Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

What evolution tells us about protein physics, and protein physics tells us about evolution

Ugo Bastolla[1], Yves Dehouck[2] and Julian Echave[3]

The integration of molecular evolution and protein biophysics is an emerging theme that steadily gained importance during the last 15 years, significantly advancing both fields. The central integrative concept is the stability of the native state, although non-native conformations are increasingly recognized to play a major role, concerning, for example, aggregation, folding kinetics, or functional dynamics. Besides molecular requirements on fitness, the stability of native and alternative conformations is modulated by a variety of factors, including population size, selective pressure on the replicative system, which determines mutation rates and biases, and epistatic effects. We discuss some of the recent advances, open questions, and integrating views in protein evolution, in light of the many underlying trade-offs, correlations, and dichotomies.

Addresses
[1] Centro de Biologia Molecular ''Severo Ochoa'', CSIC-UAM Cantoblanco, 28049 Madrid, Spain
[2] Machine Learning Group, Université Libre de Bruxelles (ULB), 1050 Brussels, Belgium
[3] CONICET and Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, San Martín, Buenos Aires, Argentina

Corresponding author: Bastolla, Ugo (ubastolla@cbm.csic.es)

Current Opinion in Structural Biology 2017, **42**:59–66

This review comes from a themed issue on **Proteins: bridging theory and experiment**

Edited by **Igor N Berezovsky** and **Ugo Bastolla**

http://dx.doi.org/10.1016/j.sbi.2016.10.020

Introduction

Proteins are designed by biological evolution to perform specific activities governed by the laws of physics. Their full understanding requires the integration of biophysical constraints and evolutionary theory [1–3,4•,5,6], which we review here. Biophysical constraints on protein evolution are reviewed in the first section, while evolutionary constraints on biophysical properties are the subject of the second section. In the third section, we treat evolutionary correlations and their influence on molecular evolution models.

Biophysical constraints on protein evolution

Function versus stability

A large body of evidence indicates that the stability of globular proteins is a target of natural selection [4•], because of the necessity to avoid aggregation prone conformations and because, except for natively unfolded proteins [7], stability is a prerequisite of function, which is the ultimate target of selection. The relationship between stability and function is not simple and trade-offs exist [8]. For instance, the ability to bind other proteins may interfere with stability against misfolding, and large functional movements may imply a stability cost. Empirically, residues at functional sites are rarely optimal for stability, so that their mutation is often less destabilizing [9], while mutations that create new functions tend to be more destabilizing than average [8,10].

Protein stability correlates well with fitness, as nicely demonstrated by a recent study of nearly 1000 mutations in beta-lactamase TEM-1 [11], or illustrated by the successful use of functional assays to identify stabilizing mutations [12]. Since modeling function requires specific assumptions, the simplest structure-aware genotype-to-phenotype mapping models fitness f as the fraction of protein correctly folded into the native functional state [13,14], P_N, which can be expressed as function of the free energy of the native (G_N), unfolded (G_U), and misfolded (G_M) states:

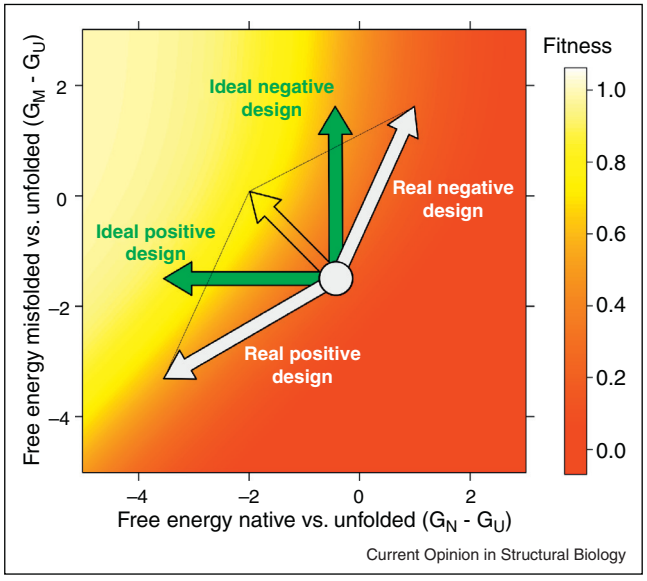
$$f = P_N = \frac{e^{-G_N/k_BT}}{e^{-G_N/k_BT} + e^{-G_U/k_BT} + e^{-G_M/k_BT}}. \quad (1)$$

Selection for stability: native and non-native aspects

The above simple model is sufficient to reveal a dichotomy between positive design, that is, the strengthening of native interactions to improve stability against unfolding, and negative design, that is, the weakening of non-native interactions to destabilize misfolded conformations [15,16]. General strategies for negative design include limiting hydrophobicity and weakening interactions between residues close in the sequence [16,17•]. However, these strategies reduce the stability of the native state with respect to the unfolded state, which may produce a trade-off between negative and positive design [16,17•], as depicted in Figure 1.

Positive design was found to dominate in small proteins, which have fewer contacts per residue and thus require stronger native interactions to compensate the loss of conformational entropy upon folding [18]. In contrast,

Figure 1



Balance between positive and negative design. Fitness (according to Equation 1) is represented as a function of the stability of the native (N) and misfolded states (M) relative to the unfolded state (U). An ideal positive design substitution stabilizes N without affecting M, while an ideal negative design substitution destabilizes M without affecting N (green arrows). However, substitutions that (de)stabilize N, for instance by increasing (decreasing) the hydrophobicity, generally have a similar effect on M, which mitigates the beneficial effect on fitness (white arrows). In consequence, the combination of both positive and negative design is often necessary to ensure sufficient fitness (transparent arrow).

negative design plays a stronger role in the evolution of large proteins [17•], in agreement with their more common tendency to fold in multiple stages, and thus to populate partially or wrongly folded states [19].

The stability of protein–protein interactions is also a target of selection, subject both to positive and negative design. Whereas the former promotes functional interactions, the latter aims at avoiding non-functional misinteractions [20–22], which may be deleterious by sequestrating the protein and inhibiting its functional activity, by interfering with other pathways, or through directly toxic effects such as the formation of aggregates [23].

Mutational robustness: protein stability

Stability improves the robustness of proteins against destabilizing mutations, enhancing their capacity to tolerate functional changes, which tend to be destabilizing [8,10]. Indeed, in a sufficiently stable protein, even a strongly destabilizing mutation may be unable to drag the fraction of correctly folded proteins out of the viable range, and would thus only have a mild effect on fitness [13,14]. Theoretical studies show that selective pressure for increasing mutational robustness is only relevant for large mutation rates [24]. In contrast, robustness is subject to selection against errors in protein translation by the error prone ribosome, in particular for highly expressed proteins whose incorrect translation can have severe consequences, which explains why they tend to evolve slowly [25]. Conversely, disordered regions [7] are less subject to selection for stability (although they must also avoid aggregation) and they generally evolve faster than ordered ones [26], consistent with the observation that residues with fewer contacts tend to evolve faster [5].

Folding kinetics and conformational dynamics

The native state must not only be stable, it must be reachable in a reasonable time. The main determinant of the folding rate is the topology of the native state, in particular its absolute contact order (ACO), that is, the mean sequence distance between residues in contact [27]. It was proposed that strengthening short-range native contacts accelerates folding [28]. Evidence for this strategy was found for proteins with large ACO that would otherwise fold very slowly [29], despite a trade-off with negative design. Indeed, strong short-range contacts tend to favor both fast folding and misfolding, in part because their formation implies a minimal cost in terms of conformational entropy. Negative design also affects folding kinetics, as non-native contacts can slow down (though sometimes accelerate) folding [30] and they may trigger pathological protein aggregation, through folding intermediates or off-pathway kinetic traps [31]. In particular, the fact that many short proteins fold with two-state kinetics might be a consequence of selective pressure to avoid intermediates [31].

Protein dynamics is a plausible target of selection, since it is essential for function, but the evidence for it is not conclusive. The intrinsic dynamics of proteins is commonly described by low-frequency normal modes (LFNM), which represent independent collective movements of large amplitude, determined by the topology of the native state [32]. LFNM correlate with large conformational changes such as those involved in allosteric regulation [32]. Functional motions correlate with LFNM more than expected based on the fact that they represent the deformations of largest amplitude of the protein, suggesting that selection plays a role [33]. The observation that LFNM are the most evolutionary conserved [34] may be considered as further evidence for selection targeting motions. However, a recent study suggests that the conservation of LFNM may arise from their robustness against mutations [35]. Native dynamics are also related to evolutionary divergence, as LFNM were observed to overlap significantly with the structural deformations within a protein superfamily [36]. Although this was considered as evidence for selection, other studies support the alternative view that structural perturbations are correlated with LFNM even for random mutations

[37], and for variability among structures of the same protein determined in different experimental conditions [38]. More research is needed to disentangle the effects of mutation and natural selection on LFNM conservation and their role in structural divergence.

Evolutionary constraints on protein biophysics

Evolutionary temperatures: population size and mutation rate

There are deep analogies between the evolution of biological populations under mutation and selection and the balance between entropy and energy in statistical physics. These analogies become precise mappings in at least two limits.

In the limit of low mutation rate, populations are genetically homogeneous and evolution can be modeled as the trajectory of the wildtype sequence in the space of amino-acid sequences as represented through the classical Moran's or Fisher–Wright's model of population genetics. At most one mutant is present at any time, and it gets either eliminated or fixed in the population, with a probability that depends on the relative fitness f_{mut}/f_{wt} and on the effective population size N. At long times, this stochastic process reaches a stationary state, in which the probability of a genotype x with fitness $f(x)$ is given by $p \propto \pi(x)e^{-\nu(-\log f(x))}$, where $\pi(x)$ is the probability of occurrence of x under mutation alone (i.e. without selective pressure), and ν is a linear function of the effective population size N depending on which evolutionary process is represented [39]. At equilibrium, the evolutionary system reaches thus a Boltzmann-like distribution with $1/\nu$ playing the role of evolutionary temperature, and $-\log f$ the role of energy [39]. Just as mean energy increases with temperature in physical systems, mean fitness increases with population size; large populations attain high fitness, while small populations are more tolerant to the exploration of low-fitness regions.

The substitution rate is predicted to be independent of population size under neutral mutations, and is expected to decrease with N when the fitness effect of mutations cannot be neglected. However, a recent study suggested that, for very low mutation rate, the substitution rate is almost independent of N even under non-neutral mutations [40]. This independence results from the compensation of two effects: for larger N, deleterious mutations are more difficult to fix, but the fitness impact of mutations is reduced due to the increased stability of the evolved proteins.

At high mutation rates, the evolutionary process cannot be described by following a single genotype. In the limit of large populations, the quasi-species theory [41] provides another analogy with statistical mechanics, in which the mutation rate plays the role of temperature. For higher mutation rate, sequences with optimal genotype represent a decreasing fraction of the population, until adaptation becomes impossible above a critical mutation rate, the error threshold, which is analogous to a phase transition [42]. In models that consider the interplay between small populations and high mutation rate, new phenomena occur due to the interactions between multiple mutations in the same clone and between multiple mutated clones in the population. For instance, slightly deleterious mutations that co-occur with advantageous ones are frequently fixed. This may explain why higher substitution rate is observed between more recently diverged species [43,44].

The role of population size is crucial in *bottleneck* events, during which the population size is drastically reduced, such as for instance in the case of obligatory endosymbionts, when a small number of bacteria are maternally transmitted [45], or under strong positive selection in tumor progression [46••]. The expected reduction of protein stability under bottlenecks was supported by computational studies of orthologous bacterial proteins [47], and it contributes to explain the accumulation of deleterious passenger mutations in cancer cells [46••].

A possible example of the relevance of population size in structural biology concerns the formation of oligomeric complexes of proteins. Oligomerization is frequent in the proteome of higher eukaryotes, which are thought to have much smaller population sizes than bacteria [48]. It was proposed that quaternary interactions in multicellular eukaryotes may have primarily arisen as a response to a stability deficit of the monomers, rather than as an adaptive trait [49•]. Indeed, the lower pressure on stability in small populations tends to generate poorly optimized protein-water interfaces, which can facilitate protein-protein binding. Low stability may also participate to the emergence of more flexible or even disordered regions, which are often found in proteins involved in the formation of large complexes such as the Centrosome [50•,51]. This macromolecular complexity would later be subject to adaptation through the evolution of cooperativity and allosteric regulation, which are likely facilitated by conformational diversity [52,53]. However, it cannot be excluded that positive selection for forming the complex is sometimes a driver of the trend towards increased flexibility.

Mutation rates and mutation bias

Although mutation and selection are sometimes conceived as independent forces, empirical data show that the properties of the mutation process can readily evolve through mutations in replicative proteins, and are thus also under selective pressure. For instance, in a landmark long-term experiment following the evolution of *E. coli* in a constant environment, a hypermutator phenotype with increased mutation rate arose while the population was

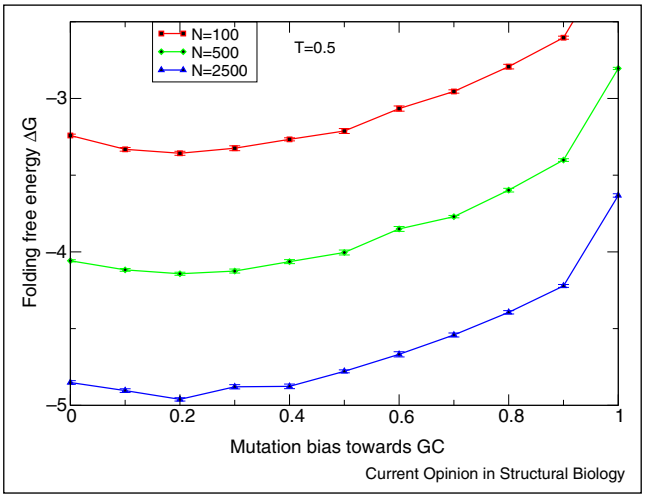
adapting [54]. Conversely, a population with hypermutator phenotype was invaded by another with reduced mutation rate, when the potential for further adaptation declined [55]. Consistently, it has been observed that bacteria tend to enhance the mutation rate in response to stress, a response with possible adaptive significance [56].

The mutation bias, that is, the preferential generation of certain types of nucleotide mutations through the replicative machinery, has been observed to vary in experimental evolution of the foot-and-mouth disease virus in presence of the mutagenic agent ribavirin [57]. The evolution of endosymbiotic and parasitic bacteria from a free-living lifestyle to an obligate intracellular lifestyle with small effective population was also accompanied most of the time by a change in mutation bias favoring AT over GC, mediated through the extensive loss of error-correcting genes [58]. Like the mutation rate, the mutation bias is modulated by the replication machinery of the cell, and is not selectively neutral. Indeed, because of the structure of the genetic code, genomes rich in AT (poor in CG) result in more hydrophobic proteins. Even under strong selective pressure, a mutational bias towards AT tends to produce proteins that are more hydrophobic and more prone to misfolding, suggesting that the mutation bias is an important determinant of the balance between positive and negative design [47,59]. Interestingly, simulations of protein evolution suggest that, for any given population size, there is an optimal mutation bias that maximizes long-term fitness and protein stability [60]. The dependence of simulated protein stability on the mutation bias is shown in Figure 2. The optimal mutation bias is towards AT, consistent with the bias observed in intracellular bacteria. On the other hand, disordered proteins are enriched in polar and charged amino acids, preferentially coded by the nucleotides C and G, which suggests that mutational biases may participate to their prevalence in the proteome of complex organisms.

Mutational robustness: chaperones

Chaperones are proteins that enhance robustness by facilitating the folding of other proteins [61]. They have been suggested to act as capacitors of phenotypic evolution, buffering the phenotypic effects of mutations, and allowing the accumulation of latent variations that give rise to morphological changes when some environmental stress reduces their buffering power [62]. The robustness provided by the overexpression of chaperones has been exploited to accelerate the experimental evolution of new enzymatic functions, demonstrating another important link between robustness and evolvability [63]. Accordingly, a positive correlation was demonstrated between chaperone binding and evolutionary rate, in various groups of proteins [64,65•,66]. It is noteworthy that this robustness is under genetic control, as chaperone expression can be increased to face conditions that are unfavorable for protein stability, such as reduced effective population size through endosymbiotic lifestyle [67] or through experimental population bottlenecks [68]. However, the expression of chaperones is metabolically costly, and it rapidly reverts in experimental evolution [69].

Figure 2



Effect of mutation bias on protein stability. Simulations of the protein evolution model Equation 1 in the low mutation rate limits have been run with different mutation bias and different population size. One can see that there is an optimal mutation bias at which the folding free energy ΔG is minimized.

Evolutionary correlations

Stability depends on residue-residue interactions and fitness depends non-linearly on stability. As a result evolutionary *epistatic* interactions may arise.

Magnitude of epistasis

Epistasis occurs when the joint effect of multiple mutations leads to either higher (positive epistasis) or lower (negative epistasis) fitness than expected from their individual effects. Recent high-throughput experiments have confirmed previous observations that, for random mutations, epistatic interactions are relatively rare but can happen even between positions that are distant in the native state [12,70,71]. Negative epistasis often results from the combination of two destabilizing mutations. Even if their effect on stability is additive, epistasis can originate in the non-linear dependence of fitness on stability, Equation 1: a destabilizing mutation occurring in a stable wildtype may still maintain the fitness in a viable range, but its combination with a similar mutation might not [8,10,14]. Cases of positive epistasis appear to be rare (in good part because most mutations are deleterious), but of potential evolutionary importance: most deleterious mutations were found to be beneficial in at least one different mutational background [71]. In particular, compensatory mutations (destabilizing followed by

stabilizing with positive epistasis) can be confounded with positive selection, as shown in the recent simulation by Dasmeh *et al.* [72].

Even if epistasis is rare for random mutations, the opposite holds for mutations fixed under purifying selection [73]. Most fixed mutations have a nearly neutral effect on fitness at their time of fixation, however many would have been deleterious in other sequence backgrounds: they are *contingent* on the previous fixation of permissive mutations [73]. Afterwards, further substitutions adapt the protein to the fixed mutations, making their reversion increasingly deleterious: they become *entrenched* by epistasis [73,74••].

Time-variation of amino acid preferences versus approximate sites-independence

The prevalence and magnitude of epistasis are critical for the development of the probabilistic models of molecular evolution used for phylogenetic inference. The very existence of epistasis contradicts the assumption made by most molecular evolution models that protein sites evolve independently. The incorporation of pairwise interactions among sites has been proposed in some models [75,76], but at the expense of preventing the analytic computation of the likelihood function. In an effort to improve phylogenetic inferences under the independent-sites assumption, several approaches have been proposed to effectively account for selection on stability without having to deal with correlations. These approaches adopt site-specific substitution processes, but differ in the way they parameterize them: (1) from the structural characteristic of each site in the native state [77]; (2) from sequence data [78,79], (3) from structurally constrained evolutionary simulations [80], (4) from mean-field approximations of evolutionary models [81], or (5) from high-throughput mutagenesis experiments [82]. These models do not account for the possibility that epistasis may induce a strong time dependence of the amino acid preferences at any given site [74••,83]. Indeed, the entranchment of a deleterious mutation over time would make the new amino acid, and similar ones, more acceptable at that position. However, the importance of this effect remains a subject of debate, and several works suggest that propensities diverge slowly and are mainly conserved among homologous proteins [84••,85••,86], consistent with the idea that most deleterious fixations reverse before entrenchment takes place [84••]. A recent step forward in the resolution of this dichotomy was provided by the demonstration that reversion rates decrease strongly with time for sites involved in epistatic interactions. Reversions are thus likely after a deleterious fixation, but if they do not happen immediately, they may become entrenched [87••].

Correlated substitutions

Epistatic interactions are responsible of the correlations observed between columns of multiple sequence alignments (MSA). Since most epistatic interactions are expected to take place between sites that are in contact in the native state, the detection of evolutionary correlations was proposed long ago as a method for predicting native contacts and even protein-protein interactions, but it was not until recently that technical advances, reviewed in [88•], improved the detection of these correlations, raising a strong interest. When thousands of homologous sequences are available, these methods allow accurate predictions of protein structures without templates [89]. An interesting development consisted in extending these methods to predict not just one native structure, but a conformational ensemble representative of the functional dynamics of the protein [90–92]. However, correlations can also be found between pairs of residues distant in the native state, consistent with the observation of long-range epistatic interactions [71], and with the interpretation that correlated mutations identify functional, but not necessarily structural, modules of the protein [93]. Another possible interpretation of correlations between positions far apart in the native state is that they can arise from non-native conformational ensembles relevant for the stability or the folding kinetics of the protein. Consistently, evolutionary correlations between structurally remote pairs have been shown to improve folding rate predictions [94•].

The notion that amino acid correlations across sites are attributable to substitutions that are correlated in time has been challenged in a recent study that found that reconstructed substitutions at correlated sites tend to occur in different branches of the phylogenetic tree [95]. This result is consistent with the stability model Equation 1, which implies that amino acids that interact advantageously tend to co-occur in neighboring positions even if they are not the result of substitutions that are very close in time.

Conclusions and outlook

The integration of the biophysical (folding stability and kinetics, conformational ensembles, functional motions, interactions) and the evolutionary (population genetics, molecular evolution) characterization of proteins is producing a dramatic change in the way these once separated disciplines are conceived. At the crossroads, methodological developments for ancestral sequence reconstruction and biophysical characterization are providing new avenues for the experimental study of evolution [96], while the dialog between new computational models and experiments is improving our understanding of proteins in structure and sequence space, producing exciting advances.

Acknowledgements

UB gratefully acknowledge previous collaborators on this subject, in particular Markus Porto, Alberto Pascual-Garcí a, Raul Méndez and Miguel Arenas. UB was supported by the Spanish Ministry of Economy of Spain, grant no. BFU2012-40020. JE was supported by UNSAM and CONICET (Argentina). Research at the CBMSO is facilitated by the Fundación Ramón Areces.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Bastolla U, Porto M, Roman HE, Vendruscolo M (Eds): *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Berlin: Springer-Verlag; 2007.

2. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning AP, Dokholyan NV, Echave J *et al.*: **The interface of protein structure, protein biophysics, and molecular evolution**. *Protein Sci* 2012, **21**:769-785.

3. Wilke CO: **Bringing molecules back into molecular evolution**. *PLoS Comput Biol* 2012, **8**:e1002572.

4. • Sikosek T, Chan HS: **Biophysics of protein evolution and evolutionary protein biophysics**. *J R Soc Interface* 2014, **11**:20140419.

A very detailed review covering many synergies between protein biophysics and protein evolution studies.

5. Echave J, Spielman SJ, Wilke CO: **Causes of evolutionary rate variation among protein sites**. *Nat Rev Gen* 2016, **17**:109-121.

6. Chi PB, Liberles DA: **Selection on protein structure, interaction, and sequence**. *Protein Sci* 2012, **25**:1168-1178.

7. Dunker AK, Silman I, Uversky VN, Sussman JL: **Function and structure of inherently disordered proteins**. *Curr Opin Struct Biol* 2008, **18**:756-764.

8. Tokuriki N, Stricher F, Serrano L, Tawfik DS: **How protein stability and new functions trade off**. *PLoS Comput Biol* 2008, **4**:e1000002.

9. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M: **PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality**. *BMC Bioinformatics* 2011, **12**:151.

10. Bloom JD, Arnold FH: **In the light of directed evolution: pathways of adaptive protein evolution**. *Proc Natl Acad Sci U S A* 2009, **106S1**:9995-10000.

11. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G *et al.*: **Capturing the mutational landscape of the beta-lactamase TEM-1**. *Proc Natl Acad Sci U S A* 2013, **110**:13067-13072.

12. Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S: **A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function**. *Proc Natl Acad Sci U S A* 2012, **109**:16858-16863.

13. Goldstein RA: **The evolution and evolutionary consequences of marginal thermostability in proteins**. *Proteins* 2011, **79**:1396-1407.

14. Serohijos AW, Shakhnovich EI: **Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics**. *Curr Opin Struct Biol* 2014, **26**:84-91.

15. Berezovsky IN, Zeldovich KB, Shakhnovich EI: **Positive and negative design in stability and thermal adaptation of natural proteins**. *PLoS Comput Biol* 2007, **3**:e52.

16. Noivirt-Brik O, Horovitz A, Unger R: **Trade-off between positive and negative design of protein stability: from lattice models to real proteins**. *PLoS Comput Biol* 2009, **5**:e1000592.

17. • Minning J, Porto M, Bastolla U: **Detecting selection for negative design in proteins through an improved model of the misfolded state**. *Proteins* 2013, **81**:1102-1112.

Through a statistical mechanical model of the misfolded state, the authors detect evidence of negative (and positive) design in natural protein sequences with respect to reshuffled sequences: sequences are neither too hydrophobic nor too hydrophilic, short-range contacts cannot be too strong even if they are present in the native state, and pairs of contacts that tend to co-occur in protein structures are penalized.

18. Bastolla U, Demetrius L: **Stability constraints and protein evolution: the role of chain length, composition, and disulphide bonds**. *Prot Eng Des Sel* 2005, **18**:405-415.

19. Privalov PL: **Intermediate states in protein folding**. *J Mol Biol* 1996, **258**:707-725.

20. Liberles DA, Tisdell MD, Grahnen JA: **Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy**. *Proc Biol Sci* 2011, **278**:1930-1935.

21. Yang JR, Liao BY, Zhuang SM, Zhang J: **Protein misinteraction avoidance causes highly expressed proteins to evolve slowly**. *Proc Natl Acad Sci U S A* 2012, **109**:E831-840.

22. Levy ED, De S, Teichmann SA: **Cellular crowding imposes global constraints on the chemistry and evolution of proteomes**. *Proc Natl Acad Sci U S A* 2012, **109**:20461-20466.

23. De Groot NS, Torrent M, Villar-Piqué A, Lang B, Ventura S, Gsponer J, Babu MM: **Evolutionary selection for protein aggregation**. *Biochem Soc Trans* 2012, **40**:1032-1037.

24. Van Nimwegen E, Crutchfield JP, Huynen M: **Neutral evolution of mutational robustness**. *Proc Natl Acad Sci U S A* 1999, **96**:9716-9720.

25. Drummond DA, Wilke CO: **Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution**. *Cell* 2008, **134**:341-352.

26. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW: **Evolution and disorder**. *Curr Opin Struct Biol* 2011, **21**:441-446.

27. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV: **Contact order revisited: influence of protein size on the folding rate**. *Protein Sci* 2003, **12**:2057-2062.

28. Oztop B, Ejtehadi MR, Plotkin SS: **Protein folding rates correlate with heterogeneity of folding mechanism**. *Phys Rev Lett* 2004, **93**:208105.

29. Bastolla U, Bruscolini P, Velasco JL: **Sequence determinants of protein folding rates: positive correlation between contact energy and contact range indicates selection for fast folding**. *Proteins* 2012, **80**:2287-2304.

30. Zarrine-Afsar A, Wallin S, Neculai AM, Neudecker P, Howell PL, Davidson AR, Chan HS: **Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding**. *Proc Natl Acad Sci U S A* 2008, **105**:9999-10004.

31. Isogai Y: **Native protein sequences are designed to destabilize folding intermediates**. *Biochemistry* 2006, **45**:2488-2492.

32. Bahar I, Lezon TR, Yang LW, Eyal E: **Global dynamics of proteins: bridging between structure and function**. *Annu Rev Biophys* 2010, **39**:23-42.

33. Dos Santos HG, Klett J, Méndez R, Bastolla U: **Characterizing conformation changes in proteins through the torsional elastic response**. *Biochim Biophys Acta* 2013, **1834**:836-846.

34. Maguid S, Fernandez-Alberti S, Echave J: **Evolutionary conservation of protein vibrational dynamics**. *Gene* 2008, **422**:7-13.

35. Echave J: **Why are the low-energy protein normal modes evolutionarily conserved**. *Pure Appl Chem* 2012, **84**:1931-1937.

36. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR: **An analysis of core deformations in protein superfamilies**. *Biophys J* 2005, **88**:1291-1299.

37. Echave J: **Evolutionary divergence of protein structure: the linearly forced elastic network model**. *Chem Phys Lett* 2008, **457**:413-416.

38. Echave J, Fernandez FM: **A perturbative view of protein structural variation**. *Proteins* 2010, **78**:173-180.

39. Sella G, Hirsh AE: **The application of statistical physics to evolutionary biology**. *Proc Natl Acad Sci U S A* 2005, **102**:9541-9546.

40. Goldstein RA: **Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability**. *Gen Biol Evol* 2013, **5**:1584-1593.

41. Eigen M, McCaskill J, Schuster P: **The molecular quasi-species**. *Adv Chem Phys* 1989, **75**:149-263.

42. Tarazona P: **Error thresholds for molecular quasispecies as phase transitions: from simple landscapes to spin-glass models**. *Phys Rev A* 1992, **45**:6038.

43. Ho SY, Phillips MJ, Cooper A, Drummond AJ: **Time dependency of molecular rate estimates and systematic overestimation of recent divergence times**. *Mol Biol Evol* 2005, **22**:1561-1568.

44. Peterson GI, Masel J: **Quantitative prediction of molecular clock and ka/ks at short timescales**. *Mol Biol Evol* 2009, **26**:2595-2603.

45. Moran NA, Baumann P: **Bacterial endosymbionts in animals**. *Curr Opin Microbiol* 2000, **3**:270-275.

46. •• McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA: **Impact of deleterious passenger mutations on cancer progression**. *Proc Natl Acad Sci U S A* 2013, **110**:2910-2915.
It is shown here that passenger mutations accumulate in positively selected cancer clones due to high mutation rate and low effective population size. It is demonstrated that they are often deleterious for protein stability and represent a significant mutational burden, and it is proposed that they can play the role of an Achilles-heel of the disease.

47. Bastolla U, Moya A, Viguera E, van Ham RC: **Genomic determinants of protein folding thermodynamics in prokaryotic organisms**. *J Mol Biol* 2004, **343**:1451-1466.

48. Lynch M: *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates; 2007.

49. • Fernandez A, Lynch M: **Non-adaptive origins of interactome complexity**. *Nature* 2011, **474**:502-505.
This paper proposes that, in the proteomes of multi-cellular eukaryotes, quaternary interactions leading to oligomerization were selected to rescue surface defects resulting from the lower pressure on stability due to the small effective population size. According to this view, the advantageous consequences of interaction complexity (cooperativity, allosteric regulation) would be not an adaptation but a side-effect (exaptation, in the evolutionary jargon).

50. • Marsh JA, Teichmann SA: **Protein flexibility facilitates quaternary structure assembly and evolution**. *PLoS Biol* 2014, **12**:e1001870.
In this paper, the authors report a statistical analysis of protein complexes, demonstrating a positive correlation between the flexibility of a protein and the number of protein–protein interactions in which it is involved. Furthermore, subunits that were included later in the evolution of a protein complex are observed to be more flexible, suggesting a clear evolutionary link.

51. Nido GS, Méndez R, Pascual-Garcí a A, Abia D, Bastolla U: **Protein disorder in the centrosome correlates with complexity in cell types number**. *Mol Biosyst* 2012, **8**:353-367.

52. Keskin O, Ma B, Rogale K, Gunasekaran K, Nussinov R: **Protein–protein interactions: organization, cooperativity and mapping in a bottom-up systems biology approach**. *Phys Biol* 2005, **2**:S24.

53. James LC, Tawfik DS: **Conformational diversity and protein evolution — a 60-year-old hypothesis revisited**. *Trends Biochem Sci* 2003, **28**:361-368.

54. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF: **Genome evolution and adaptation in a long-term experiment with *Escherichia coli***. *Nature* 2009, **461**:1243-1247.

55. Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D: **Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load**. *Proc Natl Acad Sci U S A* 2013, **110**:222-227.

56. Tenaillon O, Denamur E, Matic I: **Evolutionary significance of stress-induced mutagenesis in bacteria**. *Trends Microbiol* 2004, **12**:264-270.

57. Agudo R, Ferrer-Orta C, Arias A, de la Higuera I, Perales C, Pérez-Luque R, Verdaguer N, Domingo E: **A multi-step process of viral adaptation to a mutagenic nucleoside analogue by modulation of transition types leads to extinction-escape**. *PLoS Pathog* 2010, **6**:e1001072.

58. Moran NA: **Microbial minimalism: genome reduction in bacterial pathogens**. *Cell* 2002, **108**:583-586.

59. Bastolla U, Porto M, Roman HE, Vendruscolo M: **A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank**. *BMC Evol Biol* 2006, **6**:43.

60. Mendez R, Fritsche M, Porto M, Bastolla U: **Mutation bias favors protein folding stability in the evolution of small populations**. *PLoS Comput Biol* 2010, **6**:e1000767.

61. Hartl FU, Bracher A, Hayer-Hartl M: **Molecular chaperones in protein folding and proteostasis**. *Nature* 2011, **475**:324-332.

62. Queitsch C, Sangster TA, Lindquist S: **Hsp90 as a capacitor of phenotypic variation**. *Nature* 2002, **417**:618-624.

63. Tokuriki N, Tawfik DS: **Chaperonin overexpression promotes genetic variation and enzyme evolution**. *Nature* 2009, **459**:668-673.

64. Bogumil D, Dagan T: **Chaperonin-dependent accelerated substitution rates in prokaryotes**. *Genome Biol Evol* 2010, **2**:602-608.

65. • Williams TA, Fares MA: **The effect of chaperonin buffering on protein evolution**. *Genome Biol Evol* 2010, **2**:609-619.
A careful review by one of the most active groups in the study of the evolutionary consequences of chaperones expression, in particular in endosymbiotic bacteria with reduced population size, with possible consequences for the evolution of cancer.

66. Lachowiec J, Lemus T, Borenstein E, Queitsch C: **Hsp90 promotes kinase evolution**. *Mol Biol Evol* 2015, **32**:91-99.

67. Kupper M, Gupta SK, Feldhaar H, Gross R: **Versatile roles of the chaperonin GroEL in microorganism–insect interactions**. *FEMS Microbiol Lett* 2014, **353**:1-10.

68. Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E: **Endosymbiotic bacteria: groEL buffers against deleterious mutations**. *Nature* 2002, **417** 398–398.

69. Sabater-Mu noz B, Prats-Escriche M, Montagud-Martí nez R, López-Cerdán A, Toft C, Aguilar-Rodrí guez J, Wagner A, Fares MA: **Fitness trade-offs determine the role of the molecular chaperonin GroEL in buffering mutations**. *Mol Biol Evol* 2015, **32**:2681-2693.

70. Melamed D, Young DL, Gamble CE, Miller CR, Fields S: **Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein**. *RNA* 2013, **19**:1537-1551.

71. Olson CA, Wu NC, Sun R: **A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain**. *Curr Biol* 2014, **24**:2643-2651.

72. Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI: **The influence of selection for protein stability on dN/dS estimations**. *Gen Biol Evol* 2014, **6**:2956-2967.

73. Shah P, McCandlish DM, Plotkin JB: **Contingency and entrenchment in protein evolution under purifying selection**. *Proc Natl Acad Sci U S A* 2015, **112**:E3226-E3235.

74. •• Pollock DD, Thiltgen G, Goldstein RA: **Amino acid coevolution induces an evolutionary Stokes shift**. *Proc Natl Acad Sci U S A* 2012, **109**:E1352-1359.
Through simulations of protein evolution with structural constraints, this paper demonstrates that epistatic interactions, although rare in random mutations, are frequent in selected mutations. Initially slightly destabilizing substitutions progressively become better adapted to their new background and more unlikely to revert (entrenchment), a phenomenon that is assimilated to the Stokes effect in atomic radiation.

75. Kleinman CL, Rodrigue N, Lartillot N, Philippe H: **Statistical potentials for improved structurally constrained evolutionary models**. *Mol Biol Evol* 2010, **27**:1546-1560.

76. Bordner AJ, Mittelmann HD: **A new formulation of protein evolutionary models that account for structural constraints**. *Mol Biol Evol* 2014, **31**:736-749.

77. Koshi JM, Goldstein RA: **Models of natural mutations including site heterogeneity**. *Proteins* 1998, **32**:289-295.

78. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies**. *Mol Biol Evol* 1998, **15**:910-917.

79. Le SQ, Dang CC, Gascuel O: **Modeling protein evolution with several amino acid replacement matrices depending on site rates**. *Mol Biol Evol* 2012, **29**:2921-2936.

80. Parisi G, Echave J: **Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes**. *Gene* 2005, **345**:45-53.

81. Arenas M, Sanchez-Cobos A, Bastolla U: **Maximum likelihood phylogenetic inference with selection on protein folding stability**. *Mol Biol Evol* 2015, **32**:2195-2207.

82. Bloom JD: **An experimentally determined evolutionary model dramatically improves phylogenetic fit**. *Mol Biol Evol* 2014, **31**:1956-1978.

83. Goldstein RA, Pollock DD: **The tangled bank of amino acids**. *Protein Sci* 2016, **25**:1354-1362.

84. •• Ashenberg O, Gong LI, Bloom JD: **Mutational effects on stability are largely conserved during protein evolution**. *Proc Natl Acad Sci U S A* 2013, **110**:21071-21076.
This paper reports simulations similar to those of Pollock *et al.* (2012) and find that most deleterious mutations are reversed immediately after fixation and that the entrenchment observed by Pollock *et al.* is rare.

85. •• Risso VA, Manssour-Triedo F, Delgado-Delgado A, Arco R, Barroso-delJesus A, Ingles-Prieto A, Godoy-Ruiz R, Gavira JA, Gaucher EA, Ibarra-Molero B, Sanchez-Ruiz JM: **Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history**. *Mol Biol Evol* 2015, **32**:440-455.
Through the reconstruction of ancestral thioredoxins with molecular evolution models and their experimental study, Risso *et al.* found results similar to those by the group of Bloom: the effect of mutations and the site-specific amino acid preferences varied little throughout evolution.

86. Doud MB, Ashenberg O, Bloom JD: **Site-specific amino acid preferences are mostly conserved in two closely related protein homologs**. *Mol Biol Evol* 2015, **32**:2944-2960.

87. •• McCandlish DM, Shah P, Plotkin JB: **Epistasis and the dynamics of reversion in molecular evolution**. *Genetics* 2016, **203**:1335-1351.
This theoretical paper rationalizes some conflicting results on epistatic interactions, showing that the two most likely outcomes of destabilizing substitutions in proteins is either an almost immediate reversion, or a progressive adaptation of the sequence background, resulting in a monotonic decrease of the probability of reversion.

88. • de Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution**. *Nat Rev Genet* 2013, **14**:249-261.
An extensive review of methods based on correlated substitutions for predicting physical interactions in proteins. This group originally proposed both methods for detecting interactions between protein residues through the correlations between the corresponding columns of multiple sequence alignments and methods to detect protein-protein interactions through the correlations between their temporal pattern of substitutions.

89. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C: **Protein 3D structure computed from evolutionary sequence variation**. *PLoS ONE* 2011, **6**:e28766.

90. Morcos F, Jana B, Hwa T, Onuchic JN: **Coevolutionary signals across protein lineages help capture multiple protein conformations**. *Proc Natl Acad Sci U S A* 2013, **110**:20533-20538.

91. Sutto L, Marsili S, Valencia A, Gervasio FL: **From residue coevolution to protein conformational ensembles and functional dynamics**. *Proc Natl Acad Sci U S A* 2015, **112**:13567-13572.

92. Sfriso P, Duran-Frigola M, Mosca R, Emperador A, Aloy P, Orozco M: **Residues coevolution guides the systematic identification of alternative functional conformations in proteins**. *Structure* 2016, **24**:116-126.

93. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families**. *Science* 1999, **286**:295-299.

94. • Mallik S, Kundu S: **Co-evolutionary constraints of globular proteins correlate with their folding rates**. *FEBS Lett* 2015, **589**:2179-2185.
The authors found that, besides the absolute contact order (ACO), the average sequence separation of evolutionarily correlated residues, can improve the prediction of protein folding rates. Notably, the importance of evolutionary correlations between residues that are distant in the native state suggest that non-native contacts play a significant role in folding.

95. Talavera Lovell, Whelan: **Covariation is a poor measure of molecular coevolution**. *Mol Biol Evol* 2015, **32**:2456-2468.

96. Harms MJ, Thornton JW: **Evolutionary biochemistry: revealing the historical and physical causes of protein properties**. *Nat Rev Genet* 2013, **14**:559-571.