REVIEW

THE PROTEIN SOCIETY | WILEY

Simple mechanisms for the evolution of protein complexity

Arvind S. Pillai[1,2] | Georg K.A. Hochberg[3,4] | Joseph W. Thornton[1,5]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

[2]Institute for Protein Design, University of Washington, Seattle, WA, USA

[3]Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[4]Department of Chemistry, Center for Synthetic Microbiology, Philipps University Marburg, Marburg, Germany

[5]Departments of Human Genetics and Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

Correspondence
Joseph W. Thornton, Departments of Human Genetics and Ecology and Evolution, University of Chicago, Chicago, IL, USA.
Email: joet1@uchicago.edu

Funding information
National Institute of General Medical Sciences, Grant/Award Numbers: R01GM131128, R01GM139007, R35-GM145336

Review Editor: John Kuriyan

Abstract

Proteins are tiny models of biological complexity: specific interactions among their many amino acids cause proteins to fold into elaborate structures, assemble with other proteins into higher-order complexes, and change their functions and structures upon binding other molecules. These complex features are classically thought to evolve via long and gradual trajectories driven by persistent natural selection. But a growing body of evidence from biochemistry, protein engineering, and molecular evolution shows that naturally occurring proteins often exist at or near the genetic edge of multimerization, allostery, and even new folds, so just one or a few mutations can trigger acquisition of these properties. These sudden transitions can occur because many of the physical properties that underlie these features are present in simpler proteins as fortuitous by-products of their architecture. Moreover, complex features of proteins can be encoded by huge arrays of sequences, so they are accessible from many different starting points via many possible paths. Because the bridges to these features are both short and numerous, random chance can join selection as a key factor in explaining the evolution of molecular complexity.

1 | GRADUALISM AND PROTEIN COMPLEXITY

To understand how living things acquired their complex features—structures and functions that arise from specific interactions among differentiated parts—has been a central aim of biology for centuries.[1] Darwin supplanted divine agency with the evolutionary view: complexity arises through "numerous successive, slight modifications" under the influence of natural selection, because each step enhances functions that contribute to fitness.[2–4] This scenario of gradual elaboration and optimization is well-supported in numerous cases.[2,4–7] The most famous

is the modern vertebrate eye, which evolved from a simple light-sensitive precursor by sequentially adding cell types and more complicated relationships among tissues, each of which improved visual sensitivity or acuity.[8,9]

In the last half-century or so, a pageant of intricate forms has been revealed at a tiny new scale. Every protein is itself a complex system, because its physical and functional features depend on a large number of interactions among its many constituent amino acids. For example, a protein's ability to fold into its native tertiary structure depends on complementary steric, electrostatic, and hydrophobic interactions among scores or hundreds of residues.[10] The same is true of quaternary structure:

most proteins assemble with other molecules into specific multimeric complexes, and the interfaces that hold these complexes together often involve dozens of tightly packed residues with a high degree of electrostatic and steric complementarity.[11] Another form of complexity is allostery—changes in a protein's function caused by binding an effector molecule—which typically involves many amino acids to bind the effector and coupleing of binding to the active site.[12,13]

During the last ~3.8 billion years, evolution has generated proteins with thousands of different folds,[10] unique multimeric interactions,[11,14,15] and varying modes of allosteric regulation.[11,16] This diversity presents a molecular version of the classic question about the evolution of biological complexity: how did the stepwise processes of evolution repeatedly produce complicated systems from simpler precursors? Darwin's model of gradual adaptive elaboration was developed to explain morphological and physiological complexity, but it has been assumed to apply as well to the evolution of complex molecular features.[2,3,17,18] Intuitively, the view that protein complexity always evolves by long, consistently adaptive trajectories may seem sensible or even necessary, given certain assumptions. For a feature to evolve, the sequence states that encode it must arise by mutation and then be fixed in populations. Multimerization, allostery, and protein folds all involve elaborate arrays of interacting amino acids, so how else could they have been acquired if not by a long and specific series of many sequence changes? And, in turn, how could we explain the fixation of a long series of particular mutations if each step were not driven by the deterministic power of selection? It has been suggested that such features might arise neutrally,[19–21] but fixation by chance alone would be vanishingly improbable if many particular mutations are required.

Recent advances in protein biochemistry and molecular evolution call into question the assumptions that underlie the argument for the gradual adaptive evolution of protein complexity. Of particular note are dramatic improvements in protein design,[22–24] deep mutational scanning[25–27] (which characterizes the functions of huge numbers of protein sequence variants), and ancestral protein reconstruction[28,29] (which uses phylogenetics to infer the sequences of ancient proteins and experiments to determine the molecular functions and structures that existed in the deep past). This new body of work shows that just one or a few mutations can drive the acquisition of multimerization, allostery, and even new folds from natural precursors that lack these features; furthermore. It also explains why these short paths exist: simpler proteins often already possess most of the physical properties that underly these features. Moreover, the networks of

sequences that yield multimerization, allostery, or a given protein fold appear to be immense, and they are closely intercalated at numerous places with the sequence networks of functional proteins that lack the feature. As a result, proteins can—and do—acquire new complex features by neutral processes. Contrary to the metaphor underlying the gradualist view, the complex features of proteins are not singular, massive mountain peaks that an evolving protein can climb only via a long trek under the deterministic engine of natural selection. Rather, many complex features are better conceived of as innumerable wrinkles, each small enough to be mounted in a single step (or just a few), which proteins repeatedly encounter as they wander through a vast multidimensional landscape of functional amino acid sequences.

2 | SHORT PATHS TO MULTIMERIZATION

A substantial body of work shows that introducing one or a few mutations into naturally occurring proteins can confer on them the capacity to form new higher-level complexes. A selection of relevant studies are listed in Table 1, and we discuss a few highlights below.

2.1 | Engineering new multimeric interactions

For years, protein engineers have been reworking the surfaces of natural proteins to yield new molecular complexes and interactions via just a few mutations.[23,31–33,35] A 2008 study, for example, conferred new homomeric interactions on five different proteins by introducing between one and four mutations, which replaced polar residues on the surface with hydrophobic amino acids and improved the steric fit between the target surfaces. These mutations conferred dimerization on a monomer, tetramerization on a dimer, and octamerization on a tetramer[23] (Figure 1a).

A recent study found that large multimeric assemblies are almost shockingly easy to evolve by mutation and provided a biophysical explanation for this behavior.[24] The authors were motivated by the observation that in sickle cell disease, a single glutamate-to-valine mutation on the surface of human hemoglobin-β (Hb-β) increases the affinity of hemoglobin molecules for each other; hemoglobin is itself a tetramer that contains two Hb-β and two Hb-α proteins, so this mutation appears twice in each hemoglobin complex and triggers assembly into massive disease-causing fibers.[40] Inspired by this example, this study took as starting points 12 unrelated

TABLE 1   Protein engineering and evolutionary studies in which occupancy of novel oligomeric states is conferred by one or a few mutations

Protein system	Initial state	Derived state	Mutations	Isologous?	Notes on mechanism	Type of experiment	Reference
6-phospho-Beta-galactosidase			4	Yes	Polar to hydrophobic mutations at interface	Rational mutagenesis	Grueninger et al.[23]
Urocanase			3	Yes	Polar to hydrophobic mutations pack against existing hydrophobic residues. Each mutation repeated four times in the complex due to isology.	Rational mutagenesis	Grueninger et al.[23]
L-rhamnulose-1 phosphate aldolase			1	Yes	Polar to phenylalanine mutation fits into an existing hydrophobic pocket. Mutation epeated eight times in the octamer;	Rational mutagenesis	Grueninger et al.[23]
Ancestral Hemoglobin precursor (Anc_α/β)			2	Yes	Derived tryptophan fits into a hydrophobic pocket on ancestral surface.	Ancestral sequence reconstruction	Pillai et al.[30]
B. subtilis TRAP complex			1 (deletion of 5 residues)	No	Change in angle between neighboring subunits makes space for an addititional subunit in a ring complex.	Rational mutagenesis	Chen et al.[31]
Streptococcal Gb1			1	Yes	Phenylalanine mutation packs against unchanged residues	Mutational scan	Jee et al.[32]
Streptococcal Gb1			5	Yes	Mutationsin the protein core and surface result in tetramer with interdigitated subunits.	Mutational scan	Jee et al.[32]
IgG-binding domain of protein L			1-3	Yes	Mutations stabilize a domain-swapped dimer. A single mutation produces a micromolar strength dimer. An additional two mutations yielded picomolar affinity.	Rational mutagenesis	Kuhlman et al.[155]
PLCδ1-PH + EPOR			2	No	Derived Phe in one subunit foms hydrophobic cluster with two Phes on the other;	Rational mutagenesis	Liu et al.[33]
Dlg-PID + Pins			1	No	Mutation stabilizes open conformation, facilitating interaction with protein partner	Ancestral sequence reconstruction	Anderson et al.[34]
Isoaspartyl dipeptidase			1	Yes	Derived tyrosine mediates eight isologous contacts between octamers in a fibril	Rational mutagenesis	Garcia-Seisdedos et al.[24]
L-Fucose mutarotase			1	Yes	Derived tyrosine mediates eight isologous contacts between decamers in a fibril	Rational mutagenesis	Garcia-Seisdedos et al.[24]
AsnC			2	Yes	2 derived tyrosines mediate isologous contacts with another octamer	Rational mutagenesis	Garcia-Seisdedos et al.[24]

Note: The column labeled mutations indicates the number of amino acid point mutations required, unless otherwise noted.

proteins that form small soluble homomeric complexes, then introduced a few surface mutations that increase hydrophobicity (Figure 1b). They changed no more than three sites per protein, used mutations only to leucine or tyrosine, and made no attempt to design for steric or electrostatic complementarity. In all 12 proteins, the mutations triggered assembly into long fibrils, and in four cases, a single mutation was sufficient. The affinities were high enough that massive assemblies formed at physiological concentrations—in some cases in the
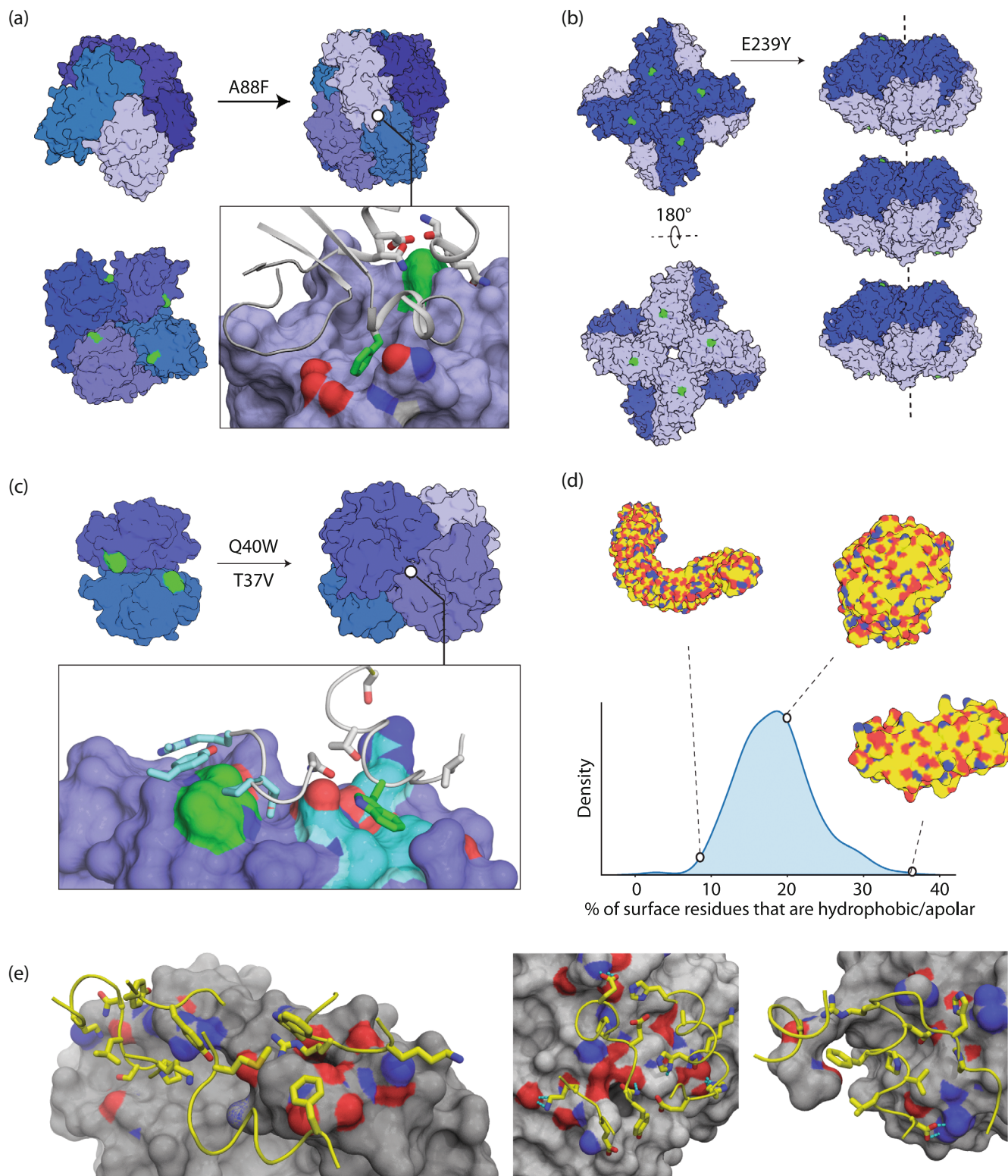

FIGURE 1 Legend on next page.

nanomolar range—and all the proteins remained folded in their native structures, indicating that a specific new interface, rather than unfolding and aggregation, mediated assembly. Altogether, the authors tested 73 mutants, of which 30 triggered formation of fibrous complexes. They concluded that most small multimers exist in evolutionary terms "on the edge of supramolecular assembly."

Other studies make clear that rational design is not necessary to find short genetic paths to the acquisition of multimerization. A small library of just ∼50 random mutants of the monomeric protein GB1 was found to contain a single mutation that yields a dimer, a quartet of substitutions that produced another structurally distinct domain-swapped dimer, and a quintet that produced an intertwined tetramer.[32,41–43] And a directed evolution study using the monomeric protein αE7 carboxylesterase found that multimeric interactions can be acquired even without direct selection: just four rounds of mutation/ shuffling and selection for increased thermal stability yielded a variant that assembled into a mixture of monomers, dimers, and tetramers.[44] This new protein differed from its ancestor by seven mutations, but individual reversion experiments showed that two of these had no effect on multimerization, indicating that it takes at most five to confer multimerization.

2.2 | Historical evolution of new multimers

Evolutionary case studies show that large-effect mutations have also played a causal role in the historical origin of biologically important protein complexes.[30,34] One study focused on the origin of hemoglobin, the primary oxygen transport and exchange protein in jawed vertebrates. Hemoglobin's paralogous subunits Hb-α and Hb-β are part of a larger family of globins, and those closely related to vertebrate Hbs are mostly monomeric.[30] The authors used ancestral sequence reconstruction to infer the sequences of ancient hemoglobin progenitors and biochemical experiments to characterize their ability to multimerize. This work established that tetramerization evolved from an ancestral dimer immediately after the gene duplication that yielded separate Hb-α and Hb-β proteins. When just two historical substitutions that occurred at the tetramerization surface of Hb-β during this interval were introduced into the ancestral dimer, they conferred high-affinity assembly into tetramers (Figure 1c). As in the biochemical study of supramolecular assembly, one of these replaced a glutamine with a large hydrophobic tryptophan, which fit into a preexisting hydrophobic cavity on the facing subunit, and this interaction occurs twice in the interfaces by which two heterodimers assemble into the hemoglobin tetramer.

Another ancestral reconstruction study addressed how a monomeric enzyme evolved to form a novel heteromeric complex that plays a key role in the organization of cells within multicellular animals.[34] The protein–protein interaction domain (PID) of the Discs Large (Dlg) protein serves as a scaffolding molecule that orients the mitotic spindle relative to cues at the cell surface in metazoans. PID descends from a much older family of monomeric enzymes, the guanylate kinases (GKenz), which produce GTP by phosphorylating ATP and do not interact with any of Dlg-PID's protein partners. Ancestral reconstruction experiments showed that the PID's ability to interact with one of its protein partners was acquired just after duplication of the common the gene duplication ancestor of GKenz and Dlg-PID, which had robust

FIGURE 1 Acquisition of multimeric interactions by one or a few mutations. (a) A single amino acid replacement in tetrameric L-rhamnulose-1-phosphate aldolase confers assembly into octamers via an isologous interface.[23] Identical subunits shown in different shades of blue. Acquired phenylalanine side chain is shown in green on each subunit. Inset: close view of interface, with one subunit shown as blue surface and the other as white cartoon and sticks. The substituted site is shown in green on each subunit. Three residues contacting the substituted side chain are colored by element on each subunit (red, oxygen; blue, nitrogen). (b) A single amino acid replacement in isoaspartyl dipeptidase, a homo-octamer (left), confers assembly into long fibrils (right, with axis of fibril assembly as a dotted line). The octamer is shown from above and below, using two shades of blue to distinguish subunits. The tyrosine mutation, which occurs eight times in the octamer, is shown in green.[24] (c) Two historical substitutions confer tetramerization on ancestral hemoglobin dimer.[30] Top: subunits of ancestral dimer and derived tetramer are shown in different shades of blue. Derived tryptophan residue in the new interface is shown in green. Inset: close view of new interface, with one subunit shown as surface and the other as cartoon and sticks. Derived tryptophan is shown in green on both subunits. Residues that contact the subunit are shown in cyan. Other conserved side chains that contribute to the interface are shown as white sticks. Red, oxygen; blue, nitrogen. (d) Many nonmultimeric proteins have hydrophobic surface that can potentially mediate new interactions. A histogram is shown of the fraction of surface-exposed residues that are hydrophobic in a dataset of monomeric protein structures.[54] Three monomers from this dataset (1o6v, 1yqs, and 1cpq) are shown on the distribution. Monomer surface is colored by atom: red, oxygen; blue, nitrogen; yellow, carbon. (e) Multimeric structures that are crystallographic packing artifacts (see Refs. 23,56,57. PDBs from left to right: 104l, 3pbg, 1b6b). In each, one subunit is shown as gray surface and the other as yellow cartoon and sticks. Hydrogen bonds are shown as cyan dashes

GTPase activity and no detectable protein-binding affinity. Introducing either one of two substitutions that occurred during this interval into the monomeric ancestral enzyme was found to confer micromolar protein-binding affinity. In this case, the mutations were in a hinge region and appear to increase the accessibility of the protein binding surface rather than changing its surface properties.

2.3 | Biophysical causes of easy interface evolution

What mechanisms explain how entirely new, high-affinity interactions can be conferred on monomeric proteins by just one or a few mutations? Two fundamental features of protein biochemistry help explain this apparently counterintuitive result.

First, the scale of the nonlinear relationship between the occupancy of a state and its free energy mean that one or two favorable mutations can increase a protein's propensity to multimerize by orders of magnitude. It has been argued that mutations that create a new hydrogen bond across an interface will contribute only weakly to the energy of binding—typically <1 kcal/mol—because the hydrogen bond also forms in the monomeric state, via interactions with water.[36,45] But other kinds of mutations have much larger effects. A single mutation that satisfies or removes an unpaired hydrogen bond donor or acceptor in a buried interface can strengthen the interaction by up to 4 kcal/mol (16 kJ/mol).[46,47] And a mutation that introduces a hydrophobic group that is complementary to the opposing interface can strengthen the interaction by a similar amount.[48–50]

Mutations with energetic effects of this magnitude can have large effects on multimerization, because occupancy is an exponential function of energy* (Figure 2a). For a protein with an initially low propensity to dimerize (e.g., 1%) a mutation that improves the energy of binding by 4 kcal/mol will increase the occupancy of the dimeric state by about two orders of magnitude (to 75%, orange rectangle). Two mutations, each contributing 4 kcal/mol, could increase occupancy of the dimer by four orders of magnitude (e.g., from 0.01 to 90%).

The second critical factor is that most homomeric interfaces are isologous: they involve the same surface region on the two subunits, rotated 180° relative to each other around an axis of symmetry (Figure 2b).[11,23,51] A single mutation in an isologous interface will change residues on both sides of the interface, doubling its energetic contribution and squaring the effect on occupancy. One favorable mutation can thus have the same effect as two mutations in a nonisologous interface, improving binding by up to 8 kcal/mol and occupancy by four orders of magnitude. For higher-order homomeric complexes, the effect is magnified further: a single affinity-increasing an
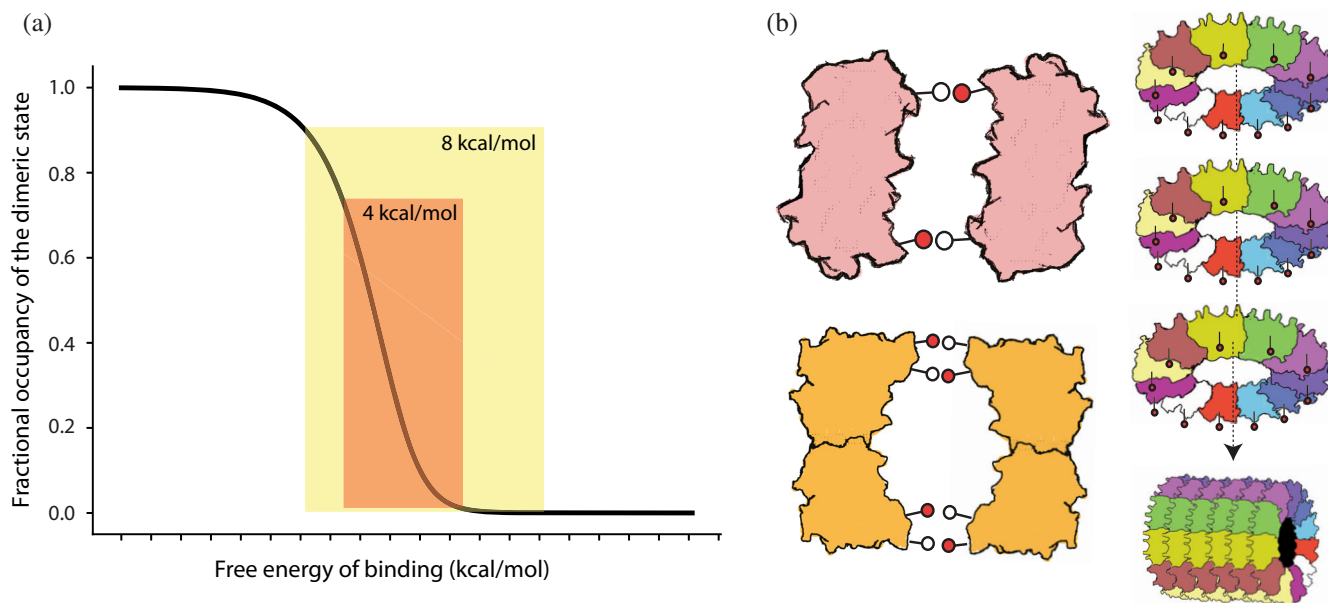

FIGURE 2 Isology and the evolution of multimerization. (a) Non-linear relationship between the free energy of assembly into a dimer, and fraction of monomers associated into the dimeric state is shown. Boxes indicate the effect on occupancy of a mutation that modifies energy by 4 kcal/mol (orange) and 8 kcal/mol (yellow). (b) Repetition of interactions in an isologous interface. In a dihedral dimer, a single favorable mutation (red) occurs twice—once on each subunit. In a tetramer (orange), it occurs four times. In a fibril with asymmetric interactions (right), a mutation on each subunit (black stick) is repeated without limit.

isologous interface occurs four times in a homotetramer (increasing occupancy by up to 8 orders of magnitude), eight times in a homo-octamer, and a vast number of times in a fibril.[23,24,30]

2.4 | Fortuitous foundations for new multimers

Another factor that makes it easy to acquire new multimers is that monomers fortuitously contain surface features that can contribute to the affinity of a new interaction. For example, replacing one polar surface residue with phenylalanine transforms the tetramer L-rhamnulose-1 phosphate aldolase (Rua) into an octamer, because the side-chain of Phe packs into a small hydrophobic cleft formed by several existing residues on the facing surface (Figure 1a). Similarly, the historical tryptophan substitution that played a key role in the evolution of hemoglobin tetramerization fits into a small cavity composed of hydrophobic residues that were already present in the ancestor. In both *Rua* and hemoglobin, these new interactions were supplemented by other favorable interactions—hydrogen bonds and hydrophobic packing—among other existing residues in the interface. On their own, these pre-existing interactions were not sufficient to drive high occupancy of the multimer, but they created a context in which the even more favorable mutation created by the focal mutation could do so.

Most monomers contain large hydrophobic patches on their surfaces that can provide a foundation for new multimeric ineractions.[52–54] One study of a large number of protein surfaces estimated that the average 1,000 Å^2 exposed patch is just two substitutions away from being as hydrophobic as the average protein–protein interface; many such patches are just one substitution away from this level of hydrophobicity, and some are already there[52] (Figure 1e). Because most of the affinity in multimeric interactions comes from hydrophobic interactions,[46] these patches have the potential to mediate fortuitous new interactions, so long as electrostatic or steric clashes do not prevent them from doing so. A mutation that resolves a clash between subunits or adds a new hydrophobic residue could therefore yield a high-affinity interaction mediated by these pre-existing hydrophobic surfaces.

The surface properties that poise proteins on the edge of multimerization appear to occur fortuitously. One line of evidence comes from the finding that proteins that do not co-occur in nature often co-assemble when mixed in the laboratory. For example, when human proteins are expressed in *Escherichia coli* cells and then characterized by in-cell NMR, they are more likely to form complexes

with the host proteins than the *E. coli* proteins do with themselves.[38] The two sets of proteins have not encountered each other for billions of years, so these interactions must be entirely fortuitous. A similar lesson comes from X-ray crystallography, which is possible because many proteins fortuitously form repeating homomultimeric structures: although particular conditions must be imposed for these crystal interactions to be realized, the interfaces involved often resemble those of biological multimers in their size and complementarity[24,39,55–57] (Figure 1e). These crystallographic interactions do not occur biologically, so they could not have arisen because of selection.

3 | SHORT PATHS TO ALLOSTERY

Allostery—changes in a protein's activity, such as ligand-binding or catalysis, caused by binding to an effector molecule[51]—arises if three necessary and sufficient conditions are present (Figure 3a): (a) a protein must bind an effector at a site that is structurally distinct from the active site; (b) the protein must be capable of occupying conformations that differ in their functional activity; and (c) effector binding must be associated with a difference in the relative stability—and therefore occupancy—of active versus inactive conformations. One might expect that the evolution of allostery from a nonallosteric precursor would require a protein to acquire each of these properties, and—because each property involves many residues in the protein—that doing so would require many mutations. In fact, recent biochemical and evolutionary studies show that allostery can be acquired via just a few genetic changes, because many natural nonallosteric proteins often already possess several of the necessary conditions for allostery. These studies are listed in Table 2, and we discuss a few highlights here.

3.1 | Engineering allostery

Protein engineering studies of four unrelated proteins with dramatically different folds and functions—green fluorescent protein, β-glycosidase, β-glucoronidase, and YeaZ—have shown that a single mutation can confer allostery by causing the protein to both bind an effector and become allosterically dependent on it[37,58] (Figure 3b, c). In every case, the same kind of mutation was involved: replacing a tryptophan with glycine >9 Å from the active site. The mutation creates a small cavity, which destabilizes the active conformation and dramatically reduces the protein's activity; however, the small molecule indole, which structurally resembles the side chain

of tryptophan, binds in the cavity, restabilizes the active conformation, and restores activity.[37,58] There was no need to create a pathway to "transmit information" from the effector site to the active site; the active site was already sensitive to the conformation and stability at the site of the mutation.

The bacterial enzyme TEM-1 beta lactamase is nonallosteric, but mutations that create a new site for effector binding are sufficient to confer allostery (Figure 3d,e). In TEM-1's active conformation, two structural regions of the protein separated by a flexible hinge are oriented closely together, and residues from both regions comprise the active site. Introducing just two histidine mutations into the hinge of a nonallosteric TEM-1 variant creates a binding site for metal ions, and metal binding allosterically downregulates catalytic activity by reducing occupancy of the active conformation.[59] TEM-1 can also be mutated to acquire positive allosteric regulation: inserting a maltose-binding protein (MBP) domain interferes with the active conformation, but binding maltose
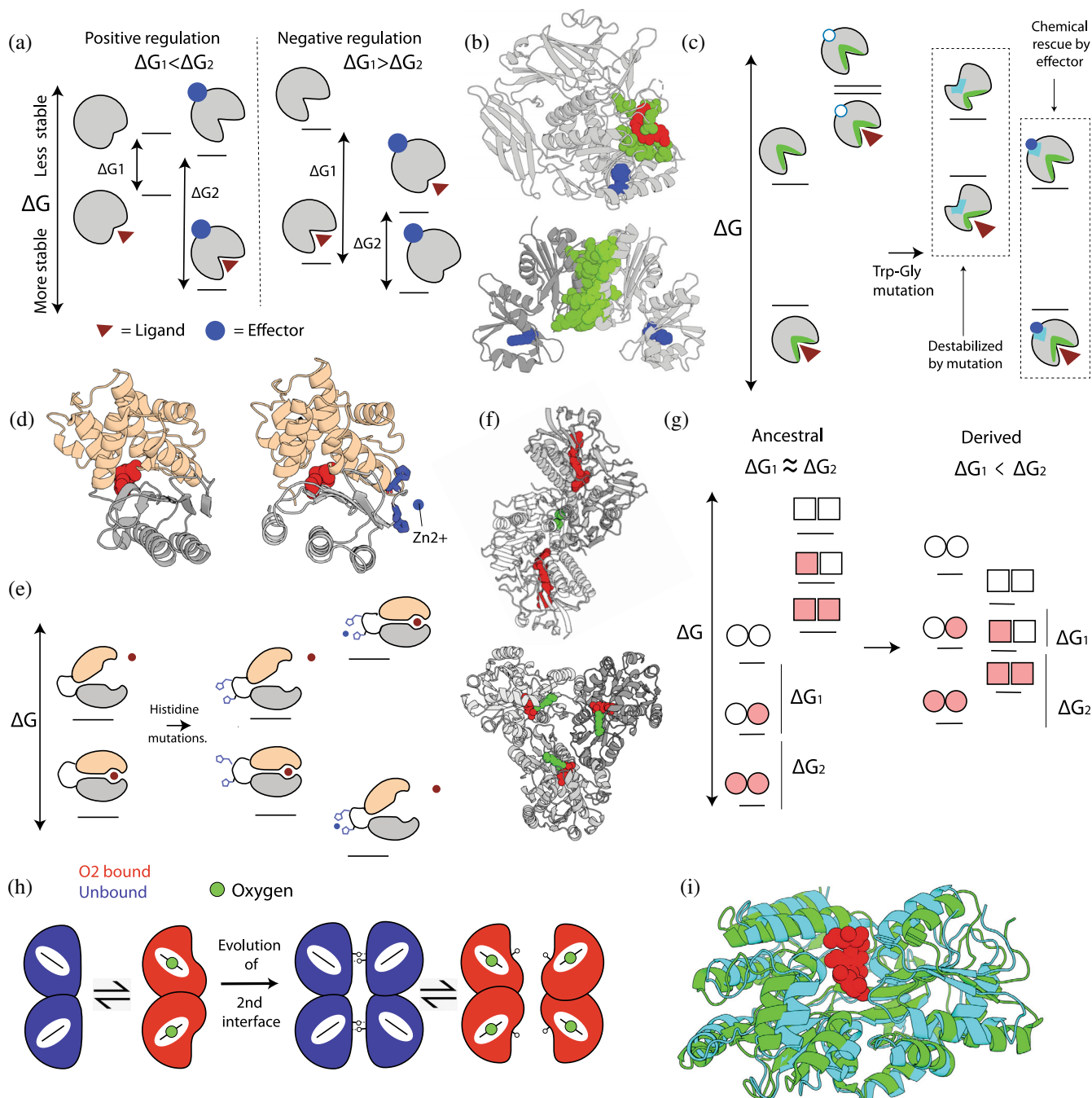

FIGURE 3 Legend on next page.

relieves this inhibition and restores activity.[60,64] Inserting an effector-binding domain confers allostery on other proteins, too, including alkaline phosphatase[70] and dihydrofolate reductase.[126] In all these cases, creating a new effector binding site by fusion or point mutation is sufficient to confer allostery because the wild-type protein already possesses two of the three properties critical for allostery—the capacity to occupy multiple conformations that differ in their activity, and sensitivity of those conformations to the state at a different location in the structure.

Sometimes nonallosteric proteins can already bind a ligand far from the active site, but this has no allosteric effect on the protein's function; in these cases, a mutation that changes the relative stability of the liganded state versus the unbound state can confer allostery. For example, in five unrelated enzymes that are noncooperative multimers with isologous interfaces, a single amino acid replacement or deletion generates strong cooperative activity.[61,62,65,66,72] (Cooperativity is a form of allostery in multimeric proteins, in which activity in one subunit increases the activity of another subunit; the ligand at the active site of one subunit serves as an effector for the other.) Homomers with isologous interfaces are particularly likely to evolve cooperativity.[63] because if the active site for one subunit is conformationally coupled to the multimerization interface, then the other must be identically coupled. In some of the enzymes studied, the causal mutation was at the interface between subunits, while in others it was in the active site (Figure 3f,g). The precise structural mechanisms were not directly identified, but an interface mutation could confer cooperativity by simply changing the multimerization affinity of a liganded subunit for an unliganded subunit, relative to the affinity when both are liganded (or both unliganded). An active-site mutation could confer cooperativity by tuning the relative stability of the active conformation when the other subunit is liganded versus when it is unliganded.

There are even cases in which all three requirements for allostery are already present in nonallosteric proteins: no effector exists in nature, but these proteins can be regulated by exogenous allosteric drugs or other effectors.[73,74]

For example, aminoglycoside phosphotransferase (3′)-IIIa (APH) is a bacterial kinase with no known natural allosteric effectors. One study created a large library of potential effectors by randomizing the residues along one surface of an ankyrin repeat protein (ARP), a stable

FIGURE 3 Acquisition of allosteric regulation. (a) Free-energy landscapes corresponding to positive (left) and negative (right) allostery. Ligand is shown as a red triangle and effector as a blue circle. ΔG_1 and ΔG_2 indicate the change in free energy upon ligand binding in the absence and presence of the effector, respectively; lower ΔG corresponds to increased stability and occupancy. Effector binding (right column) is associated with a change in conformation that increases activity. (b) A single mutation from tryptophan (dark blue) to glycine confers allosteric regulation by indole in glucuronidase (top) and yeaZ (bottom).[37,58] The wild-type tryptophan residue is shown in dark blue. In each case, indole binds at the mutated site. Green spheres show the catalytic residues in glucuronidase and the dimer interface residues in yeaZ. The substrate for glucoronidase is red. (c) Schematic of the change in the free-energy landscape caused by the Trp-Gly mutations depicted in panel b. Blue circle shows the indole effector; cyan border shows the new binding surface; red triangle shows the substrate. Before the mutation (left), activity does not depend on indole binding, which is not favorable. The mutation destabilizes the active conformation in the absence of indole, but binding indole (which is now favored) restores its stability. (d) Engineering allostery into TEM1 beta-lactamase (left: wild-type TEM-1 [PDB ID: 1erm]; right: alphafold prediction of cp-TEM-1). The protein's two domains are shown in beige and grey. Red spheres, substrate. Two mutations to histidine (blue) in the hinge between the domains result in binding a zinc ion (blue sphere), which inhibits catalytic activity.[59,60] (e) Schematic of the change in the free-energy landscape caused by the histidine mutations in panel d. Red, ligand; open blue circles, metal-coordinating histidines, and filled blue circle is the zinc ion. Binding the ion reduces the stability of the inactive conformation. (f) Engineering cooperativity into glutathione reductase[62] (top, GTR, and PDB ID 1ger) and ornithine transcarbamoylase[66] (bottom, OTC, PDB ID: 2otc. In both cases, a single mutation conferred cooperativity on a noncooperative enzyme. Subunits in each homomultimer are shown in white and gray. The ligand is red and the cooperativity-inducing mutation is a green sphere. In GTR, the mutation is near the dimerization interface; in OTC, it is near the active site. (g) Schematic for the evolution of cooperativity via changes in the free-energy landscape. In a non-cooperative multimer, the free energy difference of the first (ΔG_1) and second (ΔG_2) ligand-binding events are equal. The conformation with higher affinity for ligand (circles) is always favored over the lower affinity conformation (square). Filled and empty shapes represent ligand-bound and -unbound states, respectively. Cooperativity can be acquired by stabilizing the inactive conformation or destabilizing the active conformation (arrows). This change makes the second binding event more favorable than the first. (h) Evolution of cooperativity in hemoglobin (see Ref. 30). The precursor of the hemoglobin tetramer was a noncooperative homodimer; acquisition of tetramerization conferred cooperativity. Left: Schematic of change caused by tetramerization in the energies of active (red, high affinity for oxygen, shown as green circle) and inactive (blue, low oxygen affinity) states, both of which existed in the ancestral dimer. Hemoglobin's tetramer interface changes conformation when oxygen binds, altering its orientation relative to other subunits, causing them to also favor the conformation with high oxygen affinity. (i) Conformational heterogeneity in maltose-binding protein, which is nonallosteric. Structures of maltose-bound (blue, PDB ID: 3mbp) and unbound (green, PDB ID: 1omp) protein showing ligand-induced conformational changes. Ligand is shown in red

TABLE 2 Protein engineering and evolutionary studies in which one or a few mutations confer allostery

Protein system	Initial state	Derived state	Mutations	Notes on mechanism	
cpTEM1 beta − lactamase			2	Construction of metal binding pocket stabilizes the inactive conformation of the enzyme.	Mathieu et al.[59]
Ancestral hemoglobin precursor (Anc_α/β)			2	Evolution of a tetramerization interface in an ancestral dimer with high O2 affinity (R) stabilizes a lower-affinity conformation and confers cooperativity.	Pillai et al.[30]
TEM1-beta-lactamase + MBP			Insertion	Recombination of the two proteins yields maltose-sensitive TEM-beta lactamase activity	Guntas et al.[64]
KSS1			2-4	Design of novel phosphosites allows for inhibition or activation of KSS1 activity	Pincus et al.[156]
Beta-glycosidase Beta-Glucoronidase			1	Chemical rescue of a W>G mutant by addition of indole, allows for positive regulation by indole	Deckert et al.[37]
YeaZ			1	Chemical rescue of dimerization in a W>G mutant by addition of indole, allows for positive regulation by indole	Xia et al.[58]
Pyruvate kinase M1			One-residue deletion	Removal of a single C-terminal residue allows for FBP driven inhibition in a non-allosteric iso-form of pyruvate kinase.	Ikeda et al.[65]
Ornithine transcarbamoylase			1	Substitution at the intersubunit interface allows for cooperativity.	Kuo et al.[66]
Glutathione reductase			1	Substitution at the intersubunit interface allows for cooperativity.	Scrutton et al.[62]
Aspartate transcarbamoylase			1	Substitution at the active site allows for cooperativity.	Stebbins et al.[61]
CEBPB			3	Historical substitutions abolish ancestral positive regulation and generate novel negative phospho regulation.	Lynch et al.[67]
Tryptophan synthase			4	Historical substitutions switch ancestral negative regulation to positive regulation.	Schupfner et al.[68]
Aminoglycoside phosphotransferase			0	Binding ankyrin repeat proteins regulates the wild-type nonallosteric enzyme.	Kohl et al.[69]

Note: The column labeled mutations indicates the number of amino acid point mutations required, unless otherwise noted. Conformational ensembles associated with the initial (nonallosteric) and derived (allosteric) states are shown in schematic form.

Abbreviations: A, active conformations; I, inactive.

protein with a well-defined binding surface. Screening the library identified hundreds of ARPs that negatively regulate the wild-type enzyme, one of which bound with nanomolar binding affinity that completely abolished catalytic activity. An X-ray crystal structure showed that this effector bound the enzyme at a surface site >20 Å from the active site and stabilized several secondary elements of the protein in an inactive conformation.[69] The

number of mutations required to confer allostery on the APH protein is therefore zero.

3.2 | Natural evolution of allostery

Just as allostery can be engineered in the lab with a few mutations, several historical case studies show that allostery has been acquired during natural evolution by very short paths (Table 2).

The hemoglobin tetramer, for example, binds oxygen cooperatively. Ancestral protein reconstruction showed that hemoglobin acquired cooperativity during the same phylogenetic interval when the tetrameric architecture evolved from an ancestral, noncooperative dimer. The two historical surface substitutions that confer tetramerization on the ancestral dimer—which was not cooperative—also confer cooperativity. The structure of hemoglobin suggests why conferring tetramerization also yields this form of allostery: the surface that mediates tetramer assembly is connected by a short helix to the protein's active site, which subtly changes conformation when oxygen is bound, altering the shape of the binding interface. The association between the active site and this part of the surface simply arises from the globin fold; even myoglobin, a monomer that is necessarily noncooperative, undergoes a similar conformational change when it binds oxygen.[75,76] Allostery arises immediately when binding to an effector is conferred via the new tetramerization interface, simply because the conformation that is optimal for oxygen binding is not optimal for tetramerization, and vice versa (Figure 3h).

A few historical studies have also documented allosteric inversions, in which positive allostery has been gained from a negatively allosteric ancestor (or vice versa), by just a few mutations.[67,68] For example, ancestral reconstruction studies show that the ancestral transcription factor CEBPB was negatively regulated by phosphorylation, but allosteric activation by phosphorylation was gained on the branch leading to eutherian mammals.[67] Just three historical amino acid replacements are sufficient to recapitulate this inversion. Two of these abolish phosphorylatable serines that favored the inactive conformation when phosphorylated, while the third introduces a new serine that makes the protein's activity phosphodependent. By tuning the relative stabilities of the protein's conformations in response to phosphorylation, these mutations abolished an ancient form of regulation and—with a single new mutation—conferred a new one.

Finally, just as some nonallosteric proteins can be regulated by allosteric drugs, the yeast protein Fus3 apparently had all the properties required for an allosteric response, before its effector evolved.[77] Fus3's activity in the mating cascade in *S. cerevisiae* is allosterically upregulated by the VWA domain of a protein called Ste5. Ste5 is a relatively recent evolutionary newcomer, but Fus3 proteins from a wide variety of fungal species—even those that contain no Ste5 family members—can be regulated by heterologous copies of *S. cereivisiae* Ste5. Phylogenetic analysis shows that Fus3's capacity to be allosterically regulated by Ste5 existed in latent form long before Ste5 itself evolved. This suggests that when Ste5 first evolved, it bound immediately to Fus3—at a site where no other regulator is known to bind—and fortuitously upregulated its activity. As in the case of allosteric drugs, zero mutations in the protein itself were required to confer allostery on it.[77]

3.3 | Fortuitous foundations for acquiring allostery

Why would nonallosteric proteins possess features that are prerequisites for allostery—such as occupancy of conformations that are functionally distinct, or conformational linkages between a potential effector-binding site and the protein's active site—if the protein is not already allosterically regulated? Like the precursors of multimerization, these features arise as fortuitous features of protein architecture.

Virtually all proteins meet the first condition for allostery: they occupy an ensemble of conformations, rather than a single rigid structure.[78–83] Although physical constraints limit the number of conformations that a protein actually occupies, many degrees of freedom remain, so there are typically a vast number of subtly different conformations—differing in the angle of a helix or the rotamers of side chains, for example—with energies similar enough to allow nontrivial occupancy.[84] Even proteins that almost exclusively occupy a single conformation have the capacity to occupy others, if the energy differences between the conformations are narrowed by one or a few mutations that destabilize the major conformation or stabilize others.[79] Because of the exponential relationship between changes in the free energy of a fold and its occupancy, a single mutation with a small effect on stability can have a large effect on the protein's ensemble of conformations (Figure 2a).

The second condition—that the ensemble of folded conformations includes some that are more active than others—is also a common feature of nonallosteric proteins.[85,86] Many ligand-binding proteins exhibit conformational selection: they can occupy multiple conformations that vary in affinity, and the ligand binds predominantly to those with high affinity (Figure 3i).[80,87,88] Similarly, many unbound enzymes transiently occupy both active

and inactive conformations, with preference for the active conformation when their substrate is bound.[89,90] The functional heterogeneity of conformations arises simply because active sites have more stringent physical requirements than the fold as a whole. Affinity for a ligand can be impaired by subtle changes that disrupt the steric or electrostatic complementarity of binding surfaces, and catalytic activity involves even more precise constraints. Some of the many conformations that are occupied by a typical protein will inevitably satisfy these functional requirements better than others.

Finally, conformational changes in one region of a protein are often associated with changes elsewhere. Studies of the structural effects of mutations suggest that genetically perturbing nonallosteric proteins often changes the conformation at multiple parts of the same protein, often 10–20 Å away from the mutant sites[88,91,92] A recent deep mutational scan of PDZ and SH3 proteins found that a huge number of single mutations >5 Å away from the binding interface can change the proteins' ligand affinity, often reducing but sometimes improving it.[93] These proteins are not allosterically regulated, and the mutations do not occur in any known natural variants, so their effects cannot be attributable to selection. These couplings arise as a consequence of the constraints imposed by a protein's structures. Although proteins have extensive conformational freedom, that freedom is not infinite. A subtle change in the position of a helix, for example, will alter the most favorable position of other residues and secondary elements that interact with it, and these changes may propagate further through the protein's structure. Perturbing one part of a protein therefore typically changes the lowest-energy conformation at other locations. Sometimes the linked regions happen to be the active site and an effector-binding surface.

Nonallosteric proteins that have fortuitous properties like long-distance couplings or effector binding may therefore often be poised in sequence space on the evolutionary edge of allostery. A single mutation that adds the remaining requirement—such as effector binding or tuning the relative stabilities of active and inactive conformations—can trigger the acquisition of allosteric regulation.

4 | SHORT PATHS TO PROTEIN FOLDS

There are over 2,000 known protein folds in nature.[94] Information about the mechanisms by which these folds originated—either by descent from other folds or de novo folding from unstructured precursors—is relatively scanty, because these events occur far more rarely than multimerization or allostery are acquired. Several case studies suggest, however, that simple mutations can create entirely new folds or make simpler folds suddenly more elaborate.

4.1 | Short bridges between folds

Three experimental studies show that the genetic paths between entirely different protein folds can, in some cases at least, be surprisingly short. In a classic biochemical study, it took only two point mutations to reorganize the secondary structure and tertiary fold of the N-terminal portion of the Arc repressor[95] (Figure 4a). One of these mutations on its own resulted in partial occupancy of both folds, depending on experimental conditions. The bridge in sequence space between the folds is
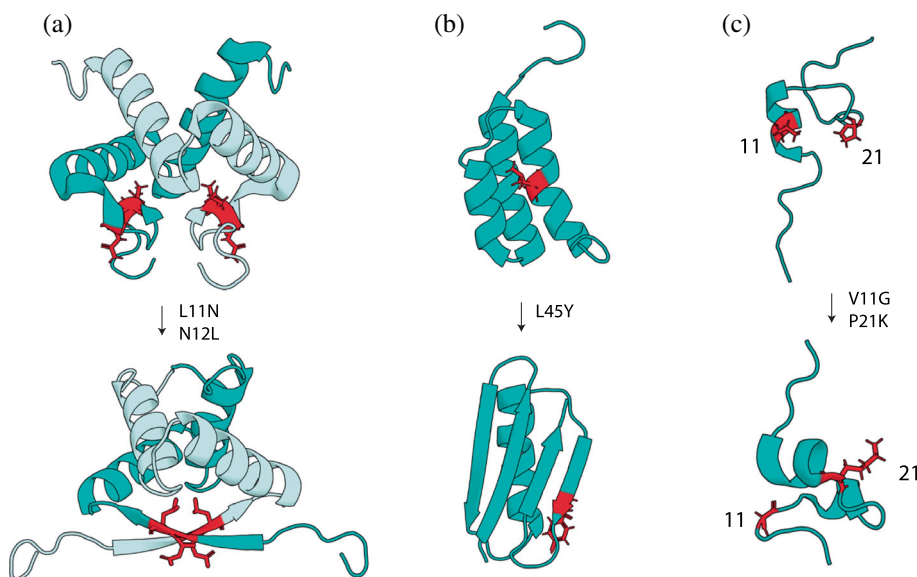

FIGURE 4 Acquisition of new folds by point mutations. 1 or 2 point mutations deliver a change in fold in (a) Arc repressor (top, PDB ID: 1qtg; bottom, PDB ID: 1arr), (b) Streptococcal protein G_A domain (top, PDB ID: 2kdl; bottom, PDB ID: 2kdm), and (c) cysteine-rich protein NW1 (top, PDB ID: 2hm6; bottom, PDB ID: 2hm3). Mutated residues are shown as red sticks. The two chains of the Arc repressor are shown in different shades of cyan

therefore only two mutations long, and the journey across that bridge does require passing through an unstructured intermediate. The implication is that, even under purifying selection, one fold could rapidly evolve into the other.

A more recent study showed that a global structural reorganization of an entire globular protein can be triggered by a single mutation (Figure 4b).[96] The A and B domains of protein G from *Streptococcus*, each ∼50 amino acids long, are unrelated in sequence and structure: domain A consists of 3 α-helices, and domain B consists of 4β strands and one helix, with a completely different tertiary fold and distinct sets of residues that make up the hydrophobic core. The authors gradually walked the two sequences towards each other in sequence space by swapping individual residues between the proteins without changing or losing the fold of either one. Ultimately, two protein sequences were achieved that differed by a single amino acid, each occupying one of the folds; the remaining mutation completely reorganized either fold into the other. Even though the "wild-type" A and B domains are as distant from each other in sequence space as is possible, a continuous network of sequences encodes each fold and, in at least this one location, the two networks are adjacent to each other.

The third example comes from two paralogous cysteine-rich domains in the cnidarian *Hydra*, which differ at 56% of their residues (Figure 4c).[97] Unlike most homologous proteins with recognizably similar sequences, these two proteins have completely different folds, mediated by distinct patterns of disulfide bridges. Exchanging just two critical residues between the two proteins—which are not cysteines but rather favor different backbone conformations—fully switches one fold into the other. When the mutations are introduced singly, the intermediate proteins occupy both conformations. The sequence networks encoded by these folds are therefore connected by at least two short bridges, neither of which involves an unstructured intermediate. Because the proteins are paralogous, the authors commented that it is likely that one fold evolved from the other via a "smooth transition."

Some natural proteins, too, can occupy two distinct folds, indicating that the sequence networks of the folds overlap. A recent bioinformatic study estimated that up to 4% of all proteins with solved structures have the capacity to occupy more than one distinct fold under different cellular conditions.[98] For example, cyanobacterial protein KaiB can adopt two different tertiary structures, depending on whether it is bound or free of its binding partner KaiC.[99] Similarly, the C-terminal domain of the transcription factor RfaH can transition between alpha-helical hairpin and a β-barrel fold in the cell, depending on whether it is free or bound to the protein's NTD.[100] In proteins like these, a single mutation that stabilizes or destabilizes one of the folds could trigger the transition to near-total occupancy of just one.

It is unknown how many protein folds in nature have originated by fold-switching mutations. Proteins with different folds are typically unalignable (though in rare cases homology between different folds can be detected[101]), so it is impossible to trace the process of descent from potential common ancestors. But this does not necessarily mean such ancestors did not exist. Sequence evolution is constrained primarily by the protein's structure: if a new protein family were born by transition into a new fold, it would immediately be subject to new constraints, and so the traces of their descent from a common ancestral fold would rapidly decay as the sequences diverge.

4.2 | Elaboration of existing folds

Many proteins have complicated structures consisting of multiple layers of secondary elements packed against each other. It is clear that these complex structures can be easily acquired from simpler, more compact folds by elongation of the primary sequence and packing of the extended region against the exterior of the ancestral fold.[102]

In the ligand-binding domain of steroid hormone receptors, for example, one group of paralogs acquired a new carboxy-terminal extension (CTE), a partially beta-containing secondary structure that creates a new layer of tertiary structure by packing against a mostly hydrophobic surface region that was exposed in the simpler ancestral structure.[103] Similarly, a series of three nested insertion events within an ancestral loop of a beta lactamase generated an internal extension, containing both alpha and beta secondary structure, which creates a novel exterior wall by packing along one side of the ancestral fold.[102]

Extensions and insertions are easy to acquire by mutation. C-terminal or N-terminal extensions can arise from point mutations that abolish stop or start codons, and mutations that move an intron splice site can insert new internal coding sequence. Sometimes these extensions and insertions will immediately contribute to the tertiary structure: the majority of random short peptides are predicted to have >25% secondary structure,[104] and, as we have discussed in this review, many protein surfaces can fortuitously bind random peptides. It should be just as likely for a peptide to fortuitously bind the surface of a protein to which it is covalently attached as it is to bind the surface of a disconnected protein the two units are disconnected.

5 | SEQUENCE DEGENERACY OF PROTEIN COMPLEXITY

The second premise of the argument for adaptive gradualism is that genotypes encoding complex features are rare.[2] For the complex features of proteins, this assumption also turns out to be wrong. Comparative structural analyses and high-throughput mutagenesis experiments have shown that a vast number of protein sequences can encode essentially equivalent forms of multimerization, allostery, and tertiary folds. These genotypes are widely dispersed across vast connected regions of sequence space (Box 1). The bridges by which complexity can be acquired are not only short but also numerous.

5.1 | Degenerate multimerization interfaces

Deep mutational scanning experiments that the number of sequences that can encode a multimeric interaction between molecules is huge. A recent study used phage display to characterize binding by several S100 proteins to a library of random peptides 12 residues long (the same length as the region of the peptides that S100s endogenously bind). They found that each S100 protein bound 7–10% of the 30,000 peptides measured, with affinity comparable to S100's biologically relevant binding partners. The vast majority of these peptides do not occur biologically, so these binding activities must be fortuitous rather than the result of selection.[106] Another study randomized 11 specificity-determining residues in the interface of the two-component signaling proteins PhoP and PhoQ and identified >500 unique functional pairs, most of which shared few or no common residues at these sites with each other or the wild-type pair.[107]

A second line of evidence comes from the record of long-term evolution in extant sequences. If interactions were subject to strict sequence constraints, the residues that mediate protein–protein interactions should be strictly constrained and therefore evolve slowly. In fact, many different amino acids are compatible with multimerization. For example, the acetyl-xylo-oligosaccharide esterase (Axe2) in thermophilic bacteria is an obligate

---

BOX 1   Multidimensional sequence space

The concept of sequence space provides a useful organizing metaphor for protein sequence evolution.[105] Sequence space consists of all possible sequences of a given length (the nodes), which are connected to each other by edges if they differ by a single mutation. The number of nodes in a protein space is vast: for an average protein 300 amino acids in length, there are $20^{300} = 10^{390}$ possible sequences, far greater than the number of subatomic particles in the visible universe.

Sequence space is multidimensional and complex. Each site represents a dimension along which a protein can change independently by mut, and (leaving aside the genetic code), there are 19 possible mutations at each site. Any protein is therefore connected to $19*300 = 5,700$ unique neighbors via single amino acid replacements. Each of those has that many neighbors, too, so the starting protein is two steps away from about 18 million other proteins, three steps from 36 billion more, and four steps from 53 trillion others. Even the most distant pairs of proteins in this massive universe, which share no residues in common, are just 300 amino acid replacements apart.

The dense connectivity of sequence space has important consequences for evolutionary processes. Although the set of all possible protein sequences is vast, a protein has the potential to explore a huge number of possible alternative sequences via short evolutionary paths. The simplest and most common model for evolution across sequence space corresponds to drift under purifying selection and a rate of evolution slow enough that individual sequence changes are fixed sequentially rather than simultaneously.[105] In this scenario, an evolving protein can move to any neighboring sequence, as long as the neighbor is functional. The set of functional sequences that are connected to each other by single mutations constitute a neutral network through which a protein may move over time. As long as functional proteins on average have more than one functional neighbor— which is clearly the case for real protein sequences—then neutral networks will be extensive, allowing substantial long-term divergence from the starting sequence. If the networks that encode proteins with different features or functions abut or overlap each other—as the studies that we review here establish—then a new property can evolve from an ancestral protein that lacks the feature, either by drift or positive selection, in as little as a single step.

octamer held together by several clusters of hydrogen bonds, but the states and sites used to mediate these bonds differ extensively among homologs from different species.[108] More generally, across a large database of proteins, the rate of sequence evolution at residues buried in interfaces is only slightly slower than the mean rate over all residues,[109] indicating that proteins can explore very large regions of sequence space while maintaining their multimeric interactions.

Not only can particular surfaces that mediate molecular interactions tolerate many different sequence states, but many different surfaces on proteins have the potential to mediate interactions. For example, members of the globin family have independently evolved to assemble into multimeric complexes multiple times in different animal lineages. On each occasion, largely nonoverlapping surfaces mediate the multimeric interactions[110–113] (Figure 5a). Similarly, the homologous urease enzyme of a plant and a bacteria have independently evolved multimeric structures, but the subunits interact using different parts of their surfaces.[115,116] And some proteins form multimers under crystallographic conditions that use different interfaces from those that they use to multimerize in solution.[117,118]

These data show that proteins can encode a multimeric interaction using a wide variety of different amino states at the same set of sites, using different sites on the same surfaces, or usientirely different parts of the protein. A consequence of this degeneracy is to dramatically increase the number of mutational paths by which an interaction can be acquired.

5.2 | Degeneracy of allostery

Allostery too can be encoded by a huge set of possible sequences, so nonallosteric proteins can acquire allostery by many possible mutations. Alignments of allosteric proteins indic that the sequence networks encoding allostery are vast. For example, LacI and PurR are homologous transcription factors with similar mechanisms of effector-driven allostery, but they share <30% sequence identity.[119–121]

Allostery can also be gained by a variety of mechanisms, increasing genetic degeneracy further. For example multimeric globins in vertebrates,[30] annelids,[112] and mollusks[122] (Figure 5a) independently acquired cooperative oxygen binding; these proteins differ from each other by up to 85%, and different surfaces and conformational mechanisms are involved in mediating the allosteric response to oxygen binding by other subunits.[123,124]

Recent DMS studies of allostery provide a more comprehensive view of the sequence degeneracy of maintaining or gaining allostery.[114,125] In a complete library of all single-site variants of the *E. coli* TetR protein (repressor of tetracycline resistance), 71% of all mutants maintained allosteric regulation—far more than the number that lost allostery (while maintaining the ability to fold and function). Five non-allosteric mutants were then re-mutagenized and screened to identify variants that regained allosteric regulation: of 960 single-site mutants tested, 216 regained allostery. The allostery-restoring mutations were scattered throughout the protein's structure, and typically involved sites other than those that caused allostery to be lost (Figure 5b). Moreover, the sites that affect allostery in the experiments are highly variable in an alignment of TetR orthologs, indicating extensive degeneracy during the historical evolution of allostery.[114]

Allostery can be maintained, acquired, or modified in so many different ways because active sites are fortuitously coupled to sites all across a protein's structure. A study in which a light-sensitive LOV2 domain was inserted at every surface site in DHFR found that 14 of the 61 insertions gained a significant, albeit modest, allosteric response to light.[126] And the DMS scan of PDZ and SH3 discussed above found that ligand affinity was altered by mutations at about half of all surface sites away from the active site.[93] This means that gaining effector binding at any one of a huge number of potential surface regions is likely to trigger an allosteric response. In a protein that already binds a potential effector, then a mutation at any third site coupled to the active site has the potential to affect the stability of active versus inactive conformations, again conferring allostery. As nonallosteric proteins evolve, they are therefore likely to encounter a lengthy menu of genetic options that, if any one of them are ordered up by mutation, can confer allostery upon them.

5.3 | Degeneracy of folds

It is well known that a single protein fold can be encoded by a huge ensembles of variant sequences.[127,128] Even the hydrophobic core of some proteins—the portion that typically evolves slowest—can be replaced with a set of entirely different hydrophobic amino acids while maintaining the native fold.[129] Protein family members that can be structurally superimposed sometimes share no more than 15–20% sequence identity (Figure 5c). The network of sequences encoding a single fold therefore extends across almost the entire "diameter" of sequence space. Moreover, these highly divergent proteins descend from a common ancestor, indicating indicates that proteins across this span are almost certainly connected by paths on which all intermediate sequences can fold and function.
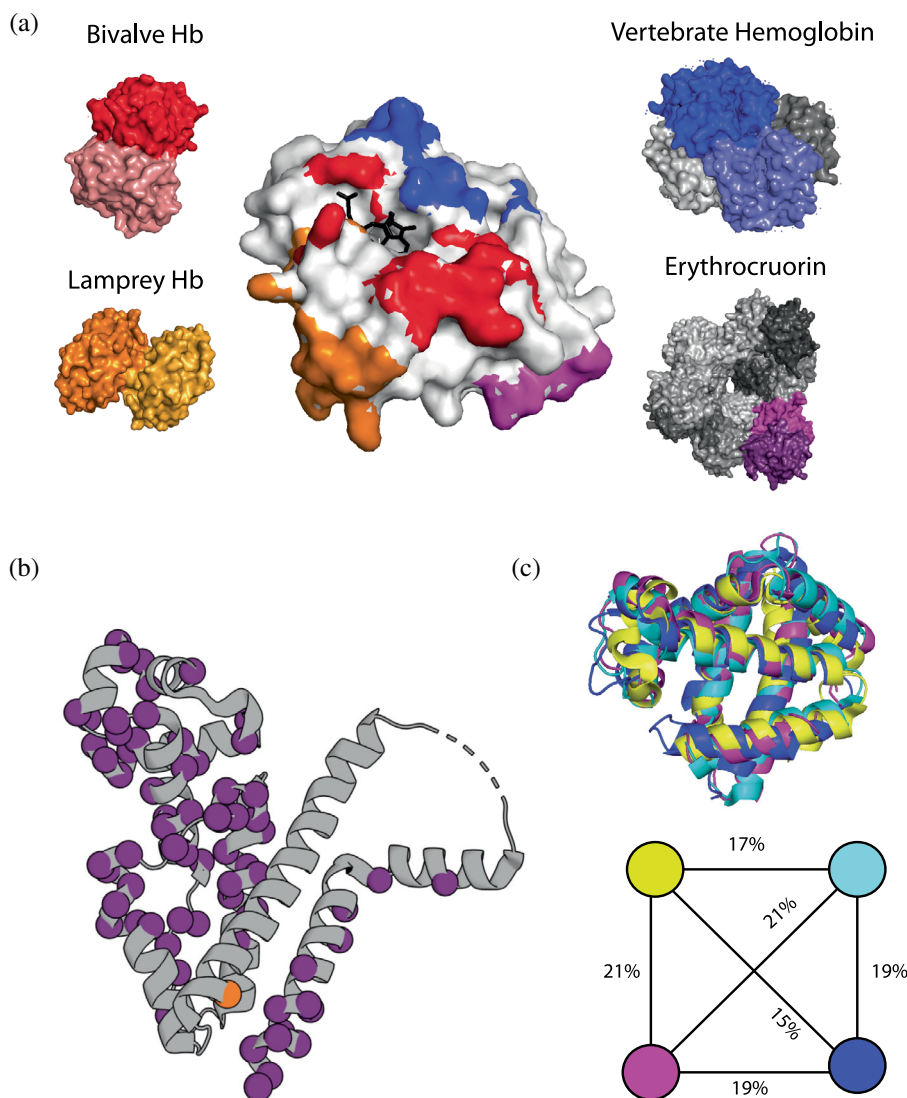
FIGURE 5 Degeneracy of interfaces, allostery, and protein folds. (a) Animal globins convergently evolved multimerization using different surface regions: bivalves (red, PDB code: 1jwn), lampreys (yellow, PDB 3lhb), gnathostome hemoglobin (blue PDB 2qsp) and annelids (purple, PDB 1x9f). In each multimer, subunits are shown with different shades. (For higher-order multimers, a particular interface is indicated, with other subunits shown in gray.) Center: the surface region that mediates assembly of each of these multimers is shown in the corresponding color on the myoglobin molecule (PDB code: 1mbn). The relative locations of these interfaces on the globin fold was determined by structurally aligning myoglobin to a monomeric component in each oligomer. (b) Spatial distribution of second-order mutations that can rescue allostery in a non-allosteric but stable mutant of *E. coli* TetR.[114] The location of the initial allostery-breaking mutation is shown with the alpha-carbon as an orange sphere. Sites at which mutations restore allostery are shown with alpha-carbons as purple spheres. (c) Structural and sequence comparison of four distantly related globins. Top: Structural alignment of neuroglobin (yellow, PDB code: 1oj6), myoglobin (purple, PDB code: 1mbn), cytoglobin (cyan, 1v5h), and lamprey Hb (blue, PDB code: 3lhb). Bottom: Sequence identity among aligned residues for each pair of protein sequences

This extraordinary degeneracy means that proteins can explore vast sequence networks as they evolve under the constraints imposed by maintaining their ancestral fold. As they drift through this network, they may occasionally encounter boundaries of the networks that encode other folds, which are also vast. These bridges may be rare, but over time evolving proteins have an extraordinary number of opportunities to win the find-a-new-fold lottery without paying a price for their losing bet, because purifying selection removes mutations that cause proteins to unfold or aggregate. Moreover, gene duplication—and the functional redundancy it allows—can weaken the constraints imposed by purifying selection to maintain the ancestral function. Along with de novo origin of simple folds, evolutionary transitions from one fold to another need not have been frequent to

explain the origin of the few thousand known protein folds that exist during the course of four billion years of massively parallel evolution.

6 | EVOLUTIONARY PROCESSES AND PROTEIN COMPLEXITY

The evolution of any feature can be pictured in two steps: first, one or more mutations that produce the feature must arise in an individual, and those mutations must then become fixed in the population under the influence of selection and/or drift. The classical view of protein complexity entails interrelated assumptions about both parts of this process. With respect to mutation, the first assumption is that complex features require many mutations that gradually increase complexity from simpler ancestral forms; the second assis that there are few such sets. The final assumption pertains to fixation: complex features are assumed to be adaptive, which allows selection to drive the serial fixation of complexity-conferring mutations. The third assumption is necessary given the first two: if acquiring a complex features requires a particular set of many mutations, it would be vanishingly improbable for them all to fix by chance. If each mutation along the path were to enhance fitness, however, then selection would reliably drive the fixation of each one, given large population sizes and long enough timescales.

The problem with the classic argument is that its premises, rather than its logic, are wrong. We know now that multimerization, allostery, and new folds can be acquired from naturally occurring proteins via paths that are just one or a few mutations long, and the number of such possible paths is often very large. These findings open the door for non-classical explanations of how complexity arises and is fixed during evolution.[19,20,130,131] When there are many complexity-conferring genotypes, and each requires just one or a few of many possible mutations, then persistent, long-term selection is not necessary to drive the smaller number of fixation events that are required. It becomes much more plausible that any one of the many possible paths to complexity would be followed under the influence of drift alone. Scenarios involving weak or transient selection, alone or in tandem with drift[132] or linkage, also become plausible.[133]

6.1 | Neutral acquisition of complex features

Neutral mutations that confer complex features would go to fixation by the same stochastic processes that cause other neutral sequence substitutions, a process that is well understood and essentially universal.[132] Most mutations— whether neutral or beneficial—are lost by chance in the first few generations after they arise, because they occur in a single copy of a gene, and that copy is often unlucky in the lottery by which genes are transmitted to the next generation.[132] Some mutations escape this fate and increase in frequency, and some —even neutral mutations— eventually reach fixation, simply because the frequency of an allele in a population fluctuates stochastically across generations. Any particular mutation is more likely to fix if it is beneficial than if it is neutral—and it takes less time for it to do so[134]—but there are so many possible neutral mutations, and they arise so regularly, that a parade of neutral substitutions is constantly evolving over time in natural populations. Neutral sequence evolution takes place at a constant rate irrespective of population size, because any neutral mutation is more likely to fix in a small population, but more such mutations are generated in large populations, and the effect of population size therefore cancels out.[135] In the end, the critical factor determining the rate of neutral substitution in a protein is the neutral mutation rate—the number of new neutral mutations in each copy of the gene coding for the protein in each generation. The resulting neutral sequence drift is the primary cause of diversity among protein sequences that have similar.

The dynamics of neutral sequence evolution can be extended to understand the evolution of neutral functional or biochemical features, like multimerization, allostery, or a new fold (see Ref. 136). If it takes only one neutral mutation to confer the feature, then the critical factor determining the rate at which the feature will evolve (and thus the probability that it will evolve in any particular period of time) is the overall mutation rate and the degeneracy of the feature—the number of possible amino acid replacements that have the potential to confer it. If instead it takes several mutations to confer the feature, the process will take longer, but the probability that such a path will be followed again increases with degeneracy—which is now the number of *sets* of mutations that can confer the feature. If the mutations are beneficial, this would speed the process,[134,137] whether the selection pressure is strong and sustained—as in the classic model—or if it is weak or intermittent. Even paths involving weakly deleterious intermediate steps can be followed, but the plausibility of this scenario depends on the population size and recombination rate.[130,136,138–140]

Our argument is not that complex protein features are never adaptive or that selection never drives their acquisition. Rather, the point is that the evolutionary forces that can generate these features are much more diverse than the classical picture in which sustained

selection is always required to drive a long series of rare mutations. Complexity may sometimes be acquired neutrally via one a few mutations; it may sometimes evolve by short paths driven by selection; and in other cases, it may involve long paths in which each step is adaptive, as envisioned by the classic model.

Supporting the idea that protein complexity sometimes evolves neutrally is the fact that many orthologous or paralogous multimers vary in these features but are no more efficient biochemically or functionally than their relatives.[30,103,131,141,142] Experiments show that some multimers can be replaced by simpler versions without any detectable costs to fitness or function.[141,143] Allostery, too, can apparently be acquired neutrally, as demonstrated by the existence of allosteric drugs, which can affect protein activity even though the protein's intrinsic capacity for allostery has no biological function.[73,74]

6.2 | Entrenchment of complex features

A rejoinder to our view might be that complex features often persist for long periods of time, and we might expect that a neutrally acquired feature would be lost as readily as it was gained. Consistent with a neutral explanation, however, complex features have in fact been repeatedly gained and lost in many protein families. For example, many multimeric enzymes exist in a wide array of stoichiometries, both between and within species, as expected if this feature continues to be neutral.[131] And allosteric regulation has been gained and lost within protein families, as well.[144,145]

But some complex features are in fact conserved in protein families over very long periods of time. Some of these may have been adaptive as soon as they were acquired, but purifying selection can preserve even neutrally acquired features. For example, complex features that are inconsequential when they originate can become almost impossible to lose if they become functionally important later—a process that clearly occurs in some cases.[77,146]

Complex features that confer no functional benefit can also persist for long periods of time if they become entrenched and are then preserved by purifying selection. Entrenchment is caused by substitutions that are compatible with the complex state but not with the simpler form; reverting to the simpler ancestral state is then deleterious and is prevented by purifying selection.[19–21,103,141,147–149] An apparently widespread mechanism for the entrenchment of multimerization is a hydrophobic ratchet: sites buried in the interfaces of multimers can neutrally accrue hydrophobic substitutions, but if multimerization were subsequently lost these residues would be exposed to solvent, causing instability and/or aggregation.[103] This process is expected to be nearly universal, because the mutational process, owing to the genetic code, has a propensity to produce hydrophobic residues at a frequency far higher than exposed surfaces on monom can tolerate. A recent analysis showed that a large majority of known multimers are likely entrenched by this mechanism and would persist whether or not they have a useful function.[103]

Functionally inconsequential elaborations of a protein's fold can also be entrenched by a similar ratchet: hydrophobic substitutions can occur neutrally in the buried regions that mediate contact between the ancestral surface and the novel decoration, which makes loss of the decoration subsequently deleterious.[103] Activation by an allosteric effector could also easily become entrenched, because mutations that reduce the stability of the active conformation in the effector's absence could accrue neutrally but make reversion to a constitutively active ancestor deleterious.

6.3 | Purifying selection against complexity

If complex features can be so easily acquired and then entrenched, why doesn't every protein contain end layers of tertiary structure, multimerize into 256-mers, and respond allosterically to hundreds of effectors at sites all across its surface? One answer is that new complex features may often be removed by purifying selection because they are deleterious. New multimeric interactions may obscure active sites, compete against other biologically important interactions, or produce toxic aggregates or fibrils,[136] as in sickle cell disease.[40] Purifying selection clearly removes many fortuitous interactions, as demonstrated by the finding that *E. coli* proteins form more complexes with heterologously expressed human proteins than they do with themselves.[38] Similarly, mutations that confer allostery may inhibit essential protein activities or upregulate them in biologically disruptive times or places. And mutations that make a protein's tertiary structure more elaborate or confer an entirely new fold will by necessity create new surfaces, potentially leading to aggregation or disruptive interactions with other molecules.

The limiting factor in the evolution of protein complexity therefore appears not to be how difficult it is for mutation to produce complex features. Rather, the major brake on the evolution of complexity may be incompatibility between new complex features and the biology of the cell or organism in which they arise. Proteins are constantly bombarded by a hail of mutations, some of which

create new interactions, modes of regulation, or changes in fold and conformation. Occasionally these may immediately enhance a biological process, but more often they will be deleterious—and so removed by purifying selection—or neutral. If the neutral features are not rapidly lost, they can become entrenched despite being functionally gratuitous or, through additional mutations in the future, become incorporated into some biologically significant function.

7 | FUTURE DIRECTIONS

The evidence that we have reviewed here establishes that proteins can acquire complex features by short paths, and in numerous historical cases they have actually done so. But the frequency with which this scenario actually occurs is not empirically known. Several lines of experimental inquiry can help to better sketch a broadly applicable account of the genetic and biophysical mechanisms underlying the evolution of protein complexity.

First, we should more thoroughly catalog historical gains and losses in complex protein features. We have a handful of case-studies in which complexity was gained along some lineage,[30,141,150] but our understanding of natural variation in multimerization and allostery among related proteins in a phylogenetic context is very sparse. Vertebrates account for most of the crystallographic data in the PDB, despite making up <5% of animal diversity and a tiny fraction of all species. With improved taxonomic sampling and biochemical characterization of proteins, it should be possible to identify far more cases in which multimerization, allostery, and changes in tertiary structure were acquired, and to isolate those changes on relatively short phylogenetic branches. Doing so would set the stage for detailed biochemical and genetic analyses of parent and child ancestral proteins to identify the mechanisms by which the complex feature was gained. There may even be variation in multimeric state, allostery, and protein structure *within* natural populations, but there has been virtually no effort to address this possibility outside of humans.[40,151]

Second, we need more data on the degeneracy of complex protein features and the number of mutational paths by which they can be acquired. A handful of studies have used deep mutational scanning and experimental evolution techniques to understand the sequence-function map around extant and ancestral proteins for a few kinds of molecular interactions and forms of allostery.[152,153] But we know of no DMS scanning methods at present for homo-oligomerization or for allosteric control of most functions. Developing high-throughput bulk assays in which protein sequence can be linked to these

features would help to illuminate these issues. An additional limitation is that most of these mutational scanning studies probe only the immediate sequence neighborhood of one or a few proteins. Advances in the throughput and speed of these assay techniques—and improvements in computational prediction of protein structure and function—may soon allow us to characterize larger fractions of sequence space. New computational approaches to predicting protein folds may also help to address a similar question about protein folds: how extensive are the sequence networks encoding individual folds, and how far are their borders from those of other folds?

Finally, there are other complex protein features, the evolutionary origins of which we would like to understand mechanistically. Historical and directed evolution studies have begun to identify relatively simple means by which catalysis can be acquired de novo.[154] There are many other protein features that may also be tractable to evolutionary and biochemical analysis, including how channels and pores evolved the capacity to pass substrates through membranes and how proteins evolved fluorescent and light-sensing capacities. Particularly exciting would be to reveal how multimeric molecular machines acquire their capacity to transform chemical energy into organized kinetic work. In each case, if a phylogenetic approach can identify intervals in which a property of interest first emerged, an historical biochemical analysis, together with DMS, may be able to identify the causal genetic and biophysical changes that mediated these events.

Tracing the origins of complexity is challenging, but it is, in part and at least for some protein systems, experimentally tractable. Studying complexity at this most fundamental level of biological organization may reveal principles that pertain at higher levels, such as the cell, tissue, organism, society, and ecosystem. Also, higher levels of biology are themselves built upon molecular innovations; understanding how and why proteins came to have their present-day properties is therefore a key step in building a mechanistic account of the evolution of the more visible forms of biological complexity. Given the pace of progress in recent years, there is reason to hope that, by working together, biochemists and evolutionary biologists can advance our understanding of one of biology's great and enduring sources of wonder.

AUTHOR CONTRIBUTIONS
Arvind S. Pillai: Conceptualization (equal); investigation (equal); writing – original draft (equal); writing – review and editing (equal). **Georg K.A. Hochberg:** Conceptualization (equal); investigation (equal); writing – original draft (equal); writing – review and editing

DATA AVAILABILITY STATEMENT

No primary data were generated for this review.

ORCID

Arvind S. Pillai https://orcid.org/0000-0002-5012-1199
Joseph W. Thornton https://orcid.org/0000-0001-9589-6994

ENDNOTE

* For a protein that can homodimerize, the fraction of molecules in the bound state is given by the second-order equation:

$$f = \frac{2}{c * e^{\left(-\frac{\Delta G}{RT}\right)}} * \left(\frac{-e^{\left(-\frac{\Delta G}{RT}\right)} + \sqrt{e^{\left(-\frac{2\Delta G}{RT}\right)} + 8c * e^{\left(-\frac{\Delta G}{RT}\right)}}}{4} \right)^2,$$

where ΔG is the difference between the Gibbs free energies of the bound and unbound states, R is the universal gas constant, c is the concentration, and T is the temperature.

REFERENCES

1. McShea DW. Complexity and evolution: What everybody knows. Biol Philos. 1991;6(3):303–324. https://doi.org/10.1007/BF00132234.
2. Dawkins R. Climbing mount improbable. New York, NY: WW Norton & Company, 1997.
3. Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray, 2009;p. 204–208 1859.
4. Bonner JT. The evolution of complexity by means of natural selection. Princeton, NJ: Princeton University Press, 1988.
5. Jones KE, Angielczyk KD, Pierce SE. Stepwise shifts underlie evolutionary trends in morphological complexity of the mammalian vertebral column. Nat Commun. 2019;10(1):5071. https://doi.org/10.1038/s41467-019-13026-3.
6. Shultz S, Opie C, Atkinson QD. Stepwise evolution of stable sociality in primates. Nature. 2011;479(7372):219–222. https://doi.org/10.1038/nature10601.
7. Suzuki TK, Tomita S, Sezutsu H. Gradual and contingent evolutionary emergence of leaf mimicry in butterfly wing patterns. BMC Evol Biol. 2014;14(1):1–13. https://doi.org/10.1186/s12862-014-0229-5.
8. Lamb TD, Collin SP, Pugh EN Jr. Evolution of the vertebrate eye: Epsins, photoreceptors, retina and eye cup. Nat Rev Neurosci. 2007;8(12):960–976. https://doi.org/10.1038/jid.2014.371.
9. Nilsson DE. Eye evolution and its functional basis. Vis Neurosci. 2013;30(1–2):5–20. https://doi.org/10.1017/S0952523813000035.
10. Thornton JM, Orengo CA, Todd AE, Pearl FMG. Protein folds, functions and evolution. J Mol Biol. 1999;293(2):333–342. https://doi.org/10.1006/jmbi.1999.3054.
11. Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct. 2000;29(1):105–153.
12. Rivalta I, Sultan MM, Lee N-S, Manley GA, Loria JP, Batista VS. Allosteric pathways in imidazole glycerol phosphate synthase. Proc Natl Acad Sci. 2012;109(22):E1428–E1436. https://doi.org/10.1073/pnas.1120536109.
13. Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol. 2003;10(1):59–69. https://doi.org/10.1038/nsb881.
14. Marsh JA, Hernández H, Hall Z, et al. Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell. 2013;153(2):461–470. https://doi.org/10.1016/j.cell.2013.02.044.
15. Ahnert SE, Marsh JA, Hernández H, Robinson CV, Teichmann SA. Principles of assembly reveal a periodic table of protein complexes. Science. 2015;350(6266):aaa2245. https://doi.org/10.1126/science.aaa2245.
16. Motlagh NH, Wrabi JO, Li J, Hilser VJ. The ensemble nature of allostery. Nature. 2014;508(7496):331–339. https://doi.org/10.1038/nature13001.The.
17. Carroll SB. The making of the fittest: DNA and the ultimate forensic record of evolution. New York, NY: WW Norton & Company, 2006.
18. Pál C, Papp B. Evolution of complex adaptations in molecular systems. Nat Ecol Evol. 2017;1(8):1084–1092. https://doi.org/10.1038/s41559-017-0228-1.
19. Stoltzfus A. Constructive neutral evolution: Exploring evolutionary theory's curious disconnect. Biol Direct. 2012;7:1–13. https://doi.org/10.1186/1745-6150-7-35.
20. Muñoz-Gómez SA, Bilolikar G, Wideman JG, Geiler-Samerotte K. Constructive neutral evolution 20 years later. J Mol Evol. 2021;89(3):172–182. https://doi.org/10.1007/s00239-021-09996-y.
21. Gray MW, Lukeš J, Archibald JM, Keeling PJ, Doolittle WF. Irremediable complexity? Science. 2010;330(6006):920–921. https://doi.org/10.1126/science.1198594.
22. Fastrez J. Engineering allosteric regulation into biological catalysts. Chembiochem. 2009;10(18):2824–2835. https://doi.org/10.1002/cbic.200900590.
23. Grueninger D, Treiber N, Ziegler MOP, Koettr JWA, Schulze M-S, Schulz GE. Designed protein-protein association. Science. 2008;319(January):206–210.
24. Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED. Proteins evolve on the edge of supramolecular self-assembly. Nature. 2017;548(7666):244–247. https://doi.org/10.1038/nature23320.
25. Fowler DM, Fields S. Deep mutational scanning: A new style of protein science. Nat Methods. 2014;11(8):801–807. https://doi.org/10.1038/nmeth.3027.

26. Sarkisyan KS, Bolotin DA, Meer MV, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016; 533(7603):397–401. https://doi.org/10.1038/nature17995.

27. Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A. 2011; 108(19):7896–7901. https://doi.org/10.1073/pnas.1016024108.

28. Harms MJ, Thornton JW. Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol. 2010;20(3):360–366. https://doi.org/10.1016/j.sbi.2010.03.005.

29. Thornton JW, Need E, Crews D. Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. Science. 2003;301(5640):1714–1717. https://doi.org/10.1126/science.1086185.

30. Pillai AS, Chandler SA, Liu Y, et al. Origin of complexity in haemoglobin evolution. Nature. 2020;581(7809):480–485. https://doi.org/10.1038/s41586-020-2292-y.

31. Chen CS, Smits C, Dodson GG, et al. How to change the oligomeric state of a circular protein assembly: Switch from 11-subunit to 12-subunit TRAP suggests a general mechanism. PLoS One. 2011;6(10):e25296. https://doi.org/10.1371/journal.pone.0025296.

32. Jee JG, Byeon IJL, Louis JM, Gronenborn AM. The point mutation A34F causes dimerization of GB1. Proteins Struct Funct Genet. 2008;71(3):1420–1431. https://doi.org/10.1002/prot.21831.

33. Liu S, Liu S, Zhu X, et al. Nonnatural protein-protein interaction-pair design by key residues grafting. Proc Natl Acad Sci U S A. 2007;104(13):5330–5335. https://doi.org/10.1073/pnas.0606198104.

34. Anderson DP, Whitney DS, Hanson-Smith V, et al. Evolution of an ancient protein function involved in organized multicellularity in animals. Elife. 2016;5:1–20. https://doi.org/10.7554/eLife.10147.

35. Guntas G, Purbeck C, Kuhlman B. Engineering a protein-protein interface using a computationally designed library. Proc Natl Acad Sci U S A. 2010;107(45):19296–19301. https://doi.org/10.1073/pnas.1006528107.

36. Pauling L, Corey RB. Configurations of polypeptide chains with favored orientations around single bonds. Proc Natl Acad Sci U S A. 1951;37(11):729–740. https://doi.org/10.1073/pnas.37.11.729.

37. Deckert K, Budiardjo SJ, Brunner LC, Lovell S, Karanicolas J. Designing allosteric control into enzymes by chemical rescue of structure. J Am Chem Soc. 2012;134(24):10055–10060. https://doi.org/10.1021/ja301409g.

38. Mu X, Choi S, Lang L, et al. Physicochemical code for quinary protein interactions in Escherichia coli. Proc Natl Acad Sci U S A. 2017;114(23):E4556–E4563. https://doi.org/10.1073/pnas.1621227114.

39. Fukasawa Y, Tomii K. Accurate classification of biological and non-biological interfaces in protein crystal structures using subtle covariation signals. Sci Rep. 2019;9(1):1–12. https://doi.org/10.1038/s41598-019-48913-8.

40. Pauling L, Itano HA, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. Science. 1949;110(2865):543–548. https://doi.org/10.1126/science.286.5444.1488.

41. Gronenborn AM, Frank MK, Clore GM. Core mutants of the immunoglobulin binding domain of streptococcal protein G: Stability and structural integrity. FEBS Lett. 1996;398(2–3): 312–316. https://doi.org/10.1016/S0014-5793(96)01262-8.

42. Frank MK, Dyda F, Dobrodumov A, Gronenborn AM. Core mutations switch monomeric protein GB1 into an intertwined tetramer. Nat Struct Biol. 2002;9(11):877–885. https://doi.org/10.1038/nsb854.

43. Byeon IJL, Louis JM, Gronenborn AM. A protein contortionist: Core mutations of GB1 that induce dimerization and domain swapping. J Mol Biol. 2003;333(1):141–152. https://doi.org/10.1016/S0022-2836(03)00928-8.

44. Fraser NJ, Liu JW, Mabbitt PD, et al. Evolution of protein quaternary structure in response to selective pressure for increased thermostability. J Mol Biol. 2016;428(11):2359–2371. https://doi.org/10.1016/j.jmb.2016.03.014.

45. Kuriyan J, Eisenberg D. The origin of protein interactions and allostery in colocalization. Nature. 2007;450(7172):983–990. https://doi.org/10.1038/nature06524.

46. Fersht AR, Shi JP, Knill-Jones J, et al. Hydrogen bonding and biological specificity analysed by protein engineering. Nature. 1985;314(6008):235–238. https://doi.org/10.1038/314235a0.

47. von Hippel P, Berg O. On the specificity of DNA-protein interactions. Proc Natl Acad Sci U S A. 1985;83:1608–1612. https://doi.org/10.1272/jnms1923.19.1032.

48. Li Y, Huang Y, Swaminathan CP, Smith-Gill SJ, Mariuzza RA. Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. Structure. 2005;13(2):297–307. https://doi.org/10.1016/j.str.2004.12.012.

49. Vallone B, Miele AE, Vecchini P, Chiancone E, Brunori M. Free energy of burying hydrophobic residues in the interface between protein subunits. Proc Natl Acad Sci U S A. 1998; 95(11):6103–6107. https://doi.org/10.1073/pnas.95.11.6103.

50. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998;280(1):1–9. https://doi.org/10.1006/jmbi.1998.1843.

51. Monod J, Wyman J, Changeux JP. On the nature of allosteric transitions: A plausible model. J Mol Biol. 1965;12(1):88–118.

52. Levy ED. A simple definition of s tructural regions in proteins and its use in analyzing interface evolution. J Mol Biol. 2010; 403(4):660–670. https://doi.org/10.1016/j.jmb.2010.09.028.

53. Munshi S, Subramanian S, Ramesh S, et al. Engineering order and cooperativity in a disordered protein. Biochemistry. 2019; 58(19):2389–2397. https://doi.org/10.1021/acs.biochem.9b00182.

54. Malleshappa Gowder S, Chatterjee J, Chaudhuri T, Paul K. Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. Sci World J. 2014;2014:1–7. https://doi.org/10.1155/2014/971258.

55. Luo J, Liu Z, Guo Y, Li M. A structural dissection of large protein-protein crystal packing contacts. Nat Sci Rep. 2015; 5(1):1–13. https://doi.org/10.1038/srep14214.

56. Heinz DW, Baase WA, Dahlquist FW, Matthews BW. How amino-acid insertions are allowed in an α-helix of T4 lysozyme. Nature. 1993;361(6412):561–564. https://doi.org/10.1038/361561a0.

57. Heinz DW, Matthews BW. Rapid crystallization of t4 lysozyme by intermolecular disulfide cross-linking. Protein Eng Des Sel. 1994;7(3):301–307. https://doi.org/10.1093/protein/7.3.301.

58. Xia Y, Diprimio N, Keppel TR, et al. The designability of protein switches by chemical rescue of structure: Mechanisms of inactivation and reactivation. J Am Chem Soc. 2013;135(50): 18840–18849. https://doi.org/10.1021/ja407644b.

59. Mathieu V, Fastrez J, Soumillion P. Engineering allosteric regulation into the hinge region of a circularly permuted TEM-1

-lactamase. Protein Eng Des Sel. 2010;23(9):699–709. https://doi.org/10.1093/protein/gzq041.

60. Ke W, Laurent AH, Armstrong MD, et al. Structure of an engineered β-lactamase maltose binding protein fusion protein: Insights into heterotropic allosteric regulation. PLoS One. 2012;7(6):e39168. https://doi.org/10.1371/journal.pone.0039168.

61. Stebbins JW, Kantrowitz ER. Conversion of the noncooperative bacillus subtilis aspartate Transcarbamoylase into a cooperative enzyme by a single amino acid substitution. Biochemistry. 1992;31(8):2328–2332. https://doi.org/10.1021/bi00123a017.

62. Scrutton NS, Deonarain MP, Berry A, Perham RM. Cooperativity induced by a single mutation at the subunit interface of a dimeric enzyme: Glutathione reductase. Science. 1992;258(5085):1140–1143.

63. Bergendahl LT, Marsh JA. Functional determinants of protein assembly into homomeric complexes. Sci Rep. 2017;7(1):4932. https://doi.org/10.1038/s41598-017-05084-8.

64. Guntas G, Mitchell SF, Ostermeier M. A molecular switch created by in vitro recombination of nonhomologous genes. Chem Biol. 2004;11:1483–1487. https://doi.org/10.1016/j.

65. Ikeda Y, Tanaka T, Noguchi T. Conversion of non-allosteric pyruvate kinase isozyme into an allosteric enzyme by a single amino acid substitution. J Biol Chem. 1997;272(33):20495–20501. https://doi.org/10.1074/jbc.272.33.20495.

66. Kuo L, Zambidis I, Caron C. Triggering of allostery in an enzyme by a point mutation: Ornithine transcarbamoylase. Science. 1989;245(4917):522–524.

67. Lynch VJ, May G, Wagner GP. Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. Nature. 2011;480(7377):383–386. https://doi.org/10.1038/nature10595.

68. Schupfner M, Straub K, Busch F, Merkl R, Sterner R. Analysis of allosteric communication in a multienzyme complex by ancestral sequence reconstruction. Proc Natl Acad Sci U S A. 2020;117(1):346–354. https://doi.org/10.1073/pnas.1912132117.

69. Kohl A, Amstutz P, Parizek P, et al. Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein. Structure. 2005;13(8):1131–1141. https://doi.org/10.1016/j.str.2005.04.020.

70. Brennan C, Christianson K, Surowy T, Mandecki W. Modulation of enzyme activity by antibody binding to an alkaline phosphatase-epitope hybrid protein. Protein Eng Des Sel. 1994;7(4):509–514. https://doi.org/10.1093/protein/7.4.509.

71. Legendre D, Soumillion P, Fastrez J. Engineering a regulatable enzyme for homogeneous immunoassays. Nat Biotechnol. 1999;17(1):67–72. https://doi.org/10.1038/5243.

72. Santamaría B, Estévez AM, Martínez-Costa OH, Aragón JJ. Creation of an allosteric phosphofructokinase starting with a nonallosteric enzyme: The case of Dictyostelium discoideum phosphofructokinase. J Biol Chem. 2002;277(2):1210–1216. https://doi.org/10.1074/jbc.M109480200.

73. Hart KM, Moeder KE, Ho CMW, Zimmerman MI, Frederick TE, Bowman GR. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. PLoS One. 2017;12(6):1–13. https://doi.org/10.1371/journal.pone.0178678.

74. Lu S, Ji M, Ni D, Zhang J. Discovery of hidden allosteric sites as novel targets for allosteric drug design. Drug Discov Today. 2018;23(2):359–365. https://doi.org/10.1016/j.drudis.2017.10.001.

75. Perutz MF, Muirhead H, Cox JM, Goaman LC. Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 a resolution: The atomic model. Nature. 1968;219(5150):131–139. https://doi.org/10.1038/219131a0.

76. Mizutani Y, Kitagawa T. Ultrafast dynamics of myoglobin probed by time-resolved resonance. Chem Rec. 2001;1(3):258–275.

77. Coyle SM, Flores J, Lim WA. Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. Cell. 2013;154(4):875–887. https://doi.org/10.1016/j.cell.2013.07.019.

78. Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins Struct Funct Genet. 2004;57(3):433–443. https://doi.org/10.1002/prot.20232.

79. Damry AM, Mayer MM, Broom A, Goto NK, Chica RA. Origin of conformational dynamics in a globular protein. Commun Biol. 2019;2(1):1–10. https://doi.org/10.1038/s42003-019-0681-2.

80. Chen Y, Hu D, Vorpagel ER, Lu HP. Probing single-molecule T4 lysozyme conformational dynamics by intramolecular fluorescence energy transfer. J Phys Chem B. 2003;107(31):7947–7956. https://doi.org/10.1021/jp022406z.

81. Wong KB, Daggett V. Barstar has a highly dynamic hydrophobic core: Evidence from molecular dynamics simulations and nuclear magnetic resonance relaxation data. Biochemistry. 1998;37(32):11182–11192. https://doi.org/10.1021/bi980552i.

82. James LC, Tawfik DS. Conformational diversity and protein evolution—A 60-year-old hypothesis revisited. Trends Biochem Sci. 2003;28(7):361–368. https://doi.org/10.1016/S0968-0004(03)00135-X.

83. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. Science. 2009;324(5924):203–207.

84. Baldwin AJ, Kay LE. NMR spectroscopy brings invisible protein states into focus. Nat Chem Biol. 2009;5(11):808–814. https://doi.org/10.1038/nchembio.238.

85. Telmer PG, Shilton BH. Insights into the conformational equilibria of maltose-binding protein by analysis of high affinity mutants. J Biol Chem. 2003;278(36):34555–34567. https://doi.org/10.1074/jbc.M301004200.

86. Sharff AJ, Rodseth LE, Spurlino JC, Quiocho FA. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the Maltodextrin binding protein involved in active transport and chemotaxis. Biochemistry. 1992;31(44):10657–10663. https://doi.org/10.1021/bi00159a003.

87. Pearlman SM, Serber Z, Ferrell JE. A mechanism for the evolution of phosphorylation sites. Cell. 2011;147(4):934–946. https://doi.org/10.1073/pnas.1318754110.

88. Daily MD, Gray JJ. Local motions in a benchmark of allosteric proteins. Proteins. 2007;67(2):385–399.

89. Kovermann M, Grundström C, Elisabeth Sauer-Eriksson A, Sauer UH, Wolf-Watz M. Structural basis for ligand binding to an enzyme by a conformational selection pathway. Proc Natl Acad Sci U S A. 2017;114(24):6298–6303. https://doi.org/10.1073/pnas.1700919114.

90. Eisenmesser EZ, Millet O, Labeikovsky W, et al. Intrinsic dynamics of an enzyme underlies catalysis. Nature. 2005; 438(7064):117–121. https://doi.org/10.1038/nature04105.

91. Clarkson MW, Gilmore SA, Edgell MH, Lee AL. Dynamic coupling and allosteric behavior in a nonallosteric protein. Biochemistry. 2006;45(25):7693–7699. https://doi.org/10.1021/bi060652l.

92. Boehr DD, Schnell JR, McElheny D, et al. A distal mutation perturbs dynamic amino acid networks in dihydrofolate reductase. Biochemistry. 2013;52(27):4605–4619. https://doi.org/10.1021/bi400563c.

93. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604(7904): 175–183. https://doi.org/10.1038/s41586-022-04586-4.

94. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins Struct Funct Genet. 1999;35(4):408–414. https://doi.org/10.1002/(SICI)1097-0134(19990601)35:4<408::AID-PROT4>3.0.CO;2-A.

95. Cordes MHJ, Burton RE, Walsh NP, McKnight CJ, Sauer RT. An evolutionary bridge to a new protein fold. Nat Struct Biol. 2000;7(12):1129–1132. https://doi.org/10.1038/81985.

96. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci U S A. 2009;106(50):21149–21154. https://doi.org/10.1073/pnas.0906408106.

97. Meier S, Jensen PR, David CN, et al. Continuous molecular evolution of protein-domain structures by single amino acid changes. Curr Biol. 2007;17(2):173–178. https://doi.org/10.1016/j.cub.2006.10.063.

98. Porter LL, Looger LL. Extant fold-switching proteins are widespread. Proc Natl Acad Sci U S A. 2018;115(23):5968–5973. https://doi.org/10.1073/pnas.1800168115.

99. Chang Y, Cohen SE, Phong C, et al. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. 2016;349(6245):324–328. https://doi.org/10.1126/science.1260031.A.

100. Burmann BM, Knauer SH, Sevostyanova A, et al. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. Cell. 2012;150(2):291–303. https://doi.org/10.1016/j.cell.2012.05.042.

101. Farías-Rico JA, Schmidt S, Höcker B. Evolutionary relationship of two ancient protein superfolds. Nat Chem Biol. 2014; 10(9):710–715. https://doi.org/10.1038/nchembio.1579.

102. Jiang H, Blouin C. Insertions and the emergence of novel protein structure: A structure-based phylogenetic study of insertions. BMC Bioinformatics. 2007;8:1–14. https://doi.org/10.1186/1471-2105-8-444.

103. Hochberg GKA, Liu Y, Marklund EG, Metzger BPH, Laganowsky A, Thornton JW. A hydrophobic ratchet entrenches molecular complexes. Nature. 2020;588(7838):503–508. https://doi.org/10.1038/s41586-020-3021-2.

104. Tretyachenko V, Vymětal J, Bednárová L, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci Rep. 2017;7(1):2–10. https://doi.org/10.1038/s41598-017-15635-8.

105. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225(5232):563–564.

106. Wheeler LC, Harms MJ. Were ancestral proteins less specific? Mol Biol Evol. 2021;38(6):2227–2239. https://doi.org/10.1093/molbev/msab019.

107. McClune CJ, Alvarez-Buylla A, Voigt CA, Laub MT. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. Nature. 2019;574(7780):702–706. https://doi.org/10.1038/s41586-019-1639-8.

108. Alalouf O, Salama R, Tal O, et al. Divergent interactions maintain the quaternary octameric structure of a new family of esterases. bioRxiv. 2018;466904:1–29. https://doi.org/10.1101/466904.

109. Jack BR, Meyer AG, Echave J, Wilke CO. Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biol. 2016;14(5):1–23. https://doi.org/10.1371/journal.pbio.1002452.

110. Honzatko RB, Hendrickson WA. Molecular models for the putative dimer of sea lamprey hemoglobin. Proc Natl Acad Sci U S A. 1986;83(22):8487–8491. https://doi.org/10.1073/pnas.83.22.8487.

111. Makino M, Sugimoto H, Sawai H, Kawada N, Yoshizato K, Shiro Y. High-resolution structure of human cytoglobin: Identification of extra N- and C-termini and a new dimerization mode. Acta Crystallogr Sect D Biol Crystallogr. 2006;62(6): 671–677. https://doi.org/10.1107/S0907444906013813.

112. Royer WE, Strand K, Van Heel M, Hendrickson WA. Structural hierarchy in erythrocruorin, the giant respiratory assemblage of annelids. Proc Natl Acad Sci U S A. 2000;97(13): 7107–7111. https://doi.org/10.1073/pnas.97.13.7107.

113. Heaslet HA, Royer WE. The 2.7 Å crystal structure of deoxygenated hemoglobin from the sea lamprey (Petromyzon marinus): Structural basis for a lowered oxygen affinity and Bohr effect. Structure. 1999;7(5):517–526. https://doi.org/10.1016/S0969-2126(99)80068-9.

114. Leander M, Yuan Y, Meger A, Cui Q, Raman S. Functional plasticity and evolutionary adaptation of allosteric regulation. Proc Natl Acad Sci U S A. 2020;117(41):25445–25454. https://doi.org/10.1073/pnas.2002613117.

115. Balasubramanian A, Ponnuraj K. Crystal structure of the first plant urease from Jack bean: 83 years of journey from its first crystal to molecular structure. J Mol Biol. 2010;400(3):274–283. https://doi.org/10.1016/j.jmb.2010.05.009.

116. Ha NC, Oh ST, Sung JY, Cha KA, Lee MH, Oh BH. Supramolecular assembly and acid resistance of helicobacter pylori urease. Nat Struct Biol. 2001;8(6):505–509. https://doi.org/10.1038/88563.

117. Santhanagopalan I, Degiacomi MT, Shepherd DA, Hochberg GKA, Benesch JLP, Vierling E. It takes a dimer to tango: Oligomeric small heat shock proteins dissociate to capture substrate. J Biol Chem. 2018;293(51):19511–19521. https://doi.org/10.1074/jbc.RA118.005421.

118. Dey S, Ritchie DW, Levy ED. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nat Methods. 2018;15(1):67–72. https://doi.org/10.1038/nmeth.4510.

119. Lewis M, Chang G, Horton NC, et al. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science. 1996;271(5253):1247–1254.

120. Schumacher MA, Choi KY, Lu F, Zalkin H, Brennan RG. Mechanism of corepressor-mediated specific DNA binding by the purine repressor. Cell. 1995;83(1):147–155. https://doi.org/10.1016/0092-8674(95)90243-0.

121. Swint-Kruse L, Matthews KS. Allostery in the LacI/GalR family: Variations on a theme. Curr Opin Microbiol. 2009;12(2): 129–137. https://doi.org/10.1016/j.mib.2009.01.009.

122. Laine JM, Amat M, Morgan BR, Royer WE, Massi F. Insight into the allosteric mechanism of Scapharca dimeric hemoglobin. Biochemistry. 2014;53(46):7199–7210. https://doi.org/10.1021/bi500591s.

123. Riggs AF. Self-association, cooperativity and supercooperativity of oxygen binding by hemoglobins. J Exp Biol. 1998;201(8):1073–1084. https://doi.org/10.1242/jeb.201.8.1073.

124. Royer WE, Zhu H, Gorr TA, Flores JF, Knapp JE. Allosteric hemoglobin assembly: Diversity and similarity. J Biol Chem. 2005;280(30):27477–27480. https://doi.org/10.1074/jbc.R500006200.

125. Tack D, Tonner P, Pressman A, et al. The genotype-phenotype landscape of an allosteric protein. Mol Syst Biol. 2021;17(3):e10179. https://doi.org/10.1101/2020.07.10.197574.

126. Reynolds KA, McLaughlin RN, Ranganathan R. Hot spots for allosteric regulation on protein surfaces. Cell. 2011;147(7):1564–1575. https://doi.org/10.1016/j.cell.2011.10.049.

127. Bashford D, Chothia C, Lesk AM. Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol. 1987;196(1):199–216. https://doi.org/10.1016/0022-2836(87)90521-3.

128. Jawad Z, Paoli M. Novel sequences propel familiar folds. Structure. 2002;10(4):447–454. https://doi.org/10.1016/S0969-2126(02)00750-5.

129. Axe DD, Foster NW, Fersht AR. Active barnase variants with completely random hydrophobic cores. Proc Natl Acad Sci U S A. 1996;93(11):5590–5594. https://doi.org/10.1073/pnas.93.11.5590.

130. Lynch M. The evolution of multimeric protein assemblages. Mol Biol Evol. 2012;29(5):1353–1366. https://doi.org/10.1093/molbev/msr300.

131. Lynch M. Evolutionary diversification of the multimeric states of proteins. Proc Natl Acad Sci. 2013;110(30):E2821–E2828. https://doi.org/10.1073/pnas.1310980110.

132. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press, 1983.

133. Betancourt AJ, Presgraves DC. Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci U S A. 2002;99(21):13616–13620. https://doi.org/10.1073/pnas.212277199.

134. Kimura M. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. Proc Natl Acad Sci U S A. 1980;77(1):522–526. https://doi.org/10.1073/pnas.77.1.522.

135. KIMURA M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47(391):713–719.

136. Zabel WJ, Hagner KP, Livesey BJ, et al. Evolution of protein interfaces in multimers and fibrils. J Chem Phys. 2019;150(22):225102. https://doi.org/10.1063/1.5086042.

137. Charlesworth B. How long does it take to fix a favorable mutation, and why should we care? Am Nat. 2020;195(5):753–771. https://doi.org/10.1086/708187.

138. Iwasa Y, Michor F, Nowak MA. Stochastic tunnels in evolutionary dynamics. Genetics. 2004;166(3):1571–1579. https://doi.org/10.1534/genetics.166.3.1571.

139. Weissman DB, Desai MM, Fisher DS, Feldman MW. The rate at which asexual populations cross fitness valleys. Theor Popul Biol. 2009;75(4):286–300. https://doi.org/10.1016/j.tpb.2009.02.006.

140. Komarova NL, Sengupta A, Nowak MA. Mutation-selection networks of cancer initiation: Tumor suppressor genes and chromosomal instability. J Theor Biol. 2003;223(4):433–450. https://doi.org/10.1016/S0022-5193(03)00120-6.

141. Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. Evolution of increased complexity in a molecular machine. Nature. 2012;481(7381):360–364. https://doi.org/10.1038/nature10724.

142. Archibald JM, Logsdon JM, Doolittle WF. Recurrent paralogy in the evolution of archaeal chaperonins. Curr Biol. 1999;9(18):1053–1056. https://doi.org/10.1016/S0960-9822(99)80457-6.

143. Bernstein HD, Zopf D, Freymann DM, Walter P. Functional substitution of the signal recognition particle 54-kDa subunit by its Escherichia coli homolog. Proc Natl Acad Sci U S A. 1993;90(11):5229–5233. https://doi.org/10.1073/pnas.90.11.5229.

144. Bridgham JT, Keay J, Ortlund EA, Thornton JW. Vestigialization of an allosteric switch: Genetic and structural mechanisms for the evolution of constitutive activity in a steroid hormone receptor. PLoS Genet. 2014;10(1):e1004058. https://doi.org/10.1371/journal.pgen.1004058.

145. Natarajan C, Signore AV, Bautista NM, et al. Evolution and molecular basis of a novel allosteric property of crocodilian hemoglobin. bioRxiv. 2022. https://doi.org/10.1101/2022.07.18.500494.

146. Bridgham JT, Carroll SM, Thornton JW. Evolution of hormone-receptor complexity by molecular exploitation. Science. 2006;312(5770):97–101. https://doi.org/10.1126/science.1123348.

147. Schulz L, Sendker FL, Hochberg GKA. Non-adaptive complexity and biochemical function. Curr Opin Struct Biol. 2022;73:102339. https://doi.org/10.1016/j.sbi.2022.102339.

148. Emlaw JR, Tessier CJG, McCluskey GD, et al. A single historical substitution drives an increase in acetylcholine receptor complexity. Proc Natl Acad Sci U S A. 2021;118(7):e2018731118. https://doi.org/10.1073/pnas.2018731118.

149. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerate mutations. Genetics. 1999;151(4):1531–1545. https://doi.org/10.1093/genetics/151.4.1531.

150. Liu R, Ochman H. Stepwise formation of the bacterial flagellar system. Proc Natl Acad Sci U S A. 2007;104(17):7116–7121.

151. Bunn F. Subunit assembly of hemoglobin: Of hematologic an important phenotype determinant. J Am Soc Hematol. 1987;2019(1):1–6.

152. Xie VC, Pu J, Metzger BPH, Thornton JW, Dickinson BC. Contingency and chance erase necessity in the experimental evolution of ancestral proteins. Elife. 2021;10:1–87. https://doi.org/10.7554/eLife.67336.

153. Starr TN, Picton LK, Thornton JW. Alternative evolutionary histories in the sequence space of an ancient protein. Nature. 2017;549(7672):409–413. https://doi.org/10.1038/nature23902.

154. Clifton BE, Kaczmarski JA, Carr PD, Gerth ML, Tokuriki N, Jackson CJ. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein article. Nat Chem Biol. 2018;14(6):542–547. https://doi.org/10.1038/s41589-018-0043-2.

155. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. Conversion of monomeric protein L to an obligate dimer by computational protein design. Proc Natl Acad Sci U S A. 2001;98(19):10687–10691. https://doi.org/10.1073/pnas.181354398

156. Pincus D, Pandey JP, Feder ZA, Creixell P, Resnekov O, Reynolds KA. Engineering allosteric regulation in protein kinases. Sci Signal. 2018;11(555):eaar3250.