THE UNIVERSITY OF CHICAGO

EVOLUTIONARY ORIGINS OF MOLECULAR COMPLEXITY IN HEMOGLOBIN

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY

ARVIND SREEKUMAR PILLAI

CHICAGO, ILLINOIS

MARCH 2021

Table of Contents

Table of Figures

Chapter 1: Introduction

In this dissertation, I trace the evolutionary origin of molecular assembly in Hemoglobin (Hb), a model protein complex, by functionally and structurally characterizing reconstructed ancestral proteins statistically inferred from a phylogeny of globin genes. I investigate how the interfaces that hold the complex together first arose during early vertebrate history, and how the elaboration of Hb's quaternary structure was tied to its essential physiological function of oxygen-transport. Specifically, I identify the numbers and effect-sizes of historical mutations that contributed to the evolution of Hb's heterotetrameric form (which is composed of paralogous α and β chains (Goodman 1981)) from a monomeric globin ancestor that showed no propensity to assemble into complexes. I measured the effects of these structure-altering mutations on Hb's key functional properties, particularly its cooperativity and its allosteric regulation by small-molecule phosphate effectors. Of key interest is whether or not these biochemically complicated features arose through many mutations of individually small effect (Fisher 1958), or via paths involving a few large-effect changes that tinker with pre-existing structural properties (Jacob 1977).

In this introduction, I briefly survey literature relevant to the study of how molecular complexes emerge during evolution. I also describe key aspects of the biology and biochemistry of hemoglobin, providing both justification for its choice as a model system and basic background for the experiments detailed in Chapter 2 and 3. A more extensive scholarly review of the implications of the work and its relationship to prior literature in the fields of population genetics, protein engineering and allostery is included in Chapter 4.

1.1 Evolution of Complexity

Biology is replete with examples of morphological(Serb et al. 2008), physiological (Berenbrink 2007), social (Kappeler et al. 2019) and molecular (Petrov et al. 2014) complexity. A great deal of work has focused on identifying the evolutionary causes of complexity and determining whether or not there has existed a general trend towards more of it during the evolutionary history of life on earth (McShea 1991). Despite enduring interest in the topic, little consensus has developed among biologists about what exactly constitutes "complexity" in living systems. Lewontin stated in 1998 that "We still await a definition of complexity that will distinguish between people and frogs and that can be employed in a rigorous theoretical scheme." (Lewontin 1998). For the purposes of this work, however, I adopt a simple definition of the term that acknowledges both the function and structure of a feature: complexity refers to the number of components in a system, as well as the specificity of the architecture they must form in order to function. Biologists have speculated about the causes and driving factors involved in the emergence of biological complexity since before Darwin – Natural theologians (and their modern intellectual descendants (Behe 1996)) attributed it to theistic design (Paley 1851), while Jean-Baptiste Lamarck offered a naturalistic explanation in which living organisms contain invisible fluids that drive their tissues, cells and organs inexorably toward higher complexity (McShea 1991). Darwin considered complex features like the eye to be the product of natural selection for improved function (Darwin 1859). He envisioned these morphological features being built through a gradual series of steps across geological time, where simpler forms were replaced sequentially by more complicated ones.

There are two ways in which biologists typically account for the origination of a complicated feature: one involves providing a description of the historical pathway of structural changes by which complexity arose in a system, while the latter is concerned with the population-genetic forces that drive its emergence. The former involves inferring a sequence of ancestral forms from which the modern-day structure arose, typically through comparisons between extant or fossil organisms (Serb et al. 2008). Such detailed accounts have been produced for the cephalopod eye, which evolved from a photosensitive skin patch via several intermediate morphological stages (eg. eyecup and pinhole camera-eye), versions of which are still represented in the anatomies of living molluscs (Serb et al. 2008).

Describing the genetic or morphological steps involved in building an organ is different from accounting for why this sequence of changes may have occurred in a lineage or population. The classical view is that complexity evolves because it is adaptive (Dawkins 1997; Darwin 1859); Complexity could be selected for because complex organs can perform new and useful functions that were unavailable to their simpler precursors. Perhaps the greater number of parts in a complex structure allows for better division of labor and functional efficiency (McShea 1991). The various earlier iterations of the eye in Cephalopod evolution may have fixed in the lineage because they granted increased light sensitivity and resolution relative to their ancestors (Serb et al. 2008).

In recent decades, Doolittle (Gray et al. 2010), Lynch (Lynch 2007) and Stolzfus (Stoltzfus 2012) have proposed that non-adaptive forces could also play a role in the origination and retention of complex features, particularly at the molecular level. A simple example of potentially neutral complexity in genomes is the profusion of partially-redundant duplicate genes. This could arise

from a process of neutral subfunctionalization (Stoltzfus 2012) – where, subsequent to gene duplication, the two descendant genes merely partition the multiple functions of the ancestor rather than expanding the total range of possible molecular functions in any way. This situation could arise because each gene copy incurs mutations that destroy one function and not the other – these mutations are tolerated because the function lost in one is preserved in the redundant, paralogous copy. Eventually, this leads to the irreversible retention of two genes that are collectively no better than their single common ancestor. Instead of selection, proponents of constructive neutral evolution emphasize (1) biased mutational variation that drives systems toward higher complexity and (2) irreversible entrenchment of traits due to epistasis as key factors in the evolution of complexity (Stoltzfus 2012).

In general, there is dearth of experimentally detailed work on how complexity emerges at any level of biological organization. This is likely because in many cases it is difficult, if not impossible, to reconstruct and manipulate the simpler ancestral forms from which modern day complex features arose (although proxies are sometimes used (Ratcliff et al. 2012)), or alternatively, to induce the evolution of novel forms of complexity in an experimental evolution setup. However, one arena in which the evolution of complexity can be explored in great empirical detail is in the evolution of protein complexes. Most proteins associate into higher-order complexes that are held together by non-covalent interactions (Marsh and Teichmann 2015). Almost every process in the cell depends on the assembly and activity of multimeric protein assemblies. Many of these are compositionally complex, with multiple structurally and functionally differentiated components (Rohl and Nierhaus 1982; Liu et al. 2007). Such "molecular machines" represent the simplest systems in which we can biochemically and

genetically dissect the evolution of complexity. Despite the biological significance and widespread nature of multimerization across proteins, there is currently no structurally or genetically detailed account of how it originated in any protein family. How were the subunits that constitute multimers brought into stable and specific association with each other over evolutionary time? Explaining how multi-subunit protein complexes and the structural interfaces that mediate their assembly came to be could offer a mechanistic window into the origin of molecular complexity.

1.2 Evolution of protein complexes

Previous work on the evolution of complexes has involved a wide array of computational and experimental approaches – here, I point to three broad approaches that were influential for the work contained in this dissertation. First, phylogenetic and comparative studies can give us a window into how the composition, size and function of protein complexes can change over time. The flagellum is a classical example of a complex molecular machine that is amenable to this sort of comparative bioinformatic analysis (Aizawa 2001). Structural comparisons show that the bacterium flagellum and the type-III secretion system share numerous homologies and are likely both derived from a common ancient apparatus (Aizawa 2001). The number of subunits involved in forming the flagellar complex are variable across species – over 50 subunits can be involved in some cases – but subunit content comparisons across species show that all flagella are derived from ancient core complex that is composed of 24 subunits (Liu et al. 2007). Further, many of these core subunits show evidence of sequence homology among them,

suggesting that they were derived by successive gene duplications from a simpler ancestor (Liu et al. 2007). Similarly, a phylogenetic profile of the subunits in the mitochondrial respiratory chain complex I showed that this 14-mer evolved from an ancestral bacterial 11-mer (Moparthi and Hägerhäll 2011). This kind of comparative work gives us valuable, if coarse-grained, information about the order and intervals in which subunits were added to specific complexes, revealing clearly that the subunit content of a molecular machine can grow over time. However, this approach cannot detail the precise genetic or evolutionary causes for those structural changes.

Another line of research applies bioinformatic tools not just to an individual protein complexes but to entire databases containing information about protein-protein interactions from different species, obtained either by X-ray crystallography (Ahnert et al. 2015) or high-throughput yeast two-hybrid experiments (Qian et al. 2011). These data can give us a global picture of how oligomers evolve and identify broad trends. For example, such large-scale analyses have revealed the rate at which new protein-protein interactions evolve in fungal genomes (Qian et al. 2011), identified the frequency with which paralogs tend to co-assemble or self-assemble (Hochberg et al. 2018), established the homomers containing even-numbered stoichiometries are more common than odd numbered ones (Marsh and Teichmann 2014), possibly because the former structures utilize closed symmetrical head-to-head interfaces that are less likely to multimerize indefinitely (M. Lynch 2013), shown that many heteromers are frequently derived by gene duplication from ancestral homomers (Mallik and Tawfik 2020) and that the order of subunit assembly in a complex is preserved by selection (Marsh et al. 2013).

Lastly, experimental protein engineering work can also give us a window into how new complexes and protein-protein interactions could be built through point mutations and indels (Grueninger et al. 2008; Chen et al. 2011). This work has previously demonstrated that a few substitutions, particularly to bulky hydrophobic groups, could be sufficient to deliver a novel oligomer with micro-molar affinity interfaces (Grueninger et al. 2008). Levy's work shows that similar tactics could be used to glue proteins into fibers via one or a few mutations (Garcia-Seisdedos et al. 2017). Although designing assembly via such mutational paths may be biophysically possible, it does not imply that natural evolution is likely to take similar routes in constructing new quaternary structures.

Collectively, this work has shown that proteins can evolve novel interactions through a variety of mechanisms – (1) by rewiring existing homomeric interactions to facilitate heteromeric ones (Mallik and Tawfik 2020; Pereira-Leal et al. 2007), (2) fusing to a protein-interaction domain (Basu et al. 2008) and (3) evolving a hydrophobic patch that mediates a stable and specific interaction (Levy 2010). Ultimately, the first two mechanisms merely involve the modification or retooling of interfaces that already exist, but the ultimate origin of an interaction must involve mutationally transforming an ancestrally solvent exposed surface into a protein binding one. Although it is now clear that such patches can be engineered on native proteins (Grueninger et al. 2008), it is unclear if the mechanisms involved in engineering them are quite the same as the ones utilized by natural evolution.

Aside from the mutational mechanisms by which proteins arise, there is the additional question of why they arise. Many researchers have proposed adaptive explanations for the origin of molecular assemblies. In many cases, the adaptive value of complex formation is relatively

obvious: for example, the assembly of actin and tubulin into fibers helps shape out the cytoskeleton (Fletcher and Mullins 2010). Interactions with other proteins can allow for allosteric regulation of a protein's function (Bergendahl and Marsh 2017). Catalytic function is in some cases, entirely dependent on protein assembly, particularly when the catalytic site lies at the interface. Oligomerization could also confer more subtle advantages in catalytic efficiency, translational efficiency or fidelity (Marsh and Teichmann 2015). Self-association could also arise in response to selection for thermostability (Fraser et al. 2016), potentially explaining why enzymes in extremophiles sometimes occupy higher oligomeric states than their mesophilic counterparts (Walden et al. 2001; Tanakai et al. 2004).

Recently, the possibility has emerged that many interactions are not simply the product of adaptive processes. Proponents of "constructive neutral evolution" (CNE) propose that such complexes could arise fortuitously in the cell, fix in populations neutrally and become structurally entrenched (Stoltzfus 2012; Hochberg et al. 2020). The apparent diversity of multimeric forms among enzymes with mostly conserved functions is perhaps evidence of this, since multimerization in these cases confers no obvious functional advantage (Lynch 2013). In some cases, replacement of a complicated complex with a simpler one does not lead to a deficit in function – one remarkable example of gratuitous molecular complexity of this sort is the 10-subunit eukaryotic yeast RNAse-P (Weber et al. 2014), which can be replaced with its monomeric counterpart in *Arabidopsis thaliana* without a major loss of fitness to the yeast. Hochberg et al. showed a mechanism by which functionally useless interfaces could persist in a lineage under drift, simply because mutations that compromise it would yield aggregation-prone proteins (Hochberg et al. 2020).

We do not currently have detailed information about the specific genetic path or evolutionary forces that led to the emergence of any particular protein complex. To explore this question, I identified a historical interval where a protein complex arose and determined the functional and structural effects of historical substitutions that created it. To do this, I relied on a statistical procedure called Ancestral Sequence Reconstruction (ASR) (Selberg et al. 2021) to resurrect ancestral protein sequences before and after the emergence of multimeric assembly in a protein lineage.

ASR is a method that allows us to interrogate the evolution of protein function along a specified phylogenetic interval in a tree (Selberg et al. 2021). ASR uses the topology and branch-lengths of a phylogenetic tree, a statistical model of sequence evolution (eg. an amino-acid substitution matrix) and an alignment of extant protein sequences to infer the most probable ancestral sequence that existed at every node in a phylogeny by maximum likelihood (Yang 2007). These hypothetical ancestral proteins can be recombinantly expressed in *Escherichia Coli* and their structures and functions characterized in the laboratory using standard biophysical and biochemical assays. This statistical approach allows us to the identify the effects of specific historical mutations in a tree. Ancestral sequence reconstruction has previously been deployed to study the evolution of a vast array of protein properties, from ligand binding specificity (Siddiq, Hochberg, and Thornton 2017) to fluorescence (Field and Matz 2010) to enzyme activity (Clifton et al. 2018). In the context of protein assembly, it was previously used to trace the evolution of a transmembrane multimer – the V0 ring of Fungal V-ATPase, which increased in subunit complexity on the lineage to modern Fungi (Finnigan et al. 2012). After Fungi

diverged from animals, this heteromeric 6 membered ring that ancestrally contained 5 subunits of one type and one of another, evolved to accommodate an additional paralogous subunit in a 4:1:1 ratio. The total number of subunits in the ring remained constant, but the diversity of subunits increased. The authors discovered that the increased complexity of the molecular machine did not confer enhanced function, consistent with the predictions of the CNE hypothesis (Stoltzfus 2012).

Prior to this thesis, ASR had been used to study the genetic elaboration of an existing complex but it had not yet been used to probe the origination of a protein complex "from scratch". To explore this evolutionary scenario, I identified a model clade of proteins that displayed a number of characteristics that would make it a useful model for studying how novel complexes are born: (1) Clear, experimentally-validated variation in oligomeric state between the members of the clade, (2) Evidence that the larger stoichiometry in the clade was derived, rather than the smaller oligomer being derived by the secondary loss of an interface in a lineage, as was reported in another ASR study (Perica et al. 2014), (3) The members of the protein family had to be highly alignable with few gaps and provide enough phylogenetic signal for a high quality reconstruction, and (4) the formation of the oligomer had to be biologically significant. Vertebrate hemoglobin proved to be an ideal system for exploring the evolution of a protein complex and its associated functions. In fact, the first paper to propose reconstructing ancestral proteins, written by Linus Pauling and Emile Zuckerkandl (Zuckerkandl 1965), actually explicitly identified Hb as a system in which it could be possible to resurrect and functionally characterize ancient sequences. In following section, I provide further background for my system of choice.

1.3 Hemoglobin as a model system for investigating the origin of complex features

Hemoglobin (Hb) is a heterotetrameric oxygen binding complex made up of 2 α and 2 β subunits, belonging to a broader family of hemoproteins that are distributed across all three domains of life (Storz, Opazo, and Hoffmann 2013). Hb's nearest paralog relatives among vertebrates, including myoglobin (Kendrew et al. 1960), cytoglobin (Lechauve et al. 2010) and Globin-E (Blank et al. 2011), are monomeric. This pattern of variation suggests that the tetrameric architecture of Hemoglobin is derived from an ancestral monomer. Vertebrate globin sequences evolve slowly enough to be highly alignable, and offer ample phylogenetic signal for inferring the evolutionary history of the protein family. The globin genes of vertebrates have been used as a model system by bioinformaticians to study the processes of gene duplication and divergence on deep timescales (Storz, Opazo, and Hoffmann 2013), as well as by researchers interested in molecular adaptation in specific lineages on shallower timescales (Natarajan et al. 2016).

Hemoglobin has been a model protein in the field of structural biology for more than a century. In 1825, J. F. Engelhart estimated the molar mass of an individual globin chain as 16 kDa (the first time the mass of a protein had ever been determined), leading to much controversy in the field owing to the fact that it was orders of magnitude larger than any known molecule at the time (Engelhart 1825). In the 1950s, Max Perutz's group obtained a crystal structure of the tetrameric complex (Bragg, Lawrence William; Perutz 1952) at around the same time Kendrew's group crystallized and solved the structure of its close monomer relative, myoglobin (Kendrew Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C. 1960). The subunit interfaces of Hb are essentially the first protein-protein interfaces to have even

been structurally characterized. The assembly of the tetramer is mediated through two structurally distinct interfaces that are both between α and β chains – which are conventionally labelled α1β1 and α1β2. In the decades since, dozens of Hb structures have been published, spanning virtually every major clade of vertebrates, from sharks (Ramesh et al. 2013) to birds (Knapp et al. 1999). This vast bedrock of structural information helped inform my interpretations and predictions of the effects of historical mutations in the Hb lineage.

Hb has also served as a model in the study of ligand-binding thermodynamics and allosteric function in proteins. In the early 20th century, Christian Bohr and his colleagues determined that the oxygen-binding equilibrium curve of Hb was sigmoidal rather than hyperbolic (Shimizu and Bucci 1974), indicating that protein subunits bind oxygen cooperatively. In order words, the binding of one subunit in the complex to oxygen increases the ability of the other subunits to bind to oxygen. This non-independence of ligand binding is important, because it allows Hb to bind oxygen in the lungs and release it to tissues over a physiologically relevant range of Oxygen partial pressures (Storz 2018). Bohr also demonstrated that the oxygen-affinity of Hb is reduced at low pH, an effect that now bears his name. Later work showed that Hb has evolved in various lineages to be responsive to wide array of other physiological effectors, including inositol hexaphosphate in birds, bicarbonate in crocodilians (Komiyama, Miyazaki, and Tamef 1995) and biphosphoglycerate in mammals (Shimizu and Bucci 1974).

Hb's oligomeric structure is key to its function. Compromising the interfaces of Hb leads to a loss in cooperativity and allosteric sensitivity (Bunn 2019). Conversely, mutations that induce even higher order assembly Hb, for instance into polymeric fibers can also be deleterious (Pauling et al. 2019). Lastly, alternative off-target stoichiometries, like α dimers (Kumar et al.

2014a) and β homotetramers (Kidd et al. 2002) are similarly nonfunctional. Much work has focused on the structural basis of cooperativity and allostery in hemoglobin. One way to study the linkage between ligand-binding and structure is to compare crystallographic structures of Hb in the oxy and deoxy states. The classical Monod-Wyman-Changeux (MWC) model of Hb's cooperativity (Monod et al. 1965) views the complex as existing in an equilibrium between two structural states – the tense conformation, which has low O2 affinity but high dimer-tetramer affinity, and the relaxed (R) conformation, which has high O2 affinity and low dimer-tetramer affinity. The structural data reveal that the transition from the T-to-R state involves (1) iron binding to oxygen, changing the position of iron relative to the heme-plane, (2) a shift in the position of a conserved histidine that coordinates the heme-iron, (3) a repositioning of the F-helix of the protein, which in turn (4) alters and collectively weakens contacts at the dimer-tetramer interface (α1β2), as well as changing the angles at which the dimers associate. This change in quaternary structure induces a gain in oxygen affinity for the other globin subunits. The origin of Hb's cooperativity is therefore necessarily tied to the evolution of the interfaces that hold the complex together – and more specifically to the dimer-tetramer interface.

I note finally that previous researchers have offered a number of speculations about the structural pathway of changes that led to today's heterotetrameric hemoglobin – Goodman (1981) proposed that the heterotetramer might have evolved from an ancestral homotetrameric complex that existed before the duplication of the α and β chains (Goodman 1981). Coates (1975), on the other hand, speculated that the ancestor of the two chains was a homodimer, perhaps similar to the structure of the lamprey hemoglobins (Coates 1975). The

results in this thesis show that Coates was partially correct – the α/β ancestor was in fact a homodimer, but it uses an interface distinct from that found in lamprey Hb to self-assemble (Pillai et al. 2020). The dimeric "hemoglobins" of lampreys are now known to represent an example of convergent evolution – its interfaces are independently derived (Schwarze et al. 2014).

1.4 Organization of this dissertation

This dissertation is organized into three chapters, the first two of which relate specifically to the evolution of multimeric assembly in hemoglobin, while the third represents a broader scholarly overview of evidence of the mechanisms and processes relating to the origin of complex protein features.

In Chapter 2, I reconstruct the phylogenetic history of globin evolution. I resurrect and functionally and structurally characterize key ancestral proteins, outlining the history of both multimeric assembly and allostery. I then identify subsets of historical changes that were causal for assembly and cooperativity.

In Chapter 3, I build upon the structural genetics work in Chapter 1 to determine the minimal genetic cause of tetramer assembly in Hb. Additionally, I identify a set of interfacial subsitutions that are sufficient to deliver specific assembly into heteromers rather than homomers. I also determine whether or not the evolution of one interface is genetically and biophysically dependent on the evolution of the other one.

In Chapter 4, I present a scholarly review of empirical data and theory relevant to how complex protein features arise, and the extent to which the historical assumptions of Darwin and Fisher

about complex traits apply to the evolution of folds, allosteric networks and multimers from scratch. I present experiments drawn from comparative biochemistry, protein engineering and high-throughput mutagenesis studies to argue that novel interfaces, allosteric networks and, in a few known cases, entirely new folds, could potentially arise though a small set of evolutionary steps, despite their apparent biophysical complexity.

Chapter 2: Origin of complexity in hemoglobin evolution

Most proteins associate into multimeric complexes with specific architectures (Ahnert et al. 2015; Marsh and Teichmann 2015), which often have functional properties like cooperative ligand binding or allosteric regulation (Monod et al. 1965). No detailed knowledge is available about how any multimer and its functions arose during historical evolution. Here we use ancestral protein reconstruction and biophysical assays to dissect the origins of vertebrate hemoglobin (Hb), a heterotetramer of paralogous α and β subunits, which mediates respiratory oxygen transport and exchange by cooperatively binding oxygen with moderate affinity. We show that modern Hb evolved from an ancient monomer and characterize the historical "missing-link" through which the modern tetramer evolved–a noncooperative homodimer with high oxygen affinity, which existed before the gene duplication that generated distinct α and β subunits. Reintroducing just two post-duplication historical substitutions into the ancestral protein is sufficient to cause strong tetramerization by creating favorable contacts with more ancient residues on the opposing subunit. These surface substitutions dramatically reduce oxygen affinity and even confer weak cooperativity, because of an ancient structural linkage between the oxygen binding site and the multimerization interface. Our findings establish that evolution can produce new complex molecular structures and functions via simple genetic mechanisms, which recruit existing biophysical features into higher-level architectures.

The interfaces that hold molecular complexes together typically involve sterically tight, electrostatically complementary interactions among many amino acids (Goodsell and Olson 2000). Similarly, allostery and cooperativity usually depend on numerous residues that connect

surfaces to active sites (Rivalta et al. 2012). Acquiring such complicated machinery would seem to require elaborate evolutionary pathways. The classical explanation, by analogy to the evolution of morphological complexity, is that multimerization conferred or enhanced beneficial functions, allowing selection to drive the many substitutions required to build and optimize new interfaces (Dawkins 1997; Goodsell and Olson 2000).

Whether this model accurately describes the evolution of any natural molecular complex requires a detailed reconstruction of the historical steps by which it evolved. The structural mechanisms that mediate Hb's multimeric assembly, cooperative oxygen binding, and allosteric regulation are well established (Perutz et al. 1960; Storz 2018). Despite considerable speculation (Goodman and Moore 1973; Coates 1975; Zuckerkandl 1965), however, virtually nothing is known about the evolutionary origin of Hb's heterotetrameric architecture and the functions that depend on it.

2.1 From monomer to homodimer.

We inferred the phylogeny of Hb and closely related globins (Fig. 2.1A, Ext. Fig. A1a,b,e). Hbα and Hbβ subunits are sister paralogs produced by a gene duplication that occurred before the last common ancestor of jawed vertebrates (Fig. 2.1a). The closest outgroups – myoglobin (Mb) (Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore 1960), globin E (Blank et al. 2011a), and globin Y (Appendix A: EFig. A1d) – are monomers. A more distant clade of agnathan "hemoglobin" and vertebrate cytoglobin includes monomers and dimers (Fago et al. 2018; Lechauve et al. 2010), but the dimers assemble through interfaces that differ from each other and from those used in Hb, indicating parallel

acquisition (Heaslet and Royer 1999; Makino et al. 2006). These observations suggest that the Hb $\alpha_2\beta_2$ heterotetramer evolved from an ancestral monomer via an unknown intermediate form.

To characterize when and how the tetramer evolved, we first reconstructed Hb of the ancestral jawed vertebrate by phylogenetically inferring the sequences of the ancestral α and β subunits (Fig. 2.1a, Appendix A: Extended Fig A1.b-c). We coexpressed and purified Ancα and Ancβ and characterized their assembly using native mass spectrometry (nMS), size-exclusion chromatography (SEC) and multi-angle light scattering. Like extant Hb, Ancα+Ancβ associate into $\alpha_2\beta_2$ heterotetramers, with a tetramer-dimer dissociation constant (K_d) of 10 μM, comparable to human Hb (15 μM, Fig. 2.1b-c, EFig. A2.a-c,f,i). Expressed in isolation, Ancα forms homodimers (Appendix A: Ext. Fig. A1.4a), and Ancβ forms homotetramers (Ext. Fig. A1.4b), just as extant Hb subunits do (Kidd et al. 2002; Kumar et al. 2014). Hb's heterotetrameric structure therefore evolved before the jawed vertebrate ancestor.

In contrast, Ancα/β, the pre-duplication ancestral protein, homodimerizes with a K_d of 9 μM measured by nMS, but the tetrameric state is unoccupied (Fig. 2.1b, Appendix A: EFig. A2d,f-g). Even at 1.4 mM, no tetramers are detectable using SEC (Appendix A: EFig. A2h). Ancα/β was therefore a homodimer, with virtually no propensity to tetramerize. This result is robust to incorporating statistical uncertainty about the ancestral sequence in an alternative construct (Appendix A: Ext. Fig. A3). This is also the most parsimonious history, because extant Hbα dimerizes and Hbβ tetramerizes when expressed in isolation (Kumar et al. 2014b; Kidd et al. 2002): a monomeric Ancα/β would imply independent gains of dimerization, and a tetramer

would require early gain of tetramerization followed by loss in Hbα (Appendix A: EFig. A3). AncMH, the common ancestor of Hb and myoglobin, is monomeric. No higher-order stoichiometries were detected using nMS of His-tagged AncMH at 70 μM (Appendix A: EFig. A4f). Even at 600 μM, only monomers are apparent using SEC (EFig. A2j). The untagged protein also does not dimerize at concentrations at which Ancα/β is predominantly dimeric, as shown using SEC and a globin-specific concentration assay on lysate from transformed cells (Ext. Fig. A4d-e). A monomeric AncMH is also the most parsimonious scenario, because its closest outgroups are all monomers (Appendix A: EFig. A3b-e).

The Ancα/β homodimer is therefore the evolutionary missing link between an ancient monomer and the Hb heterotetramer. After duplication, a novel interaction evolved, enabling these dimers to associate into tetramers.

2.2 Evolution of Hb functions.

We characterized the evolution of Hb's functional properties by assaying the ancestral proteins' oxygen binding characteristics. Modern Hb's physiological role – loading oxygen in the lungs/gills and unloading it in the periphery – is possible because Hb binds and releases oxygen cooperatively and has affinity lower than myoglobin; its affinity is further reduced by allosteric effectors (Storz 2018). Like human Hb, Ancα+Ancβ displays measurable cooperativity, and its oxygen affinity is similar to that of stripped, recombinant human Hb (Hoffman et al. 1990) (Fig. 2.1d-e). Ancα+Ancβ's affinity is reduced in the presence of the allosteric effector inositol hexaphosphate (IHP), although less so than that of human Hb (Hoffman et al. 1990). The

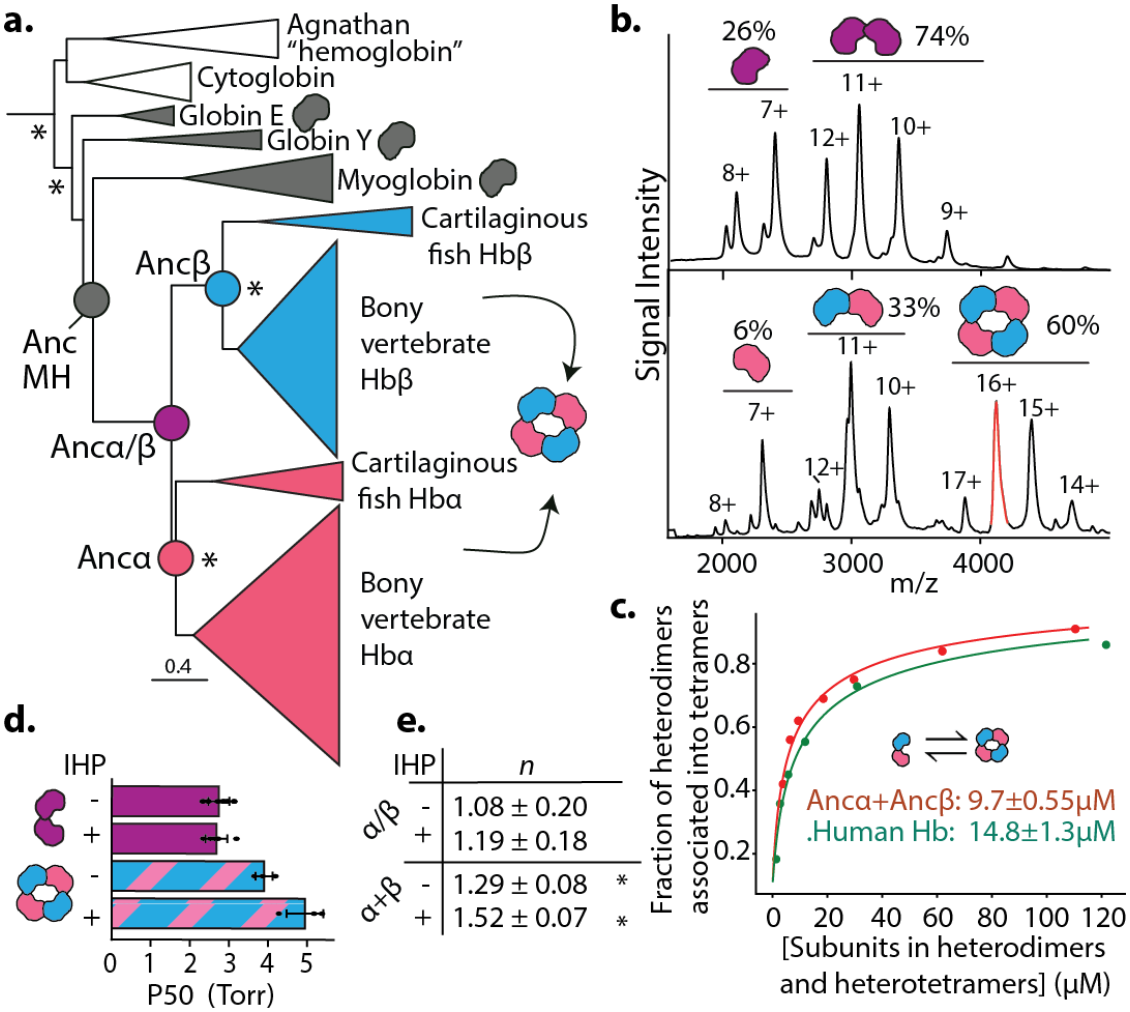


Figure 2.1 Structure and function of ancestral globins. a) Simplified phylogeny of vertebrate globins. Icons, oligomeric states. *, approximate likelihood ratio statistic >10. Complete phylogeny in Extended Figure 1a. Circles, reconstructed ancestral proteins. **b**) nMS spectra of Ancα/β (upper, purple) and Ancα+Ancβ (lower, pink+blue) at 20 μM. Charge states, stoichiometries, and occupancy (fraction of moles of subunits) shown. Red, analyzed by MSMS in Appendix A: EFig. A2e**. c)** Dimer-to-tetramer affinity of Ancα+Ancβ (red) and Human Hb (green). Circles, fraction of $\alpha+\beta$ heterodimers incorporated into $\alpha2\beta2$ tetramers, measured once by nMS. K_d (dissociation constant, with SE, in moles of subunits in heterodimers or heterotetramers) estimated by nonlinear regression. **d,e)** Oxygen affinity (P50) and cooperativity (Hill coefficient, *n*) of Ancα/β and Ancα+Ancβ. +IHP, 2x molar excess inositol hexaphosphate. Mean and 95% c.i. from 3-5 replicates (dots) shown. *, significant cooperativity ($n\neq1$, $P<0.05$, *F*-test; Appendix A: EFig. A1f).

functional characteristics of extant Hb were therefore in place by the jawed vertebrate ancestor.

In contrast, Ancα/β has oxygen affinity significantly higher than Ancα+Ancβ, and it does not display detectable cooperativity or allosteric regulation by IHP (Fig. 2.1d-e, Supplementary Discussion). The major functional characteristics of modern Hb therefore evolved between Ancα/β and Ancα+Ancβ, the same interval during which tetramerization evolved. This also represents the most parsimonious history: Hb tetramers are cooperative, but Hbα homodimers and Hbβ homotetramers are not (Tyuma, Benesch, and Benesch 1966; Kidd et al. 2002), suggesting that this property did not yet exist in their common ancestor (Appendix A: EFig. A3).

If Ancα/β lacked cooperativity, allostery, or reduced affinity, it could not have performed the physiological role that Hb now plays in oxygen exchange. Further, the first step in the evolution of Hb's tetrameric architecture – acquisition of homodimerization from a monomeric ancestor – could not have been driven by selection for Hb's major functional properties, because the homodimer did not possess any of them.

2.2 Ancestral and derived interfaces.

Hb assembly involves two distinct interfaces on each subunit: IF1 mediates α1-β1 and α2-β2 contacts, while IF2 mediates α1-β2 and α2-β1 contacts (Fig. 2.2a) (Perutz et al. 1960). To identify which interface evolved before Ancα/β, we applied hydrogen-deuterium exchange mass spectrometry (HDX-MS) to Ancα/β. We compared patterns of deuterium uptake at high versus low protein concentrations (at which dimers or monomers predominate, respectively, Appendix A: EFig. A2d,f-g). Solvent-exposed residues incorporate deuterium faster than buried

residues, so peptides that contribute to the dimer interface should exhibit higher deuterium uptake when the monomeric state predominates. We found that Ancα/β peptides with residues in IF1 incorporate significantly more deuterium under monomer-favoring than dimer-favoring conditions; no difference was observed for IF2 (Fig. 2.2b-c, Ext. Fig. A5-7). Moreover, mutating residues in IF1 substantially impairs Ancα/β dimerization, but a mutation that disrupts IF2 in human Hb (Manning et al. 1999) had no effect (Fig. 2.2d, Ext. Fig. A7c, Ext. Fig. A9). Reverting all IF1 residues in Ancα/β to the amino acid state from AncMH yielded predominantly monomers, but reverting those at IF2 had no effect (Fig. 2.2d, Ext. Fig. A7d). Ancα/β homodimers therefore assembled via IF1. After duplication, IF2 evolved, enabling assembly of dimers into tetramers (Fig. 2.2e). Corroborating this inference, extant Hbα homodimers assemble via IF1, whereas Hbβ tetramers use both IF1 and IF2, indicating inheritance of IF1 from their ancestor Ancα/β. This finding explains why Ancα/β is neither cooperative nor allosterically regulated, because both functions require IF2-mediated assembly into tetramers. (Ackers 1980)

2.4 Genetic mechanisms for the new interface.

The causal substitutions for the evolution of heterotetramers from the homodimer must have occurred on one or both of the post-duplication branches leading from Ancα/β to Ancα and to Ancβ. On the Ancα branch, there were only 3 changes, of which none were at IF2. On the Ancβ branch, there were 42 changes, including 5 at IF2 and 4 others at IF1 (Fig. 3a,b).

Introducing the IF2 substitutions into Ancα/β (Ancα/β5) confers strong assembly into tetramers, including both heterotetramers and homotetramers, when coexpressed with Ancα (Fig. 2.3c, Appendix A: EFig. A10.c,d). A version containing only 4 of these (Ancα/β4) also forms

homotetramers at 20μM but does not heteromerize with Ancα; the fifth change (h104E) therefore confers the capacity to associate with Ancα, presumably because it interacts with His104 on Ancα, forming a hydrogen bond in the heteromer but clashing in the homomer (Fig. 2.3b, Appendix A: EFig. A10.a,b).

Even a subset of just two IF2 changes (Ancα/β2) causes high-affinity assembly into homotetramers (K_d=1 μM, Figs. 1.3c,1.3e, Appendix A: EFig. A10.g). The genetic basis for the evolution of a new strong interface was therefore simple.

The IF2 substitutions are not sufficient to yield specific occupancy of the $\alpha_2\beta_2$ architecture: coexpression of Ancα/β5+Ancα forms a mixture of tetramers containing zero, one, or two α subunits (Fig.. 1.3c, Ext. Figs. A10.c,d). We hypothesized that IF1 substitutions conferred heterospecificity by favoring assembly of heterodimers across IF1, which then form $\alpha_2\beta_2$ heterotetramers across IF2. We introduced the IF1 substitutions into Ancα/β5 (Ancα/β9) and coexpressed it with Ancα. As predicted, heterotetramers and heterodimers predominated over homomers (Fig. 2.3d). Ancα/β9+Ancα is poorly soluble, preventing quantitation by nMS, but adding 5 historical substitutions at sites proximal to the interfaces (Ancα/β14+Ancα) improved solubility, and nMS confirms preferential occupancy of $\alpha_2\beta_2$ heterotetramers (K_d=6 μM, Fig. 2.3c,e, Appendix A: EFig. A10e,f).

The Hb heterotetramer therefore evolved from the Ancα/β homodimer via two sets of substitutions. Changes at IF2 created a strong new interface that conferred tetramerization; changes at IF1 yielded heterospecificity. In both cases, only a few substitutions were required.

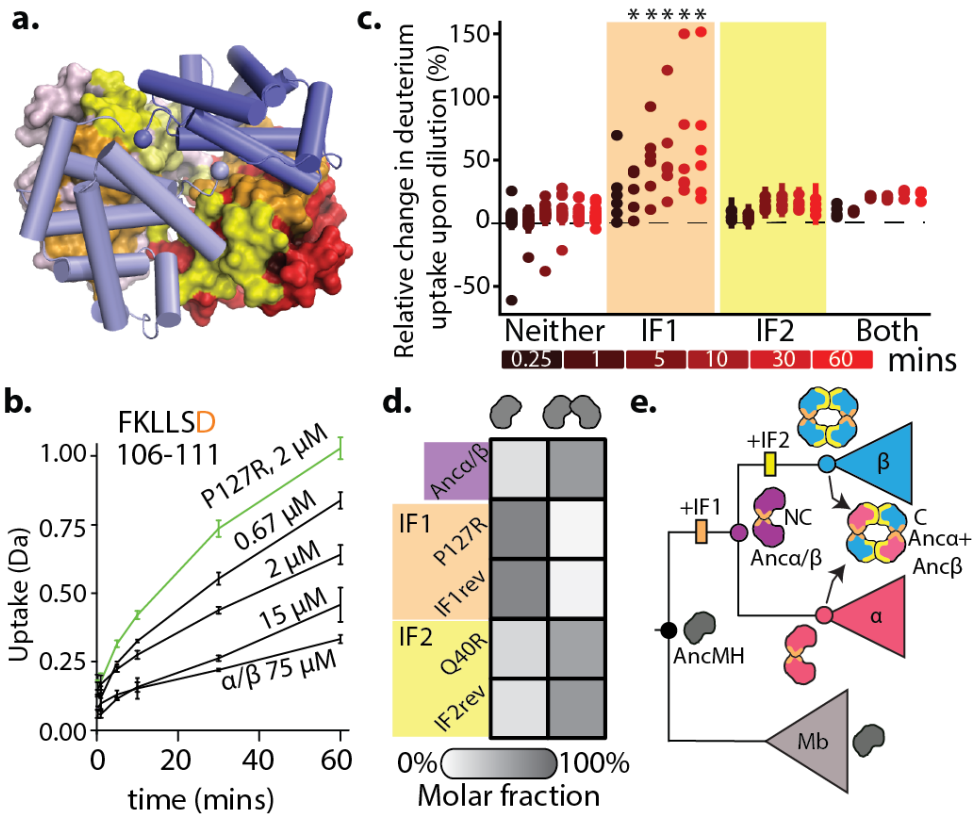


Figure 2.2. Identification of homodimerization interface in Ancα/β. a) Hb heterotetramers assemble via two interfaces (IF1, orange; IF2, yellow) on each subunit. Red and pink surfaces, α subunits; blue cartoon, β subunits. Ancα+Ancβ homology model is shown. **b)** Deuterium incorporation by an Ancα/β peptide that contributes to IF1 (Appendix A: EFig. A5g,h). Uptake (mean and SE from 3 replicates per incubation time) is shown for Ancα/β (black) and monomeric IF1 mutant P127R (green). **c)**Each circle, mean difference in deuterium uptake by one Ancα/β peptide when expressed at monomer-favoring vs. dimer-favoring concentrations (0.67 and 75 μM, 3 replicates each, with SE). Peptides are classified by the interface to which they contribute and colored by incubation time. *, mean uptake in interface category significantly different from other categories (P<0.05, permutation test, Extended Figs. 6g,7). **d)** Dimer and monomer occupancy by Ancα/β and mutants, assessed using nMS at 20 μM. P127R and Q40R disrupt IF1 and IF2, respectively. IF1rev and IF2rev revert historical substitutions to state in AncMH (spectra in Appendix A: EFig. A7c-d). **e)** Evolution of Hb tetramer. Rectangles, acquisition of IF1 and IF2. C, cooperative; NC, noncooperative. Mb, myoglobin.

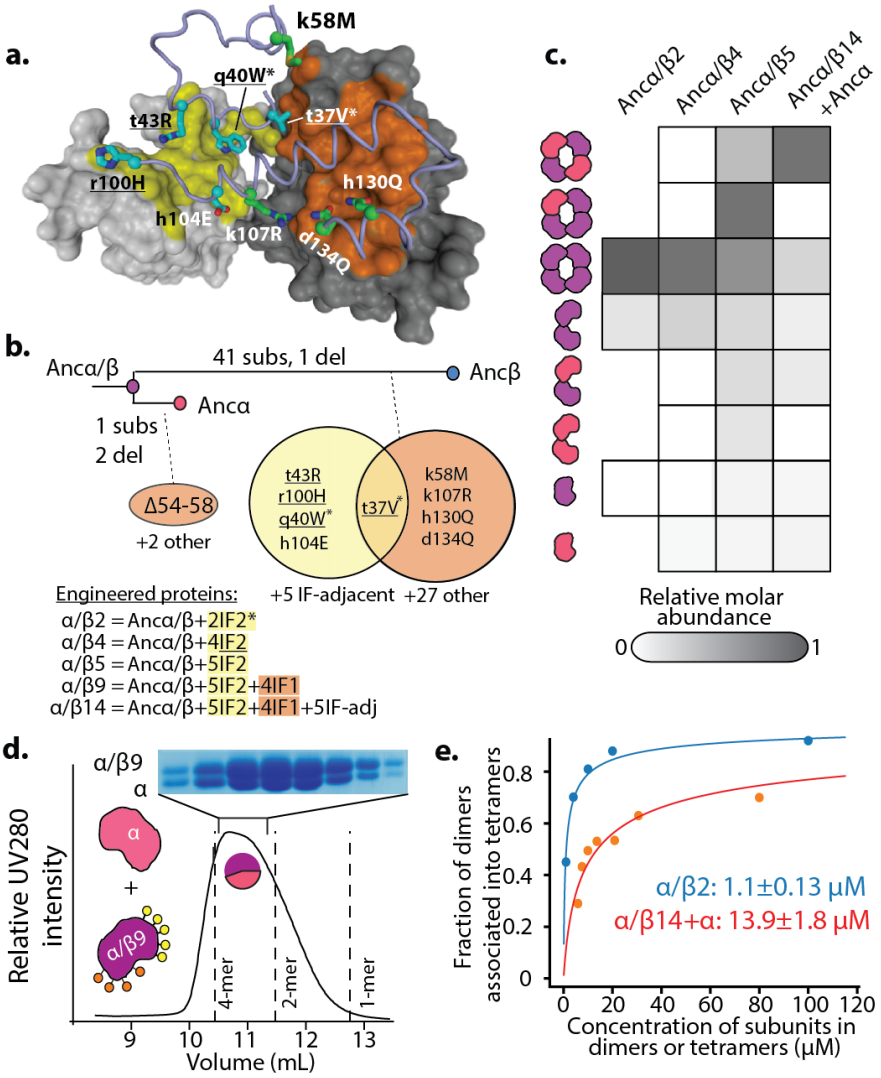


Figure 2.3. Genetic mechanisms of tetramer evolution. a) Homology model of Ancα+Ancβ tetramer with interface residues substituted between Ancα/β and Ancβ. Gray surfaces, two Ancα subunits; yellow, IF2; orange, IF1. Blue cartoon, partial backbone of one Ancβ subunit; sticks, side chains of substituted sites (IF2 cyan, IF1, green). Labels show state in Ancα/β (lower case) and Ancβ (upper). *, sites in Ancα/β2; underlined, Ancα/β4. **b)** Phylogenetic interval between Ancα/β and Ancα+Ancβ with number of substitutions and deletions per branch. Venn diagrams, sites substituted at interfaces. Below, substitutions incorporated in mutant proteins. **c)** Occupancy of multimers, measured by nMS at 20 μM, as fraction of moles of subunits in each state. Ancα/β2 was expressed in isolation, so only homomers are plotted. Spectra in Appendix A: EFig. A10. **d)** SEC of Ancα/β9+Ancα at 80 μM. Lines, elution volumes of tetramer (Ancα+Ancβ), dimer (Ancα/β), monomer (Human Mb). Pie, proportions of Ancα and Ancα/β9 subunits in tetramer-containing fraction, by denaturing MS (Appendix A: EFig. A11e). Above, electrophoresis of tetramer-containing fraction. **e)** Dimer-to-tetramer affinity of Ancα/β2 (blue) and Ancα/β14+Ancα (orange). Orange circles, fraction of Ancα/β14+Ancα heterodimers incorporated into heterotetramers; blue, fraction of Ancα/β2 homodimers in homotetramers, measured by nMS once. K_d (with SE) estimated by nonlinear regression.

2.5 Structural mechanisms of interface acquisition.

How could so few substitutions have generated a new and specific multimeric interaction? Using a homology model of the heterotetramer, we identified all favorable contacts that mediate association across the ancestral interfaces and used the phylogeny to determine when these amino acids evolved (Fig. 2.4a-c, Extended Figs. A10h-i).

The substitutions that conferred tetramerization recruited residues that already existed on the opposing surface into newly favorable interactions. All 13 residues that Ancα contributes to IF2 are unchanged from their ancestral state in Ancα/β, and many were acquired earlier (Fig. 2.4c). The IF2 substitutions on the Ancβ branch yielded new van der Waals contacts and hydrogen bonds with these ancient residues (Fig. 2.4c,d). For example, the ring of Trp40 (substituted in Ancβ from the ancestral Gln) nestles tightly in an ancient hydrophobic indentation on Ancα. Similarly, the IF1 substitutions that increase occupancy of the $\alpha_2\beta_2$ heterotetramer all modify interactions with ancient residues that were conserved on Ancα (Fig4b,e).

Both interfaces also involve favorable contacts between residues that were unchanged from their deep ancestral states in both subunits. In IF1, for example, R33 on each subunit donates two hydrogen bonds to F125 on the facing surface, and both residues evolved before AncMH. Each subunit contains both residues, and IF1 occurs twice in the tetramer, so these two sites form a total of 8 hydrogen bonds in the complex (Fig. 2.4b,e). Similarly, IF2 contains several hydrogen bonds and Van der Waals interactions between pairs of residues that originated before Ancα/β.

Because of the exponential relationship between binding energy and affinity, one substitution can dramatically increase occupancy of the multimer, if it builds on the foundation of even very

weak interactions between older residues. Satisfying an unpaired hydrogen-bond donor/acceptor or burying a hydrophobic ring can contribute up to 16 kJ/mol to an association (Fersht et al. 1985; Eisenberg and McLachlan 1986). Each interface occurs twice in Hb (Fig. 2.2a), so a substitution that confers a favorable interaction does so twice in the tetramer, doubling its effect on binding free energy and reducing K_d by up to 6 orders of magnitude. A single mutation can therefore shift occupancy of the tetramer from virtually nonexistent to the predominant species.

2.6 Mechanisms of Cooperativity.

Finally, we sought insight into the evolution of Ancα+Ancβ's cooperativity and reduced affinity. Cooperativity in extant Hb involves two conformational states that all subunits can adopt: one has higher affinity for oxygen but weaker IF2 contacts between subunits than the other (Mihailescu and Russu 2002; Ackers 1980). Cooperativity is classically thought to be mediated by an "allosteric core" – the set of residues on the helix that connect the heme to IF2, which is positioned differently in the two conformations (Gelin, Lee, and Karplus 1983).

To understand the mechanisms that triggered the evolution of cooperativity and reduced oxygen affinity, we first examined the phylogenetic history of residues in the heme pocket and allosteric core. At sites within 4 Å of the heme, no substitutions occurred during the interval when cooperativity was acquired. The vast majority were acquired prior to AncMH (Fig. 2.5a, Ext. Fig. 1c), including the "proximal histidine," which covalently binds the heme iron and transduces the movement of the heme upon oxygen binding to the allosteric core and IF2, thereby triggering the conformational shift between low- and high-affinity states in other

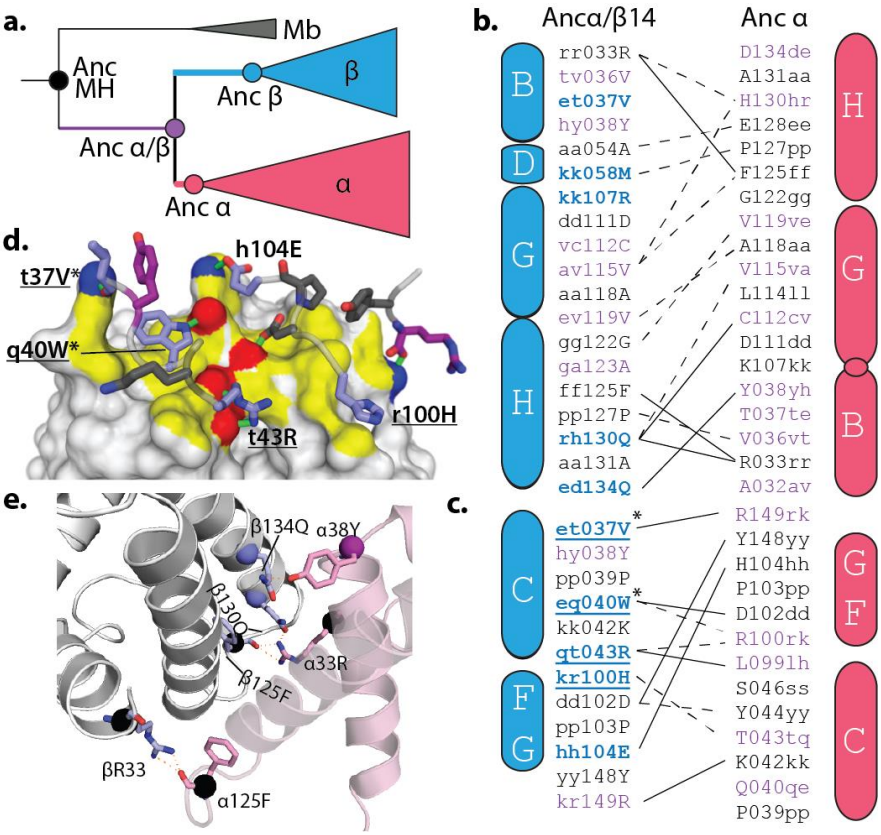


Figure 2.4. Structural mechanisms of Hb interface evolution. a) Phylogenetic classification of ancestral states and substitutions. Black, state in AncMH; purple, substituted from AncMH to Ancα/β; blue or red, substituted from Ancα/β to Ancβ or Ancα. **b,c)** Contact maps for residues buried at IF1 (**b**) and IF2 (**c**) Ancα+Ancβ**.** Residues colored by scheme in **a.** Letters, state in AncMH (outside, lower case), Ancα/β (middle, lower case) and Ancβ or Ancα (inside, upper case). Solid lines, predicted hydrogen bonds; dotted, van der Waals interactions. Underlined, substitutions in Ancα/β4; *, in Ancα/β2. Cylinders, helices (See Appendix A: EFig. A2a). Circle, deletion of helix. **d)** IF2 contacts in Ancα+Ancβ. Grey surface, Ancα, with yellow IF2; hydrogen-bonding atoms are red (oxygen) or blue (nitrogen), with bonds as green lines. Cartoon, Ancβ backbone, with IF2 interacting sidechains (sticks, colored as in **a**). **e)** Close-up of IF1 in Ancα+Ancβ model. Sticks, hydrogen-bonding residues; spheres, Cα atoms, colored by **a**.

subunits. Two substitutions occurred in Ancβ on the helix that connects IF2 to the histidine, but there were none in Ancα (Fig. 2.5a), and both subunits make the conformational transition in extant Hb. These observations suggest that the structural properties that mediate the allosteric linkage between the heme/oxygen-binding site and IF2 already existed in Ancα/β, before cooperativity and tetramerization evolved. Consistent with this idea, many of the conformational changes that mediate Hb cooperativity, such as distortion of the heme's geometry and movement of the histidine and helix upon oxygen binding, also occur in myoglobin, which is monomeric and noncooperative (Sato et al. 2007; Barends et al. 2015).

We hypothesized that, because of this ancient structural connection between the IF2 surface and the active site, evolution of the intersubunit interaction across IF2 conferred cooperativity and reduced affinity. We characterized oxygen binding by Ancα/β2, which contains only two historical substitutions at IF2. As predicted, these mutations reduce oxygen affinity by 2- to 3-fold compared to Ancα/β (Fig. 2.5b); they also confer weak but statistically significant cooperativity (Appendix A: EFig. A5b). Acquisition of the tetrameric association alone therefore changes the protein's oxygen-binding function and confers cooperative oxygen binding. The tetramer's ability to transition between high- and low-affinity states, however, is sensitive to mutation. Ancα/β4 and the Ancα/β14+Ancα heterotetramer also have reduced oxygen affinity relative to Ancα/β, but they lose the cooperativity found in Ancα/β2 (Fig. 2.5b). A likely explanation is that these additional mutations overstabilize the low-affinity conformation relative to the high-affinity state. If so, then some of the other substitutions that occurred between Ancα/β and the cooperative complex Ancα+Ancβ must have tuned this equilibrium so that both conformations are occupied, depending on the oxygen partial pressure

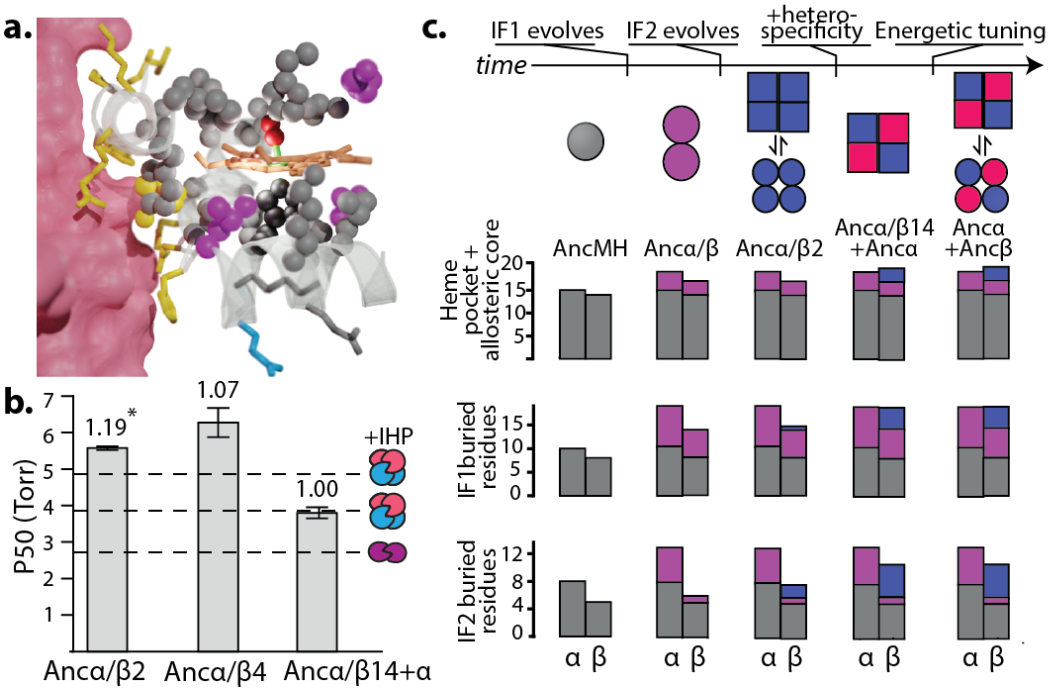


Fig. 2.5. Evolution of cooperativity by interface acquisition. a) Heme pocket and IF2 in Ancα+Ancβ. Pink surface, one Ancα. Heme (tan sticks, with green iron and red oxygen). Spheres, Ancβ residues within 4 Å of heme, colored by temporal category: grey, conserved since AncMH (dark grey, iron-coordinating histidine); purple, conserved since Ancα/β; blue, substituted between Ancα/β and Ancβ. Sticks, other residues on helix connecting histidine to IF2, colored temporally. Yellow, Ancβ residues at IF2. No changes near heme or IF2 occurred in Ancα. **b)** Oxygen binding by Ancα/β mutants with historical substitutions. Columns and error bars, P50 ± SE, with Hill coefficient *n* above, estimated by nonlinear regression under effector-stripped conditions (raw data in Extended Figure 10j). *, significant cooperativity ($n \neq 1$, $P = <0.05$, F-test, Appendix A: EFig. A1f). Dotted lines, affinities of Ancα+Ancβ and Ancα/β, which is unaffected by IHP. **c)** Evolution of the cooperative Hb heterotetramer. Circles and squares, conformations with high and low oxygen affinity, respectively. Two IF2 substitutions cause homotetramerization, cooperativity, and reduced affinity (see B). Other substitutions that confer heterotetramerization change the relative stabilities of high and low-affinity conformations, abolishing/restoring cooperativity. White box, interval in which order of substitutions is unknown. **d)** Acquisition of residues in structurally defined categories in Ancα and Ancβ, ordered as in **d**, colored by temporal category. No changed occurred in Ancα.

(Fig. 2.5c). The order in which these changes occurred cannot be resolved: the IF2 substitutions may have immediately generated a cooperative Hb-like complex, similar to Anc$\alpha/\beta2$; alternatively, cooperativity may have evolved via a low-affinity tetrameric intermediate, like Anc$\alpha/\beta4$ (Fig. 2.5c).

2.7 Evolution of molecular complexity.

Our findings establish that simple genetic changes drove the evolution of Hb's complex structure and functions from its dimeric precursor. Other molecular complexes may also have evolved by short mutational paths. Interactions between proteins and other kinds of substrates, such as DNA or small molecules, have historically evolved via one or a few historical substitutions (Siddiq, Hochberg, and Thornton 2017), and we see no reason why multimeric interactions should be more difficult to evolve. Multimers can be engineered from non-assembling precursors by one or a few mutations, (Garcia-Seisdedos et al. 2017; Grueninger et al. 2008) and naturally occurring point mutations are known to cause disease by inducing higher-order complexes (Pauling et al. 2019).

The simple mechanism by which Hb appears to have evolved its cooperativity – acquisition of binding to a molecular partner at a new interface – could explain the origin of cooperativity and allostery in other systems (Coyle, Flores, and Lim 2013; Reynolds, McLaughlin, and Ranganathan 2011). If two plausible conditions are met – the new interface is near or structurally connected to the functionally active site, and the optimal conformation for binding is different from the optimal conformation for activity – then binding will impair activity, and

vice versa. Given this tradeoff, evolution of binding will confer cooperativity or negative allostery.

Hb's history shows that complex molecular structures and functions can arise by means other than the long, gradual trajectories of functional optimization by which biological complexity has long been thought to evolve (Dawkins 1997; Darwin 1859). In principle, molecular assemblies could arise and become more complex via neutral processes (M. Lynch 2013; Finnigan et al. 2012; Gray et al. 2010), but this scenario is unlikely if many mutations are required. Our work shows that Hb's higher-level multimeric state and functional properties evolved through just a few mutations, which fortuitously built upon and interacted with ancient structural features. These older features could not have been initially acquired because of selection for the functions of the final complex, because they existed before those functions first appeared. Some likely originated and were preserved by selection for ancestral functions, while others may have transiently appeared by chance. Although evolution of any particular molecular sequence or architecture without consistent selection for those properties is vanishingly improbable, our findings suggest that proteins evolve constantly through a dense space of possibilities in which complex new interactions and functional states are easily accessible.

2.8 Methods

Sequence Data and Alignment. 177 annotated amino acid sequences of hemoglobin and related paralogs from 72 species were collected from UniPROT, Ensembl and NCBI RefSeq. Sequences were aligned using MAFFT v7 (Katoh, Rozewicki, and Yamada 2017). The ML

phylogeny and branch lengths were inferred from the alignment using PHYML v3.1 (Guindon et al. 2010) and the LG model (Le and Gascuel 2008) with gamma-distributed among-site rate variation and empirical state frequencies. This best-fit evolutionary model was selected using the Akaike Information Criterion in PROTTEST. Node support was evaluated using the approximate likelihood ratio test statistic (aLRS), which expresses the difference in likelihood between the most likely topology and the most likely topology that does not include the split of interest; aLRS has been shown to be reasonably accurate, robust, and efficient compared to other means of characterizing support (Anisimova and Gascuel 2006; Anisimova et al. 2011). The tree was rooted on neuroglobin and globin X, paralogs that are found in both deuterostomes and protostomes (Dröge et al. 2012). Tetrapods possess three paralogous Hbα genes, called Hbα(A), Hbα(D),and Hbα(Z); however, the ML phylogeny inferred from this alignment contained a weakly supported sister relationship between all Actinopterygian Hbα genes and the tetrapod Hbα(Z), to the exclusion of tetrapod Hbα(A) and Hbα(D). This is a nonparsimonious scenario, because it requires an early gene duplication and subsequent loss of the Hbα(A)/Hbα(D) lineage in Actinopterygii We therefore constrained the topology to unite the tetrapod Hbα(A), Hbα(D), and Hbα(Z) in a clade (Appendix A: EFig. A1A). PhyML v3.1 was then used to re-infer the best-fit branch topology and branch lengths given this constraint.

Ancestral sequences were reconstructed and posterior probability distributions of ancestral states were inferred using the ML method using the codeml package in PAML 4.9 (Yang 2007), given the ML constrained phylogeny and branch lengths. Historical substitutions were assigned to phylogenetic branches as differences between the maximum a posteriori amino acid states between parent and daughter nodes. The asymmetry between the branch

lengths leading from Ancα/β to Ancα and to Ancβ has been observed previously (Schwarze, Singh, and Burmester 2015) and presumably reflects there being more amino acid states shared between Hbα and the outgroups (myoglobin, globins E and Y, etc.) than between Hbβ and the outgroups. Sequences for reconstructed ancestors have been deposited in Genbank (IDs MT079112, MT079113, MT079114, MT079115).

Recombinant protein expression. Ancestral genes were codon-optimized for *E.* coli expression using CodonOpt and generated by de novo DNA synthesis (IDT gBlocks). For globin expression, coding sequences were cloned into pLIC expression vector without affinity tags and expressed under a T7 polymerase promoter. For oxygen-affinity measurements, plasmid pCOMAP (Natarajan et al. 2011), which expresses *E coli.* methionine aminopeptidase 1 (MAP1), was cotransformed to ensure efficient N-terminal methionine excision. For co-expression of two globins, sequences were expressed from a polycistronic operon in plasmid pGM, without tags and under a T7 promoter, separated by a spacer containing a stop codon and ribosome binding site. *E coli.* methionine aminopeptidase 1 (MAP1) was coexpressed from the same plasmid.

JM109 DE3 *E. coli* cells (NEB) were transformed and plated into solid Luria broth (LB) media containing 50 μg/ml carbenicillin (and 50 μg/ml kanamycin, if pCOMAP was being cotransformed). A single colony was inoculated into 50 mL of LB with appropriate antibiotics and grown overnight. 5 mL of this culture was inoculated into a larger 500 mL LB culture. Cells were grown at 37C and shaken at 225 rpm in an incubator (New Brunswick 126) until they reached an OD600 of 0.4-0.6. The culture was then supplemented with 0.5 mM isopropyl-β-d-1-thiogalactopyranoside (IPTG) and 50 mg/L of hemin (Sigma). After 4 hours of expression at 37C,

CO was bubbled through the solution for 10 minutes and cells were collected by centrifugation at 5000 g. Protein purification was carried out immediately after expression.

Protein purification by ion exchange. Ancα/β, P127R, V119A, Ancα/β4+Ancα and the alternative ancestral reconstructions were purified using ion exchange chromatography (Natarajan et al. 2011; Hoffman et al. 1990). All buffers were saturated with CO prior to purification and vacuum filtered through a 0.2 µM PFTE membrane (Omnipore) to remove particulates. After expression, cells were resuspended in 200 mL of 50 mM Tris (pH 6.8) with 2 cOMPLETE protease inhibitor tablets (Roche) and 0.5 mM DTT. The cell suspension was lysed in 50 mL batches in a glass beaker using an FB505 sonicator with a power setting of 90%, 1s on/off for 2 minutes. The lysate was then centrifuged at 30000g to eliminate cell debris, inclusion bodies and aggregates. The supernatant was further syringe-filtered used HPX Millex Durapore filters (Millipore). A HiTrap SP cation exchange (GE) column was attached to an FPLC system (AKTAprime plus) and equilibrated in 50 mM Tris (pH 6.8). Lysate was passed over the column. The SP column was washed with 200 mL of 50 mM Tris to eliminate weakly bound contaminants. Bound Hbs eluted with a 100 mL gradient of 50 mM Tris (pH 6.9) 1 M NaCl, from 0 mM to 1M. 0.5 mL fractions were collected along the length of the gradient. The 4 reddest fractions were collected and then concentrated in an Amicon µLtra-15 tube by centrifugation at 4000 g to final volume of 500 µL. The sample was injected into a Sephacryl Hiprep 16/60 S-100 HR size-exclusion column (SEC) for additional purification. The SE column was equilibrated in phosphate buffered saline (PBS) at pH 7.4. Depending on molecular weight, purified globins elute at 48-52 mL (tetramer), 56-60 ml (dimer) or 64-67 ml (monomer). The purity and identity

of isolated proteins was assessed using 20% SDS-PAGE and denaturing HRA-MS. The purified proteins were concentrated and then flash frozen with liquid nitrogen until usage.

Protein purification by zinc affinity chromatography. Ancα/β5 + Ancα, Ancα/β9+Ancα, Ancα/β14+Ancα, and Ancα+Ancβ were purified using zinc-affinity chromatography, adapted from (53). Buffers were loaded onto the metal affinity column using an AKTAprime FPLC. To prepare the zinc affinity column, nickel was removed from a HisTrap column (GE) using stripping buffer (100 mM EDTA, 100 mM NaCl, 20 mM TRIS, pH 8.0). The column was then washed with diH_2O for five column volumes. Then 0.1 M ZnSo4 was passed over the column until conductance reached a stable value. The column was then washed with 5 column volumes of water. After expression, cells were resuspended in 50 mL of lysis buffer containing 20 mM Tris and 150 mM Nacl (pH 7.4). The cells were sonicated as described in the previous section. The lysate was passed over a Zinc-affinity HisTrap column. The column was washed with 200 mL of wash buffer (20 mM Tris and 150 mM Nacl, pH 7.4). The bound Hbs were eluted with a 50 mL gradient of imidazole, upto 500 mM and 0.5 mL fractions were collected during the run. The 4 reddest fractions were collected. The Hb-containing fractions were concentrated and injected into a Sephacryl S-100 HR column for additional purification, as described above.

Purification of Globin Y. The Globin Y sequences of *Callorhincus milli* (NCBI reference sequence NP_001279719.1) and *Xenopus laevis* (NCBI reference sequence NP_001089155.1) were synthesized (IDT, Coralville, IA, USA) and cloned into a pLIC vector with an N-terminal hexahistidine tag (MHHHHHH). Expression and lysis were carried out under the same conditions as described in previous sections. The bacterial lysate was passed over a 5 mL HisTrap Nickel-affinity column (GE). The column was washed with 5 column volumes of

wash buffer (20 mM Tris and 150 mM Nacl, pH 7.4). The bound globins were eluted with a 15 mL gradient of imidazole from 0 to to 500 mM; five fractions of equal volume were collected. The 3 reddest fractions were combined. The eluted protein was concentrated to 2 mL, passed through a 0.45 µM filter, and subject to a final purification by size-exclusion chromatography using a Sephacryl S-100 HR column and an AKTA Prime FPLC system. Globin Y eluted in fractions collected between 61 and 64 mL.

Purification of his-tagged AncMH. The sequence of AncMH was codon-optimized for expression in *E. coli,* synthesized, and cloned into a pLIC vector with an N-terminal hexahistidine tag, because untagged AncMH was not readily purifiable. Recombinant expression, cell lysis, and purification were carried out under the conditions described for GbY.

Characterization of protein stability. Protein stability was measured by circular dichroism (CD) using a JASCO 1500 CD spectrophotometer. Experiments were conducted at protein concentration of 10 µM (50 mM Sodium fluoride, 20 mM Sodium phosphate buffer) in a 0.2 mm path length quartz cell. CD spectra were collected at 2°C intervals (10 minutes each) as the temperature was increased from 25°C to 95°C. Molar ellipticity at 222 nm was measured four times at each temperature; the mean was then divided by the value of molar ellipticity at 222 nm at room temperature (25°C) to estimate the fraction of unfolded protein. To estimate the melting point (T_m) of each protein, a custom script was written to find the best fit parameters (T_m and slope) for the Boltzmann sigmoid function: Fraction unfolded = $1/ (1+e^{(t-Tm)/slope)}$). All three ancestral proteins were stable, with T_m> 60°C (Appendix A: EFig.1c).

High resolution denaturing mass spectrometry. 200 µL of purified proteins were placed in Slide-A-Lyzer MINI dialysis unit that was suspended in 500 mL of 50 mM Ammonium Acetate.

The solution was stirred overnight at 4C. After dialysis, the proteins were transferred to a microfuge tube and centrifuged at 30,000g to eliminate aggregates. Concentration was adjusted to 20 µM. 0.5 µL of sample was sprayed using an Agilent 6224 Tof Mass Spectrometer at fragment voltage 200V. Protein masses were estimated by maximum entropy mass deconvolution implemented in MassHunter (Agilent).

Size exclusion chromatography (SEC) and multi-angle light scattering (MALS). All proteins were converted to the CO-bound form by adding sodium dithionite to 5 mg/ml, desalting on a Sephadex G-25 desalting column equilibrated with CO-saturated PBS (150 mM NaCl, pH 7.4), and then passing CO through the eluent. Protein concentration was measured by UV absorbance at 280 nm (Tryptophan) and 419 nm (HbCO-specific) using a Nanodrop 2000c (Thermo-scientific). For analytic size exclusion chromatography, a Superdex 75 10/300 GL column (GE) was equilibrated in CO-saturated PBS, and then injected with 500 µL of sample, using a 500 µL injection loop on an AKTAprime and monitored by absorbance at 280 nm. For SEC coupled with multi-angle light scattering (MALS), a Superdex 200 10/300 GL column was injected with 150 µL of sample on the AKTAprime; refractive index and light scattering of eluent was measured using a Dawn Helios-II (Wyatt) light scattering detector and Optilab T-rEX refractometer respectively. Molar mass fitting was carried using Astra software.

Globin concentration assay. After protein expression, cells harvested by centrifugation from one 500 mL culture were resuspended in 15 mL PBS and sonicated as described above. Cell debris and aggregate were removed by centrifugation at 20,000g. Remaining lysate was concentrated to 5 mL in Amicon µLtra-15 centrifuge concentrators (3,000 NMWL). 500 µL of this sample was injected into a superdex-75 10/300 GL column. 0.2 mL fractions of eluent were

collected. 50 µL was taken from each fraction and added to 150 µL of Hemoglobin Assay kit reagent (Sigma) in one well of a 96 well plate. In each plate, 50 µL of a 100 mg/dl calibrator (Sigma) was also added to 150 µL of Hemoglobin Assay kit reagent (Sigma) in one well. 50 µL of PBS added to the 150 µL reagent was used as a blank. Absorbance was measured at 400 nm using a Victor x5 plate reader (PerkinElmer). Heme concentration in each fraction was measured using the following equation: Concentration = 62.5* (OD_{sample}- OD_{blank})/ ($OD_{callibrator}$- OD_{blank}) µM.

Oxygen affinity and cooperativity. Purified proteins were deoxygenated using sodium dithionite at 10 mg/ml and immediately passed through a PD-10 desalting column (GE Healthcare) equilibrated with 25 ml of 0.01 M HEPES/0.5 mM EDTA (pH 7.4). Eluted proteins were concentrated using Amicon µLtra-4 Centrifugal Filter Units (Millipore). Equilibrium oxygen-binding assays were performed at 25°C using a Blood Oxygen Binding System (Loligo Systems), using 0.1 mM protein (heme concentration) dialyzed in 0.1 M HEPES/0.5 mM EDTA buffer. Protein solution was sequentially equilibrated at three to five different oxygen tensions (PO_2) yielding 30 to 70% saturation while continually monitoring absorbance at 430 nm (deoxy peak) and 421 nm (oxy/deoxy isosbestic point). Plots of fractional saturation vs PO_2 were constructed from these measurements, and the Hill equation was fit to each plot using OriginPro 2016, yielding estimates of P50 (PO_2 at half-saturation) and the cooperativity coefficient (*n,* the slope at half saturation in the Hill plot, n50). 95% confidence intervals on parameter estimates were calculated by multiplying the standard error of the mean over replicate experiments by 1.96 (Figs. 1d,e). Statistical significance of cooperativity was assessed

by using an F-test to compare the fit of the data to a model in which *n* is a free parameter to a null model in which *n*=1.

To assess the potential for ancestral proteins to have been regulated by allosteric effectors, assays were performed in stripped medium or with inositol hexaphosphate (IHP) added at 0.5 mM. Although IHP may not have been the physiological effector in ancestral organisms, it has been shown to allosterically regulate Hbs of representatives from all major vertebrate lineages, whereas other organic phosphates like 2,3-biphosphoglycerate (BPG), ATP, and GTP have more lineage-specific effects (Bonaventura and Bonaventura 1980; Weber, RE., Jensen 1988; Isaacks and Harkness 1980). IHP therefore serves as a useful "all-purpose" polyanion to test the allosteric regulatory capacity of the ancestral Hb. There is ample precedent for using IHP to study Hb allostery irrespective of whether it is the authentic physiological effector (Benesch, R. 1968; Imai 1982; Imaizumi et al. 1979; Grispo et al. 2012). This is because IHP modulates Hb-O_2 affinity in a manner that is qualitatively similar to other effectors, including BPG, ATP, GTP, and IPP (Storz 2018; Imai 1982). These molecules all share the same mechanism of action, reversibly binding a set of cationic residues in the cleft between β_1 and β_2 subunits, thereby stabilizing the low-affinity T conformation via electrostatic interactions (Richard, Dodson, and Mauguen 1993; Arnone 1972; Arnone A 1974).

Native Mass Spectrometry (nMS). Proteins were buffer exchanged into 200 mM ammonium acetate with a centrifugal desalting column (Micro Bio-Spin P-6, BioRad) and loaded into a gold-coated glass capillary. Samples were ionized for MS measurement by electrospray ionization. MS and MS/MS ion isolation were performed on a Synapt G1 HDMS instrument (Waters Corporation) equipped with a radio frequency generator to isolate higher *m/z* species

(up to 32k) in the quadrupole, and a temperature-controlled source chamber as previously described (Cong et al. 2016). Instrument parameters were tuned to maximize signal intensity for MS and MS/MS while preserving the solution state of the protein complexes. All samples were sprayed at room temperature. Instrument settings were: source temperature of 50 °C, capillary voltage of 1.7kV, sampling cone voltage of 100V, extractor cone voltage of 5V, trap collision energy of 25V, argon flow rate in the trap was set to 7 ml/min (5.6 x 10^{-2} mbar), and transfer collision energy set to 15V. The T-wave settings were for trap (300 ms^{-1}/1.0V), IMS (300 ms^{-1}/20V) and transfer (100 ms^{-1}/10V), and trap DC bias (30V). For MS/MS, ion isolation was achieved using the same settings as described above, with the quadrupole LM resolution was set to 6. Activation of protein complexes for individual monomer identification was achieved by increasing the trap collision voltage to 120V in MS/MS mode, with all other settings unchanged. Analysis of the MS and MS/MS data to estimate masses and relative abundances was performed with the software program Unidec (Marty et al. 2015).

Occupancy of each stoichiometric state was calculated as the proportion of globin subunits in that state, based on the summed areas under the corresponding peaks in the spectrum. To estimate K_d of the monomer-to-dimer transition Ancα/β, we performed nMS at variable protein concentrations. At each concentration, the observed fraction of subunits incorporated into dimers (F_d) was estimated as $F_d = \frac{2x_d}{x_m+2x_d}$, where x_m and x_d are the sum of the signal intensities of all peaks corresponding to the monomeric and dimeric stoichiometries, respectively. This procedure was repeated at a range of protein concentrations. Nonlinear regression was then used to find the best-fit value of K_d using the equation: $F_d = \frac{1}{P_{tot}} *$

$\frac{(4P_{tot}+K_d)-\sqrt{((4P_{tot}+K_d)^2-16{P_{tot}}^2)}}{4}$, where P_{tot} is the total protein concentration (expressed in terms of monomer) estimated by UV absorbance at 280 nm. The resulting K_d is expressed in terms of the concentration of globin subunits. We observed no higher stoichiometries.

To estimate K_d of the heterodimer-heterotetramer transition in Ancα+Ancβ (or mutant ancestral globins) we performed nMS at variable protein concentrations. Because nMS directly quantifies the abundance of all species in solution, we were able to extract molarities for the α_1/β_1 heterodimer and α_2/β_2 heterotetramers and directly calculate the K_d of their association/dissociation equilibrium, without having to fit a large number of K_ds as part of a coupled set of many equilibria across many homomeric and heteromeric forms. At each concentration, we first calculated the total fraction of subunits that were incorporated into heme-bound heterodimers, including both free heterodimers and heterodimers assembled into heterotetramers, as

$F_{\alpha\beta} = \frac{2x_{\alpha1\beta1}+4x_{\alpha2\beta2}}{x_{\alpha1}+x_{\beta1}+2x_{\alpha1\beta1}+2x_{\alpha2}+x2_{\beta2}+4x_{\alpha2\beta2}+2y_{\mathrm{apo}-\alpha1\beta1}}$, where x is the sum of the signal intensities of all peaks corresponding to the stoichiometry indicated by the subscript. $y_{\text{apo-}\alpha1\beta1}$ is the signal-intensity of the peaks corresponding to heterodimers that are only partially heme-bound and cannot associate into tetramers. The concentration of all heme-bound subunits incorporated into heterodimers (free heterodimers or assembled into heterotetramers) was calculated as $C_{\alpha\beta} = F_{\alpha\beta} \times P_{tot}$. The fraction of all heterodimers incorporated into heterotetramers was calculated as $F_{\alpha2\beta2} = \frac{4x_{\alpha2\beta2}}{2x_{\alpha1\beta1}+4x_{\alpha2\beta2}}$. Assembly of heterodimers into heterotetramers as concentration increases was then analyzed to find the best-fit value of K_d using nonlinear

regression and the following equation: $F_{\alpha 2\beta 2} = \frac{1}{C_{\alpha\beta}} * \frac{4C_{\alpha\beta}+K_d-\sqrt{(4C_{\alpha\beta}+K_d)^2-16C_{\alpha\beta}{}^2}}{4}$. The resulting K_d is expressed in terms of the concentration of globin subunits contained in heterodimers and heterotetramers.

For homotetramerization of globins expressed in isolation, the K_d of the dimer-tetramer transition was calculated using a similar approach. The fraction of all subunits incorporated into homodimers (including both free homodimers and those associated into homotetramers) was calculated as $F_d = \frac{2x_d+4x_t}{x_m+2x_d+4x_t}$, and the concentration of all dimers was calculated as $C_d = F_d \, x \, P_{tot}$. The fraction of all dimers that were incorporated into tetramers was calculated as $F_t = \frac{4x_t}{2x_d+4x_t}$. Nonlinear regression was then used to fit K_d to the data using the equation $F_t = \frac{1}{C_d} * \frac{4C_d+K_d-\sqrt{(4C_d+K_d)^2-16C_d{}^2}}{4}$. The resulting K_d is expressed in terms of the concentration of globin subunits contained in homodimers and homotetramers. For Fig. 3c, Ancα/β4 was coexpressed with Ancα, fractionated by SEC, and the tetrameric fraction analyzed by nMS.

Native MS spectra for human Hb and Ancα/β14+Ancα at high concentrations contained peaks corresponding to dimers that had lost one or both hemes. In these cases, wWe calculated K_ds by both including and excluding these species. For the fits shown in main figures (Figs. 1d, 3d), these peaks were excluded from the analysis; for the fits shown in Appendix A: EFig. A2k, they were included. Both approaches yielded K_d estimates of the same order, although the fit to the data was much better in the former case. Spectra for Ancα+Ancβ included twinned peaks, which represent cesium iodide adducts on tetramers. For the fits shown in the main figures (Figs. 1c, 3d), these peaks were excluded; for the fits in Appendix A: EFig. A2i , they were

included. Both approaches gave almost identical K_d estimates, although the fit to the data was better in the former case.

Hydrogen/deuterium exchange mass spectrometry (HDX-MS). All chemicals and reagents were purchased from Sigma Aldrich (Gilligham, UK). Native equilibration buffer contained 100 mM PBS (H_2O), pH 7.4. Labelling buffer contained 100 mM PBS (D_2O), pD 7.4. Quench buffer contained 100 mM potassium phosphate (H_2O), pH 1.9, with 1 M guanadinium chloride. 5 µL of protein sample was diluted into 55 µL of a deuterated buffer of the same composition and corresponding pD. This results in a labelling solution ~92 % D_2O. Samples were incubated between 15 s and 1 hour at 20° C before quenching with an ice-cold H_2O buffer (pH 1.9) of equal volume. The quenched solution pH was ~2.5 at 0 °C. This was quickly injected into an on-line HDX manager (Waters, Milford, MA, USA). The sample was injected on to a 50 µL sample loop at 0 °C before passing over an immobilised pepsin column (Enzymate Pepsin 5 µm, 2.1 mm × 30mm, Waters) at 20 °C using an isocratic H_2O (0.1 % v/v) formic acid solution (200 µL/min). Peptide products were collected on a trapping column (BEH C18, 1.7 µm, 2.1 mm × 5 mm, Waters) held at 0 °C. After 2 minutes of collection, and de-salting, peptides were eluted from the trap column on to an analytical column (BEH C18, 1.7 µm, 1 mm × 100 mm, Waters) for separation using a reverse-phase gradient with a flow rate of 40 µL/min. The elution profile using a H_2O/MeCN (+0.1% formic acid v/v) gradient was as follows: 1-7 minutes 97 % water to 65 % water, 7-8 minutes 65 % water to 5 % water, 8-10 minutes hold at 5 % water. The analytical flow rate was 40 µL/min and the eluate was electrosprayed directly into a Synapt G2Si (Waters, Wilmslow, UK) Q-ToF instrument for mass analysis.

Sample handling was semi-automated using a robotic liquid handling HDX system (LEAP technologies, Ringwood, Australia) to ensure reproducibility in timings. A blank and cleaning injection cycle was performed between each labelling experiment. Mass spectrometry conditions were as follows: capillary 2.8 kV, sample cone 30 V, source offset 30 V, trap activation 4 V, transfer activation 2 V. The source temperature was set to 80 °C and cone gas flow 80 L/hr, the desolvation temperature was 150 °C and the desolvation gas flow of 250 L/hr. LeuEnk was used as an internal calibrant and acquired every 30 s. For reference, back-exchange was estimated separately using lyophilised samples of angiotensin II. Angiotensin II was dissolved into D_2O (pH 4.0) and left for 48 hours. After which the sample was loaded onto the same robotic and UPLC system and analysed after 2 minutes of trapping to give a back-exchange of 31.8 ±0.2 %.

Peptides were identified, in the absence of labelling, by data-independent MS/MS analysis (MS^E) of the eluted peptides and subsequent database searching in the Protein Lynx Global server 3.0 software (Waters). Peptide fragments were generated in the trap region through collisions with Ar gas (0.4 mL/min). Peptide identifications were filtered according to fragmentation quality (minimum fragmentation products per amino acid: 0.2), mass accuracy (maximum [MH]+ error: 5 ppm), and reproducibility (peptides identified in all MS^E repeats) before their integration into HDX analysis. HDX-MS data were processed in DynamX 3.0 software (Waters), and all automated peptide assignments were manually verified, with noisy and overlapping spectra discarded. External python scripts were written to generate and analyse the Woods plots from data outputs of DynamX.

Sample concentration was varied to control the relative populations of monomeric and dimeric species of Ancα/β. After dilution into the labelling buffer Ancα/β concentrations were 0.67, 2, 15, and 75 μM; to avoid significant sample over-loading of the column when using high concentrations of Ancα/β, samples were diluted during quenching to give an injection quantity of ~15 pmol. To ensure back-exchange occurred equally across all diluted samples, the final ratio of H_2O:D_2O after quenching was kept constant at 54:46 and the pH of the quench buffer adjusted to pH 2.5. This allowed for all concentrations to be compared without correcting for back-exchange. All automated peptide assignments were manually verified, with noisy and overlapping spectra discarded. After processing a sequence coverage of 91% was achieved with a redundancy of 5.3.

Statistical comparison of peptides. For each peptide in the dilution experiment, The difference in deuterium uptake between different conditions was normalized by dividing the difference by the absolute uptake in the dimeric condition (75 μM). In Fig. 2.2c, Peptides that incorporated deuterons in the monomeric condition at quantities statistically indistinguishable from zero (P<0.01) were excluded. For peptide locations and alternative normalization methods, see EFig. A6-7. A permutation test was used to determine if relative deuterium uptake by residues at IF1 (or IF2) was significantly different from that of other residues. To eliminate statistical non-independence arising from the fact that many peptides overlap, we constructed a non-overlapping peptide set by subsampling without replacement from the total set of peptides, requiring that selected peptides do not share any residues. 1000 such non-overlapping peptide sets were constructed, and a p-value was estimated for each set using the following permutation test. Peptides in the nonoverlapping set were partitioned into those

containing residues mapping to IF1 and those containing no IF1 residues; a similar approach was used to test for a difference between peptides containing IF2 residues and those containing none; peptides containing residues contributing to both interfaces were excluded. The mean of the measured relative uptake difference over peptides in each partition was calculated, and the difference between the means of the two partitions was determined. A null distribution was then estimated by randomly partitioning peptides in the nonoverlapping set into two categories (without changing the size of the categories) and calculating the difference in means between the two randomly permuted peptide partitions. The p-value was calculated as the proportion of random partitions in which the difference between peptide category means was greater than or equal to that of the difference for the empirical categories. Appendix A: EFig. A5 displays the distribution of p-values calculated this way for 1000 non-overlapping peptide sets. A interface category was identified as having significantly increased uptake if the mean p-value from this analysis was <0.05.

Homology models for Anc α/β IF1 and IF2. Structural modelling of the Ancα/β monomer was performed using SWISS-MODEL. A deoxy structure of an Hbα monomer contained in recombinantly expressed human hemoglobin (1A3N) was used as the template. Hbα was used because its sequence similarity to Ancα/β is greater than that of any other extant globin. Further, both Hbα and Ancα/β form homodimers in isolation, unlike Hbβ (which is a mixture of dimers and tetramers at similar concentrations) or myoglobin. EMBO PISA (Krissinel and Henrick 2007) was used to identify sites in 1A3N subunits that buried >50% of their surface area at the interfaces or formed intersubunit hydrogen-bonding or salt bridge contacts at either IF1 or IF2. The HADDOCK 2.2 webserver was used to dock two Ancα/β monomers along an IF1

or an IF2 orientation by specifying the corresponding homologous residues (1a3n). The best scoring docked complex was used for all subsequent analyses and visualizations.

Homology models, interface burial, and contact maps for Ancα+Ancβ and Anc α/β14. Structural modelling was performed using SWISS-MODEL. A deoxy structure of recombinantly expressed human hemoglobin (PDB 1A3N) was used as the template for Ancα+Ancβ and for Ancα/β14 +Ancα. The extant Hbα and Hbβ were used as templates because they have higher sequence identity to to Ancα and Ancα/β14, respectively, than any other globin paralogs. EMBO PISA was used to estimate residue burial at the interfaces and predict hydrogen bonds across interfaces. Residues were classified as contributing to an interface if its solvent-accessible surface area was reduced by >10% in the assembled form relative to the nonassembled form. Van der waals contacts were identified as pairs of cross-interface atoms with center-to-center distances <3.5, using a custom script. PyMOL v4.19 was used to visualize and render protein structures. The similarity of interfaces in the homology model to those in X-ray crystal structures of extant hemoglobins was assessed by aligning the Ancα/β14 +Ancα tetramer to Hb of human (1A3N) and rainbow trout (*Oncorhynchus mykiss* 2R1H) (Appendix A: EFig. A10).

Data and code availability. Reconstructed ancestral sequences have been deposited in Genbank (IDs MT079112, MT079113, MT079114, MT079115). Alignment and inferred phylogeny, raw mass spectra, oxygen-binding data, and homology model coordinates have been deposited at https://doi.org/10.5061/dryad.w0vt4b8mx. HDX-MS data are available through doi: 10.5287/bodleian:5zRrdMB7E. Scripts for analysis for the HDX permutation

analysis and identification of contacts between subunits in modeled structures have been deposited at https://github.com/JoeThorntonLab/Hb_evolution.

Acknowledgments. We thank Chandrashekar Natarajan for technical advice and the Hb co-expression plasmid and members of the Thornton Lab for technical advice and comments on the manuscript. Supported by NIH R01-GM131128 and R01-GM121931 (JWT), NIH R01-HL087216 and NSF OIA-1736249 (JFS), NIH T32-GM007197 (CRC), a Chicago Fellowship (GKAH), BBSRC BB/L017067/1 and Waters Corp. (JLPB).

Author contributions. AP identified and developed the model system. AP, GH, and JT coordinated the project, interpreted the data, and led writing of the manuscript. AP performed and interpreted phylogenetic analyses and biochemical assays. AS and JS performed and interpreted oxygen binding assays. YG and AL performed and interpreted native mass spectrometry experiments. SC and JB performed and interpreted HDX experiments. CC performed and interpreted biochemical assays. All authors contributed to writing the manuscript.

Competing interests. The authors declare no competing interests.

Chapter 3: Contingency and specificity in the evolution of a protein complex

Abstract

The formation of stable and specific interfaces between proteins is key to virtually all biological processes in the cell. The genetic mechanisms that allow such interfaces to arise and mediate specific binding to other proteins remains an open question. To address this, we genetically dissected the historical substitutions that built the interfaces of a model multimer: vertebrate hemoglobin (Hb), a tetrameric complex made up of paralogous α and β globin subunits. Hemoglobin's assembly into a 2α:2β tetramer is mediated by two structurally distinct heteromeric subunit interfaces. Hb's subunits are descended from an ancestrally homodimeric protein that evolved to assemble into tetramers by acquiring a new interface after the duplication event that yielded the α and β subunits. To address the mechanisms by which multimerization and specificity arose during this interval, we pursued three objectives: (1) isolation of the minimal mutational set responsible for the evolution of tetramerization, (2) determining if the evolution of the derived interface was biophysically contingent upon the prior evolution of homodimerization (3) identifying the mutations and mechanisms by which the α and β chains evolved to bind one another rather than themselves, despite both chains inheriting a homomeric interaction from their duplication ancestor. We show that a single substitution is both necessary and sufficient for the evolution of a micro-molar affinity dimer-dimer interface that holds the Hb tetramer together. The efficacy of this mutation in yielding a protein-protein interaction is, however, historically and biophysically contingent on the earlier evolution of dimer interface. We demonstrate that five β substitutions at the ancestral homodimeric interface are sufficient to confer specific assembly into heterodimers.

The energetic preference for the heteromer over the potential homomers evolved through the optimization of heteromeric interaction, rather that through weakening of the α-α and β-β off-target interactions. These findings indicate that although the formation and specification of tight interfaces during evolution may be achieved through one or a small number of affinity-building changes, the effects of these trigger mutations are highly contingent upon the sequence and the quaternary structure of the protein backgrounds in which they arise.

3.1 Introduction

Most cellular processes rely on the assembly and activity of multi-subunit protein complexes (Goodsell and Olson 2000; Ahnert et al. 2015; Marsh and Teichmann 2015). Multimeric assembly is ubiquitous among proteins, and is often mediated through protein-protein interfaces that form strong and specific interactions. These (generally) non-covalent interactions involve many residues with complementary electro-static and steric properties. Identifying the genetic mechanisms (Pillai et al. 2020) by which such stable protein-protein interfaces arise during evolution and mediate subunit assembly into specific stoichiometries is a key question for both molecular evolution and biochemistry (Laub 2016; Pillai et al. 2020). In this paper, we use vertebrate Hemoglobin (Hb) as a system for exploring the number, effect sizes and structural mechanisms of mutations responsible for creating both affinity and specificity at an interface. We further investigate the genetic and structural preconditions that allow such interfaces to be accessible during history (Pillai et al. 2020).

Hb is an ideal system for exploring how novel and specific interfaces could arise during evolution (Pillai et al. 2020). Hb is a tetramer of paralogous globin subunits, α and β, that associate in a 2:2 ratio via two heteromeric interfaces that are both stable, specific (Perutz et al. 1960) and originated in a historical interval that is now well-characterized (Perutz et al. 1960) (**Fig 3.1a,b**). The tetramer-forming globin subunits belongs to a larger clade of vertebrate-specific globin-genes, including myoglobin, Globin E and Globin Y that are monomeric (Pillai et al. 2020). The two subunits are the result of gnathostome-specific gene duplication that occurred 450 million years ago, prior to the divergence of cartilaginous and bony fishes (Schwarze et al. 2014; Hoffmann, Opazo, and Storz 2011), but after the split between gnathostomes and agnathans. Previous work used ancestral sequence reconstruction to resurrect and structurally characterize ancestral proteins from the interval where Hb evolved its tetrameric architecture (Pillai et al. 2020). This work showed that hemoglobin evolved from an ancestral homodimer (**Fig 3.1d**) that existed prior to the gene-duplication that produced the α and β genes, which we term Ancα/β. This gene evolved, in turn, from an earlier monomeric protein, AncMH: the ancestor of Hb and Mb. The evolution of the complex therefore involved gaining two interfaces sequentially, which we shall call IF1 (Interface #1, or canonically the α1β2 interaction) and IF2 (Interface #2, canonically the α1β2 interaction) (Fig 3.1b). IF1 arose before the duplication and mediated the assembly of monomers into a homodimer (Fig 3.1a). IF2 arose after the duplication and, by the time of the ancestral gnathostome, allowed specific assembly between the ancestral α (Anc α) and ancestral β (Anc β) into heterotetramers (**Fig 3.1c**). Ancα, and its modern descendants, continue to self-assemble into homodimers, when expressed in isolation, retaining the same IF1 interface as Ancα/β (Kumar et al. 2014a)

(**Appendix B: Extended figure B1**). β chains, on the other hand, assemble into homotetramers when isolated from α (A.S. Pillai et al. 2020; Kidd et al. 2002). In this paper, we address three key questions pertaining to the emergence of multimeric protein complexes, using vertebrate hemoglobin as a model system.

First, we investigated the structural and biophysical effects of the mutations that generated the IF2 interaction. In Hb, it was previously shown that two substitutions on the β branch were sufficient to build a micromolar-strength tetramer. One of these mutations, t37V, straddles the IF1 and IF2 interfaces – with nearly all of the surface area of the valine side chain buried in the former - while the other sits directly at IF2 and slots into a hydrophobic cleft on the opposing surface (q40W) (**Fig 3.1e**). In this work we disentangle the individual effects of these substitutions, and measure the extent to which their effects are pleiotropic across interfaces, or restricted to just one. This section of the work reveals whether or not a single historical substitution could deliver a stable interface or if synergistic epistasis between the substitutions was necessary to construct it.

Second, we investigated whether the interfaces that hold the tetramer together could have arisen in any order, or if the IF2 interface was historically contingent on the formation of an older dimeric interface. We pursued this question by reverting the evolutionary changes that created IF1 in Ancα/β, and then determining if IF2 remains genetically accessible via one or both of the mutations described above. We also mutationally disrupted IF1 by other means and tested to see if IF2 was still competent to form interactions. The existence of such a dependency would imply that the mutational accessibility of useful interfaces can be facilitated or stopped by the previous evolution of other non-overlapping interactions: a form of historical

contingency that reduces the repeatability of structural evolution. We outline the possible mechanistic causes for this dependency.

In addition to its implications for repeatability during evolution, this kind of dependent accessibility could explain why some complexes exhibit ordered assembly in vivo. There are two plausible evolutionary explanations for why the interfaces in a complex may be biophysically dependent upon one another, such that certain subcomplexes must form before the final stoichiometry is attained, as has been observed in a number of complexes (Levy et al. 2008; Rohl and Nierhaus 1982; Peterson et al. 2018). Assembly order could evolve in ancestral complexes that do not initially exhibit any order in interface formation; in such a case, order could arise because it maximizes the speed of assembly (Marsh et al. 2013), facilitates regulation, or because mutations fixed by drift that favored a particular assembly sequence. Alternatively, it could also have arisen because it was intrinsic to the mutations that initially created the interaction, rather than something that was secondarily gained during evolution.

Lastly, we investigated how hemoglobin evolved to assemble specifically into a $\alpha_2\beta_2$ heterotetramer rather than any of the possible off-target homomeric complexes (Kidd et al. 2002; Kumar et al. 2014). Specificity appears to be encoded at the very first step of Hb's assembly, where monomers associate specifically into αβ dimers. We addressed how ancestrally homomeric IF1 evolved to mediate a specific αβ heteromeric interaction (Siddiq, Hochberg, and Thornton 2017; Fersht et al. 1985; Pereira-Leal et al. 2007) between α and β subunits, by identifying historical mutations that occurred after the α/β duplication that prevented assembly into non-target dimers: namely, α_2 and β_2 homodimers. Previous work spanning many protein families has shown that ancestrally homomeric interfaces frequently

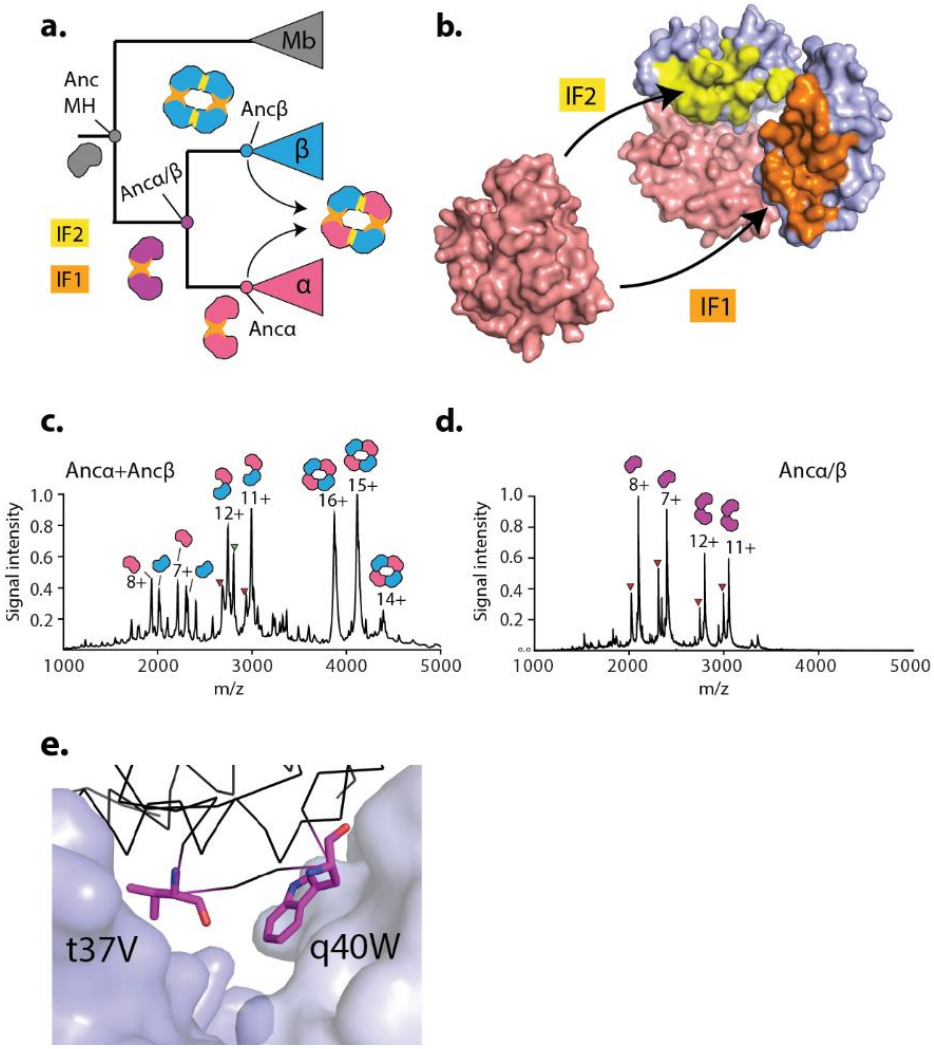


Figure 3.1. Evolution of the stoichiometry of vertebrate hemoglobin, a. Evolution of Hemoglobin's interfaces. Circles, reconstructed ancestral proteins; Icons, oligomeric states. Yellow and Orange surfaces, IF2 and IF1 respectively. **b.** Structure of extant hemoglobin (2qsp) with distinct heteromeric interfaces shown as yellow (IF2) and orange (IF1) surfaces. Blue subunits, β chain ; Pink subunits, α chain. Single α subunit has been separated from the complex to allow visualization of the interfaces. **c, d.** Native mass spectrometry on Ancα+Ancβ and Ancα/β. Icons, oligomeric states associated with each peak series. Charge states for each major peak shown. Red triangles, heme-bound variants. Green triangle, variant with 258 Da adduct attached. **e.** Causal mutations for tetramerization shown in a structural homology model of Ancα/B2. Purple sticks, V37 and 40W residues. Ribbon, backbone of one subunit. Blue surfaces, two receiving subunits.

evolve to support heteromeric interactions between paralogs (Pereira-Leal et al. 2007). It remains an open question, in biophysical terms, how this sort of specificity arises among structurally similar subunits during evolution. Does it emerge simply through mutations that selectively weaken of the homomeric off-products, or they typically optimize the heteromeric interaction? Do mutations that change the affinity of the heteromer, necessarily change the affinity of the homomers? To answer this question, we measured and compared the effects of historical mutations on the affinities of both heteromers and homomers.

3.2 Results

Identification of a historical assembly-triggering substitution. In order to identify the minimal possible genetic cause for the evolution of tetramerization in Hb, we isolated the substitutions that occurred at the IF2 surface after the duplication of Ancα/β, which were necessary for the formation of the dimer-dimer interaction. These substitutions occurred on the β branch, while the α surface remained static (A.S. Pillai et al. 2020a). Previous work identified two historical IF2 substitutions clustered on the C-helix of Hb that confer assembly into a homotetramer when introduced into the ancestral homodimer (Ancα/β2, **Fig. 3.1d, 2a**). One of these mutations (q40W) is situated at tetrameric IF2, while the other resides at the junction of the two interfaces (v37T) (**Fig. 3.1e**).

To study the individual thermodynamic effects of these mutations on tetramerization, we used site-directed mutagenesis to introduce each of them separately into ancestral α/β and then measured the occupancy of the tetrameric form by these mutants using both nMS and size-exclusion chromatography (**Fig. 3.2a, b, e**). We found that introducing q40W by itself is

sufficient to induce assembly into tetramers, while v37T seems to have little or no direct influence on dimer-to-tetramer affinity by nMS (**Fig. 3.2a,e**). To test the extent to which the W40 residue was necessary for tetramer-formation by the time of Ancα + Ancβ, we mutationally reverted the β tryptophan in the derived heterotetramer back to its ancestral state in Ancα/β. The resulting mixture of Ancα + AncβW38Q is entirely dimeric, with no detectable tetramer signal at 100 μM by Size exclusion chromatography (SEC) (**Fig. 3.2c**). These results demonstrate that a single β change is both sufficient and necessary for the evolution of a tetramer from a dimer (**Fig. 3.2b**). The tetramer-dimer and monomer-dimer affinities of both point-mutants was measured by estimating the proportion of tetramer, dimer and monomer at successive protein concentrations for each construct by nMS and fitting a dissociation constant. Q40W generates an affinity at IF2 (Kd = 10μM) comparable to human hemoglobin (Pillai et al. 2020) (13μM) as well as Ancα+Ancβ (9 μM), but significantly weaker than α/β2 (1.6 μM) (**Fig. 3.2f**).

To further test if t37V induces even weak millimolar affinity at IF2, we measured its oligomeric size at 1 mM using SEC – a method that is amenable to analyzing stoichiometry at much higher concentrations than nMS. We find no evidence for significant tetramer formation from this experiment even at these very high concentrations (**Fig. 3.2d**). However, t37V does appreciably increase the affinity of IF1, reducing the Kd of the interaction by more than two orders of magnitude (**Fig. 3.2e**). Additionally, although the single mutant does not itself contribute to IF2, it does induce a significant gain in tetramer-dimer affinity (6-fold change in Kd) when introduced in combination with q40W. (**Fig. 3.2f**)

Our dissection of the IF2 changes demonstrates that a single substitution to a side chain bearing a bulky aromatic ring is sufficient to generate a protein-protein of substantial strength. Despite this dramatic structural effect and its proximity to IF1, the q40W substitution does not impinge on the affinity of the old interface, which stays virtually identical to its value in Ancα/β (~9 μM) (**Fig. 3.2e**). t37V does not build IF2 when introduced singly, but nonetheless stabilizes the tetrameric architecture when coupled with q40W in the Ancα/β2 construct in two ways: (1) It increases the affinity of IF1, thereby increasing the pool of dimers that are available to tetramerize, so that the a/b tetramer reaches 50% molar occupancy at protein of concentration of 1 μM rather than 15 μM as in t37V (2) it offers an energetic boost to the dimer-tetramer interface, by indirectly stabilizing interactions there created by q40W.

Order of the two interfaces. To determine whether or not IF2's evolution was contingent on the earlier emergence of IF1, we introduced the IF2-forming mutations identified in the previous senction into ancestral and engineered protein backgrounds that do not display IF1 mediated assembly and determined if IF2 can still form. First, we introduced both of the tetramer-stabilizing mutations v37T and q40W into ancMH (AncMH2), the monomeric common ancestor of myoglobin and hemoglobin, which existed before ancα/β. If IF2 is not contingent on the substitutions that occurred between AncMH and Ancα/β, including those that built IF1, then these substitutions would be sufficient to induce ancMH to form IF2-mediated dimers. By nMS, we show that AncMH2 remains monomeric and does not form IF2 dimers, showing that changes in the interval between ancMH and ancα/β were necessary before the IF2 mutations could induce assembly (**Fig. 3.3a,b**), although it does not reveal if these "potentiating" mutations occurred at IF2, IF1 or outside of these surfaces entirely.

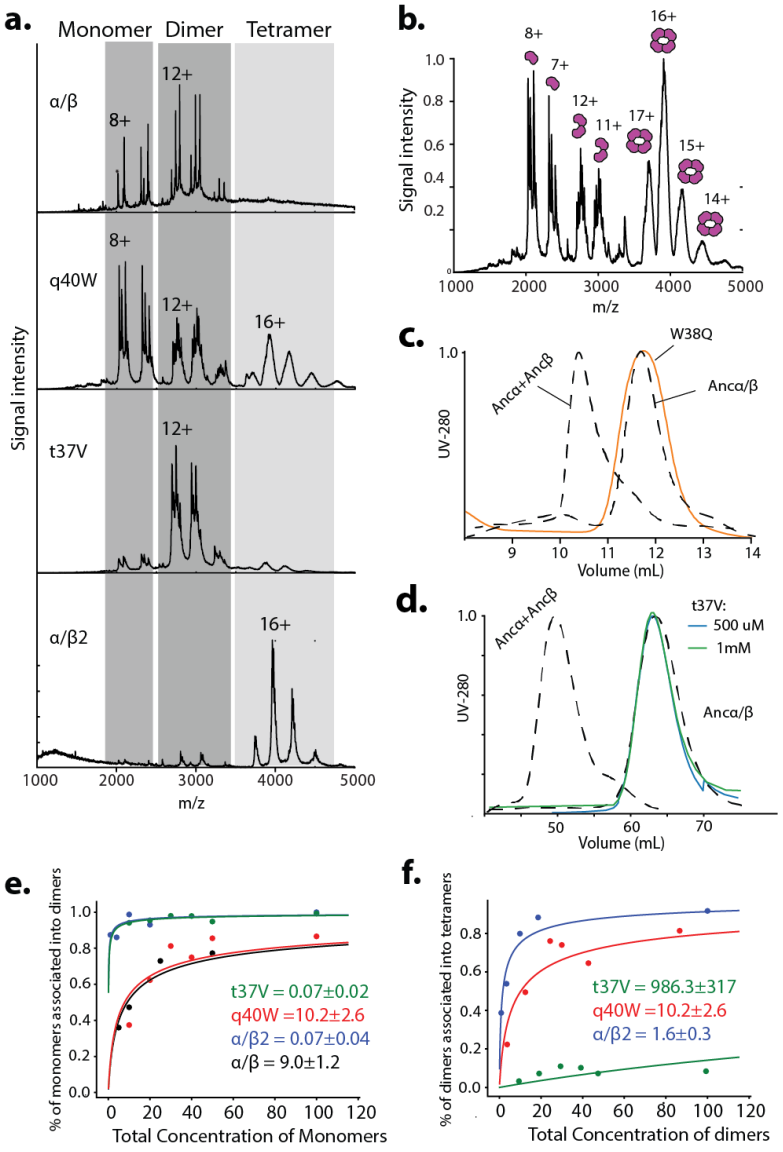


Figure 3.2. Genetic dissection of Ancα/β2 a. nMS on Ancα/β, q40W, t37V and Ancα/β2. Shading indicates peak series associated with monomer, dimers and tetramers. Charge states associated with major peaks is shown. All spectra were collected at 20 µM **b.** nMS on q40W at 50 µM. Icons, oligomeric states associated with each peak. Charge-states for each peak are shown. **c.** Size exclusion chromatogram of dimeric AncA/β, tetrameric Ancα+Ancβ (dashed lines) and W40Q at 100 µM. **d.** Size exclusion chromatogram of dimeric Ancα/β, Ancα+Ancβ (dashed) and t37V at 0.5 mM (blue) and 1mM (green). **e.** Monomer-dimer affinity for AncA/β, q40W, t37V and Ancα/β2. Circles, fraction of monomers assembled into dimers; Kd values and 95% confidence intervals estimated from these data by non-linear regression are shown for each construct. **f.** dimer-tetramer affinity for q40W, t37V and Ancα/β2. Circles, fraction of dimers assembled into tetramers; Kd values and 95% confidence intervals estimated from these data by non-linear regression are shown for each construct.

To determine if the IF1 substitutions along the AncMH-Ancα/β interval were necessary for the later formation of IF2, we first reverted the states in Ancα/β at IF1 back to their ancestral state in ancMH. This revertant protein (Ancα/β_IF1rev) is a monomer, as was previously demonstrated, confirming that sequence changes at IF1 between AncMH and Ancα/β were necessary for its formation (**Fig. 3.3e**) (Pillai et al. 2020). We then introduced mutations that build affinity at IF2 into this background; if these mutations induce dimerization via IF2, then this is clear evidence that IF2 can emerge without the prior evolution of IF1. If they do not, then some or all of the IF1 mutations are specifically necessary for IF2 to be accessible. We show that introducing q40W singly or in combination with v37t (**Fig. 3.3d**) into this background was not sufficient to yield dimers via IF2, despite the fact that these mutations can build IF2 once IF1 evolves. These experiments demonstrate that the accessibility of IF2 is shaped by the previous evolution of IF1.

To further test the dependency of the tetramer-forming substitutions on the prior evolution of IF1, we mutationally disrupted IF1 and measured the occupancy of IF2-mediated dimers. To break IF1, we introduced a mutation, P127R, at a site that remains unchanged during the MH-α/β interval, which was previously shown to break the IF1 interaction (by clashing with an opposing arginine at site 33) without substantially disrupting the tertiary structure of the protein (Bunn 2019; A.S. Pillai et al. 2020) (**Fig. 3.3c**). Introducing the triggering IF2 mutations, t37V and q40W, into this background does not yield dimers by nMS. This indicates that physically breaking IF1 is sufficient to block the biophysical accessibility of IF2 (**Fig. 3.3c**).

Does the evolution of IF2 remain accessible at later points in history? The descendant gene Ancα, as well as modern day Hb-α proteins continue to form homodimers like anc α/β, albeit

with an apparent loss in self-affinity, while being incapable of self-assembling into tetramers. To test the accessibility of the tetramer at this subsequent point in history, we introduced the IF2-forming mutations into Ancα, and demonstrate that these changes remain sufficient to induce assembly into tetramers at 20 μM. These experiments show that tetramer formation remains accessible even at this later point in history, despite a loss of affinity at the IF1 interface (**Appendix B: Extended Figure B1**).

Taken collectively, these experiments demonstrate that the formation of one interface is both historically and biophysically dependent on the formation of another.

Evolution of specificity at IF1. Lastly, we investigated how the α and β proteins evolved to interact specifically with one another rather than with themselves. We did this by measuring the effects that substitutions on these branches had on both affinity for self and the other subunit. We focused our genetic and biophysical experiments on the evolution of specificity at IF1 (which mediates the first step of assembly into dimers), rather than IF2 for several reasons. First, previous work showed that the post-duplication IF2 substitutions on β are collectively not sufficient to deliver specific assembly into heterotetramers when recombinantly co-expressed with α in *E coli.*, generating instead a mixture of αβ3, α2β2 and β4 forms (Pillai et al. 2020). Secondly, when Ancα+Ancβ is diluted until it populates almost entirely the dimeric state, there are only trace quantities of alternative dimers observed by nMS – either α_2 nor β_2 – indicating that the heterodimer is specific (**Fig. 3.4a**). Assuming that IF2 cannot form between monomers, as the work shown in the previous section suggests, this would indicate that IF1 is specific. Thirdly, the evolution of specificity at the ancestral IF1 is both necessary and theoretically sufficient for assembly into heterotetramers. To illustrate this, we performed a

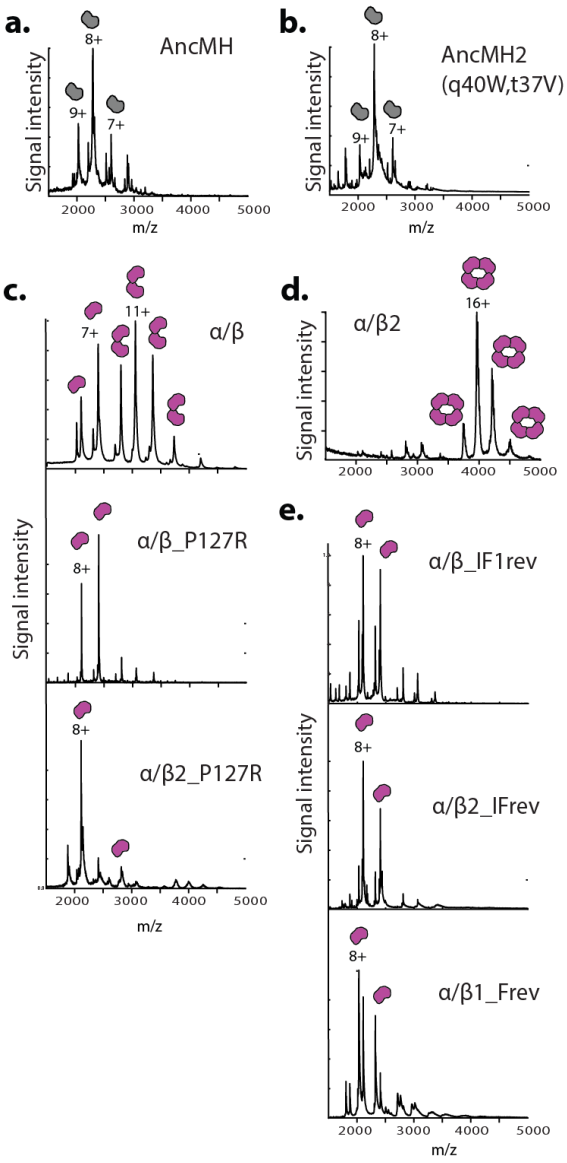


Figure 3.3. Genetic dissection of Ancα/β2 a, b. nMS on AncMH and AncMH2. Icons, oligomeric state. Charge states of peaks shown. **c,** nMS on ancα/β (top), P127R (middle) and Ancα/β2_P127R. Icons, oligomeric state. A single charge state is shown for each charge series. All spectra collected at 20 µM. **d.** nMS on anc α/β2. Icons, oligomeric state. Single charge state for tetramer-series is shown. **e.** nMS on ancα/β_IF1rev, ancα/β2_IF1rev, q40W_IF1rev. Icons, oligomeric state. A single charge state is shown for each charge series. All spectra collected at 20 µM.

theoretical calculation of the relative molar fraction of heterotetramers expected to form under three scenarios – one where IF1 is the only specific interface, IF2 is the only specific interface and where both interfaces are specific. The subunit assembly equilibria utilized to make these calculations are shown in **Fig. 3.4b**. In a scenario in which the descendant α and β proteins (each at 50 μM) retain the ancestral affinity of ancα/β and remain non-specific at IF1, while a specific heteromeric IF2 arises, with tetramer-dimer affinity comparable to Anc α +Ancβ (9 μM), the resultant mixture is predicted to contain only <43% Heterotetramers (remaining fraction includes homotetramers, dimers and monomers) if only IF2 were specific, as opposed to the 66% Heterotetrameric fraction expected if both interfaces were specific. On the other hand, increasing specificity at IF1 alone allows the heterotetramer fraction in mixture to asymptotically approach 66%. (**Fig. 3.4b,c**), demonstrating that IF1 is theoretically sufficient for assembly. Finally, probing affinity at IF1 is technically tractable, because this interface can be reconstituted by mixing separately expressed subunits at defined concentrations, unlike IF2, which does not form without co-expression and/or co-folding of the two subunits. This allows us to perform simple titration series and quantitatively estimate and compare the strengths of heteromeric and homomeric versions of IF1.

We determined how IF1-mediated heterodimer specificity arose by measuring the effects on specificity of substitutions that occurred after the α-β gene duplication on both descendant branches. Hetero-specificity at IF1 could have arisen in three ways: through the weakening of self-interactions, through the optimization of the heteromeric interactions, or some combination of the two processes.

To test whether or not specificity could have arisen solely through weakening the off-target assemblies, we first measured the affinity of independently expressed β and α homodimers by nMS. α homomers (Kd = 21 μM) are significantly weaker than α/β dimers (Kd = 9 μM), which is consistent with the fact that the deletion of the D-helix (a synapomorphy of the α clade) removes a small set of contacts at IF1 **(Fig. 3.5a)**. β homomers, on the other hand, are significantly stronger than their ancestor (Kd = 0.72, 6-fold decrease in Kd) (**Fig. 3.5a**). This strengthening could be caused – in part – by the substitution V37t, which was shown to increase dimer affinity in the background of ancAlpha/β. These results indicate that specificity in this system must have relied on the optimization of the interaction between the two chains, rather than simply the dual weakening of self-interactions of the homomers, since one of the descendant subunits became a stronger homomer than anc α/β.

To quantify the affinity of Anc α for Anc β, we performed an nMS titration experiment. We measured the strength of the heterodimer interaction relative to the homomers by mixing increasing concentrations of the Anc α subunit with 50 μM of the Anc β subunit, estimating the molar proportions of dimer forms in each mixture by nMS, and obtaining the best-fit heterodimeric Kd to these observed proportions by nonlinear regression. If there were no subunit specificity, we would expect 50% of the signal in an equimolar mixture of subunits to come from homomeric complexes. At equimolar concentrations (50 μM each), the fraction of homomeric off-pathway complexes in a mixture of Anc α and Anc β is small (<19% total) (**Fig. 3.4d**). The heteromeric affinity obtained from the fitting procedure indicates that the IF1-affinity of Anc α for Anc β (Kd = 0.4+/-0.04 μM) is stronger than either homomeric interaction (**Fig. 3.5a**), as well as the ancestral α/β interaction. This experiment demonstrates that the

emergence of heteromeric specificity did involve the improvement – in absolute terms - of the heteromeric interaction over the ancestral homomeric one. We use the specificity observed in the αβ heterodimers as a benchmark when comparing the specificities other combinations of ancestral and engineered proteins.

To identify the impact of the interfacial substitutions on the evolution of IF1, we introduced all 5 IF1 substitutions that occurred after the duplication into ancα/β. This engineered protein was mixed with ancα/β in equimolar quantities and the relative proportions of different dimeric species was estimated by nMS. Since ancα/β and ancα/β_IF1 are very close in their masses, we added an N-terminal hexahistidine-tag to the ancα/β subunit to help clearly resolve the masses of the dimers – the N-terminus of Hb does not participate in interactions at IF1 and we show that this tag does not compromise the capacity of Ancα/β to dimerize (**Appendix B: Extended Figure B1c**). Ancα/β and ancα/β_IF1 co-assemble to yield a mixture dominated by heterodimers, with <22% of the dimer signal coming from the two types of homomers (**Fig. 3.4e**). The strength of this heterodimer – as estimated from a titration series - is comparable to Anc α- Anc β heterodimers (**Fig. 3.5a**). These interface mutations also increase the strength of the β-β homodimer (Kd = 1.43 μM) relative to the ancestral α/β, by a Kd-factor of over 6 – likely in part because they include the affinity-enhancing t37V mutation described earlier (**Fig. 3.5a**). Thus, the five interfacial substitutions along the β lineage are sufficient to recapitulate the evolution of heterospecificity. This implies that, after the duplication, a few mutations on only one of the descendant genes was sufficient to produce hetero-specific assembly – no coevolution between the genes was strictly necessary to produce a complementary interface.

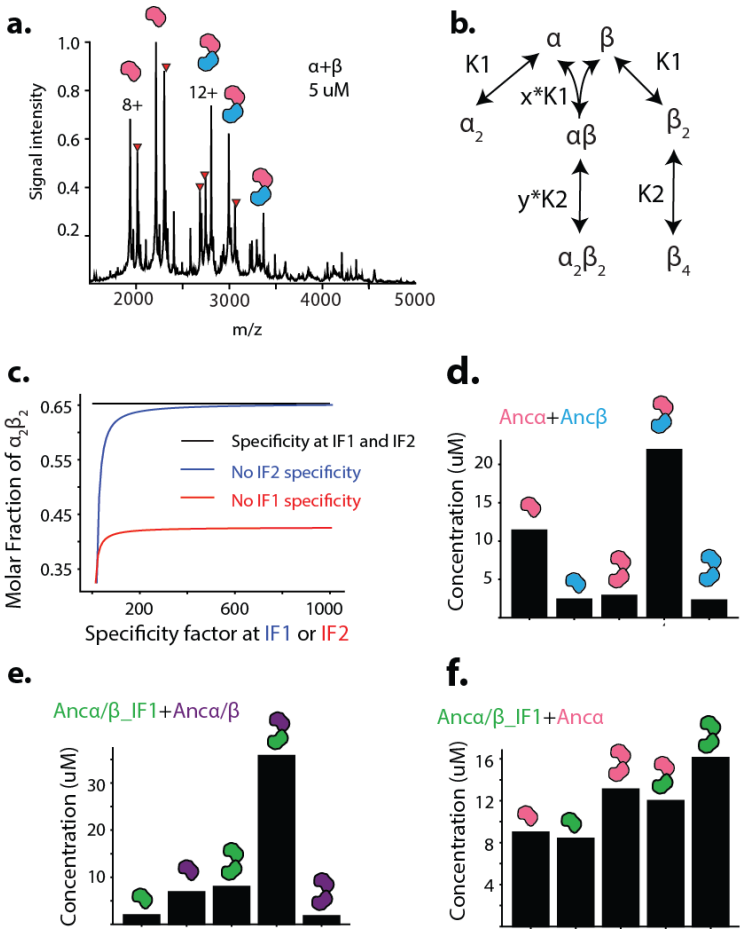


Figure 3.4. Evolution of specificity in hemoglobin a. nMS of AncA+AncB at 5 µM, icons, oligomeric states associated with each peak. Red triangles, heme variants. A single charge state is shown for each charge series. **b.** Assembly pathway of α and β subunits with oligomer equilibria and dissociation constant variables used in the specificity model displayed. In our model, specificity at IF1 and IF2 were quantified respectively as x, the fold difference between the heterodimer equilibria and the two homomer equilibria (which are assumed to be equal; K1) and y, fold difference between the two dimer-tetramer equilibria. **c,** Effect of thermodynamic specificity on heterotetramer occupancy at 50 µM. Lines, fraction of heterotetramers in solution when the specificity of IF1 (blue) and IF2 (red) are separately increased. Black, heterotetramer fraction when both IF1 and IF2 are highly specific (x=y=1000). **d.** Concentrations of different oligomers formed when isolated ancα and ancβ are mixed at 50 µM each. Icons, oligomeric state. **e.** Concentrations of oligomers formed when isolated ancα/β and ancα/β_IF1 are mixed at 50 µM each. Icons, oligomeric state. **f.** Concentrations of oligomers formed when isolated ancα and ancα/β_IF1 are mixed at 50 µM each. Icons, oligomeric state.

Although the IF1 β substitutions may have been sufficient to account for heterospecificity, they may not encapsulate the full array of changes that affected specificity after the duplication of Ancα/β. Sequence changes also occurred on the α lineage, as well as outside of the β IF1; and subsets of these may have contributed to or diminished specificity. Since Ancα/β_IF1+Ancα/β and Ancβ +Ancα show comparable heterospecificity, the effects of these mutations must either be individually negligible or compensate for one another such that their net effect on specificity is effectively zero. Additionally, these mutational effects may have had additive effects on affinity or alternatively, they may have been epistatically dependent on mutations at other sites (including sites in their binding partner).

To determine how the α changes affected heteromerization, we mixed α with anc α/β and then with Ancα/β_IF1. In the first case, α combines with Ancα/β to yield heteromers that are stronger than either homomeric interaction (Kd = 4 μM): 5x smaller Kd with respect to Ancα homodimers and 2x with respect to Ancα/β_IF1 homodimers, but still weaker than the Ancα+Ancβ heterodimers (**Fig. 3.5a**). The five IF1 β mutations are not, however, sufficient to confer specific assembly with α, in contrast to their effect on binding ancα/β. Ancα/β_IF1 prefers binding to itself over Ancα (5 μM), which is less comparatively stable (**Fig. 3.5a**). At equimolar concentrations, nMS signal from Ancα/β_IF1 homomers actually exceeds that from the other two dimers when it is mixed with Ancα (**Fig. 3.4f**). Since Ancα+Ancβ forms specific heterodimers, the implication of this is that the non-IF1 β substitutions must confer added affinity for α. This demonstrates that residues outside of an interface can contribute to the evolution of specificity. Further, it shows that the emergence of a specific interface by the

Ancα+Ancβ ancestor did involve mutations with compensatory effects on both subunits, and that this compensation involves substitutions outside of the interface. Additionally, the α mutations raise the affinity of the heterodimer in one context, but they diminish specificity in another, demonstrating that intergenic epistasis can play a role in the early genesis of a specific interaction.

3.3 Discussion

Our finding that a micro-molar affinity interface could have originated through a single historical substitution adds additional evidence to the notion that new and stable oligomers are highly accessible during natural evolution. This finding is consistent with previous protein engineering and directed evolution studies that have discovered synthetic single-amino acid mutations that cause protein assembly into symmetrical complexes or fibers. The energetic threshold for forming an oligomer in vivo is further eased in cases, like Hemoglobin, where high cellular concentrations (~1 mM) ensure that even a weak micro-molar interface might be sufficient for high occupancy of the new oligomer in vivo. Additionally, the easy accessibility of higher-order assembly implies that it could arise and be maintained under a range of possible scenarios other than gradual selection and optimization, as has sometimes been assumed for complicated multi-residue features, including drift and episodic selection. In population terms, the tetrameric architecture could have arisen from “biochemical scratch” through the fixation of a single allele, implying that (1) multimers can arise rapidly during the course of evolution, including on microevolutionary time-scales, and (2) a derived, stable higher-order symmetrical oligomer could coexist with a monomer as a single-amino-acid protein polymorphism within the same population. To our knowledge, in known cases of oligomeric variation within a

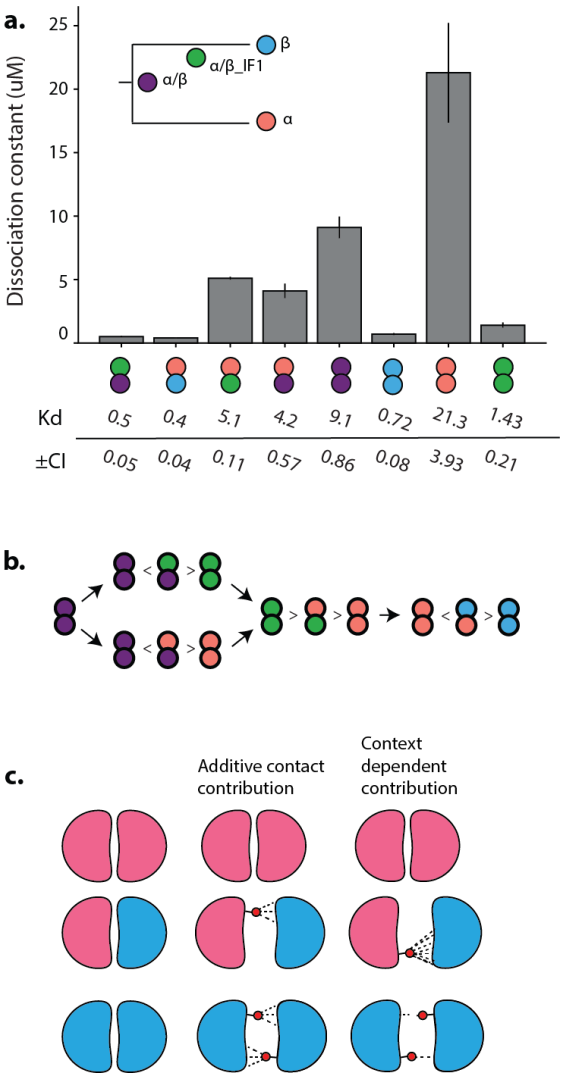


Figure 3.5. Mechanisms of the evolution of specificity a. Monomer-Dimer affinities of homodimers and heterodimers. Kd is estimated by non-linear regression from measurements of dimer occupancy. These were in turn estimated from 4-6 nMS spectra of subunits mixed at concentrations between 10 µM and 50 µM. Bars, Kd values; Error bars, 95% confidence intervals on the parameter estimates. Inset: phylogeny indicating genotypes whose assembly was tested. Below: Numerical values of Kd and confidence interval **b.** The effects of mutations on specificity. Arrows indicate historical sets of substitutions. Inequality signs indicate relative strengths of interactions, as determined from Fig4a. **c.** Illustration of the effect of a contact forming residue change at a para-isologous interface. Left: 3 possible dimer species that existed prior to mutation. Middle: Scenario in which side-chain interactions are not conditional on the stoichiometry. Right: Scenario in which side-chain interactions are different between dimeric species. Red circle, mutant residue. Number of dashed lines indicates strength of contact.

population, the derived oligomer is either smaller in size (i.e. has lost an interface) or assembles into an open fibre (Pauling et al. 2019). The q/w change in the vertebrate lineage provides a historical instance of at least one such mutational trigger in the past. Once it arose, the tryptophan became a fixture of the interface in the subsequent 450 million years of its evolution; virtually invariant across vertebrates, from sharks to humans, and mutations away from that state became deleterious.

The observation that q40W and t37V have individual energetic impacts that are restricted to IF1 and IF2 respectively has implications for the evolutionary modularity of interfaces. The pre-existing interfaces of a complex do not necessarily need to be altered in order to accommodate the new interface, nor do large-effect mutations at the new interface have to modify affinity at the old. In Hb's case, this is true despite the fact that the two interfaces are close by and even overlap at one residue. If many such "modular" interface-strengthening mutations are available to a complex during its history, then the emergence of stable oligomerization is considerably simplified, since there would exist a large pool of affinity-building mutations that do not have negative pleiotropic effects on the old interface.

The dependency of IF2's origin on IF1 demonstrates that the interfaces in a complex are not wholly independently evolving units. The accessibility of an interaction during evolution can be shaped by the what interactions a protein already participates in. This sort of dependent origination of interfaces could account for why ordered assembly arises in protein systems. The in vivo dependency of one interface on another could arise because the effects of the mutations

that initially built it during evolution were biophysically dependent on the formation of the

earlier interface. The order in which interfaces form in the cell is then a simple consequence of the order in which they evolved (Marsh et al. 2013). This finding weighs against the alternative explanation, which is that order evolves in complexes that do not initially exhibit any assembly order, either by chance or because of selection for efficient assembly. Our results imply that it could instead be an intrinsic feature of assembly that existed since the moment of its inception, rather than having arisen secondarily.

What is the mechanism of IF2's biophysical dependency on IF1? A simple factor that could account for this dependency relates to isology and the symmetry of the tetramer-structure (Monod et al. 1965). Every contact at IF2 is presented 4 times in a symmetrical homotetramer and twice in a heterotetramer (Monod et al. 1965; Kidd et al. 2002). When IF1 is eliminated, these contacts are only presented twice in a putative IF2 homodimer (and just once in an IF2 heterodimer). This halving in the number of presented contacts per oligomer could maximally increase the effective Kd associated with the IF2 interaction by three orders of magnitude. The true loss of affinity may be less than this theoretical factor because there of substantially favorable entropy gain associated with dissociation into dimers. A direct implication of the multiplicative boost provided by the isologous presentation of contacts is that relatively weak interface-forming mutations could easily deliver new stoichiometries in proteins that already oligomerize through one interface – and the effect is further enhanced in tetramers, octamers etc. Additionally, this sort of dependency between interfaces could also allow, in some cases, for two interfaces to arise simultaneously through a single substitution– as the comparison of α/β2 and α/β2_P127R reveals. The removal of a steric clash might not only create an interface at the site of the mutation, but could theoretically reveal another incipient interface that was

too weak to form without the avidity contributed by formation of the first interface. An additional mechanistic explanation for dependency (that is not mutually exclusive to the first), is that the formation of IF1 induces conformational changes at IF2 that make it competent to form interactions, however further experiments are required to determine if this is the case.

With respect to the evolution of specificity, our results suggest that the genetic route from homodimerization to specific heterodimerization is short, requiring no more than five interfacial changes, and possibly fewer. Many heteromeric protein complexes are made up of paralogs are derived from ancestral homomers, and this work implies that cross-specificity between paralogs could arise quickly after duplication, consistent with previous work. Aside from identifying a subset of causal mutations, our results add crucial information as to how these mutations affect the stability of the off-target homomer assemblies, thereby providing biophysical insight into the mechanisms of specificity. The mutations that generate specificity α and β must perform at least one of two tasks: (1) they must collectively add favorable contacts to the heteromer and/or (2) introduce negative clashes into the homomers. The specificity-inducing β IF1 substitutions do the former, but not the latter. A structural homology model indicates that all of these mutations form contacts with opposing residues that remain unchanged in both α and β. This reveals that an evolutionary scenario involving coevolutionary tit-for-tat substitutions on two subunits is not necessary to create a specific interface between them. Instead, exploiting the existing side-chains on a protein surface appears sufficient. However, this history raises a question about how these IF1 mutations could have strengthened contacts at the IF1 interface of αβ, without having an even greater stabilizing effect in the β2 dimer due to isology – the symmetrical repetition of contacts in the homomer.

Although isology may assist in the evolution of strong homomeric interfaces by (maximally) doubling the energetic effect of each interaction-forming substitution (Monod et al. 1965), it can also act as a constraining factor in the evolution of hetero-specificity for the very same reason (**Fig. 3.5c**). Each positive contact in the IF1 mutational set is occurs twice in the homomer, and only once in the heteromer. Our nMS experiments show that heteromer is stronger than the homodimer despite this factor. These data suggest that the energetic contribution of an individual residue change must differ in the homomer and the heteromeric context, even though the residues it is surrounded by from the opposing subunit remain the same. The implication of Ancα/β _IF1+Ancα/β 's specificity is that at least some of the side-chain contacts formed in the heteromer are more favorable than their doubly presented counterparts in the homomer (**Fig. 3.5c**). This could arise because of small conformational differences in how the same contacts are presented in the three possible dimeric assemblies, such that hydrogen bonds end up being stronger or water better occluded in the heteromeric interface.

Lastly, our findings indicate that compensatory mutations and epistasis play a role in the emergence and maintenance of specificity. We show that the IF1 substitutions have background dependent effects on specificity. Immediately after the gene duplication, the IF1 substitutions would have been sufficient to induce heterospecificity if introduced into one of the duplicated genes, while the sister gene remained unchanged. However, the very same mutations compromise specificity if the other gene undergoes the changes to α. Thus, although coevolution between the subunits is not strictly necessary for the evolution of specificity, mutations on the cognate subunit can change not only the magnitude but the sign of the effect

of a mutation (**Fig. 3.5b**). Our finding that sequence changes outside of the IF1 surface on the β protein ultimately "compensate" for the diminished hetero-specificity of Ancα/β_IF1+Ancα, by raising the affinity of the heteromer-interface, indicates that substitutions outside of an interface can impact the evolution of specificity; selectively strengthening one assembly over the others (**Fig. 3.5b**). Such long-range effects of substitutions on interfaces have been observed in the evolution of other complexes (Perica et al. 2014; Marsh and Teichmann 2015). In summary, this data shows that, when studying the specific heteromeric interface in an extant complex, one cannot assume that (1) all of the side-chains involved in mediating interactions had historically positive contributions to specificity when they first arose, (2) that the effects of these substitutions remained unchanged as the sequence background changed during history (3) that the directly interacting side-chains represent the only relevant contributing factors to the evolution of its specificity.

This work contributes to the growing body of evidence that molecular assemblies and specificities can be generated though small sets of genetic changes during evolution. However, their accessibility through a given set of mutations may be highly conditional on earlier, distal structural features and genetic states fixing in a lineage. This property of accessibility coupled with biophysical dependency could explain why higher-order assembly appears to be highly evolvable but typically involves quite different structural encodings (Royer et al. 2005a) – i.e. different structural interfaces and contacting residues – when it arises independently in different lineages. Part of the explanation for this could be that the mutations that generate specific interfaces in one lineage may be inaccessible in another, because of their epistatic

dependency on structural features both within and outside of that interface.

3.4 Methods

Recombinant protein expression

Ancestral genes were codon-optimized for expression in Escherichia coli using CodonOpt and generated by de novo DNA synthesis (IDT gBlocks). For globin expression, coding sequences were cloned into a pLIC expression vector without affinity tags and expressed under a T7 polymerase promoter. For oxygen-affinity measurements, plasmid pCOMAP49, which expresses E. coli methionine aminopeptidase 1 (MAP1), was cotransformed to ensure efficient N-terminal methionine excision. For co-expression of two globins, sequences were expressed from a polycistronic operon in plasmid pGM, without tags and under a T7 promoter, separated by a spacer containing a stop codon and ribosome binding site. E. coli methionine aminopeptidase 1 (MAP1) was coexpressed from the same plasmid.

JM109 DE3 E. coli cells (NEB) were transformed and plated into solid Luria broth (LB) medium containing 50 μg/ml carbenicillin (and 50 μg/ml kanamycin, if pCOMAP was being cotransformed). A single colony was inoculated into 50 ml LB with appropriate antibiotics and grown overnight. Five millilitres of this culture was inoculated into a larger 500-ml LB culture. Cells were grown at 37 °C and shaken at 225 rpm in an incubator (New Brunswick 126) until they reached an optical density at 600 nm (OD600) of 0.4–0.6. The culture was then supplemented with 0.5 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) and 50 mg/l hemin (Sigma). After 4 h of expression at 37 °C, CO was bubbled through the solution for 10 min and

cells were collected by centrifugation at 5,000g. Protein purification was carried out immediately after expression.

Protein purification by ion exchange

Ancα/β, t37V, q40W, P127R, Ancα/β2, Ancα/β2_P127R, Ancα/β2_IF1, Ancα/β_IF1rev, Ancα/β1_IF1rev, Ancα/β2_IF1rev, Ancα and Ancα2 were purified using ion exchange chromatography20,49. All buffers were saturated with CO before purification and vacuum filtered through a 0.2 µM PFTE membrane (Omnipore) to remove particulates. After expression, cells were resuspended in 200 ml of 50 mM Tris (pH 6.8) with 2 cOMPLETE protease inhibitor tablets (Roche) and 0.5 mM DTT. The cell suspension was lysed in 50-ml batches in a glass beaker using an FB505 sonicator with a power setting of 90%, 1 s on/off for 2 min. The lysate was then centrifuged at 30,000g to eliminate cell debris, inclusion bodies and aggregates. The supernatant was further syringe-filtered used HPX Millex Durapore filters (Millipore). A HiTrap SP cation exchange (GE) column was attached to an FPLC system (AKTAprime plus) and equilibrated in 50 mM Tris (pH 6.8). Lysate was passed over the column. The SP column was washed with 200 ml of 50 mM Tris to eliminate weakly bound contaminants. Bound Hbs eluted with a 100-ml gradient of 50 mM Tris (pH 6.9) 1 M NaCl, from 0 mM to 1 M. Fractions (0.5 ml) were collected along the length of the gradient. The four reddest fractions were collected and then concentrated in an Amicon µLtra-15 tube by centrifugation at 4,000g to a final volume of 500 µl. The sample was injected into a Sephacryl Hiprep 16/60 S-100 HR size-exclusion column (SEC) for additional purification. The column was equilibrated in phosphate buffered saline (PBS) at pH 7.4. Depending on molecular weight, purified globins elute at 48–52 ml (tetramer), 56–60 ml (dimer) or 64–67 ml (monomer). The purity and identity of isolated proteins was

assessed using 20% SDS–PAGE and denaturing HRA-MS. The purified proteins were concentrated and then flash frozen with liquid nitrogen until use.

Protein purification by zinc affinity chromatography

Ancα + Ancβ was purified using zinc-affinity chromatography, adapted from a published method50. Buffers were loaded onto the metal affinity column using an AKTAprime FPLC. To prepare the zinc affinity column, nickel was removed from a HisTrap column (GE) using stripping buffer (100 mM EDTA, 100 mM NaCl, 20 mM TRIS, pH 8.0). The column was then washed with diH2O for five column volumes. Then 0.1 M ZnSo4 was passed over the column until conductance reached a stable value. The column was then washed with five column volumes of water. After expression, cells were resuspended in 50 ml lysis buffer containing 20 mM Tris and 150 mM Nacl (pH 7.4). The cells were sonicated as described above. The lysate was passed over a zinc-affinity HisTrap column. The column was washed with 200 ml wash buffer (20 mM Tris and 150 mM Nacl, pH 7.4). The bound Hbs were eluted with a 50-ml gradient of imidazole, up to 500 mM, and 0.5-ml fractions were collected during the run. The four reddest fractions were collected. The Hb-containing fractions were concentrated and injected into a Sephacryl S-100 HR column for additional purification, as described above.

Native Mass Spectrometry

Protein samples were buffer exchanged into 200mM ammonium acetate using either a centrifugal buffer exchange device (Micro Bio-Spin P-6 Gel, Bio-Rad) or a dialysis device (Slide-A-Lyzer MINI Dialysis Unit, 10000 MWCO, Thermo) prior to native MS experiments. Native MS was performed on a Synapt G1 HDMS instrument (Waters corporation) equipped with a 32k RF

generator. The instrument was set to a source pressure of 5.47 mbar, capillary voltage of 1.75 kV, sampling cone voltage of 20 V, extractor cone voltage of 5.0 V, trap collision voltage of 10 V, collision gas (Argon) flow rate of 2 ml/min (2.65 x 10-2 mbar), and T-wave settings (velocity/height) for trap, IMS and transfer of 100 ms-1/0.2 V, 300 ms-1/16.0 V, and 100 ms-1/10.0 V, respectively. The source temperature (70 oC) and trap bias (30 V) were optimized. Analysis of the MS data to estimate masses and relative abundances was performed with the software program Unidec.

Estimation of homomeric and heteromeric affinities

The occupancy of each oligomeric state in solution was calculated as the proportion of globin subunits in that state, based on the summed areas under the corresponding peaks in the spectrum. To estimate Kd of the monomer-to-homodimer transition of isolated chains of Ancα/β, Ancα, Ancβ, Ancα/β_IF1, t37V, q40W and Ancα/β2 we performed nMS at variable protein concentrations. At each concentration, the observed fraction of subunits incorporated into dimers (Fd) was estimated as Fd = (2xd+4xt)/ (xm + 2xd+4xt), where xm, xd and xt are the sums of the signal intensities of all peaks corresponding to the monomeric, dimeric and tetrameric stoichiometries, respectively. This procedure was repeated at a range of protein concentrations. Nonlinear regression was then used to find the best-fit value of Kd using the equation: $F_d = \frac{1}{P_{tot}} * \frac{(4P_{tot}+K_d)-\sqrt{((4P_{tot}+K_d)^2-16{P_{tot}}^2)}}{4}$

For homotetramerization of globins expressed in isolation, the Kd of the dimer–tetramer transition was calculated using a similar approach. The fraction of all subunits incorporated into homodimers (including both free homodimers and those associated into homotetramers) was

calculated as Fd = (2xd + 4xt)/ (xm + 2xd + 4xt), and the concentration of all dimers was calculated as Cd = Fd × Ptot. The fraction of all dimers that were incorporated into tetramers was calculated as Ft = 4xt/ (2xd + 4xt). Nonlinear regression was then used to fit Kd to the data using the equation: $F_{\alpha 2\beta 2} = \frac{1}{C_{\alpha\beta}} * \frac{4C_{\alpha\beta}+K_d-\sqrt{\left(4C_{\alpha\beta}+K_d\right)^2-16{C_{\alpha\beta}}^2}}{4}$

The resulting Kd is expressed in terms of the concentration of globin subunits contained in homodimers or homotetramers.

To determine the Kd of heterodimerization between various combinations of subunits, we performed a titration series where one subunit concentration was held constant, while the other was mixed with it in increasing concentrations. We estimated the proportion of the heterodimer and the two possible homodimers as Faa = 2xaa/ (2xaa + 2xab+2xbb+xa+xb), Fab = 2xab/ (2xaa + 2xab+2xbb+xa+xb) and Fbb = 2xbb/ (2xaa + 2xab+2xbb+xa+xb), where x represents the signal intensity of the all peaks corresponding to the species denoted in the subscript. The concentration of each dimer was then estimated by multiplying each fraction with the total monomer concentration C, yielding variables Caa, Cab and Cbb. Nonlinear regression was then used to fit Kd to the data using the equation:

$$C_{\alpha\beta} = \frac{C_{\alpha\alpha} * C_{\beta\beta} * K2 * K3}{K1}$$

Where K1 is the equilibrium constant of the heterodimer, while K2 and K3 are the equilibrium constants of the homodimers; the latter two constants are estimated separately in a protein dilution series involving only isolated subunits.

Chapter 4: Origin of complex features during protein evolution

Proteins fold into elaborate structures, organize themselves into complexes and are functionally regulated by many allosteric inputs. Although such complex structural features are typically viewed as the product of long, gradual trajectories of adaptive evolution, here we argue that such paths could involve small numbers of substitutions that exploit pre-existing, latent properties in ancestral proteins. These paths could be available because protein sequence space is host to many biophysically and genetically distinct but functionally equivalent forms of complexity, such that some of these solutions are stochastically accessible to proteins through short mutational pathways at some point during their evolution.

4.1 Introduction

The ability of proteins to fold, assemble into complexes and respond to allosteric signals depends on the cooperative action of many specific and spatially disparate amino acids (**Fig. 4.1a-c**). The apparent sequence-specificity of these features poses a key question for evolutionary biologists: How does evolution find such rare functional solutions in a vast space of possible (and overwhelmingly nonfunctional) sequences? The evolutionary path to producing complex protein structures and functions *ex nihilo* would appear to be long and complicated, simply because so many amino-acid states must co-occur and so many biophysical conditions simultaneously satisfied before a given interface, fold or allosteric switch could be functional. For instance, the interaction between two proteins may rely on dozens of donor-acceptor pairs and complementary residues that together form a molecular interface (Goodsell and Olson

2000) (**Fig 4.1c**). Similarly, many residues may be responsible for allosterically linking a protein's interaction with a ligand to its interaction with a regulatory molecule (Rivalta et al. 2012; Süel et al. 2003) (**Fig 4.1b**). Lastly, the topology of a fold itself can depend on stabilizing interactions between hundreds of residues (Thornton et al. 1999) (**Fig 4.1a**). Such complicated structural features can be often be destroyed by one or a few mutations that alter a set of apparently fine-tuned amino-acid couplings (Kortemme, Kim, and Baker 2004) (**Fig 4.1d**). Despite their apparent biophysical complexity, however, evolution has been able to generate thousands of proteins with unrelated folds (Thornton et al. 1999), subunit interactions (Marsh et al. 2013) and regulatory modes (Goodsell and Olson 2000).

This paper deals with the process by which evolution invents new protein features – including new protein assemblies, allosteric switches and tertiary structures. We draw an admittedly fuzzy distinction here between invention, which involves the construction of a new functional property with no obvious prior analogue, and innovation, which involves modification of an existing function to yield new variant properties. To illustrate by morphological analogy: feathers could be considered a morphological invention, while variant feathers in different bird lineages could be considered innovations. Molecular evolutionists have tended to focus their empirical work on understanding the latter process and have delivered numerous compelling examples of how an existing protein architecture can be tweaked or retooled slightly to deliver new or enhanced function (Jacob 1977). Such studies have demonstrated how a binding pocket could be modified to accommodate a slightly new ligand or the how the catalytic efficiency of an enzyme could be improved by selection (Salverda and Barlow 2010). This work reveals how evolution could "tinker" with a pre-existing function but it does not reveal the mutational arc

by which those functions were initially built. Francis Jakob argued that this sort of tinkering was actually the fundamental substance of molecular evolution, and that the "period of true biochemical creativity occurred early during evolution" (Jacob 1977).

The classical explanation for the origin of complicated features is that they arise through a lengthy series of incremental mutational steps under sustained selection for improved function (Tetsuya Yomo 1999; Dawkins 1997; Darwin 1859) (**Fig 4.2e**). Although this explanation could certainly pertain to proteins, the role of gradualist, selection-driven step-wise evolution in producing complicated protein features is hard to empirically validate or falsify for a number of reasons. Firstly, proteins do not leave behind a fossil record (with some minor exceptions (Wadsworth and Buckley 2014)), forcing us to rely on indirect inferences of history based on comparative phylogenetics (Moparthi and Hägerhäll 2011; Liu et al. 2007; Aizawa 2001), which may not yield sufficient information to reconstruct a well-resolved history of genetic and functional change. Extinction may have erased many key intermediates in the functional history of a complex feature. Further, detailed structural knowledge of proteins is heavily weighted toward proteins from a small array of model-organisms, limiting our capacity to use parsimony to reliably infer history in many cases. Secondly, if the process of building novel protein features is sufficiently gradual, it may not occur in the context of real-time laboratory evolutionary experiments, even with relatively fast-replicating E. coli and yeast, to be observable in a lab setting. Lastly, the evolution of a new architecture may have involved many imperceptibly small changes stretching across millions or billions of years. If the evolution of complex protein features involved this degree of incrementalism, then – in some sense – there was no "moment

of origination" for these features, just a continuum of scarcely distinguishable transitional forms. The difference between an innovation and an invention is merely one of time.

Despite the dearth of direct molecular evidence for adaptive gradualism in the emergence of novel protein features, it remains an attractive idea because it is theoretically appealing (Fisher 1958) and because, until recently (Stoltzfus 2012), there has been little in the way of sound evolutionary alternatives. In the last two decades, however, protein scientists have used rational mutagenesis to engineer totally new properties into molecules, opening the door to understanding how similar novelties could arise during natural evolution. Before we address the plausibility of adaptive gradualism in generating multi-amino-acid features and the various alternatives one might propose, we will turn to consider in detail why both gradualism and adaptation have come to be viewed as a sufficient explanation for the emergence of complex features.

4.2 The tension between gradualism and punctuation

Complicated features have always been understood to be unlikely to arise from random processes. Many pre-Darwinian thinkers reasoned that complex biological features must arise from a teleologically-directed process, whether that be naturalistic or supernatural. Lamarck proposed that organisms possess an innate biological striving for complexity. He proposed that organisms are filled with an invisible fluid that tends to build and elaborate their tissues and organs, yielding ever more complicated forms (Burkhardt 2013). The natural theologians of 19th century Europe viewed the fine-tuned complexity of biological features to be clear evidence of divine design - an argument echoed by their intellectual descendants today (Behe 1996). William Paley, for example, compared the probability of complexity arising spontaneously

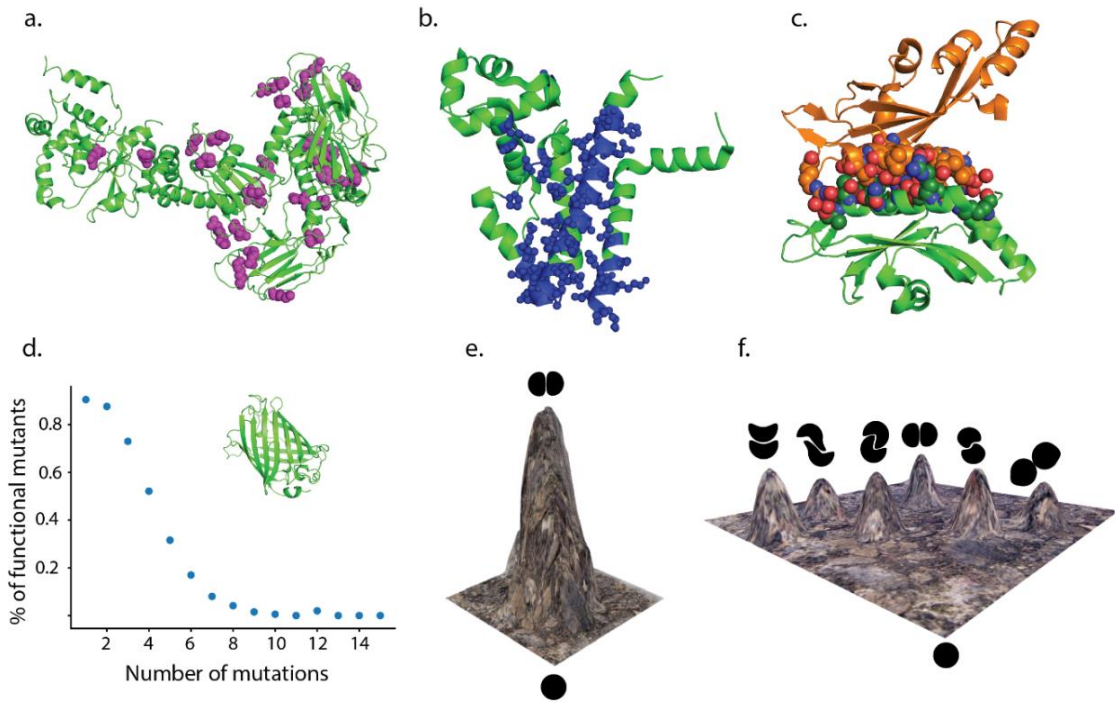


Figure 4.1. The sequence-specificity of complex protein features. (a) Map of residues in Yeast HSP90 that cannot be altered without negative fitness consequences. Green cartoon, Yeast HSP90 backbone; Purple spheres, residues where all tested mutations yield reduced growth rate **(b)** Allosteric network of a transcriptional repressor, TetR-B, gleaned from a mutational scan. Green Cartoon, TetR-B backbone; Blue, sites where the majority of mutations yield variants with defective allostery. **(c)** Interaction between PhoP (green) and PhoQ (orange); spheres, atoms that are within 4A of the opposing protein surface. **(d)** The proportion of folded and fluorescent GFP proteins in the immediate sequence space of of wild-type GFP in a recent mutational scanning study. Mutations are randomly distributed. **(e)** Dawkins' mount improbable, where evolution must scale up a single global fitness under selection in order to generate an oligomer from a monomer. **(f)** A multi-peaked fitness landscape where many alternative dimers are accessible to an ancestral monomer.

without a designer to the likelihood of finding a watch assembled via wholly naturally processes lying in a heath (Paley 1851). In 1969, Salisbury –a biologist who shared Paley's creationist sensibilities - stated that, in the case of a DNA or a protein sequence, the implausibility of random origination could actually be quantified. Based on some loose biochemical assumptions, he intuited that the probability of a functional enzyme arising by chance in a pool of random proteins might be in the range of one sequence in 10 thousand to one in 1 trillion (Salisbury 1969). Decades later, empirical work on a random sequence library indicated that the real number could be closer to his lower estimate, even for a relatively 'simple' protein function like phosphate binding (Keefe and Szostak 2001). John Maynard Smith responded to this problem in a foundational work of molecular evolution, arguing that evolution does not produce protein features by sampling random sequences (Smith 1970). Rather, genes generally evolve by navigating through a network of mutationally-connected functional sequences in stepwise-fashion (Smith 1970). Protein functions are distributed in the vast, connected network of sequences in which many mutational steps are functionally neutral, while some connect genotypes with differing functional properties (Wagner 2014). In this way, JMS provided a scheme in which evolution could make the journey from a simple (perhaps subfunctional) ancestral phenotype to a more complex derived one – say, from a monomeric protein to one that assembles into complexes, or a constitutive receptor to one that can be regulated – though discrete mutational steps. Nonetheless, in order for evolution to arrive at those rare genotypes compatible with complex protein functions, there must be a force that sifts away the overwhelmingly complexity-decreasing or non-functionalizing outcomes produced by random mutation, and guides systems towards higher complexity.

In the classical view, this directing force is natural selection, which gradually elaborates and optimizes systems by sequentially adding new parts or interactions that improve function (Dawkins 1997; Darwin 1859). This explanation has been invoked to explain complexity at practically all levels of biological organization: the emergence of the primordial proteins (Yomo 1999), multicellularity (Herron et al. 2019) and various kinds of morphological complexity (Dawkins 1997). The modern vertebrate eye, for example, is understood to have been derived from a simple eye-patch precursor in the pre-Cambrian, through several successive modifications that improved visual sensitivity and acuity, by adding new cell types and components into the original structure (Lamb et al. 2007). It is plausible that the complex multi-amino-acid properties of proteins are progressively built in much the same way. A new protein-protein interface, for example, could be established gradually through a string of mutations that gradually improve stability and specificity of a useful interaction.

Adaptive gradualism has been the mainstream explanation for the origin of complex features for most of the modern history of evolutionary biology, but it has always existed in tension with the (minority) view that significant novelty could arise through sudden jumps (Theißen 2009). Darwin firmly believed in gradualism, and even stated that evidence of such sudden jumps would actually count as evidence against his broader evolutionary ideas (Theißen 2009). Huxley viewed the assumption of gradualism as an unnecessary burden for Darwin's theory, and argued that evolution could sometimes proceed in jumps (All 1995). In the early 20th century, Goldschmidt argued that there were many discontinuities between species that could not be explained simply through the fixation of minutely different alleles (Goldschmidt 1982). Crossing such "bridgeless gaps" required mutations of large-effect – ones, for example, that rewrote a

fundamental developmental programme in an organism, producing a new anatomical feature in a single sweep (Goldschmidt 1982). In response, Fisher (Fisher 1958) and Charlesworth (Charlesworth 1982) argued that such mutations of large effect are not likely to be relevant for evolution, because large mutations are likely to have negative pleiotropic effects. Furthermore, the current absence of intermediate forms between species that differ substantially in morphology does not prove that a string of such intermediates did not occur historically (Charlesworth 1982).

The tension between gradualism and punctuated evolution persists in various forms to the present day, but the focus historically has rested primarily in the realm of developmental biology or paleontology (Theißen 2009). The debate between the two camps could apply equally to the level of proteins as well – do novel protein structures arise through a long line of intermediate forms, or can they be produced by leaps? Gradualism could apply to the evolution of complex protein features for two reasons: (1) steps that confer small structural changes are much more likely to occur mutationally than large-effect steps and (2) large steps towards a new structural feature are likely to have negative pleiotropic effects. The first factor may result simply from a physio-chemical limit, where single amino-acid changes that immediately produce productive fits between complicated arrays of atoms are either rare or nonexistent. Such large-effect mutations are therefore unlikely to be sampled during evolution relative to mutations with subtle effects, especially in small populations with low mutation supply, owing simply to the vastly higher frequency of the latter. Conveniently, the availability of such mutations in sequence space can actually be tested using the tools of biochemistry and protein

mutagenesis. The second (and more classical) argument in favor of gradualism stems from the view that large-effect mutations that have a beneficial effect on one phenotype are more likely to have large negative effects on a host of other phenotypes. If a protein sequence can be interpreted as an entity in a high dimensional biochemical space, where (a) each axis represents a different biochemical 'factor' that is relevant for its function and (b) the fitness associated with a protein variant is defined by its phenotypic distance from a single adaptive peak, then Fisher's arguments about gradualism in organismal adaptation may apply to the emergence of protein novelties. Large effect mutations that allow a protein to suddenly populate a new fold, multimeric form or allosteric conformation, may be possible but they may be more likely to lead to conflict with other facets of protein function in a cellular context - including solubility, specificity, stability or activity, than small-effect mutations.

To summarize, the perceived necessity of adaptive gradualism as an explanation for complexity arises from the assumption that useful, novel protein features are rare and poorly accessible in genotype space and that the fixation of many small-scale variants constitutes a tractable path to building them from ancient precursors. In this paper we argue that it is biochemically feasible to imagine an alternative to gradualism, where novel functional folds, allostery and interfaces, could arise via a few steps from ancestors that do not possess detectable rudiments of those functions (**Fig 4.1f.**). This is mainly possible if genotype space contains many diverse sequences that can confer that feature – enough that they are sometimes quickly reachable during a protein lineage's history without any prior selection specifically for that feature.

4.3. The origin of new protein features via short paths

Evidence from a wide array of protein engineering studies from the last two decades shows that significant molecular invention is possible through short mutational paths from naturally occurring proteins. We restrict our focus here to experiments where the mutations used reflect changes that are at-least mechanistically plausible under standard molecular-evolutionary conditions, including point mutations and small indels – but excluding elaborate fusions or recombinations. Although large gene-fusions may have facilitated the evolution of folds, allostery and interfaces in the past, it is difficult to know whether or not the specific recombinations used in protein-design experiments constitute plausible routes in natural evolution. We also focused on outcomes that can fairly be described as de-novo acquisitions of new protein properties, rather than incremental changes to pre-existing traits, or ones that only confer marginal occupancy of new states.

Perhaps the clearest examples of such shifts exist in the literature on designed protein-protein interactions. Grueninger et al. 2008 show that one or a few point mutations (upto five) on a protein surface, typically to bulky, hydrophobic residues, is sufficient to confer strong self-association (nanomolar to low micromolar affinity) into higher order oligomers in a wide variety of protein backgrounds (Grueninger et al. 2008) (**Fig 4.2b**). In a similar vein, another group was able to convert monomeric GB1 into a micromolar affinity dimer in one substitution, and into a tetramer in five substitutions (Jee et al. 2008). Further, another, structurally distinct dimer was available in 4 substitutions from GB1 (Jee et al. 2008), indicating that distinct dimer-solutions were available in the local sequence space around the wild-type proein. In symmetrical homomeric interfaces, each interface-forming site is presented twice. This multiplicity of

contacts makes it easy for proteins to be linked into self-associations via point substitutions that contribute twice to the free energy of interface-formation (Monod et al. 1965). This multiplicative effect is further amplified if the ancestral protein in question is already assembled into a symmetrical complex – in a tetramer, for example, it is possible for a surface mutation that confers affinity for another tetramer to occur four times. Garcia-Seisdedos et al. 2017 further demonstrated that introducing small sets of point mutations into several modern proteins is sufficient to confer self-association into soluble fibers (Garcia-Seisdedos et al. 2017) (**Fig 4.2c**). A small insertion, allows for the cyclic 11-mer TRAP complex in *B. subtilis* to incorporate an additional subunit, indicating that rings can be enlarged with similarly limited genetic modifications (Chen et al. 2011). These observations are not restricted to symmetrical homomeric complexes: two entirely unrelated proteins, PLC1-PH and EPOR, could be induced to form micromolar affinity heterodimers through a single point mutation (Liu et al. 2007). These data demonstrate that higher order oligomers and interfaces can be created through short genetic paths(Keightley et al. 2009). Such short paths to new interfaces have also been shown to exist even in more explicitly evolutionary studies (Anderson et al. 2016). Two ancestral reconstruction studies demonstrated cases where a strong protein-protein interaction was initiated through one or two substitutions, with large subsequent effects on evolutionary history (Anderson et al. 2016; Pillai et al. 2020).

In a similar vein, naturally non-allosteric proteins have been converted into allosteric molecules via a small set of mutations as well. Substituting one or two buried residues in fibronectin with ionizable histidines was sufficient to make the conformational equilibrium sensitive to protonation and therefore pH sensitive (Heinzelman et al. 2015) (the allostery in this case is tied

to reversible self-assembly into fibrils). Nonallosteric Pyruvate kinase M was made allosteric by transferring a few states from an allosteric relative (Ikeda, Tanaka, and Noguchi 1997), and Dictyostelium PFK by a small deletion (Santamaría et al. 2002). Phosphosites engineered with a few point substitutions at various locations on the surface of yeast KSS1 are individually sufficient to confer phosphoregulation by a kinase (PKA) that does not regulate it in vivo (Pincus et al. 2018). Point mutants in β-glycosidase and β-glucorodinase are sufficient to produce de novo positive regulation by indole (Deckert et al. 2012) (**Fig 4.2a**). The unique ability of crocodilian Hb to respond to bicarbonate can be transplanted into human Hb by transferring just five states from crocodile Hb to human Hb (Komiyama, Miyazaki, and Tamef 1995). The gradual construction of a physically connected allosteric network was not essential in any of these cases. In fact, the mechanisms underpinning its evolution may actually be quite simple. For example, the disruption of protein stability and its subsequent chemical rescue by an allosteric binder may be sufficient to confer positive heterotropic allostery (Deckert et al. 2012).

Lastly protein engineering has given us some insight into how proteins could navigate between different folds. In one study, it was shown that a single mutation was sufficient to convert a sequence coding for an one fold (3α) into one that codes for another (α + 4β) (Alexander et al. 2009). In another study a single point mutation was sufficient to change the secondary structure and side-chain packing of a substantial portion of the Arc repressor (Cordes et al. 2000). Two mutations can deliver a major fold change in the tertiary structure of the nematocyst protein in *Hydra* (Meier et al. 2007). Thus, highly distinct secondary and tertiary structures can be connected to one another by small steps in sequence space.

4.4. The latent evolutionary potential of proteins

The trajectories related above involve small numbers of mutations that interact with large numbers of sites that remain unchanged in terms of their amino acid identity. This means that pre-existing, latent features of a protein can frequently contribute to the construction of a new structural feature. There is substantial evidence for such latent potential in proteins. In the case of interfaces, a computational analysis showed that soluble protein surface patches of 1000 A2 are, on average, two substitutions away from having a sequence composition characteristic of the residues buried in a protein-protein interface (Levy 2010). Many previously non-interacting amino-acid states could help constitute the "rim" of an interface once it emerges (Levy 2010). A recent experimental study of the mobility of human proteins in E Coli cells shows that they are slower and "stickier" than their E Coli homologs (Mu et al. 2017). This suggests both that human proteins have the latent potential to form weak, random, non-adaptive interactions with proteins that they have not co-evolved with, and that such weak, random interactions are so ubiquitous that E Coli proteins have actually accrued changes that suppress them because they are net deleterious (Mu et al. 2017). One computational analysis found that a computational library comprising random marginally-stable proteins contained surface pockets that showed significant correspondence to many biological functions, despite no selection having been imposed on the pool for any of these functions. This suggests that the set of random stable proteins has atleast some latent functional binding potential prior to selection (Skolnick and Gao 2013).

Similarly, the amino-acid networks that underpin allosteric regulation may precede selection for allostery. Proteins may possess latently allosteric features even before they accrue

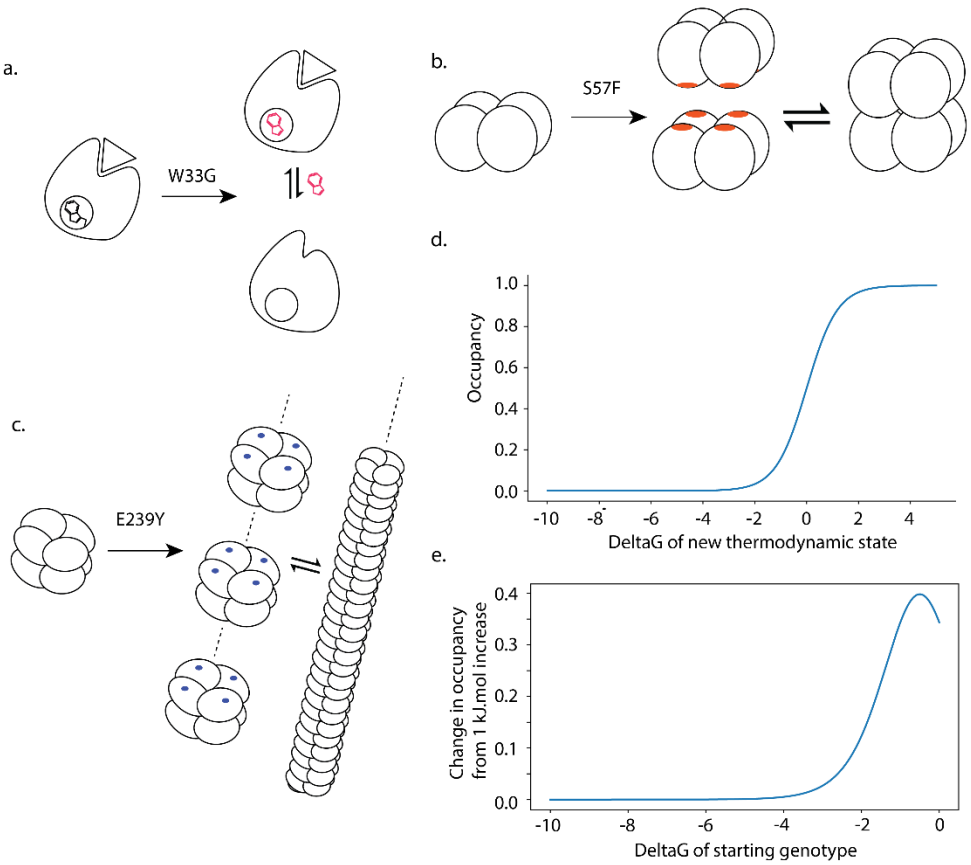


Figure 4.2. Accessibility of novel functions via short paths (a) Replacement of a single buried tryptophan with glycine in an enzyme (B -glycosidase) produces a compromised enzyme. This enzyme can, however, be chemically rescued by the addition of indole – a structural analog for tryptophan. **(b)** Rua-A can be converted into an octamer through a single substitution. Each mutation to Phe occurs four times in one tetramer owing to symmetry. **(c)** Single substitution to tyrosine induces fiber formation in vivo in Isoaspartyl Dipeptidase. **(d)** Relationship between deltaG and occupancy of a given thermodynamic state, where the reference state is taken to be 0 kJ/mol. **(e)** Change in occupancy produced by a mutation that adds a 10 kJ/mol interaction when the deltaG of the starting starts out being 0 to -10 kJ/mol away from the wild-type.

substitutions that optimize their response to a particular effector. The fungal allosteric effector Ste5 evolved to regulate the MAP-Kinase protein Fus3 in Saccharomycetes, but Ste5 has been shown to the modulate MAPK protein activity in other fungal lineages in which this effector never evolved (Coyle, Flores, and Lim 2013). This suggests that Ste5 activation evolved by tapping into an already existing allosteric network. A similar principle may have applied to the evolution of a classic model of allosteric function: vertebrate hemoglobin. The cooperativity of tetrameric hemoglobin relies on a structural linkage between oxygen binding to the heme-iron and alterations at the subunit interface that holds the two dimers together (Gelin, Lee, and Karplus 1983). This linkage is conveyed through the motion of the F helix. Similar perturbations of the F-helix during linkage-binding are observed even in the non-cooperative homologous monomer, myoglobin, suggesting that the linkage is more ancient than the evolution of cooperativity (Barends et al. 2015). A recent ancestral reconstruction study demonstrated that the evolution of cooperativity in Hemoglobin did not involve altering any of the many residues involved in binding to the heme and further, that it was possible to construct a cooperative tetramer from a noncooperative, ancestral dimer without changing the residues in the F-helix (Pillai et al. 2020). These data suggest that a similar recruitment of ancestral, latent potential also underpinned the evolution of allostery in the globin family.

Based on these studies, the evolution of allostery appears to involve exploiting pre-existing, long-range thermodynamic couplings between sites at the ligand-binding pocket and sites at the new effector binding site. This is clearly revealed in the case of allosteric drug targets: drugs can induce changes in enzyme activity by binding to surface sites distal to the substrate binding pocket, implying that these sites are fortuitously linked to active sites (Lu et al. 2018). Not only

are these linkages not the result of selection, they are actively deleterious, since they led to inhibition by a drug. Such latent structural couplings between sites have also been demonstrated via mutagenesis in proteins that do not actually exhibit allosteric regulation – small mutational perturbations can easily influence the structural environment at sites that are 16A away, as revealed by NMR (Davidson et al. 2008), suggesting that proteins already exhibit the long-range couplings between sites that could contribute to regulation. Collectively, these studies show that allosteric connections within proteins can exist without any direct prior selection for an effector interaction.

In the case of folds, many proteins (perhaps as many as 5% of solved structures) already exhibit the capacity to occupy more than one fold, albeit under varying cellular conditions (Porter and Looger 2018). Completing the transition to exclusively occupying one fold could therefore involve a stabilizing a conformation that already exhibits partial, if marginal, occupancy in an ancestral genotype.

4.5. Degeneracy

What then, explains the existence of these latent features – why are proteins so well positioned in sequence space to serve as platforms upon which novel features can be built?

We identify two explanations for this apparent latent potential: the first is selection for evolvability (Wagner, n.d.), and the latter is that proteins are chemically versatile enough that this potential could arise by chance (**Fig 4.3e,f**). Under the first explanation, selection pushes proteins into regions of functional sequence space where complex phenotypes are accessible via a few short steps. This is theoretically possible, but it is unclear if the rather exotic

population-genetic conditions under which this could be the case are actually in place in most naturally evolving populations (Michael Lynch 2007). It requires indirect selection, not for organismal phenotype, but for the propensity of a population to yield new phenotypes. The alternative is that this latent potential was arrived with at no direct selection at all, through neutral sequence wandering or selection for unrelated features, which brings with it its own host of plausibility issues.

The plausibility of chance emergence as an explanation for this potential depends on the distribution of complex phenotypes in functional sequence space. If biophysically complex structures are sufficiently sparse in sequence space, then genotypes that are mutationally adjacent to them may almost never actually be sampled by drift within a neutral network. Alternatively, sequence space may actually be quite "degenerate" (as in, multiple different sequence solutions for a given biological problem) with respect to many protein functions, which might explain how proteins manage to occassionally wander within striking range of a new and more complex phenotype during history.

One way to probe degeneracy is to comprehensively sample random sequence space for 0functional sequences. The most complete version of this experiment is to make all possible amino acid sequences of a defined length and test them for all possible functions. This experiment will be impossible for the foreseeable feature, both because random sequence space is enormous, and the true range of biologically relevant protein functions is unknown. Molecular biologists have however created large libraries to sample the sequence space around proteins for functionality, either comprehensively at a subselected set of sites in a protein (Starr, Picton, and Thornton 2017), or coarsely for an entire protein of defined length (Fowler

and Fields 2014a). In a random library of 80 amino-acid proteins, Szostak's group found that there are multiple possible phosphate binding proteins in random sequence space, which share no significant protein sequence identity, and are structurally unrelated (Keefe and Szostak 2001). In another study, in which four residues in the DNA-binding helix of a transcription factor were randomized, it was shown that there are alternative DNA-binding helices in the steroid receptor family, which share no sequence identity at four essential sites, that can nonetheless confer specific DNA binding to the response element (Starr, Picton, and Thornton 2017). McClune et al. (2019) showed that it is possible to engineer alternative, specific two-component protein systems in bacteria which do not share sequence identity with any existing two-component system at the cores of their protein-protein interfaces (McClune et al. 2019). Allostery, too, can be induced in a non-allosteric protein through a vast array of possible mutations at various sites that tune the relative stabilities of conformations so as to generate a functional switch upon the addition of an effector (Leander et al. 2020). These data show that many genetically and biophysically quite distinct sequence solutions are possible when compared to those that actually fixed during natural history.

Further evidence for functional degeneracy comes from the fact that many naturally existing biochemical structures that differ radically in sequence nonetheless perform the same function. Life has managed to reinvent biochemical wheels on multiple occasions using highly divergent structural solutions (Liljas and Laurberg 2000). Evolution has produced oxygen-binding globin multimers on multiple occasions, but utilized very different structural interfaces and allosteric mechanisms to do so (Qiu et al. 2000). Allosteric control can also be encoded in multiple different ways in the same protein family: MAP kinases have evolved to be regulated by

phosphorylation at multiple structurally disparate sites for example, suggesting that the sequence constraints on encoding phosphoregulation are not highly specific (Pincus et al. 2018). Multiple distinct independently evolved interfaces are capable of conferring equivalent oxygen-binding cooperativity among different globins (Qiu et al. 2000) (**Fig 4.3a,b**). Selection for similar functions can frequently result in the exploitation of totally different protein folds in different lineages. The carbonic anhydrases of algae, animals and bacteria (Liljas and Laurberg 2000) (**Fig 4.3c**), the S-crystallins (Tan et al. 2016) of Cephalopods and α-crystallins (Wistow 2012) of vertebrates, the globins (Mouche, Boisset, and Penczek 2001) of earthworms and the hemocyanins (Kato et al. 2018) of molluscs, are all structurally and genetically unrelated, despite performing essentially the same respective functions.

At the broadest level, the complement of distinct sequences that deliver folded proteins may be larger than was previously appreciated. The potential of random sequences to yield folded – or at least – soluble proteins is key to the accounting for how new folds arises. For example, Salisbury considered the probability of a random sequence yielding function as being vanishingly minute (Salisbury 1969) – however, there is evidence that new, functional genes have actually arisen during evolution from stretches of non-coding DNA in organisms as diverse as Yeast (Cai et al. 2008) and Primates (Ruiz-Orera et al. 2015) on relatively shallow time-scales. This implies that the functional density of sequence space is high enough that random sequences can sometimes produce functional proteins, which bear some marks of organized secondary and tertiary structure (Bungard et al. 2017), contrary to Salisbury's intuitions. De novo exonization may be a mechanism by which totally new folds arise. It remains an open

question, however, whether new folds arise de novo from non-coding sequences, or by a process of descent with modification from other folds.

These observations imply that sequence space is amply stocked with distinct sequence and structural solutions for a given biological problem. This functional degeneracy has a number of implications for protein evolution: (1) Temporal accessibility: if many possible solutions ("local optima") are possible for an initially neutrally evolving protein lineage, then one or another such solution will eventually be accessible through a small path (2) Genomic accessibility: if many functionally-equivalent biochemical solutions are feasible in sequence space, then many genic starting points, including genes from different different protein families could plausibly supply the starting point for the evolution of a radically new innovation. This is clear in cases where totally distinct folds, for example, were recruited to perform the same metabolic reaction (Galperin and Koonin 2012). (3) Repeatability: When the same form of complexity emerges from different genetic starting points, then its genetic and biophysical basis might be quite different, given that there are many local structural optima, rather than one or a few to which all populations gravitate toward under long-term selection.

What is the ultimate cause of degeneracy? Perhaps one explanation could be that proteins are polymers composed of enough different and chemically diverse monomers that they can accommodate numerous functions in numerous ways - more so than other biological polymers, which are made of fewer and less diverse monomers. It is not surprising then, that synthetic biologists have aimed to expand the range of available amino acids with the understanding that this will grant them access to even more chemical functions (Link, Mock, and Tirrell 2003). Although we lack specific data that could address the early evolution of protein biology, It is

possible that the genetic code of life evolved to code for 20 amino-acids as a compromise between having a wide enough range of chemical properties to support metabolism and ensuring translation fidelity (Koonin and Novozhilov 2017; Bedian 1984).

What we observe in extant genomes is only a subset of possibilities that fixed during history and gives us an impoverished picture of the range of functional genotypes that were actually possible during history. It is a record of "winners". Random libraries, protein engineering studies and the study of structural convergence can give us a broader picture of how degenerate sequence space is with respect to different kinds of biologically relevant functions and structures.

4.6. Gradualism in theory:

Even if new structural features reside only a few steps away from extant proteins, such steps may never actually be taken during evolution because they are disruptive to the protein's broader function. Gradualists in the past argued that organisms with dramatic alterations are likely to be "hopeless" monsters that are discarded by evolution, even if the generation of such mutant organisms is clearly possible in a laboratory context (Charlesworth 1982). Is this the fate we might expect from the laboratory-generated protein mutants discussed in section 3?

By Fisher's logic, even if large-effect mutations that induce complex innovation are not rare, they may come with serious pleiotropic costs (Fisher 1958) - then they are unlikely to be fixed during evolution. Some practical examples of this principle applied to evolutionary biochemistry could include: (a) a single mutation that builds a high affinity interaction with a target protein, but also yields strong off-target interactions with other related paralogs that are net

deleterious. (b) an interaction that binds so tightly to a target that it cannot be displaced by effector interactions (c) In the case of fold evolution, a mutation that stabilizes a new and useful fold, could also stabilize other deleterious thermodynamic states (d) a case where overstabilizing a fold compromises activity. If large-effect mutations are enriched for such negative pleiotropic consequences relative to small-effect ones, then they may be unlikely to fix. In these cases, gradualism may constitute the only tractable path to complexity. In the examples provided above, however, the mutant forms are soluble and stable enough to be amenable to in vitro biochemical assays, suggesting that their negative pleiotropic effects were not so serious as to eliminate the integrity of the protein fold. We take this to be provisional evidence that these engineered molecules should not be dismissed as “hopeless monsters”, destined for elimination by purifying selection were they to arise in nature.

Additionally, we note that work since Fisher’s time has illustrated that the strictest form of gradualism, involving infinitesimally small steps is not theoretically viable. Although small effect mutations may have fewer pleiotropic costs, they are also be more liable to be stochastically lost due to drift, as indicated by work by Kimura and Orr (Orr 2005). In order to fix, a mutation must not only confer a beneficial phenotypic effect, but its size of its phenotypic effect must be sufficient in order to escape elimination by drift. On this basis, Kimura argued that evolution is actually not likely to progress exclusively through infinitesimally-small micromutations as Fisher envisioned, since these actually have a low probability of fixation (Orr 2005). Overcoming drift is especially consequential in the specific case of generating a totally new fold or interface, during the early phases of an adaptive walk. In a simple case, where the emergence of a complex structure relies on the perfectly equal and additive energetic contributions of 5 new

hydrogen bonds, the first two steps would have marginal effects on the thermodynamic occupancy of the new structure, while the final two would have much larger effects. This arises simply from the nonlinear relationship between occupancy and free energy. It could also arise from epistasis among the residues that confer a novel protein property as well, where the formation of a fold – for instance – relies on the synergistic contributions of several amino acids that have no effect on free-energy when introduced singly. Selection must be strong enough to resolve the fitness differences between the early genotypes in an adaptive walk. These early steps could confer such small differences in occupancy (<1%) of a new complex state that they are likely to be eliminated by drift when they arise, since they are nearly functionally neutral (**Fig 4.2a,b**). Unless the linkage between organismal fitness and a particular protein novelty is quite substantial, these early-arising alleles are likely to be invisible to selection.

A body of theoretical work by Orr and Gillespie actually establishes a role for relatively large effect mutations during adaptative walks (Orr 2005, 2002; Gillespie 2014). For example, larger step sizes are favored when the adaptive optimum is phenotypically far away (Orr 2005). An adaptive walk in a population is expected to involve diminishing returns as it approaches an optimum, with the early steps conferring much greater changes in phenotype than those that occur later. In a model that explicitly accounts for the discrete nature of genotypic space (as opposed to the continuous phenotypic space imagined by Fisher), where there exist a finite number of accessible alleles with a distribution of fitness effects, Gillespie (2014) was able to show that adaptive walks at the molecular level are expected to be short, and that bulk of the fitness gain is due to a small number of substitutions (Gillespie 2014).

4.7 Selection and the production of complex protein features

The possibility that chemically ornate protein functions could be generally reachable through a few steps opens the door to explanations for their origination that do not involve persistent, long-term selection. The various proponents of gradualism over the past century have always argued that functional complexity is produced by directional selection and maintained by purifying selection for function. In this section, we investigate different possible evolutionary conditions under which complexity could arise (or be suppressed), by relaxing the assumptions of the gradualist evolutionary model with which we began this paper.

Strong and sustained selection is necessary to generate a feature when a vanishingly small subset of sequences in a network of accessible genotypes could confer it. A good example of such a feature might be hyperstability (Coleman and Sharp 2010) (i.e. maintaining foldedness at temperatures greater than 100°C) – which is understood to be exceedingly rare in sequence space – and only likely to be produced by thermal adaptation (Coleman and Sharp 2010) or human design (Anil, Craig-Schapiro, and Raleigh 2006; Wunderlich and Schmid 2006). The previous sections demonstrates that not all novel protein features show such a rarified distributed in genotype space, opening the door to both episodic selection and drift as alternative explanations for how modern day forms of protein complexity arise.

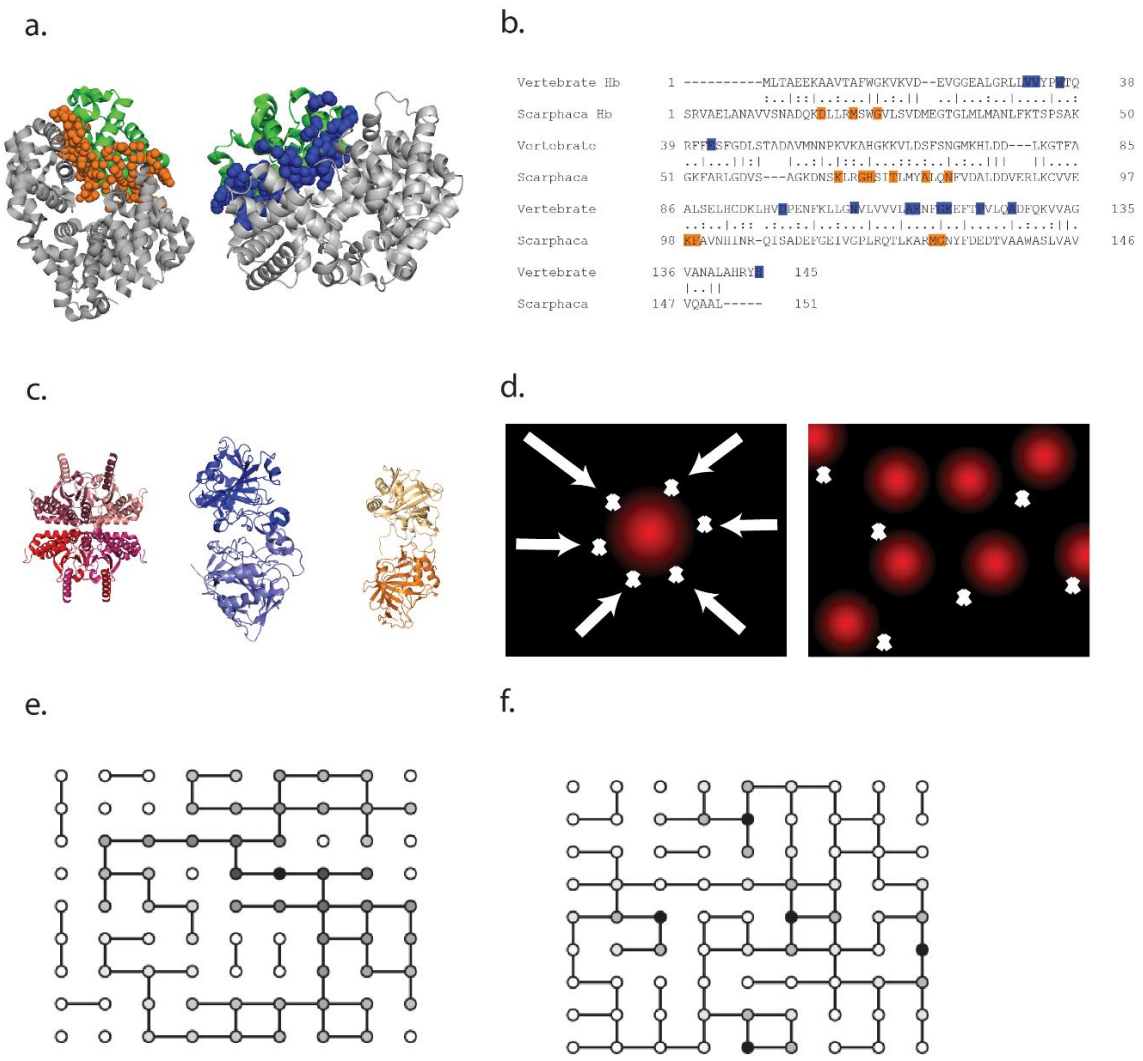


Figure 4.3. Functional degeneracy of sequence space. (a) Heterotetrameric, oxygen-transporting, cooperative globin proteins evolved independently in molluscs (left) and vertebrates (right). Orange and Blue spheres indicate mollusc and vertebrate interfacial residues respectively. Single monomer shown in green. **(b)** Interfacial sites for mollusc and vertebrate globins shown in a pair-wise alignment between the two globins. **(c)** Non-homologous carbonic anhydrases from algae (right), bacteria (middle) and animals (right). **(d)** Fitness landscape visualization of two hypotheses explaining the evolvability of new protein features. Red represents occupancy of the new feature. Either selection drives proteins to be close to new, but rare functions (left) or proteins wander close neutrally to numerous alternative possible encodings of the same function. **(e)** Sequence network where there is a single optimal genotype with several steps connecting it to lower-fitness variants that exhibit less of that phenotype **(f)** Sequence network where there are multiple genotypes that confer a phenotype, but they are connected to sequences that do not exhibit that phenotype.

Episodic selection: If we maintain the classical assumption that complexity has adaptive value, but accept that many alternative complex solutions are available and that some of these are therefore accessible through short paths during history (**Fig 4.3f**), then the efficacy of adaptation in generating complicated features is substantially increased. Complexity is more evolvable in this scenario, and selection need not be as strong or as persistent as it would have to be to cross the troughs and valleys of a fitness landscape in which there is only one poorly-accessible complex solution. In fact, observable, adaptively-significant variation in protein complexity could exist in a single population, where a segregating allele bearing a large-effect mutation codes for a protein variant that exhibits a novel form of allostery or multimeric assembly. Additionally, selection need not be as intense in order to bring a complex feature into existence. Rather than one gently-sloped peak in a fitness landscape, there are several steep peaks that can be surmounted in a single bound. As a result of this genetic diversity of solutions, the sequence outcome of selection for complexity, as a result, could be quite unpredictable.

Drift: If complex features can arise in single steps, and many such single mutations are available during evolution, then they could also plausibly fix by drift (Stoltzfus 2012). As a result, complexity may arise even when it does not yield substantially increased fitness relative to the ancestral genotype. There is evidence to indicate that this may be the case for some derived protein complexes, which are replaceable with simpler forms with no serious loss to viability. A striking example comes from the yeast RNAse P, which is a 10 subunit ribonucleoprotein that can be replaced with the much simpler Arabidopsis single-subunit RNAse P with no measurable fitness cost (Eswara et al. 2007). In a scheme where complex phenotypes are encoded by many

possible sequences, these sequences could be sampled during a neutral random walk, which would be highly implausible if complexity were instead rare.

Biophysical rachets: Accessibility alone, however, does not explain how neutral forms of complexity could persist once they arise through one or a few chance steps. Under a neutral scheme, where sequences conferring complexity are intercalated within the neutral network of genotypes that encode for simple phenotypes, complex interfaces should be lost as easily as they are gained because they are connected by short neutral paths. However, simple and complex phenotypes could exist in connected but ultimately distinct neutral networks. This could arise because the sequence-constraints that derived complex phenotypes evolve under are quite different from those that existed in a simpler ancestral context, permitting the fixation of previously deleterious variation. For example, complex features may evolve under different compositional constraints across sites than their precursors. In the case of a newly emergent protein-protein interface, which may have arisen through a single point mutation, the exclusion of water at a surface may allow previously deleterious hydrophobic substitutions to accumulate at that interface. These mutations may have caused nonspecific aggregation in an earlier ancestor, but they may be tolerated at a hydrophobic interface (Hochberg et al. 2020). Regardless of whether an interface arose neutrally or under selection, after a period of hydrophobically-biased accumulation, it is possible that a short neutral path back to a monomeric precursor is no longer available, because exposing these hydrophobic side chains would be deleterious (Hochberg et al. 2020). This mechanism could entrench some molecular complexes even without direct selection for the function of a complex.

Deleterious complexity: In the gradualist adaptive model where the only path to complexity involves many small changes, a novel form of complexity would never arise if it were inimical to fitness – in other words, the fixation of many mutations involved would be practically impossible if they were collectively deleterious. If point mutations can, in some cases, single-handedly induce complex features, then it is possible for deleterious forms of complexity to transiently segregate in populations, be weeded out or – on rare occasions - fix in populations (KIMURA 1962; Kimura 1980). Although it appears likely that interfaces, folds and regulatory responses can be generated through short mutational paths, it may be that – under most conditions – these forms of structural complexity are generally neutral or even deleterious because they conflict with established cellular roles. Mutations that confer new interfaces may obscure access to a ligand binding site, or occlude other important interactions, like PTMs. The best-documented polymorphism in quaternary structure causes sickle cell anemia, and is deleterious under most circumstances (Pauling et al. 2019). Many mutations that permit positive allostery may be yield inactive proteins that insufficiently stable in the absence of the allosteric effector. Novel forms of allostery could confer inhibitory interactions with small-molecules that may be deleterious (in the case of small-molecule inhibitor drugs, for example). New folds may be aggregation prone, as has been suggested by studies that computationally investigated the stability and aggregation propensity of newly emergent orphan genes. A chief limitation then, may not be the genetic complexity of delivering a switch or complementary interface, but rather making it compatible with protein function in an organismal context.

4.8 Future directions

The lines of evidence related above establish the plausibility of short mutational trajectories to novel forms of complexity in proteins- including allostery, multimerization and folding. However, the frequency with which evolution actually follows such paths during history, as well as their relevance to adaptive evolution, remains to be empirically established. In order to definitively address this question, we propose three lines of experimental inquiry:

First, to what extent do multimerization and allosteric capacity actually vary within families of related proteins? Answering this question requires the phylogenetically-unbiased characterization of protein variation across the tree of life. Such comprehensive sampling can reveal heretofore unknown structural and functional malleability within a protein family. For example, the characterization of globins from many animal clades allowed biochemists to the detect variation in the underlying mechanisms of allostery and multimerization in the globin family, and propose evolutionary mechanisms for their origination (Royer et al. 2005b). For most protein families, however, high quality structural and biochemical data is unavailable across a wide distribution of species - vertebrates, for example, account for most of the crystallographic data in the PDB, despite making up <5% of animal diversity (Krissinel and Henrick 2007). With enough sampling, however, it may be possible to isolate short phylogenetic branches along which novel protein features first arose, and even understand the mechanisms involved. Further it may reveal that the overwhelming conservation of protein structure is an illusory result of our limited sequence sampling.

Secondly, what proportion of mutations actually confer new protein features? To evaluate this, we suggest scanning point mutagenesis (Fowler and Fields 2014a) on extant or ancestral proteins to determine the accessibility of new forms of allostery and interfaces – baseline systems for analyzing large numbers of protein sequences for both allosteric regulation and protein-protein interactions have already been developed. This high-throughput approach would help us measure the fraction of point mutants that can induce structural innovation – a key variable in determining its evolvability.

Lastly, to what extent do proteins exhibit variation in oligomeric state or allostery *within* natural populations? If point mutants that encode new structural innovations are widespread, then such protein variants ought to be observed segregating in populations – in most cases, these may be segregating at low frequencies, because they are deleterious, but a small set of them may be selectively neutral, and an even smaller cadre may confer new functions. Observations at this scale may help us understand the reasons complex protein features spread or are extinguished in populations.

Although tracing the ultimate origins of complexity in protein structure and function remains a daunting prospect, it is at least an experimentally tractable one - it will involve deploying existing technologies and methodological frameworks, rather than developing them from scratch. Once it is resolved, however, evolutionary biologists can claim to have addressed a crucial piece of the bigger puzzle of how biological complexity originates.

Conclusion

This thesis provides the first detailed experimental description of the evolutionary origin of a crucial protein complex in deep time. By applying a combination of biophysical and biochemical techniques like native mass spectrometry and oxygen affinity assays on putative ancestral globin, we were able to isolate small sets of substitutions that were causal for Hb's unique molecular properties. These include its cooperativity, multimerization and heteromeric specificity. In addition to shedding light on the early history of essential protein complex, we identified several phenomena that could apply generally to the historical emergence of novel protein complexes. These include (but are not limited to) (1) the capacity of small numbers of changes to produce high-affinity, stable interfaces (2) the construction of new and useful interactions and allosteric capacities through the exaptation of pre-existing protein features. (3) the formation of one protein-protein interface in a complex can be biophysically and evolutionarily contingent of the origination of another. This work demonstrates that the biophysical complexity of a feature does not necessary mean that the evolutionary path required to build it involves many, small, individually adaptive steps. In this concluding section I outline several additional research directions and questions that could be pursued to follow up on this work.

Although many heteromeric complexes involve paralogous subunits, the majority of them do not (Ahnert et al. 2015). It remains an open question whether or not the principles we uncovered would apply also to the evolution of interactions between nonhomologous proteins – for example, the multiplier effect of mutations associated with isology (Monod et al. 1965) discussed in Chapters 1 and 2, would not apply to non-paralogous complexes, since their

interfaces are asymmetric. However, such interactions could still evolve – as we observed in Hb – through a small number of changes that occur primarily or entirely on one binding partner, while the other member of the complex remains evolutionarily static. One way to address this question would be to identify a phylogenetic interval during which two non-homologous proteins first evolved to interact and reconstruct ancestors before and after that period. Unlike the work described in Chapter 1, all of the relevant ancestral sequences would not occur in the same phylogenetic tree (since they are not related by gene duplication), but require separate phylogenetic analyses on different protein families. I propose that Hemoglobin could also be used as a system for exploring how non-homologous proteins associate into complexes as well. After the tetramer arose, it evolved to interact with a number of other non-globin proteins. It's interaction with haptoglobin, for example, evolved after the divergence of lampreys from gnathostomes. Haptoglobin binds free Hb in plasma with sub-nanomolar affinity and facilitates its degradation, thereby preventing oxidative damage to tissues (Polticelli et al. 2008). Haptoglobin evolved from a family of serine proteases involved in the immune system (Kurosky et al. 1980). The serine proteases and haptoglobins are highly alignable and slow-evolving, so it could be possible to trace the evolution of a novel Hb-binding Haptoglobin from an ancestral protease. Another novel interaction with Hb arose during early mammal evolution: after the split between monotremes and placental mammals, Hb-α became structurally dependent on a mammal-specific chaperone called A-Hemoglobin Stabilizing protein (AHSP). The 130 amino-acid chaperone appears to bear no obvious homology to any other protein family, and appears to either arisen de novo or undergone a period of extreme sequence divergence from any putative paralogs (Costa and Favero 2011). No orthologous equivalent for AHSP is known in

birds, reptiles or fish, suggesting that its acquisition was a mammalian novelty; Here one could the origin of the A-AHSP interaction, by reconstructing Hbs before and after the rise of the chaperone, and testing which changes on Hb allowed it to interact with AHSP, or if AHSP exploited a surface on Hb that predated its emergence.

A criticism that could be levied against the work in this dissertation is that Hemoglobin represents a rather simple molecular complex. It comprises only two genetically distinct subunits and two structurally distinct interfaces. In this sense, the parsing apart the history of hemoglobin represents only a first step. A major task for the field is to explain how even larger and more structurally elaborate complexes could arise. Even within the broader globin family, there exist vastly larger oxygen-transporting complexes- Eryhrocruorin, for example, is a 144-mer composed of both globin and non-globin subunits that transports oxygen in annelids (Royer et al. 2000). Complexes of similar size to Erythrocruorin and much greater subunit diversity exist throughout cellular biology. Unpacking the evolutionary, genetic and biochemical mechanisms by which ancient riboprotein complexes like the ribosome (Petrov et al. 2014) and transmembrane molecular machines like the flagellum (Liu et al. 2007) arose during history present important future frontiers for evolutionary biochemistry.

The finding that novel and stable oligomers can arise through a single amino substitution implies that distinct oligomeric variants could exist as polymorphisms within a single population. Patterns of allelic variation at loci that code for protein that differ in stoichiometry (or perhaps even the quantitative strength of their oligomerization) could reveal whether or not quaternary structures evolve under positive or balancing selection or if they segregate neutrally under drift. Currently, the only examples of oligomeric variation within populations represent

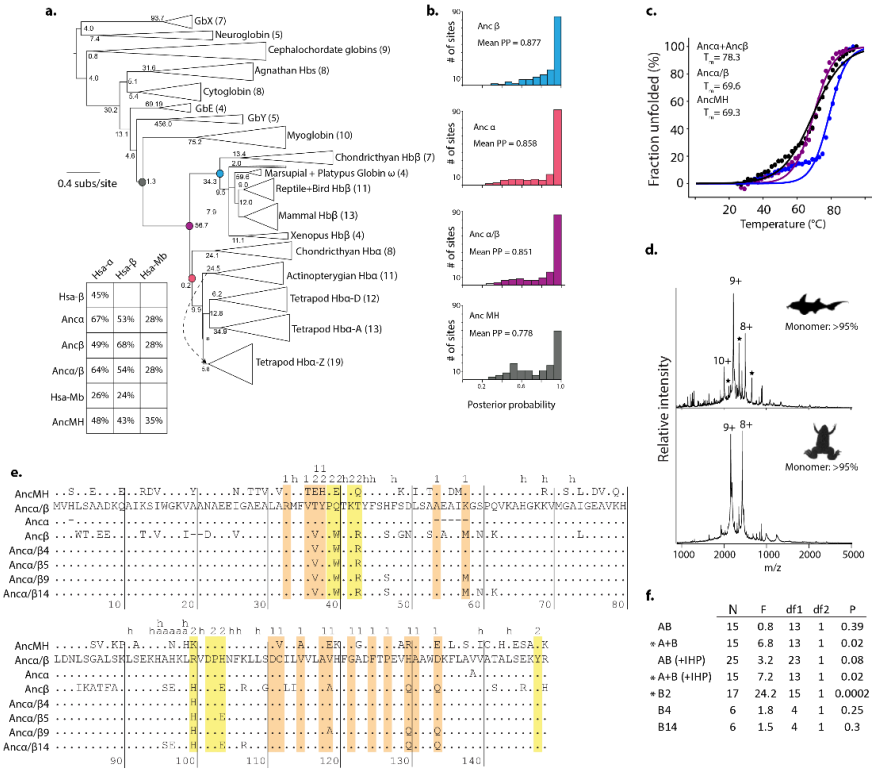
secondary losses (Williamson and Bentley 1983) or fiber assemblies (Pauling et al. 2019) rather than the gain of a novel, closed complex. We currently have no direct knowledge of the population genetic forces that drive oligomer variation or fixation. One way to assess this is to pick a model protein that is amenable to rapid characterization of stoichiometry and determine the extent to which variation exists within a population by assaying thousands of individuals. This could give us a more direct window into the extent of protein-stoichiometric variation in populations, as well as the forces that sustain or alter it.

The reconstruction of ancient complexes can tell us about the mutations that built protein-protein interfaces during history but it does not tell how rare such mutations are, or how many alternative genetic routes to forming an interface exist. To determine this, we would have to characterize the quaternary structures of all proteins one or a few mutations away from a wild-type sequence. This would tell us if oligomerization-inducing mutations are exceedingly rare, or relatively abundant. It is possible that deep mutational scanning techniques could be used to probe the local accessibility of novel complexes from a given (extant or ancestral) protein background. This would require constructing a library of point mutants and rapidly characterizing their capacity to self-assemble or not. The structural techniques described in this dissertation are low-to-medium-throughput since they typically rely on purified proteins, so an alternative, (likely cell-fluorescence based) screening method would be required to rapidly assay the enormous numbers of variants required to map even local sequence-space (Fowler and Fields 2014). One possibility is to test if any of the point mutants of a soluble, monomeric protein are capable of binding to another, unrelated monomeric protein using a high-

throughput yeast 2-hybrid system. An experiment of that sort could inform us as to the true degree to which wiring two monomers into a heterodimer is mutationally easy or hard

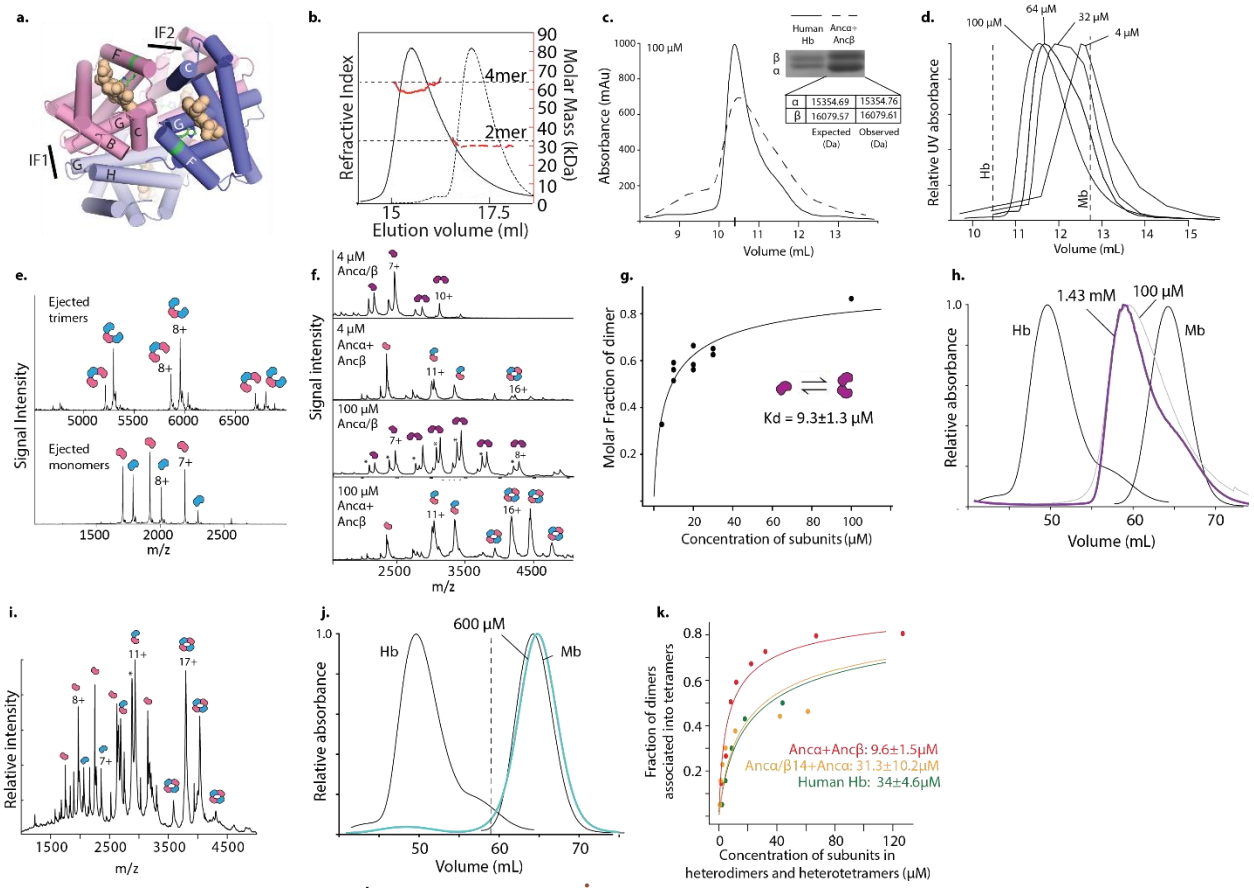
It is my hope that the work contained in this dissertation has contributed, however incrementally, to our mechanistic understanding of how protein complexes arise during evolution.

Appendix A: Extended Figures for Chapter 2



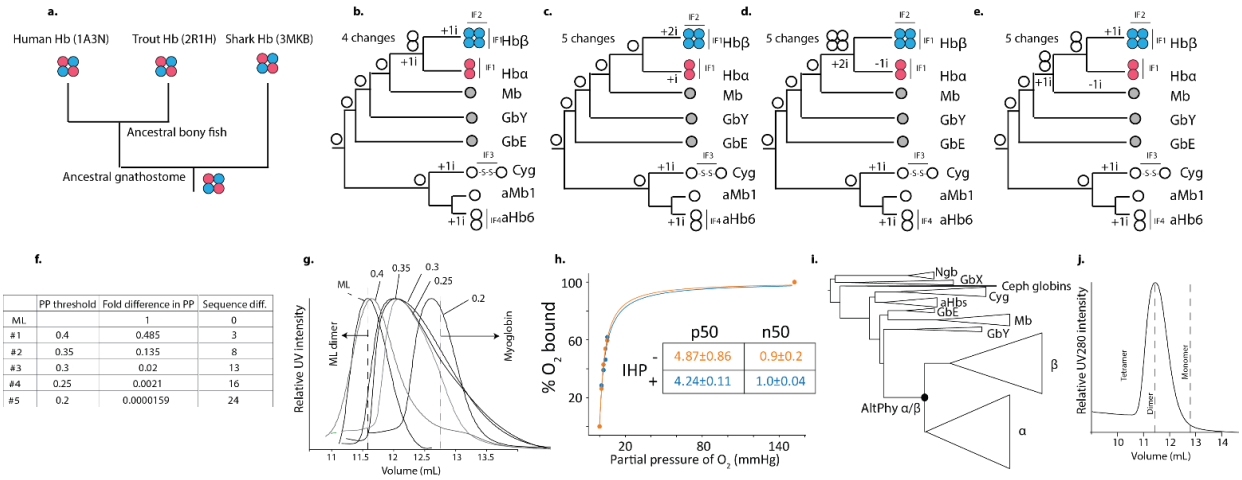
Extended Figure A1. Reconstruction of ancestral hemoglobin and precursors. **a)** Phylogeny of Hb and related globins. Node supports are shown as approximate likelihood ratio statistic[58,59]. Numbers of sequences in each group are shown in parentheses. Ancestral sequences reconstructed in this study are shown as colored circles. Arrow, branch swap that differentiates this phylogeny from the unconstrained maximum likelihood phylogeny, which requires additional gene gains/losses. The tree is rooted on neuroglobin and globin X, paralogs that duplicated prior to the divergence of deuterostomes and protostomes[60]. Inset: Pairwise sequence identities among extant (human, Hsa) and reconstructed ancestral globins. **b)** Distribution across sites of the posterior probabilities (PP) of maximum a posteriori states for reconstructed ancestral proteins. **c)** Thermal stability of ancestral globins. Points, fraction of secondary structure lost as temperature increases in Ancα/β (purple), Ancα+Ancβ (blue) and AncMH (black), measured by circular dichroism spectroscopy at 222 nm, relative to signal at 23°C. Estimated Tm and SE from nonlinear regression and the best-fit curve (lines) are shown. Each point is the mean of 4 measurements. **d)** Native mass spectra (nMS) of Globin Y (GbY) from elephant shark (*Callorhinchus milii*) and African clawed frog (*Xenopus laevis*) at 30 µM. Charge states of heme-bound monomer shown. Asterisk, cleavage products. Spectra were collected once. **e)** Sequence alignment of reconstructed ancestral globins. Dots, states identical to Ancα/β. Yellow, IF2 sites; Orange, IF1 sites; h, sites 4 Å away from the heme; a, sites that link the heme-coordinated proximal histidine (H95) to IF2. **f)** Statistical test of cooperativity of

(**Extended Figure A1.** continued) oxygen binding for ancestral proteins and mutants. An F-test was used to compare the fit of a model in which the Hill coefficient (n) is a free parameter to a null model with no cooperativity (n=1). Computed *P*-value and degrees of freedom (df) are shown. N, number of concentrations measured. *, $P<0.05$. Data were pooled across replicate experiments for nonlinear regression.



Extended Figure A2. Stoichiometric characterization of ancestral globin complexes. a) Homology model of Ancα+Ancβ (template 1A3N) showing heme (tan spheres). Blue cartoon, Ancβ subunits; red, Ancα. Helices and interfaces are labelled. Green, proximal histidine. **b)** Size exclusion chromatography and multiangle light scattering of Ancα/β (90 µM) and Ancα + Ancβ (60 µM). Black, relative refractive index. Red, estimated molar mass. Dotted lines, expected mass for dimers and tetramers. **c)** SEC of human Hb (dashed) and Ancα+Ancβ (solid) at 100 µM. Inset, SDS-PAGE of these complexes, with bands corresponding to α and β subunits. Inset, masses estimated by denaturing MS of Ancα+Ancβ, compared to expected masses based on primary sequence. **d)** SEC of Ancα/β across a series of concentrations. Dotted lines, elution peak volumes of human hemoglobin tetramer and myoglobin monomer. **e)** Tandem MS of the heterotetrameric peak in the Ancα + Ancβ nMS (indicated Fig. 2.1b). Ejected monomer and trimer charge series and the subunits they contain are shown. **f)** nMS of Ancα+Ancβ and Ancα/β at 4 µM and 100 µM. Charge series and fitted stoichiometries are indicated. Dotted peaks represent apo-chains. **g)** Monomer-dimer association by Ancα/β. Abundance of monomer and dimer were characterized using nMS across a range of concentrations. Circles, fraction of all subunits that were assembled into dimers as a function of the concentration of subunits in all states. Nonlinear regression (line) was used to estimate the dissociation constant (Kd, with standard error). **h)** SEC of Ancα/β at high concentrations (purple and gray lines). SEC traces of human Hb, myoglobin (Mb) are shown for comparison. **i)** nMS of Human Hb at 50 µM. **j)** SEC of AncMH (cyan) at a high concentration. SEC of human Hb and myoglobin (black) are shown for reference. Dashed line, Ancα/β dimer elution peak volume (see f). **k)** Alternative estimation of

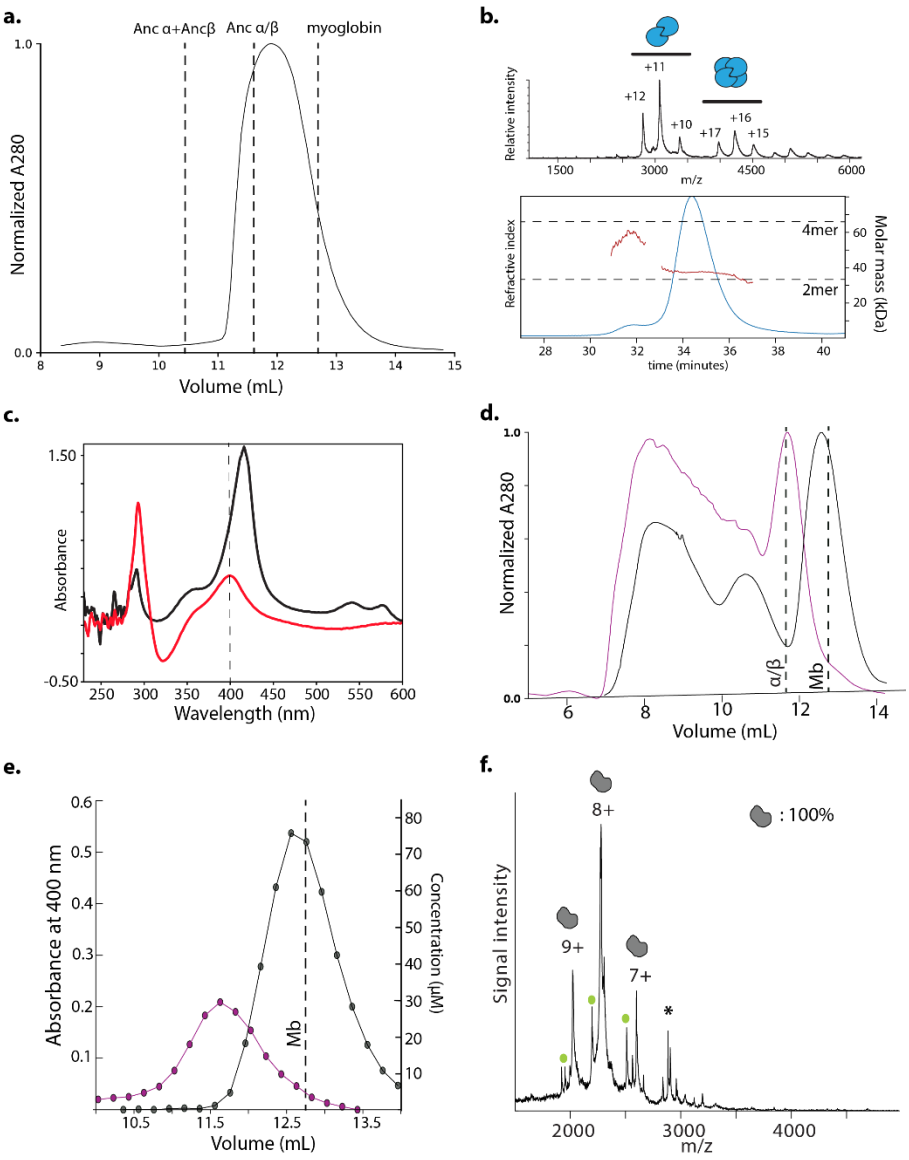
(**Extended Figure A2.** continued) affinity of dimer-tetramer association by nMS. For human Hb (blue) and Ancα/β14+Ancα (orange), the fraction of heterodimers incorporated into heterotetramers includes both heme-deficient and holo-heterodimers. For Ancα+Ancβ (red), cesium iodide adduct were included. Compare to Figs. 1d and 3d. Kds (with SE) were estimated by nonlinear regression (lines). All concentrations are expressed in terms of monomer. All nMS and SEC experiments were performed once at each concentration.



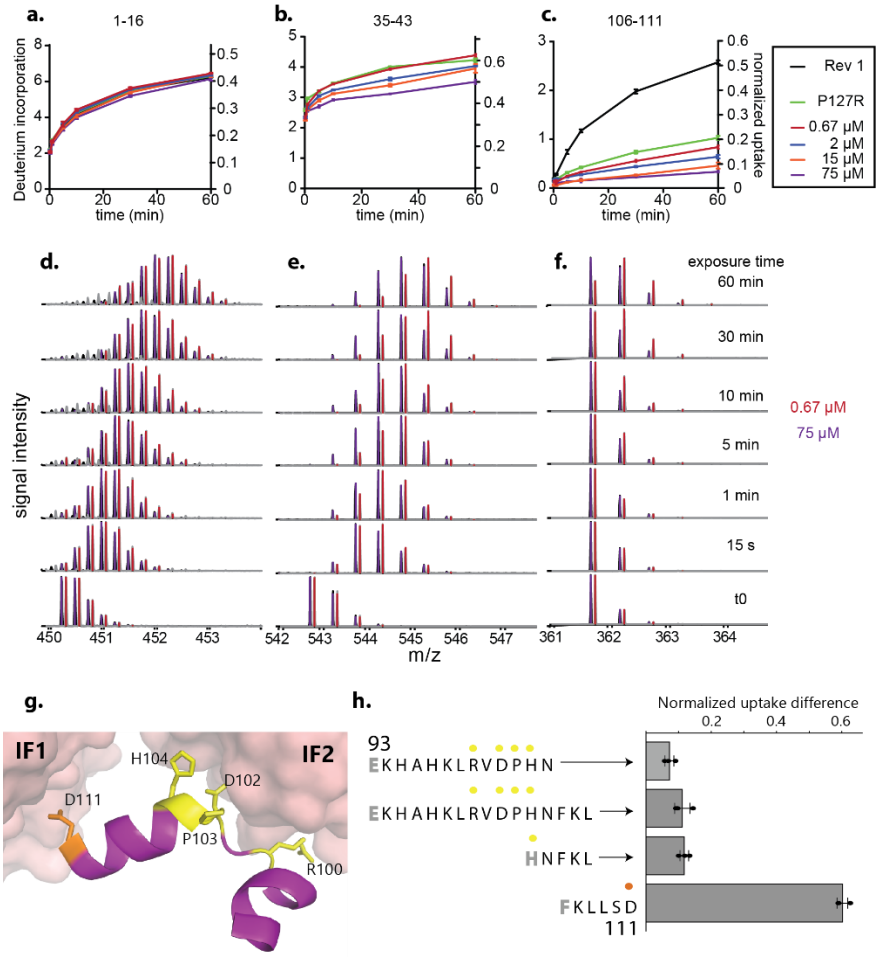
	PP threshold	Fold difference in PP	Sequence diff.
ML		1	0
#1	0.4	0.485	3
#2	0.35	0.135	8
#3	0.3	0.02	13
#4	0.25	0.0021	16
#5	0.2	0.0000159	24

Extended Figure A3. Biochemical inferences about ancestral Hbs are robust to uncertainty in sequence reconstructions. a-e) Maximum parsimony inferences of ancestral stoichiometry and interface loss/gains based on the distribution of stoichiometries among extant globins. **a,** Hbs in all extant lineages of jawed vertebrates are heterotetramers, supporting the inference that AncHb was heterotetrameric. Stoichiometries from representative species' Hbs are shown with PDB IDs. **b-e,** Each panel shows a hypothetical set of ancestral stoichiometries, plotted on the phylogeny of extant Hb subunits and closely related globins, with the minimal number of changes required by each scenario. **b)** The most parsimonious reconstruction is that Ancα/β was a homodimer and AncMH was a monomer. **c)** For Ancα/β to have been a tetramer, early gain and subsequent loss of IF2 in Hbα would be required. **d)** For Ancα/β to have been a monomer, IF1 would have been independently gained in Hbα and Hbβ. **e)** For AncMH to have been a dimer, IF1 would have been lost in lineages leading to the monomers myoglobin (Mb) and globin E (GbE) (Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore 1960; Blank et al. 2011b). The dimeric globins most closely related to Hb -- agnathan "hemoglobin" (aHb) and cytoglobin (Cyg) -- use interfaces that are structurally distinct from those in Hb (Lechauve et al. 2010; Heaslet and Royer 1999), indicating independent acquisition. **f-j**) Alternative reconstructions of Ancα/β are biochemically similar to ML reconstruction. **f)** Alternative ancestral versions of Ancα/β were constructed, each containing the the ML state at every unambiguously reconstructed site and the second most likely state at all ambiguously reconstructed sites, using different thresholds of ambiguity. For each alternative reconstruction, the table shows the threshold posterior probability (PP) used to define an ambiguous site, as well as the fold-difference in total PP of the entire sequence and the number of sites different from the ML reconstruction. **g)** SEC of ML α/β and alternative ancestors at 75 μM. Dotted lines show elution peak volumes for the dimeric ML α/β and monomeric human myoglobin. Constructs that elute between the expected volumes for dimer and monomer indicate dimers that partially dissociate during the run. None tetramerize; all form predominantly dimers, except AltAll (PP >0.2), which is ~62,000 times less probable than ML , which is mostly monomeric. UV traces were collected once for each construct. **h)** Oxygen binding curves of Ancα/β-AltAll (0.25), the dimeric AltAll with the lowest PP, with and without 2x IHP. Dissociation constant (P50, with SE) estimated by nonlinear regression is shown. Lack of (continued) cooperativity is indicated by the Hill coefficient (n50=~1.0). Oxygen binding at each

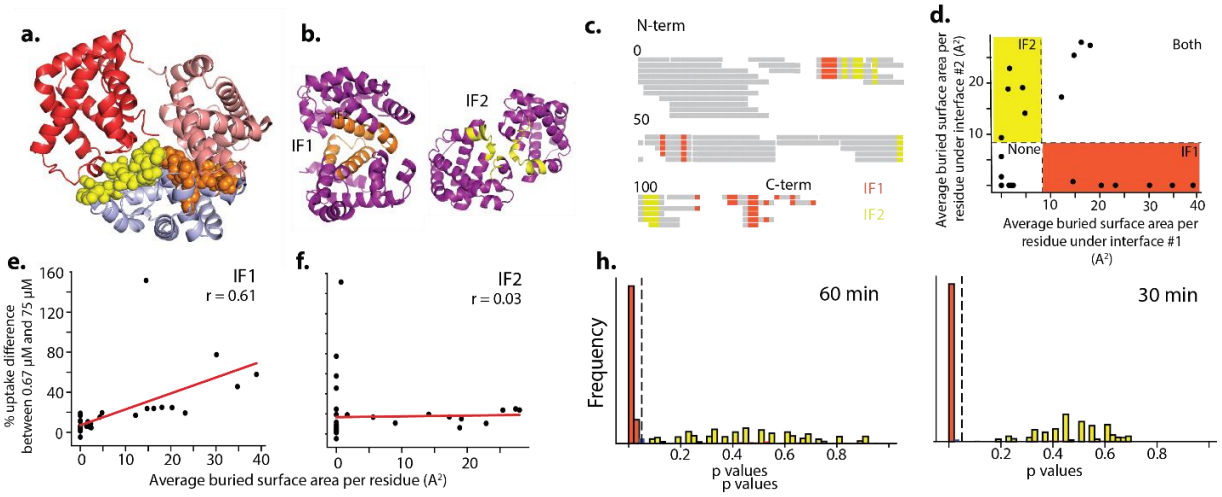
(**Extended Figure A3.** continued) concentration was measured once. **i)** Alternate globin phylogeny that is more parsimonious than the ML topology with respect to gene duplications and synteny but has lower likelihood given the sequence data. A version of Ancα/β (Ancα/β-AltPhy) was reconstructed on this phylogeny. **j)** SEC of Ancα/β-AltPhy. Dotted lines show expected elution volumes for various stoichiometric forms.



Extended Figure A4. Stoichiometric analysis of Ancα, Ancβ, and AncMH. **a)** SEC of Ancα at 75 μM. **b)** nMS spectra (top, at 20 μM) and SEC-MALS (bottom) of Ancβ. **c)** Colorimetric hemoglobin concentration assay. Absorbance spectra before (black) and after (red) adding 150 uL Triton/NaOH reagent to 50 uL of purified Ancα/β. In the presence of reagent, globins absorb at 400 nm. **d)** SEC of crude cell lysate after expression of AncMH (purple) and Ancα/β (black). Dashed lines, expected elution volumes for monomer (human myoglobin) and dimer (Ancα/β). **e)** Colorimetric hemoglobin concentration assay on collected SEC fractions of crude lysate (panel e**)** containing AncMH (purple) and Ancα/β (black). **f)** nMS of His-tagged AncMH at 70 μM, with monomer charge series indicated. *, cleavage product. Green, apo. Fractional occupancy of the monomeric form is shown. All experiments were performed once.

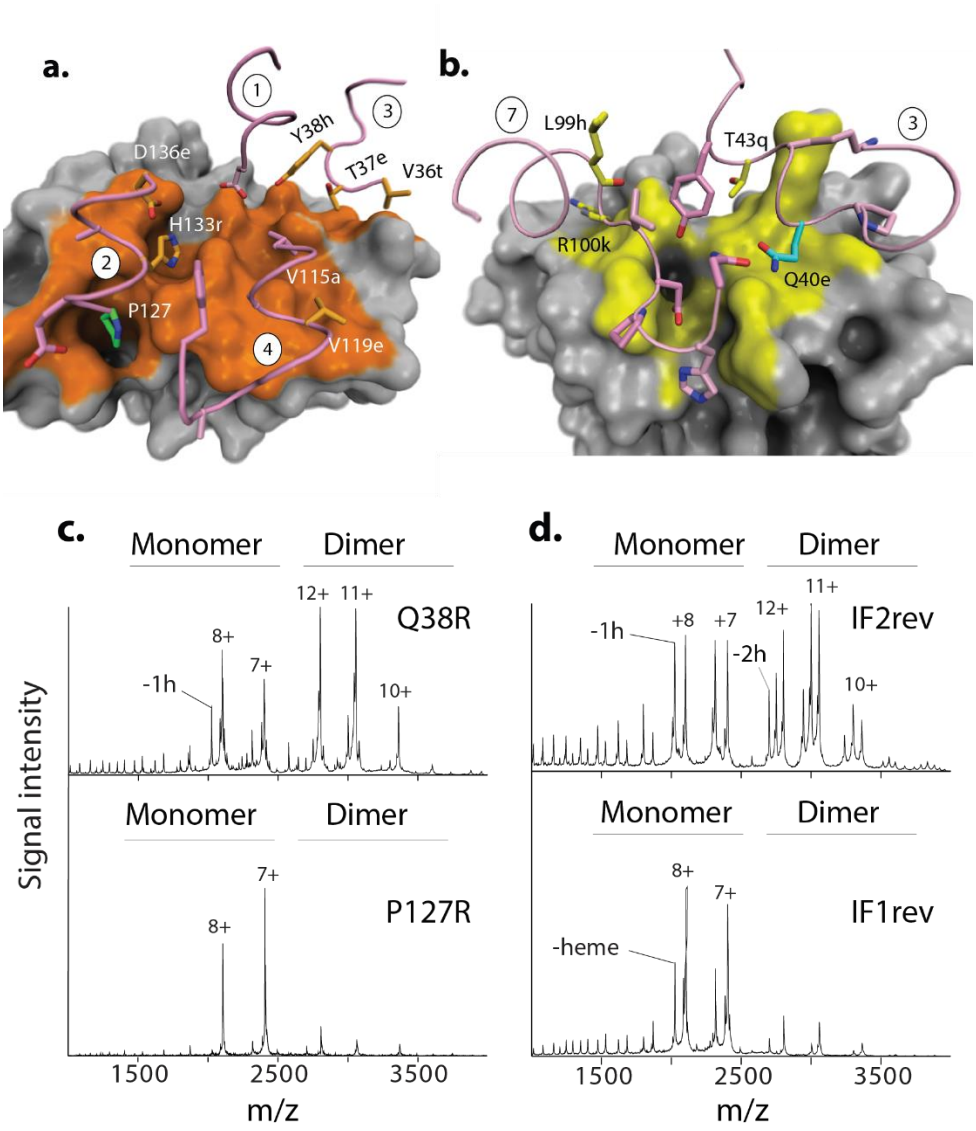


Extended Figure A5. HDX-MS of Ancα/β. a-c) Deuterium uptake measurements across time for three peptides, left vertical axis, raw deuterium incorporation; right vertical axis, deuterium incorporation divided by the total number of exchangeable amide hydrogens per peptide. Uptake curves for four concentrations by mutants IF1rev and P127R are shown. Each point shows mean with SE of 3 replicate measurements. **d-f)** Raw MS spectra for the peptides shown in a-c at 0.67 µM (red, at which the protein is monomeric), and 75 µM (purple, at which it is entirely dimeric: see Extended Figure 2). The traces are slightly offset to allow visualization. One replicate at each incubation time is shown. **g)** Amino acids 99 to 111 contact IF1 (orange) or IF2 (yellow). The homology model of one chain of Ancα/β (cartoon and sticks), was aligned to the α subunit of human Hb (PDB 1A3N); β subunits in are shown as surfaces. **h**) Normalized deuterium uptake difference (mean and SE from 3 replicates), defined as the uptake difference between monomer and dimer, divided by the uptake of the monomer, observed for peptides containing amino acids 99-111. Gray N-terminal residues do not contribute to uptake. Amino acid sequences are aligned and labeled (orange dots, IF1; yellow, IF2).

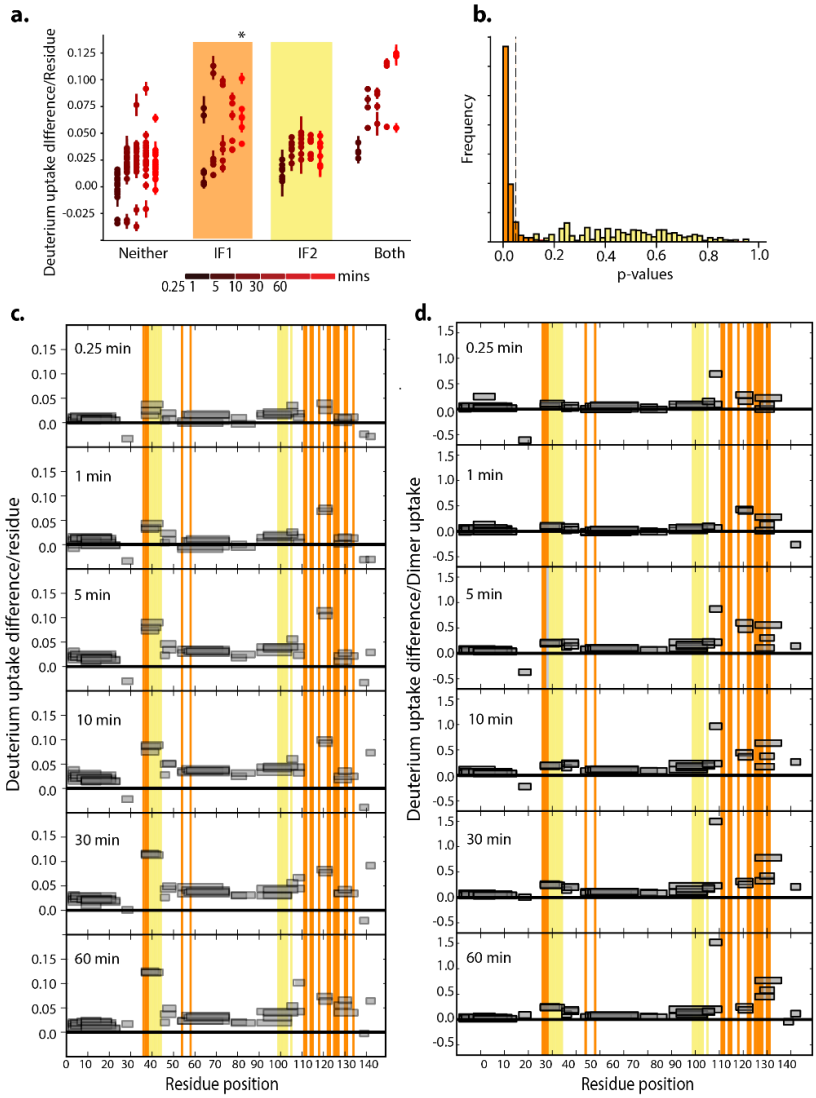


Extended Figure A6. Statistical analysis of HDX-MS results by peptides containing interface residues. a) Residues in human Hb (PDB 1A3N) that bury at least 50% of their surface area in either IF1 (orange) or IF2 (yellow) are shown as spheres. A subunits, red and pink; β, blue. **b)** Homology models of Ancα/β dimer across IF1 (left) or IF2 (right). Two subunits of Ancα/β were computationally docked using HADDOCK using the $\alpha 1/\beta 1$ inerface (IF1, left) or $\alpha 1/\beta 2$ subunits (IF2, right) of human Hb (1A3N) as a template. **c)** Coverage of peptides produced by trypsinization of Ancα/β, assessed by MS. Sites that bury surface area at IF1 and IF2 in the modeled dimeric structures are orange or yellow, respectively. **d)** Classification of trypsin-produced peptides that contribute to IF1 or IF2. Each circle represents one peptide, plotted by average surface area per residue buried at each interface (total buried area divided by total number of residues). Dashed line, cutoffs to classify peptides as contributing to IF1 (orange zone) or IF2 (yellow). **e,f)** Correlation between change in deuterium uptake and burial of surface area at IF1 or IF2. Each point is one of 47 peptides, plotted according to the normalized difference in deuterium uptake between concentrations at which monomer or dimer predominate (0.67 or 75 µM, normalized by uptake at 75 µM) and average buried surface area at IF1 or IF2. r, Pearson correlation coefficient. **g)** Permutation test to evaluate the difference in deuterium uptake at two time points by peptides containing IF1 vs. all other peptides (orange), or IF2 vs. all other peptides (yellow). To avoid non-independence, the experimental data were reduced to a set of nonoverlapping peptides by sampling without replacement. Peptides were categorized by whether they contained residues at IF1, IF2, or neither; peptides contributing to both IFs were excluded. For each interface, the mean uptake by peptides contributing to the interface was calculated, as was the mean uptake by peptides not in that category, and the difference in means was recorded. Peptide assignment to categories was then randomized, and the difference in mean uptake recorded; this permutation process was repeated until all possible randomized assignment schemes for those peptides had been sampled once. *P-value*, fraction of permuted assignment schemes with a difference in mean uptake between

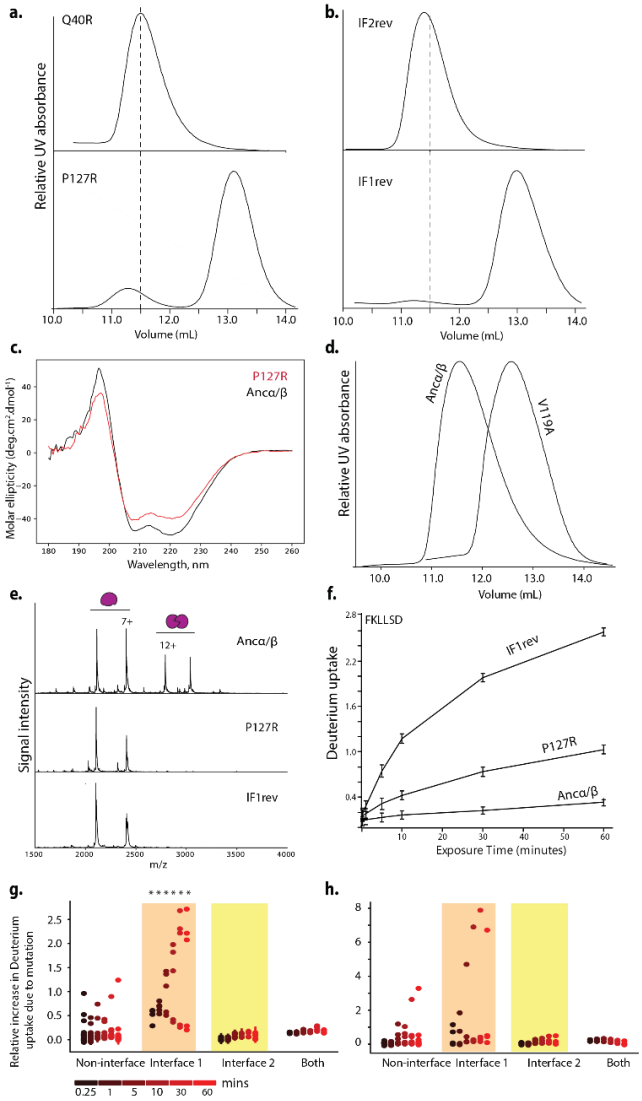
(**Extended Figure A6.** continued) categories greater than or equal to that from the true scheme. This process was repeated for 1000 nonoverlapping peptide sets; the histogram shows the frequency of *P-values* across these sets. Dotted line, *P*=0.05.



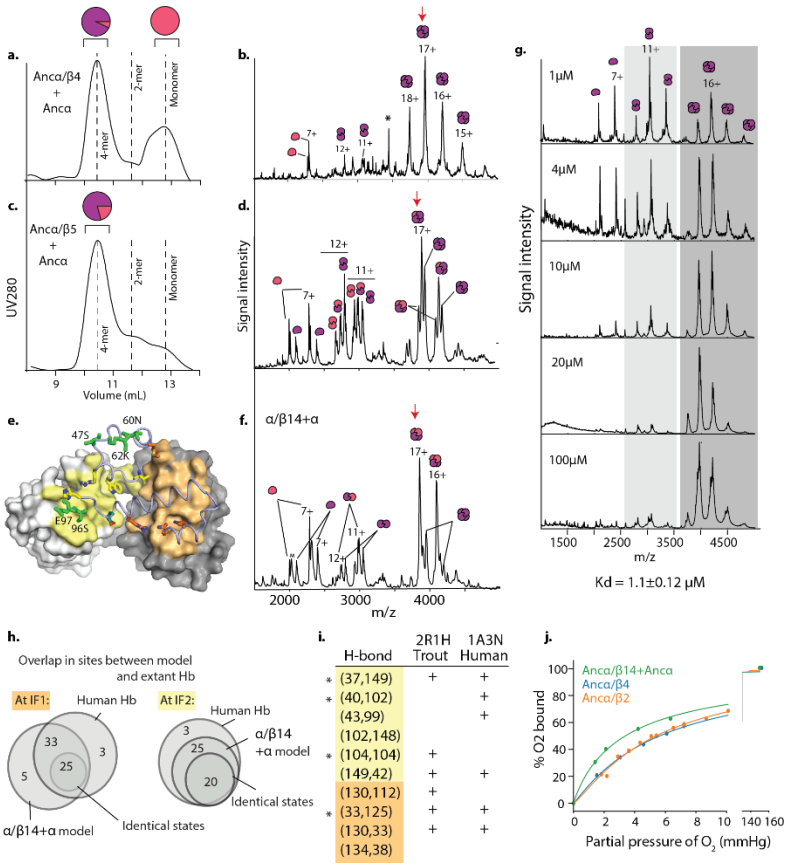
Extended Figure A7: Dissection of IF1 and IF2 by HDX-MS and mutagenesis. a,b) Peptides with residues contributing to IF1 (a) or IF2 (b) that have the largest relative uptake difference upon dimerization are shown as purple tubes. Sticks, side chains predicted to contact the other subunit (orange surface, IF1; yellow IF2). Side chains are colored orange or yellow (IF1 or IF2) if they were substituted between AncMH and Ancα/β; purple, unchanged in that interval; green, site for targeted mutation P127; blue, Q40. Circled numbers show the rank of each peptide among all peptides for the normalized difference in deuterium uptake between monomer and dimer conditions. Homology models of the Ancα/β dimer using half-tetramers of human Hb (1A3N) are shown. In panel a, the dimer is modeled using the α1/β1 subunits; in b, it is modeled on the α1/β2 subunits. **c,d**) nMS of interface mutants Q40R (at IF2) and P127R (at IF1) and for mutants IF1 and IF2, in which interface residues in Ancα/β were reverted to their states in AncMH. All assays at 20 µM. Stoichiometries and charge states are labelled. Unhemed peak series due to heme ejection during nMS is labeled. Spectra were collected once.



Extended Figure A8. Alternative methods to normalize deuterium uptake. a) Deuterium uptake difference between monomer (0.67µM) and dimer (75µM) at each time point was normalized by the length of each peptide. Peptides were categorized by the interface to which they contribute, as in Fig. 2c. *, interface peptide sets that have significantly increased uptake upon dilution when compared to peptides outside of that interface, as determined by a permutation test (see Extended Fig. 6). Each point shows the mean and SE from 3 replicates. **b)** Permutation test to evaluate the difference in deuterium uptake at 60 minutes by peptides at each interface, when uptake difference per peptide is normalized by length (using the methods described in Extended Fig. 6g). Orange, peptides with IF1-containing residues vs. those with no IF1 residues. Yellow, IF2-containing peptides vs. those with no IF2 residues. Dotted line, *P*=0.05. **c, d)** Average deuterium uptake difference per residue (**c**) and uptake difference normalized by dimer uptake (**d**) for peptides at different time points. IF1 sites (Orange), IF2 sites (Yellow). Each rectangle shows the position of the peptide in the linear sequence and its uptake (mean of 3 replicates).

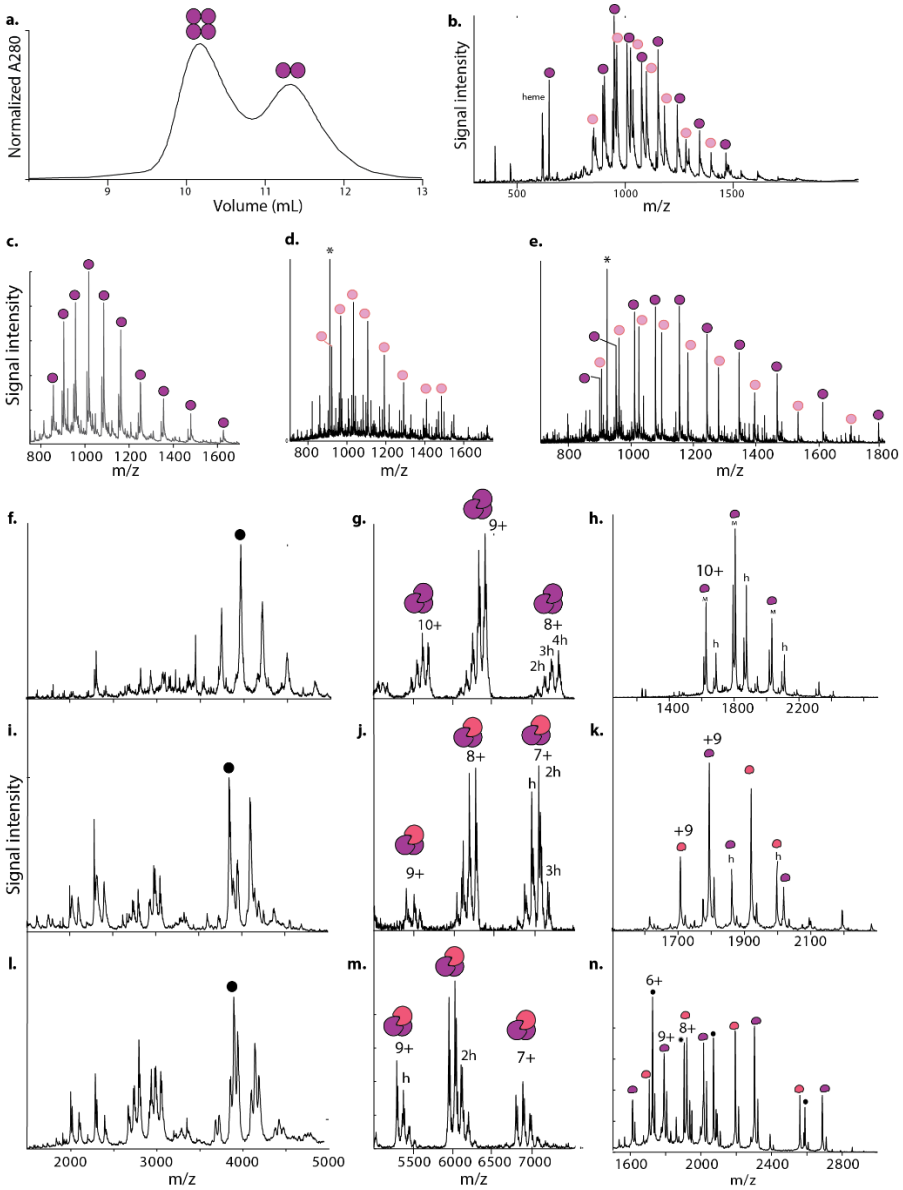


Extended Figure A9. Effect of interface-disrupting mutations on Ancα/β. a,b) SEC of mutants at IF2 (Q40R and IF2rev, which reverts all substitutions that occurred between AncMH and Ancα/β at IF2 sites) and at IF1 (P127R and IF1rev), at 100 µM. Dashed line, elution peak volume for Ancα/β. **c)** Circular dichroism spectra for P127R and Ancα/β, showing comparable helical structure. **d)** SEC from IF1 mutant V119A at 64 µM. **e)** nMS of Ancα/β, P127R and IF1rev at 10 µM. Stoichiometries and charges are shown. For a-d, nMS and SEC experiments were performed once per concentration. **f)** Normalized deuterium uptake by IF1-containing peptide 106-111 in HDX-MS of Ancα/β (75 µM) and mutants P127R (2 µM) and IF1rev (2 µM). Points and error bars, mean and SE of 3 replicates. **g,h).** Difference between deuterium uptake by each peptide in Ancα/β and uptake by the same peptide IF1 mutants P127R (**g**) and IF1 rev (**h**), both at 2 µM, normalized by uptake in Ancα/β. Peptides are classified by interface category. Circles and error bars, mean and SE of 3 replicates. *, peptide sets that have significantly increased relative uptake (by permutation test, see Extended Fig. 6) compared to all other peptides (peptides containing both IF1 and IF2 residues excluded).



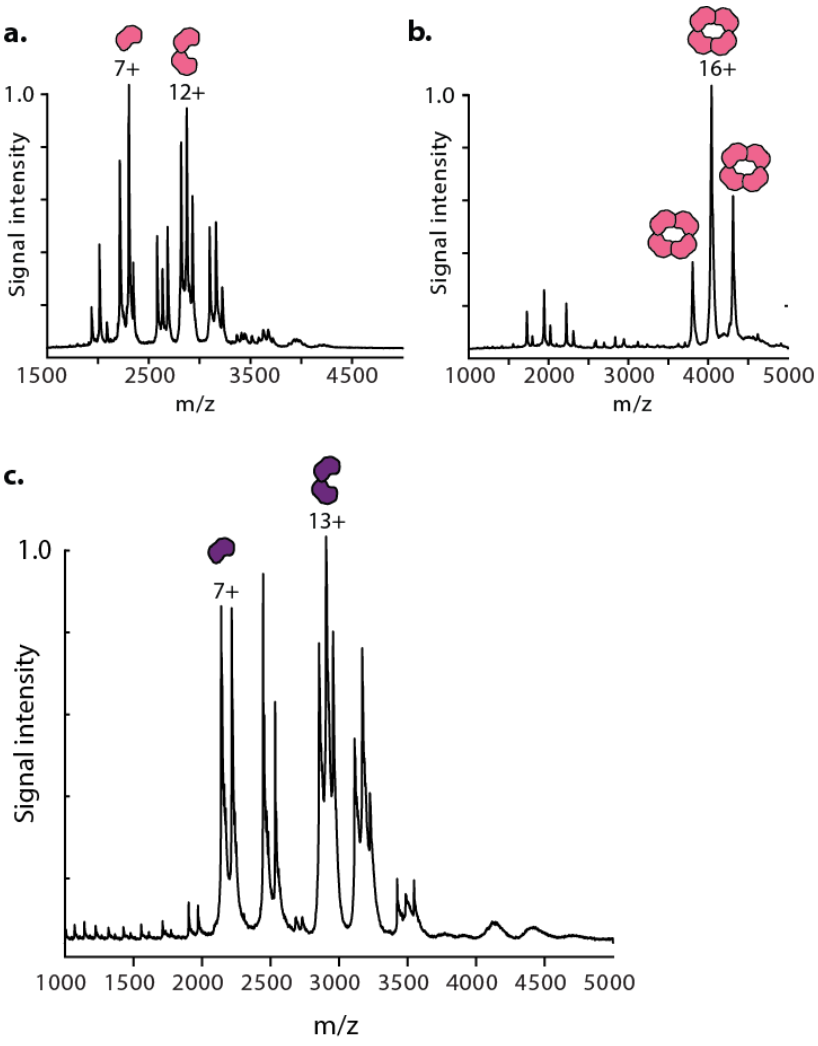
Extended Figure A10. Genetic mechanisms of tetramer evolution. a,c) SEC of Ancα/β containing sets of historical substitutions, when coexpressed and purified with Ancα. Vertical lines, elution volumes of known stoichiometries (4mer, Ancα +Ancβ; 2mer, Ancα/β; monomer, human myoglobin). Pie graphs, relative proportions of α (pink) and α/β mutant (purple) subunits in fractions corresponding to each peak, as determined by high resolution MS (Extended Figure 11). **b)** nMS of tetrameric fraction in **a.** at 20 μM (monomer concentration). Together, a and b show that tetramers formed by coexpression of Ancα/β4+ Ancα incorporate virtually no α subunit. Occupancy from this experiment is shown in Fig. 3b. **d, f)** nMS of unfractionated purified protein complexes of Ancα/β5+α and Ancα/β14+α at 20 μM. Charge series, stoichiometries indicated. *, apparent impurity. **e)** Homology model of Ancα/β14+α using Human Hb (1a3n) as template. Yellow and cyan sticks, Ancβ-lineage substitutions on IF2, orange sticks, Ancβ substitutions on IF1. Yellow surface, α IF2; Orange surface, α IF1. Green, 5 β substitutions close to the interfaces included in Ancα/β14+α. Red arrows, peaks isolated for further characterization by tandem MS (Ext. Fig. 11). **g)** nMS of Ancα/β2 across concentrations. Charge series and stoichiometries indicated. **h)** Similarity between interfaces in Ancα/β14+Ancα homology model and X-ray crystal structure of Human Hb. Venn diagrams show sites buried at IF1 and IF2 in one or both structures. Small circle, number of shared interface sites with identical amino acid state. **i)** Hydrogen-bond contacts at interfaces in Ancα/β14+α homology model are also found in X-ray crystal structures of extant hemoglobins. Residue pairs hydrogen-bonded in Ancα/β14+α IF2 (yellow) and IF1 (orange) are listed; +, also present in crystal

(**Extended Figure A10.** continued) structure. *, interactions discussed in the text of this paper. PDB identifiers are shown. **j.** Oxygen equilibrium curves of Ancα/β14+α, Ancα/β4, Ancα/β2. All experiments were performed once per concentration. Lines, best-fit curve by nonlinear regression.



Extended Figure A11. Stoichiometric characterization of Ancα/β containing historical substitutions. a) SEC of Ancα/β5. Circles show stoichiometry associated with each peak's elution volume. **b)** High-resolution accuracy mass spectrometry (HRA-MS) of Ancα/β5 + α. Purple circles label peaks associated with Ancα/β5; pink, Ancα. *, 922 m/z reference standard. **c)** HRA-MS of tetramer-containing SEC fraction of Ancα/β4+Ancα. **c)** HRA-MS of monomer-containing SEC fraction of Ancα/β4+Ancα. **e)** HRA-MS of Ancα/β9+Ancα. **f)** nMS of tetramer-containing SEC fraction of Ancα/β4+Ancα (see Fig. 3a,b). Black circle, most abundant peak used for tandem MS. **g)** Tandem MS of isolated most-abundant peak in **f**, showing trimer-containing peaks. Charge states and number of hemes (h) in the 8+ peak are indicated. **h)** monomer-containing peaks. **I, j ,k)** nMS (**i**) and tandem MS (**j, k**) of Ancα/β14+Ancα (see Fig. 3f). **l, m, n).** nMS and tandem MS of Ancα/β5+Ancα (see Fig. 3c,d). Black dots in (**n)** mark charge species produced by cleavage of Ancα/β5. All experiments were performed once.

Appendix B: Extended Figure for Chapter 3



Extended Figure B1. a. nMS of Ancα at 20 µM. Numbers, charge states of major peaks. Icons, Oligomeric states. **b.** nMS of Ancα+2 at 20 µM. Numbers, charge states of major peaks. Icons, Oligomeric states. **c.** nMS of Ancα/β with a His-tag at 20 µM. Numbers, charge states of major peaks. Icons, Oligomeric states.

BIBLIOGRAPHY

Ackers, G. K. 1980. “Energetics of Subunit Assembly and Ligand Binding in Human Hemoglobin.” *Biophysical Journal* 32 (1): 331–46. https://doi.org/10.1016/S0006-3495(80)84960-5.

Ahnert, Sebastian E., Joseph A. Marsh, Helena Hernández, Carol V. Robinson, and Sarah A. Teichmann. 2015. “Principles of Assembly Reveal a Periodic Table of Protein Complexes.” *Science* 350 (6266). https://doi.org/10.1126/science.aaa2245.

Aizawa, Shin Ichi. 2001. “Bacterial Flagella and Type III Secretion Systems.” *FEMS Microbiology Letters* 202 (2): 157–64. https://doi.org/10.1016/S0378-1097(01)00301-9.

Alexander, Patrick A., Yanan He, Yihong Chen, John Orban, and Philip N. Bryan. 2009. “A Minimal Sequence Code for Switching Protein Structure and Function.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (50): 21149–54. https://doi.org/10.1073/pnas.0906408106.

All, U T C. 1995. “The Origins of T . H . Huxley ’ s Saltationism : History in Darwin ’ s Shadow Author (s): Sherrie L . Lyons Source : Journal of the History of Biology , Autumn , 1995 , Vol . 28 , No . 3 (Autumn , Published by : Springer Stable URL : Https://Www.Jstor.” 28 (3): 463–94.

Anderson, Douglas P., Dustin S. Whitney, Victor Hanson-Smith, Arielle Woznica, William Campodonico-Burnett, Brian F. Volkman, Nicole King, Joseph W. Thornton, and Kenneth E. Prehoda. 2016. “Evolution of an Ancient Protein Function Involved in Organized Multicellularity in Animals.” *ELife* 5 (JANUARY2016): 1–20. https://doi.org/10.7554/eLife.10147.

Anil, Burcu, Rebecca Craig-Schapiro, and Daniel P. Raleigh. 2006. “Design of a Hyperstable Protein by Rational Consideration of Unfolded State Interactions.” *Journal of the American Chemical Society* 128 (10): 3144–45. https://doi.org/10.1021/ja057874b.

Anisimova, Maria, and Olivier Gascuel. 2006. “Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative.” *Systematic Biology* 55 (4): 539–52. https://doi.org/10.1080/10635150600755453.

Anisimova, Maria, Manuel Gil, Jean Franois Dufayard, Christophe Dessimoz, and Olivier Gascuel. 2011. “Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-Based Approximation Schemes.” *Systematic Biology* 60 (5): 685–99. https://doi.org/10.1093/sysbio/syr041.

Arnone A, Perutz M. 1974. “Structure of Inositol Hexaphosphate-Human Deoxyhaemoglobin Complex” 249 (1969): 195–97.

Arnone, Arthur. 1972. “X-Ray Diffraction Study of Binding of 2,3-Diphosphoglycerate to Human Deoxyhaemoglobin.” *Nature* 237 (5351): 146–49. https://doi.org/10.1038/237146a0.

Barends, Thomas R M, Lutz Foucar, Albert Ardevol, Karol Nass, Andrew Aquila, Sabine Botha, R

Bruce Doak, et al. 2015. “Direct Observation of Ultrafast Collective Motions in CO Myoglobin upon Ligand Dissociation.” *Science* 350 (6259).

Basu, Malay Kumar, Liran Carmel, Igor B. Rogozin, and Eugene V. Koonin. 2008. “Evolution of Protein Domain Promiscuity in Eukaryotes.” *Genome Research* 18 (3): 449–61. https://doi.org/10.1101/gr.6943508.

Bedian, Vahe. 1984. “The Origin of the Genetic Code.” *International Journal of Quantum Chemistry* 26 (11 S): 87–89. https://doi.org/10.1002/qua.560260711.

Behe, M. J. 1996. *Darwin’s Black Box: The Biochemical Challenge to Evolution*. Free Press New York.

Benesch, R., Benesch RE. 1968. “The Interaction of Hemoglobin and Its Subunits with 2,3-Diphosphoglycerate.” *Biochemistry* 61: 1102–6.

Berenbrink, Michael. 2007. “Historical Reconstructions of Evolving Physiological Complexity: O 2 Secretion in the Eye and Swimbladder of Fishes.” *Journal of Experimental Biology* 210 (9): 1641–52. https://doi.org/10.1242/jeb.003319.

Bergendahl, L. Therese, and Joseph A. Marsh. 2017. “Functional Determinants of Protein Assembly into Homomeric Complexes.” *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-05084-8.

Blank, Miriam, Laurent Kiger, Anke Thielebein, Frank Gerlach, Thomas Hankeln, Michael C Marden, and Thorsten Burmester. 2011a. “Oxygen Supply from the Bird’s Eye Perspective: Globin E Is a Respiratory Protein in the Chicken Retina” 286 (30): 26507–15. https://doi.org/10.1074/jbc.M111.224634.

Bonaventura, Celia, and Joseph Bonaventura. 1980. “Anionic Control of Function in Vertebrate Hemoglobins.” *Integrative and Comparative Biology* 20 (1): 131–38. https://doi.org/10.1093/icb/20.1.131.

Bragg, Lawrence William; Perutz, Max Ferdinand; 1952. “The Structure of Haemoglobin.” *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 213 (July): 425–35.

Bungard, Dixie, Jacob S. Copple, Jing Yan, Jimmy J. Chhun, Vlad K. Kumirov, Scott G. Foy, Joanna Masel, Vicki H. Wysocki, and Matthew H.J. Cordes. 2017. “Foldability of a Natural De Novo Evolved Protein.” *Structure* 25 (11): 1687–1696.e4. https://doi.org/10.1016/j.str.2017.09.006.

Bunn, Franklin. 2019. “Subunit Assembly of Hemoglobin: Of Hematologic An Important Phenotype Determinant.” *The Journal of the American Society of Hematology* 1 (1987): 1–6.

Burkhardt, Richard W. 2013. “Lamarck, Evolution, and the Inheritance of Acquired Characters” 194 (August): 793–805. https://doi.org/10.1534/genetics.113.151852.

Cai, Jing, Ruoping Zhao, Huifeng Jiang, and Wen Wang. 2008. “De Novo Origination of a New

Protein-Coding Gene in Saccharomyces Cerevisiae." *Genetics* 179 (1): 487–96. https://doi.org/10.1534/genetics.107.084491.

Charlesworth, Brian. 1982. "Review : Hopeful Monsters Cannot Fly Reviewed Work (s): The Material Basis of Evolution . by Richard B . Goldschmidt" 8 (4): 469–74.

Chen, Chao Sheng, Callum Smits, Guy G. Dodson, Mikhail B. Shevtsov, Natalie Merlino, Paul Gollnick, and Alfred A. Antson. 2011. "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* 6 (10). https://doi.org/10.1371/journal.pone.0025296.

Clifton, Ben E., Joe A. Kaczmarski, Paul D. Carr, Monica L. Gerth, Nobuhiko Tokuriki, and Colin J. Jackson. 2018. "Evolution of Cyclohexadienyl Dehydratase from an Ancestral Solute-Binding Protein Article." *Nature Chemical Biology* 14 (6): 542–47. https://doi.org/10.1038/s41589-018-0043-2.

Coates, Michael L. 1975. "Hemoglobin Function in the Vertebrates: An Evolutionary Model." *Journal of Molecular Evolution* 6 (4): 285–307. https://doi.org/10.1007/BF01794636.

Coleman, Ryan G., and Kim A. Sharp. 2010. "Shape and Evolution of Thermostable Protein Structure." *Proteins: Structure, Function and Bioinformatics* 78 (2): 420–33. https://doi.org/10.1002/prot.22558.

Cong, Xiao, Yang Liu, Wen Liu, Xiaowen Liang, David H. Russell, and Arthur Laganowsky. 2016. "Determining Membrane Protein-Lipid Binding Thermodynamics Using Native Mass Spectrometry." *Journal of the American Chemical Society* 138 (13): 4346–49. https://doi.org/10.1021/jacs.6b01771.

Cordes, Matthew H.J., Randall E. Burton, Nathan P. Walsh, C. James McKnight, and Robert T. Sauer. 2000. "An Evolutionary Bridge to a New Protein Fold." *Nature Structural Biology* 7 (12): 1129–32. https://doi.org/10.1038/81985.

Costa, Fernando Ferreira, and Maria Emília Favero. 2011. "Alpha-Hemoglobin-Stabilizing Protein: An Erythroid Molecular Chaperone." *Biochemistry Research International* 2011. https://doi.org/10.1155/2011/373859.

Coyle, Scott M., Jonathan Flores, and Wendell A. Lim. 2013. "Exploitation of Latent Allostery Enables the Evolution of New Modes of MAP Kinase Regulation." *Cell* 154 (4): 875–87. https://doi.org/10.1016/j.cell.2013.07.019.

Daniel N. Osherson, Don Scarborough, Saul Sternberg. n.d. *An Invitation to Cognitive Science*.

Darwin, Charles. 1859. "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London." *On the Origin of Species by Means of Natural Selection*, no. February 2009: 204–8.

Davidson, E H, D H Erwin, K S Zaret, G A Wray, B Li, M Carey, J L Workman, et al. 2008. "Surface Sites for Engineering Allosteric Control in Proteins." *Science* 322 (October): 438–42.

Dawkins, Richard. 1997. *Climbing Mount Improbable*. WW Norton & Company.

Deckert, Katelyn, S. Jimmy Budiardjo, Luke C. Brunner, Scott Lovell, and John Karanicolas. 2012. "Designing Allosteric Control into Enzymes by Chemical Rescue of Structure." *Journal of the American Chemical Society* 134 (24): 10055–60. https://doi.org/10.1021/ja301409g.

Dröge, Jasmin, Amit Pande, Ella W. Englander, and Wojciech Makałowski. 2012. "Comparative Genomics of Neuroglobin Reveals Its Early Origins." *PLoS ONE* 7 (10). https://doi.org/10.1371/journal.pone.0047972.

Eisenberg, D, and A D McLachlan. 1986. "Solvation Energy in Protein Folding and Stability." *Nature* 319 (6050): 199–203.

Engelhart, Johann Friedrich. 1825. *EnglishDeutschImpressum Commentatio de Vera Materiae Sanguini Purpureum Colorem Impertientis Natura*.

Eswara, Manoja B K, Andrew T Mcguire, Jacqueline B Pierce, Dev Mangroo, Ema Kikovska, Staffan G Svärd, Leif a Kirsebom, et al. 2007. "Eukaryotic RNase P RNA Mediates Cleavage in the Absence of Protein." *Proceedings of the National Academy of Sciences of the United States of America* 104 (7): 2062–67. https://doi.org/10.1073/pnas.0607326104.

Fago, Angela, Kim Rohlfing, Elin E. Petersen, Agnieszka Jendroszek, and Thorsten Burmester. 2018. "Functional Diversification of Sea Lamprey Globins in Evolution and Development." *Biochimica et Biophysica Acta - Proteins and Proteomics* 1866 (2): 283–91. https://doi.org/10.1016/j.bbapap.2017.11.009.

Fersht, Alan R., Jian Ping Shi, Jack Knill-Jones, Denise M. Lowe, Anthony J. Wilkinson, David M. Blow, Peter Brick, Paul Carter, Mary M.Y. Waye, and Greg Winter. 1985. "Hydrogen Bonding and Biological Specificity Analysed by Protein Engineering." *Nature* 314 (6008): 235–38. https://doi.org/10.1038/314235a0.

Field, Steven F., and Mikhail V. Matz. 2010. "Retracing Evolution of Red Fluorescence in GFP-like Proteins from Faviina Corals." *Molecular Biology and Evolution* 27 (2): 225–33. https://doi.org/10.1093/molbev/msp230.

Finnigan, Gregory C., Victor Hanson-Smith, Tom H. Stevens, and Joseph W. Thornton. 2012. "Evolution of Increased Complexity in a Molecular Machine." *Nature* 481 (7381): 360–64. https://doi.org/10.1038/nature10724.

Fisher, Ronald Aylmer. 1958. *The Genetical Theory of Natural Selection.* Рипол Классик.

Fletcher, Daniel a, and R Dyche Mullins. 2010. "Cell Mechanisms and Cytoskeleton." *Nature* 463 (7280): 485–92. https://doi.org/10.1038/nature08908.Cell.

Fowler, Douglas M., and Stanley Fields. 2014a. "Deep Mutational Scanning: A New Style of Protein Science." *Nature Methods* 11 (8): 801–7. https://doi.org/10.1038/nmeth.3027.

———. 2014b. "Deep Mutational Scanning: A New Style of Protein Science." *Nature Methods* 11 (8): 801–7. https://doi.org/10.1038/nmeth.3027.

Fraser, Nicholas J., Jian Wei Liu, Peter D. Mabbitt, Galen J. Correy, Chris W. Coppin, Mathilde Lethier, Matthew A. Perugini, et al. 2016. "Evolution of Protein Quaternary Structure in

Response to Selective Pressure for Increased Thermostability." *Journal of Molecular Biology* 428 (11): 2359–71. https://doi.org/10.1016/j.jmb.2016.03.014.

Galperin, Michael Y., and Eugene V. Koonin. 2012. "Divergence and Convergence in Enzyme Evolution." *Journal of Biological Chemistry* 287 (1): 21–28. https://doi.org/10.1074/jbc.R111.241976.

Garcia-Seisdedos, Hector, Charly Empereur-Mot, Nadav Elad, and Emmanuel D. Levy. 2017. "Proteins Evolve on the Edge of Supramolecular Self-Assembly." *Nature* 548 (7666): 244–47. https://doi.org/10.1038/nature23320.

Gelin, Bruce R., Angel Wai Mun Lee, and Martin Karplus. 1983. "Hemoglobin Tertiary Structural Change on Ligand Binding Its Role in the Co-Operative Mechanism." *Journal of Molecular Biology* 171 (4): 489–559. https://doi.org/10.1016/0022-2836(83)90042-6.

Gillespie, John H. 2014. "Molecular Evolution Over the Mutational Landscape" 38 (5): 1116–29.

Goldschmidt, Richard. 1982. *The Material Basis of Evolution.* Yale University Press,.

Goodman, Morris. 1981. "Globin Evolution Was Apparently Very Rapid in Early Vertebrates: A Reasonable Case against the Rate-Constancy Hypothesis." *Journal of Molecular Evolution* 17 (2): 114–20. https://doi.org/10.1007/BF01732683.

Goodman, Morris, and G. William Moore. 1973. "Phylogeny of Hemoglobin." *Systematic Zoology* 22 (4): 508–32. https://doi.org/10.2307/2412957.

Goodsell, David S, and Arthur J Olson. 2000. "Structural Symmetry and Protein Function." *Annual Review of Biophysics and Biomolecular Structure* 29 (1): 105–53.

Gray, Michael W., Julius Lukeš, John M. Archibald, Patrick J. Keeling, and W. Ford Doolittle. 2010. "Irremediable Complexity?" *Science* 330 (6006): 920–21. https://doi.org/10.1126/science.1198594.

Grispo, Michael T., Chandrasekhar Natarajan, Joana Projecto-Garcia, Hideaki Moriyama, Roy E. Weber, and Jay F. Storz. 2012. "Gene Duplication and the Evolution of Hemoglobin Isoform Differentiation in Birds." *Journal of Biological Chemistry* 287 (45): 37647–58. https://doi.org/10.1074/jbc.M112.375600.

Grueninger, Dirk, Nora Treiber, Mathias O. P. Ziegler, Jochen W. A. Koettr, Monika-Sarah Schulze, and Georg E. Schulz. 2008. "Designed Protein-Protein Association." *Science* 319 (January): 206–10.

Guindon, Stéphane, Jean Franois Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3): 307–21. https://doi.org/10.1093/sysbio/syq010.

Heaslet, Holly A., and William E. Royer. 1999. "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* 7 (5): 517–26. https://doi.org/10.1016/S0969-

2126(99)80068-9.

Heinzelman, Pete, John Krais, Eliza Ruben, and Robert Pantazes. 2015. “Engineering PH Responsive Fibronectin Domains for Biomedical Applications.” *Journal of Biological Engineering* 9 (1): 1–11. https://doi.org/10.1186/s13036-015-0004-1.

Herron, Matthew D., Joshua M. Borin, Jacob C. Boswell, Jillian Walker, I. Chen Kimberly Chen, Charles A. Knox, Margrethe Boyd, Frank Rosenzweig, and William C. Ratcliff. 2019. “De Novo Origins of Multicellularity in Response to Predation.” *Scientific Reports* 9 (1): 1–9. https://doi.org/10.1038/s41598-019-39558-8.

Hochberg, Georg K.A., Yang Liu, Erik G. Marklund, Brian P.H. Metzger, Arthur Laganowsky, and Joseph W. Thornton. 2020. “A Hydrophobic Ratchet Entrenches Molecular Complexes.” *Nature* 588 (7838): 503–8. https://doi.org/10.1038/s41586-020-3021-2.

Hochberg, Georg K.A., Dale A. Shepherd, Erik G. Marklund, Indu Santhanagoplan, Matteo T. Degiacomi, Arthur Laganowsky, Timothy M. Allison, et al. 2018. “Structural Principles That Enable Oligomeric Small Heat-Shock Protein Paralogs to Evolve Distinct Functions.” *Science* 359 (6378): 930–35. https://doi.org/10.1126/science.aam7229.

Hoffman, S J, D L Looker, J M Roehrich, P E Cozart, S L Durfee, J L Tedesco, and G L Stetler. 1990. “Expression of Fully Functional Tetrameric Human Hemoglobin in Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (21): 8521–25.

Hoffmann, Federico G., Juan C. Opazo, and Jay F. Storz. 2011. “Differential Loss and Retention of Cytoglobin, Myoglobin, and Globin-E during the Radiation of Vertebrates.” *Genome Biology and Evolution* 3 (1): 588–600. https://doi.org/10.1093/gbe/evr055.

Ikeda, Yoshitaka, Takehiko Tanaka, and Tamio Noguchi. 1997. “Conversion of Non-Allosteric Pyruvate Kinase Isozyme into an Allosteric Enzyme by a Single Amino Acid Substitution.” *Journal of Biological Chemistry* 272 (33): 20495–501. https://doi.org/10.1074/jbc.272.33.20495.

Imai, K. 1982. *Allosteric Effects in Haemoglobin*. Cambridge, UK: Cambridge University Press.

Imaizumi, Kazuhiko, Kiyohiro Imai, Itiro Tyuma, and J. Biochem. 1979. “The Linkage between the Four-Step Binding of Oxygen and the Binding of Heterotropic Anionic Ligands in Hemoglobin.” *Journal of Biochemistry* 86 (6): 1829–40. https://doi.org/10.1093/oxfordjournals.jbchem.a132705.

Isaacks, Russell E., and Donald R. Harkness. 1980. “Erythrocyte Organic Phosphates and Hemoglobin Function in Birds, Reptiles, and Fishes.” *Integrative and Comparative Biology* 20 (1): 115–29. https://doi.org/10.1093/icb/20.1.115.

Jacob, F. 1977. “Evolution and Tinkering.” *Science*.

Jee, Jun Goo, In Ja L. Byeon, John M. Louis, and Angela M. Gronenborn. 2008. “The Point Mutation A34F Causes Dimerization of GB1.” *Proteins: Structure, Function and Genetics* 71

(3): 1420–31. https://doi.org/10.1002/prot.21831.

Kappeler, Peter M., Tim Clutton-Brock, Susanne Shultz, and Dieter Lukas. 2019. “Social Complexity: Patterns, Processes, and Evolution.” *Behavioral Ecology and Sociobiology* 73 (1): 1–6. https://doi.org/10.1007/s00265-018-2613-4.

Kato, Sanae, Takashi Matsui, Christos Gatsogiannis, and Yoshikazu Tanaka. 2018. “Molluscan Hemocyanin: Structure, Evolution, and Physiology.” *Biophysical Reviews* 10 (2): 191–202. https://doi.org/10.1007/s12551-017-0349-4.

Katoh, Kazutaka, John Rozewicki, and Kazunori D. Yamada. 2017. “MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization.” *Briefings in Bioinformatics*, no. July: 1–7. https://doi.org/10.1093/bib/bbx108.

Keefe, A. D., and J. W. Szostak. 2001. “Functional Proteins from a Random-Sequence Library.” *Nature* 410 (6829): 715–18. https://doi.org/10.1038/35070613.

Keightley, Peter D., Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L. Blaxter. 2009. “Analysis of the Genome Sequences of Three Drosophila Melanogaster Spontaneous Mutation Accumulation Lines.” *Genome Research* 19 (7): 1195–1201. https://doi.org/10.1101/gr.091231.109.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C. 1960. “Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution.” *Nature* 185 (4711): 422.

Kendrew Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C., J C. 1960. “Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution.” *Nature* 185 (4711): 422.

Kidd, Richard D., Heather M. Baker, Antony J. Mathews, Thomas Brittain, and Edward N. Baker. 2002. “ Oligomerization and Ligand Binding in a Homotetrameric Hemoglobin: Two High-Resolution Crystal Structures of Hemoglobin Bart’s (γ 4), a Marker for α-Thalassemia .” *Protein Science* 10 (9): 1739–49. https://doi.org/10.1110/ps.11701.

Kimura, M. 1980. “Average Time until Fixation of a Mutant Allele in a Finite Population under Continued Mutation Pressure: Studies by Analytical, Numerical, and Pseudo-Sampling Methods.” *Proceedings of the National Academy of Sciences of the United States of America* 77 (1): 522–26. https://doi.org/10.1073/pnas.77.1.522.

KIMURA, M. 1962. “On the Probability of Fixation of Mutant Genes in a Population.” *Genetics* 47 (391): 713–19.

Knapp, James E., Marcos A. Oliveira, Qiang Xie, Stephen R. Ernst, Austen F. Riggs, and Marvin L. Hackert. 1999. “The Structural and Functional Analysis of the Hemoglobin D Component from Chicken.” *Journal of Biological Chemistry* 274 (10): 6411–20. https://doi.org/10.1074/jbc.274.10.6411.

Komiyama, N. Hennakao, Gentaro Miyazaki, and Jeremy Tamef. 1995. “Transplanting a Unique

Allosteric Effect from Crocodile into Human Haemoglobin." *Nature* 373 (6511): 244–46. https://doi.org/10.1038/373244a0.

Koonin, Eugene V., and Artem S. Novozhilov. 2017. "Origin and Evolution of the Universal Genetic Code." *Annual Review of Genetics* 51 (2): 45–62. https://doi.org/10.1146/annurev-genet-120116-024713.

Kortemme, Tanja, David E. Kim, and David Baker. 2004. "Computational Alanine Scanning of Protein-Protein Interfaces." *Science's STKE : Signal Transduction Knowledge Environment* 2004 (219): 1–9. https://doi.org/10.1126/stke.2192004pl2.

Krissinel, Evgeny, and Kim Henrick. 2007. "Inference of Macromolecular Assemblies from Crystalline State." *Journal of Molecular Biology* 372 (3): 774–97. https://doi.org/10.1016/j.jmb.2007.05.022.

Kumar, Kaavya Krishna, David A. Jacques, J. Mitchell Guss, and David A. Gell. 2014a. "The Structure of α-Haemoglobin in Complex with a Haemoglobin-Binding Domain from Staphylococcus Aureus Reveals the Elusive α-Haemoglobin Dimerization Interface." *Acta Crystallographica Section F:Structural Biology Communications* 70 (8): 1032–37. https://doi.org/10.1107/S2053230X14012175.

Kurosky, A., D. R. Barnett, T. H. Lee, B. Touchstone, R. E. Hay, M. S. Arnott, B. H. Bowman, and W. M. Fitch. 1980. "Covalent Structure of Human Haptoglobin: A Serine Protease Homolog." *Proceedings of the National Academy of Sciences of the United States of America* 77 (6 I): 3388–92. https://doi.org/10.1073/pnas.77.6.3388.

Laub, Michael T. 2016. "Promiscuous Intermediates" 163 (3): 594–606. https://doi.org/10.1016/j.cell.2015.09.055.Evolving.

Le, Si Quang, and Olivier Gascuel. 2008. "An Improved General Amino Acid Replacement Matrix." *Molecular Biology and Evolution* 25 (7): 1307–20. https://doi.org/10.1093/molbev/msn067.

Leander, Megan, Yuchen Yuan, Anthony Meger, Qiang Cui, and Srivatsan Raman. 2020. "Functional Plasticity and Evolutionary Adaptation of Allosteric Regulation." *Proceedings of the National Academy of Sciences of the United States of America* 117 (41): 25445–54. https://doi.org/10.1073/pnas.2002613117.

Lechauve, Christophe, Cédric Chauvierre, Sylvia Dewilde, Luc Moens, Brian N. Green, Michael C. Marden, Chantal Célier, and Laurent Kiger. 2010. "Cytoglobin Conformations and Disulfide Bond Formation." *FEBS Journal* 277 (12): 2696–2704. https://doi.org/10.1111/j.1742-4658.2010.07686.x.

Levy, Emmanuel D. 2010. "A Simple Definition of s Tructural Regions in Proteins and Its Use in Analyzing Interface Evolution." *Journal of Molecular Biology* 403 (4): 660–70. https://doi.org/10.1016/j.jmb.2010.09.028.

Levy, Emmanuel D., Elisabetta Boeri Erba, Carol V. Robinson, and Sarah A. Teichmann. 2008. "Assembly Reflects Evolution of Protein Complexes." *Nature* 453 (7199): 1262–65.

https://doi.org/10.1038/nature06942.

Liljas, A., and M. Laurberg. 2000. “A Wheel Invented Three Times. The Molecular Structures of the Three Carbonic Anhydrases.” *EMBO Reports* 1 (1): 16–17. https://doi.org/10.1093/embo-reports/kvd016.

Link, A. James, Marissa L. Mock, and David A. Tirrell. 2003. “Non-Canonical Amino Acids in Protein Engineering.” *Current Opinion in Biotechnology* 14 (6): 603–9. https://doi.org/10.1016/j.copbio.2003.10.011.

Liu et al., Howard. 2007. “Stepwise Formation of the Bacterial Flagellar System.” *Proceedings of the National Academy of Sciences* 104 (17): 7116–21.

Liu, Sen, Shiyong Liu, Xiaolei Zhu, Huanhuan Liang, Aoneng Cao, Zhijie Chang, and Luhua Lai. 2007. “Nonnatural Protein-Protein Interaction-Pair Design by Key Residues Grafting.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (13): 5330–35. https://doi.org/10.1073/pnas.0606198104.

Lu, Shaoyong, Mingfei Ji, Duan Ni, and Jian Zhang. 2018. “Discovery of Hidden Allosteric Sites as Novel Targets for Allosteric Drug Design.” *Drug Discovery Today* 23 (2): 359–65. https://doi.org/10.1016/j.drudis.2017.10.001.

Lynch, M. 2013. “Evolutionary Diversification of the Multimeric States of Proteins.” *Proceedings of the National Academy of Sciences* 110 (30): E2821–28. https://doi.org/10.1073/pnas.1310980110.

Lynch, Michael. 2007. “The Frailty of Adaptive Hypotheses for the Origins of Organismal Complexity.” *In the Light of Evolution* 1 (Table 1): 83–103. https://doi.org/10.17226/11790.

Makino, Masatomo, Hiroshi Sugimoto, Hitomi Sawai, Norifumi Kawada, Katsutoshi Yoshizato, and Yoshitsugu Shiro. 2006. “High-Resolution Structure of Human Cytoglobin: Identification of Extra N- and C-Termini and a New Dimerization Mode.” *Acta Crystallographica Section D: Biological Crystallography* 62 (6): 671–77. https://doi.org/10.1107/S0907444906013813.

Mallik, Saurav, and Dan S Tawfik. 2020. “Loss of Homomeric Interactions and Heteromers Formation Is the Long-Term Fate of Duplicated Homomers,” 1–22.

Manning, Lois R., Antoine Dumoulin, W. Terry Jenkins, Robert M. Winslow, and James M. Manning. 1999. “Determining Subunit Dissociation Constants in Natural and Recombinant Proteins.” *Methods in Enzymology* 306: 113–29. https://doi.org/10.1016/S0076-6879(99)06008-5.

Marsh, Joseph A., Helena Hernández, Zoe Hall, Sebastian E. Ahnert, Tina Perica, Carol V. Robinson, and Sarah A. Teichmann. 2013. “Protein Complexes Are under Evolutionary Selection to Assemble via Ordered Pathways.” *Cell* 153 (2): 461–70. https://doi.org/10.1016/j.cell.2013.02.044.

Marsh, Joseph A., and Sarah A. Teichmann. 2014. “Protein Flexibility Facilitates Quaternary

Structure Assembly and Evolution." *PLoS Biology* 12 (5). https://doi.org/10.1371/journal.pbio.1001870.

Marty, Michael T., Andrew J. Baldwin, Erik G. Marklund, Georg K.A. Hochberg, Justin L.P. Benesch, and Carol V. Robinson. 2015. "Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles." *Analytical Chemistry* 87 (8): 4370–76. https://doi.org/10.1021/acs.analchem.5b00140.

McClune, Conor J., Aurora Alvarez-Buylla, Christopher A. Voigt, and Michael T. Laub. 2019. "Engineering Orthogonal Signalling Pathways Reveals the Sparse Occupancy of Sequence Space." *Nature* 574 (7780): 702–6. https://doi.org/10.1038/s41586-019-1639-8.

McShea, Daniel W. 1991. "Complexity and Evolution: What Everybody Knows." *Biology and Philosophy* 6 (3): 303–24. https://doi.org/10.1007/BF00132234.

Meier, Sebastian, Pernille R. Jensen, Charles N. David, Jarrod Chapman, Thomas W. Holstein, Stephan Grzesiek, and Suat Özbek. 2007. "Continuous Molecular Evolution of Protein-Domain Structures by Single Amino Acid Changes." *Current Biology* 17 (2): 173–78. https://doi.org/10.1016/j.cub.2006.10.063.

Mihailescu, M.-R., and I. M. Russu. 2002. "A Signature of the T -> R Transition in Human Hemoglobin." *Proceedings of the National Academy of Sciences* 98 (7): 3773–77. https://doi.org/10.1073/pnas.071493598.

Monod, J., Wyman, J., & Changeux, J. P. 1965. "On the Nature of Allosteric Transitions: A Plausible Model." *Jounal of Molecular Biology* 12 (1): 88–118.

Moparthi, Vamsi K., and Cecilia Hägerhäll. 2011. "The Evolution of Respiratory Chain Complex i from a Smaller Last Common Ancestor Consisting of 11 Protein Subunits." *Journal of Molecular Evolution* 72 (5–6): 484–97. https://doi.org/10.1007/s00239-011-9447-2.

Mouche, Fabrice, Nicolas Boisset, and Pawel A Penczek. 2001. "Lumbricus Terrestris Hemoglobin — The Architecture of Linker Chains and Structural Variation of the Central Toroid." *Journal of Structural Biology* 192 (133): 176–92. https://doi.org/10.1006/jsbi.2001.4362.

Mu, Xin, Seongil Choi, Lisa Lang, David Mowray, Nikolay V. Dokholyan, Jens Danielsson, and Mikael Oliveberg. 2017. "Physicochemical Code for Quinary Protein Interactions in Escherichia Coli." *Proceedings of the National Academy of Sciences of the United States of America* 114 (23): E4556–63. https://doi.org/10.1073/pnas.1621227114.

Natarajan, Chandrasekhar, Federico G. Hoffmann, Roy E. Weber, Angela Fago, Christopher C. Witt, and Jay F. Storz. 2016. "Predictable Convergence in Hemoglobin Function Has Unpredictable Molecular Underpinnings." *Science* 354 (6310): 336–39. https://doi.org/10.1126/science.aaf9070.

Natarajan, Chandrasekhar, Xiaoben Jiang, Angela Fago, Roy E. Weber, Hideaki Moriyama, and Jay F. Storz. 2011. "Expression and Purification of Recombinant Hemoglobin in Escherichia Coli." *PLoS ONE* 6 (5): 1–7. https://doi.org/10.1371/journal.pone.0020176.

Orr, H. Allen. 2002. “The Population Genetics of Adaptation: The Adaptation of DNA Sequences.” *Evolution* 56 (7): 1317–30. https://doi.org/10.1111/j.0014-3820.2002.tb01446.x.

Paley, William. 1851. *Natural Theology: Or, Evidences of the Existence and Attributes of the Deity, Collected from the Appearances of Nature*.

Pauling, Linus, Harvey A Itano, S J Singer, Ibert C Wells, Linus Pauling, Harvey A Itano, S J Singer, and Ibert C Wells. 2019. “Sickle Cell Anemia , a Molecular Disease.” *Science* 110 (2865): 543–48.

Pereira-Leal, Jose B., Emmanuel D. Levy, Christel Kamp, and Sarah A. Teichmann. 2007. “Evolution of Protein Complexes by Duplication of Homomeric Interactions.” *Genome Biology* 8 (4). https://doi.org/10.1186/gb-2007-8-4-r51.

Perica, Tina, Yasushi Kondo, Shya P. Tiwari, Stephen H. McLaughlin, Katherine R. Kemplen, Xiuwei Zhang, Annette Steward, Nathalie Reuter, Jane Clarke, and Sarah A. Teichmann. 2014. “Evolution of Oligomeric State through Allosteric Pathways That Mimic Ligand Binding.” *Science* 346 (6216). https://doi.org/10.1126/science.1254346.

Perutz, By M F, M G Rossmann, A N N F Cullis, Hilary Muirhead, Georg Will, and A C T North. 1960. “Structure of Haemoglobin.” *Nature* 1: 416–22.

Peterson, Lenna X., Yoichiro Togawa, Juan Esquivel-Rodriguez, Genki Terashi, Charles Christoffer, Amitava Roy, Woong Hee Shin, and Daisuke Kihara. 2018. “Modeling the Assembly Order of Multimeric Heteroprotein Complexes.” *PLoS Computational Biology* 14 (1): 1–30. https://doi.org/10.1371/journal.pcbi.1005937.

Petrov, A. S., C. R. Bernier, C. Hsiao, A. M. Norris, N. A. Kovacs, C. C. Waterbury, V. G. Stepanov, et al. 2014. “Evolution of the Ribosome at Atomic Resolution.” *Proceedings of the National Academy of Sciences* 111 (28): 10251–56. https://doi.org/10.1073/pnas.1407205111.

Pillai, A.S., S.A. Chandler, Y. Liu, A.V. Signore, C.R. Cortez-Romero, J.L.P. Benesch, A. Laganowsky, J.F. Storz, G.K.A. Hochberg, and J.W. Thornton. 2020a. “Author Correction: Origin of Complexity in Haemoglobin Evolution (Nature, (2020), 581, 7809, (480-485), 10.1038/S41586-020-2292-Y).” *Nature* 583 (7816). https://doi.org/10.1038/s41586-020-2472-9.

Pincus, David, Jai P. Pandey, Zoë A. Feder, Pau Creixell, Orna Resnekov, and Kimberly A. Reynolds. 2018. “Engineering Allosteric Regulation in Protein Kinases.” *Science Signaling* 11 (555): 1–12. https://doi.org/10.1126/scisignal.aar3250.

Polticelli, F., A. Bocedi, G. Minervini, and P. Ascenzi. 2008. “Human Haptoglobin Structure and Function - A Molecular Modelling Study.” *FEBS Journal* 275 (22): 5648–56. https://doi.org/10.1111/j.1742-4658.2008.06690.x.

Porter, Lauren L., and Loren L. Looger. 2018. “Extant Fold-Switching Proteins Are Widespread.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (23): 5968–73. https://doi.org/10.1073/pnas.1800168115.

Qian, Wenfeng, Xionglei He, Edwin Chan, Huailiang Xu, and Jianzhi Zhang. 2011. “Measuring the Evolutionary Rate of Protein - Protein Interaction.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (21): 8725–30. https://doi.org/10.1073/pnas.1104695108.

Qiu, Yang, David H Maillett, James Knapp, John S Olson, Austen F Riggs, Hendrickson Honzatko, W A Proc Natl, and Acad Sci. 2000. “Lamprey Hemoglobin” 275 (18): 13517–28.

Ramesh, Pandian, S. S. Sundaresan, Pon Sathya Moorthy, M. Balasubramanian, and M. N. Ponnuswamy. 2013. “Structural Studies of Haemoglobin from Pisces Species Shortfin Mako Shark (Isurus Oxyrinchus) at 1.9 Å Resolution.” *Journal of Synchrotron Radiation* 20 (6): 843–47. https://doi.org/10.1107/S0909049513021572.

Ratcliff, William C., R. Ford Denison, Mark Borrello, and Michael Travisano. 2012. “Experimental Evolution of Multicellularity.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (5): 1595–1600. https://doi.org/10.1073/pnas.1115323109.

Reynolds, Kimberly A., Richard N. McLaughlin, and Rama Ranganathan. 2011. “Hot Spots for Allosteric Regulation on Protein Surfaces.” *Cell* 147 (7): 1564–75. https://doi.org/10.1016/j.cell.2011.10.049.

Richard, V., G. G. Dodson, and Y. Mauguen. 1993. “Human Deoxyhaemoglobin-2,3-Diphosphoglycerate Complex Low-Salt Structure at 2·5 Å Resolution.” *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1993.1505.

Rivalta, I., M. M. Sultan, N.-S. Lee, G. A. Manley, J. P. Loria, and V. S. Batista. 2012. “Allosteric Pathways in Imidazole Glycerol Phosphate Synthase.” *Proceedings of the National Academy of Sciences* 109 (22): E1428–36. https://doi.org/10.1073/pnas.1120536109.

Rohl, R., and K. H. Nierhaus. 1982. “Assembly Map of the Large Subunit (50S) of Escherichia Coli Ribosomes.” *Proceedings of the National Academy of Sciences of the United States of America* 79 (3 I): 729–33. https://doi.org/10.1073/pnas.79.3.729.

Royer, William E., Kristen Strand, Marin Van Heel, and Wayne A. Hendrickson. 2000. “Structural Hierarchy in Erythrocruorin, the Giant Respiratory Assemblage of Annelids.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (13): 7107–11. https://doi.org/10.1073/pnas.97.13.7107.

Royer, William E., Hao Zhu, Thomas A. Gorr, Jason F. Flores, and James E. Knapp. 2005a. “Allosteric Hemoglobin Assembly: Diversity and Similarity.” *Journal of Biological Chemistry* 280 (30): 27477–80. https://doi.org/10.1074/jbc.R500006200.

———. 2005b. “Allosteric Hemoglobin Assembly: Diversity and Similarity.” *Journal of Biological Chemistry* 280 (30): 27477–80. https://doi.org/10.1074/jbc.R500006200.

Ruiz-Orera, Jorge, Jessica Hernandez-Rodriguez, Cristina Chiva, Eduard Sabidó, Ivanela Kondova, Ronald Bontrop, Tomàs Marqués-Bonet, and M. Mar Albà. 2015. “Origins of De Novo Genes in Human and Chimpanzee.” *PLoS Genetics* 11 (12): 1–24. https://doi.org/10.1371/journal.pgen.1005721.

Salisbury, Frank B. 1969. “Natural Selection and the Complexity of the Gene.” *Nature* 224 (5217): 342–43. https://doi.org/10.1038/224342a0.

Salverda, Merijn L M, and Miriam Barlow. 2010. “Natural Evolution of TEM-1 b-Lactamase: Experimental Reconstruction and Clinical Relevance.” *FEMS Microbiology Reviews,* 34 (6): 1015–1036. https://doi.org/10.1111/j.1574-6976.2010.00222.x.

Santamaría, Belén, Antonio M. Estévez, Oscar H. Martínez-Costa, and Juan J. Aragón. 2002. “Creation of an Allosteric Phosphofructokinase Starting with a Nonallosteric Enzyme: The Case of Dictyostelium Discoideum Phosphofructokinase.” *Journal of Biological Chemistry* 277 (2): 1210–16. https://doi.org/10.1074/jbc.M109480200.

Sato, Akira, Ying Gao, Teizo Kitagawa, and Yasuhisa Mizutani. 2007. “Primary Protein Response after Ligand Photodissociation in Carbonmonoxy Myoglobin.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (23): 9627–32. https://doi.org/10.1073/pnas.0611560104.

Schwarze, Kim, Kevin L. Campbell, Thomas Hankeln, Jay F. Storz, Federico G. Hoffmann, and Thorsten Burmester. 2014. “The Globin Gene Repertoire of Lampreys: Convergent Evolution of Hemoglobin and Myoglobin in Jawed and Jawless Vertebrates.” *Molecular Biology and Evolution* 31 (10): 2708–21. https://doi.org/10.1093/molbev/msu216.

Schwarze, Kim, Abhilasha Singh, and Thorsten Burmester. 2015. “The Full Globin Repertoire of Turtles Provides Insights into Vertebrate Globin Evolution and Functions.” *Genome Biology and Evolution* 7 (7): 1896–1913. https://doi.org/10.1093/gbe/evv114.

Selberg, Avery G A, Eric A Gaucher, and David A Liberles. 2021. “Ancestral Sequence Reconstruction : From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond.” *Journal of Molecular Evolution*, no. 0123456789. https://doi.org/10.1007/s00239-021-09993-1.

Serb, Jeanne M., and Douglas J. Eernisse. 2008. “Charting Evolution’s Trajectory: Using Molluscan Eye Diversity to Understand Parallel and Convergent Evolution.” *Evolution: Education and Outreach* 1 (4): 439–47. https://doi.org/10.1007/s12052-008-0084-1.

Shimizu, Katsuiko, and Enrico Bucci. 1974. “Allosteric Effectors of Hemoglobin. Interaction of Human Adult and Fetal Hemoglobins with Poly(Carboxylic Acids).” *Biochemistry* 13 (4): 809–14. https://doi.org/10.1021/bi00701a026.

Siddiq, Mohammad A., Georg KA Hochberg, and Joseph W. Thornton. 2017. “Evolution of Protein Specificity: Insights from Ancestral Protein Reconstruction.” *Current Opinion in Structural Biology*. https://doi.org/10.1016/j.sbi.2017.07.003.

Skolnick, Jeffrey, and Mu Gao. 2013. “Interplay of Physics and Evolution in the Likely Origin of Protein Biochemical Function.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (23): 9344–49. https://doi.org/10.1073/pnas.1300011110.

Smith, John Maynard. 1970. “Natural Selection and the Concept of a Protein Space.” *Nature*.

Starr, Tyler N., Lora K. Picton, and Joseph W. Thornton. 2017. “Alternative Evolutionary Histories in the Sequence Space of an Ancient Protein.” *Nature* 549 (7672): 409–13. https://doi.org/10.1038/nature23902.

Stoltzfus, Arlin. 2012. “Constructive Neutral Evolution: Exploring Evolutionary Theory’s Curious Disconnect.” *Biology Direct* 7: 1–13. https://doi.org/10.1186/1745-6150-7-35.

Storz, Jay F. 2018. *Hemoglobin: Insights into Protein Structure, Function, and Evolution.* Oxford University Press.

Storz, Jay F, Juan C Opazo, and Federico G Hoffmann. 2013. “Molecular Phylogenetics and Evolution Gene Duplication , Genome Duplication , and the Functional Diversification of Vertebrate Globins.” *Molecular Phylogenetics and Evolution* 66 (2): 469–78. https://doi.org/10.1016/j.ympev.2012.07.013.

Süel, Gürol M., Steve W. Lockless, Mark A. Wall, and Rama Ranganathan. 2003. “Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins.” *Nature Structural Biology* 10 (1): 59–69. https://doi.org/10.1038/nsb881.

Tan, Wei Hung, Shu Chun Cheng, Yu Tung Liu, Cheng Guo Wu, Min Han Lin, Chiao Che Chen, Chao Hsiung Lin, and Chi Yuan Chou. 2016. “Structure of a Highly Active Cephalopod S-Crystallin Mutant: New Molecular Evidence for Evolution from an Active Enzyme into Lens-Refractive Protein.” *Scientific Reports* 6 (July): 1–9. https://doi.org/10.1038/srep31176.

Tanakai, Yoshikazu, Kouhei Tsumoto, Yoshiaki Yasutake, Mitsuo Umetsu, Min Yao, Harumi Fukada, Isao Tanaka, and Izumi Kumagai. 2004. “How Oligomerization Contributes to the Thermostability of an Archaeon Protein: Protein L-Isoaspartyl-O-Methyltransferase from Sulfolobus Tokodaii.” *Journal of Biological Chemistry* 279 (31): 32957–67. https://doi.org/10.1074/jbc.M404405200.

Tetsuya Yomo, Seiji Saito and Masaki Sasai. 1999. “Gradual Development of Protein-like Global Structures through Functional Selection.” 5: 743–46.

Theißen, Günter. 2009. “Saltational Evolution: Hopeful Monsters Are Here to Stay.” *Theory in Biosciences* 128 (1): 43–51. https://doi.org/10.1007/s12064-009-0058-z.

Thornton, Janet M., Christine A. Orengo, Annabel E. Todd, and Frances M.G. Pearl. 1999. “Protein Folds, Functions and Evolution.” *Journal of Molecular Biology* 293 (2): 333–42. https://doi.org/10.1006/jmbi.1999.3054.

Trevor D. Lamb, Shain P. Collin, Edward N. Pugh Jr. 2007. “Evolution of the Vertebrate Eye: Epsins, Photoreceptors, Retina and Eye Cup.” *Nature Reviews Neuroscience* 8 (12): 960–76. https://doi.org/10.1038/jid.2014.371.

Tyuma, Itiro, Ruth E. Benesch, and Reinhold Benesch. 1966. “The Preparation and Properties of the Isolated α and β Subunits of Hemoglobin A.” *Biochemistry* 5 (9): 2957–62. https://doi.org/10.1021/bi00873a027.

Wadsworth, Caroline, and Mike Buckley. 2014. “Proteome Degradation in Fossils: Investigating

the Longevity of Protein Survival in Ancient Bone." *Rapid Communications in Mass Spectrometry* 28 (6): 605–15. https://doi.org/10.1002/rcm.6821.

Wagner, Andreas. n.d. "Neutralism and Selectionism: A Network-Based Reconciliation."

Walden, Helen, Graeme S. Bell, Rupert J.M. Russell, Bettina Siebers, Reinhard Hensel, and Garry L. Taylor. 2001. "Tiny TIM: A Small, Tetrameric, Hyperthermostable Triosephosphate Isomerase." *Journal of Molecular Biology* 306 (4): 745–57. https://doi.org/10.1006/jmbi.2000.4433.

Weber, RE., Jensen, FB. 1988. "Functional Adaptations In Hemoglobins From Ectothermic Vertebrates." *Annual Review of Physiology* 50 (1): 161–79. https://doi.org/10.1146/annurev.physiol.50.1.161.

Weber, Christoph, Andreas Hartig, Roland K. Hartmann, and Walter Rossmanith. 2014. "Playing RNase P Evolution: Swapping the RNA Catalyst for a Protein Reveals Functional Uniformity of Highly Divergent Enzyme Forms." *PLoS Genetics* 10 (8). https://doi.org/10.1371/journal.pgen.1004506.

Williamson, John H., and Michael M. Bentley. 1983. "Comparative Properties of Three Forms of Glucose-6-Phosphate Dehydrogenase in Drosophila Melanogaster." *Biochemical Genetics* 21 (11–12): 1153–66. https://doi.org/10.1007/BF00488467.

Wistow, Graeme. 2012. "The Human Crystallin Gene Families." *Human Genomics* 6 (1): 1–10. https://doi.org/10.1186/1479-7364-6-26.

Wunderlich, Michael, and Franz X. Schmid. 2006. "In Vitro Evolution of a Hyperstable Gβ1 Variant." *Journal of Molecular Biology* 363 (2): 545–57. https://doi.org/10.1016/j.jmb.2006.08.034.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

Zuckerkandl, E. 1965. "The Evolution of Hemoglobin." *Scientific American* 212 (5): 110–18.