excited-state conformation with an altered helical segment (blue; left). This T4 lysozyme variant could be further transformed via a single G113A substitution into a bi-stable protein that also populates a new folded structure in which the local structure of the helical segment is modified (red; right). An additional R119P substitution on this L99A, G113A variant then leads to a protein that adopts the conformation on the right as its essentially unique native structure [173]. (*b*) Wild-type Arc repressor is a homo-dimeric protein. Each monomeric unit contributes a β-strand to form a two-stranded antiparallel β-sheet (blue; left). This shared configuration becomes bi-stable with the introduction of a single N11L substitution to each of the monomeric units. The mutated sequence now populates the original structure as well as a new structure with the β-strands changed into two short helices (red; right). An additional L12N substitution on each of the monomeric units results in a sequence that adopts the new configuration on the right as its essentially unique native structure [65]. (*c*) The cysteine-rich domain NW1 forms a stable structural element (blue; left) with three disulfide bonds (yellow sticks) between the residue pairs (8,20), (12,25) and (16,24). A single K21P substitution results in a bi-stable mutant that also populates a structure with a different overall conformation (red, right) and an alternate disulfide-bonding pattern, now between residue pairs (8,24), (12,20) and (16,25). Introduction of a single G11V substitution on this bi-stable mutant results in a sequence that adopts the conformation on the right as its essentially unique native structure [174]. (*d*) Two domains of streptococcal Protein B, named G_A and G_B, with a 3α and a 4β + α-fold, respectively, and no significant sequence similarity, were transformed into each other by a series of point mutations that resulted in a structure pair G_A98 and G_B98 that allows the switch between the two structures with just a single L45Y mutation. G_A98 exhibits a small 4β + α population and thus may also be regarded as bi-stable [175]. (*e*) The viral P22 Cro and λ Cro are DNA-binding proteins. Encoded by different sequences, they have structurally very similar helical N-terminal domains (represented by the yellow and green ribbon, respectively) but have structurally distinct C-terminal domains. P22 Cro has a helical C-terminal, whereas the C-terminal of the homo-dimeric λ Cro forms a β-sheet. A 24-residue chameleon sequence created largely by mixing residues from the helical and sheet-forming C-termini adopts different secondary structure depending on whether it is inserted in the P22 or λ context [176]. Sequence and structural information presented in this figure was taken from the cited original references.

that changes the structure and/or dynamics of the protein's active, functional site positioned at a distance from the allosteric site [185]. Allostery is important for biological function and its malfunction is implicated in disease processes [186,187]. Mutations affect allostery. Mutational effects on allostery can be subtle because allosteric communication between the allosteric and active sites can be underpinned by multiple mechanisms [188,189]. Nonetheless, mutational effects on allostery can be rationalized by computational approaches in some instances [190].

Conformational flexibility, dynamics of protein folded states and allosteric transitions often can be deduced to a reasonable degree from the structure(s) of the protein in question using elastic network models for folded-state dynamics [191–194] or native-centric Gō-like potentials [195] with multiple folding basins ([196]; reviewed in [177]). Similar to the aforementioned case for the probable existence of geometric/topological constraints on the evolution of folding stability and cooperativity (§2.5), the success of structure-based native-centric modelling in rationalizing conformational dynamics and allosteric transitions suggests that there are significant structural constraints on the evolution of functional folded-state dynamics. The computational efficiency of elastic network models also allows enzymes that are dissimilar in sequence and structure yet probably perform similar functions to be detected by their similar dynamic properties [194,197], making it possible for relationships between evolutionary conservation and conformational dynamics to be explored [198].

Allostery is envisioned to have evolved by oligomerization, gene fusion and/or recruitment of unused/flexible parts of a pre-existing protein structure (reviewed in [199]). The latter evolutionary route may proceed by positive selection of opportunistic binding of excited-state conformations. The mechanism of such binding may lie anywhere between the 'conformational selection' and 'induced fit' scenarios [177,200]. Evolution has apparently exploited latent allosteric potentials entailed by conformational dynamics in this manner, as in the case of Ste5 activators that target MAP kinases in yeast [201].

Opportunistic binding of excited-state conformations can also facilitate evolution of new functions that are not necessarily allosteric [122,202,203]. During such an evolutionary process, a sequence with a multi-basin energy landscape can serve as an evolutionary bridge. In particular, the evolutionary intermediate of two sequences each encoding for a

different dominant structure can be a bi-stable sequence that folds to both structures with equal or similar probabilities [161,178,204].

2.7.2. Bi-stable proteins and conformational switches

Experiments in several laboratories have found cases where a single mutation was able to either create a bi-stable protein from a uniquely folding protein or completely switch one uniquely folding protein to another with a new native structure [65,173–175,205,206]. Although these cases of mutation-induced structure switches were artificially engineered, they demonstrated that it is generally possible for bi-stable proteins to arise through mutations during natural evolution.

An early example of mutation-induced structure switching was the Arc repressor, which is a homodimer with a two-stranded inter-unit β-sheet. Experiments by Cordes *et al*. [205] showed that the β-sheet in the wild-type protein can be changed to a pair of 3_{10}-helices by two amino acid substitutions that swap the neighbouring sequence positions of an asparagine and a leucine. A subsequent experiment indicated that a mutant with a single asparagine-to-leucine substitution has approximately equal populations of the β-sheet and helical forms, and thus may be regarded as an evolutionary bridge [65] (figure 4*b*). A recent study showed further that if two more polar or charged to hydrophobic substitutions are introduced, the resulting triple mutant adopts an octamer configuration with approximately half the helical content of wild-type Arc, indicating that new protein–protein interactions and novel oligomeric states can readily result from a small number of mutations [207].

Experimental mutagenesis has uncovered a similar behaviour in the cysteine-rich domains (CRD) of cnidarian nematocyst proteins. Different CRDs fold to either one of two structures with different disulfide-bonding arrangements despite high sequence similarity and identical sequence patterns for their cysteines. Meier *et al*. [174] found that a CRD sequence that folds to one disulfide arrangement can be converted to another disulfide arrangement by only two amino acid substitutions, one from lysine to proline and the other from glycine to valine, whereas the single-substitution mutant with only the lysine-to-proline mutation behaves as an evolutionary bridge that populates both disulfide arrangements (figure 4*c*). This finding again underscores that large structural changes can be effected by minimal changes in the amino acid sequence.

The study by Alexander *et al*. [175] of the G_A/G_B system showed that a single leucine-to-tyrosine substitution can convert a sequence encoding for an albumin-binding 3α (G_A) structure to a sequence encoding for an immunoglobulin-binding 4β + α (G_B) structure (figure 4*d*). A subsequent experiment on two other mutants identified two additional 3α ↔ 4β + α structure switches induced by a single amino acid substitution [206]. Interestingly, a mutant with a conformational ensemble that is 95% 3α and only 5% 4β + α when measured in isolation nevertheless binds immunoglobulin but not albumin [206], providing an excellent example of how protein–protein interactions can dramatically shift the conformational distributions of the binding partners [200].

Another recent example of an artificial 'evolutionary intermediate' is a 24-residue sequence that can adopt either the α-helical or β-sheet C-terminal conformations, respectively, of transcription factors P22 Cro and λ Cro, depending on whether the designed sequence is fused with the N-terminal domain of P22 Cro or λ Cro [176] (figure 4*e*). In this case, the naturally occurring wild-type 24-residue C-terminal sequences of P22 Cro and λ Cro have only five identical amino acid positions, whereas the amino acid residues of the designed sequence at all but four positions are either identical to that in the wild-type P22 Cro or in the wild-type λ Cro. This finding underscores the critical role of tertiary context in determining secondary structure in proteins [208]. Although the designed sequence is nine and 14 substitutions away from the corresponding sequences in wild-type P22 Cro and λ Cro, respectively, the successful design of a structurally ambivalent 'chameleon' sequence in this experiment suggests that a smooth evolution transition from one Cro fold to another is possible [176].

Computation-assisted design of conformational switches has seen notable success [209,210]; but it is still a challenge to apply our current biophysical knowledge to provide a fundamental physical rationalization for experimentally observed conformational switching. For the G_A/G_B system, a mutation-induced gradual stabilization of one structure over another was demonstrated using a common software for $\Delta\Delta G$ prediction (§2.1) [178]. However, the mutation-induced G_A/G_B conformational switching was not reproduced in atomistic molecular dynamics simulations [67], even though a part of the simulated energetics is consistent with experiment [68].

The structural plasticity in bi-stable and multi-stable proteins probably plays an important role in protein evolution [122,211,212]. Conformational switches and bridge sequences facilitate evolution by allowing continuous or near-continuous transition from one folded structure to another. The experiments in figure 4 suggest that, under certain circumstances, multi-functional proteins can be created by only a few mutations that stabilize certain hidden or excited states. A situation where it is advantageous to take such a route is the coevolution of pathogens and their hosts, a highly competitive evolutionary process that demands frequent change of protein shapes and functions. It is thus unsurprising that bi-stability and multi-specificity are exhibited in antibodies [213], antimicrobial peptides in natural plant defence [214] and antiviral proteins [215].

2.8. Intrinsic disorder

When structural plasticity is extreme, one might expect a multi-stable sequence to morph into one without a discrete set of clearly discernible favoured conformations. This in itself is not surprising because an overwhelming majority of polypeptides with random amino acid sequences do not fold to a unique structure [116]. What is remarkable, in the context of our decades-long near-exclusive focus on proteins with well-ordered structures, is the existence of many functional proteins with such extreme conformational diversity. Although our main concern here is evolution of globular proteins, it is important to recognize that intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) play key roles in cellular processes [216–222].

2.8.1. Any protein conformational state can potentially have biological function

With the discovery of functional IDPs/IDRs, it has become abundantly clear that biology can exploit any protein conformational state that it finds useful. In this respect, an

intriguing recent suggestion is that although avoidance of amyloid-like aggregation has apparently been a driving force of protein evolution [223] (§2.2), it is possible that modern protein folds have an amyloid origin in evolution [224]. For IDPs/IDRs, current understanding of the evolution of the triplet genetic code [225] suggests that the amino acid composition of primordial polypeptides was conducive to more disordered conformations before the modern genetic code for a 20-letter amino acid alphabet was completed [222]. However, surveys of modern proteomes indicate that IDPs/IDRs are more common in eukaryotes than in prokaryotes: more than 32% of amino acid residues in eukaryotic proteins are in IDPs/IDRs whereas the corresponding percentage is less than 27% for prokaryotic proteins. This pattern suggests that the proteins in the last universal ancestor were probably well structured and emergence of the IDPs/IDRs observed today was relatively late [222], perhaps coinciding with an evolutionary trend that has witnessed a general decrease in protein hydrophobicity [226].

According to one estimate, more than 30% of eukaryotic proteins have IDRs of more than 50 consecutive residues [216], consisting of more proline, glycine and charged residues but fewer hydrophobic residues [227,228]. IDPs/IDRs are involved in fundamental processes such as transcription, translation and cell cycle regulation that, when they malfunction, can lead to cancer. The essential role of IDPs/IDRs in mediating biological regulation suggests that, in some situations, they have certain advantages over folded proteins in recognition and binding [229]. For instance, their ability to flexibly bind to many different partners has allowed them to occupy hub-like roles in protein–protein interaction networks [230,231]. They can also encode relatively larger intermolecular interfaces to economize genome and cell sizes [218]. Protein–protein interactions for some IDPs/IDRs entail significant folding upon binding [219], while others undergo only restricted local ordering at the binding site with other parts of the protein remaining disordered, thus forming a dynamic 'fuzzy' complex [220,232–236].

2.8.2. Biophysical constraints on evolution of intrinsically disordered proteins and regions

What can be expected of the biophysical constraints on the evolution of IDPs/IDRs? IDPs/IDRs do not fold to a unique structure. Therefore, in contrast to many globular proteins, the energy landscapes of IDPs/IDRs are not funnel-like [222]. As far as near-neutral mutations [237,238] are concerned, one might expect less biophysical constraints on IDP/IDR evolution than on globular protein evolution because for IDPs/IDRs there is no need to maintain an essentially unique folded structure. However, it can also be argued that evolution of certain IDPs/IDRs may be subject to even more restrictive constraints because of their requirement to bind to multiple partners. As a result, these IDP/IDRs may suffer from low mutational robustness similar to that of bi-stable globular proteins that play the role of an evolutionary bridge between two folded structures [178,239]. Nevertheless, even in such cases, IDPs/IDRs in a neutral net might only need to conserve certain functional residues that are compatible with multiple binding partners while imposing few constraints on mutations at amino acid sites in the rest of the protein.

These expectations are largely consistent with database studies and experiments. Phylogenetic analyses indicate that IDRs generally evolved faster than ordered regions of proteins, but some IDRs such as DNA-binding regions evolved slower [240,241]. For proteins that have both ordered and disordered regions, mutations in IDRs lead to smaller stability changes than in ordered regions. Thus, IDPs/IDRs may enhance protein evolvability and the development of new functions [242], as evolutionary changes in protein sequence and structure are often correlated with local flexibility and disorder [243].

The biophysical constraints on IDP/IDR evolution [244] are quite different from those on folded protein evolution [12]. In fact, the accepted amino acid substitutions in IDPs/IDRs resemble those in solvent-exposed loops and turns of globular proteins [244]. Chemical composition defined as the fraction of positive, negative, polar, hydrophobic and special (proline and glycine) residues is often maintained across IDR orthologues that otherwise exhibit little conservation [245]. This observation is in line with the finding that whether an IDP is elastomeric or amyloidic depends largely on the relative compositions of proline and glycine [228], and is consistent with the central role of aromatic composition in a set of IDP interactions that are presumably underpinned by cation–π attraction [236]. Relative to the substitution matrices for globular proteins, substitution matrices for IDPs/IDRs entail a generally higher probability of evolutionary changes, but some residues such as tryptophan and tyrosine tend to be highly conserved in IDPs/IDRs, perhaps because of their critical role in protein–protein interfaces [244,246].

It should be recognized that IDP/IDR conformations are far from random. Biological functions of proteins are always underpinned by conformational structures. In this respect, the difference between IDPs/IDRs and ordered proteins is that the IDP/IDR function is conferred by a much more diverse conformational ensemble than for globular proteins. The transient, 'fuzzy' tertiary contacts in IDP/IDR conformations are often important for the function; hence mutations that disrupt such contacts can be extremely detrimental to function. An example of how a single mutation can disrupt IDP function is the threonine-to-arginine mutation at position 45 of the cyclin-dependent kinase inhibitor Sic1 [234,247]. This amino acid substitution leads to a dramatic increase in its hydrodynamic radius [234] and, at the same time, a serious disruption of its biological function in regulating the cell cycle [247]. Current biophysical understanding of this and other mutational effects on IDP/IDR conformational distribution is limited. Much remains to be discovered about the evolution of these proteins.

2.9. Protein dynamics and phenotypic plasticity: what is a molecular phenotype?

In the study of molecular evolution, the term genotype is used for the inheritable part of genetic information; whereas phenotype refers to the biomolecules of interest that are produced based on the genotypic information. In theoretical studies of protein evolution, as a *modelling simplification*, the genotype may be identified with the amino acid sequence because as far as *in vitro* protein folding is concerned, it contains essentially the same information as the nucleic acid sequence that encodes it. This is a simplified approach that neglects *in vivo* complexities such as the fact that synonymous mutations can lead to altered cellular folding pathways (§2.2). In principle, the molecular phenotype should encompass all

properties—including but not limited to biological functions—of the protein encoded by the genotype. In practice, molecular phenotypes in theoretical and experimental investigations are defined, and thus are restricted, by the question being addressed. However, an oversimplified view of molecular phenotypes that is too restrictive can hinder understanding of important principles of protein evolution.

For globular proteins that have an essentially unique folded structure, a practical and seemingly natural definition of molecular phenotype of a given amino acid sequence is its structure as deposited in the Protein Data Bank (PDB). This practice is useful for constructing a neutral net of sequences that encode uniquely for the same protein structure and the evolution from one such phenotype to another [137,161]. However, this simplistic view of molecular phenotype neglects the dynamic nature of proteins. Recent advances in experimental techniques, especially those using NMR, have enabled detailed characterizations of the dynamic properties of proteins [173,248–250] and, in conjunction with computation, allowed for the construction of ensembles of diverse conformations of disordered proteins based on NMR and other experimental measurements [251,252]. As a result of these experimental advances and the theoretical energy landscape perspective [34,253,254], our view of how protein molecules function has undergone a drastic change in the past two decades, with increasing recognition of the biophysical, biological and evolutionary significance of protein dynamics [255,256].

Because of the role of dynamics in protein function (§2.7 and 2.8), identifying a protein's molecular phenotype only with its native folded structure is often too restrictive. Ideally, the molecular phenotype of an amino acid sequence should correspond to the totality of its biologically relevant properties. Although it may not be practical to enumerate many properties of a protein, for many applications the molecular genotype should at least be understood as an ensemble of conformations with a sequence-specific and environment-dependent distribution. Within this ensemble, certain phenotypic properties, such as the presence of a secondary structure in the protein conformation, are not necessarily fixed but can undergo thermal fluctuations or environment-induced changes. This phenomenon is referred to as single-genotype phenotypic fluctuation or phenotypic plasticity, which can underpin important evolutionary responses to environmental changes [257].

Phenotypic plasticity tends to enhance evolvability. This trend can be seen clearly in an experimental evolution study of *E. coli* cells that express mutants of green fluorescence protein. In this experiment, mutants leading to a larger fluctuation in fluorescence among cells containing the same green fluorescence protein gene were found to exhibit a higher rate of evolution [258]. A positive correlation between single-genotype phenotypic fluctuation and evolvability has also been rationalized recently by computational models of proteins [259] and RNA [260]. As mentioned in §2.7.2 above, plasticity of molecular phenotype (i.e. conformational diversity) is generally conducive to higher evolution rates that can be beneficial to organisms in rapidly changing environments [261,262] through selection of 'moonlighting' [263] promiscuous functions [264]. We will elaborate below on how the relatively new view of protein dynamics and conformational distribution [202,203,265] has enriched our understanding of evolution.

3. Applications of biophysics-based models to understand protein evolution

To gain insights into how proteins evolve, various models of the mapping from protein sequence (genotype) space to structure (phenotype) space have been constructed. Here, we focus largely on models with an explicit representation of the protein chain and biophysics-based interactions because these models provide a better delineation of what is physically plausible for real proteins than other models of molecular evolution that postulate such a mapping in the absence of or with little biophysical considerations.

3.1. Protein sequence and structure spaces: evolution meets biophysics

Recent genomic and proteomic initiatives have greatly advanced our knowledge of the global sequence–structure relationship and the evolution of the primary, secondary, supersecondary, domain, tertiary and quaternary/complex structures of proteins [266–269]. An earlier study suggested that α/β folds appeared later in evolution than other structural classes [266]. However, more recent investigations indicate that the globular protein architectures observed today had emerged during evolution roughly in the following order: α/β (e.g. TIM-barrel), $\alpha+\beta$, all-α, all-β, then multi-domain proteins [268]. Consistent with this perspective, a core of functional diversity corresponding mostly to the more ancient α/β folds in the protein structure space has been identified [270]. Interestingly, a computer modelling study of the dynamic properties of protein fold space suggests that α/β folds are also more stable than other fold classes [271].

In the space of all possible amino acid sequences, an overwhelming majority of sequences do not have a biological function. It has been estimated that the probability of finding a functional protein among random amino acid sequences is approximately 10^{-11} [272]. Evolutionarily, natural protein sequences are still diverging from one another today, albeit at a slow rate because biophysical and functional constraints allow only about 2% of amino acid sites to be mutated [273]. This ongoing divergence means that the coverage of the space of all possible sequences by biologically viable sequences has been and still is increasing; i.e., there has been a continuing expansion of the 'protein universe' since the beginning of life on the Earth [273].

There is consensus among researchers that the repertoire of globular protein folds is probably finite. The SCOP classification [274] currently identifies about 1200 different folds in the PDB. Estimates for the total number of possible folds range from about 1000 [275], 2000 [276] to 10 000 [277]. Fold classifications are inherently difficult and estimation of the total number of possible folds is sensitive to the definition of a fold (see [278] and references therein). Nonetheless, in most cases, structures of recently sequenced proteins are related to known folds [279], suggesting that the existing PDB structures are probably a near-complete representation of all biologically viable globular protein folds.

A broader issue is whether the observed natural globular protein folds constitute a relatively small subset selected by evolution from a much larger collection of all physically possible compact conformations. Biophysics and polymer physics

have shed light on this question. A hallmark of most globular proteins is their helical and/or sheet-like organization. These secondary structures facilitate backbone–backbone hydrogen bonding in the folded protein core (reviewed in [12]). Secondary structures are conducive to tight tertiary packing as well. It has been shown that secondary-structure-like chain organization is enhanced by conformational compactness [280–282], but in the absence of hydrogen bonding such structures exhibit deviations from sharply defined α-helices and β-sheets [283–285]. These findings suggest that biophysical constraints of conformational compactness in conjunction with hydrogen bonding can go a long way in accounting for the basic architecture of globular protein folds.

A more recent study using computational sampling of homopolypeptide conformations suggested further that the current repertoire of globular protein folds is nearly complete in its coverage of all physically possible compact folds [286]. However, studies by three other groups have found instead that the current fold repertoire represents only a small fraction of all possible folds [287–289]. In particular, an investigation of the compact conformations of 60-residue polyvaline concluded that known protein folds constitute only approximately 5% of all physically possible folds, and that on average the natural folds have more local intrachain contacts (i.e. lower contact orders [146]) than the set of all possible folds, suggesting an evolutionary preference for structures with lower contact orders [288]. In response, a recent study argued that inasmuch as an appropriate criterion for matching simulated compact conformations and natural folds is applied, the existing library of single-domain PDB structures is probably complete in covering all physically possible folds [290]. A separate study by the same group indicated that computer-generated compact conformations tend to contain cavities resembling binding pockets in natural proteins as well, even in the absence of selection, suggesting that 'many features of biochemical function arise from the physical properties of proteins that evolution likely fine-tunes to achieve specificity' [291]. While the degree to which evolution has shaped the space of known protein folds remains to be further elucidated, the investigative effort described above is an excellent illustration of how explicit-chain biophysical models can be harnessed to address fundamental questions in protein evolution.

3.2. Simple exact models and other explicit-chain coarse-grained models of protein evolution

Sequence and structure spaces of proteins are vast. Coarse-grained explicit-chain models are valuable tools in the study of protein evolution. Currently, modelling the sequence–structure mapping by energetically and structurally high-resolution representations is often not practical, especially for addressing large-scale evolutionary changes involving many different protein folds. In this regard, lattice models—wherein conformations of model proteins are configured on two- or three-dimensional (2D or 3D) lattices—are particularly useful because of their computational tractability [292–294]. In view of their historical and current utility for investigating fundamental evolutionary issues (see, e.g., recent applications of 2D lattice models to study the basis of homology modelling [295], adaptive conflict [239] (§3.6) and long-term survivability [296]), a detailed assessment of the biophysical foundation of these models is in order.

3.2.1. Conformational enumeration

Among lattice models of protein evolution, simple exact models allow for exhaustive enumeration of all possible sequences and structures in the model [292]. These models include the two-letter 2D hydrophobic-polar (HP) model that uses a reduced alphabet consisting of only two types of residues, hydrophobic (H) and polar (P), to capture the prominent effects of hydrophobic interactions in protein energetics [137,161,162,297–299] (figure 5), two-letter variants of the 2D HP model (e.g. [293,301,302]), and a four-letter 2D model that also includes residues behaving somewhat like positive and negative charges [303]. Some 20-letter 2D models may also be considered as exact, because they consider all possible mutations in the immediate sequence-space neighbourhood of a given sequence [40,153].

Other types of lattice models have been used to study protein evolution as well. These models either restrict chain conformations to be maximally compact so as to allow model proteins with longer chain lengths to be studied [152,160,304], or rely to various degrees on sequence-space sampling instead of considering all possible sequences because of their usage of a 20-letter amino acid alphabet that entails many more sequences (e.g. [40,153,305]) or both (e.g. [306,307]). Restricting model structure space to maximally compact conformations in 2D [306] or 3D [307,308] reduces computation drastically because such conformations constitute only a tiny fraction of all possible conformations [280,309–312]. For instance, whereas the total number of all distinguishable conformations (not related by rotations and reflections) for a chain with 25 residues configured on the 2D square lattice is 5 768 299 665 [313], the number of maximally compact 25-residue conformations restricted to a 5×5 square, as considered in the study of Taverna & Goldstein [306], is only 1081 [309]. In 3D, the total number of all possible conformations (not related by rotations and inversions) on the simple cubic lattice for a chain with 27 residues is 11 447 808 041 780 409 [313,314], but the number of maximally compact 27-residue conformations restricted to a $3 \times 3 \times 3$ cube, as considered by Deeds & Shakhnovich [307], is only 103 346 [280].

However, from a biophysical standpoint, it is important to keep in mind that real protein chains are not restricted to be maximally compact. Although behavioural trends predicted by models that use only maximally compact conformations may sometimes correlate with models that consider the full conformational ensemble, significant distortions of protein folding energetics are introduced by this approach [158,292]. In particular, for a given model sequence with a physically plausible interaction, the lowest-energy structure among maximally compact conformations may not be the true ground state structure, which is often less than maximally compact [299,315,316].

3.2.2. Model interactions and their biophysical basis

We next consider the physicality of the model interaction potential. Several examples presented below to illustrate recent biophysical insights into basic principles of protein evolution are based on the 2D HP model (figure 5). We choose the 2D HP model for this purpose because owing to

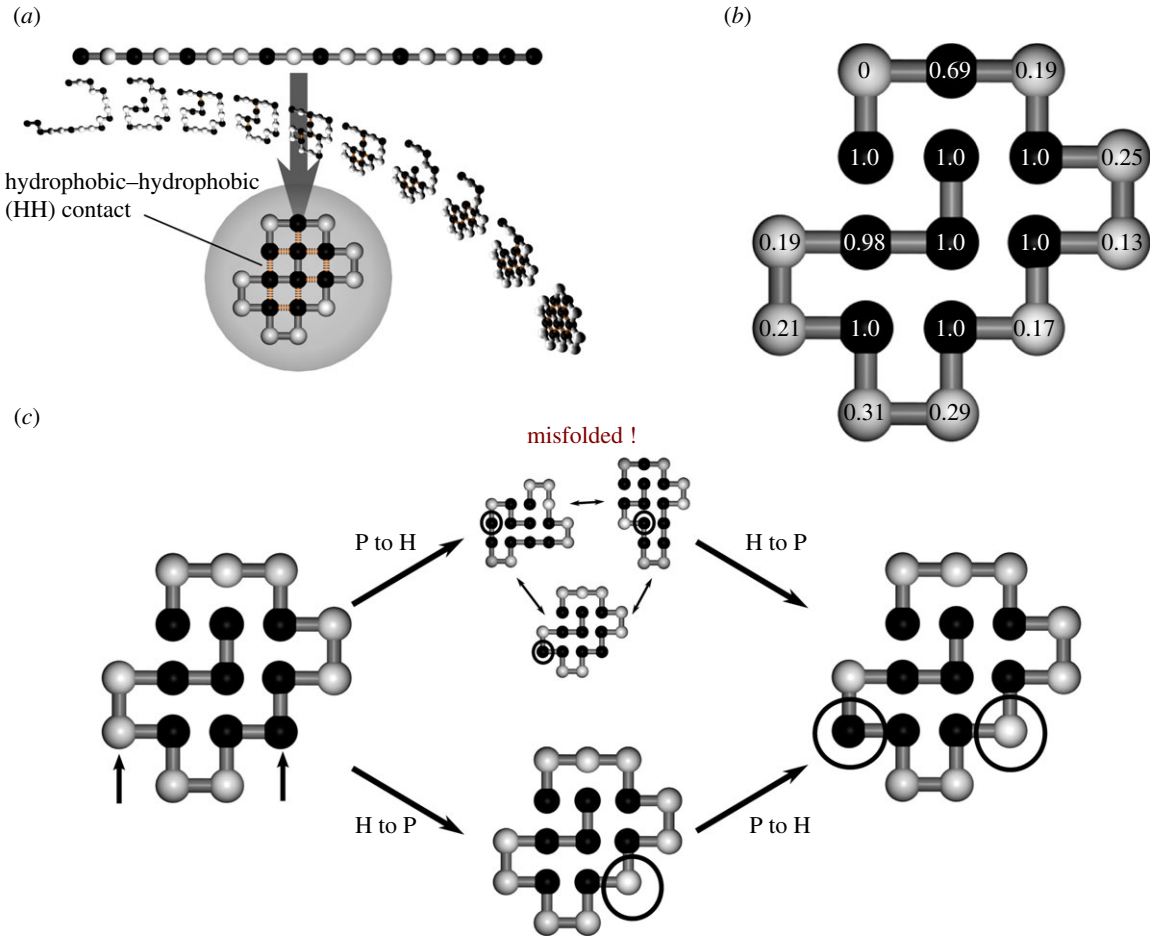

Figure 5. The simple exact 2D HP lattice model is a useful tool for studying evolution across entire sequence and structure spaces. (*a*) An example HP model sequence of length 18. Hydrophobic (H) and polar (P) residues are depicted, respectively, as black and grey beads. A favourable energy is assigned to each hydrophobic–hydrophobic (HH) contact (as indicated by the orange connections between two black beads), other contact types are neutral (carry a zero energy). The total energy of a conformation is proportional to the number of HH contacts. For a given sequence, the energy of every conformation can be computed accordingly. The schematic drawing below the sequence shows the folding funnel of the sequence. Conformations with more HH contacts are placed lower, because they are energetically more favourable than conformations with fewer HH contacts. The conformation encased in the grey circle is the lowest-energy (native) conformation of the given sequence. (*b*) Conserved and variable sites in an HP model structure. Among all the HP sequences that fold to this structure, the sequence shown here is the one that provides the highest native stability. The number on each bead is the relative frequency of occurrence of an H residue at the given sequence position among the entire neutral set of 48 HP sequences encoding for this structure [137]. Most core hydrophobic positions cannot be mutated (1.0H, i.e. 100% H) without losing the native structure; one surface position must be polar (0.0H, i.e. 100% P); but most other surface positions could be mutated to either polar or hydrophobic. This means that a hydrophobic-to-polar substitution has a very different structural effect depending on its location (surface or core). (*c*) Epistasis in the 2D HP lattice model. The sequence on the right is a double mutant of the sequence on the left (mutated positions indicated by circles and arrows, respectively). As for a real protein, here the first mutation could occur in either of the two positions to produce a single-substitution intermediate sequence. One of the intermediate sequences is viable. It folds uniquely to the same structure (lower middle drawing in (*c*)), whereas the other intermediate sequence is misfolded: in addition to the original structure, this sequence adopts two other equally likely lowest-energy conformations as shown in the upper middle drawing in (*c*). Since the protein core is conserved, these epistatic interactions occur at the surface. In approximately 90% of such epistatic double mutants in the model (as calculated for the given example structure), the non-viable mutation is a P to H substitution. A real-world example of this form of epistasis is found in adenylate kinase [300].

the model's simplicity, its biophysical underpinning is transparent and intuitive; yet, the same underpinnings are closely related to those of more complex 2D lattice models with a biophysics-based 20-letter alphabet (see discussion below in this section). In fact, the biophysical criteria used to evaluate the 2D HP model may be applied as well to assess various 3D lattice models of protein evolution, including those that are being developed to study the impact of protein–protein interactions on evolution in model cells (§3.2.3).

One advantage of the minimalist HP construct is that, within the model, it allows for an exact, complete description of the sequence–structure mapping for short model proteins of lengths up to approximately 25 residues [316,317]. However, the extreme coarse-graining of both the sequence and structure spaces in this model means that energetic heterogeneity (diversity) among 20-letter real protein sequences with the same two-letter HP pattern is ignored; and mutations among 20-letter sequences with the same HP pattern are not considered. Obviously, the correspondence between model lattice conformations and real protein structures can only be intuitive.

Nevertheless, short-chain 2D HP models do capture a number of essential features of the sequence–structure mapping of real globular proteins. First, only a small fraction (approximately 2%) of HP sequences with chain lengths less than or equal to 30 have a unique lowest-energy structure

[158,162,299], consistent with experimental observations that only a tiny percentage of random amino acid sequences (much lower than 2%) can fold and/or have a biological function [116,272]. Experimentally, it is very rare for folded and/or functional sequences to arise in random sequence search, but binary HP patterning can help with artificial design of such sequences [318,319]. Second, the small fraction of model HP sequences that have a unique lowest-energy 2D structure exhibit statistics of HP patterns similar to that of real proteins [320,321]. Third, many of the lowest-energy 2D HP structures are highly compact but not maximally compact, as for real globular proteins [158,299], and a significant fraction of compact structures are not encodable by any HP sequence as its unique native structure [158,162]. The latter observation may bear on the question of whether the currently known set of globular protein folds is nearly complete in its coverage of all physically possible compact folds, as discussed in §3.1 [286–290], but one has to also keep in mind that the HP model interaction potential is less heterogeneous, and thus entails fewer encodable structures, than model potentials that contain repulsive interactions or otherwise more heterogeneous interactions [158,322,323]. Fourth, the 2D HP lattice model provides sequences that act like evolutionary bridges [161,178,204,239,299] (see §3.5), encode for autonomous folding units [292,316,324], and exhibit homology-like behaviours [295], all similar to properties observed in real proteins.

Two likely physical reasons underlie the similarity between the sequence–structure mapping of the 2D HP model and that of real globular proteins. First, the HP model potential captures the prominent effect of hydrophobic interactions in protein folding [44]. Similar to the hydrophobic effects operating in real proteins, the model potential leads to folded structures with a hydrophobic core and mostly polar residues on the surface (see examples in figure 5). Because the surface-core ratio of folded structures on the 2D square lattice with chain lengths approximately 16 is similar to the surface-core ratio of 3D folded structures with chain length approximately 150 [317], the energetics of short-chain 2D HP models should bear resemblance to that of real proteins with approximately 100 amino acid residues. Second, as has been argued [324], although the simple potentials of the HP model and its two-letter variants, and for that matter even 20-letter lattice potentials, are not sufficient to capture more detailed thermodynamic [325] and/or kinetic [326] properties of protein folding, the HP potential may still provide a useful caricature of the *mapping* between sequence and folded structure of real globular proteins because of the 'consistency principle' [327] or 'principle of minimal frustration' [328]. These principles stipulate consistency or near-consistency among various energy terms that contribute to the stability of natural proteins. Therefore, in this perspective, the folded state of a protein—at least for a natural, evolved protein—is expected to be a lowest-energy structure for a hydrophobic interaction potential similar to the one prescribed by the HP model, although other interaction types may provide it with additional stabilization.

Besides the HP and HP-like models, many of the 20-letter models adopt interaction energies derived from the PDB-based Miyazawa–Jernigan (MJ) statistical potential [329]. The original MJ potential (Table V of [329]), as was used in [40,153,296,330], embodies the hydrophobic effect, thus it tends to place non-polar residues in the folded protein core as in real proteins (reviewed in [116,204]). However, certain modified, 'shifted' forms of MJ potentials that are similar to table VI of [329] with prominent repulsive energies [331] do not embody this biophysical property (reviewed in [158,204]), making it problematic to interpret some of the predictions from such models. For instance, in some cases, a shifted-MJ potential may lead to nominally charged residues instead of hydrophobic residues occupying the core of a model protein [114,204,331].

These limitations notwithstanding, all aforementioned lattice models provided useful evolutionary insights. Different approaches are often complementary because they tackle different aspects of protein evolution. However, caution should always be used to take these models' limitations into account so as not to over-interpret model predictions. Major earlier progress of lattice models of protein evolution can be found in several reviews [292,332–334]. We now recall briefly some of the significant early efforts before highlighting recent advances.

3.2.3. Predictions and rationalizations

In several 2D [137,161,293,299,303,304,323] and 3D [305,335] lattice models as well as an off-lattice model [304], protein sequence space was found to be organized as multiple neutral nets. A neutral net is a network of sequences that are connected by single-point substitutions and encoding for the same folded structure. In a few studies that addressed the global connectivity of sequence space, it has been observed further that neutral nets for different folded structures are interconnected to form a dominant 'supernet' covering most of the sequence space [298,303,324] in a manner similar to that envisioned by Maynard Smith [336]. Consistent with experiment [139–142] (§2.4.2), mutations on encoding sequences often result in sequences encoding for the same structure [137,297,305], especially if the mutation site is on the surface of the folded protein [297]. In other words, many sequences are stable against mutation, a property referred to as mutational robustness [306].

Several models indicate that the topology of a protein neutral net has a superfunnel organization [137], wherein sequences encoding for the same structure tend to centre around a prototype sequence with maximum mutational robustness as well as maximum thermodynamic stability for the encoded native structure [137,293,304,323,335]. Mutational stability is generally correlated with thermodynamic stability [114,137], such that sequences at the edge of the neutral net have lower native stability [137,299,337]. Consistent with this trend, a 20-letter 2D lattice model predicts that, with increasing number of amino acid substitutions, the probability for a protein to retain its original structure declines exponentially [40].

Evolution of protein function has been explored using 2D square-lattice [296,301,302,330] and 3D diamond-lattice [335] model sequences that encode folded structures with a binding site. In the 2D square-lattice model studies of Bloom *et al*. [330,338], folded-state stability was found to promote evolvability of new binding functions, consistent with experimental observations. These simulated evolution processes allow for extensive exploration of sequence space. However, no structural change to the fold of the evolving lattice protein was seen during the simulation. Thus, the scope of protein evolution in these models [330,338] is largely limited to the

development of new binding abilities through modifications of surface residues, while the evolved protein retains its original structural scaffold. For larger structural changes, one expects that it would be harder for a more stable protein to evolve to a specific new fold, because a more stable sequence is farther away from the edge of the neutral net where it can switch to another fold (§2.7.2).

More recently, the binding between 2D square-lattice model proteins and lattice ligands was used to investigate short-term versus long-term evolutionary success. Interestingly, this study by Feldman and co-workers indicate that although the model evolution process is stochastic, long-term evolutionary success—as determined by stability of the evolving protein and its binding with a given ligand—is 'surprisingly predictable' from the founding sequence of a lineage. In this lattice protein model, long-term survivability is only partially determined by short-term fitness, i.e. short-term adaptive success does not guarantee long-term survival of a lineage [296].

2D lattice models have been applied to compare the evolutionary effects of point mutations and recombination [152,324,335]. An initial study showed that crossover of two encoding sequences is an effective way of producing a new encoding sequence, suggesting that local sequence patterns are important for determining whether the full protein sequence can fold to a unique structure [324]. This theoretical prediction is consistent with subsequent experiments on β-lactamase indicating that for a given number of amino acid substitutions, recombined variants are much more likely to retain function than variants generated by random point mutations [339]. In another simulation study, evolutionary dynamics that admits both point mutations and recombination leads to a much higher concentration of population in the prototype sequence than if evolution proceeds via point mutations alone, suggesting a significant role of recombination in the prototype-like behaviours of natural proteins [152].

Lattice models have also been used to investigate how mutagenesis data may be exploited to improve forcefields for protein structure prediction [340,341]. In addition, they have provided insights into the relationship between the native contact pattern of a target structure and its designability [342], the degree to which evolutionary selection at the molecular and/or organism level has led to the observed scale-free distribution of protein structure similarities [307,334], and the biophysical basis [343,344] of methods that use evolutionary/mutational information as a probe to reveal long-range energetic coupling in proteins [26].

An exciting recent development of 3D maximally compact lattice protein models is their application to the study of fitness and population dynamics of model cells containing multiple proteins. One model assumes that all proteins in a model cell are essential and that the fitness of the cell depends on the stability of the least stable proteins it contains. Real-life-like properties such as preferred folds and protein families readily emerge from these simple assumptions [308]. Introduction of various protein–protein interactions to this modelling set-up by Shakhnovich and co-workers [345] has provided further rationalization for the emergence of species-like collections of model cells with very similar sequence make-ups, an increased rate of mutation in stress response [346], as well as a trade-off between strengthening functional interactions and avoidance of misinteractions as observed in experimental proteomic data [347]. More recently, Zhang and co-workers [99] developed a related lattice approach to model evolution of protein–protein interactions that offers an explanation of slow evolution of highly expressed proteins in terms of stronger constraints on these proteins to avoid misinteractions.

3.3. Energy, entropy, fitness, protein neutral nets and fitness/mortality landscapes

Protein evolution can be formulated in terms of population dynamics on a fitness or adaptive landscape in which a fitness function assigns a fitness value to each protein sequence, with evolving populations migrating to areas of higher fitness over time [348–350]. Certain parallels may be drawn between fitness landscapes defined on sequence spaces and energy landscapes defined on protein conformational spaces. Mathematically, both protein sequence and conformational spaces are high-dimensional. As for protein energy landscapes, fitness landscapes are intrinsically high-dimensional constructs, even though fitness and energy landscapes are often depicted as 1D profiles or 2D surfaces for metaphorical, conceptual visualization [34,254,317,348,351]. With this in mind, it is important to focus on the quantitative fitness function itself and not to over-interpret the picturesque 1D and 2D landscape representations [351].

The biological fitness concept has been compared to the physical quantity of negative energy or negative free energy [352]. There is an obvious analogy between the sequence-space distribution of steady-state evolutionary population and the conformational-space distribution of equilibrium population. Just as lower-energy states and higher-entropy macrostates are more favoured in statistical mechanical systems, higher-fitness sequences and phenotypes encoded by more sequences are expected to be more populated during evolution. Under certain limiting conditions, a direct quantitative correspondence can be made between fitness and energy, as well as between population size and inverse temperature [352]. In general, however, there is an important difference between the analogous roles of energy and fitness: insofar as a statistical mechanical system is ergodic [353] and total population is conserved, the equilibrium population of a given state is determined by its energy in accordance with the Boltzmann distribution. By contrast, for evolution, total population can be increased by reproduction and decreased by lethal mutations. As a result, the steady-state population of a given sequence in a large evolving population is determined not only by reproductive fitness of the sequence but also by its connectivities to neighbouring viable sequences. This issue will be addressed in the discussion on mutational robustness in §3.4.

Studies of molecular evolution indicate that evolutionary pathways on fitness landscapes are subjected to various constraints [351,354]. The topographies of fitness landscapes of real proteins are far different, on average, from those postulated theoretically by random assignments of fitness [351]. This observation underscores the importance of biophysical considerations in constructing model fitness landscapes. As an illustration, figure 6 shows the sequence space of a 2D HP model (figure 6*a*), part of its neutral net organization (figure 6*b*), and a model fitness landscape that identifies fitness with native stability (figure 6*c*). As discussed above, the biophysics embodied by the HP model imparts it with several essential protein-like properties. For instance,

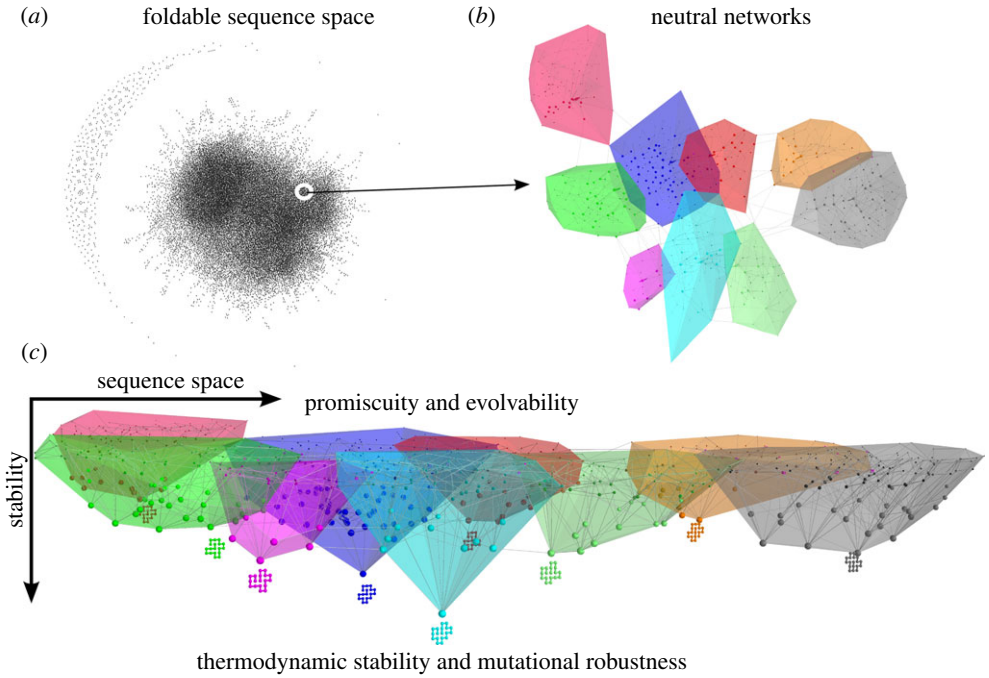

Figure 6. The space of foldable HP sequences gives a glimpse of what real protein sequence space might look like. (*a*) A representation of the adjacency network of approximately 25 000 HP sequences of length 18, each with either a unique ground-state structure or up to six different but equally populated ground-state structures. Only a small minority of sequences is not connected to the dominant 'supernet'. (*b*) Local sequence neighbourhoods form clusters of sequences that fold to the same ground-state structure. These clusters are called neutral nets or neutral networks, each of which is shown in a different colour here. Some of the neutral nets overlap. (*c*) Thermodynamic stabilities of the folded ground states of the sequences (stability increases in the downward direction) are plotted using a schematic planar of sequence space as in (*b*). A distinct funnel-like organization for each neutral net is apparent. These sequence-space funnels are referred to as *superfunnels* [137]. Typically, there is a single *prototype* sequence per neutral net that is thermodynamically maximally stable and also tolerates the highest number of neutral mutations (i.e. those resulting in a sequence within the neutral net). The ground-state structure in the HP model is shown for each neutral net. Sequences with high stability (at the bottom of the landscape) have a clear tendency to also have more neutral mutations and thus higher mutational robustness. In contrast, promiscuity and multi-functionality is likely encoded in less stable sequences, including sequences with two or more ground-state structures. These sequences are located further away from the prototype sequence but closer to other neutral nets; hence they have a higher evolvability—meaning a higher probability to acquire new functions through further mutations (see also figure 8). The network layouts here and in subsequent figures were constructed in such a way that the lengths of the edges roughly reflect the Hamming distances between the sequences connected by the edges [239]. Because of the large number of sequences depicted, the network drawings in the present figure convey only a heuristic impression of the sequence connections. More detailed descriptions of some of the neutral nets shown in figures 5–9 are provided in [137,239].

consideration of all model HP sequences that fold uniquely to the native structure in figure 5*a* [137,204] shows that certain sites are conserved, and that the probability that a mutation is viable is site-dependent (figure 5*b*), echoing the context-dependent substitution rates of amino acids in real proteins (§1). It is also noteworthy that because the main biophysical driving forces in protein folding are different from that in RNA secondary structure formation, the global organization of neutral nets of globular proteins is probably very different [161,355] from that of RNA [356–359].

It is conventional to associate increasing fitness with upward movement on fitness landscapes [348]. For the model fitness landscape in figure 6*c*, however, we have chosen to represent increasing fitness with downward movement. As discussed elsewhere, such landscapes may be referred to as 'inverse-fitness' or 'mortality' landscapes [324]. We prefer such depictions because our inescapable experience with gravity on Earth makes it easier for us to appreciate natural driving forces that point downward than ones that point upward. In this respect, mortality landscapes may offer a better metaphor for the drive by natural selection towards lower mortality and thus higher fitness. In analogy with the most favoured protein conformation having the lowest free energy in the conformational energy landscape, the fittest sequence is seen as situated at the bottom of an attractive basin with lowest mortality in the sequence-space mortality landscape [324].

3.4. Mutational robustness, sequence-space topology and population evolution

Experiments showed that many natural globular proteins are robust to mutations [40,53,55,202,360]. The observation that proteins with diverse primary sequences can be grouped into families with very similar structure and function gives further illustration of this robustness [361]. A folded structure that has a larger neutral net is more designable [160,322]. Compared to a less designable structure, it is expected to be more robust against mutations and thus more likely to be populated by evolution. From a biophysical standpoint, robustness may be viewed as a form of entropy in sequence space [306]. As discussed above, an idea of sequence-space 'entropy' (§3.3) is useful for analysing possible origins of marginal stability of natural globular proteins [135] (§2.4).

There are at least two biophysical reasons for the observed mutational robustness of natural globular proteins. First, a significant part of the stabilization of specific secondary structures in protein comes from backbone hydrogen

bonding. With proline as the only exception, amino acid substitutions do not change the ability of backbone atoms to form hydrogen bonds. While the nature of the amino acid sidechain has an effect on the preferred backbone dihedral angles and interactions with surrounding atoms, an α-helix or a β-sheet can be formed by many different combinations of amino acids along the primary sequence [362,363]. Second, the biophysics of hydrophobicity results in a rough grouping of a globular protein's amino acids into non-polar residues in a largely conserved hydrophobic core and polar surface residues. Surface residues with higher solvent accessibility are more tolerant to mutations [22] and thus contribute to mutational robustness because, on an individual basis, they are less crucial for stability (see lattice model example in figure 5*b*).

Mutational robustness is not determined solely by the number of sequences encoding for a certain structure or property. Another important factor that contributes to mutational robustness of any given sequence is its connectivity with other viable sequences. The pattern of mutational connections among sequences has also been referred to as sequence-space topology [137,151]. Among sequences with the same fitness, analytical and simulation studies have shown that evolving populations tend to prefer sequences that have more viable mutants. Fundamentally, this relative concentration of steady-state evolutionary population at sequences that are mutationally more stable (i.e. robust) arises from the fact that they lose less population to lethal mutations than sequences that are mutationally less robust [137,151,292]. As long as the evolving population N is sufficiently large and satisfies $\mu N \gg 1$, where μ is the mutation rate [151], the phenomenon is independent of μ and N [137,151]. A lattice-model example that illustrates how sequence-space topology affects evolutionary population under the $\mu N \gg 1$ condition is provided in figure 7. It should be noted, however, that when the $\mu N \gg 1$ condition is not satisfied, the significance of sequence-space topology on evolutionary population diminishes [151,153]. In the limit of $\mu N \ll 1$, evolutionary population dynamics on a neutral net becomes a random walk without regard to the relative abundance of connectivities of different sequences [151].

Figure 7*a* demonstrates that the role of fitness in steady-state evolutionary population distribution does not correspond directly to that of energy in equilibrium population distribution. Microstates having the same energy have equal equilibrium populations in a canonical ensemble [353]. In contrast, although all sequences are assumed to be equally fit in figure 7*a*; their steady-state evolutionary populations are different when $\mu N \gg 1$ because of sequence-space topology, with the prototype sequence that has maximum native stability and is also mutationally most robust achieving the highest population among sequences belonging to this sequence-space superfunnel [137]. However, this relative concentration of population at the prototype sequence is weak [137,152,153,335]. Thus, sequence-space topology alone is probably insufficient to account for the experimentally observed dominance of prototype-like sequences in natural proteins [139–142]. This recognition led to the argument in §2.4.2 that the distribution of evolutionary population in natural proteins has been driven significantly by selection for native thermodynamic and/or kinetic stability. Here, using the same explicit-chain lattice modelling set-up, we show in figure 7*b* how a selection for stability

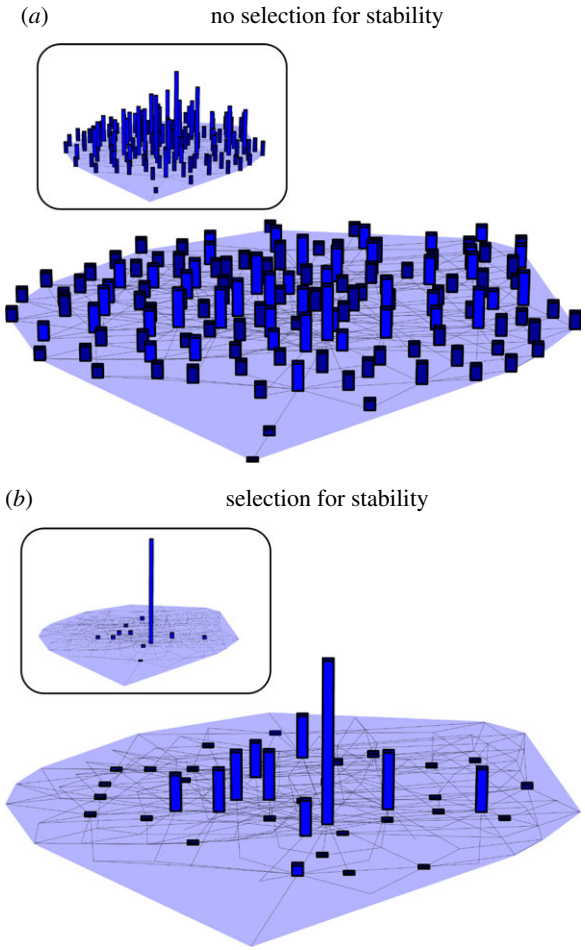

Figure 7. Interplay of network-topology and fitness effects on evolutionary population. In this figure, the 132 HP model sequences in an extended neutral net are depicted as nodes on a planar representation of sequence space (light blue surface), with each edge denoting a single-point mutation. The sequences either fold to the native structure in figure 5*b* uniquely or fold to that structure as well as at most five other ground-state structures [137]. Mutations leading to a sequence outside the net are deemed lethal in the present consideration. Here, steady-state distribution of evolutionary populations (i.e. at mutation–selection balance) of an infinite, asexually reproducing sequence population within the neutral net is computed without (*a*) or with (*b*) positive selection for native stability. The distributions of steady-state populations are shown by the histograms in logarithmic scales with the insets showing the same distributions in linear scales. Fractional populations of sequences that fold uniquely and sequences that fold to multiple ground-state structures are shown by vertical bars in lighter and darker colours, respectively. To better highlight the overall trend, different vertical scales are used in (*a*,*b*); and only fractional populations more than 0.1% are plotted. (*a*) With no selection for native stability, all sequences in the net have equal fitness. In this case, the uneven distribution in steady-state populations originates entirely from network topology. Among all 132 sequences, the prototype sequence has the highest steady-state population that encompasses, however, only 3.31% of total population. (*b*) When fitness is proportional to native stability, the majority of the steady-state population (78.3%) is concentrated at the prototype sequence. The fitness functions used for (*a*) and (*b*) here correspond to the single-structure fitness function defined in [239] with $\theta = 0$; $\tau = 1$ and $\theta = \tau = 1$, respectively.

may work in concert with sequence-space topology to beget a dominant population at the prototype sequence. This example also illustrates how mutational robustness of natural globular proteins may have arisen by selection for native stability without a direct selection on robustness itself [364].

3.5. Evolvability, hidden states and promiscuous functions

Mutational robustness is often discussed in conjunction with evolvability, which characterizes the ability of a biological system to evolve new traits [164,365–368]. The relationship between robustness and evolvability has been seen as opposing evolutionary forces, with the former impeding and the latter promoting evolutionary innovation. However, a network-based view of mutational robustness and evolvability indicates that they are not mutually exclusive [369,370]. Although mutational robustness implies that the sequence-space distance to any specific alternative phenotype is large (§§2.7.2 and 3.2.3), general evolvability can be enhanced by the mutational robustness afforded by a larger neutral network because different positions on such a network are likely to be mutationally close to many different phenotypes [294,371,372], as has been demonstrated experimentally for the evolution of enzyme functions [55,202,265,338].

These observations led to an understanding that seemingly neutral mutations can dramatically alter the future potential of a protein to evolve towards new functions. The hidden evolutionary potential of a neutral network is embodied in a wide variety of latent phenotypes that were not under selection originally but are present biophysically, because these mutational variations do not affect the main function of the protein. Several studies have shown how the co-option of neutrally evolved properties can allow adaptation towards new functions under the shadow of a dominant existing function. For example, enzyme promiscuity—which refers to low-affinity binding of molecules resembling an enzyme's main ligand target [373–375]—has been demonstrated as a potent mechanism for adaptations [264]. In the same vein, latent evolvable traits have also been identified in the evolution of steroid receptor specificity [376], allostery [201], gene regulation [377] and metabolic networks [378]. The term 'exaptation' (or 'spandrels' [138]; §2.4.1) has been coined for such latent traits that arise by chance and may or may not evolve to have a new function [379]. Apart from point mutations, mobile genetic elements are likely to play a crucial role in providing exaptations [380–382]. Each genome appears to constantly produce transcribed and translated 'proto-genes' that arise by chance, some of which may be exapted by evolution for a certain function [383].

It follows that a major part of the enhancement of evolvability by mutational robustness is based on the evolutionary potential provided by conformational dynamics at the level of a single sequence when excited-state structures of a protein [173,384–386] (§2.9) with beneficial function come under natural selection. Selection of such a promiscuous function rewards mutations that further stabilize the beneficial excited state. In this scenario, a protein can retain its original ground-state native structure while at the same time stabilizing an excited-state structure incrementally, thus maintaining continuous viability during evolutionary transitions. Eventually, the protein may first become bi-stable then switch to the selected excited-state conformation as its dominant structure (see lattice-model example in figure 8) or switch to the new dominant structure directly, as has been observed in protein design experiments (figure 4).

Consistent with experiment [202], lattice-model studies [178,203] indicate that an evolutionary process that takes

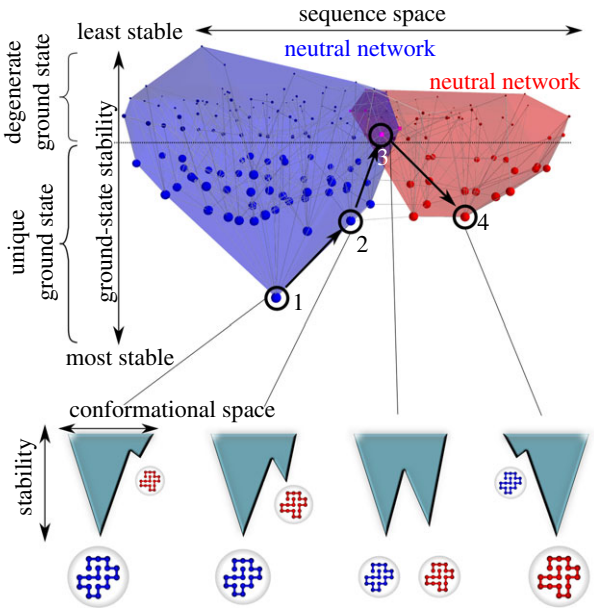

Figure 8. Mutational paths can be guided by selection for hidden conformational states: a more detailed view of the sequence-space fitness/mortality landscapes (top) of two of the HP model neutral nets in figure 6, with the conformational energy landscapes (bottom) of four adjacent sequences highlighted to illustrate the concept of excited-state selection [178,203, 239]. Idealized folding funnels at the bottom provide a schematic depiction of the relative populations of the blue and red structures. From left to right: sequence 1 folds predominantly into the blue native structure with only a negligible population (less than 1%) of the red structure. Now, a mutation that does not change the ground-state structure of sequence 1 produces sequence 2, which has an increased population of the red structure, to around 4%. At this population level, the red structure may be able to perform some new selectable function. Such selection can promote further mutations that produce a sequence that populates the blue and red structures with equal probabilities, i.e. a bi-stable protein (sequence 3; see figure 4). One further mutation can then switch the relative stabilities of the red and blue structure in sequences 1 and 2, finally leaving the red structure as the new native state (sequence 4). This theoretical perspective offers a biophysical rationalization for several recent results from directed evolution and NMR experiments [173,202]. Sequences 1–4 in this model have different fitness values if fitness is strictly proportional to ground state native stability. However, if fitness does not increase above a certain level of native stability, sequences 1–4 can be equally fit or nearly so [239] (figure 9).

advantage of excited-state selection is much more efficient than a process that applies selection pressure only on the dominant function [203]. In this perspective, bi-stable or multi-stable proteins at the peripheral of neutral networks—as exemplified by the overlapping regions in the lattice-model example in figure 6*b*—and proteins that sacrifice stability for functional promiscuity, as is the case in some antibodies [387], should be more evolvable towards new functions underpinned by significantly different native structures than sequences with high mutational robustness and thermodynamic stability. The latter sequences, however, can be more evolvable towards new functions that are still based upon the original structural scaffold [330,338].

3.6. Adaptive conflicts: evolution under constraints

While selection of latent traits can be an efficient route to new function, such an evolutionary process raises a basic

question regarding biophysical limits on the degree of multi-functionality or promiscuity that can be carried within one protein molecule. Multi-functionality also bears the danger of creating an adaptive conflict. Such a conflict can emerge whenever adaptation on the same gene is driven by two or more different or even mutually exclusive functional requirements. In the extreme case of viruses with severely constrained genome sizes, adaptive conflict can arise from overlapping open reading frames encoding for different proteins within the same DNA sequence [388]. For a multi-functional protein, adaptive conflict arises when enhancing one subfunction impedes another subfunction.

As far as adaptive conflicts in a protein is concerned, if multi-functionality is realized by the presence of different binding interfaces on the protein surface, a small number of such interfaces may coexist, limited by such factors as the protein size and/or surface area and the number of surface hydrophobic residues that can be tolerated without causing misfolding. Binding interfaces can also overlap, using some but not all of the same residues for different ligands [128,129]. In that case, an adaptive conflict may be anticipated since increasing the binding affinity for one interface through mutations may interfere with the binding affinity of the second interface, and vice versa. If multi-functionality is underpinned by bi-stability or multi-stability, i.e. the coexistence of and dynamic inter-conversion between alternative functional conformations (§2.7.2), it is expected biophysically that only a narrow capacity for accommodating several distinct conformations can exist in the lifetime of a globular protein and perhaps a somewhat higher capacity for doing so in disordered proteins. During evolution, a general mechanism for resolving adaptive conflict is offered by gene duplication and subfunctionalization [389], which we will discuss briefly in §3.7.

3.7. Protein divergence driven by gene duplication and mutational robustness

Evolution of bi-stable and more generally multi-stable proteins is an efficient means to meet new functional needs; but multi-functional proteins often serve as evolutionary intermediates rather than long-term solutions. Explicit-chain biophysical models suggest that multi-stable globular proteins are mutationally much less robust than globular proteins with an essentially unique native structure [161,178,239]. This trend is readily seen in the 2D HP lattice model example in figure 9, which shows that the sequence-space 'entropy' of bi-stable sequences (magenta area) is much smaller than that of sequences folding uniquely for either one of the two structures encoded by the bi-stable sequences (blue and red areas). This is a prediction that should be testable experimentally, for example, by using the recently designed bi-stable proteins [65,174,206] (figure 4). In this picture, the short-term advantage of bi-stability/bi-functionality is expected to give way eventually to an alternative sequence-space arrangement that is mutationally more robust, provided that the gene encoding for the bi-stable protein is duplicated at some point in the evolutionary process.

Subfunctionalization, or functional divergence after gene duplication, is a ubiquitous phenomenon in evolution [389,390]. For example, an experimental study of the reconstructed common ancestor of the fluorescent proteins in corals that emit either red or green light was found to emit

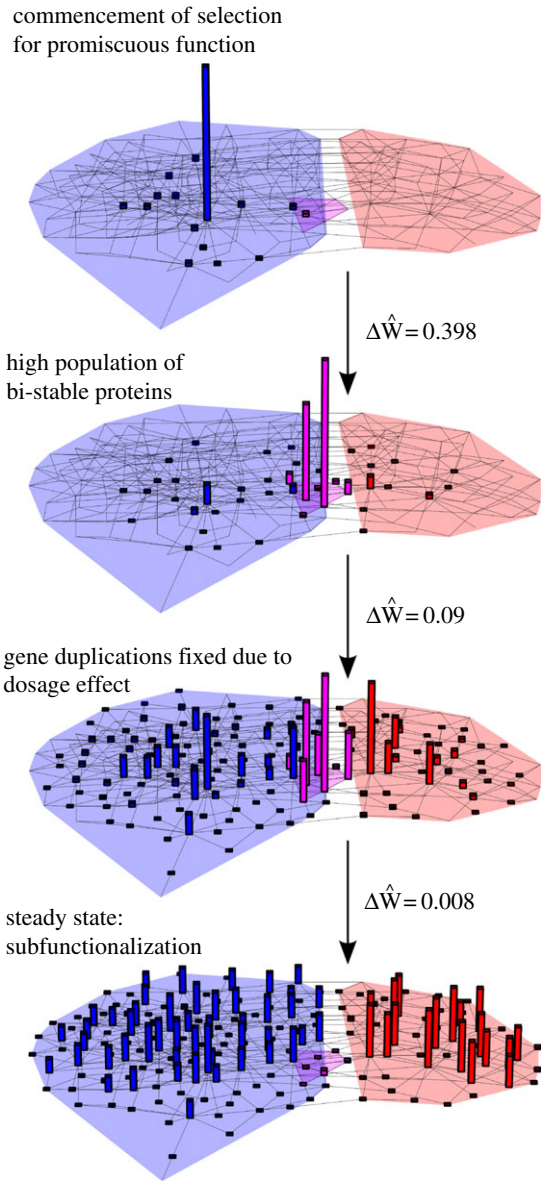

Figure 9. The simulated evolution, with gene duplication, of an essentially infinite HP sequence population under an adaptive conflict of two selection pressures. Here, four stages of the evolutionary dynamics are shown by representative changes in the distribution of evolutionary population and average population fitness $\Delta\hat{W}$ from one stage to the next. The two adjacent neutral nets (blue and red) are the same as those in figure 8. Distributions of population are plotted by logarithmic scales in the same style as figure 7. Initially, before the native structure of the red network is selected, nearly the entire population occupies the most stable HP sequence of the blue network. After selection pressure is imposed simultaneously for both the blue and red structures (figure 8), the red structure becomes a selectable promiscuous function in the model. After a number of generations, bi-stable proteins (magenta) appear as high-fitness evolutionary intermediates that are prone to undergo gene duplication. Duplicated bi-stable sequences that have maximum fitness in the model (because of an assumed beneficial dosage increase) then slowly give way to equally fit subfunctionalized (functionally diverged) gene pairs occupying those regions of the neutral networks with the highest mutational robustness. In other words, this model shows that some subfunctionalization processes can be non-adaptive in that subfunctionalization can be driven by sequence-space topology even when the total fitness of a duplicated pair of bi-stable proteins is the same as that of a pair of prototype sequences each folding uniquely to one or the other selected structures, as in this model. Consistent with this perspective, and as indicated by the last $\Delta\hat{W}$ values in this figure, the final-stage optimization of population distribution around the two prototype sequences provides only a very small increase in fitness value. Details of this model can be found in [239].

both red and green light [391]. Subfunctionalization of a duplicated multi-functional gene is probably a more efficient evolutionary route than neofunctionalization, which necessitates evolution of new function in a duplicated gene from scratch. However, in the subfunctionalization route, this process can be facilitated by selection on latent traits (§3.6) before gene duplication [239]. In the lexicon of fitness landscape, gene duplication amounts to doubling the number of dimensions of sequence space and thus may be viewed as an 'extradimensional bypass mechanism' for resolving adaptive conflicts [392].

Functional or structural divergence can be driven by an increase in fitness when a pair of identical bi-stable sequences is transformed into two subfunctionalized sequences. Generally speaking, such an increase in fitness is biophysically plausible because each of the subfunctionalized sequences may afford a higher kinetic stability [81] to one or the other functional structure than that provided by a bi-stable sequence [239]. However, divergence does not always have to be adaptive. Even if the fitness of a subfunctionalized pair is identical to the fitness of a pair of bi-stable sequences, the inherent tendency of protein evolution towards higher mutational robustness can still drive subfunctionalization. Biophysically, an ensemble of protein structures subjected to more overlapping functional constraints is likely to be more restricted in sequence space, resulting in low mutational robustness, as is exemplified in figure 9. When the constraints are lifted through gene duplication or changes in the environmental selection pressures, evolution will naturally favour mutations that result in higher mutational robustness even if there is no gain in functional fitness for the proteins in the process. This scenario is illustrated by the model evolutionary dynamics in figure 9. Simulation results summarized in this figure indicate that after an adaptive pressure to simultaneously select two structures is imposed, the evolving proteins first attempt a short-term resolution of the adaptive conflict using bi-stable sequences with low mutational robustness. Subsequently, upon gene duplication, a process of divergence that is essentially neutral ensues, with each copy of the originally bi-stable protein evolving towards the central, high-robustness region of one of the two neighbouring neutral networks [239] (figure 9).

3.8. Epistasis and co-evolution of interacting amino acid residues

The phenomenon of epistasis, referring originally to non-additivity of genetic effects caused by gene interactions (e.g. [393]), can also manifest within a protein molecule. Mutational effects on stability or function at different sites of a protein can be non-additive when the sites are energetically coupled [394]. A consequence of this biophysical property is that the overall evolutionary effect of multiple mutations can depend on the order in which the mutations are made. For the same given set of mutations, it may be that one temporal order of mutations is evolutionarily favoured because it entails a monotonic increase in fitness, whereas another order of mutations is disfavoured because it involves an intermediate step that decreases fitness. Several studies have demonstrated this type of epistatic behaviour in proteins and its constraints on evolutionary pathways [42,53,300,395–398]. For instance, experiments on adenylate kinase indicate that a double mutant with higher stability can only be obtained via one mutation path [300]. When the order of the two mutations was changed the protein could not fold. This type of behaviour is readily observed in simple biophysical models of protein evolution, as is illustrated by the HP model example in figure 5*c*.

An implication of epistasis is that the propensity for a viable mutation at a certain site in a protein structure may not be fixed. Rather it should depend on the preceding mutations at other sites. This phenomenon is illustrated by a recent simulation study of the evolution of purple acid phosphatase by generating random mutations in the structure [58]. Based on stability calculations in the model, the propensities for all amino acids at selected solvent-exposed and buried sites were determined. The results indicated considerable variations of these propensities because stability effects of amino acid substitutions at a given site change with time as the evolutionary process progresses. The simulation showed that an enforced destabilizing mutation (which could arise in real proteins owing to functional constraints such as the need to preserve an active site) can be compensated by subsequent mutations, thus increasing the future viability of the already-mutated residue at that same site and rendering the reverse mutation detrimental and therefore less probable [58]. Although this model probably overestimated native stability as well as stability effects of mutations and underestimated the probability for misfolding because of its simplified treatment of the unfolded state (see discussion above on marginal stability; §2.4.2), it offers an excellent elucidation of the 'holistic' nature of intra-protein interactions and the biophysical forces that govern the mutational effect on stability and how it may depend on the temporal order in which the mutations occur [58].

Epistatic effects are common but not universal. Strong epistasis arising from significant evolutionary shifts in the stability effects of mutations as envisioned in [58] may even be rare [57]. In the example of the influenza nucleoprotein, an experimental analysis of mutations in a set of homologues showed that stability effects of mutations with no clear functional benefit are largely conserved across homologues, mostly additive and exhibit no aforementioned [58] strong dependence on temporal order [57]. It has been argued that mutations in viral proteins in general—which probably have evolved to buffer deleterious mutations—are not likely to exhibit strong epistasis [399]. Whenever the functional benefits outweigh the cost of destabilization of a mutation, strong epistatic effects are more likely to follow [21]. Nonetheless, a weaker form of epistasis can occur even if stability effects of mutations are conserved because different temporal orders of stability changes can result in drastically different survivabilities. For instance, a recent experimental study of the 39 mutations on the nucleoprotein of the influenza virus between years 1968 and 2007 identified several mutations that decrease the stability of the protein significantly when introduced individually to the starting 1968 protein, thus suggesting strongly that these mutations were preceded by 'enabling' mutations that increase native stability. An inferred evolutionary trajectory was constructed based on the stability constraints [42]. Epistasis has also been revealed by studying disease-causing single mutations in humans and comparing them with compensated mutations that do not cause disease in other species [400–402]. One estimate indicates that 80% of pathological mutations result in protein stability changes [403]. When the compensated pathological mutations are compared against uncompensated pathological mutations, compensated mutations are mostly found at solvent-exposed positions and the amino acid substitutions are 'milder',

entailing, for example, less changes in hydrophobicity [400–402]. These experimental trends are consistent with the biophysical principles of protein structure and stability expounded here. A more detailed discussion of the biophysical/structural basis of epistasis and compensatory evolution is available in a recent review [404].

3.9. Fitness landscapes for multiple phenotypic properties

Natural protein evolution takes place in a highly wired, interacting molecular system. Ultimately, therefore, studies of protein evolution have to take into account a complex molecular context [392] (see §3.11). The expanded concept of molecular phenotype discussed in §2.9 is an attempt towards a better account of this biophysical reality. In this regard, highly simplified yet promising explicit-chain biophysical models have recently been developed to study protein evolution in the context of protein–protein interactions [99,347] (§3.2.3) We have also summarized explicit-chain biophysical models that take into consideration the ensemble nature of a protein's conformational phenotype, and how these models can provide insights into selection of promiscuous function, bi-stability and structural divergence (§3.2).

With the understanding that the fitness of a protein sequence depends not only on its ground-state native structure, but also on its entire conformational distribution as well as potential functional interactions and detrimental misinteractions with other molecules, a critical issue in modelling is how to assess contributions from different phenotypic properties to the overall fitness. For example, in the bi-stable fitness landscape in figures 8 and 9, the combined fitness is taken as the sum of two fitness values, one for each structure [178,239]. Although this modelling scheme is useful for illustrating general principles, it would be too simplistic when applied to real-life situations. Different forms of multi-functionality may require different rules for combining fitness contributions. Ideally, a fitness function should include not only positive contributions from selected biophysical properties, but also account for the negative effects on folding rates and misfolding, as well as aggregation and misinteraction. Thus, protein evolution in general entails a multi-factorial optimization problem where only something like a Pareto optimality can be achieved, i.e. a satisfaction of multiple optimality criteria just above a minimum threshold of optimality for each criterion [405]. One simple example would be the trade-off between thermodynamic stability of a folded protein versus the need for conformational dynamics in biochemical functions and degradation. Both these criteria probably cannot be fully satisfied, but a Pareto optimality may be achieved such that the protein is stable enough to maintain the same fold yet flexible enough to allow binding.

Constructing more realistic fitness functions will be a challenging task. Genomic information is abundant; but pinpointing mutational impact on cellular function by experiment is often daunting. Theoretical/computational investigations can assist greatly in this endeavour by developing more comprehensive models that account for various biophysical constraints on protein evolution. Two recent examples will be discussed in §3.10 and 3.11 to illustrate how incorporating information about protein–protein interactions into biophysical models can advance understanding of experimentally observed evolutionary patterns.

3.10. Biophysical links between protein expression level and evolutionary rate

A fundamental evolutionary question is why different proteins evolve at different rates. What makes some proteins less likely to accept new mutations than others? Can the different evolutionary rates be explained in terms of the biophysical constraints on mutations as outlined in figure 3?

One hypothesis is that proteins that are functionally more important are more conserved, because the cell cannot risk their function to be compromised in any way, even slightly. Some authors have linked the evolutionary rate to the position of a protein in the protein–protein interaction network, finding that 'hub' proteins involved in many interactions evolve more slowly [406]. Empirically, evolutionary rate was found to be most strongly anticorrelated with the expression level [407]. This trend is not inconsistent with the functional argument. A protein is essential to an organism if the organism fails to survive when the gene encoding for the protein is deleted from its genome [408,409]. Many essential proteins are highly expressed, as the cell needs a constant supply for its most basic and vital functions. However, is the slower evolutionary rate of highly expressed proteins a result of the importance of their functions or a more direct consequence of their high concentrations in the cell? Several biophysical mechanisms have been proposed for the latter scenario. Here we summarize briefly two mechanisms that are based, respectively, on protein misfolding and protein misinteraction, noting however that an explanation in terms of mRNA folding rate has also been put forth recently [410]. Multiple mechanisms can be at play because the proposed mechanisms are not mutually exclusive.

All proteins have to avoid misfolding. Taking the population of a protein sequence as a whole, a protein that is abundantly populated provides more opportunity for the formation of detrimental misfolded structures than a protein that is sparsely populated; thus the constraint imposed by misfolding avoidance is stronger on protein sequences with higher populations. A similar consideration applies to the misinteractions, which will be discussed further below. Accordingly, the need to prevent or at least minimize misfolding caused by translation error has been proposed as a major constraint on the evolution of highly expressed proteins [411–414], leading to slower evolution. Consistent with this picture, highly expressed proteins are selected to be more robust against translation errors by using synonymous codons with the smallest chance of producing non-synonymous changes as a result of translation errors. Apart from translation errors, the need to avoid misfolding of the properly translated protein also constitutes a strong evolutionary constraint on highly expressed proteins, resulting in preferential usage of amino acid residues that minimize misfolding [413]. These restrictive requirements lead to proteins that are both slowly evolving and thermodynamically more stable [414,415].

Another probable biophysical constraint behind the anti-correlation between protein expression level and evolutionary rate is the need for a protein to avoid misinteraction with other proteins [99]. This selection pressure affects primarily surface residues that can potentially participate in interactions between the protein and other molecules. Thus, its effects are to some degree complementary to that arising from the need for folding stability, kinetic accessibility of the folded structure

and avoidance of misfolding. The latter selection pressures affect primarily buried residues but can also affect surface residues. Misinteractions may be caused by the same exposed hydrophobic surface residues that are part of the functional protein–protein interactions, leading to an adaptive conflict between increasing the strength of functional interactions and avoiding misinteractions [347]. This conflict can result in further constraints that limit the viability of mutants of highly expressed proteins. As a consequence of these biophysical constraints, evolution might have increased the proportion of functional monomeric proteins with hydrophilic surfaces, reduced the abundance of functional multi-chain complexes, weakened the strengths of functional interactions, or increased the degree of disordered protein interactions to minimize exposed hydrophobics while still allowing many interaction partners [230,347]. In other words, such strategies might have contributed to the evolution of interaction network topologies that can better alleviate the conflict between functional interactions and misinteractions [110].

3.11. Protein evolution in the context of functional networks

As emphasized above, proteins do not act in isolation in living organisms; hence a full understanding of the function and evolution of a protein should take into account its interactions with other biomolecules and metabolites [392,416]. It would be daunting to account for these interactions in all their complexity at the molecular level. Using simple explicit-chain protein models, conceptual advances were made in elucidating how the biophysical constraint of misinteraction avoidance might impact protein evolution [99,347] (§3.10). However, investigators have to rely upon abstract descriptions of protein interactions, using model parameters extracted from experimental data on binding and on the effects of enzymatic activities on biochemical reaction rates. With an increasing repertoire of genomic data, this approach has produced significant progress. For instance, the recent mapping of a realistic network of DNA sequences bound by the same transcription factor [417] has afforded new evidence in support of the idea that a large genotype network enhances both mutational robustness and evolvability (§3.5).

Important advances have also been made by taking an abstract approach in the study of metabolic networks [418]. Notably, a reductive evolution algorithm was applied to determine minimal viable genomes for *E. coli* [419]. In principle, the effect of a mutation on metabolism is difficult to predict, because it affects not only the activity of the mutated protein but also many downstream events. Yet metabolic networks are often found to be robust against perturbations such as gene deletions and loss-of-function mutations because of 'distributed robustness', i.e. an ability of the network to compensate for the local perturbation by systemic adjustments [420]. *In silico* metabolic networks have also shed light on the evolution of specialist versus generalist enzymes. By analysing a model network of *E. coli* metabolism, it was found that specialists are very efficient at catalysing single metabolic reaction steps, responsible for a high metabolic throughput, and often essential to the cell. These functional roles necessitate more regulation of its activity. Consequently, specialists require a much higher degree of maintenance than generalists that are promiscuous and multi-functional. This model study thus offered an explanation for why specialists have not replaced all the generalists in real organisms [421].

A more recent computational study used flux balance analysis [422] and random re-wiring of a realistic model metabolic network [423] to study evolution of a model cell under a selection pressure to survive on a given carbon source [378]. The simulation showed that selection for one carbon source also allows the model cell to survive on a number of other carbon sources that were not selected for. This finding demonstrated that metabolic systems embody latent evolutionary potentials, and that beneficial traits can arise non-adaptively through exaptations [379] in the absence of selection at the level of metabolic network [378], as is the case we have seen at the molecular level (§3.5).

Although it is currently not feasible to apply explicit-chain models of proteins in the simulation of a realistic cellular metabolic network, a recent evolutionary population dynamics study was able to incorporate energetic information of explicit-chain continuum (off-lattice) models for 10 proteins from the folate biosynthetic pathway [59]. The study considers a population of 1000 model cells. The fitness of each model cell is taken to be the total metabolic output of the model biosynthetic pathway minus the number of misfolded proteins, with both of these quantities dependent upon the thermodynamic folded-state stabilities of the proteins. Protein stabilities are in turn computed using a biophysical potential function. Simulation results from this model provide a protein-based molecular biophysical rationalization for the distribution of stabilizing and destabilizing mutations and other experimentally observed patterns of polymorphisms [59].

4. Outlook: enriching the biophysics of protein evolution

As this review emphasizes, evolution is ultimately a physicochemical process that cannot be fully comprehended without biophysics. Likewise, because evolution happened under biophysical constraints, evolutionary information can help decipher aspects of protein biophysics that are still too complex or too costly to be tackled by first-principles physical or chemical methods. In closing, we provide a few further examples to showcase the productive research directions in which future progress will probably be made by this synergistic approach.

4.1. Synergy between biophysics and the study of protein evolution

Perhaps the most direct way to access the change in biophysical properties during the long evolutionary history of natural proteins is to perform experiments on 'resurrected' ancestral proteins. Recently, several putative ancestral protein sequences have been constructed computationally using common phylogenetic methods and then synthesized in the laboratory [82,391,424–428]. We have mentioned thioredoxin as one of the proteins that were studied in this manner in the discussion on kinetic stability (§2.2). Another interesting case is the reconstruction of the evolution of steroid receptors [376,429], revealing that ancient steroid receptors were already able to bind to the hormone aldosterone. But aldosterone only became available to

mammalian cells much later during evolution, indicating once again that latent functions, or exaptations (§3.5), can play important roles in protein evolution [376]. The study of these putative ancestral proteins indicates further that epistatic interactions within the structure of hormone receptors have led to surprisingly irreversible evolutionary pathways [430].

The utility of an ancestral reconstruction, however, is only as good as the accuracy of the phylogenetic relationships it assumes. As mentioned earlier, mutation rates depend not only on amino acid residue type, but also on the conformational context of the site being mutated; but many common phylogenetic methods do not account for this dependence (§1). One of the structural properties that exhibit significant correlation with mutation rate is conformational diversity. The conformational diversity exhibited by a single protein sequence is also reflected in the conformational diversity among the sequences in the family to which the protein belongs [431]. Local packing density of an amino acid position, which correlates negatively with local backbone conformational diversity (flexibility), was also found to correlate negatively with evolution rate. In other words, amino acid positions that have a lower local packing density and are more flexible locally tend to evolve faster. The correlation of evolutionary rate with local packing density is even stronger than that with solvent accessibility [432]. More recently, it has been suggested further that this strong correlation is a reflection of a fundamental relationship between evolutionary rate and the energetic stress caused by random mutations since average mutational stress is proportional to local packing density in an idealized elastic network model of protein structure [433]. This general trend is also consistent with the observation that IDRs typically evolve faster than globular proteins except when the IDR site is involved in the binding of multiple partners (§2.8.2). Taken together, the findings summarized above contradict another group's earlier finding that evolutionary rate is negatively correlated with conformational diversity [434,435]. The correlation coefficients computed from the limited dataset considered in the earlier study were, however, all very small [434].

As mentioned in §1, a recent method makes use of the information about solvent exposure of residues in a known protein structure to identify sites in a protein-coding gene that have undergone positive or negative evolutionary selection [23,24]. Although synonymous mutations are not necessarily neutral [95–97,436], the ratio of non-synonymous over synonymous codon replacement rates, $\omega = dN/dS$, is commonly used to detect adaptation in a given phylogenetic tree of related gene sequences. In such studies, most mutations do not show any signs of adaptation, with $\omega \approx 1$, and are thus considered 'neutral'; $\omega > 1$ is usually taken to indicate positive selection but negative selection is difficult to decipher from ω values. The predictive power of this method hinges on an accurate discrimination between neutral, adaptive and conservative ω signatures. Recent studies have shown that by considering solvent-exposed areas of mutation sites, this discrimination can be improved in a protein-specific manner. In this new approach, instead of using the same underlying assumptions for every protein family, as has been commonly practiced, biophysical consideration of solvent exposure is used to customize and improve the accuracy of the model for the specific protein under investigation (figure 1). This enriched methodology provides more accurate evolutionary inference than approaches that do not consider the conformational context of the mutation sites, but it requires the 3D structure of the protein in question. Hopefully, with better protein structure prediction techniques [437,438], it may be possible to apply structure-based phylogenetic reconstruction methods routinely by starting with sequence information alone.

4.2. Evolutionary protein biophysics: evolutionary information benefiting biophysical studies of proteins

The advances summarized above exemplify how biophysics can assist in the study of evolution. In the following, we describe briefly three examples in which evolutionary data have assisted in biophysical studies of proteins. The first example concerns prediction of mutational change in protein stability, $\Delta\Delta G$. As described in §2.1, a number of biophysical methods for $\Delta\Delta G$ prediction exist but are limited in various respects. In this context, it has been shown recently that a Bayesian method for inferring $\Delta\Delta G$ values of individual mutations from the evolutionary information embodied in homologous sequences can achieve accuracy exceeding pure biophysical methods and sequence-based consensus approaches. The method was applied to predict stabilizing mutations for influenza haemagglutinin. Subsequent experiment demonstrated that some of the mutations do allow a temperature-sensitive virus to grow at a higher temperature, attesting to the utility of this evolution-based method in improving biophysical understanding [439].

Another example that we have mentioned briefly is the detection of protein sectors from coevolution data (§1). Evolutionarily, protein sectors are largely independent of one another even though they are parts of the same protein. Amino acid residues within a sector are physically connected in the folded structure and are correlated evolutionarily [27] (figure 2). Sectors constitute sparse networks of co-evolving amino acid residues comprising only a minority of the residues in a protein. A recent high-throughput saturation point mutagenesis study of a PDZ domain (1577 mutations were tested) showed that sector positions are functionally less tolerant to mutation than non-sector positions [440]. These observations suggest that coevolution data can be used in general to gain insight into the biophysics of functional binding.

Coevolution data have also been applied to predict biophysical interactions in proteins, as mentioned briefly in §1 [28–32]. A computational algorithm has recently been developed to use pure sequence information to predict contacts within a protein structural domain. The approach is useful for predicting the native structure when sequence data are abundant but a structure has not been determined experimentally for a protein [31]. Even more interestingly, and going beyond earlier seminal findings [26], it was found that sequence information can reveal residue interactions that are not present in the PDB structure, including interactions between structural domains [31] as well as interactions involved in alternative conformational states with evolutionarily conserved functional significance [29]. Most recently, coevolutionary information of several protein families has been applied to determine a theoretical sequence-space

'selection temperature' T_{sel} that can be related to the T_f/T_g ratio between the folding and glass-transition temperatures in protein folding [441]. Because T_f/T_g is a biophysical measure of folding cooperativity [36,254], this latest analysis [441] demonstrates remarkably how fundamental biophysical principles can be revealed by evolutionary data. All in all, the examples in this section illustrate a general productive research approach of evolutionary protein biophysics, in which the current deluge of sequence-based evolutionary data is harnessed to extract important biophysical/structural information to improve understanding of protein function.

4.3. Role of theory and computation

Theoretical/computational methods are an integral part of the biophysical study of protein evolution (§3.2–3.4). Quantitative biophysical modelling is indispensable in the formulation of concepts, rationalization of existing experimental data, discovery of novel hypotheses and provision of predictions for subsequent experimental testing. Recent models not only addressed evolution of individual proteins but have also begun to take into account the interaction and metabolic networks in model organisms (§3.11). More complex models tend to be richer in that they have the capacity to provide non-trivial predictions that are not immediately obvious from the modelling set-up. Nonetheless, even simple explicit-chain models that serve largely to confirm expected trends are useful for conceptualizing how known evolutionary behaviours might have arisen from the physical forces that govern protein properties.

Despite improved model sophistication and the tremendous recent increase in computational power, models of protein evolution have to rely on representations that are coarse-grained and thus the models are inevitably highly simplified caricatures of the complex real situation they seek to mimic. In general, model predictions are sensitive to the assumed parameter values; but a precise correspondence between these values and physical reality often cannot be readily established. With these limitations in mind, theoreticians should strive to perform more exploration, as controls, of multiple parameter sets and alternative modelling set-ups that are biophysically plausible. Enhanced efforts in scenario classification are needed in general to better delineate the logical relationship between the assumptions and predictions of any given model.

4.4. Evolution within and across protein families and superfamilies

As outlined above, much progress has been made in experimental and theoretical studies of evolution within a protein family or superfamily. Compared to mutational changes that convert one protein fold to another, mutational changes that maintain essentially the same folded structure are computationally less costly to simulate; their study is more amenable to experimental techniques such as directed evolution, and can also benefit from the availability of abundant genomic data. It is more challenging to study fold-altering protein evolution. Nonetheless, notable recent experimental advances have been made in the design and structural characterization of bi-stable proteins and conformational switches (§2.7; figure 4). Understanding structural innovation entails biophysical accounting of conformational diversity—which underpins functional promiscuity—as well as the plasticity, or ensemble nature of molecular phenotype (§2.9). Here, we have placed considerable emphasis on this broader perspective of protein evolution, although theoretical investigation in this area is only in its infancy (§§2.7 and 3.5–3.7; figures 4 and 8). We hope to witness more advances in this direction. It is exciting to better understand not only how proteins evolved within a family or a superfamily but also, more fundamentally, how the structural families originated in the first place.

5. Concluding remarks

The aim of this review is to provide a broad sketch of the fundamental biophysical forces that both enable and constrain protein evolution. Starting with the effects of mutations on protein stability, folding kinetics, interactions, functional dynamics, promiscuous functions, conformational switch and conformational disorder, these findings are then linked to broader evolutionary themes including the global and local organization of protein sequence and structure space, simple models of the protein sequence–structure mapping, fitness/mortality landscapes, sequence-space topology and mutational robustness, adaptive conflict and its possible resolution by selection of promiscuous function and subfunctionalization driven by mutational robustness, evolvability, epistasis and intra-cellular networks. We have highlighted advances made through computational models, especially simple exact and other explicit-chain models of protein evolution, because many insightful discoveries in biophysics of protein evolution were pioneered through simple, coarse-grained modelling of biological or biophysical processes that are too complex to be studied in atomistic details. As far as simplified models are concerned, explicit-chain models with biophysics-based interactions enjoy a clear physical advantage over theories that contain little or no biophysical consideration of protein structure and dynamics. We have also summarized several recent experimental advances that bear on the biophysics of evolution, as many questions that have arisen from theory and simulation can only be answered definitely by further experiments. Even so, this review touches upon only a small fraction of all the exciting discoveries that have been made lately. Looking into the future, we expect to witness increasing collaboration between the fields of biophysics and evolution as well as between theory/computation and experiment to decipher many aspects of the evolutionary forces that have been shaping the biological roles of proteins.

Acknowledgements. We thank Jesse Bloom, Xavier de la Cruz, Julie Forman-Kay, Alessandro Laio, Austin Meyer, Marc Ostermeier, Jose Sanchez-Ruiz, Andreas Wagner and Claus Wilke for helpful discussions. H.S.C. wishes to take this opportunity to thank Erich Bornberg-Bauer specially for a fruitful and pleasurable collaboration on evolutionary studies over many years. Part of this work was presented at the 2014 Meeting of the Society for Molecular Biology and Evolution (San Juan, Puerto Rico) by T.S., who gratefully acknowledges a travel award he received from the Canadian Institutes of Health Research (CIHR) Training Program in 'Protein Folding and Interaction Dynamics: Principles and Diseases' at the University of Toronto.

Funding statement. This work was supported by a CIHR grant to H.S.C. and the computational resource provided by SciNet of Compute Canada.

References

1. Dunham I *et al*. 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. (doi:10.1038/nature11247)
2. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004 Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216. (doi:10.1016/j.sbi.2004.03.011)
3. Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. 2008 Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.* **33**, 444–451. (doi:10.1016/j.tibs.2008.05.008)
4. Bornberg-Bauer E, Albà MM. 2013 Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.* **23**, 459–466. (doi:10.1016/j.sbi.2013.02.012)
5. Söding J, Lupas AN. 2003 More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays* **25**, 837–846. (doi:10.1002/bies.10321)
6. Höcker B, Claren J, Sterner R. 2004 Mimicking enzyme evolution by generating new $(\beta\alpha)_8$-barrels from $(\beta\alpha)_4$-half-barrels. *Proc. Natl Acad. Sci. USA* **101**, 16 448–16 453. (doi:10.1073/pnas.0405832101)
7. Carbone MN, Arnold FH. 2007 Engineering by homologous recombination: exploring sequence and function within a conserved fold. *Curr. Opin. Struct. Biol.* **17**, 454–459. (doi:10.1016/j.sbi.2007.08.005)
8. Höcker B. 2013 Engineering chimaeric proteins from fold fragments: 'hopeful monsters' in protein design. *Biochem. Soc. Trans.* **41**, 1137–1140. (doi:10.1042/BST20130099)
9. Smith MA, Romero PA, Wu T, Brustad EM, Arnold FH. 2013 Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein Sci.* **22**, 231–238. (doi:10.1002/pro.2202)
10. Henikoff S, Henikoff JG. 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10 915–10 919. (doi:10.1073/pnas.89.22.10915)
11. Dayhoff M, Schwartz R, Orcutt B. 1978 A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (ed MO Dayhoff), pp. 345–352. Washington, DC: National Biomedical Research Foundation.
12. Worth CL, Gong S, Blundell TL. 2009 Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **10**, 709–720. (doi:10.1038/nrm2762)
13. Wilke CO. 2012 Bringing molecules back into molecular evolution. *PLoS Comput. Biol.* **8**, e1002572. (doi:10.1371/journal.pcbi.1002572)
14. Liberles DA *et al*. 2012 The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* **21**, 769–785. (doi:10.1002/pro.2071)
15. Studer RA, Dessailly BH, Orengo CA. 2013 Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.* **449**, 581–594. (doi:10.1042/BJ20121221)
16. Harms MJ, Thornton JW. 2013 Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571. (doi:10.1038/nrg3540)
17. Bordner AJ, Mittelmann HD. 2014 A new formulation of protein evolutionary models that account for structural constraints. *Mol. Biol. Evol.* **31**, 736–749. (doi:10.1093/molbev/mst240)
18. Rodrigue N, Philippe H. 2010 Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet.* **26**, 248–252. (doi:10.1016/j.tig.2010.04.001)
19. Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010 Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.* **27**, 1546–1560. (doi:10.1093/molbev/msq047)
20. Le SQ, Gascuel O. 2010 Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst. Biol.* **59**, 277–287. (doi:10.1093/sysbio/syq002)
21. DePristo MA, Weinreich DM, Hartl DL. 2005 Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687. (doi:10.1038/nrg1672)
22. Franzosa EA, Xia Y. 2009 Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26**, 2387–2395. (doi:10.1093/molbev/msp146)
23. Scherrer MP, Meyer AG, Wilke CO. 2012 Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol. Biol.* **12**, 179. (doi:10.1186/1471-2148-12-179)
24. Meyer AG, Wilke CO. 2013 Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.* **30**, 36–44. (doi:10.1093/molbev/mss217)
25. Stevens J, Corper AL, Basler CF, Taubenberger JK, Palese P, Wilson IA. 2004 Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* **303**, 1866–1870. (doi:10.1126/science.1093373)
26. Lockless SW, Ranganathan R. 1999 Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299. (doi:10.1126/science.286.5438.295)
27. Halabi N, Rivoire O, Leibler S, Ranganathan R. 2009 Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786. (doi:10.1016/j.cell.2009.07.038)
28. Morcos F *et al*. 2010 Modeling conformational ensembles of slow functional motions in Pin1-WW. *PLoS Comput. Biol.* **6**, e1001015. (doi:10.1371/journal.pcbi.1001015)
29. Morcos F, Jana B, Hwa T, Onuchic JN. 2013 Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl Acad. Sci. USA* **110**, 205 33–205 38. (doi:10.1073/pnas.1315625110)
30. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. 2009 High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl Acad. Sci. USA* **106**, 22 124–22 129. (doi:10.1073/pnas.0912100106)
31. Morcos F *et al*. 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301. (doi:10.1073/pnas.1111471108)
32. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. 2012 Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc. Natl Acad. Sci. USA* **109**, E1733–E1742. (doi:10.1073/pnas.1201301109)
33. Leopold PE, Montal M, Onuchic JN. 1992 Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl Acad. Sci. USA* **89**, 8721–8725. (doi:10.1073/pnas.89.18.8721)
34. Dill KA, Chan HS. 1997 From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19. (doi:10.1038/nsb0197-10)
35. Onuchic JN, Wolynes PG. 2004 Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75. (doi:10.1016/j.sbi.2004.01.009)
36. Chan HS, Zhang Z, Wallin S, Liu Z. 2011 Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. *Annu. Rev. Phys. Chem.* **62**, 301–326. (doi:10.1146/annurev-physchem-032210-103405)
37. Koshland DE. 1958 Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA* **44**, 98–104. (doi:10.1073/pnas.44.2.98)
38. Csermely P, Palotai R, Nussinov R. 2010 Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546. (doi:10.1016/j.tibs.2010.04.009)
39. Bellotti V, Stoppini M, Mangione PP, Fornasieri A, Min L, Merlini G, Ferri G. 1996 Structural and functional characterization of three human immunoglobulin kappa light chains with different pathological implications. *Biochim. Biophys. Acta* **1317**, 161–167. (doi:10.1016/S0925-4439(96)00049-X)
40. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005 Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA* **102**, 606–611. (doi:10.1073/pnas.0406744102)
41. Mayer S, Rüdiger S, Ang HC, Joerger AC, Fersht AR. 2007 Correlation of levels of folded recombinant p53 in *Escherichia coli* with thermodynamic stability *in vitro*. *J. Mol. Biol.* **372**, 268–276. (doi:10.1016/j.jmb.2007.06.044)
42. Gong LI, Suchard MA, Bloom JD. 2013 Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631. (doi:10.7554/eLife.00631)
43. Kauzmann W. 1959 Some factors in the interpretation of protein denaturation. *Adv. Protein*

Chem. **14**, 1–63. (doi:10.1016/S0065-3233(08)60608-7)

44. Dill KA. 1990 Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155. (doi:10.1021/bi00483a001)
45. Rost B. 2001 Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134**, 204–218. (doi:10.1006/jsbi.2001.4336)
46. Guerois R, Nielsen JE, Serrano L. 2002 Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387. (doi:10.1016/S0022-2836(02)00442-4)
47. Capriotti E, Fariselli P, Casadio R. 2005 I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, W306–W310. (doi:10.1093/nar/gki375)
48. Parthiban V, Gromiha MM, Schomburg D. 2006 CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **34**, W239–W242. (doi:10.1093/nar/gkl190)
49. Yin S, Ding F, Dokholyan NV. 2007 Eris: an automated estimator of protein stability. *Nat. Methods* **4**, 466–467. (doi:10.1038/nmeth0607-466)
50. Wang Q, Canutescu AA, Dunbrack RL. 2008 SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.* **3**, 1832–1847. (doi:10.1038/nprot.2008.184)
51. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. 2009 Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: poPMuSiC-2.0. *Bioinformatics* **25**, 2537–2543. (doi:10.1093/bioinformatics/btp445)
52. Kellogg EH, Leaver-Fay A, Baker D. 2011 Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838. (doi:10.1002/prot.22921)
53. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. 2006 Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932. (doi:10.1038/nature05385)
54. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. 2007 The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* **369**, 1318–1332. (doi:10.1016/j.jmb.2007.03.069)
55. Bershtein S, Goldin K, Tawfik DS. 2008 Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**, 1029–1044. (doi:10.1016/j.jmb.2008.04.024)
56. Bloom JD, Nayak JS, Baltimore D. 2011 A computational-experimental approach identifies mutations that enhance surface expression of an oseltamivir-resistant influenza neuraminidase. *PLoS ONE* **6**, e22201. (doi:10.1371/journal.pone.0022201)
57. Ashenberg O, Gong LI, Bloom JD. 2013 Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl Acad. Sci. USA* **110**, 21 071–21 076. (doi:10.1073/pnas.1314781111)
58. Pollock DD, Thiltgen G, Goldstein RA. 2012 Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl Acad. Sci. USA* **109**, E1352–E1359. (doi:10.1073/pnas.1120084109)
59. Serohijos AWR, Shakhnovich EI. 2014 Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. *Mol. Biol. Evol.* **31**, 165–176. (doi:10.1093/molbev/mst189)
60. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. 2009 Predicting free energy changes using structural ensembles. *Nat. Methods* **6**, 3–4. (doi:10.1038/nmeth0109-3)
61. Willis JR, Briney BS, DeLuca SL, Crowe JE, Meiler J. 2013 Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput. Biol.* **9**, e1003045. (doi:10.1371/journal.pcbi.1003045)
62. Howell SC, Inampudi KK, Bean DP, Wilson CJ. 2014 Understanding thermal adaptation of enzymes through the multistate rational design and stability prediction of 100 adenylate kinases. *Structure* **22**, 218–229. (doi:10.1016/j.str.2013.10.019)
63. Cordes MHJ, Sauer RT. 1999 Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Sci.* **8**, 318–325. (doi:10.1110/ps.8.2.318)
64. Gu H, Doshi N, Kim DE, Simons KT, Santiago JV, Nauli S, Baker D. 1999 Robustness of protein folding kinetics to surface hydrophobic substitutions. *Protein Sci.* **8**, 2734–2741. (doi:10.1110/ps.8.12.2734)
65. Cordes MHJ, Burton RE, Walsh NP, McKnight CJ, Sauer RT. 2000 An evolutionary bridge to a new protein fold. *Nat. Struct. Biol.* **7**, 1129–1132. (doi:10.1038/81985)
66. Seeliger D, de Groot BL. 2010 Protein thermostability calculations using alchemical free energy simulations. *Biophys. J.* **98**, 2309–2316. (doi:10.1016/j.bpj.2010.01.051)
67. Allison JR, Bergeler M, Hansen N, van Gunsteren WF. 2011 Current computer modeling cannot explain why two highly similar sequences fold into different structures. *Biochemistry* **50**, 10 965–10 973. (doi:10.1021/bi2015663)
68. Hansen N, Allison JR, Hodel FH, van Gunsteren WF. 2013 Relative free enthalpies for point mutations in two proteins with highly similar sequences but different folds. *Biochemistry* **52**, 4962–4970. (doi:10.1021/bi400272q)
69. Roy A, Perez A, Dill KA, Maccallum JL. 2014 Computing the relative stabilities and the per-residue components in protein conformational changes. *Structure* **22**, 168–175. (doi:10.1016/j.str.2013.10.015)
70. Baker D. 2000 A surprising simplicity to protein folding. *Nature* **405**, 39–42. (doi:10.1038/35011000)
71. Brockwell DJ, Radford SE. 2007 Intermediates: ubiquitous species on folding energy landscapes? *Curr. Opin. Struct. Biol.* **17**, 30–37. (doi:10.1016/j.sbi.2007.01.003)
72. Matthews CR, Hurle MR. 1987 Mutant sequences as probes of protein folding mechanisms. *BioEssays* **6**, 254–257. (doi:10.1002/bies.950060603)
73. Shortle D. 1989 Probing the determinants of protein folding and stability with amino acid substitutions. *J. Biol. Chem.* **264**, 5315–5318.
74. Jackson SE, ElMasry N, Fersht AR. 1993 Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry* **32**, 11 270–11 278. (doi:10.1021/bi00093a002)
75. Lawrence C, Kuge J, Ahmad K, Plaxco KW. 2010 Investigation of an anomalously accelerating substitution in the folding of a prototypical two-state protein. *J. Mol. Biol.* **403**, 446–458. (doi:10.1016/j.jmb.2010.08.049)
76. Viguera AR, Vega C, Serrano L. 2002 Unspecific hydrophobic stabilization of folding transition states. *Proc. Natl Acad. Sci. USA* **99**, 5349–5354. (doi:10.1073/pnas.072387799)
77. Zarrine-Afsar A, Wallin S, Neculai AM, Neudecker P, Howell PL, Davidson AR, Chan HS. 2008 Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding. *Proc. Natl Acad. Sci. USA* **105**, 9999–10 004. (doi:10.1073/pnas.0801874105)
78. Debès C, Wang M, Caetano-Anollés G, Gräter F. 2013 Evolutionary optimization of protein folding. *PLoS Comput. Biol.* **9**, e1002861. (doi:10.1371/journal.pcbi.1002861)
79. Mirny LA, Shakhnovich EI. 1999 Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196. (doi:10.1006/jmbi.1999.2911)
80. Di Nardo AA, Larson SM, Davidson AR. 2003 The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333**, 641–655. (doi:10.1016/j.jmb.2003.08.035)
81. Sanchez-Ruiz JM. 2010 Protein kinetic stability. *Biophys. Chem.* **148**, 1–15. (doi:10.1016/j.bpc.2010.02.004)
82. Inglés-Prieto A, Ibarra-Molero B, Delgado-Delgado A, Perez-Jimenez R, Fernandez JM, Gaucher EA, Sanchez-Ruiz JM, Gavira JA. 2013 Conservation of protein structure over four billion years. *Structure* **21**, 1690–1697. (doi:10.1016/j.str.2013.06.020)
83. Godoy-Ruiz R, Ariza F, Rodriguez-Larrea D, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM. 2006 Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J. Mol. Biol.* **362**, 966–978. (doi:10.1016/j.jmb.2006.07.065)
84. Gsponer J, Hopearuoho H, Whittaker SB-M, Spence GR, Moore GR, Paci E, Radford SE, Vendruscolo MH. 2006 Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7. *Proc. Natl Acad. Sci. USA* **103**, 99–104. (doi:10.1073/pnas.0508667102)
85. Kato H, Feng H, Bai Y. 2007 The folding pathway of T4 lysozyme: the high-resolution structure and folding of a hidden intermediate. *J. Mol. Biol.* **365**, 870–880. (doi:10.1016/j.jmb.2006.10.047)

86. Dalessio PM, Boyer JA, McGettigan JL, Ropson IJ. 2005 Swapping core residues in homologous proteins swaps folding mechanism. *Biochemistry* **44**, 3082–3090. (doi:10.1021/bi048125u)
87. Valastyan JS, Lindquist S. 2014 Mechanisms of protein-folding diseases at a glance. *Dis. Model Mech.* **7**, 9–14. (doi:10.1242/dmm.013474)
88. Ciryam P, Tartaglia GG, Morimoto RI, Dobson CM, Vendruscolo MH. 2013 Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell Rep.* **5**, 1–10. (doi:10.1016/j.celrep.2013.09.043)
89. Jahn TR, Parker MJ, Homans SW, Radford SE. 2006 Amyloid formation under physiological conditions proceeds via a native-like folding intermediate. *Nat. Struct. Mol. Biol.* **13**, 195–201. (doi:10.1038/nsmb1058)
90. Surguchev A, Surguchov A. 2010 Conformational diseases: looking into the eyes. *Brain Res. Bull.* **81**, 12–24. (doi:10.1016/j.brainresbull.2009.09.015)
91. Das P, King JA, Zhou R. 2011 Aggregation of γ-crystallins associated with human cataracts via domain swapping at the C-terminal β-strands. *Proc. Natl Acad. Sci. USA* **108**, 10 514–10 519. (doi:10.1073/pnas.1019152108)
92. Ji F, Jung J, Koharudin LMI, Gronenborn AM. 2013 The human W42R γD-crystallin mutant structure provides a link between congenital and age-related cataracts. *J. Biol. Chem.* **288**, 99–109. (doi:10.1074/jbc.M112.416354)
93. Baskakov IV, Legname G, Prusiner SB, Cohen FE. 2001 Folding of prion protein to its native α-helical conformation is under kinetic control. *J. Biol. Chem.* **276**, 19 687–19 690. (doi:10.1074/jbc.C100180200)
94. Lobkovsky AE, Wolf YI, Koonin EV. 2010 Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc. Natl Acad. Sci. USA* **107**, 2983–2988. (doi:10.1073/pnas.0910445107)
95. Ciryam P, Morimoto RI, Vendruscolo MH, Dobson CM, O'Brien EP. 2013 *In vivo* translation rates can substantially delay the cotranslational folding of the *Escherichia coli* cytosolic proteome. *Proc. Natl Acad. Sci. USA* **110**, E132–E140. (doi:10.1073/pnas.1213624110)
96. Sander IM, Chaney JL, Clark PL. 2014 Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.* **136**, 858–861. (doi:10.1021/ja411302m)
97. Tsai C-J, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. 2008 Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J. Mol. Biol.* **383**, 281–291. (doi:10.1016/j.jmb.2008.08.012)
98. Nooren IMA, Thornton JM. 2003 Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492. (doi:10.1093/emboj/cdg359)
99. Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. 2012 Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl Acad. Sci. USA* **109**, E831–E840. (doi:10.1073/pnas.1117408109)
100. Ingram VM. 1957 Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* **180**, 326–328. (doi:10.1038/180326a0)
101. Pauling L, Itano H, Singer S, Wells I. 1949 Sickle cell anemia, a molecular disease. *Science* **110**, 543–548. (doi:10.1126/science.110.2865.543)
102. Meyer V *et al*. 2014 Single mutations in tau modulate the populations of fibril conformers through seed selection. *Angew. Chem. Int. Ed. Engl.* **53**, 1590–1593. (doi:10.1002/anie.201308473)
103. Schuster-Böckler B, Bateman A. 2008 Protein interactions in human genetic diseases. *Genome Biol.* **9**, R9. (doi:10.1186/gb-2008-9-1-r9)
104. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. 2012 Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164. (doi:10.1038/nbt.2106)
105. Ellis RJ, Minton AP. 2006 Protein aggregation in crowded environments. *Biol. Chem.* **387**, 485–497. (doi:10.1515/BC.2006.064)
106. Gershenson A, Gierasch LM. 2010 Protein folding in the cell: challenges and progress. *Curr. Opin. Struct. Biol.* **21**, 32–41. (doi:10.1016/j.sbi.2010.11.001)
107. Dill KA, Ghosh K, Schmit JD. 2011 Physical limits of cells and proteomes. *Proc. Natl Acad. Sci. USA* **108**, 17 876–17 882. (doi:10.1073/pnas.1114477108)
108. Levy ED, De S, Teichmann SA. 2012 Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl Acad. Sci. USA* **109**, 20 461–20 466. (doi:10.1073/pnas.1209312109)
109. Sarkar M, Smith AE, Pielak GJ. 2013 Impact of reconstituted cytosol on protein stability. *Proc. Natl Acad. Sci. USA* **110**, 19 342–19 347. (doi:10.1073/pnas.1312678110)
110. Johnson ME, Hummer G. 2011 Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc. Natl Acad. Sci. USA* **108**, 603–608. (doi:10.1073/pnas.1010954108)
111. Yue K, Dill KA. 1992 Inverse protein folding problem: designing polymer sequences. *Proc. Natl Acad. Sci. USA* **89**, 4163–4167. (doi:10.1073/pnas.89.9.4163)
112. Isogai Y. 2006 Native protein sequences are designed to destabilize folding intermediates. *Biochemistry* **45**, 2488–2492. (doi:10.1021/bi0523714)
113. Sali A, Shakhnovich EI, Karplus M. 1994 Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636. (doi:10.1006/jmbi.1994.1110)
114. Broglia RA, Tiana G, Roman HH, Vigezzi E, Shakhnovich EI. 1999 Stability of designed proteins against mutations. *Phys. Rev. Lett.* **82**, 4727–4730. (doi:10.1103/PhysRevLett.82.4727)
115. Chan HS, Shimizu S, Kaya H. 2004 Cooperativity principles in protein folding. *Methods Enzymol.* **380**, 350–379. (doi:10.1016/S0076-6879(04)80016-8)
116. Chan HS. 1999 Folding alphabets. *Nat. Struct. Biol.* **6**, 994–996. (doi:10.1038/14876)
117. Tompa P, Tusnády GE, Cserzo M, Simon I. 2001 Prion protein: evolution caught en route. *Proc. Natl Acad. Sci. USA* **98**, 4431–4436. (doi:10.1073/pnas.071308398)
118. Baldwin AJ *et al*. 2011 Metastability of native proteins and the phenomenon of amyloid formation. *J. Am. Chem. Soc.* **133**, 14 160–14 163. (doi:10.1021/ja2017703)
119. Harrison PM, Chan HS, Prusiner SB, Cohen FE. 2001 Conformational propagation with prion-like characteristics in a simple model of protein folding. *Protein Sci.* **10**, 819–835. (doi:10.1110/ps.38701)
120. Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. 2013 Protein–protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. *PLoS Comput. Biol.* **9**, e1003369. (doi:10.1371/journal.pcbi.1003369)
121. Andreeva A, Murzin AG. 2006 Evolution of protein fold in the presence of functional constraints. *Curr. Opin. Struct. Biol.* **16**, 399–408. (doi:10.1016/j.sbi.2006.04.003)
122. Tokuriki N, Tawfik DS. 2009 Protein dynamism and evolvability. *Science* **324**, 203–207. (doi:10.1126/science.1169375)
123. Yomo T, Saito S, Sasai M. 1999 Gradual development of protein-like global structures through functional selection. *Nat. Struct. Biol.* **6**, 743–746. (doi:10.1038/11512)
124. Nagao C, Terada TP, Yomo T, Sasai M. 2005 Correlation between evolutionary structural development and protein folding. *Proc. Natl Acad. Sci. USA* **102**, 18 950–18 955. (doi:10.1073/pnas.0509163102)
125. Perica T, Chothia C, Teichmann SA. 2012 Evolution of oligomeric state through geometric coupling of protein interfaces. *Proc. Natl Acad. Sci. USA* **109**, 8127–8132. (doi:10.1073/pnas.1120028109)
126. Chen J, Sawyer N, Regan L. 2013 Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* **22**, 510–515. (doi:10.1002/pro.2230)
127. Levy ED. 2010 A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* **403**, 660–670. (doi:10.1016/j.jmb.2010.09.028)
128. Davis FP, Sali A. 2010 The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput. Biol.* **6**, e1000668. (doi:10.1371/journal.pcbi.1000668)
129. Dasgupta B, Nakamura H, Kinjo AR. 2011 Distinct roles of overlapping and non-overlapping regions of hub protein interfaces in recognition of multiple partners. *J. Mol. Biol.* **411**, 713–727. (doi:10.1016/j.jmb.2011.06.027)
130. Levin KB, Dym O, Albeck S, Magdassi S, Keeble AH, Kleanthous C, Tawfik DS. 2009 Following evolutionary paths to protein–protein interactions with high affinity and selectivity. *Nat. Struct. Mol. Biol.* **16**, 1049–1055. (doi:10.1038/nsmb.1670)
131. Privalov PL, Gill SJ. 1988 Stability of protein structure and hydrophobic interaction. *Adv. Protein*

Chem. **39**, 191–234. (doi:10.1016/S0065-3233(08)60377-0)

132. Pace CN. 2001 Polar group burial contributes more to protein stability than nonpolar group burial. *Biochemistry* **40**, 310–313. (doi:10.1021/bi001574j)
133. Zavodszky P, Kardos J, Svingor A, Petsko GA. 1998 Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl Acad. Sci. USA* **95**, 7406–7411. (doi:10.1073/pnas.95.13.7406)
134. Beadle BM, Shoichet BK. 2002 Structural bases of stability–function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285–296. (doi:10.1016/S0022-2836(02)00599-5)
135. Taverna DM, Goldstein RA. 2002 Why are proteins marginally stable? *Proteins* **46**, 105–109. (doi:10.1002/prot.10016)
136. Goldstein RA. 2011 The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* **79**, 1396–1407. (doi:10.1002/prot.22964)
137. Bornberg-Bauer E, Chan HS. 1999 Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl Acad. Sci. USA* **96**, 10 689–10 694. (doi:10.1073/pnas.96.19.10689)
138. Gould SJ, Lewontin RC. 1979 The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B* **205**, 581–598. (doi:10.1098/rspb.1979.0086)
139. Shortle D, Stites WE, Meeker AK. 1990 Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* **29**, 8033–8041. (doi:10.1021/bi00487a007)
140. Green SM, Meeker AK, Shortle D. 1992 Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. *Biochemistry* **31**, 5717–5728. (doi:10.1021/bi00140a005)
141. Meeker AK, Garcia-Moreno B, Shortle D. 1996 Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry* **35**, 6443–6449. (doi:10.1021/bi960171+)
142. Itzhaki LS, Otzen DE, Fersht AR. 1995 The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288. (doi:10.1006/jmbi.1995.0616)
143. Serrano L, Day AG, Fersht AR. 1993 Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J. Mol. Biol.* **233**, 305–312. (doi:10.1006/jmbi.1993.1508)
144. Zhao H, Arnold FH. 1999 Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Eng. Des. Sel.* **12**, 47–53. (doi:10.1093/protein/12.1.47)
145. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. 2003 Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368. (doi:10.1126/science.1089427)
146. Plaxco KW, Simons KT, Baker D. 1998 Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994. (doi:10.1006/jmbi.1998.1645)
147. Miller EJ, Fischer KF, Marqusee S. 2002 Experimental evaluation of topological parameters determining protein-folding rates. *Proc. Natl Acad. Sci. USA* **99**, 10 359–10 363. (doi:10.1073/pnas.162219099)
148. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D. 2012 Principles for designing ideal protein structures. *Nature* **491**, 222–227. (doi:10.1038/nature11600)
149. Shortle D, Chan HS, Dill KA. 1992 Modeling the effects of mutations on the denatured states of proteins. *Protein Sci.* **1**, 201–215. (doi:10.1002/pro.5560010202)
150. Arnold FH, Wintrode PL, Miyazaki K, Gershenson A. 2001 How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **26**, 100–106. (doi:10.1016/S0968-0004(00)01755-2)
151. Van Nimwegen E, Crutchfield JP, Huynen M. 1999 Neutral evolution of mutational robustness. *Proc. Natl Acad. Sci. USA* **96**, 9716–9720. (doi:10.1073/pnas.96.17.9716)
152. Xia Y, Levitt M. 2002 Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc. Natl Acad. Sci. USA* **99**, 10 382–10 387. (doi:10.1073/pnas.162097799)
153. Bloom JD, Raval A, Wilke CO. 2007 Thermodynamics of neutral protein evolution. *Genetics* **175**, 255–266. (doi:10.1534/genetics.106.061754)
154. Watters AL, Deka P, Corrent C, Callender D, Varani G, Sosnick T, Baker D. 2007 The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell.* **128**, 613–624. (doi:10.1016/j.cell.2006.12.042)
155. Zhang Z, Chan HS. 2010 Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proc. Natl Acad. Sci. USA* **107**, 2920–2925. (doi:10.1073/pnas.0911844107)
156. Zhang Z, Chan HS. 2009 Native topology of the designed protein Top7 is not conducive to cooperative folding. *Biophys. J.* **96**, L25–L27. (doi:10.1016/j.bpj.2008.11.004)
157. Badasyan A, Liu Z, Chan HS. 2008 Probing possible downhill folding: native contact topology likely places a significant constraint on the folding cooperativity of proteins with∼40 residues. *J. Mol. Biol.* **384**, 512–530. (doi:10.1016/j.jmb.2008.09.023)
158. Chan HS, Dill KA. 1996 Comparing folding codes for proteins and polymers. *Proteins* **24**, 335–344. (doi:10.1002/(SICI)1097-0134(199603)24:3<335::AID-PROT6>3.0.CO;2-F)
159. Govindarajan S, Goldstein RA. 1995 Searching for foldable protein structures using optimized energy functions. *Biopolymers* **36**, 43–51. (doi:10.1002/bip.360360105)
160. Li H, Helling R, Tang C, Wingreen NS. 1996 Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669. (doi:10.1126/science.273.5275.666)
161. Bornberg-Bauer E. 1997 How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393–2403. (doi:10.1016/S0006-3495(97)78268-7)
162. Chan HS, Dill KA. 1991 'Sequence space soup' of proteins and copolymers. *J. Chem. Phys.* **95**, 3775–3787. (doi:10.1063/1.460828)
163. Kim YE, Hipp MS, Bracher A, Hayer-Hartl M, Hartl FU. 2013 Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* **82**, 323–355. (doi:10.1146/annurev-biochem-060208-092442)
164. Wagner GP, Altenberg L. 1996 Perspective: complex adaptations and the evolution of evolvability. *Evolution* **50**, 967–976. (doi:10.2307/2410639)
165. Tokuriki N, Tawfik DS. 2009 Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* **459**, 668–673. (doi:10.1038/nature08009)
166. Wyganowski KT, Kaltenbach M, Tokuriki N. 2013 GroEL/ES buffering and compensatory mutations promote protein evolution by stabilizing folding intermediates. *J. Mol. Biol.* **425**, 3403–3414. (doi:10.1016/j.jmb.2013.06.028)
167. Bogumil D, Dagan T. 2010 Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biol. Evol.* **2**, 602–608. (doi:10.1093/gbe/evq044)
168. Warnecke T, Hurst LD. 2010 GroEL dependency affects codon usage–support for a critical role of misfolding in gene evolution. *Mol. Syst. Biol.* **6**, 340. (doi:10.1038/msb.2009.94)
169. O'Brien EP, Vendruscolo M, Dobson CM. 2014 Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nat. Commun.* **5**, 2988. (doi:10.1038/ncomms3988)
170. Cetinbaş M, Shakhnovich EI. 2013 Catalysis of protein folding by chaperones accelerates evolutionary dynamics in adapting cell populations. *PLoS Comput. Biol.* **9**, e1003269. (doi:10.1371/journal.pcbi.1003269)
171. Kim H, Abeysirigunawarden SC, Chen K, Mayerle M, Ragunathan K, Luthey-Schulten Z, Ha T, Woodson SA. 2014 Protein-guided RNA dynamics during early ribosome assembly. *Nature* **506**, 334–338. (doi:10.1038/nature13039)
172. Seo M-H, Park J, Kim E, Hohng S, Kim H-S. 2014 Protein conformational dynamics dictate the binding affinity for a ligand. *Nat. Commun.* **5**, 3724. (doi:10.1038/ncomms4724)
173. Bouvignies G *et al*. 2011 Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature* **477**, 111–114. (doi:10.1038/nature10349)
174. Meier S, Jensen PR, David CN, Chapman J, Holstein TW, Grzesiek S, Özbek S. 2007 Continuous molecular evolution of protein-domain structures by single amino acid changes. *Curr. Biol.* **17**, 173–178. (doi:10.1016/j.cub.2006.10.063)
175. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. 2009 A minimal sequence code for switching protein structure and function. *Proc. Natl Acad. Sci.*

USA **106**, 21 149–21 154. (doi:10.1073/pnas.0906408106)

176. Anderson WJ, Van Dorn LO, Ingram WM, Cordes MHJ. 2011 Evolutionary bridges to new protein folds: design of C-terminal Cro protein chameleon sequences. *Protein Eng. Des. Sel.* **24**, 765–771. (doi:10.1093/protein/gzr027)
177. Zhuravlev PI, Papoian GA. 2010 Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Q. Rev. Biophys.* **3**, 1–38. (doi:10.1017/S0033583510000119)
178. Sikosek T, Bornberg-Bauer E, Chan HS. 2012 Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. *PLoS Comput. Biol.* **8**, e1002659. (doi:10.1371/journal.pcbi.1002659)
179. Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF. 2008 Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl Acad. Sci. USA* **105**, 5057–5062. (doi:10.1073/pnas.0709518105)
180. Luo X, Tang Z, Xia G, Wassmann K, Matsumoto T, Rizo J, Yu H. 2004 The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nat. Struct. Mol. Biol.* **11**, 338–345. (doi:10.1038/nsmb748)
181. Andersen JF, Ding XD, Balfour C, Shokhireva TK, Champagne DE, Walker FA, Montfort WR. 2000 Kinetics and equilibria in ligand binding by nitrophorins 1–4: evidence for stabilization of a nitric oxide-ferriheme complex through a ligand-induced conformational trap. *Biochemistry* **39**, 10 118–10 131. (doi:10.1021/bi000766b)
182. Ådén J, Verma A, Schug A, Wolf-Watz M. 2012 Modulation of a pre-existing conformational equilibrium tunes adenylate kinase activity. *J. Am. Chem. Soc.* **134**, 16 562–16 570. (doi:10.1021/ja3032482)
183. Burmann BM, Knauer SH, Sevostyanova A, Schweimer K, Mooney RA, Landick R, Artsimovitch I, Rösch P. 2012 An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **150**, 291–303. (doi:10.1016/j.cell.2012.05.042)
184. Di Russo NV, Estrin DA, Martí MA, Roitberg AE. 2012 pH-Dependent conformational changes in proteins and their effect on experimental pK_as: the case of Nitrophorin 4. *PLoS Comput. Biol.* **8**, e1002761. (doi:10.1371/journal.pcbi.1002761)
185. Monod J, Wyman J, Changeux J-P. 1965 On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118. (doi:10.1016/S0022-2836(65)80285-6)
186. Hilser VJ, Wrabl JO, Motlagh HN. 2012 Structural and energetic basis of allostery. *Annu. Rev. Biophys.* **41**, 585–609. (doi:10.1146/annurev-biophys-050511-102319)
187. Nussinov R, Tsai C-J. 2013 Allostery in disease and in drug discovery. *Cell* **153**, 293–305. (doi:10.1016/j.cell.2013.03.034)
188. Laskowski RA, Gerick F, Thornton JM. 2009 The structural basis of allosteric regulation in proteins. *FEBS Lett.* **583**, 1692–1698. (doi:10.1016/j.febslet.2009.03.019)
189. Zayner JP, Antoniou C, French AR, Hause RJ, Sosnick TR. 2013 Investigating models of protein function and allostery with a widespread mutational analysis of a light-activated protein. *Biophys. J.* **105**, 1027–1036. (doi:10.1016/j.bpj.2013.07.010)
190. Weinkam P, Chen YC, Pons J, Sali A. 2013 Impact of mutations on the allosteric conformational equilibrium. *J. Mol. Biol.* **425**, 647–661. (doi:10.1016/j.jmb.2012.11.041)
191. Keskin O, Jernigan RL, Bahar I. 2000 Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys. J.* **78**, 2093–2106. (doi:10.1016/S0006-3495(00)76756-7)
192. Micheletti C, Lattanzi G, Maritan A. 2002 Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures. *J. Mol. Biol.* **321**, 909–921. (doi:10.1016/S0022-2836(02)00710-6)
193. Zheng W, Brooks BR, Thirumalai D. 2006 Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl Acad. Sci. USA* **103**, 7664–7669. (doi:10.1073/pnas.0510426103)
194. Zen A, Carnevale V, Lesk AM, Micheletti C. 2008 Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. *Protein Sci.* **17**, 918–929. (doi:10.1110/ps.073390208)
195. Taketomi H, Ueda Y, Gō N. 1975 Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7**, 445–459. (doi:10.1111/j.1399-3011.1975.tb02465.x)
196. Schug A, Whitford PC, Levy Y, Onuchic JN. 2007 Mutations as trapdoors to two competing native conformations of the Rop-dimer. *Proc. Natl Acad. Sci. USA* **104**, 17 674–17 679. (doi:10.1073/pnas.0706077104)
197. Micheletti C. 2013 Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. *Phys. Life Rev.* **10**, 1–26. (doi:10.1016/j.plrev.2012.10.009)
198. Liu Y, Bahar I. 2012 Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.* **29**, 2253–2263. (doi:10.1093/molbev/mss097)
199. Peracchi A, Mozzarelli A. 2011 Exploring and exploiting allostery: models, evolution, and drug targeting. *Biochim. Biophys. Acta Proteins Proteomics.* **1814**, 922–933. (doi:10.1016/j.bbapap.2010.10.008)
200. Boehr DD, Nussinov R, Wright PE. 2009 The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796. (doi:10.1038/nchembio.232)
201. Coyle SM, Flores J, Lim WA. 2013 Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell* **154**, 875–887. (doi:10.1016/j.cell.2013.07.019)
202. Amitai G, Gupta RD, Tawfik DS. 2007 Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1**, 67–78. (doi:10.2976/1.2739115)
203. Wroe R, Chan HS, Bornberg-Bauer E. 2007 A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J.* **1**, 79–87. (doi:10.2976/1.2739116)
204. Chan HS, Kaya H, Shimizu S. 2002 Computational methods for protein folding: scaling a hierarchy of complexities. In *Current topics in computational molecular biology* (eds T Jiang, Y Xu, MQ Zhang), pp. 403–447. Cambridge, MA: The MIT Press.
205. Cordes MHJ, Walsh NP, Mcknight CJ, Sauer RT. 1999 Evolution of a protein fold *in vitro*. *Science* **284**, 325–327. (doi:10.1126/science.284.5412.325)
206. He Y, Chen Y, Alexander PA, Bryan PN, Orban J. 2012 Mutational tipping points for switching protein folds and functions. *Structure* **20**, 283–291. (doi:10.1016/j.str.2011.11.018)
207. Stewart KL, Dodds ED, Wysocki VH, Cordes MHJ. 2013 A polymetamorphic protein. *Protein Sci.* **22**, 641–649. (doi:10.1002/pro.2248)
208. Minor DL, Kim PS. 1996 Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730–734. (doi:10.1038/380730a0)
209. Ambroggio XI, Kuhlman B. 2006 Design of protein conformational switches. *Curr. Opin. Struct. Biol.* **16**, 525–530. (doi:10.1016/j.sbi.2006.05.014)
210. Dagliyan O *et al*. 2013 Rational design of a ligand-controlled protein conformational switch. *Proc. Natl Acad. Sci. USA* **110**, 6800–6804. (doi:10.1073/pnas.1218319110)
211. Meier S, Özbek S. 2007 A biological cosmos of parallel universes: does protein structural plasticity facilitate evolution? *BioEssays* **29**, 1095–1104. (doi:10.1002/bies.20661)
212. Bryan PN, Orban J. 2010 Proteins that switch folds. *Curr. Opin. Struct. Biol.* **20**, 482–488. (doi:10.1016/j.sbi.2010.06.002)
213. James LC, Roversi P, Tawfik DS. 2003 Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362–1367. (doi:10.1126/science.1079731)
214. Franco OL. 2011 Peptide promiscuity: an evolutionary concept for plant defense. *FEBS Lett.* **585**, 995–1000. (doi:10.1016/j.febslet.2011.03.008)
215. Caines MEC, Bichel K, Price AJ, McEwan WA, Towers GJ, Willett BJ, Freund SMV, James LC. 2012 Diverse HIV viruses are targeted by a conformationally dynamic antiviral. *Nat. Struct. Mol. Biol.* **19**, 411–416. (doi:10.1038/nsmb.2253)
216. Dunker AK *et al*. 2001 Intrinsically disordered protein. *J. Mol. Graph Model* **19**, 26–59. (doi:10.1016/S1093-3263(00)00138-8)
217. Tompa P. 2002 Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533. (doi:10.1016/S0968-0004(02)02169-2)
218. Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R. 2003 Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* **28**, 81–85. (doi:10.1016/S0968-0004(03)00003-3)
219. Dyson HJ, Wright PE. 2005 Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208. (doi:10.1038/nrm1589)

220. Fuxreiter M, Simon I, Bondos S. 2011 Dynamic protein-DNA recognition: beyond what can be seen. *Trends Biochem. Sci.* **36**, 415–423. (doi:10.1016/j.tibs.2011.04.006)
221. Marsh JA, Teichmann SA, Forman-Kay JD. 2012 Probing the diverse landscape of protein flexibility and binding. *Curr. Opin. Struct. Biol.* **22**, 643–650. (doi:10.1016/j.sbi.2012.08.008)
222. Uversky VN. 2013 A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci.* **22**, 693–724. (doi:10.1002/pro.2261)
223. Monsellier E, Chiti F. 2007 Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* **8**, 737–742. (doi:10.1038/sj.embor.7401034)
224. Greenwald J, Riek R. 2012 On the possible amyloid origin of protein folds. *J. Mol. Biol.* **421**, 417–426. (doi:10.1016/j.jmb.2012.04.015)
225. Trifonov EN. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151. (doi:10.1016/S0378-1119(00)00476-5)
226. Mannige RV, Brooks CL, Shakhnovich EI. 2012 A universal trend among proteomes indicates an oily last common ancestor. *PLoS Comput. Biol.* **8**, e1002839. (doi:10.1371/journal.pcbi.1002839)
227. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. 2001 Sequence complexity of disordered protein. *Proteins* **42**, 38–48. (doi:10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3)
228. Rauscher S, Baud S, Miao M, Keeley FW, Pomès R. 2006 Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. *Structure* **14**, 1667–1676. (doi:10.1016/j.str.2006.09.008)
229. Liu Z, Huang Y. 2014 Advantages of proteins being disordered. *Protein Sci.* **23**, 539–550. (doi:10.1002/pro.2443)
230. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. 2005 Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148. (doi:10.1111/j.1742-4658.2005.04948.x)
231. Cumberworth A, Lamour G, Babu MM, Gsponer J. 2013 Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.* **454**, 361–369. (doi:10.1042/BJ20130545)
232. Borg M, Mittag T, Pawson T, Tyers M, Forman-Kay JD, Chan HS. 2007 Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl Acad. Sci. USA* **104**, 9650–9655. (doi:10.1073/pnas.0702580104)
233. Tompa P, Fuxreiter M. 2008 Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **33**, 2–8. (doi:10.1016/j.tibs.2007.10.003)
234. Mittag T, Orlicky S, Choy W-Y, Tang X, Lin H, Sicheri F, Kay LE, Tyers M, Forman-Kay JD. 2008 Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl Acad. Sci. USA* **105**, 17 772–17 777. (doi:10.1073/pnas.0809222105)
235. Mittag T, Marsh JA, Grishaev A, Orlicky S, Lin H, Sicheri F, Tyers M, Forman-Kay JD. 2010 Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**, 494–506. (doi:10.1016/j.str.2010.01.020)
236. Song J, Ng SC, Tompa P, Lee KAW, Chan HS. 2013 Polycation-π interactions are a driving force for molecular recognition by an intrinsically disordered oncoprotein family. *PLoS Comput. Biol.* **9**, e1003239. (doi:10.1371/journal.pcbi.1003239)
237. Kimura M. 1968 Evolutionary rate at the molecular level. *Nature* **217**, 624–626. (doi:10.1038/217624a0)
238. Ohta T. 1973 Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98. (doi:10.1038/246096a0)
239. Sikosek T, Chan HS, Bornberg-Bauer E. 2012 Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *Proc. Natl Acad. Sci. USA* **109**, 14 888–14 893. (doi:10.1073/pnas.1115620109)
240. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002 Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110. (doi:10.1007/s00239-001-2309-6)
241. Nilsson J, Grahn M, Wright APH. 2011 Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol.* **12**, R65. (doi:10.1186/gb-2011-12-7-r65)
242. Huang H, Sarai A. 2012 Analysis of the relationships between evolvability, thermodynamics, and the functions of intrinsically disordered proteins/regions. *Comput. Biol. Chem.* **41**, 51–57. (doi:10.1016/j.compbiolchem.2012.10.001)
243. Marsh JA, Teichmann SA. 2014 Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays* **36**, 209–218. (doi:10.1002/bies.201300134)
244. Brown CJ, Johnson AK, Daughdrill GW. 2010 Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.* **27**, 609–621. (doi:10.1093/molbev/msp277)
245. Moesa HA, Wakabayashi S, Nakai K, Patil A. 2012 Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol. Biosyst.* **8**, 3262–3273. (doi:10.1039/c2mb25202c)
246. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. 2011 Evolution and disorder. *Curr. Opin. Struct. Biol.* **21**, 441–446. (doi:10.1016/j.sbi.2011.02.005)
247. Nash P, Tang X, Orlicky S, Chen Q, Gertler FB, Mendenhall MD, Sicheri F, Pawson T, Tyers M. 2001 Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* **414**, 514–521. (doi:10.1038/35107009)
248. Mittermaier A, Kay LE. 2006 New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224–228. (doi:10.1126/science.1124964)
249. Henzler-Wildman K, Kern D. 2007 Dynamic personalities of proteins. *Nature* **450**, 964–972. (doi:10.1038/nature06522)
250. Lange OF *et al*. 2008 Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471–1475. (doi:10.1126/science.1157092)
251. Choy WY, Forman-Kay JD. 2001 Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **308**, 1011–1032. (doi:10.1006/jmbi.2001.4750)
252. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo MH. 2005 Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128–132. (doi:10.1038/nature03199)
253. Frauenfelder H, Sligar SG, Wolynes PG. 1991 The energy landscapes and motions of proteins. *Science* **254**, 1598–1603. (doi:10.1126/science.1749933)
254. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995 Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195. (doi:10.1002/prot.340210302)
255. Wrabl JO, Gu J, Liu T, Schrank TP, Whitten ST, Hilser VJ. 2011 The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys. Chem.* **159**, 129–141. (doi:10.1016/j.bpc.2011.05.020)
256. Bastolla U, Porto M, Roman HE. 2013 The emerging dynamic view of proteins: protein plasticity in allostery, evolution and self-assembly. *Biochim. Biophys. Acta* **1834**, 817–819. (doi:10.1016/j.bbapap.2013.03.016)
257. Chevin L-M, Lande R, Mace GM. 2010 Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biol.* **8**, e1000357. (doi:10.1371/journal.pbio.1000357)
258. Sato K, Ito Y, Yomo T, Kaneko K. 2003 On the relation between fluctuation and response in biological systems. *Proc. Natl Acad. Sci. USA* **100**, 14 086–14 090. (doi:10.1073/pnas.2334996100)
259. Chen T, Vernazobres D, Yomo T, Bornberg-Bauer E, Chan HS. 2010 Evolvability and single-genotype fluctuation in phenotypic properties: a simple heteropolymer model. *Biophys. J.* **98**, 2487–2496. (doi:10.1016/j.bpj.2010.02.046)
260. Wagner A. 2014 Mutational robustness accelerates the origin of novel RNA phenotypes through phenotypic plasticity. *Biophys. J.* **106**, 955–965. (doi:10.1016/j.bpj.2014.01.003)
261. Rosenberg SM. 2001 Evolving responsively: adaptive mutation. *Nat. Rev. Genet.* **2**, 504–515. (doi:10.1038/35080556)
262. Earl DJ, Deem MW. 2004 Evolvability is a selectable trait. *Proc. Natl Acad. Sci. USA* **101**, 11 531–11 536. (doi:10.1073/pnas.0404656101)
263. Jeffery CJ. 1999 Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11.
264. Khersonsky O, Tawfik DS. 2010 Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505. (doi:10.1146/annurev-biochem-030409-143718)
265. Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, Tawfik DS. 2005 The 'evolvability' of

promiscuous protein functions. *Nat. Genet.* **37**, 73–76. (doi:10.1038/ng1482)

266. Hou J, Sims GE, Zhang C, Kim S-H. 2003 A global representation of the protein fold space. *Proc. Natl Acad. Sci. USA* **100**, 2386–2390. (doi:10.1073/pnas.2628030100)
267. Sippl MJ. 2009 Fold space unlimited. *Curr. Opin. Struct. Biol.* **19**, 312–320. (doi:10.1016/j.sbi.2009.03.010)
268. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. 2009 The origin, evolution and structure of the protein world. *Biochem. J.* **417**, 621–637. (doi:10.1042/BJ20082063)
269. Caetano-Anollés G, Kim KM, Caetano-Anollés D. 2012 The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein biosynthesis. *J. Mol. Evol.* **74**, 1–34. (doi:10.1007/s00239-011-9480-1)
270. Osadchy M, Kolodny R. 2011 Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc. Natl Acad. Sci. USA* **108**, 12 301–12 306. (doi:10.1073/pnas.1102727108)
271. Minary P, Levitt M. 2008 Probing protein fold space with a simplified model. *J. Mol. Biol.* **375**, 920–933. (doi:10.1016/j.jmb.2007.10.087)
272. Keefe AD, Szostak JW. 2001 Functional proteins from a random-sequence library. *Nature* **410**, 715–718. (doi:10.1038/35070613)
273. Povolotskaya IS, Kondrashov FA. 2010 Sequence space and the ongoing expansion of the protein universe. *Nature* **465**, 922–926. (doi:10.1038/nature09105)
274. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540. (doi:10.1006/jmbi.1995.0159)
275. Wolf YI, Grishin NV, Koonin EV. 2000 Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897–905. (doi:10.1006/jmbi.2000.3786)
276. Govindarajan S, Recabarren R, Goldstein RA. 1999 Estimating the total number of protein folds. *Proteins* **35**, 408–414. (doi:10.1002/(SICI)1097-0134(19990601)35:4<408::AID-PROT4>3.0.CO;2-A)
277. Coulson AFW, Moult J. 2002 A unifold, mesofold, and superfold model of protein fold use. *Proteins* **46**, 61–71. (doi:10.1002/prot.10011)
278. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. 2013 On the universe of protein folds. *Annu. Rev. Biophys.* **42**, 559–582. (doi:10.1146/annurev-biophys-083012-130432)
279. Godzik A. 2011 Metagenomics and the protein universe. *Curr. Opin. Struct. Biol.* **21**, 398–403. (doi:10.1016/j.sbi.2011.03.010)
280. Chan HS, Dill KA. 1990 The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**, 3118–3135. (doi:10.1063/1.458605)
281. Chan HS, Dill KA. 1990 Origins of structure in globular proteins. *Proc. Natl Acad. Sci. USA* **87**, 6388–6392. (doi:10.1073/pnas.87.16.6388)
282. Maritan A, Micheletti C, Trovato A, Banavar JR. 2000 Optimal shapes of compact strings. *Nature* **406**, 287–290. (doi:10.1038/35018538)
283. Gregoret LM, Cohen FE. 1991 Protein folding. Effect of packing density on chain conformation. *J. Mol. Biol.* **219**, 109–122. (doi:10.1016/0022-2836(91)90861-Y)
284. Hunt NG, Gregoret LM, Cohen FE. 1994 The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search. *J. Mol. Biol.* **241**, 214–225. (doi:10.1006/jmbi.1994.1490)
285. Yee DP, Chan HS, Havel TF, Dill KA. 1994 Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. *J. Mol. Biol.* **241**, 557–573. (doi:10.1006/jmbi.1994.1531)
286. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich EI, Skolnick J. 2006 On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl Acad. Sci. USA* **103**, 2605–2610. (doi:10.1073/pnas.0509379103)
287. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I. 2009 Probing the 'dark matter' of protein fold space. *Structure* **17**, 1244–1252. (doi:10.1016/j.str.2009.07.012)
288. Cossio P, Trovato A, Pietrucci F, Seno F, Maritan A, Laio A. 2010 Exploring the universe of protein structures beyond the Protein Data Bank. *PLoS Comput. Biol.* **6**, e1000957. (doi:10.1371/journal.pcbi.1000957)
289. Dai L, Zhou Y. 2011 Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. *J. Mol. Biol.* **408**, 585–595. (doi:10.1016/j.jmb.2011.02.056)
290. Skolnick J, Zhou H, Brylinski M. 2012 Further evidence for the likely completeness of the library of solved single domain protein structures. *J. Phys. Chem. B* **116**, 6654–6664. (doi:10.1021/jp211052j)
291. Skolnick J, Gao M. 2013 Interplay of physics and evolution in the likely origin of protein biochemical function. *Proc. Natl Acad. Sci. USA* **110**, 9344–9349. (doi:10.1073/pnas.1300011110)
292. Chan HS, Bornberg-Bauer E. 2002 Perspectives on protein evolution from simple exact models. *Appl. Bioinform.* **1**, 121–144.
293. Xia Y, Levitt M. 2004 Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins* **55**, 107–114. (doi:10.1002/prot.10563)
294. Greenbury SF, Johnston IG, Louis AA, Ahnert SE. 2014 A tractable genotype-phenotype map modelling the self-assembly of protein quaternary structure. *J. R. Soc. Interface* **11**, 20140249. (doi:10.1098/rsif.2014.0249)
295. Moreno-Hernández S, Levitt M. 2012 Comparative modeling and protein-like features of hydrophobic-polar models on a two-dimensional lattice. *Proteins* **80**, 1683–1693. (doi:10.1002/prot.24067)
296. Palmer ME, Moudgil A, Feldman MW. 2013 Long-term evolution is surprisingly predictable in lattice proteins. *J. R. Soc. Interface* **10**, 20130026. (doi:10.1098/rsif.2013.0026)
297. Lau KF, Dill KA. 1990 Theory for protein mutability and biogenesis. *Proc. Natl Acad. Sci. USA* **87**, 638–642. (doi:10.1073/pnas.87.2.638)
298. Lipman DJ, Wilbur WJ. 1991 Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B* **245**, 7–11. (doi:10.1098/rspb.1991.0081)
299. Holzgräfe C, Irbäck A, Troein C. 2011 Mutation-induced fold switching among lattice proteins. *J. Chem. Phys.* **135**, 195101. (doi:10.1063/1.3660691)
300. Miller C, Davlieva M, Wilson C, White KI, Couñago R, Wu G, Myers JC, Wittung-Stafshede P, Shamoo Y. 2010 Experimental evolution of adenylate kinase reveals contrasting strategies toward protein thermostability. *Biophys. J.* **99**, 887–896. (doi:10.1016/j.bpj.2010.04.076)
301. Hirst JD. 1999 The evolutionary landscape of functional model proteins. *Protein Eng. Des. Sel.* **12**, 721–726. (doi:10.1093/protein/12.9.721)
302. Blackburne BP, Hirst JD. 2001 Evolution of functional model proteins. *J. Chem. Phys.* **115**, 1935. (doi:10.1063/1.1383051)
303. Burke S, Elber R. 2011 Super folds, networks, and barriers. *Proteins* **80**, 463–470. (doi:10.1002/prot.23212)
304. Noirel J, Simonson T. 2008 Neutral evolution of proteins: the superfunnel in sequence space and its relation to mutational robustness. *J. Chem. Phys.* **129**, 185104. (doi:10.1063/1.2992853)
305. Bastolla U, Roman HE, Vendruscolo MH. 1999 Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* **200**, 49–64. (doi:10.1006/jtbi.1999.0975)
306. Taverna DM, Goldstein RA. 2002 Why are proteins so robust to site mutations? *J. Mol. Biol.* **315**, 479–484. (doi:10.1006/jmbi.2001.5226)
307. Deeds EJ, Shakhnovich EI. 2005 The emergence of scaling in sequence-based physical models of protein evolution. *Biophys. J.* **88**, 3905–3911. (doi:10.1529/biophysj.104.051433)
308. Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI. 2007 A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. *PLoS Comput. Biol.* **3**, e139. (doi:10.1371/journal.pcbi.0030139)
309. Chan HS, Dill KA. 1989 Compact polymers. *Macromolecules* **22**, 4559–4573. (doi:10.1021/ma00202a031)
310. Pande VS, Joerg C, Grosberg AY, Tanaka T. 1994 Enumerations of the Hamiltonian walks on a cubic sublattice. *J. Phys. A Math. Gen.* **27**, 6231–6236. (doi:10.1088/0305-4470/27/18/030)
311. Lee JH, Kim S-Y, Lee J. 2011 Parallel algorithm for calculation of the exact partition function of a lattice polymer. *Comput. Phys. Commun.* **182**, 1027–1033. (doi:10.1016/j.cpc.2011.01.004)
312. Schram RD, Schiessel H. 2013 Exact enumeration of Hamiltonian walks on the $4\times4\times4$ cube and applications to protein folding. *J. Phys. A Math. Theor.* **46**, 485001. (doi:10.1088/1751-8113/46/48/485001)

313. Lee J. 2004 Exact partition function zeros of two-dimensional lattice polymers. *J. Korean Phys. Soc.* **44**, 617–620. (doi:10.3938/jkps.44.617)
314. Clisby N, Liang R, Slade G. 2007 Self-avoiding walk enumeration via the lace expansion. *J. Phys. A Math. Theor.* **40**, 10 973–11 017. (doi:10.1088/1751-8113/40/36/003)
315. Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995 A test of lattice protein folding algorithms. *Proc. Natl Acad. Sci. USA* **92**, 325–329. (doi:10.1073/pnas.92.1.325)
316. Irbäck A, Troein C. 2002 Enumerating designing sequences in the HP model. *J. Biol. Phys.* **28**, 1–15. (doi:10.1023/A:1016225010659)
317. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. 1995 Principles of protein folding—a perspective from simple exact models. *Protein Sci.* **4**, 561–602. (doi:10.1002/pro.5560040401)
318. Kamtekar S, Schiffer J, Xiong H, Babik J, Hecht M. 1993 Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680–1685. (doi:10.1126/science.8259512)
319. Urvoas A, Valerio-Lepiniec M, Minard P. 2012 Artificial proteins from combinatorial approaches. *Trends Biotechnol.* **30**, 512–520. (doi:10.1016/j.tibtech.2012.06.001)
320. Irbäck A, Sandelin E. 2000 On hydrophobicity correlations in protein chains. *Biophys. J.* **79**, 2252–2258. (doi:10.1016/S0006-3495(00)76472-1)
321. Irbäck A, Peterson C, Potthast F. 1996 Evidence for nonrandom hydrophobicity structures in protein chains. *Proc. Natl Acad. Sci. USA* **93**, 9533–9538. (doi:10.1073/pnas.93.18.9533)
322. Buchler NE, Goldstein RA. 1999 Effect of alphabet size and foldability requirements on protein structure designability. *Proteins* **34**, 113–124. (doi:10.1002/(SICI)1097-0134(19990101)34:1<113::AID-PROT9>3.0.CO;2-J)
323. Wroe R, Bornberg-Bauer E, Chan HS. 2005 Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys. J.* **88**, 118–131. (doi:10.1529/biophysj.104.050369)
324. Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. 2002 Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl Acad. Sci. USA* **99**, 809–814. (doi:10.1073/pnas.022240299)
325. Chan HS. 2000 Modeling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins* **40**, 543–571. (doi:10.1002/1097-0134(20000901)40:4<543::AID-PROT20>3.0.CO;2-0)
326. Chan HS. 1998 Protein folding. Matching speed and locality. *Nature* **392**, 761–763. (doi:10.1038/33808)
327. Gō N. 1983 Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210. (doi:10.1146/annurev.bb.12.060183.001151)
328. Bryngelson JD, Wolynes PG. 1987 Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA* **84**, 7524–7528. (doi:10.1073/pnas.84.21.7524)
329. Miyazawa S, Jernigan RL. 1985 Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552. (doi:10.1021/ma00145a039)
330. Bloom JD, Wilke CO, Arnold FH, Adami C. 2004 Stability and the evolvability of function in a model protein. *Biophys. J.* **86**, 2758–2764. (doi:10.1016/S0006-3495(04)74329-5)
331. Abkevich VI, Gutin AM, Shakhnovich EI. 1994 Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10 026–10 036. (doi:10.1021/bi00199a029)
332. Xia Y, Levitt M. 2004 Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.* **14**, 202–207. (doi:10.1016/j.sbi.2004.03.001)
333. Goldstein RA. 2008 The structure of protein evolution and the evolution of protein structure. *Curr. Opin. Struct. Biol.* **18**, 170–177. (doi:10.1016/j.sbi.2008.01.006)
334. Zeldovich KB, Shakhnovich EI. 2008 Understanding protein evolution: from protein physics to Darwinian selection. *Annu. Rev. Phys. Chem.* **59**, 105–127. (doi:10.1146/annurev.physchem.58.032806.104449)
335. Blackburne BP, Hirst JD. 2005 Population dynamics simulations of functional model proteins. *J. Chem. Phys.* **123**, 154907. (doi:10.1063/1.2056545)
336. Maynard Smith J. 1970 Natural selection and the concept of a protein space. *Nature* **225**, 563–564. (doi:10.1038/225563a0)
337. Bastolla U, Vendruscolo MH, Roman HE. 2000 Structurally constrained protein evolution: results from a lattice simulation. *Eur. Phys. J. B* **15**, 385–397. (doi:10.1007/s100510051140)
338. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. 2006 Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874. (doi:10.1073/pnas.0510098103)
339. Drummond DA, Silberg JJ, Meyer MM, Wilke CO, Arnold FH. 2005 On the conservative nature of intragenic recombination. *Proc. Natl Acad. Sci. USA* **102**, 5380–5385. (doi:10.1073/pnas.0500729102)
340. Cui Y, Wong W. 2000 Multiple-sequence information provides protection against mis-specified potential energy functions in the lattice model of proteins. *Phys. Rev. Lett.* **85**, 5242–5245. (doi:10.1103/PhysRevLett.85.5242)
341. Nanda V, DeGrado WF. 2005 Automated use of mutagenesis data in structure prediction. *Proteins* **59**, 454–466. (doi:10.1002/prot.20382)
342. England JL, Shakhnovich BE, Shakhnovich EI. 2003 Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl Acad. Sci. USA* **100**, 8727–8731. (doi:10.1073/pnas.1530713100)
343. Noivirt-Brik O, Unger R, Horovitz A. 2009 Analysing the origin of long-range interactions in proteins using lattice models. *BMC Struct. Biol.* **9**, 4. (doi:10.1186/1472-6807-9-4)
344. Liu Z, Chen J, Thirumalai D. 2009 On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: illustrations using lattice model. *Proteins* **77**, 823–831. (doi:10.1002/prot.22498)
345. Heo M, Kang L, Shakhnovich EI. 2009 Emergence of species in evolutionary 'simulated annealing'. *Proc. Natl Acad. Sci. USA* **106**, 1869–1874. (doi:10.1073/pnas.0809852106)
346. Heo M, Shakhnovich EI. 2010 Interplay between pleiotropy and secondary selection determines rise and fall of mutators in stress response. *PLoS Comput. Biol.* **6**, e1000710. (doi:10.1371/journal.pcbi.1000710)
347. Heo M, Maslov S, Shakhnovich EI. 2011 Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc. Natl Acad. Sci. USA* **108**, 4258–4263. (doi:10.1073/pnas.1009392108)
348. Wright S. 1932 The roles of mutations, inbreeding, crossbreeding and selection in evolution. In *Proc. 6th Int. Congr. Genet*, vol. 1, pp. 356–366. Menasha, WI: Brooklyn Botanical Garden.
349. Kauffman S, Levin S. 1987 Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45. (doi:10.1016/S0022-5193(87)80029-2)
350. Voigt CA, Kauffman S, Wang ZG. 2000 Rational evolutionary design: the theory of *in vitro* protein evolution. *Adv. Protein Chem.* **55**, 79–160. (doi:10.S0065-3233(01)55003-2)
351. Carneiro M, Hartl DL. 2010 Adaptive landscapes and protein evolution. *Proc. Natl Acad. Sci. USA* **107**, 1747–1751. (doi:10.1073/pnas.0906192106)
352. Sella G, Hirsh AE. 2005 The application of statistical physics to evolutionary biology. *Proc. Natl Acad. Sci. USA* **102**, 9541–9546. (doi:10.1073/pnas.0501865102)
353. Pathria R. 1980 *Statistical mechanics*. Oxford, UK: Pergamon Press.
354. Lobkovsky AE, Wolf YI, Koonin EV. 2013 Quantifying the similarity of monotonic trajectories in rough and smooth fitness landscapes. *Mol. Biosyst.* **9**, 1627–1631. (doi:10.1039/c3mb25553k)
355. Ferrada E, Wagner A. 2012 A comparison of genotype–phenotype maps for RNA and proteins. *Biophys. J.* **102**, 1916–1925. (doi:10.1016/j.bpj.2012.01.047)
356. Fontana W, Schuster P. 1987 A computer model of evolutionary optimization. *Biophys. Chem.* **26**, 123–147. (doi:10.1016/0301-4622(87)80017-0)
357. Fontana W, Stadler P, Bornberg-Bauer E, Griesmacher T, Hofacker I, Tacker M, Tarazona P, Weinberger E, Schuster P. 1993 RNA folding and combinatory landscapes. *Phys. Rev. E* **47**, 2083–2099. (doi:10.1103/PhysRevE.47.2083)
358. Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994 From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B* **255**, 279–284. (doi:10.1098/rspb.1994.0040)
359. Ancel LW, Fontana W. 2000 Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* **288**, 242–283. (doi:10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-0)
360. Guo HH, Choe J, Loeb LA. 2004 Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**, 9205–9210. (doi:10.1073/pnas.0403255101)
361. Punta M *et al*. 2012 The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301. (doi:10.1093/nar/gkr1065)

362. Thorne JL, Goldman N, Jones DT. 1996 Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**, 666–673. (doi:10.1093/oxfordjournals.molbev.a025627)
363. Goldman N, Thorne JL, Jones DT. 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458.
364. Massey SE. 2010 Pseudaptations and the emergence of beneficial traits. In *Evolutionary biology—concepts, molecular and morphological evolution* (ed. P Pontarotti), pp. 81–98. Berlin, Germany: Springer.
365. Kitano H. 2004 Biological robustness. *Nat. Rev. Genet.* **5**, 826–837. (doi:10.1038/nrg1471)
366. Masel J, Siegal ML. 2009 Robustness: mechanisms and consequences. *Trends Genet.* **25**, 395–403. (doi:10.1016/j.tig.2009.07.005)
367. Masel J, Trotter MV. 2010 Robustness and evolvability. *Trends Genet.* **26**, 406–414. (doi:10.1016/j.tig.2010.06.002)
368. Rorick MM, Wagner GP. 2011 Protein structural modularity and robustness are associated with evolvability. *Genome Biol. Evol.* **3**, 456–475. (doi:10.1093/gbe/evr046)
369. Wagner A. 2008 Neutralism and selectionism: a network-based reconciliation. *Nat. Rev. Genet.* **9**, 965–974. (doi:10.1038/nrg2473)
370. Wagner A. 2008 Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B* **275**, 91–100. (doi:10.1098/rspb.2007.1137)
371. Draghi JA, Parsons TL, Wagner GP, Plotkin JB. 2010 Mutational robustness can facilitate adaptation. *Nature* **463**, 353–355. (doi:10.1038/nature08694)
372. Bornberg-Bauer E, Kramer L. 2010 Robustness versus evolvability: a paradigm revisited. *HFSP J.* **4**, 105–108. (doi:10.2976/1.3404403)
373. Nobeli I, Favia AD, Thornton JM. 2009 Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167. (doi:10.1038/nbt1519)
374. Babtie A, Tokuriki N, Hollfelder F. 2010 What makes an enzyme promiscuous? *Curr. Opin. Chem. Biol.* **14**, 200–207. (doi:10.1016/j.cbpa.2009.11.028)
375. Schreiber G, Keating AE. 2011 Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol.* **21**, 50–61. (doi:10.1016/j.sbi.2010.10.002)
376. Bridgham JT, Carroll SM, Thornton JW. 2006 Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**, 97–101. (doi:10.1126/science.1123348)
377. Rebeiz M, Jikomes N, Kassner VA, Carroll SB. 2011 Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proc. Natl Acad. Sci. USA* **108**, 10 036–10 043. (doi:10.1073/pnas.1105937108)
378. Barve A, Wagner A. 2013 A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203–206. (doi:10.1038/nature12301)
379. Gould S, Vrba E. 1982 Exaptation-a missing term in the science of form. *Paleobiology* **8**, 4–15.
380. Brosius J, Gould SJ. 1992 On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl Acad. Sci. USA* **89**, 10 706–10 710. (doi:10.1093/oxfordjournals.molbev.a025627)
381. Brosius J. 1999 RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**, 115–134. (doi:10.1016/S0378-1119(99)00227-9)
382. Krull M, Brosius J, Schmitz J. 2005 Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* **22**, 1702–1711. (doi:10.1093/molbev/msi164)
383. Carvunis A-R *et al*. 2012 Proto-genes and de novo gene birth. *Nature* **487**, 370–374. (doi:10.1038/nature11184)
384. Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T. 2009 Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673. (doi:10.1038/nature08615)
385. Vallurupalli P, Bouvignies G, Kay LE. 2012 Studying 'invisible' excited protein states in slow exchange with a major state conformation. *J. Am. Chem. Soc.* **134**, 8148–8161. (doi:10.1021/ja3001419)
386. Sekhar A, Kay LE. 2013 NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl Acad. Sci. USA* **110**, 12 867–12 874. (doi:10.1073/pnas.1305688110)
387. Dimitrov JD, Kaveri SV, Lacroix-Desmazes S. 2014 Thermodynamic stability contributes to immunoglobulin specificity. *Trends Biochem. Sci.* **39**, 221–226. (doi:10.1016/j.tibs.2014.02.010)
388. Sabath N, Wagner A, Karlin D. 2012 Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.* **29**, 3767–3780. (doi:10.1093/molbev/mss179)
389. Conant GC, Wolfe KH. 2008 Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* **9**, 938–950. (doi:10.1038/nrg2482)
390. Soskine M, Tawfik DS. 2010 Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet.* **11**, 572–582. (doi:10.1038/nrg2808)
391. Ugalde JA, Chang BSW, Matz MV. 2004 Evolution of coral pigments recreated. *Science* **305**, 1433. (doi:10.1126/science.1099597)
392. Gutiérrez J, Maere S. 2014 Modeling the evolution of molecular systems from a mechanistic perspective. *Trends Plant Sci.* **19**, 292–303. (doi:10.1016/j.tplants.2014.03.004)
393. Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. 2013 Genetic incompatibilities are widespread within species. *Nature* **504**, 135–137. (doi:10.1038/nature12678)
394. Wells JA. 1990 Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509–8517. (doi:10.1021/bi00489a001)
395. Weinreich DM, Delaney NF, Depristo MA, Hartl DL. 2006 Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114. (doi:10.1126/science.1123539)
396. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. 2007 Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**, 1544–1548. (doi:10.1126/science.1142819)
397. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012 Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538. (doi:10.1038/nature11510)
398. Soylemez O, Kondrashov FA. 2012 Estimating the rate of irreversibility in protein evolution. *Genome Biol. Evol.* **4**, 1213–1222. (doi:10.1093/gbe/evs096)
399. Pollock DD, Goldstein RA. 2014 Strong evidence for protein epistasis, weak evidence against it. *Proc. Natl Acad. Sci. USA* **111**, E1450. (doi:10.1073/pnas.1401112111)
400. Ferrer-Costa C, Orozco M, de la Cruz X. 2004 Sequence-based prediction of pathological mutations. *Proteins* **57**, 811–819. (doi:10.1002/prot.20252)
401. Ferrer-Costa C, Orozco M, de la Cruz X. 2007 Characterization of compensated mutations in terms of structural and physico-chemical properties. *J. Mol. Biol.* **365**, 249–256. (doi:10.1016/j.jmb.2006.09.053)
402. Baresić A, Hopcroft LEM, Rogers HH, Hurst JM, Martin ACR. 2010 Compensated pathogenic deviations: analysis of structural effects. *J. Mol. Biol.* **396**, 19–30. (doi:10.1016/j.jmb.2009.11.002)
403. Wang Z, Moult J. 2003 Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins* **53**, 748–757. (doi:10.1002/prot.10522)
404. Ivankov DN, Finkelstein AV, Kondrashov FA. 2014 A structural perspective of compensatory evolution. *Curr. Opin. Struct. Biol.* **26**, 104–112. (doi:10.1016/j.sbi.2014.05.004)
405. Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U. 2012 Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336**, 1157–1160. (doi:10.1126/science.1217405)
406. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002 Evolutionary rate in the protein interaction network. *Science* **296**, 750–752. (doi:10.1126/science.1068696)
407. Pál C, Papp B, Hurst L. 2001 Highly expressed genes in yeast evolve slowly. *Genetics* **71**, 416–417.
408. Giaever G *et al*. 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391. (doi:10.1038/nature00935)
409. Zhang R, Lin Y. 2009 DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* **37**, D455–D458. (doi:10.1093/nar/gkn858)
410. Park C, Chen X, Yang J, Zhang J. 2013 Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **110**, E678–E686. (doi:10.1073/pnas.1218066110)
411. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005 Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14 338–14 343. (doi:10.1073/pnas.0504070102)
412. Drummond DA, Wilke CO. 2008 Mistranslation-induced protein misfolding as a dominant constraint

on coding-sequence evolution. *Cell* **134**, 341–352. (doi:10.1016/j.cell.2008.05.042)

413. Yang J-R, Zhuang S-M, Zhang J. 2010 Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* **6**, 421. (doi:10.1038/msb.2010.78)
414. Serohijos AWR, Rimas Z, Shakhnovich EI. 2012 Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* **2**, 249–256. (doi:10.1016/j.celrep.2012.06.022)
415. Drummond DA, Wilke CO. 2009 The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **10**, 715–724. (doi:10.1038/nrg2662)
416. Papp B, Notebaart RA, Pál C. 2011 Systems-biology approaches for predicting genomic evolution. *Nat. Rev. Genet.* **12**, 591–602. (doi:10.1038/nrg3033)
417. Payne JL, Wagner A. 2014 The robustness and evolvability of transcription factor binding sites. *Science* **343**, 875–877. (doi:10.1126/science.1249046)
418. McCloskey D, Palsson BØ, Feist AM. 2013 Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.* **9**, 661. (doi:10.1038/msb.2013.18)
419. Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. 2006 Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670. (doi:10.1038/nature04568)
420. Wagner A. 2005 Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays* **27**, 176–188. (doi:10.1002/bies.20170)
421. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO. 2012 Network context and selection in the evolution to enzyme specificity. *Science* **337**, 1101–1104. (doi:10.1126/science.1216861)
422. Varma A, Palsson BO. 1994 Metabolic flux balancing: basic concepts, scientific and practical use. *Nat. Biotechnol.* **12**, 994–998. (doi:10.1038/nbt1094-994)
423. Barve A, Rodrigues JFM, Wagner A. 2012 Superessential reactions in metabolic networks. *Proc. Natl Acad. Sci. USA* **109**, E1121–E1130. (doi:10.1073/pnas.1113065109)
424. Konno A, Kitagawa A, Watanabe M, Ogawa T, Shirai T. 2011 Tracing protein evolution through ancestral structures of fish galectin. *Structure* **19**, 711–721. (doi:10.1016/j.str.2011.02.014)
425. Perez-Jimenez R *et al*. 2011 Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, 592–596. (doi:10.1038/nsmb.2020)
426. Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, Verstrepen KJ. 2012 Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol.* **10**, e1001446. (doi:10.1371/journal.pbio.1001446)
427. Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, Arcus VL. 2012 On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus. *Mol. Biol. Evol.* **29**, 825–835. (doi:10.1093/molbev/msr253)
428. Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. 2013 Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902. (doi:10.1021/ja311630a)
429. Harms MJ, Eick GN, Goswami D, Colucci JK, Griffin PR, Ortlund EA, Thornton JW. 2013 Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proc. Natl Acad. Sci. USA* **110**, 11 475–11 480. (doi:10.1073/pnas.1303930110)
430. Bridgham JT, Ortlund EA, Thornton JW. 2009 An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515–519. (doi:10.1038/nature08249)
431. Friedland GD, Lakomek N-A, Griesinger C, Meiler J, Kortemme T. 2009 A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput. Biol.* **5**, e1000393. (doi:10.1371/journal.pcbi.1000393)
432. Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. 2014 Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol. Biol. Evol.* **31**, 135–139. (doi:10.1093/molbev/mst178)
433. Huang T-T, del Valle Marcos ML, Hwang J-K, Echave J. 2014 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. (doi:10.1186/1471-2148-14-78)
434. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. 2013 Protein conformational diversity correlates with evolutionary rate. *Mol. Biol. Evol.* **30**, 1500–1503. (doi:10.1093/molbev/mst065)
435. Juritz E, Palopoli N, Fornasari MS, Fernandez-Alberti S, Parisi G. 2013 Protein conformational diversity modulates sequence divergence. *Mol. Biol. Evol.* **30**, 79–87. (doi:10.1093/molbev/mss080)
436. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014 A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592. (doi:10.1093/molbev/msu081)
437. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. 2011 Assessment of template based protein structure predictions in CASP9. *Proteins* **79**, 37–58. (doi:10.1002/prot.23177)
438. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. 2011 Critical assessment of methods of protein structure prediction (CASP)–round IX. *Proteins* **79**, 1–5. (doi:10.1002/prot.23200)
439. Bloom JD, Glassman MJ. 2009 Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput. Biol.* **5**, e1000349. (doi:10.1371/journal.pcbi.1000349)
440. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. 2012 The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142. (doi:10.1038/nature11500)
441. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. 2014 Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl Acad. Sci. USA* (doi:10.1073/pnas.1413575111)

INTERFACE

rsif.royalsocietypublishing.org

Correction

Cite this article: Sikosek T, Chan HS. 2015 Correction to 'Biophysics of protein evolution and evolutionary protein biophysics'. *J. R. Soc. Interface* **12**: 20150915.
http://dx.doi.org/10.1098/rsif.2015.0915

Correction to 'Biophysics of protein evolution and evolutionary protein biophysics'

Tobias Sikosek and Hue Sun Chan

J. R. Soc. Interface **11**, 20140419 (2014; Published online 27 August 2014) (doi:10.1098/rsif.2014.0419)

The following typographical errors should be corrected:

Page 11, right column, line 39—'threonine-to-arginine mutation' should read 'threonine-to-alanine mutation'.

Page 34, ref. 407—the volume and page numbers should be '**158**, 927–931'.

THE ROYAL SOCIETY
PUBLISHING