THE UNIVERSITY OF CHICAGO

MECHANISMS FOR THE EVOLUTION OF PROTEIN MULTIMERIZATION AND ALLOSTERY

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

GRADUATE PROGRAM IN CELL AND MOLECULAR BIOLOGY

BY

CARLOS R. CORTEZ

CHICAGO, ILLINOIS

JUNE 2025

Copyright 2025 by Carlos R. Cortez

Dedication

To my family—my parents, brothers, my wife, and our beloved pets—whose support and unconditional love were the safety I needed to become the person I am today.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This work is the culmination of almost 7 years of work, which was only possible through the continuous support of a large collection of individuals I met during my time in graduate school. I have changed as a scientist and person during my time in Chicago, in large part of the love I have received from people. This acknowledgement is a thank you to all that were part of my journey.

I joined Joe Thornton's lab at the end of my first year of graduate school. The scientists I met were amazing, some of the most dedicated and kind people. I learned how to do great science and how to be a wonderful person at the same time. I also want to thank my advisor, Joe Thornton, who I have worked with closely throughout my entire PhD. His perspective on science is a huge influence on my approach and how I would like to lead. I believe that my success is directly a result of his patience and willingness to help.

I would like to thank my Chicago friends, those that I met at the start and during my journey in graduate school. I have lived a beautiful life because of the memories that we made together, I cherish them fondly. As professionals, I admire your accomplishments and talent, it feels so good to be surrounded by friends who push me to be a better scientist. I cannot wait to see how your lives develop, and I will carry you all in my heart forever.

To my family in Los Angeles, who has supported me from afar throughout this journey. Thank you for being a reminder of my roots. No matter how far I have gotten, and how much further I will go in this life, I appreciate that everyone will always call and ask about what my life looks like. These calls are a reminder of the quieter moments in my life, something that I have learned to value more as I grow older.

To my wife, Suzy, you have been the most important person during this time in my life. Our world is a result of the dedication, commitment, and love that we share between each other. You

have taught me, through words and actions, to keep going even when our lives were at their most difficult. We now share a beautiful home with multiple pets, which gives me so much excitement every time I step foot inside. You have taught me scientific rigor, how to think through problems, and how to write more clearly. You are a constant inspiration in my life, and I cannot wait to see where our journey goes from here.

Chapter 1

Introduction

1.1 Overview

Proteins are fundamental for organismal viability. Understanding how their structures and functions are determined by their amino acid sequences is a central goal of biochemistry (1, 2). Unifying principles may emerge if we can define how residue changes lead to the evolution of novel features and structures across diverse experimental systems. Evolution serves as nature's longest-running experiment in protein design, generating a vast diversity of forms and functions through residue-level changes over time. By leveraging the evolutionary history of protein and using biochemical techniques, we can deepen our understanding of how amino acid encode protein structure and function.

Studying evolutionary history extends the sequence-structure-function paradigm by incorporating evolutionary history to understand how ancient residue changes led to our modern proteins. By examining the ancestor-descendent relationships of proteins, we can contextualize how residues led to the emergence of new features and how the ancestral structure contributed to their evolution (3). Despite a providing a powerful framework by which to understand proteins, our studies on ancient proteins remain limited and how particular features emerged or were maintained during history are not known.

This thesis spans multiple topics which can be broken down into two distinct sections: the following two chapters describe the mechanisms for the acquisition of three ubiquitous protein feature. Chapter 2 addresses multimerization and paralog specificity. Multimerization is the ability for multiple subunits of the same protein to associate into intricate quaternary complexes through binding at hydrophobic or electrostatically charged patches located on the surface.

Paralog specificity is the ability for multiple subunits of genes related via duplication to preferentially associate with each other rather than themselves. Chapter 3 addresses allostery, intra-protein functional regulation, where binding at a distal site within the tertiary structure regulates the functional ability at another site through coupled stabilization of distinct structural states within the protein. Chapter 4 addresses how allosteric regulation is maintained as proteins undergo sequence drift during history, which allows us to understand how residues involved in allosteric function are distributed across a protein.

By describing the underlying mechanisms in each case and identifying which features are ancestral to the new elaborations, we uncover biophysical and structural principles intrinsic to the proteins we study. These principles highlight how a small number of residues can drive evolutionary changes, if particular structural features are already present at the time of the residue changes.

This introduction will explore how evolution intersects with biochemistry and protein design. In each chapter, the introduction will address the central question motivating the research within the context of evolutionary biology, and the discussion will highlight the broader implications for the fields.

1.2 Evolutionary biology to determine when novel protein features emerged

Understanding how the proteins we observe today evolved to their current structures and functions is a central question in biochemistry (2). One way to address this question is by studying the evolutionary history of proteins and determining how those features emerged. Evolutionary biology accounts for the characteristics of proteins in terms of their histories and allows us to understand how sequence changes occur across time (4). By tracing the evolutionary paths of protein families, we can pinpoint key residues that have been conserved or

modified over time, shedding light on how structural and functional innovations arose (5). Many tools can help us trace the history of proteins. For the purposes of this thesis, there are two that are critical for understanding how new features are built: molecular phylogenetics and ancestral sequence reconstruction (6, 7).

Molecular phylogenetics examines the evolutionary relationships of molecular sequences (DNA, RNA, and proteins) by grouping them based on shared homology, where genes that share a recent common ancestor are likely to have many conserved residues (6, 8). These relationships are represented by nodes in a phylogenetic tree, where genes that share a common ancestor are connected by these nodes, and represent hypothetical ancestral genes. Importantly, the positioning of these nodes can provide a sense of time, deeper nodes indicate earlier divergences between proteins (9, 10). By using molecular phylogenetics to understand the history of proteins, we can estimate when specific protein features may have emerged during evolution.

To understand how novel features emerged, we must be able to explicitly test sequences during evolutionary history and determine which residue changes caused the emergence of novel functions. Here, ancestral sequence reconstruction (ASR) is critical because it statistically infers sequences at each node of a phylogenetic tree (11). These sequences can be synthesized and then tested explicitly in lab. ASR is a probabilistic method that leverages three inputs: multiple sequence protein alignments, the maximum-likelihood phylogenetic tree, and the best fit evolution model (11). Each of these inputs are carefully chosen through traditional evolutionary biology or rigorous statistical approaches to reduce uncertainties in the reconstruction process or discordance with known evolutionary relationships.

Together, molecular phylogenetics and ancestral sequence reconstruction (ASR) make it possible to trace the evolutionary history of proteins. By experimentally characterizing ancestral proteins,

we can pinpoint the sequence changes responsible for the emergence of novel features. The logic in these approaches will be explored in the next section.

1.3 Evolutionary biochemistry to determine how novel protein features emerged

Evolutionary biochemistry is a burgeoning sub-field of biochemistry that adds conceptually to the classic sequence-structure-function paradigm by incorporating evolutionary history as a key variable (4).

Since the 1960's, scientists have been interested in reconstructing and experimentally characterizing the history of proteins (7). With the development of maximum likelihood phylogenetic, it has now become possible to reconstruct ancestral protein sequences with strong statistical rigor. Coupled with the low cost of DNA synthesis, this makes it feasible to experimentally test the structures and functions of these reconstructed proteins (9, 11).

Evolutionary biochemistry is the experimental study of the history of proteins, using phylogenetics and ASR to define the ancestor-descendent relationships within a gene family and biochemistry techniques to test these relationships in the lab (3, 4). By focusing on discrete time intervals defined by these ancestor-descendent relationships, we can experimentally pinpoint when new features arose, which narrows down the candidate substitutions that drove changes in structure and function to those specific historical intervals. From here, the relationships between sequence, structure, and function can be tested explicitly.

Identifying when functional features arose during evolution allows us to infer causal relationships between ancestral and descendant proteins (3, 4, 12). Substitutions that occur along the branch between an ancestor and its descendant, within the historical interval in which a new feature emerges, drove that innovation. But branches typically contain many substitutions, so determining which caused the change in function requires explicit testing of smaller sets of

substitutions. This can be done by establishing which changes were necessary for the change and which were sufficient.

To test necessity, candidate residues are reverted in their ancestral state in the descendant protein and tested for the feature, if the feature reverts to its ancestral state, then the candidate residues are necessary. To test sufficiency, candidate residues are substituted to their derived state in the ancestor, if the feature is conferred, then the candidate residues are sufficient. Together, we can explicitly determine the number of residues involved in novel features and how each of them contributed to the novel feature.

The evolutionary biochemistry framework has proven powerful for determining the evolutionary history of proteins, discovering how novel features arose in history, and furthering our understanding of how sequence determines structure and function. One way it has helped is by making epistatic interactions, nonadditive changes in residue effects based on introduction of other residues, part of biochemical mechanisms for changes in function (13-18). A well-known example is the shift in ligand specificity of the ancestral glucocorticoid receptor (19). Two key substitutions were responsible for the new specificity, but their effects depended on five "permissive" substitution. These permissive changes did not directly affect ligand binding but stabilized the protein in a way that tolerated the introduction of the function-switching mutations (19). For this reason, in present-day proteins epistasis also obscures the mechanisms underlying the emergence of new protein functions (3). In the present day, if we introduce those two function-switching mutations into the mineralocorticoid receptor, which do not contain the 5 permissive residues, there would be no change ligand specificity.

Evolutionary biochemistry has also shown how changes or the emergence of features within proteins occurred. Much of evolutionary biochemistry has described shifts in molecular

specificity—whether for small ligands or DNA—revealing that such functional transitions often require surprisingly few amino acid changes (19-21). These findings suggest that the genetic basis of molecular specificity can be relatively simple. However, epistatic interactions have consistently proven essential to enabling these functional changes. In some cases, they amplify the effects of key functional mutations; in others, they relieve steric clashes or otherwise stabilize new conformational states (14, 19, 22). Moreover, changes in molecular complex stoichiometry—such as the formation of higher-order oligomers—do not always arise due to direct functional advantage. Instead, epistasis may constrain the evolutionary paths available, making it difficult to revert or remove interfaces once they have formed (23).

Biochemistry aims to understand how sequence determines structure and function by testing the effects of residue changes in present-day proteins (1). However, it is important to recognize that all nature proteins are a product of evolutionary history, and these histories represent an underutilized resource for uncovering the principles that link sequence to structure and function. Today, the UniProt database contains over 250 million natural protein sequence (24). Although many of these are most likely homologs of the same proteins, they still encompass a vast reservoir of functional and structural innovations – many of which remain unexplored.

1.4 Mechanisms for novel protein features and design

A major component of this thesis concerns the mechanisms by which new protein assemblies, and functions evolve. The ability to engineer novel structures and activities is a central goal of protein design (25). While there has been notable success in designing multimeric and allosteric proteins, these efforts face limitations. Incorporating principles uncovered through evolutionary biochemistry, the role of historical substitutions and epistatic interactions, offers a path to overcome some of the constraints and improve the rational design of complex protein functions.

Multimerization has seen the most success: early rational design studies showed that introducing small number of mutations could strongly stabilize multimeric assemblies (26-29). De novo design takes from these studies and builds multimeric assemblies through symmetric interfaces (30). The advent of computational design and denovo assemblies further demonstrated that rules of symmetry and interface complementarity could be used to build novel homomeric proteins from scratch (30-32). Recent machine learning approaches have pushed this further by generating entirely novel multimeric proteins with atomic-level accuracy, though the success rates remain variable depending on interface geometry and stoichiometry (31, 33).

Specific assembly has been more difficult to design. Key work has revealed that relatively small sets of residues can drive the gain of specific protein-protein interactions (34-38). Yet predictive models for specificity remain limited, largely because there is a lack of large-scale datasets measuring specificity from non-specific starting points, which is essential for predicting specificity in mixed systems. De novo design has had some success, by building hydrogen bonding interactions between proteins, but still requires extensive testing (39, 40).

Allosteric regulation design still requires more research, as the development of engineered and de novo design are few. There have been successful engineering efforts: primarily by building ligand binding pockets or large insertions of regulatory domains into non-allosteric proteins (41-47). De novo design has produced a single study that has been able to create an allosteric protein, which did not focus on the residue changes themselves but instead whole concerted movements within the structures (48). Furthermore, no current machine learning models can predict or design multiple protein conformations, largely because datasets that connect sequence variation to conformational dynamics and function are few (49).

Together, these efforts highlight a limitation across protein design: we must know what specific residues are required to generate new assemblies and allostery on as many proteins as possible to be able to build them efficiently. As machine learning and de novo design continue to generate new proteins, their success will depend on datasets that capture the functional consequences of defined mutations in diverse backgrounds, especially those that can differentiate functional from non-functional variants in different structural contexts (50). Evolutionary biology can offer a complementary strategy to improve these efforts: by analyzing how many natural proteins evolved and identifying historical substitutions that shaped function, we can map the causal architecture to identify residues with mechanistic importance in diverse backgrounds. These principles can be incorporated into new experimental strategies that can further improve engineering or designed protein efforts.

For this reason, chapters 2 and 3 of my thesis focus on identifying the number and effect of residues that confer multimerization, heteromeric specificity, and allosteric regulation in the model system hemoglobin. Chapter 4 shifts to a second model system, the allosteric repressor TetR, to examine how drift shapes its underlying genetic architecture, with particular attention to the epistatic interactions that emerge and influence evolutionary trajectories. Together, these findings offer directly address outstanding questions in evolutionary biology and can offer insight into design principles of multimeric and allosteric proteins by providing an empirical foundation to improve rational and machine learning based design efforts.

1.5 Hemoglobin as a model system for the origination of multimerization, specificity, and allostery

Hemoglobin is one of the most influential model systems in the field of biochemistry. Multimerization, heteromeric specificity, and allosteric regulation are well understood in

hemoglobin (51, 52). The function of the protein is to coordinate reversable oxygen binding within each subunit using an iron atom coordinated within a porphyrin ring (53). Hemoglobin is a heterotetramer, containing two α and two β subunits, allowing for binding of up to four oxygen molecules per tetramer. Assembly occurs through two interfaces, and each is specific for the other subunit (ie. α always binds to β and vice versa) (54-57). The heterotetrameric structure is critical for the regulation of oxygen binding affinity within the heme which allows for efficient offloading at peripheral tissues (52). Oxygen binding affinity in hemoglobin is regulated by two forms of allostery. The first, cooperativity, is an intramolecular interaction where oxygen binding to one subunit increases the affinity of the remaining subunits and occurs through changes in the tetrameric interface (58-61). The second, heterotropic allosteric regulation, occurs when a ligand binds to the central cavity – a pocket that appears between the two beta subunits that occurs only in the tetramer – and decreases oxygen affinity (62, 63).

Residue variation in hemoglobin has been studied extensively to understand which residues drive multimerization, heteromeric specificity, and allosteric regulation (64). The residues involved in multimerization and specificity are well known, as they are mostly conserved across jawed vertebrates and found at the interfaces (65). The strength of allosteric effect to effector binding varies across extant organisms, and identity of effector molecules which bind to hemoglobin or their specificity have changed amongst clades (51, 64). This has led to uncertainty over which residues confer allosteric function and when they arose. Not all interface residues contribute equally to the energetics of assembly, raising the possibility that a small number of large-effect mutations may have played a key role in establishing multimerization and specificity.

The evolutionary history of hemoglobin has been of particular interest because it represents a large innovation in the regulatory capacity of oxygen at the dawn of jawed vertebrates (51). It is

thought that the appearance of hemoglobin may have played a role in the origin of jawed vertebrates and subsequent radiations are driven by changes in hemoglobin regulatory capacities (66). Phylogenetic analysis supports these claims as hemoglobin appears at the same time as jawed vertebrates and further sequence changes have been shown to change allosteric regulation and oxygen affinity in animals occupying various environments (51, 64, 66, 67).

If hemoglobin has been key in many physiological innovations in jawed vertebrates, then understanding how hemoglobin's structure and allosteric regulation initially emerged is critical. It is known that hemoglobin descends from other globin genes that are monomeric and non-allosteric: the outgroups of are myoglobin, globin E, and globin Y (67, 68). Using ancestral sequence reconstruction and biochemical experiments, a previous study showed that extant hemoglobin appeared in two steps: at the time of duplication from all myoglobin and hemoglobin genes there was a duplication and, on the branch leading to the base of all hemoglobin genes, the ancestral monomer became a dimer. Another duplication occurred that gave rise to the alpha and beta genes, and on the following branches the dimer because an allosteric heterotetramer (69). Together, we know when the structure and allosteric function of hemoglobin emerged during history.

Parts of my thesis continue this work. In chapter 2, I carefully dissect the residues that are implicated in the dimer to tetramer transition, how specificity is distributed between the two interfaces in the heterotetramer, and the gain of specificity between the alpha and beta subunits after gene duplication. In chapter 3, I discover the set of residue changes that confer allosteric regulation and then dissect all residues involved to uncover the full historical allosteric landscape. In all cases, we find that the number of residues required for these large innovations

in Hb were few, suggesting that hemoglobin as we know today could have appeared quickly and that other proteins may have done the same.

1.6 TetR as a model system for allostery

Tetracycline repressor proteins (TetR) are an allosteric transcription factors known to confer antibiotic resistance to tetracycline (70). Understanding allosteric regulation within this protein has been of considerable interest, because it may offer ways to prevent bacteria from gaining antibiotic resistance. TetR is an obligate repressor that binds to a 15 base pair palindromic sequence of the TetA promoter via a DNA binding domain, preventing the transcription of the gene (71). The allosteric transition occurs upon tetracycline binding in the ligand binding domain of the protein and causes de-repression of TetR. Upon releasing from the DNA, transcription of TetA occurs and pumps out tetracycline thereby conferring antibiotic resistance.

The residues involved in the TetR allosteric response are not well understood. Crystal structures of ligand-bound and unbound states of TetR do not undergo large scale conformational changes, suggesting that large residue movements are unlikely to underlie the allosteric mechanisms (71). Alternative models have proposed mechanisms such as alterations in dynamics or changes in the populations within conformational ensembles (72-74). However, these models do not clearly identify which residues are essential for allosteric regulation. Recent studies have shown that residue mutations that inactivate allostery are distributed across the entire protein, suggesting that residues contributing to allostery may be spread throughout the structure.

In chapter 4, I used phylogenetic analysis, ASR, and large-scale mutational scans to investigate how the genetic architecture of TetR allosteric response drifts through sequence space. By introducing all single substitutions across distinct historical intervals, we examine how tolerated mutations changed during history. We discover that epistasis, the change in mutational effects

when other mutations are introduced, play a larger role than previous appreciated. This finding helps explain why pinpointing the residues responsible for allosteric responses has remained a persistent challenge in the field.

Chapter 2

Symmetry facilitated the evolution of heterospecificity and high-order stoichiometry in vertebrate hemoglobin

2.1. Abstract

Many proteins form paralogous multimers – molecular complexes in which evolutionarily related proteins are arranged into specific quaternary structures. Little is known about the mechanisms by which they acquired their stoichiometry (the number of total subunits in the complex) and heterospecificity (the preference of subunits for their paralogs rather than other copies of the same protein). Here we use ancestral protein reconstruction and biochemical experiments to study historical increases in stoichiometry and specificity during the evolution of vertebrate hemoglobin (Hb), a a_2b_2 heterotetramer that evolved from a homodimeric ancestor after a gene duplication. We show that the mechanisms for this evolutionary transition were simple. One hydrophobic substitution in subunit b after the gene duplication was sufficient to cause the ancestral dimer to homotetramerize with high affinity across a new interface. During this same interval, a single-residue deletion in subunit a at the older interface conferred specificity for the heterotetrameric form and the *trans*-orientation of subunits within it. These sudden transitions in stoichiometry and specificity were possible because the interfaces in Hb are isologous – involving the same surface patch on interacting subunits, rotated 180° relative to each other. This architecture amplifies the impacts of individual mutations on stoichiometry and specificity, especially in higher-order complexes, and allows single substitutions to differentially affect heteromeric vs homomeric interactions. Our findings suggest that elaborate and specific symmetrical molecular complexes may often evolve via simple genetic and physical mechanisms.

2.2. Introduction

Protein multimers – associations of multiple protein subunits arranged in specific quaternary architectures – carry out most biochemical functions in living cells (75, 76). The mechanisms by which these complexes evolved their stoichiometry and specificity present some puzzling questions (27, 36, 76-82). Multimers assemble via interfaces that typically contain dozens of sterically and electrostatically complementary residues, and higher-than-dimeric stoichiometries (tetramers, octamers, etc.) use several such interfaces on each subunit (26). This seems to imply that many sequence substitutions would be required for a new multimeric assembly to originate during evolution.

A second complication is that many multimers are composed of paralogs -- proteins related to each other by gene duplication (83). Paralogs are genetically and structurally indistinguishable when generated by duplication, so initially they assemble indiscriminately into homomers and heteromers. Most complexes, however, have evolved specificity for either the homomeric or heteromeric form, with the latter being the most common outcome (83, 84). How specificity evolves is unclear, because mutations that affect multimerization are expected to cause correlated effects on the affinities of homomerization and heteromerization(80, 83, 85). The structural similarity of paralogs seems to imply that substitutions in both paralogs are required to confer any specificity at all. This complication is magnified for higher-order paralogous multimers, in which one might expect that every interface must evolve specificity to mediate assembly into the complex's particular architecture.

A critical factor in the evolution of specificity and high-order stoichiometry may be whether a multimer assembles through symmetrical interfaces. In many complexes, identical or paralogous subunits bind each other using an isologous interface – a form of symmetry in which a surface

patch on one subunit binds to the same patch on its partner but rotated 180 degrees relative to each other (1). Isologous complexes might, in principle, have the potential to evolve changes in stoichiometry and specificity through simpler mechanisms than nonisologous head-to-tail interfaces. A single substitution appears twice across the interface(s) of an isologous homodimer or heterotetramer, four times in a homotetramer, etc. (Fig. 2.1A). Mutations that weakly affect affinity on their own can therefore confer large effects on the assembly of isologous multimers (27, 69, 75, 79, 86, 87). Isology also changes the way that mutations can affect specificity. In a nonisologous interface, specificity requires mutations on both surfaces so that the tails are recognizably different from each other and each head prefers one tail over the other. In an isologous interface, however, a substitution on the surface of just one subunit has the potential to differentially affect the affinity of each kind of complex, because it will appear twice in the interface of a homomer, once in the heteromer, and not at all in the other homomer (Fig. 2.1A). Little is known about the historical evolution of heterospecific complexes or the role of symmetry in this process, especially in high-order complexes. Biochemical and protein engineering studies have addressed the determinants of binding affinity in both homomeric and heteromeric interfaces of extant proteins (29, 35, 37, 38, 88-90). But the genetic and structural mechanisms by which those interactions were acquired long ago are often different from their derived forms in the present (3). Ancestral sequence reconstruction (ASR) can address this limitation by experimentally characterizing the effects of historical sequence changes when introduced into ancestral proteins. ASR has been used to understand the evolution of specificity after duplication in head-to-tail paralogous heteromers (91, 92) and in multimers composed of unrelated proteins, which are by definition asymmetrical (90). But we know of no studies that have addressed how isologous heteromers historically evolved their specificity or how specificity

in high-order complexes was acquired. A recent in silico analysis predicted that it should be possible for specificity in heterodimers to evolve rapidly after gene duplication through small perturbations in binding energy (93), but the underlying mechanisms and historical relevance of this phenomenon are unknown.

Here we use ASR to study the evolution of higher-order stoichiometry and specificity in vertebrate hemoglobin (Hb), the major carrier of oxygen in the blood of jawed vertebrates. Hb is a paralogous $\alpha_2\beta_2$ heterotetramer ((69), Fig. 2.1B), assembly of which is mediated by two distinct and isologous interface patches (IF1 and IF2). Each subunit of the tetramer uses its IF1 to bind IF1 of a paralogous subunit; two of these heterodimers bind to each other using the IF2 on each subunit to make the tetramer ((53), Fig. 2.1B). Hbα and Hbβ descend from a gene duplication deep in the vertebrate lineage (Fig. 1C), and their sequences retain sufficient phylogenetic signal to allow high-confidence reconstruction of ancestral Hb protein sequences. Using ASR, we recently showed experimentally that extant Hb evolved its heterotetrameric architecture in two phases from a monomeric precursor via a homodimeric intermediate (69). In the first phase, prior to the gene duplication that yielded paralogous α and β lineages, a monomeric ancestor evolved the capacity to homodimerize with moderate affinity across IF1. In the second phase – after the gene duplication but before the last common ancestor of all vertebrates – binding across IF2 was acquired, yielding the tetrameric stoichiometry, and specificity for the heteromeric form $\alpha_2\beta_2$ also evolved (Fig. 2.1C).

Here we characterize the genetic and physical mechanisms that mediated the evolutionary transition from homodimer to heterotetramer in this second phase. By experimentally characterizing reconstructed ancestral hemoglobin subunits and the effects of historical sequence changes on them, we address the following questions: 1) How many substitutions were required

to confer tetrameterization across IF2, and what thermodynamic and structural mechanisms mediated their effects? 2) Did the evolution of specificity for the heterotetrameric form require sequence changes at one or both interfaces, in one or both subunits, and what physical mechanisms drove the acquisition of this specificity? 3) How did the symmetry of Hb's two interfaces affect this evolutionary transition to a high-order, heterospecific architecture? 4) Does a mutational propensity favor increased molecular complexity during the evolution of isologous complexes?

2.1 Evolution of tetrameric stoichiometry

We first sought to identify the historical sequence changes that conferred tetramerization after duplication of the ancestral homodimer Ancαβ. We focused on the branch leading from the duplication of Ancαβ to Ancβ (the Hbβ subunit in the last common ancestor of jawed vertebrates), because Ancβ heterotetramerizes with Ancα (the Hbα subunit in the jawed vertebrate ancestor) and, like extant Hbβs, also homotetramerizes with itself. We previously found two amino acid replacements that occurred on the branch which, if introduced together into Ancαβ, are sufficient to confer high-affinity assembly into homotetramers (69). One of these (q40W) is buried in the IF2 interface, whereas the other (t37V) makes contacts across both IF1 and IF2 (Fig. 2.1D. 17, using lower and upper case to denote ancestral and derived amino acids, respectively). W40 is strictly conserved in Hbb subunits throughout the jawed vertebrates, and V37 is conserved in Hbb of most taxa (Fig. A1.1).

Here we isolated the individual contributions of each amino acid changer by introducing them singly into Ancαβ and characterizing their effect on assembly into tetramers using size-exclusion chromatography (SEC) and native mass spectrometry (nMS) (94, 95). We found that q40W alone is sufficient to recapitulate the evolution of Hb's tetrameric stoichiometry. Ancαβ forms

only dimers in SEC at 100 μM of total protein subunits; by contrast, the mutant Ancαβ$_{q40W}$ is tetrameric, with occupancy of the tetramer similar to that observed in the derived Ancα + Ancβ complex and human Hb (Fig. 2.1E). We used nMS across a titration series to measure the affinity with which dimers associate into tetramers and found that the tetramerization affinity of Ancαβ$_{q40W}$ (Kd 10 μM) is stronger than that of Ancα + Ancβ (29 μM) and human Hb (34 μM) (Fig. 2.1F). The conclusion that q40W is sufficient to confer tetramerization is robust to statistical uncertainty about the ancestral sequence: similar experiments using a different reconstruction of Ancαβ that incorporates alternative residues at all ambiguously reconstructed sites yields almost identical results (Fig. A1.2).

The other historical replacement, t37V, is not sufficient to confer tetramerization. Mutant Ancαβ$_{t37V}$ confers no detectable tetramer occupancy by SEC, even at 1 mM (Fig. 2.1G), and it displays no measurable affinity to form tetramers using nMS (Fig. 1H). When combined with q40W, however, t37V does increase affinity of the dimer-tetramer transition by a factor of 6 compared to the effect of q40W alone (Fig. 2.1D; Fig. A1.3).

In principle, a sequence change could also facilitate tetramerization by increasing affinity of the monomer-to-dimer transition; by increasing the effective concentration of dimers, more tetramers would be produced at a given protein concentration, even if affinity of the dimer-tetramer transition were unchanged. Using nMS, we found that t37V improves the monomer-dimer affinity of Ancαβ by >100-fold (Fig. 2,1H; Fig. A1.3). Substitution q40W, in contrast, has no effect on monomer-dimer affinity. These findings are consistent with the structural location of these residues -- t37V contributes to both IF1 and IF2 and q40W to IF2 only -- and they explain why t37V does not confer tetramerization on its own but enhances the impact of q40W.

A likely physical mechanism for the effect of q40W is that tryptophan's bulky hydrophobic side chain nestles into a hydrophobic divot on the IF2 surface of the facing subunit, and is further strengthened by a hydrogen bond to 102D (96). To test this hypothesis, we identified alternative amino acid replacements with similar biochemical properties and measured whether they also could have caused Ancαβ to evolve into a tetramer. Like tryptophan, the bulky hydrophobic residues phenylalanine or tyrosine at this position also confer tetramerization, albeit at affinity slightly weaker than q40W but similar to that of Ancα+Ancβ and human Hb. The greater affinity of tryptophan may be due to its longer side chain, which buries more hydrophobic surface area across the interface; the hydrogen bond with 102D could make a small contribution but is not necessary, because phenylalanine confers tetramerization but provides no hydrogen bond donor. Leucine, in contrast, which has a smaller volume and no hydrogen bonding capacity, confers no measurable tetramerization (Fig. 2.1I). High-affinity homotetramerization could therefore have evolved via any of three different aromatic replacements at site 40.

Taken together, these data indicate that replacing the amino acid at a single residue position was sufficient to confer tetramerization during historical Hb evolution, and several alternative replacements at the same site could also have caused the acquisition of this higher-order stoichiometry.

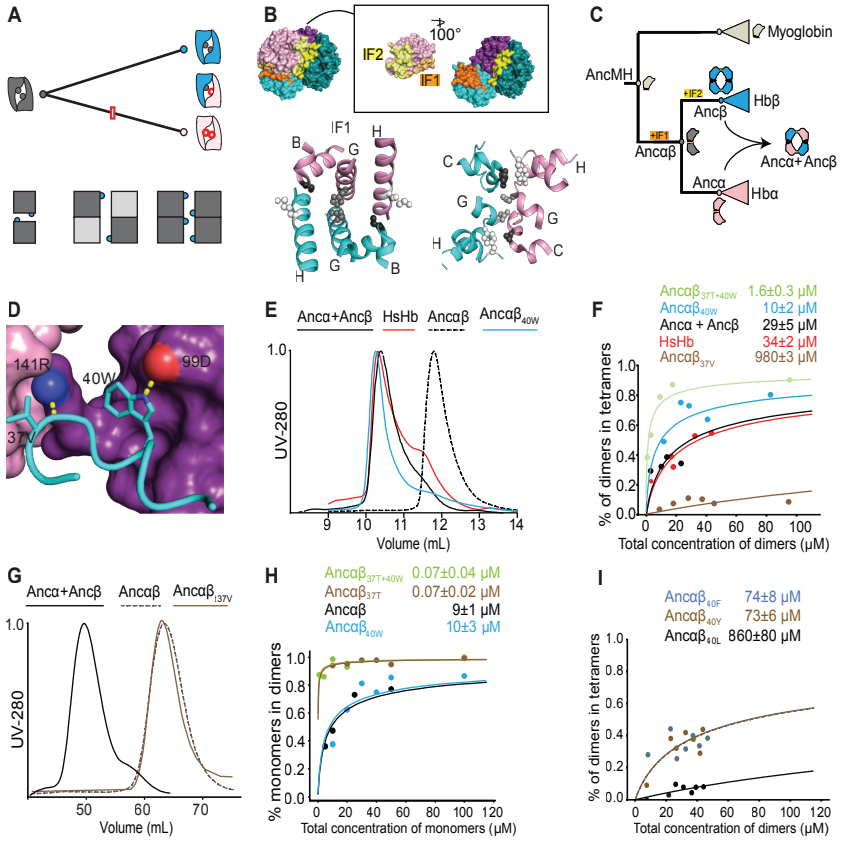

Figure 2.1. A single substitution confers tetramerization on an ancestral dimer. (A) A substitution in one subunit can potentially affect specificity and stoichiometry in an isologous interface. *Top*: After duplication of an isologous homodimer (gray), a substitution that occurs in one paralog (red box) appears twice in the interface of a homodimer (red circles), once in a heterodimer, and not at all in the other homodimer (blue). *Bottom*: One substitution (blue circle) in an isologous interface appears twice in a homodimer (*left*), twice in a heterotetramer (*middle*), and four times in a homotetramer (*right*), multiplying its effects on affinity. Dark and light gray, paralogous subunits. **(B)** *Top*: Interfaces in the human Hb heterotetramer (PDB 4HHB). Pink, Hbα; blue, Hbβ; α_1 and β_1 are in lighter hues than α_2 and β_2. IF1 surfaces (orange) mediate α_1-β_1 and α_2-β_2 interactions; yellow surfaces (IF2) mediate α_1-β_2 and α_2-β_1 interactions. Only interfaces involving α_1 are shown. Inset, α_1 subunit rotated away from the rest of the tetramer to show IF1 and IF2. *Bottom*: Isology of IF1 and IF2. Helices contributing to each interface are shown and labeled. Balls and sticks: on each helix, one residue's side chain is shown to visualize symmetry. **(C)** Evolution of tetrameric stoichiometry on the phylogeny of Hb and related globins. Icons, oligomeric states determined by experimental characterization of reconstructed ancestral proteins (69). Acquisition of interfaces of IF1 and IF2 is shown (69). **(D)** Key residues V37 and W40 that were substituted in Ancb. Cyan cartoon helix, b_1 subunit. Pink and violet surfaces, a subunits that interact with b_1 via IF1 and IF2, respectively. Dotted lines to red or blue spheres, hydrogen bonds to oxygen or nitrogen atoms, respectively (PDB 4HHB). **(E,G)** Effect of historical substitutions on stoichiometry, as measured by size exclusion chromatography. The ancestral dimer Ancαβ and the tetramers Ancα+Ancβ and human hemoglobin (HsHb) are shown for comparison. Protein concentration at 100 mM (E) or 1 mM (G). **(F)** Dimer-to-tetramer affinity of reconstructed ancestral Hb subunits containing historical substitutions q40W and t37V, measured

by native mass spectrometry across a titration series. Points, fraction of dimers that are incorporated into tetramers. Lines, best-fit binding curves. Estimated Kd and 95% confidence interval are shown. **(H)** Effect of historical substitutions on monomer-dimer affinity measured by native MS. **(I)** Effect on dimer-tetramer affinity of nonhistorical hydrophobic mutations in at residue 40, measured by native MS.

2.3. Isology facilitated IF2 evolution

How could a single amino acid replacement cause such a dramatic change in stoichiometry? The Hb tetramer can be viewed as two heterodimers, each of which is mediated by isologous assembly across IF1 (the larger interface); these heterodimers then bind to each other isologously across IF2. We hypothesized that this doubly symmetrical architecture allowed substitution q40W to confer the dimer-tetramer evolutionary transition, because isology causes the derived amino acid to appear four times in the homotetramer and twice in the heterotetramer.

If this hypothesis is correct, then assembly across IF2 by the derived Hb protein should require assembly across IF1 to multiply the intrinsic affinity of IF2 (Fig. 2.1A). We tested this prediction by introducing q40W into Ancαβ but doing so under conditions that prevent assembly across IF1. We first compromised dimerization across IF1 genetically by reverting the IF1 surface to the ancestral states of the monomeric ancestor AncMH; these mutations abolish dimer occupancy, leaving a monomers-only population at 20 μM (Fig. 2.2A). We then introduced q40W into these IF1-ablated mutants and assessed stoichiometry using nMS. As predicted, these proteins do not form any observable dimers or tetramers (detection limit ~1 μM) (Fig. A1.4). Similar results are found when we used the combination of t37V/q40W to confer association across IF2 or the mutation P127R – which introduces unsatisfied positive charges into IF1 -- to compromise IF1. (Fig. 2.2B). The IF2 mutations do not compromise heme binding or solubility, because the mutant proteins are purifiable and heme-bound in nMS.

We also tested whether assembly across IF2 could have been historically acquired before dimerization across IF1 evolved. We introduced t37V/q40W into the ancestral monomer AncMH – which existed before the evolution of dimerization -- and tested whether dimer assembly across IF2 can be conferred in this background. As predicted, only monomers were observed, with no dimers or higher stoichiometries detected (Fig. 2.2C). Acquisition of multimerization across IF2 by q40W and by the pair t37V/q40W therefore depends on the prior evolution of dimerization via IF1.

These observations can be explained by a simple model in which the two symmetrical interfaces contribute independently to the energy of binding. A single iteration of IF2 is too weak to confer measurable binding of two monomers into a dimer; however, IF1 is stronger and mediates assembly of dimers. Each IF1-mediated dimer presents two iterations of the IF2 surface patch, doubling the total energy of IF2-mediated assembly of dimers into tetramers. Because of the exponential relationship between energy and occupancy, a weak IF2 can therefore confer high-affinity binding but only if IF1 is already present. The affinities that we measured are consistent with this simple model. If the energy of dimer-to-tetramer assembly is twice that of monomer-dimer binding using the same interface, then the Kd of IF2-mediated tetramerization should be the square of the Kd of IF2-mediated dimerization (Fig. 2.2D). The Kd of the dimer-tetramer transition by Ancαβ $_{t37V/q40W}$ across IF2 is 1 mM, which predicts that the affinity of IF2-mediated monomer-dimer transition when IF1 is compromised should be ~ 1mM. Consistent with this prediction, we detected no dimer occupancy by Ancαβ $_{t37V/q40W;\ IF1reverted}$ using an assay that can quantify Kd up to 400 μM (see Methods). This simple additive model therefore explains most – and possibly all -- of the difference in affinity conferred when IF2 is doubled in the symmetrical tetramer. The dependence of assembly across IF2 upon the presence of IF1 does not imply any

direct physical interaction between the interfaces or any conformational change in one interface caused by binding at the other. We cannot rule out the possibility that IF1 binding may also allosterically modify IF2 and increase its affinity beyond the additive effect conferred by isologous repetition alone; however, any such effect must be relatively small.

Taken together, these data indicate that the isologous architecture of IF1 and IF2 facilitated the evolution of the Hb tetramer via substitution q40W. Without this doubly symmetrical architecture, IF2 would have been too weak to mediate multimerization. The dependence of q40W's effect on the presence of IF1 also creates contingency and order-dependence in the evolution of the Hb complex. We previously showed that IF1 evolved before the duplication of the dimeric ancestor Ancαβ (69). Our present results show that if that IF1-mediated dimer had never evolved, substitution q40W at IF2 would not have been sufficient to drive the acquisition of the tetrameric stoichiometry, and the ancestral Hb protein would have remained a monomer. If events had occurred in the opposite order – with the affinity-enhancing substitution at IF2 occurring first – this intermediate ancestor would have been a monomer; when the substitutions that confer binding across IF1 did occur, they would have triggered an immediate evolutionary transition from monomer to tetramer.

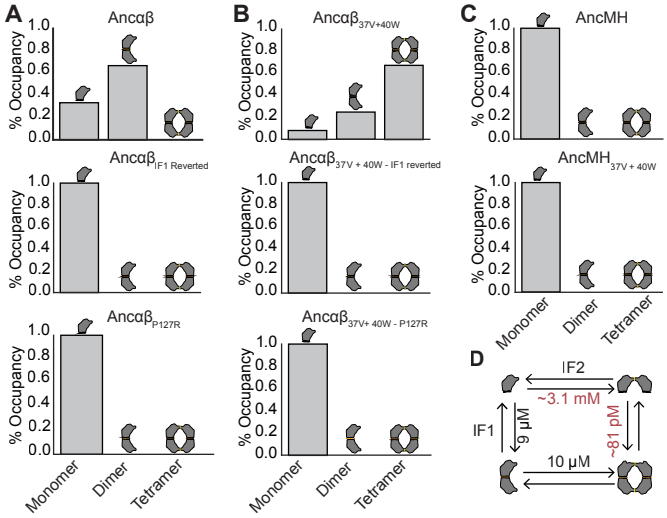

Figure 2.2. Multimerization across IF2 requires IF1. (A) IF1-mediated dimerization can be compromised by mutations. Relative occupancy of each stoichiometry as measured by native MS at at 20 mM total protein is shown for the ancestral dimer Ancαβ (top), $Anc\alpha\beta_{IF1\ reverted}$ (middle, a variant of Ancαβ in which all IF1 residues are reverted to the ancestral state found in AncMH), and Ancαβ-P127R (bottom, in which a mutation known to compromise IF1-mediated dimerization has been introduced). **(B)** Compromising IF1 prevents assembly across IF2. Relative occupancy of $Anc\alpha\beta_{40W + 37V}$ with and without mutations that compromise IF1-mediated dimerization. **(C)** AncMH, which does not dimerize across IF1, cannot multimerize across IF2, even when mutations sufficient to confer IF2-mediated mutimerization in Ancαβ are introduced. **(D)** Observed (black) and expected (red) affinities of Ancαβ +q40W interfaces. Expected Kd of a single iteration of IF2 (top) equals the square root of the measured apparent Kd when two iterations are present (bottom). Expected apparent Kd of two iterations of IF1 (right) equals the square of the measured Kd of a single IF1 (left).

2.4 Heteromeric specificity evolved at a single interface

We next focused on understanding the evolution of Hb's specificity for the heterotetrameric form, which was acquired during the same phylogenetic interval after the duplication of Ancαβ. Our first question was whether specificity for heteromeric interactions was conferred by sequence changes at IF1, IF2, or both. Our previously published experiments suggest that evolutionary changes at IF2 confer no specificity: when all historical substitutions that occurred at the IF2 surface during the post-duplication interval are introduced into Ancαβ and this protein is coexpressed with Ancα, an indiscriminate mixture of homotetramers, $\alpha_1\beta_3$ heterotetramers, and $\alpha_2\beta_2$ heterotetramers is produced (69). We therefore hypothesized that heterospecificity of the Hb tetramer is encoded entirely by IF1, such that Ancα and Ancβ specifically heterodimerize across IF1, and these heterodimers then bind to each other via a nonspecific IF2, yielding $\alpha_2\beta_2$ heterotetramers.

This hypothesis makes two predictions: 1) IF1 mediates specific assembly of α and β subunits into heterodimers, and 2) this specificity is sufficient to account for the heterospecificity of $\alpha_2\beta_2$ heterotetramer. To test the first hypothesis, we characterized the specificity of hetero- vs homodimer assembly by IF1 under two different conditions in which no binding across IF2

occurs. First, we diluted a coexpressed mixture of Ancα and Ancβ to concentrations at which dimers rather than tetramers assemble: at 50 μM, only heterodimers and heterotetramers form; at 5 μM, only heterodimers are observed (Fig. 2.3A). IF2 does not mediate assembly of monomers into dimers in the absence of IF1 (Fig. 2.2A & B), so these heterodimers must be IF1-mediated, indicating that IF1 is heterospecific (Fig. 2.3A). Second, we expressed Ancα and Ancβ separately and mixed them at equal and moderate concentration; because tetramerization requires co-folding, only IF1 dimers form (97), and these are predominantly heterodimers (Fig. 2.3B, Fig A1.5). Finally, we engineered protein Ancβ' – a variant of Ancβ in which all IF2 residues that were substituted between Ancαβ and Ancβ are reverted to the ancestral state, thus abolishing binding across IF2– and found that it also forms predominantly heterodimers when mixed with Ancα (Fig. 2.3C, Fig. A1.5). Together, these data indicate that the derived IF1 is specific, preferentially mediating assembly into heterodimers.

To test the second prediction – that heterospecificity mediated by IF1 is sufficient to drive specific assembly of $\alpha_2\beta_2$ heterotetramers even if IF2 is nonspecific – we measured the affinities of homomerization and heteromerization across IF1 and used these measurements to predict their effects on tetramer specificity in the absence of any specificity at IF2. Using nMS and Ancβ', we found that IF1's heterodimerization affinity (Kd=0.6 μM) is slightly worse than its homodimerization affinity (0.2 μM), but both are far better than the Ancα homodimer (24 μM) (Fig. 3D, Fig A1.6, A1.7, A1.8, & A1.9). We then predicted occupancy of each stoichiometry as the concentration of Hb subunits changes, given these affinities at IF1 and assuming that IF2 has a dimer-to-tetramer affinity of 30 μM, as measured in Ancα + Ancβ, with no preference for homomeric or heteromeric binding (Fig. 2.1D). At low concentrations, the system produces almost exclusively IF1-mediated heterodimers. The predominance of heterodimers is attributable

to Ancα's weak homodimerization affinity; the excess of unbound Ancα subunits causes Ancβ subunits to preferentially heterodimerize rather than homodimerize at equilibrium, even though Ancβ's homodimerization affinity is slightly stronger than its heterodimerization affinity (Fig. 2.3D). As protein concentration increases, these dimers begin to assemble with each other across IF2 into tetramers. The excess of heterodimers over homodimers means that the vast majority of the tetramers are heterotetramers, even though IF2 itself does not distinguish between subunit types. At physiologically relevant concentrations of 3mM total Hb subunits (98), the population is dominated by $\alpha_2\beta_2$ heterotetramers, with a small fraction of heterodimers and virtually no homotetramers (Fig. 3D; right panel).

Taken together, these data establish that the measured specificity of IF1 alone mediates highly specific assembly of Ancα+ Ancβ into heterotetramers, even when IF2 is entirely nonspecific -- which our previous experiments suggest is the case – because IF1 is a much stronger interface than IF2. The historical acquisition of heterospecificity across IF1 after the Ancαβ gene duplication is therefore sufficient to account for the evolution of Hb's heterotetrameric architecture.

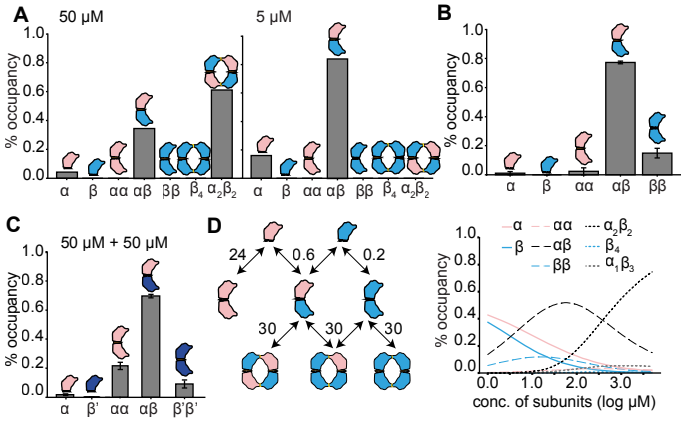

Figure 2.3. Heterotetramer specificity is conferred by specificity at IF1. (A) Occupancy (as fraction of all Hb subunits) when Ancα +Ancβ are coexpressed, measured by native MS. At 50 uM total protein, heterotetramers and heterodimers predominate (left). At 5 uM (right) – at which assembly occurs only across the high-affinity interface (IF1) -- all dimers are heterodimers. (**B)**

Occupancy of subunits in stoichiometries as measured by nMS when Ancα and Ancβ are separately expressed and then mixed at 50 mM each; IF2-mediated tetramer assembly does not occur under these conditions, and dimers are predominantly heterodimers. Error bars represent SEM for three replicates. **(C)** Percent occupancy of stoichiometries when Ancα and Ancβ' (Ancβ with all derived IF2 surface residues reverted to the state in Ancαβ) are expressed separately and then mixed at 50 uM. **(D)** Predicted occupancy of multimeric stoichiometries if IF1 is specific and IF2 is nonspecific. Left: binding scheme with experimentally estimated Kds (in mM) for IF1 and IF2-mediated multimerization by Ancα + Ancβ, assuming that all IF2 Kds are equal (for Kds, see Fig. 4D and 1D). Right: expected occupancies of each monomer, dimer, and tetramer, given the binding scheme at left. Occupancies are expressed as the fraction of all subunits in each species.

2.5 Heteromeric specificity evolved primarily by reducing homodimerization affinity of Ancα

Given our finding that heterospecificity evolved at the IF1 interface, we next sought to characterize whether the acquisition of specificity was driven by evolutionary changes in the α subunit, the β subunit, or both.

The heterospecificity of a pair of dimerizing proteins can be quantified in energetic terms as the difference in the ΔG of binding between the heterodimer and the mean of the two homodimers ($\Delta\Delta G_{spec}$; see methods for calculation). If $\Delta\Delta G_{spec} = 0$, then the fractional occupancy of the heterodimer at saturating and equal concentrations of subunits will be 50%, as will the sum of the homodimers; this is true even if the homodimer ΔGs are very different from each other, as long as the heterodimer ΔG is halfway between them. By contrast, if $\Delta\Delta G_{spec}<0$, then heterodimers will account for the majority of dimers; conversely, if $\Delta\Delta G_{spec}>0$, homodimers together will predominate (Fig. 2.4A-C). Hetero- or homospecificity thus arises when two paralogs contribute nonadditively to dimerization. Whether or not the system is hetero- or homospecific, the two homodimers will have equal occupancies to each other at saturating conditions: irrespective of the fraction of subunits assembled into heterodimers, the remaining subunits will be at equal concentrations and by definition will be well above the homodimerization Kds (93).

We quantified the heterospecificity of Ancα and Ancβ at IF1 by estimating $\Delta\Delta G_{spec}$. We used nMS to measure the homodimer and heterodimer affinities of Ancα and Ancβ', which contains all substitutions that occurred along the Ancβ branch except those that mediate tetramerization across IF2, allowing us to prevent tetramerization and thus isolate the specificity effects at IF1. From these affinities, we calculated the ΔG of binding and the expected fractional occupancy of each dimer at high and equal concentration of subunits. For Ancα+Ancβ', we found that $\Delta\Delta G_{spec} = -1.3$ (in units of kT) and heterodimer occupancy of 82% (Fig. 2.4D). This represents the total specificity acquired by the two diverging paralogs after the duplication of of Ancαβ, which by definition had no specificity. This specificity was acquired because of evolutionary changes in all three relevant affinities. Relative to the ancestral dimerization affinity of Ancαβ, Ancα's energy of homodimerization became worse ($\Delta\Delta G = 0.9$) while homodimerization by Ancβ' improved substantially ($\Delta\Delta G = -3.7$). The heterodimer affinity improved by $\Delta\Delta G = -2.7$, substantially more than the average of the two homodimers, yielding the observed strong preference for the heterodimer.

We next sought to isolate the contribution of the evolutionary changes that occurred along each of the two branches. To measure the specificity acquired along the branch leading to Ancα, we measured affinities and calculated $\Delta\Delta G_{spec}$ when Ancα is mixed with the deeper ancestor Ancαβ. This pair of proteins is heterospecific, with $\Delta\Delta G_{spec} = -1.2$ (expected heterodimer occupancy 76%). Changes in the α subunit alone therefore account for >90% of the total $\Delta\Delta G_{spec}$ that was acquired by the entire Ancα+Ancβ' system. This specificity is acquired via a 2.6-fold reduction in Ancα's homodimerization affinity compared to the Ancαβ ancestor and a 1.8-fold improvement in heterodimer affinity (Fig. 2.4E; Fig. A1.7B & D).

To isolate the contribution to IF1 specificity of evolutionary changes that occurred along the branch to Ancβ, we measured affinities when Ancβ' is mixed with Ancαβ. This pair of proteins is weakly heterospecific, with $\Delta\Delta G_{spec} = -0.3$ and expected heterodimer occupancy of just 58%. The specificity is weak because both the homodimer and heterodimer affinity improved, but the deviation of the heterodimer from the average of the homodimers is very small (Fig. 2.4F; Fig. A1.8A & C).

Finally, we assessed whether the evolutionary changes in the Hbα subunit and those in the Hbβ subunit interacted with each other nonindependently. If the changes affect specificity entirely independently, $\Delta\Delta G_{spec}$ should equal the sum of the $\Delta\Delta G_{spec}$ acquired on each of the two branches ($-1.2 + -0.3 = -1.5$). The observed $\Delta\Delta G_{spec} = -1.3$, suggesting a very weak negative interaction between changes in the two subunits, which makes the complex slightly less heterospecific than expected if the substitutions were independent (Fig. 4G).

Taken together, these data indicate that the IF1 specificity acquired by the derived complex Ancα + Ancβ is primarily attributable to substitutions in the α subunit, with substitutions in the β subunit making a much smaller contribution; nonadditive interactions between the two sets of changes did not contribute to the evolution of heterospecificity. The most important factor was that Ancα became much worse at binding itself than at binding Ancβ. Ancβ, by contrast, became slightly worse at binding Ancα than binding itself (Fig. 2.4G).

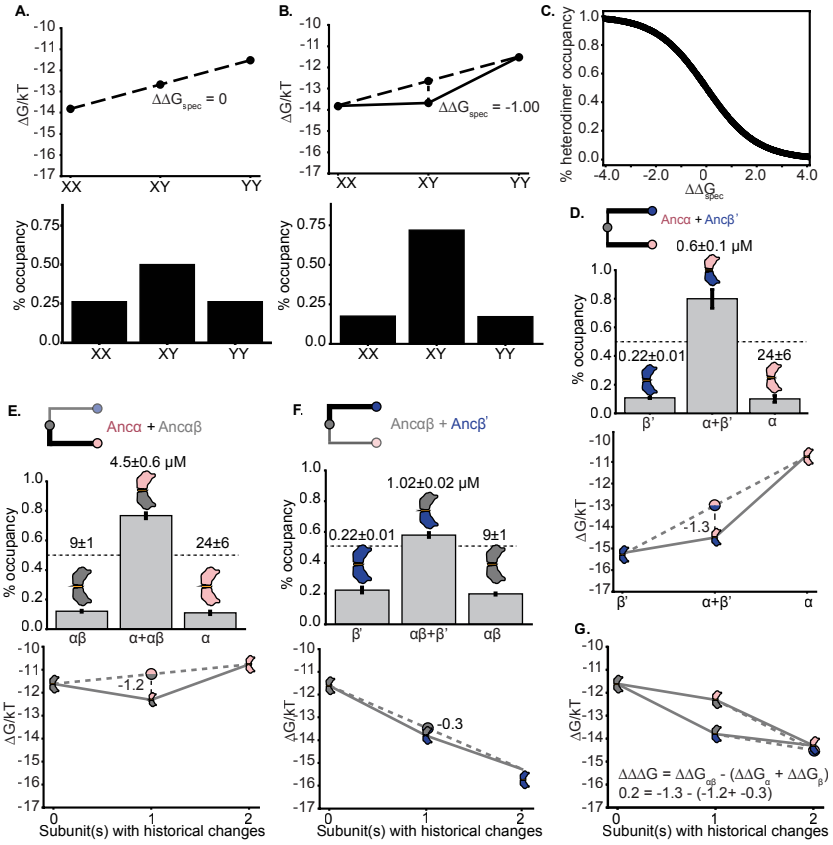

Figure 2.4. Contribution of historical changes in each subunit to the acquisition of heterospecificity. (A) Theoretical example of affinities and occupancy in a system of dimers with no specificity. *Top*: ΔG of dimerization for homodimers (XX and YY) and heterodimers (XY), in units of kT. In the absence of specificity, ΔG of the heterodimer equals the average of the homodimers (dotted line). *Bottom*: expected fractional occupancies of dimers at 1 mM per subunit and dissociation constants (Kd), given the ΔGs in the top panel. In the absence of specificity, heterodimer occupancy = 50%. **(B)** Example of a system with preference for the heterodimer. ΔΔG (the deviation of the heterodimer ΔG from the average of the homodimers) is shown. *Bottom*: Kd and predicted occupancy of each dimer at 1 mM. **(C)** Relationship between ΔΔG and heteromeric occupancy at 1 mM per subunit, assuming the ΔGs of homodimerization for as shown in panel A. **(D)** Specificity of IF1 dimerization in system of Ancα+Ancβ'. *Top*: expected fractional occupancies at 1 mM, given Kds assessed by nMS (shown above each bar, with 95% confidence interval). *Bottom*: ΔGs and ΔΔG given measured Kds. Dotted line, expected occupancies in the absence of specificity. **(E)** Specificity of IF1 acquired on the branch leading from Ancαβ to Ancα, shown as occupancy and ΔGs of the Ancαβ + Ancα system. The number of subunits that contain historical changes in each dimer is shown relative to the Ancαβ homodimer. **(F)** Specificity of IF1 acquired on the branch leading from Ancαβ to Ancβ, shown as occupancy and ΔGs of Ancαβ + Ancβ'. **(G)** Interaction effect on specificity when evolutionary changes leading from Ancαβ to Ancα (pink) and Ancβ' (blue) are combined. Homodimer of Ancαβ (gray) and each heterodimer are plotted by their ΔG. The observed ΔΔG of each heterodimer in combination Ancαβ is shown (see panels D-F). If the specificity acquired in the two subunits affects heterodimerization independently, then ΔΔG of Ancα+Ancβ will equal the sum of the ΔΔGs, yielding a parallelogram. The deviation from this expectation is shown.

2.6 A one-residue deletion was the primary evolutionary cause of heterospecificity

We next sought to identify the particular historical substitutions in Ancα that conferred this heteromeric specificity on IF1. Only three sequence changes occurred on the branch from Ancαβ to Ancα: a single-residue deletion of a histidine at site 3 (ΔH3), a five-residue deletion in helix D (ΔD), and an amino acid replacement (v140A). ΔH3 is on the protein's N-terminal loop near IF1, and ΔD directly contributes to the interface. Substitution v140A is biochemically conservative and far away from the interface. The deletions are conserved to the present in Hba subunits throughout the jawed vertebrates, including humans, whereas the amino acid at site 140 varies (Fig. A1.1). We therefore focused first on the effects of the deletions.

To isolate the contribution of each deletion to the evolution of specificity, we introduced each one singly into Ancαβ and measured its effect on affinity and specificity when the mutant protein is mixed with Ancαβ. We found that introducing ΔH3 alone confers substantial specificity, recapitulating >80% of the Ancα's acquired heterospecificity for Ancαβ ($\Delta\Delta G_{spec}$ = –1.0 out of a total $\Delta\Delta G_{spec}$ = –1.2 acquired along this branch) and >75% of the specificity acquired along both branches by the entire Ancα+Ancβ' complex ($\Delta\Delta G_{spec}$ = –1.3, Figs. 2.5A, C). ΔH3 enhances specificity by improving heterodimer affinity and reducing homodimer affinity, with both Kds very close to those of Ancα (Fig. 2.5A; Fig. A1.10A & C).

The other deletion, ΔD, removes several residues that directly interact with the other subunit across IF1, but introducing this change into Ancαβ had a much weaker effect on specificity ($\Delta\Delta G_{spec}$ = –0.4, Fig. 2.5B; Fig. A1.10B & D). When the contributions of ΔH3 and ΔD to specificity are added together, they slightly exceed the specificity of Ancα (–1.4 rather than –1.3), suggesting the possibility of a very weak negative epistatic interaction between them or a small countervailing effect of the third change v104A. Taken together, these results indicate that

ΔH3 was a large-effect historical sequence change that accounted for most of the specificity historically acquired by the derived Hb complex.

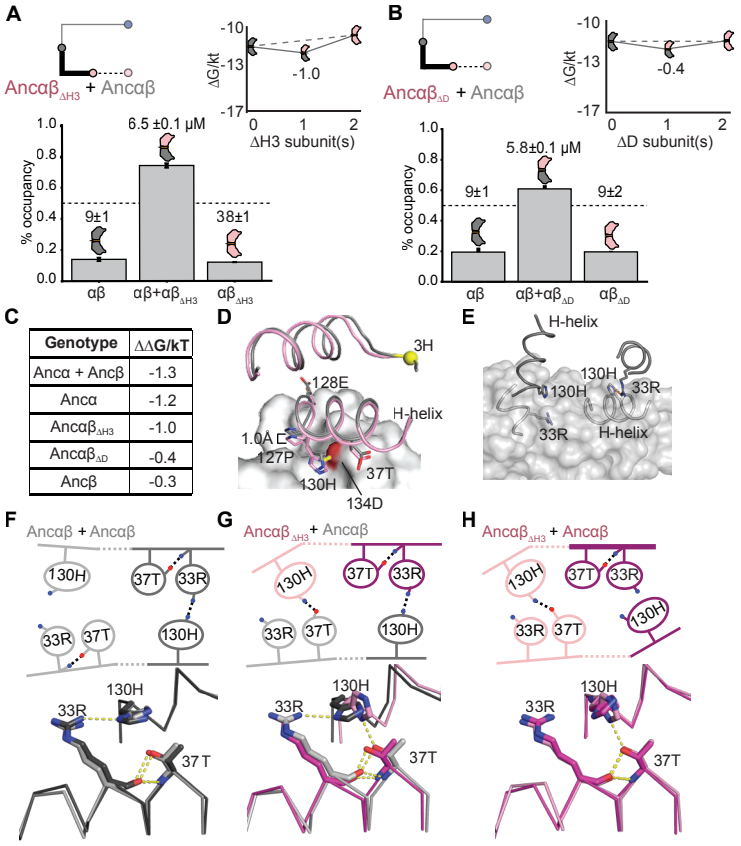

Figure 2.5. Effect of historical sequence changes on specificity. (A) Specificity of Ancαβ$_{ΔH3}$ with Ancαβ, as in Fig. 4. **(B)** Specificity of Ancαβ$_{ΔD}$ with Ancαβ. **(C)** Gain in specificity caused by various sets of historical mutations, relative to Ancαβ. Ancα+Ancβ, all changes on both post-duplication branches. Ancα, all changes on the branch leading to Ancα. ΔH2 and ΔD, deletions that occurred on the Ancα branches. **(D)** Models of Ancαβ homodimer and Ancαβ$_{ΔH3}$ + Ancαβ heterodimer. The N-terminal helix and the portion of IF1 involving helix H is shown. Grey surface, Ancαβ subunit common to both models. Grey cartoon, other Ancαβ subunit in the homodimer; pink cartoon, Ancαβ$_{ΔH3}$ subunit in the heterodimer. Yellow, 3H residue deleted in Ancαβ$_{ΔH3}$. Helix H side chains in the interface are shown as sticks. The hydrogen bond in the heterodimer from 130H to 37T (red surface) is shown (dotted line). **(E)** A portion of IF1 in the Ancαβ homodimer model, showing the isologous interactions with imperfect symmetry between 130H and 33R. Orange dashed-line, hydrogen bond. The two subunits are colored different shades of gray. The surface of the light-gray subunit is shown. **(F ,G, H)** Key residues in IF1 with hydrogen bonds that are affected by ΔH3 in the homodimers and heterodimer of Ancαβ and Ancαβ$_{ΔH3}$. Top, cartoon of key contacts. The two iterations of these interactions across the isologous interface are shown, one each in light or dark hue. Blue and red, nitrogen and oxygen atoms, respectively. Dotted lines, hydrogen bonds. The change in position of the H-helix caused by ΔH2 is shown. Bottom, structural alignment of the two iterations of the isologous interface in each dimer. Each dimer structure was duplicated exactly and then aligned to the original by

targeting one subunit of the copy to align to the other subunit of the original. Hues correspond to the isologous iterations in the cartoon above.

2.7 Structural mechanisms for the gain in specificity

We next considered the structural mechanisms by which ΔH3 conferred specificity by increasing heterodimer affinity and reducing homodimer affinity. For a mutation to have these opposite effects, it must yield favorable interactions when introduced into one side of the interface (in the heterodimer) but have deleterious effects when introduced twice (in the homodimer). In principle, two kinds of mechanisms could cause these opposite effects. Either 1) the mutated residue could interact directly with the same residue on the other subunit favorably when one is in the derived state but unfavorably when both are, or 2) the symmetry of the interface could be imperfect, such that introducing the mutation on one side of the interface is favorable but introducing it again onto the other side is net-unfavorable. The first scenario does not pertain in this case. Residue H3 is part of the N-terminal loop, which does not participate directly in IF1 but instead packs against helix H, which does contribute to IF1. But neither helix H nor the N-terminal loop contact the same elements in the other subunit across the interface (Fig. 2.5D). Asymmetry in the interface is therefore the likely of cause ΔH3's differential effects on heterodimer vs. homodimer affinity.

To gain insight into the possible nature of this asymmetry and potential mechanisms by which DH3 affects specificity, we modeled the structures of the Ancαβ homodimer, the Ancαβ$_{\Delta H2}$ homodimer, and the heterodimer of these two proteins. The modeled Ancαβ homodimer contains a subtle asymmetry: on one end of IF1, residue 130H on helix H sits close to 33R on the opposite subunit, which allows a cross-interface hydrogen bond to form; on the other end of the interface, the two residues are slightly further away from each other, leaving their hydrogen-bonding potential unsatisfied when bound (Fig. 2.5E & F)**.** In the heterodimer, deleting His2 from one

subunit repairs this unfavorable interaction. Specifically, the deletion shortens the N-terminal loop and changes its packing interaction against helix H, which causes helix H to slide along the interface by ~1 Å compared to its position in the unmutated Ancαβ homodimer (Figs. 2.5D, 2.5G). 130H moves closer to 37T on the other subunit, allowing it to form a new hydrogen bond across the interface, and several other interactions across the interface are also enhanced. On the other end of the isologous interactions, the favorable interactions found in the homodimer remain intact. This provides a potential structural explanation for how ΔH3 improves heterodimer affinity (Figs. 2.5D, G).

The modeled Ancαβ$_{\Delta H3}$ homodimer structure is notably asymmetric and suggests a possible mechanism by which introducing ΔH3 into both subunits reduces affinity (Fig. 2.5H). One side displays the favorable new cross-interface interactions caused by ΔH3 in the heterodimer, including the 130H-37T hydrogen bond. On the other side, however, the effect of the deletion is very different: ΔH3 again causes helix H to slide along the interface, but on this side the movement of 130H breaks the ancestral 130H-33R hydrogen bond, and 37T is also too far away to interact favorably. This leaves the side chains of both 130H and 33R unsatisfied, reducing homodimer affinity. In total, the homodimer of Ancαβ$_{\Delta H3}$ contains three unsatisfied hydrogen-bond donors/acceptors at these sites, whereas only one and two are unsatisfied in the heterodimer and the ancestral homodimer, respectively.

These hypothesized mechanisms appear to have persisted over time. The same pattern of interactions are found in the modeled structures of the hetero- and homodimers of Ancα + Ancβ (Fig. A1.11). High-resolution crystal structures of extant hemoglobin also show notable asymmetries in the multimerization interfaces, which exceed the deviation expected given the resolution of the structures (99). These structures include some of the particular asymmetrical

interactions observed in our ancestral models: in the human Hb heterotetramer, 33R hydrogen bonds across IF1 to residue 130, but this interaction is again lacking in the homodimer of human Hba, leaving 33R unsatisfied, potentially explaining the weak homomeric affinity of Hba (Fig. A1.11). At least some of the mechanisms of heterodimer specificity suggested by the structural models of the ancestral proteins are therefore present in the known structures of its present-day descendants. Structural models are prone to error, and the asymmetries we observed are subtle; further research will be required to definitively characterize potential asymmetries in the ancestral multimers.

2.8 Multiple historical sets of substitutions could have conferred heterospecificity

If specificity in an isologous interface can evolve simply by causing nonadditive impacts on the binding energies of heterodimer and homodimers, then there should be many mutations that have the potential to make the interface specific in one direction or another. Indeed, if the interface's symmetry is imperfect, then most mutations that affect affinity should impart specificity to some degree.

To test this hypothesis, we measured the effect on specificity of subsets of changes that occurred along the Ancβ lineage, which the results above show had strong effects on affinity when introduced all together. First, we tested the five substitutions that that occurred at the IF1 surface (Fig. 2.5E & 5F). We introduced these changes into Ancαβ (creating protein $Anc\alpha\beta_{IF1}$) and measured affinity and specificity when this protein is mixed with Ancαβ. These substitutions yield a highly heterospecific complex ($\Delta\Delta G_{spec}$ = –2.2, heterodimer occupancy 90%, Fig. 2.6A; Fig. A1.12A, C, & E). Unlike the Ancα substitutions, the $Anc\beta_{IF1}$ substitutions confer heterospecificity by improving both homodimer and heterodimer affinity, but they improve the latter by more than the former.

Because Ancαβ$_{IF1}$ is specific in complex with Ancαβ, we wondered whether it would also be specific with Ancα. We found that this complex is barely heterospecific ($\Delta\Delta G_{spec}$ = –0.1, Fig. 2.6B), implying that other substitutions on the branch leading to Ancβ but not on the interface must have contributed to the evolution of specificity between Ancαβ$_{IF1}$ and Ancα. We therefore introduced an additional set of five historical substitutions that occurred in Ancβ but one structural layer away from IF1 (69). This protein (Ancαβ$_{IF1+Adjacent}$) has strong heterospecificity when mixed with Ancα ($\Delta\Delta G_{spec}$ = –2.0, heterodimer occupancy >85%, Fig. 2.6D; Fig. A1.12B, D, & F), because these mutations together improve both heterodimer and homodimer affinity, but with a larger improvement in the heterodimer. It is also moderately heterospecific when mixed with Ancαβ ($\Delta\Delta G_{spec}$ = –0.9).

Finally, we tested the effect of the adjacent substitutions on their own and found that they confer specificity when mixed with Ancαβ ($\Delta\Delta G_{spec}$ = –0.9). These mutations impart specificity by causing almost identical changes in homo- and heterodimer affinity. They also confer some heterospecificity when Ancαβ$_{Adjacent}$ is mixed with Ancα ($\Delta\Delta G_{spec}$ = –0.6, Fig. A1.13A-D). There are therefore several distinct sets of substitutions that occurred during history, and which can be sufficient to confer heterospecificity on their own (and in various combinations), and they do so via distinct patterns of effects on affinity. This degeneracy of mechanisms for evolving specificity arises because there are many ways in which the energy of binding can change nonadditively between heterodimer and homodimer. In every case, heteromeric specificity rather than preference for the homomer was the result.

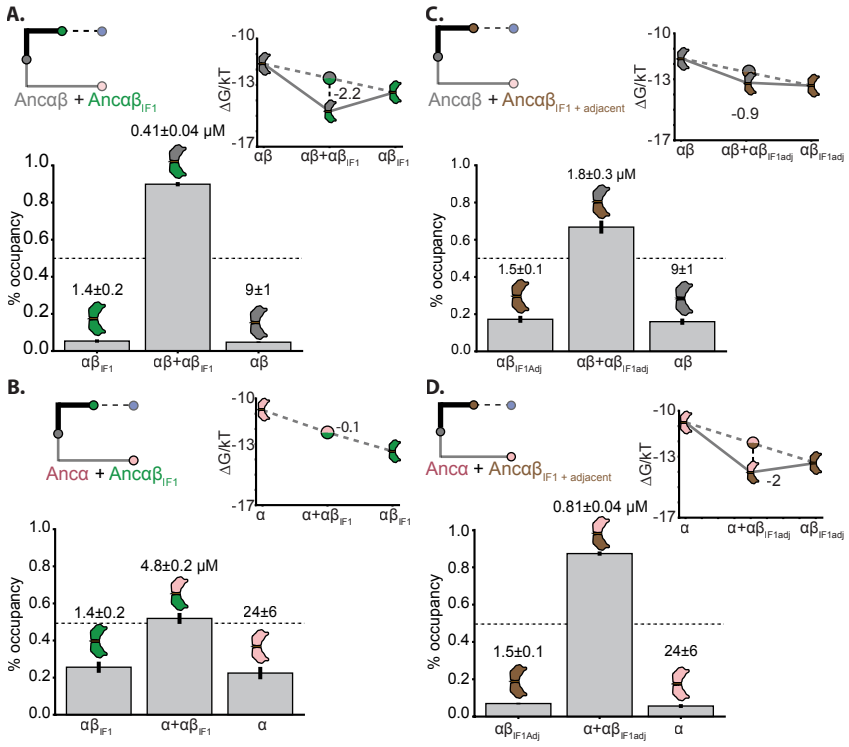

Figure 2.6. Other subsets of historical substitutions confer heterospecificity on IF1. Affinities measured by nMS, predicted occupancy based on those Kds at 1 mM each subunit, and $\Delta\Delta G_{spec}$ are shown for A) Ancαβ + Ancαβ$_{IF1}$, which contains the five substitutions at the IF1 surface that occurred in the Ancβ lineage; B) Ancαβ + Ancαβ$_{IF1 + adjacent}$, which also includes 4 additional substitutions in Ancβ near but not on the interface; C) Ancα + Ancαβ$_{IF1}$, and D) Ancα + Ancαβ$_{IF1 + adjacent}$.

2.9 Discussion

This work provides a mechanistic history of the evolutionary transition from the ancestral Ancαβ homodimer to the derived Hb heterotetramer (summarized in Fig. A1.1). Each transition was driven by a very simple genetic mechanism: a single substitution at IF2 conferred high affinity tetramerization, and a single amino acid deletion at IF1 conferred heteromeric specificity. These two key sequence changes have remained conserved in the descendant Hba and Hbb subunits of all extant jawed vertebrates (Fig. A1.1). These transitions were both facilitated by the isologous architecture of Hb's two interfaces, which creates a propensity for mutation to produce high-order complexes and heterospecificity.

2.10 Symmetry facilitated evolution of the tetrameric stoichiometry

We found that tetramerization across IF2 was driven primarily by a single replacement to a bulky hydrophobic amino acid (q40W). In biochemical studies of extant protein interfaces, much of the free energy change in protein-protein binding is attributable to interactions of bulky hydrophobic residues with hydrophobic surface indentations (100), and mutations to bulky hydrophobic amino acids can drive assembly into high-order multimers (27, 101-104). Similar substitutions during history may have been driving mechanisms during the evolution not only of Hb but of other molecular complexes, as well.

The majority of complexes assemble through isologous interfaces (105). It has been suggested that this must imply that isology confers some selective benefit by improving protein function (75). Our results suggest an alternative explanation. If mutations are much more likely to produce isologous complexes than nonisologous ones, then isologous complexes will predominate in nature, even if there is no systematic fitness difference between the two types of multimer. We found that although IF2 is intrinsically weak and mutation q40W cannot confer dimerization on its own, it can drive tetramerization if its effects are multiplied in an isologous higher-order complex. By contrast, If the interfaces were non-isologous -- with q40W interacting with a hydrophobic divot on some other surface of the facing subunit – then this favorable interaction would appear only once, and it would be insufficient to substantially improve binding energy and confer meaningful tetramer occupancy. By this explanation, isologous complexes are abundant because they are easier to produce by mutation than head-to-tail multimers, not more likely to be fixed by selection.

It has been observed that in high-order multimers, the interface with higher affinity usually evolves before the lower-affinity interface(s) (106-109). Hb evolution displays this pattern, with the stronger interface IF1 evolving before IF2 (17). It has been suggested that this pattern is

attributable to selection: by this hypothesis, selection favors evolutionary intermediates that contain the high-affinity interface, because those containing only the low-affinity interface assemble slowly and/or misassemble into anomalous complexes (106, 107). Our work here suggests a different explanation: it is easier for mutations to generate a new interface that confers a higher stoichiometry if a strong interface is already present, because the affinity of the new interface is multiplied by iteration in an isologous complex. By contrast, low-affinity interfaces do not confer multimerization on their own, so if the low-affinity interface were to evolve first, then the effects of mutations on the second interface would not be multiplied. In complexes with multiple interfaces, the stronger interface tends to be older not because such trajectories improve fitness but because mutation is more likely to build elaborate complexes in this historical order.

2.11 One interface confers specificity on a higher-order multimer

Our experiments show that evolutionary change at just one of Hb's interfaces was sufficient to confer specific assembly into heterotetramers. Specificity at IF1 alone was sufficient to mediate the heterospecificity of the tetramer because this interface is so much stronger than IF2: IF1 mediates the specific assembly of heterodimers, which assemble into heterotetramers across IF2, even though IF2 itself confers little or no specificity.

The specificity of IF1 and the isology of the complex also explains the *trans* conformation of Hb's quaternary structure, in which each Hba subunit binds one Hbβ subunit across IF1 and a different Hbβ across IF2. The alternative *cis* conformation -- in which Hbα is paired with an Hbα (and Hbβ with Hbβ) across one of the interfaces – is never observed. Although IF2 imposes little or no specificity, its isologous orientation necessarily means that the two IF1-mediated heterodimers must be rotated 180° relative to each other, placing each Hbα across IF2 from the Hbβ of the other heterodimer. In the *cis* conformation, the heterodimers would not be rotated

180° relative to each other, and all the favorable interactions that IF2 comprises would not form; residue 40W, for example, would not face the hydrophobic divot on IF2 across the interface. Given the heterospecificity of IF1, isology constrains the Hb tetramer to its *trans* a_2b_2 architecture.

These observations suggest a simple and potentially general mechanism for the evolution of specificity in the quaternary structures of high-order multimers. Specificity need not evolve at every interface in the complex, especially if the interfaces are isologous. Rather, mutations need only make the stronger interface specific to confer assembly into particular high-order architectures.

2.12 Symmetry allowed specificity to evolve in one subunit

We found that a single genetic change in one paralog – a one residue deletion in Anca -- was sufficient to confer IF1's heterospecificity. This result contrasts with prior studies of nonisologous complexes, in which heterospecificity evolved because of genetic changes in both interacting subunits (36, 37, 84, 90-92).

This difference in historical genetic mechanism reflects the opportunities presented by the two different types of multimeric architecture. In nonisologous complexes, a mutation in the "head" of one duplicate gene will not be sufficient to distinguish between its own tail and that of its paralog (unless it somehow changes the conformation of both distinct surfaces). In an isologous complex like Hb, however, a change in one subunit can confer specificity, because it makes the interface different between the heterodimer, the mutated homodimer, and the unmutated homodimer.

Acquiring specificity in an isologous interface requires the mutation to nonadditively change the affinity of the heterodimer relative to the homodimers. If the symmetry of such interfaces were

perfect, a mutation in one subunit would affect interactions across the interface identically on each side of the interface, resulting in additive effects on affinity. Nonadditivity would arise only if mutations affect sites that interact with each other across the rotated interface. This would require either a mutation at the precise axis of rotational symmetry or multiple mutations at several sites.

If the symmetry is imperfect, however, a single mutation can affect interactions differently when it appears twice in the homodimer versus when it occurs once in the heterodimer. Imperfect asymmetry could facilitate the evolution of specificity in many complexes. Virtually all isologous interfaces contain subtle asymmetries (110). This imperfection arises for two reasons: perfect symmetry is entropically unfavorable, and amino acids near the axis seldom face each other with perfect symmetry, because each amino acid itself is asymmetrical, and this asymmetry propagates elsewhere in the interface (110, 111).Extant human hemoglobin is one of many examples of isologous interfaces in which asymmetry is imperfect (59, 99). Isologous interfaces therefore provide a starting point for homo- or heterospecificity to be acquired by substitutions in a single subunit.

2.13 Specificity evolved through a single mutation

We found that a single mutation – deletion of residue His2 in the alpha subunit – conferred most of the heterospecificity of Ancα + Ancβ. This simple mechanism was possible because only a small change in relative binding energy is required to yield substantial changes in specificity. The IF1 of Ancα+Ancβ' mediates 80% heterodimer occupancy at equal and saturating concentrations, but its specificity is quite moderate $\Delta\Delta G_{spec} = -1.3$). This difference in binding energy is less than that associated with a typical hydrogen bond or burial of a large hydrophobic residue. The structural differences in physical interactions across the homodimer vs. heterodimer

interfaces in our modeled structures could easily yield energetic differences of this magnitude, although the particular form of asymmetry in ancestral hemoglobin complexes remains uncertain. Recent in silico work also found that small differences in ΔG can cause large differences in occupancy between homodimers and heterodimers (93).

Why do such subtle differences in energy have such large impacts on specificity? Mutations that cause a modest deviation in binding energies can cause large changes in occupancy because of the nonlinear Boltzmann relationship between these quantities (Fig. 2.4C). Moreover, specificity is determined by the deviation from additivity in the heterodimer relative to the homodimers, so small differences in the free energy of binding will propagate into even larger changes in specificity. Because of this intrinsic sensitivity, we predict that the evolution of specificity in paralogous complexes with symmetrical interfaces will often be attributable to one or a few genetic changes with relatively small energetic effects and subtle structural mechanisms. That specificity can evolve so easily also implies that paralog interference after gene duplication (112, 113) may often be easily resolved through one or a few mutations.

If specificity can be acquired by small deviations from energetic additivity in either direction, one might expect that homomeric and heteromeric specificity would be equally likely to evolve. But empirical observations suggest that heteromers evolve much more frequently after gene duplication (83, 84). Our findings suggest a plausible explanation for this pattern. The modeled structures suggest that the critical mutation for conferring specificity on Hb does so because imperfect asymmetry in the interface creates a kind of antagonistic pleiotropy: a favorable interaction occurs when the mutation is introduced once in the heteromer, but it fails to produce the same favorable contact and even disrupts a different favorable contact when introduced again on the other side of the interface in the homomer. Heterospecificity will result whenever

asymmetry causes antagonistic pleiotropy like this, such that a favorable interaction can be optimized when it is iterated once but not twice. In contrast, homomeric specificity requires a mutation to be even more favorable the second time it is introduced on the other side of an interface. For this to occur, imperfect symmetry must synergistically enhance the interactions caused by the two iterations of the mutation in the homodimer. This scenario seems far less likely than an antagonistic effect, because favorable interactions are constrained in many ways, requiring fairly precise compatibility of polarity, size, angle, etc. The imperfect symmetry of isologous interfaces may therefore create a mutational propensity that favors the evolution of heteromeric over homomeric specificity.

Taken together, our observations contribute to a growing body of evidence that complex multimeric complexes can evolve through simple genetic mechanisms (44, 79, 86, 101, 114-117). In Hb evolution, a single substitution in one of the duplicated genes was sufficient to cause a doubling in stoichiometry from dimer to tetramer, and a single-residue deletion at one interface in the other subunit was sufficient to confer strong preference for the $\alpha_2\beta_2$ heterotetrameric form. Although other substitutions enhanced these effects, and others may have permitted or entrenched them (23, 79), our data indicate that discrete evolutionary increases in complexity can occur by very short mutational paths from simpler ancestral forms. The major effects of these small sequence changes was possible because they took place in the context of an isologous complex, and it is likely that its symmetry was slightly imperfect. Many multimers share these structural properties, so we predict that, when other multimeric complexes are studied in detail, simple mechanisms will be found to have driven their historical elaboration.

2.14 Methods

Sequence data, alignment, phylogeny, and ancestral sequence reconstruction

The reconstructed ancestral sequences used here are the same as those reported previously (69). Briefly, 177 amino acid sequences of hemoglobin and related paralogs were collected and aligned. The maximum likelihood (ML) phylogenetic tree was inferred using the AIC best-fit model, LG+G+F. The phylogeny was rooted using as outgroups neuroglobin and globin X, which are found in both deuterostomes and protostomes and diverged prior to the gene duplications that produced vertebrate myoglobin and the hemoglobin subunits. Ancestral sequence reconstruction was performed using the empirical Bayes method (11), given the alignment, ML phylogeny, ML branch lengths, and ML model parameters. Reconstructed ancestors that were used in this study have been deposited previously in GenBank (IDs MT079112, MT079113, MT079114, MT079115).

The historical mutations that we introduced into those ancestral proteins are the following. For the set *IF1-reverted*, all sites in IF1 that were substituted on the branch leading to Ancβ are reverted to the ancestral state found in Ancαβ; the mutations introduced are V36t, Y38h, V115a, V119e, H130r, D134e. For the set *IF2-reverted*, all sites that were substituted in IF2 on the branch leading to Ancβ are reverted to the ancestral state found in Ancαβ; the mutations introduced are T37v, W40q, R43t, H100r, E104h. For the set *IF1*, all sites at IF1 that were substituted between Ancαβ and Ancβ are changed to the derived state found in Ancβ; the mutations introduced are t37V, k58M, r107K, h130Q, d134Q4. For the set *Adjacent*, five sites adjacent to IF1 that were substituted between Ancαβ and Ancβ are changed to the derived state found in Ancβ; the mutations introduced are h47S, s60N, q62K, a96S, h97E. The set *IF1+Adjacent* is the union of the sets *IF1* and *Adjacent*. Deletion ΔD deletes residues a54, e55, a56, i57, and k58 from Ancαβ.

Recombinant protein expression

Coding sequences for reconstructed ancestral proteins were optimized for expression in *Escherichia coli* using IDT Codon Optimization and synthesized *de novo* as gBlocks (IDT). Coding sequences were cloned by Gibson assembly into vector pLIC (118) under control of a T7 polymerase promoter. For co-expression of Ancα+Ancβ, a polycistronic operon was constructed under control of a T7 promoter and separated by a spacer containing a stop codon and ribosome binding site, as described in (119).

BL21 (DE3) *Esherichia coli* cells (New England Biolabs) were heat-shock transformed and plated onto Luria broth (LB) containing 50 ug/mL carbenicillin. For the starter culture, a single colony was inoculated into 50 mL of LB with 1:1000 dilution of working-stock carbenicillin and grown overnight. 5 mL of the starter culture were inoculated into a larger 500-mL terrific broth (TB) mixture containing the appropriate antibiotic concentration. Cells were grown at 37° C and shaken at 225 rpm in an incubator until they reached an optical density at 600 nm of 0.6-0.8.

For expression of single globin proteins, 100 uM of isopropyl-β-D-1-thiogalactopyranoside (IPTG) and 25 mg/500 mL of hemin were added to each culture. Expression of single proteins in culture were done overnight at 22° C. Cells were collected by centrifugation at 4,000*g* and stored at -80° C until protein purification. Coexpressed proteins were induced using 500 mM IPTG expression with 25 mg/500 mL hemin for 4 hours at 37°C. Cells were collected by centrifugation at 4,000*g*, immediately followed by purification.

Human hemoglobin was bought commercially (Sigma-Aldridge) and resuspended in PBS.

We attempted to co-express and purify Ancα$\beta_{\Delta H3}$ in complex with Ancαβ_{40W}, but we were not able to identify conditions at which the two species could be expressed and purified to near-equal concentrations.

Protein purification by ion exchange

All singly expressed proteins (all ancestral globins except Ancα+Ancβ) were purified using ion exchange chromatography. All buffers were vacuum filtered through a 0.2 μM PFTE membrane (Omnipore). After expression, cells were resuspended in 30 mL of 50 mM Tris-Base (pH 6.88). The resuspended cells were placed in a 10 mL falcon tube and lysed using a FB505 sonicator (1s on/off for three cycles, each 1 minute). The lysate was saturated with CO, transferred to a 30 mL round bottom tube, and centrifuged at 20,000*g* for 60 minutes to separate supernatant from non-soluble cell debris. The supernatant was collected and syringe-filtered using HPX Millex Durapore filters (Millipore) to further remove debris. A HiTrap SP cation exchange (GE) column was attached to an FPLC system (Biorad) and equilibrated in 50 mM Tris-Base (pH 6.88). The lysate was passed over the column. 50 mL of 50 mM Trise-Base (pH 6.88) was run through the SP column to remove weakly bound non-target soluble products. Elution of bound ancestral Hbs was performed with 100-mL gradient of 50mM Tris-Base 1 M NaCl (pH 6.88) buffer which was run through the column from 0% to 100%. 1.5 mL fractions were captured during the gradient process, all fractions containing red eluant were put into an Amicon ultra-15 tube and concentrated by centrifugation at 4,000g to a final volume of 1 mL. For additional purification, concentrated sample was injected into a HiPrep 16/60 Sephacryl S-100 HR size exclusion chromatography (SEC) column. The column was equilibrated in phosphate buffered saline (PBS) at pH 7.4. Purified ancestral globins elute at different volumes depending on the protein's complex stoichiometry: 48-52 for tetramers, 56-60 for dimers, and 65-67 for monomers. The purified proteins were concentrated as mentioned above and then flash frozen with liquid nitrogen.

For experiments in which two proteins were singly expressed and purified and then mixed together, expression and purification of each protein were performed as described above. The concentration of each protein was then quantified using the Hemoglobin Assay Kit (Sigma). Proteins were then mixed together at 50 mM each for nMS. This procedure was performed in triplicate to assess technical error introduced by the quantification and mixing process.

Protein purification by zinc affinity chromatography

Coexpressed proteins Ancα + Ancβ were purified using zinc-affinity chromatography, which was performed using a HisTrap metal affinity column (GE) on a Biorad NGC Quest. Nickle ions were stripped from the column (buffer 100 mM EDTA, 100 mM NaCl, 20 mM TRIS, pH 8.0), followed by five column volumes of water. To attach zinc to the column, 0.1 M $ZnSO_4$ was passed over until conductance was stable, approximately 5 column volumes, followed by five column volumes of water. After expression, cells were resuspended in a 50 mL lysis buffer (20 mM Tris, 150 mM Nacl, 10% glycerol (v/v), 1mM BME, 0.05% Tween-20, and 1 Roche Protease EDTA-free inhibitor tablet, pH 7.40), sonicated as described above, and the lysate passed through the prepared column. To remove non-specifically bound protein, the column was washed with 50 mL of lysis buffer. Bound protein was then eluted across a gradient of imidazole concentrations (0 to 500 mM) in a total of 100 mL elution buffer (20 mM Tris, 150 mM NaCl, 500 mM imidazole, 10% glycerol, and 1 mM BME, pH 7.4). 1 mL fractions were collected. The fraction corresponding to the second peak of UV absorbance at 280 nm has a visible red color and was collected and concentrated as described above. The concentrated solution was injected into a Biorad ENrich 650 10 x 300 columns for additional purification and eluted in PBS buffer.

Size exclusion chromatography assay

For protein concentrations from 0 to 500 μM, size exclusion chromatography was performed using a Superdex 75 increase 10/300 GL column (GE) equilibrated in PBS, then injected with 250 μL of sample using a 2 mL injection loop on an Biorad NGC Quest FPLC and monitored by absorbance at 280 nm. For proteins at concentration 1 mM, a HiPrep 16/60 Sephacryl S-100 HR was equilibrated in PBS using an AKTAprime FPLC, then injected with 1mL sample and monitored by absorbance at 280 nm.

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. Samples were loaded into gold-coated glass capillaries made in-house and introduced to Synapt G1 HDMS instrument (Waters corporation) equipped with a 32k RF generator (94). 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 °C) and trap bias (30 V) were optimized. Part of the native MS experiments were conducted by Thermo Scientific Exactive Plus Orbitrap with Extended Mass Range (EMR) with tuning as follow: source DC offset of 15 V, injection flatapole DC to 13 V, inter flatapole lens to 5, bent flatapole DC to 4, transfer multipole DC to 3 and C trap entrance lens to 0, trapping gas pressure to 5.0 with the CE to 10, spray voltage to 1.50 kV, capillary temperature to 100 °C, maximum inject time to 100 ms. Mass spectra were

acquired with a setting of 8750 resolution, microscans set to 1 and averaging set to 100. Mass spectra were deconvoluted using Unidec (120).

Calculating multimerization affinity of homodimers

To estimate Kd of the monomer-to-homodimer transition of singly expressed proteins, we performed nMS at variable protein concentrations (P_{tot}). The occupancy of each oligomeric state at each concentration was calculated as the proportion of all globin subunits in that state, based on the summed areas under the corresponding peaks in the native MS spectrum. The fraction of subunits assembled into dimers (*Fd*) includes dimers and tetramers and is defined as

$$Fd = \frac{2x_d+4x_t}{(x_m + 2x_d+4x_t)},$$

where x_m, x_d, and x_t are the total signal intensities of all peaks corresponding to the monomeric, dimeric and tetrameric stoichiometries, respectively. Nonlinear regression was used to find the best-fit value of Kd of dimerization using the equation:

$$F_d = \frac{1}{P_{tot}} * \frac{(4P_{tot} + K_d) - \sqrt{(4P_{tot} + K_d)^2 - 16{P_{tot}}^2}}{4}$$

As an alternative model of homodimerization, we also used a version of the Hill-Langmuir equation:

$$F_d = \frac{P_{tot}}{P_{tot} + K_d}$$

The Hill-Langmuir model, which is typically used for ligand-receptor binding, does not account for depletion of free monomeric subunits as homodimerization takes place and is therefore not a valid model in this case; we used it solely to determine the robustness of the estimated Kd to the specific binding equation used.

Calculating multimerization affinity of homtetramers

To estimate the Kd of the homodimer–homotetramer transition, the fraction of subunits assembled into tetramers is defined as

$$F_t = \frac{4x_t}{(2x_d + 4x_t)}.$$

The concentration of all dimers is defined as

$$P_d = F_d \times P_{tot}.$$

Nonlinear regression was then used to find the Kd of tetramerization using the equation:

$$F_t = \frac{1}{P_d} * \frac{(4P_d + K_d) - \sqrt{(4P_d + K_d)^2 - 16{P_d}^2}}{4}$$

Parameters were estimated using the curvefit script in the Scipy package (121). The 95% confidence interval on the K_d was estimated as 1.96 times the estimated standard error.

Calculating multimerization affinity of heteromers

To determine the Kd of heterodimerization, we used nMS to measure stoichiometries across a titration series in which one protein's concentration was held constant at 50 mM and the other was added at variable concentration (1 to 50 mM). From the nMS spectrum, we estimated the proportion of the heterodimer and the two homodimers as

$$F_{\alpha\alpha} = \frac{2x_{\alpha\alpha}}{\left(2x_{\alpha\alpha} + 2x_{\alpha\beta} + 2x_{\beta\beta} + x_\alpha + x_\beta\right)}$$

$$F_{\alpha\beta} = \frac{2x_{\alpha\beta}}{\left(2x_{\alpha\alpha} + 2x_{\alpha\beta} + 2x_{\beta\beta} + x_\alpha + x_\beta\right)}$$

$$F_{\beta\beta} = \frac{2x_{\beta\beta}}{\left(2x_{\alpha\alpha} + 2x_{\alpha\beta} + 2x_{\beta\beta} + x_\alpha + x_\beta\right)}$$

where each *x* represents the signal intensity of all peaks corresponding to the species denoted in the subscript. The dissociation constant for each dimer is defined as $Kd_1 = \frac{x_\alpha{}^2}{x_{\alpha\alpha}}$, $Kd_2 = \frac{x_\beta{}^2}{x_{\beta\beta}}$, and $Kd_3 = \frac{x_\alpha x_\beta}{x_{\alpha\beta}}$. By substitution, $F_{\alpha\beta}$ can be expressed as

$$F_{\alpha\beta} = \frac{\sqrt{Kd_1 * Kd_2 * F_{\alpha\alpha} * F_{\beta\beta}}}{Kd_3}$$

Kd_3 was estimated using this equation by nonlinear regression, where $F_{\alpha\alpha}, F_{\alpha\beta}$ and $F_{\beta\beta}$ were measured using the titration series, and the affinities Kd_1 and Kd_2 were assigned the values estimated in the homodimerization experiments described above.

Prediction of homodimer and heterodimer occupancy at high concentrations

The occupancy of each dimer at physiologically relevant concentrations (1 mM total globin subunits) was predicted as follows, because nMS is limited to concentrations <100mM. In a mixture of two types of globins *A* and *B*, the total concentration of each subunit can be expressed in terms of the concentration of monomers [A] and [B] in the mixture:

$$[A]_{tot} = [A] + [AB] + 2[AA] = [A] + \frac{[A][B]}{Kd_3} + \frac{2[A]^2}{Kd_1}$$

$$[B]_{tot} = [B] + [AB] + 2[BB] = [B] + \frac{[A][B]}{Kd_3} + \frac{2[B]^2}{Kd_2}$$

We used these equations to predict [A] and [B] at any value of C_A and C_B given the experimentally estimated Kds. The concentration of each dimer was then estimated using the equations $[AA] = \frac{[A]^2}{Kd_1}$, $[BB] = \frac{2[A][B]}{Kd_3}$, and $[BB] = \frac{[B]^2}{Kd_2}$.

Establishing the upper limit of IF2 Kd

We estimated the minimum Kd of assembly across IF2 by Ancαβ $_{37V+40W;\ IF1\ removed}$, because no homotetramer was observed using nMS at a protein concentration of 20 mM. The minimum detection limit for dimers in the nMS assay is 1 mM. Kd is defined as $Kd = \frac{[M]^2}{[D]}$, where $[M]$ and $[D]$ are the equilibrium concentrations of monomer and dimer, respectively. Therefore

$$Kd_{min} = \frac{(20 * 10^{-6})^2 M}{1 * 10^{-6} M} = 400\ uM$$

Determining ΔΔG of specificity

Specificity for heterodimer assembly between two paralogs can be defined as the difference between the additive affinity of the heterodimer and the measured affinity of the heterodimer, using ΔGs derived measured dimerization affinity for two homodimers and their respective heterodimer. The additive affinity of the heterodimer is defined as the averaged ΔG of both homodimers:

$$\Delta G^{additive}_{heterodimer} = \frac{\Delta G_{homodimer\ 1} + \Delta G_{homodimer\ 2}}{2}$$

Specificity is then the difference between the additive and measured heterodimer ΔG.

$$\Delta\Delta G_{spec} = \Delta G^{measured}_{heterodimer} - \Delta G^{additive}_{heterodimer}$$

This metric is analogous to the coupling energy, which expresses the deviation of the measured DG for a double mutant from that expected given the DGs of two single mutants assuming additivity (122-124).

Quantifying non-additive effect on specificity between Ancα and Ancβ

The non-additive effect on specificity can be defined as the difference between the predict and measured ΔΔG of the derived complex Ancα + Ancβ.

$$\Delta\Delta\Delta G = \Delta\Delta G_{\alpha+\beta} - \left(\Delta\Delta G_{\alpha} + \Delta\Delta G_{\beta}\right).$$

Prediction of monomer, dimer, and tetramer occupancies with no IF2 specificity

The occupancy of monomers, dimers, and tetramers between 1 mM and 4 mM predicted was calculated as follows. The concentration of subunit in each stoichiometric species can be expressed in terms of the concentration of monomers [A] and [B]:

$$[\mathrm{A}]_{\mathrm{tot}} = [\mathrm{A}] + [\mathrm{AB}] + 2[\mathrm{AA}] + [\mathrm{ABBB}] + 2[\mathrm{AABB}]$$

$$= [\mathrm{A}] + \frac{[\mathrm{A}][\mathrm{B}]}{\mathrm{Kd}_3} + \frac{2[\mathrm{A}]^2}{\mathrm{Kd}_1} + \frac{\frac{[\mathrm{A}][\mathrm{B}]^3}{\mathrm{Kd}_2 * \mathrm{Kd}_3}}{\mathrm{Kd}_4} + \frac{\frac{2[\mathrm{A}]^2[\mathrm{B}]^2}{\mathrm{Kd}_2^2}}{\mathrm{Kd}_4}$$

$$[\mathrm{B}]_{\mathrm{tot}} = [\mathrm{B}] + [\mathrm{AB}] + 2[\mathrm{BB}] + 2[\mathrm{AABB}] + 3[\mathrm{ABBB}] + 4[\mathrm{BBBB}]$$

$$= [\mathrm{B}] + \frac{[\mathrm{A}][\mathrm{B}]}{\mathrm{Kd}_3} + \frac{2[\mathrm{B}]^2}{\mathrm{Kd}_2} + \frac{\frac{2[\mathrm{A}]^2[\mathrm{B}]^2}{\mathrm{Kd}_2^2}}{\mathrm{Kd}_4} + \frac{\frac{3[\mathrm{A}][\mathrm{B}]^3}{\mathrm{Kd}_2 * \mathrm{Kd}_3}}{\mathrm{Kd}_4} + \frac{\frac{4[\mathrm{B}]^4}{\mathrm{Kd}_3^2}}{\mathrm{Kd}_4}$$

We used these equations to predict [A] and [B] across a range of $[\mathrm{A}]_{\mathrm{tot}}$ and $[\mathrm{B}]_{\mathrm{tot}}$ values given previously measured equilibrium constants. Predicted [A] and [B] concentrations were used to calculate the concentration of homodimers and heterodimers as described above, and the concentration of tetramers were calculated using the following equations:

$$[\mathrm{BBBB}] = \frac{\frac{[\mathrm{B}]^4}{\mathrm{Kd}_2^2}}{\mathrm{Kd}_4}$$

$$[\mathrm{ABBB}] = \frac{\frac{[\mathrm{A}][\mathrm{B}]^3}{\mathrm{Kd}_2 * \mathrm{Kd}_3}}{\mathrm{Kd}_4}$$

$$[\mathrm{AABB}] = \frac{\frac{[\mathrm{A}]^2[\mathrm{B}]^2}{\mathrm{Kd}_3^2}}{\mathrm{Kd}_4}$$

where [BBBB] corresponds to the concentration of homotetramer, [ABBB] is concentration of $\alpha_1\beta_3$ tetramers, and [AABB] is the concentration of $\alpha_2\beta_2$ heterotetramers.

Homology models

SWISS-Model was used to generate a structural model of the Anc$\alpha\beta_{q40W}$ homotetramer using the crystal structure of the human Hbβ homotetramer (PDB 1CMB) as template, which was then refined using Rosetta's Fast Relax protocol, which energetically minimizes the initial structure via small adjustments to the backbone and side chain torsion angles (125). PyMOL V2.1 was used to visualize the proteins and capture images.

IF1-mediated homodimers were generated by the same procedure, except for homodimers of Ancα or Anc$\alpha\beta_{\Delta D}$, for which the homodimer of human Hbα (PDB 3S48) was used as template. IF1-mediated heterodimers were generated by the same procedure but using the heterotetramer of human Hb (PDB 4HHB).

Chapter 3

Evolutionary origins of allostery in vertebrate hemoglobin

3.1 Abstract

Allostery is a fundamental property of many proteins that enables functional regulation by stabilizing one conformation from an ensemble through effector binding at a distant site. The mechanism by which allostery originated historically from a non-allosteric ancestor is unknown. Here, we use ancestral sequence reconstruction and biochemical experiments to investigate how allosteric regulation of oxygen affinity by organic phosphates emerged in vertebrate hemoglobin, focusing on the historical transition from a non-allosteric homodimer to the allosteric $\alpha_2\beta_2$ heterotetramer following gene duplication. We show that allostery emerged through simple genetic mechanisms: either of two independent substitutions on the branch leading to extant β subunits could have conferred allostery in the homotetramer, but in a non-physiological direction that increases oxygen affinity upon effector binding. Two additional substitutions altered conformational stabilities to correct the allosteric response, decreasing oxygen affinity when effector binds. Crucially, the homotetramer was inherently capable of sampling multiple conformations, even in the absence of allosteric regulation, due to an ancient helix movement at the tetramer interface which is triggered by oxygen binding. Our results show that allostery can evolve through tuning of pre-existing conformational ensembles, enabling diverse forms and mechanisms of allostery to arise from a common structural foundation.

3.2 Introduction

Allosteric regulation, the modulation of a protein's activity by effector binding at sites distant from the active site, is a fundamental feature of many protein and essential to their physiological functions (59, 62). Allostery within proteins occur when effector binding stabilizes one functional conformation from an ensemble of conformations. Despite its prevalence, no study has described the mechanisms by which allostery emerged from a non-allosteric ancestor in history.

Allostery appears complex because it relies on multiple functional features: the ability to adopt multiple conformations, to acquire a new effector binding site, and to couple effector binding to a shift in conformational equilibrium (126). Given this apparent complexity, the transition from a non-allosteric to an allosteric protein may have taken numerous substitutions because each feature would have to be built into a non-allosteric protein (127).

However, protein often contain features that are compatible with allosteric regulation, potentially reducing the substitutions needed to produce allostery. Indeed, extant proteins contain latent structural features that have been exploited to engineer allosteric regulation in non-allosteric proteins (42, 43, 128, 129). To date, historical studies have focused on the inversion of allosteric regulation, such as transitions from inhibitory to activating effects, or vice versa, but no study had described the origin of the feature (43, 128-132). Identifying the specific substitutions that conferred allostery, and distinguishing which features were novel versus pre-existing – is essential to understanding how complex regulatory architectures emerged during history.

Here we use ancestral sequence reconstruction (ASR) to investigate the origin of organic-phosphate mediated allostery in vertebrate hemoglobin (Hb), the major carrier of oxygen in the blood of jawed vertebrates. Hb is an $\alpha_2\beta_2$ heterotetramer whose oxygen affinity is decreased by

organic-phosphate effectors like inositol hexaphophate (IHP), which binds between beta subunits in a pocket known as the central cavity (CC) (52, 63, 133). Using ancestral sequence reconstruction and biochemical experiments, we previously showed that allostery in hemoglobin arose after the duplication of Ancαβ, a non-allosteric homodimer and the last common ancestor of all α- and β-globin genes, during the interval leading to the separate ancestors of the α (Ancα) and β (Ancβ) lineages. (Figure 3.1A & B, Figure A2.1 & 2; (69)). Here we describe the mechanisms by which organic phosphate mediated allosteric regulation emergence in Hb. We focus on understanding which portions of this functions were pre-existing features, and which were established through substitutions.

3.3 **Evolution of organic-phosphate mediated allostery**

What substitutions could have conferred allosteric regulation? Previous studies have described a binding pocket for effectors that is between the beta subunits within the heterotetramer, called the central cavity (CC) (134). Because CC mediates organic phosphate binding, we hypothesized that allostery could emerge from changes in this region, but only in the tetramer because it creates the binding pocket.

Starting from the homodimer Ancαβ, we identified five substitutions on the branch leading to Ancβ that we hypothesized could be sufficient to confer allostery. Four of these substitutions (s85K, t142S, e146R, and r149H) occurred at CC. Some of the substituted residues are known to contact organic phosphates directly in extant Hb, and others are involved in modulating relative conformational stabilities (Figure 3.1D (51)). The fifth substitution, q40W, has been shown to have conferred tetramerization in Ancαβ, which is required for allosteric regulation (Figure 3.1E (59, 63)).

To test whether these five substitutions are sufficient to confer allostery, we introduced them into Ancαβ, called Ancαβ$_{q40W + CC}$. We measured the P50 of oxygen binding in Ancαβ$_{q40W + CC}$ without and with 500 uM of IHP using an oximeter, where allostery occurs if oxygen affinity decrease in the presence of effector (135). In the stripped condition, the oxygen affinity is high, increased two-fold relative to Ancαβ (Figure 3.1B & 1F). Upon addition of IHP, oxygen affinity has decreased significantly, with a stronger allosteric response than Ancα + Ancβ (Figure 3.1C & 1F). We also tested the response to ATP, a distinct organic phophsate, to determine robustness of this allostery to other organic effectors and observed a detectable allosteric effect, albiet less than the response with IHP (Fig. A2.3). These results demonstrate that that just five substitutions on the branch from Ancαβ to Ancβ are sufficient to confer allosteric regulation from a non-allosteric homodimeric ancestor.

But were these CC changes sufficient to confer allostery in a heterotetramer, which is the derived complex seen in history? Ancα + Ancαβ$_{14}$ is a stable heterotetramer that was previously identified but is not allosteric (69). Under both stripped and IHP conditions, the protein has identical oxygen affinity (Figure 3.1G). We tested CC if was sufficient to confer allostery by introduce these substitutions into this heterotetramer (Ancα + Ancαβ$_{14 + CC}$) and measured oxygen affinity in both codition. Relative to the stripped condition, oxygen affinity in the presence of IHP decreased oxygen affinity (Figure 3.1G). Introducing CC into a heterotetramer was sufficient to confer allosteric regulation.

Together, these results show that the evolution of allostery in hemoglobin required only five substitutions: one substitution to create the tetramer (q40W) and four in the central cavity. These results are robust to multiple organic phosphate effectors and can occur in both homotetramer

and heterotetramer backgrounds. The evolution of allosteric regulation in Hb is genetically simple.

3.4 Tetramerization and central cavity substitutions are necessary but not sufficient to confer allostery in Hb

These 5 substitutions can be separated into distinct categories on the structure of Hb, q40W is at the tetrameric interface and the other 4 are at the central cavity. Introducing each group of substitutions on their own will help us understand how each group of substitutions helped to contribute to allostery and whether each of these sets could have conferred allostery on their own. To test this, we created two Ancαβ variants - Ancαβ$_{q40W}$ which can tetramerize but has no changes at the central cavity and Ancαβ$_{CC}$ which cannot tetramerize but has changes at the central cavity. Each one ancestral protein variant will tell us if the changes introduced were sufficient to confer allostery and whether the other set of changes not introduced into the genotype were necessary.

What was the effect of q40W alone and could it have conferred allosteric regulation in Hb? Ancαβ$_{q40W}$ decreased oxygen affinity in the stripped condition relative to Ancαβ (Figure 3.1F). The substitution q40W did not confer allostery on its own: in the presence of 500 uM of IHP the oxygen affinity of the protein is equal to that of the stripped condition. Then, q40W decreases oxygen affinity relative to Ancαβ but is not sufficient to confer allosteric regulation.

What was the effect of the CC changes, and could they have conferred? Ancαβ$_{CC}$ increased its oxygen affinity in the stripped condition relative to Ancαβ (Figure 3.1F). In the presence of IHP, the oxygen affinity of the protein is equal to that of the stripped condition. The effect of all 4 CC substitution was to increase oxygen affinity, but they were not sufficient to confer allostery regulation.

Together, each set of substitutions – q40W and CC- are not sufficient on their own to confer allostery, because each set contributes a critical part of the allosteric mechanism for the protein. Introducing q40W into the dimeric ancestor Ancαβ creates a tetramer and allows for population of a low affinity state by decreasing oxygen affinity 2-fold (Figure 3.1C & 1F). The CC substitutions contributed to allosteric regulation by allowing population of a high affinity state in the stripped condition by increasing oxygen affinity 1.3-fold (Figure 3.1C & 1H). Together, they allow for allosteric regulation by organic phosphate effectors.

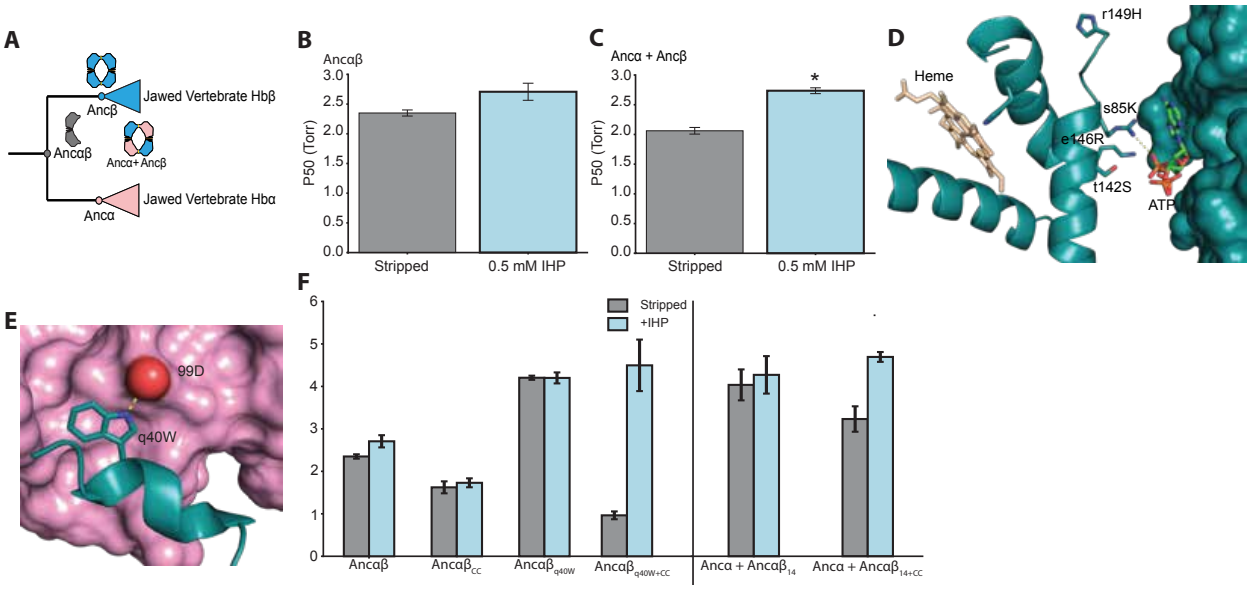

Figure 3.1. Simple genetic mechanism conferred allostery in ancestral hemoglobin. (A) Evolution of allosteric regulation on the HB phylogeny. Icons, oligomeric states determined by experimental characterization of reconstructed ancestral proteins. **(B)** Bar graph of oxygen affinity of Ancαβ in the presence and absence of IHP. In grey, stripped condition, where no IHP is in solution. In blue, IHP condition, where 500 μM of IHP is added to solution. Error bars represent standard error of measurement, n = 5. **(C)** Oxygen affinity bar graph of Ancα +Ancβ in the presence and absence of IHP. Colors and error bars same as mentioned above. Stars represent P-value: <0.05 between conditions as done via student t-test. **(D)** Alphafold-3 models of central cavity within the Ancα +Ancβ tetramer. Blue surface and helices shown correspond to Ancβ subunits. In wheat, heme corresponding to subunit shown as helix with proximal histidine that coordinates heme iron shown as sticks. In green, ATP molecule bound to residues in central cavity. Residue substitutions on the branch from and Ancαβ to Ancβ that occurred in the central cavity shown as sticks. The hydrogen bonds between central cavity residues and ATP shown (dotted line). **(E)** Alphafold-3 model of q40W interaction across IF2 within the AncA + AncB tetramer. Blue helix corresponds beta subunit and pink surface is alpha subunit. The resiude 40W

shown as stick interacting with ancestral residue 99D. Dark blue atom correnods to nitrogen and red surface correpons to oxygen on 99D sidechain. Hydrogen bond between 40W and 99D shown as dotted line. **(F)** *Right:* affinity bar graph of single ancestral gene variants: Ancαβ, Ancαβ$_{CC}$, Ancαβ$_{q40W}$, and Ancαβ $_{q40W\ CC}$. Colors and errors bar same as mentioned above. *Left:* affinity bar graph for protein variants containing two ancestral genes: Ancα + Ancαβ$_{14}$ and Ancα + Ancαβ$_{14\ CC}$.

3.5 Complete binary landscape of central cavity substitutions reveals degeneracy in the evolution of allosteric regulation

Given our previous analysis, where both q40W and CC substitutions were found to be necessary for conferring allosteric regulation, we were interested in understanding their effect on oxygen affinity and how that confers allostery. Given our current results and previous study, we understand the effects of q40W on allosteric regulation - it was the key substitution in conferring tetramerization, and the effect of this change is to decrease oxygen affinity relative to the ancestral protein Ancαβ (69). But we know little about the effect of the CC substitutions, on their own and in groups, and how they contributed to the evolution of organic phosphate regulation.

To investigate how the CC substitutions contribute to allostery in hemoglobin, we constructed a complete binary substitution landscape incorporating all possible combination of the four CC substitutions—s85K, t142S, e146R, and r149H—on the Ancαβ$_{q40W}$ background. We analyzed their effects both in the stripped condition and in the presence of IHP to determine how each individual substitution, as well as their combinations, influence allosteric regulation. Surprisingly, 3 of 4 triple CC substitution combinations were sufficient to confer allostery. The triple mutant, Ancαβ$_{q40W\ +\ s85K\ +\ t142S\ +\ e146R}$, had the strongest allosteric effect with an oxygen affinity of 1.5 torr in the stripped condition and 3.5 torr in the presence of IHP (Figure 3.2A). The other two allosteric mutants, Ancαβ$_{q40W\ +\ s85K\ +\ e146R\ +\ r149H}$ and Ancαβ$_{q40W\ +\ t142S\ +\ e146R\ +\ r149H}$, had weaker decreases in oxygen affinity in the presence of IHP but still confer allosteric

regulation. Of all central cavity changes, only 146R was necessary, as the only triple mutant with no allostery does not contain that substitution (Figure 3.2A). But no set of triple mutants has an allosteric response as strong as Anc$\alpha\beta_{40W + CC}$.

What were the effects of pairwise substitutions in the central cavity? One mutant Anc$\alpha\beta_{q40W + t142S + e146R}$ had a weak allosteric effect on oxygen affinity in the presence of IHP (Figure 3.2B). Another mutant, Anc$\alpha\beta_{q40W + e146R + r149H}$, showed an inverse allosteric effect. All other mutants were non-allosteric and each of them varied in oxygen affinity, ranging from 1.9 torr to 4 torr.

Finally, we examined the effect of single substitutions on oxygen affinity and allosteric regulation. No single mutant allosterically decreased oxygen affinity in the presence of IHP (Figure 3.2C). But, two single mutant showed inversed allosteric signal, increasing oxygen affinity relative to the stripped condition. The mutant Anc$\alpha\beta_{q40W + t142S}$ had the strongest oxygen binding increase in the presence of IHP, going from 3.8 torr to 1.9 torr and Anc$\alpha\beta_{q40W + t149H}$ also showed an inversed allosteric response (Figure 3.2C). All the single substitution mutants increase oxygen affinity in the stripped and IHP conditions. The substitution with the strongest oxygen affinity increase in the stripped condition was e146R.

Together, our results show that the evolution of allostery in Hb was degenerative. Across all 16 possible combinations of substitution at the central cavity, four were allosterically regulated by IHP (Figure 3.2A & B). Two single mutants are also allosterically regulated but in the opposite direction than what has been seen in history. Of the ones that are allosterically regulated in the correct direction, we notice that only e146R is necessary to confer regulation and all other substitutions are not necessary as there is at least one allosteric variant in which each substitution does not appear. But, on its own e146R is not sufficient to confer allostery. These results

emphasize the interaction between substitution and how they, together, result in many variants that conferred allostery during Hb evolution.

3.6 The evolution of Hb allostery occurred because of additive and epistatic genetic effects

We found that all allosteric variants required combination of central cavity substitutions, although only 1 was necessary in all backgrounds. These results emphasize the interdependency of substitution effects. A question that arises from this is whether the evolution of stripped and IHP conditions are a result of additive genetc effects or if epistatic interactions, the interdependence of mutational effects, were required (13).

We determined how well the single mutant effects predicted the pairwise, third order, and fourth order mutant oxygen affinities in the stripped and +IHP conditions. To do so, we estimated the $\log_{10}$ oxygen affinities for all pairwise and higher mutants using only the additive effects of each single mutant that comprises the higher order variants and calculated the pearson correlation coefficient to quantify how well the predicted affinities correlate with the observed $\log_{10}$ oxygen affinities. The stripped condition predicted effects correlate strongly with the observed effects, with a Pearson correlation of 0.97 (Figure 3.2D). In the +IHP condititon, however, have a Pearson correlation of -0.27 showing the predicted effects poorly correlated with the observed effects. These results suggest that epistatic interactions are involved in + IHP conditions. Together, these results show that both additive and epistatic effects caused the evolution of Hb allostery. The tetramerization caused by q40W decreases oxygen affinity in both the stripped and IHP conditions, but the protein is not allosteric (Figure 3.1F). The substitutions at the central cavity additively increased the oxygen affinity of Hb (Figure 3.2D). These same substitutions, however, epistatically interact with IHP because the single substitutions on their own are sufficient to confer allostery, but in the opposite direction of allosteric response in the WT Hb

proteins (Figure 3.1C & 2F). To create an allostery in Hb, there must be sign epistasis, which inverts the mutational effect and decreases oxygen affinity in the presence of IHP.

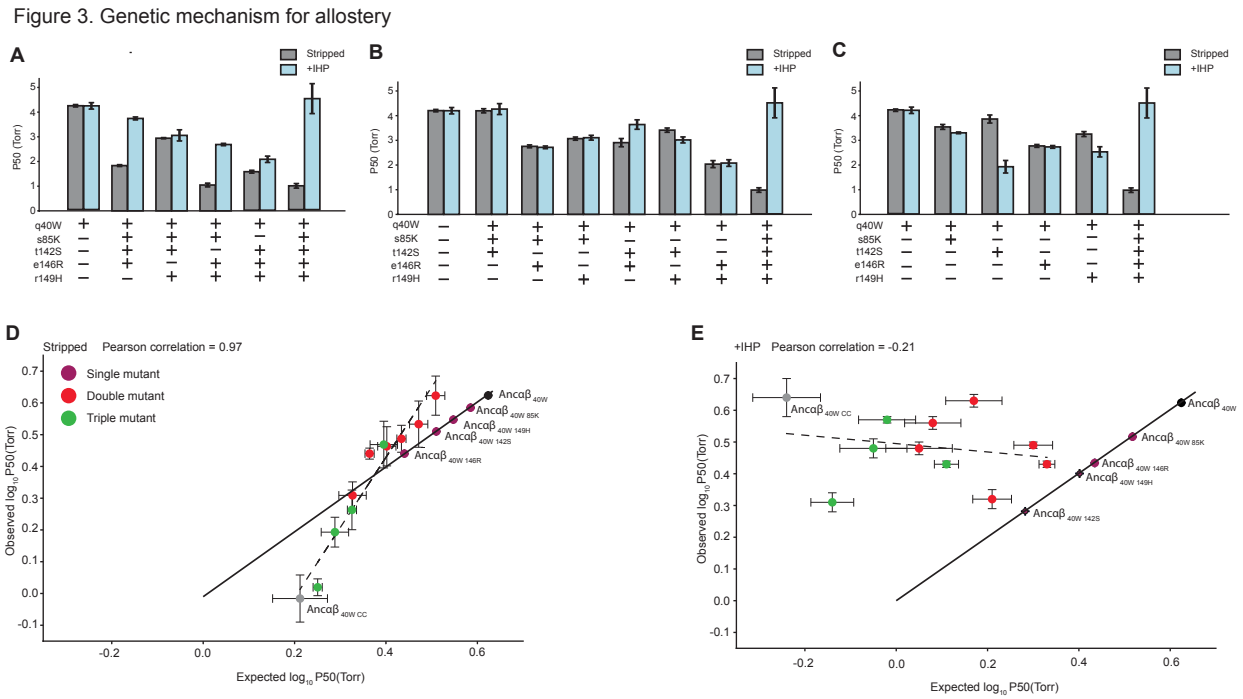

Figure 3.2. The genetic mechanism for allostery is degenerative. (A) Bar graph of oxygen affinity of Ancαβ$_{q40W\ CC}$ and triple central cavity mutants, where each variant was tested in the presence and absence of IHP. In grey, stripped condition, where no IHP is in solution. In blue, IHP condition, where 500 μM of IHP is added to solution. Error bars are standard error of measurement, n = 3. Bottom legend corresponds to the variant specified above, where (+) means that this substitution is present in the variant and (-) means that the substitution is absent in the variant. **(B)** Oxygen affinity bar graph of all double central cavity mutants, each variant was tested in the presence and absence of IHP. Bar colors and errors bars same as mentioned above. Bottom legend works as previously mentioned. **(C)** Oxygen affinity bar graphs of all single central cavity mutants. Tested conditions, bar color, and error bars are same as mentioned above. Bottom legend same as previously mentioned. **(D)** Scatter plot of stripped condition observed log10 oxygen affinity of pairwise, 3rd order, and 4th order central cavity variants vs log10 predicted oxygen affinity of higher-order mutants from single mutant effects. In black and purple, Ancαβ$_{q40W}$ and single mutant variants. Red, green, and grey correspond to double mutant, triple mutants, and Ancαβ$_{q40W\ CC}$ respectively. Solid black line corresponds higher order mutants are completely additive relative to affinity predicted from lower order effects. Dash line is the Pearson correlation relationship between observed and predicted values. **(E)** Scatter plot of IHP condition observed log10 oxygen affinity of higher-order mutants vs log10 predicted oxygen affinity of higher-order mutants from single mutant effects. All scatter point colors and lines same as mentioned above.

3.7 Evolution of multiple conformational states in Hb coopted an ancient helix movement

A critical question in understanding the evolution of allostery is determining which residues are conferring the protein's ability to exist in multiple conformations and whether toggling between states can exist without allosteric regulation.

The structural basis of hemoglobin's conformational change is well-characterized: upon oxygen binding, the quaternary structure undergoes a rotation across IF2, burying exposed interface residues. These changes have been detected using fluorescence emission scans at 280 nm, where deoxygenated human Hb has a higher fluorescence relative to the oxygenated condition (136-138).

We determined that the ability to toggle between oxygenated and deoxygenated conformations through IF2 evolved during the historical interval between Ancαβ and Ancα + Ancβ by measuring fluorescence emissions of each protein. Ancαβ is a dimer that does not assemble through IF2 and should not be able to toggle between oxy and deoxy states. Indeed, when fluorescence emission is measured in oxy and deoxy conditions, we see similar fluorescent values (Fig 3.3A). Ancα + Ancβ is a tetramer that assembles through IF2 and should be able to toggles between the two conformational states. When fluorescence emission is measured in oxy and deoxy conditions, we see that deoxygenated fluorescence is higher by 11%, in the direction seen in present-day Hb, albiet with lower total relative fluorescence (Fig 3.3B, Figure A2.4). The ability to occupy oxy and deoxy states emerged during the historical interval between Ancαβ and Ancα + Ancβ.

What substitutions are sufficient to confer toggling between conformational states in Hb and does the protein need to be allosteric for this to occur? Previous studies have shown that myoglobin, the sister to all Hb genes, can undergo motion at the FG-corner upon the release of

carbon monoxide, which has been noted as the critical structural position at IF2 that allows for occupancy for the T and R states in Hb (139, 140). If this movement is ancestral, as suggested by homology between Hb and myoglobin, then the ability to toggle between conformations would emerge as soon as the tetramer was established.

We tested this hypothesis by introducing the substitution q40W into Ancαβ, the dimeric ancestor that does not assemble through IF2. As mentioned previously, this substitution conferred tetramerization, but was not allosteric, allowing us to assess whether tetramerization alone was sufficient for conformation switching (Figure 3.1F). Fluorescence emission measurement show that $Anc\alpha\beta_{q40W}$ exhibits higher fluorescence in the deoxygenated states than in the oxygenated states, consistent with the conformation changes observed in extant Hb (Figure 3.3C, Fig. A2.4). The tetramer alone could occupy multiple conformational states because of movements at the FG corner.

If the FG-corner movement occupancy of multiple conformational states, just how old is this structural feature? Upon ligand binding, the corresponding helices in Hb and myoglobin will move to accommodate the bound oxygen, suggesting origins deeper than the last common ancestor of Hb and myoglobin genes (Figure 3.3E). Indeed, neuroglobin, one of the earliest diverging vertebrate globin genes, has been noted to rotate at the FG-corner upon oxygen binding (Figure 3.3F (141)). Thus, the movement of the FG-corner, which dictates conformational states within hemoglobin has been present since the dawn of all vertebrates. Together, the evolution of conformational switching in Hb emergence through cooption of an ancient feature of the globin fold. The globin fold must accommodate oxygen binding and do so by altering the structural position of the helices around the heme (139). The emergence of conformational switching between T and R then occurred immediately upon the evolution of the

tetramer, because conformational switching exploits the preexisting structural feature of the globin fold. The stabilization of the discrete states is then a consequence of tetramerization, arising direction from favorable contacts across IF2, rather than because of allosteric regulation.

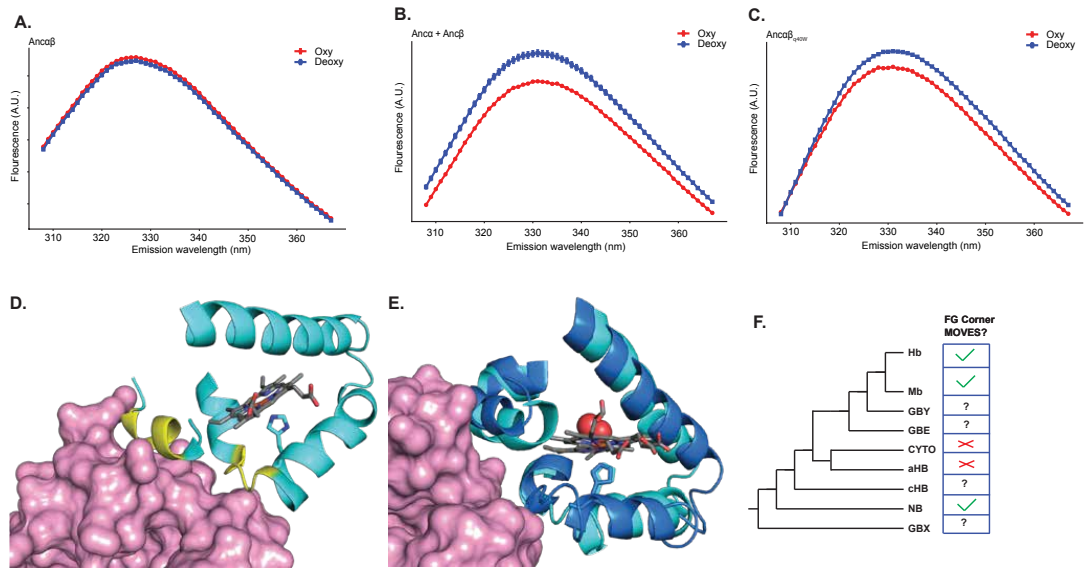

Figure 3.3 Conformational toggling exploited an ancient structural feature of globin fold. **(A)** Fluorescence emissions scans of Ancαβ when excited at 280 nm. In red, protein is oxygenated. In blue, the protein is deoxygenated. Error bars represent standard error of measurement, n = 10. **(B)** Flourescence emission scans of Ancα + Ancβ when excited at 280 nm. Colors and error bars are same as mentioned above. **(C)** Fluorescence emission scan of Ancαβ$_{q40W}$ when excited at 280 nm. Color and error bars same as mentioned above. **(D)** Heme pocket and IF2 in Ancα + Ancβ. Pink surface and blue helices represent alpha subunit and beta subunit, respectively. Grey sticks are heme, red and blue atoms represent oxygen and nitrogen atoms. Blue stick represents proximal heme. Yellow helices are IF2. **(E)** Human Hb conformational states move across IF2 when in the oxygenated (PDB extension) and deoxygenated conditions (PDB extension). Pink surface is alpha subunit. Light blue represents oxygenated condition of beta subunit and dark blue represent deoxygenated condition. Grey sticks is heme, red sphere is O2 atoms, red stick is oxygen atom and blue stick is nitrogen atom. Blue sticks represent proximal histidine. **(F)** Graphical representation of vertebrate globin phylogeny and whether they contain movement at the FG corner.

3.8 Hb allosteric regulation may evolved through interactions that alter conformational stability

Conformational toggling existed in Anc$\alpha\beta_{q40W}$, but the protein is not allosteric. So how does allosteric regulation evolve?

In principle there are four conditions that need to be satisfied for Hb to have become allosteric. The protein must be able to occupy multiple conformations, which are functionally distinct from each other. Hb must be able to bind an effector, and the binding is preferentially associated with one conformation. Finally, because hemoglobin is a tetramer, different quaternary conformations are involved and functionally distinct tertiary conformations are associated with distinct quaternary conformations, and the effector preferentially binds to one of these. What substitutions satisfy each of these conditions?

We know that conformational toggling that q40W is sufficient to allow for occupancy of multiple conformations but that this protein is allosteric (Figure 3.3C). The addition of the other substitutions must satisfy the rest.

Introducing t142S on the background of Anc$\alpha\beta_{q40W}$ allows for binding and preferentially associates binding with a distinct conformation. The protein is now allosteric and is regulated by IHP (Figure 3.2C). However, the function of the protein is to increase oxygen affinity, in contrast to how the mechanisms of allostery work in Hb (Figure 3.2A). We were interested in understanding whether the functional linkage between tertiary and quaternary conformations remained consistent and tested this using fluorescence emission scans in the oxy and deoxy condition. Surprisingly, we see that the oxygenated condition is higher than the deoxy condition, indicative of the quaternary structure and the tertiary structure occupying a new state (Figure 3.4A). This suggests that although two substitutions can create an allosteric system, in hemoglobin, more substitutions are required for allosteric regulation to occur in the correct direction.

What substitutions help to invert allosteric response in the correct direction and does this alter the quaternary conformational dynamics? We know that allostery requires an inversion of the oxygen affinity as a result of interactions between the substitutions (Figure 3.2E). The triple substitution on the background of q40W does cause the allosteric effect of 142S to be in the correct direction of allostery. We measured the double mutant effect on conformational toggling by measuring the fluorescence emission of the protein in the oxy and deoxy condition. Indeed, we see that the deoxy fluorescence is higher than the oxy fluorescence, consistent with the directionality of the conformations that has been seen in the literature (Figure 3.4B).

We were interested in understanding if the double mutants were sufficient to change the effect of 142S on the quaternary structure's conformational toggling. We introduced the 142S on the background of the double mutant and measure the fluorescence emission of the protein. We see that the deoxy condition of the protein is higher now than that of the oxy condition and that this protein is now decrease oxygen affinity in the presence of IHP (Figure 3.2A & 4B, Figure A2.5B & C). The ability to decrease oxygen affinity from 142S arises because of substitution interaction that alter the conformational states of the protein.

Indeed, we see the same pattern with the other allosterically single mutant, 149H, which on its own decrease's oxygen affinity in the presence of IHP and populates an alternative conformation when measured using fluorimetry (Figure 3.4B, Fig. A2.5A). Introducing the double mutant of 85K and 146R also creates an allosteric protein that occupies the WT conformations (Figure 3.4B, Fig. A2.5D).

Together, these results show us a that the method for allosteric evolution arises from epistatic interactions due to changes in conformational occupancy in the oxy and deoxy conditions. 142S or 149H can bind to IHP and is allosteric but increases oxygen affinity when bound because

these proteins are in an alternative conformation that stabilizes high-oxygen affinity binding upon IHP binding (Figure 3.2C & 4B). Introducing 85K and 146R then helps to stabilize the WT conformations while retaining the ability to allosterically regulate oxygen binding affinity of Hb. The evolution of allostery within hemoglobin is genetically simple, a single substitution that causes tetramerization allows for conformational toggling between oxy and deoxy states to occur, another substitution then introducing ligand binding and allows for preferential association for one of the functionally distinct conformations. But this conformation is inversed and causes an increase in oxygen affinity upon ligand binding. Epistasis between the substitution and the two others, is what then allows for IHP binding to decrease oxygen affinity because these substitutions rearrange the conformational stability putting the Hb system back into one where allosteric regulation decreases oxygen affinity is the preferred bound state after ligand binding.

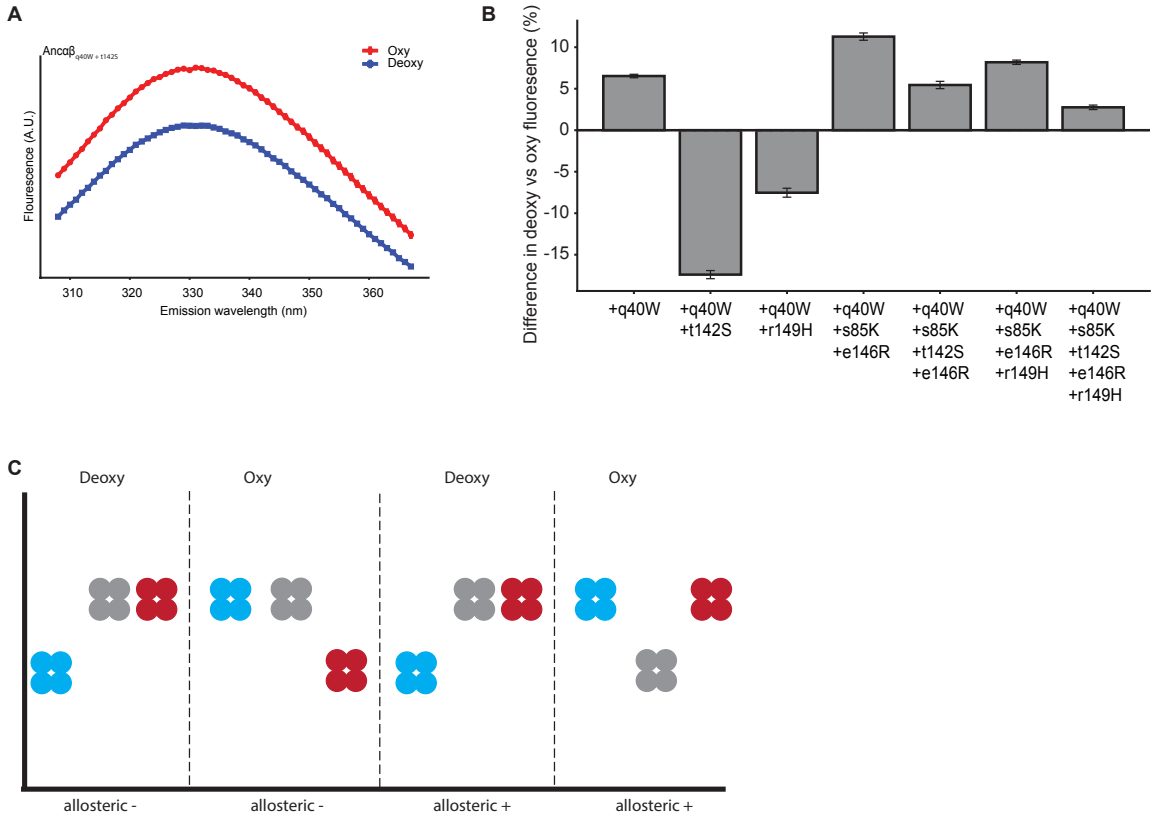

Figure 3.4. Mechanisms for the emergence of allosteric regulation. (A) Fluorescence emissions scans of Ancαβ $_{q40W\ t142S}$ when excited at 280 nm. In red, protein is oxygenated. In blue, the protein is deoxygenated. Error bars represent standard error of measurement, n = 10. **(B)** Bar graphs of percent difference of deoxygenated vs oxygenated conditions of ancestral Hb variants. Error bars are standard error of measurements, n = 10. **(C)** Simple conformational model explaining the evolution of allosteric regulation in Hb. Blue represent deoxygenated condititon, red is oxygenated condititon, grey are alternative conformations that have seen via substitutions that occur at the central cavity.

Discussion

3.9 Simple genetic mechanisms lead to the emergence of allosteric regulation

Our results show that just 5 substitutions led to the emergence of allosteric regulation on Hb's oxygen affinity. The simplicity of the genetic mechanism agrees with prevailing research within the allosteric field, where rational design uses latent properties, deep mutational scans can inactivate and residue allosteric function through single mutations, and historical studies can invert allosteric regulation through single changes (48, 142, 143). Instead, we found that focusing on a few key structural positions within the Hb architecture and exploiting ancient features of the globin fold could immediately confer allosteric regulation (127).

3.10 Alpha subunit as a case of useless complexity in hemoglobin

All the substitutions required for acquisition of allosteric regulation in Hb occurred on the branch leading to Ancβ, showing that the alpha gene was not necessary to acquire allostery. These results contrast much of the Hb allostery literature, which has postulated that the deletion on the alpha subunit within the structure are critical for the mechanisms of allosteric regulation (144, 145). Indeed, the alpha gene maybe critical in the present-day for allosteric regulation, but we have shown that alpha, and the asymmetry that it gives to the tetrameric structure, are not necessary for the acquisition of allosteric regulation.

If oxygen binding of the globin fold is an ancient feature of the protein, and all the features required for regulation of oxygen affinity via allosteric regulation were caused by the changes

along the beta lineage, then what was the initial purpose of the alpha subunit for HB's function? We believe that the specificity contributed by the alpha subunit into the tetrameric structure may be an example of useless complexity (23). We have shown that specificity for heteromerization can occur through a single residue deletion that occurs at IF1 along the branch leading to AncA. It is reasonable to assume then, that the heteromeric structure may arise through a simple change and that the heteromeric structure did not do anything different from the homomeric structure. Consistent with this, Ancα + Ancαβ_{14+CC}, a heteromeric structure that includes alpha is also allosterically regulated by organic phosphate effectors (Figure 3.1G).

If the homotetramer can do everything, then why does the heteromer persists? One explanation could be that the heteromeric structure becomes required for allosteric function at some point in history, which would mean that the homotetramer could no longer confer regulation of function. Indeed, we see that allosteric regulation within the beta subunit is lost by the time of the derived protein Ancβ suggesting that coevolution entrenched alpha to become necessary for function at some point along this historical interval (Fig. A2.6). More work will be required to understand what substitutions caused the necessity of the heterotetrametric structure, and whether the change occurred immediately through a small number of substitution effects or if it was a gradual degradation of beta's homomeric allosteric function.

3.10 Epistatic interactions on conformational states lead to the evolution of an allosteric protein

Our work is a direct historical example of epistasis that arises from changes in relative conformational stabilities. The substitutions q40W and t142S were sufficient to confer allosteric regulation that increased oxygen affinity in the presence of IHP and the substitutions s85K and e146R invert the regulation, decreasing oxygen affinity in the presence of effector. This epistasis

seems to be occurring through changes in conformational stability, although we cannot rule out that epistatic interaction with ligand affinity could also play a role in the strength of allosteric regulation.

These interaction between substitutions in Hb affecting conformational stability agrees with many studies with extant proteins. Single mutations have been shown to create new conformations in non-allosteric protein (146-148). In allosteric proteins, mutations can alter allosteric function by changing the conformational landscapes of the proteins (149). Substitutions that conferred allosteric regulation did alter conformational stabilities but also altered the structural connection between tertiary function and quaternary structure, implying that a mutation can have multiple effects on the conformational stabilities within the proteins.

The genetic pathways lead to allostery then can be many but may also arise because of unexpected interactions between sets of mutations. Epistasis has been shown to facilitate the increase in functions, and to help facilitate new structural and functional landscapes (15, 17, 150). Here we show that allostery is too facilitated by epistasis.

3.11 Ancient spandrel of the globin fold are coopted for allosteric evolution

The evolution of hemoglobin's allosteric regulation was made possible by the pre-existing conformational flexibility of the globin fold, particularly the movement of the FG corner. This movement, which occurs upon oxygen binding and unbinding, has been widely observed in globins and originates from structural shifts in the heme group (140, 141). Photodissociation studies in myoglobin demonstrate that heme displacement upon oxygen release drives helix movement, suggesting that this flexibility is an intrinsic feature of the fold (139). Rather than evolving de novo, hemoglobin coopted this pre-existing structural property, allowing the tetramer to adopt multiple conformational states and enabling allosteric regulation.

Cooption of helix movement seems to be a general strategy of other globins to confer allostery as well. For example, agnathan hemoglobin form allosteric multimers through a distinct mechanism, where oxygen binding induces heme movements across the E-helix (151). In these proteins, we see that multimerization occurs across the E-helix, stabilizing cooperative oxygen binding. Similarly, many non-vertebrate globins utilize helix movements triggered by oxygen binding to enable allosteric control, suggesting that leveraging intrinsic flexibility for regulation may be a widespread evolutionary principle (152-154).

Hemoglobin is not unique; nearly all proteins exhibit some degree of conformational flexibility, and many function as multimers (155). Just as hemoglobin evolved allosteric regulation by co-opting pre-existing structural movements, other proteins may have followed similar evolutionary trajectories. Whether at localized residues or entire structure, structural motions can serve as evolutionary footholds for the emergence of allosteric regulation, allowing proteins to evolve complex regulatory mechanisms from inherent flexibility.

3.12 Material and Methods

Sequence data, alignment phylogeny, and ancestral sequence reconstruction. The reconstructed ancestral sequences used here are the same as those reported previously (69). Briefly, 177 amino acid sequences of hemoglobin and related paralogs were collected and aligned. The maximum likelihood (ML) phylogenetic tree was inferred using the AIC best-fit model, LG+G+F (156). The phylogeny was rooted using as outgroups neuroglobin and globin X, which are found in both deuterostomes and protostomes and diverged prior to the gene duplications that produced vertebrate myoglobin and the hemoglobin subunits. Ancestral sequence reconstruction was performed using the empirical Bayes method, given the alignment, ML phylogeny, ML branch lengths, and ML model parameters (11). Reconstructed ancestors that were used in this study

have been deposited previously in GenBank (IDs MT079112, MT079113, MT079114, MT079115).

For the set of historical mutations *Central Cavity (CC)*, all sites at central cavity that were substituted between AncAB and AncB are changed to the derived state found in AncB; the mutations introduced are s85K, t142S, e146R, and r149H. The substitution q40W, refers to a single substitution that occurs on the branch leading to AncB that is sufficient to confer tetramerization on the ancestral background AncAB. *Q40W+CC* is the union of the single substitution q40W and the set *Central Cavity.*

Recombinant protein expression

Coding sequences for reconstructed ancestral proteins were optimized for expression in Escherichia coli using IDT Codon Optimization and synthesized de novo as gBlocks (IDT). Coding sequences were cloned by Gibson assembly into vector pLIC under control of a T7 polymerase promoter (118). For co- expression of Ancα+Ancβ, a polycistronic operon was constructed under control of a T7 promoter and separated by a spacer containing a stop codon and ribosome binding site, as described in (119).

JM109 (DE3) Esherichia coli cells (New England Biolabs) were heat-shock transformed and plated onto Luria broth (LB) containing 50 ug/mL carbenicillin. For the starter culture, a single colony was inoculated into 50 mL of LB with 1:1000 dilution of working- stock carbenicillin and grown overnight. 5 mL of the starter culture were inoculated into a larger 500-mL terrific broth (TB) mixture containing the appropriate antibiotic concentration. Cells were grown at 37° C and shaken at 225 rpm in an incubator until they reached an optical density at 600 nm of 0.6-0.8. For expression of single globin proteins, 100 uM of isopropyl-β-D-1- thiogalactopyranoside (IPTG) and 25 mg/500 mL of hemin were added to each culture. Expression of single proteins in

culture were done overnight at 22° C. Cells were collected by centrifugation at 4,000g and stored at -80° C until protein purification. Coexpressed proteins were induced using 500 mM IPTG expression with 25 mg/500 mL hemin for 4 hours at 37°C. Cells were collected by centrifugation at 4,000g, immediately followed by purification.

Human hemoglobin was bought commercially (Sigma-Aldridge) and resuspended in PBS.

Protein purification by ion exchange

Purification of singly expressed ancestral hemoglobin proteins were done as previously described (65). All singly expressed proteins (all ancestral globins except Ancα+Ancβ) were purified using ion exchange chromatography, with variation in buffer pH. All buffers were vacuum filtered through a 0.2 μM PFTE membrane (Omnipore). After expression, cells were resuspended in 30 mL of 50 mM Tris-Base (pH 6.88). The resuspended cells were placed in a 10 mL falcon tube and lysed using a FB505 sonicator (1s on/off for three cycles, each 1 minute). The lysate was saturated with CO, transferred to a 30 mL round bottom tube, and centrifuged at 20,000g for 60 minutes to separate supernatant from non-soluble cell debris. The supernatant was collected and syringe-filtered using HPX Millex Durapore filters (Millipore) to further remove debris. A HiTrap SP cation exchange (GE) column was attached to an FPLC system (Biorad) and equilibrated in 50 mM Tris-Base (pH 6.88). The lysate was passed over the column. 50 mL of 50 mM Trise-Base (pH 6.88) was run through the SP column to remove weakly bound non-target soluble products. Elution of bound ancestral Hbs was performed with 100-mL gradient of 50mM Tris-Base 1 M NaCl (pH 6.88) buffer which was run through the column from 0% to 100%. For the construct $AncAB_{q40WCC}$, all buffer pH for purification were 6.47. 2 mL fractions were captured during the gradient process, all fractions containing red eluant were put into an Amicon ultra-15 tube and concentrated by centrifugation at 4,000g to a final volume of 1

mL. For additional purification, concentrated sample was injected into a HiPrep 16/60 Sephacryl S-100 HR size exclusion chromatography (SEC) column. The column was equilibrated in phosphate buffered saline (PBS) at pH 7.4. Purified ancestral globins elute at different volumes depending on the protein's complex stoichiometry: 48-52 for tetramers, 56-60 for dimers, and 65-67 for monomers. The purified proteins were concentrated as mentioned above and then flash frozen with liquid nitrogen.

Protein purification by zinc affinity chromatography. Coexpressed proteins Ancα + Ancβ were purified using zinc-affinity chromatography, which was performed using a HisTrap metal affinity column (GE) on a Biorad NGC Quest. Nickle ions were stripped from the column (buffer 100 mM EDTA, 100 mM NaCl, 20 mM TRIS, pH 8.0), followed by five column volumes of water. To attach zinc to the column, 0.1 M ZnSO4 was passed over until conductance was stable, approximately 5 column volumes, followed by five column volumes of water. After expression, cells were resuspended in a 50 mL lysis buffer (20 mM Tris, 150 mM Nacl, 10% glycerol (v/v), 1mM BME, 0.05% Tween-20, and 1 Roche Protease EDTA-free inhibitor tablet, pH 7.40), sonicated as described above, and the lysate passed through the prepared column. To remove non-specifically bound protein, the column was washed with 50 mL of lysis buffer. Bound protein was then eluted across a gradient of imidazole concentrations (0 to 500 mM) in a total of 100 mL elution buffer (20 mM Tris, 150 mM NaCl, 500 mM imidazole, 10% glycerol, and 1 mM BME, pH 7.4). 1 mL fractions were collected. The fraction corresponding to the second peak of UV absorbance at 280 nm has a visible red color and was collected and concentrated as described above. The concentrated solution was injected into a Biorad ENrich 650 10 x 300 columns for additional purification and eluted in PBS buffer.

Oxygen affinity and allostery

Purified proteins were deoxygenated using 10 mg/mL sodium dithionite and immediately desalted via a PD-10 column (GE Healthcare) equilibrated with 25 mL of 10 mM HEPES, 0.5 mM EDTA (pH 7.4). Eluted proteins were concentrated using Amicon Ultra-4 centrifugal filters (Millipore).

Equilibrium oxygen-binding assays were conducted at 25°C using a Blood Oxygen Binding System (Loligo Systems) with 0.1 mM protein (heme concentration) dialyzed in 100 mM HEPES, 0.5 mM EDTA buffer. Protein solutions were sequentially equilibrated at 3–5 oxygen tensions (PO_2), achieving 30–70% saturation, while continuously monitoring absorbance at 430 nm (deoxy peak) and 421 nm (oxy/deoxy isosbestic point). Fractional saturation was plotted against PO_2, and the Hill equation was fitted to each dataset using OriginPro 2016 to estimate P_{50} (PO_2 at half-saturation) and the Hill coefficient (n_{50}, slope at half-saturation). Confidence intervals (95%) were calculated by multiplying the standard error of the mean from replicate experiments by 1.96.

To assess potential allosteric regulation by organic phosphate effectors, assays were conducted under three conditions: without effectors (stripped), with 0.5 mM IHP, and with 0.5 mM ATP. Most experiments were performed with IHP due to its stronger allosteric effect compared to ATP and their qualitatively similar modulation of Hb-O_2 affinity (63, 133). Although IHP may not have been the physiological effector in ancestral organisms, it has been shown to allosterically regulate hemoglobins across major vertebrate lineages, whereas effectors such as 2,3-bisphosphoglycerate (BPG), ATP, and GTP exhibit lineage-specific effects. Thus, IHP serves as a general polyanion for assessing the allosteric capacity of ancestral Hb, a well-established approach in hemoglobin studies (51, 69).

Fluorescence emission scan

Purified proteins were deoxygenated with 10 μM sodium dithionite to remove oxidized hemoglobin, then desalted using a PD-10 column (Cytiva) equilibrated with PBS. If necessary, eluted proteins were concentrated using Amicon Ultra-15 filters prewashed with PBS.

To evaluate conformational states as a function of oxygenation, proteins were aliquoted into two 150 μL samples at 160 μM concentration. Buffer-only aliquots were prepared for fluorescence background correction. Oxygenated hemoglobin samples were verified by UV-Vis spectroscopy, confirming the characteristic double peak at 480 nm. Samples were transferred into a sealed sub-micro quart fuorometer cell (Starna Cells, Inc.). Fluorescence scans were performed using a Fluorolog-3 (Horiba) with excitation at 280 nm and emission recorded from 307–500 nm in 1 nm increments. Slit widths for excitation and emission ranged from 4–7 nm. To mitigate inner-filter effects of hemoglobin, a front-facing mirror setup was used, an established approach for Hb fluorescence studies (157).

For deoxygenated samples, an aliquot was incubated in a custom anaerobic chamber saturated with nitrogen for two hours, with periodic resuspension to facilitate oxygen exchange. At the time of measurement, the sample was transferred to the same sealed sub-micro quartz fluorometer cell used in oxygenation condition. Fluorescence scans were conducted immediately after chamber removal.

Isothermal Calorimetry.

Purified ancestral proteins were dialyzed overnight in ITC buffer (50 mM MES, 100 mM NaCl, pH 6.5). Protein concentrations were quantified post-dialysis and, if necessary, adjusted to 200 μM using Amicon Ultra-15 filters. Ligand buffer was prepared by supplementing dialysate with 1.5 mM IHP. All buffers and protein samples were degassed before ITC measurements.

ITC experiments were conducted using a MicroCal ITC 200 (GE). The sample cell and syringe were pre-cleaned with 10% Contrad overnight, followed by five 3-minute washes with methanol and deionized water, with air drying after the final rinse. Before loading, the sample cell was equilibrated with 300 µL degassed ITC buffer for five minutes and emptied using a 500 µL glass Hamilton syringe. The cell was then loaded with 250–300 µL of protein, ensuring the removal of air bubbles. The syringe was filled with 100 µL of ligand buffer.

Experiments were performed with a reference power of 10 µcal/sec, using 20 injections: an initial 0.4 µL injection followed by 19 injections of 2.0 µL, each spaced 150 seconds apart to allow baseline stabilization. The syringe was rotated at 750 RPM. Data were analyzed using Origin 7.0 (OriginLab).

Chapter 4

Rampant epistasis during the evolution of an allosteric transcription factor reveals shifting genetic basis

4.1 Abstract

Allostery requires multiple residues to work compatibly, enabling effector binding at one site to alter function at a distant site. Yet, the genetic mechanisms that underlie and relay allosteric regulation across protein families remain poorly understood. One way to dissect these mechanisms is by studying historical sequence drift: as allosteric proteins evolve, the residue changes they accumulate can reveal when and how new functional constraints arise. We experimentally reconstructed sequence evolution in the tetracycline repressor family by introducing individual historical substitutions across the evolutionary interval separating the allosteric protein TetR(B) and its closest paralog, TetR(D). We found extensive epistasis: 40% of the 107 historical substitutions in this interval disrupted allostery when introduced individually into at least one historical background. Among three tested historical intermediates, none shared a common set of inactivating mutations, indicating continuous turnover in the mutational effects across evolutionary time. These epistatic interactions are structurally diffuse and can span large distances—over 50 Å—within the protein structure. Further, we observed similar turnover across the broader Proteobacteria TetR clade, suggesting that epistasis is a rampant feature shaping allosteric evolution. Our results show that the shifting genetic architecture of allostery—driven by long-range, evolving epistatic interactions—helps explain why the genetic basis of allosteric regulation is difficult to describe across protein families.

4.2 Introduction

Allostery, in which regulatory binding at a distal site modulates protein function and occurs through changes in the structural conformation in the proteins, or by adjusting residue interactions throughout the protein (43, 129, 131, 158-160). Sequence drift, where gradual sequence change occurs as the protein maintain their existing functions, explains how genetic diversity arises during history (161). How the physical and genetic architecture constrains sequence drift of an allosteric protein is not known.

The evolution of allostery of particular interest because the ability to regulate function through distal binding would seem to require, in some cases, most of a protein to work (43, 62, 126, 129, 131, 142, 143, 149, 155, 158, 162, 163). Significant progress in understanding the evolution of allosteric came by leveraging conservation in large protein alignments (164). In this, conserved pairwise interactions within allosteric proteins were uncovered, where mutations of these residues cause inactivation of allosteric regulation which can be rescued by mutations at nearby sites (43, 131, 158, 163). Yet other studies have shown that conservation need not exist within the alignment for residues to be important for allosteric regulation (165). So, exactly which residues define the evolutionary paths that allosteric proteins remain unknown.

But what if the residues and positions that define allostery within a protein change throughout history? Evidence suggests that this is true for even the most mutationally constrained residues in allosteric proteins. A recent study performed a saturating mutagenesis on the allosteric transcriptional repressor E. coli tetracycline repressor-B (E.c. TetR(B)) and found allosteric inactivating residues span the whole protein (165). Positions within TetR(B) were highly-allosterically constrained residues - where any residue mutation will inactivate allosteric

regulation. These high constrained residues were not conserved in large sequence alignments, suggesting that epistasis has alter the functional effect of residues and allowed for substitutions to replace those sites. Yet, how often epistasis changes the allosteric effect of residues, whether previously inactivating residues are substituted, and the mechanisms by which changes in function occurred remain unknown.

Here, we use TetR and ancestral sequence reconstruction to reconstruct the history of the TetR family and study evolutionary drift of allosteric regulation. The tetracycline repressor protein (TetR) family provides an excellent model system for studying the impact of epistasis on allostery. TetR's primary function is to repress gene activity by binding to a palindromic sequence upstream of a gene; upon binding to its effector, tetracycline, TetR allosterically unbinds, allowing transcription (71). Ancestral sequence reconstruct can leverage the natural history of the protein to establish directionality in sequence changes that occurred throughout history, allowing us to study how sequence drift affects mutational paths that could be taken during evolution (3, 4). By leveraging the previously published mutagenesis dataset, phylogenetic analysis, and high-throughput experiments, we aim to determine how the sequence architecture constrains drift in an allosteric protein.

4.3 Historical trajectory of TetR protein

To characterize sequence drift of allosteric protein, we inferred the maximum likelihood phylogeny of TetR(B) and its paralog, TetR(D), within enterobacteriales, rooting with betaproteobacteria as an outgroup (Figure 4.1A). We used ancestral sequence reconstruction to infer the last common ancestor of E.c. TetR(B) and E.c. TetR(D) proteins (AncBD), along with the last common ancestors of the TetR(B) (AncB) and TetR(D) (AncD) clades respectively. We

identified substitutions as differences between the most probably reconstructions between ancestral and descendent nodes.

Along this entire trajectory, there have been a total of 107 substitutions at 40 sites across the entire protein. Most of the sites are high confidence, and every ancestral protein has a mean posterior probability across all sites of >0.89 (Fig. A3.1). Notably, the aLRT score of AncBD is low, which suggest that an alternative topology is also likely but will most likely not affect the overall conclusions (Fig. A3.2).

4.4 Ancestral TetR proteins have retained allosteric ability despite containing known deactivating amino acids

To understand whether known allosterically inactivating amino acids have occurred during history, we compared the amino acids of AncBD, AncB, AncD, and Ec-TetR(D) with the previous deep mutational scan of Ec-TetR(B). Across all proteins of interest, we identified up to 13 mutations that, when introduced individually into Ec-TetR(B), disrupt the protein's allosteric response to tetracycline (Fig. 4.1B) (165).

We were interested in whether AncBD, AncB, AncD, and TetR(D) were able to respond allosterically to tetracycline despite containing allosteric inactivating states. To do so, we created isogenic E. coli reporter strains that allow us to measure the allosteric ability of TetR using a plate reader (74). Each strain contains two plasmids: one expresses the TetR protein of interest, and the other contains GFP under the control of a promoter the contains a putative TetR binding elements. When the protein is expressed, GFP fluorescence is repressed presumably because the TetR protein is bound to binding element and preventing transcription. The addition of tetracycline to solution will cause depression of TetR and higher expression of GFP. The

allosteric response, then, the change in fluorescence in the absence and presence of ligand (Figure 4.1C).

We measured the allosteric response of AncBD, AncB, AncD, E.c TetR(D), and E.c. TetR(B). All proteins showed low normalized GFP fluorescence in the absence of tetracycline, indicating repression of the reporter gene. In the presence of tetracycline, all proteins increased their GFP fluorescence indicating de-repression of upstream element binding (Figure 4.1D). These results show that all proteins, even if they contain known inactivating residues, are allosteric.

We quantify the strength of binding as the relative change in GFP fluorescence in Tet (-) and Tet (+) conditions. E.c. TetR(B) has the strongest allosteric response, with a fold change of ~70. The lowest allosteric response is E.c. TetR(D) with a fold difference of 10. All ancestral proteins have an allosteric response between these two extant proteins: AncBD is 15, AncB is 35, and AncD is 20-fold (Figure 4.1D). Proteins that contain allosteric inactivating residues have similar strength of binding.

Together, AncBD, AncB, AncD, and E.c. TetR(D) are allosteric, despite containing known allosterically inactivating resides. These results show that the epistatic effects are constant throughout the phylogeny.

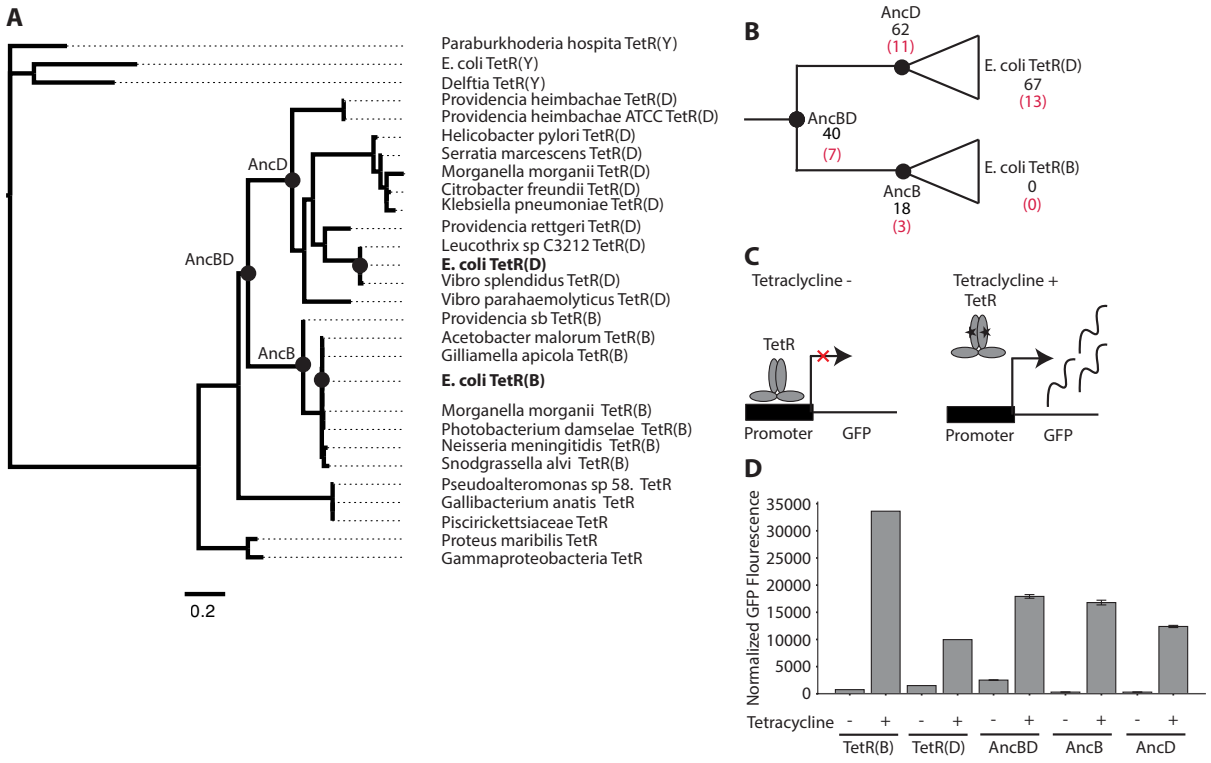

Figure 4.1 The evolutionary history of TetR(B) and TetR(D). (A) Phylogeny of TetR(B) and TetR(D) family. Circles are reconstructed ancestral proteins. Scale bar, substitutions per site. **(B)** simplified phylogeny of E. coli TetR(B), E. coli TetR(D), the ancestors of each clade, and the last common ancestor of all TetR(B) and TetR(D) proteins. Black numbers, pairwise difference in amino acid composition relative to E. coli TetR(B). Red numbers are the number of amino acids within each protein that were identified as allosteric breaking states in the E. coli TetR(B) DMS. **(C)** The molecular mechanism of TetR repression in fluorescence experiments. *Left:* TetR bound to promoter repression transcription of GFP reporter in the absence of allosteric effect, tetracycline. Ovals represent protein, black bar is promoter and gene represented by line. *Right:* TetR binds to tetracycline, derepressing and allowing for transcription of GFP. Stars represent tetracycline and squiggly line is GFP transcripts. Ovals, black bar, and line is same as mentioned above. **(D)** Fluorescence of extant TetR and ancestors in the presence and absence of tetracycline. Error bars represent standard error of mean, n = 3.

4.5 Continuous turnover in allosteric function during history

While it is clear that epistasis has shaped the effects of allosterically inactivating substitutions over time, the extent to which turnover in allosteric function occurred during the evolution of TetR sequences remains unknown.

One way to address this question is to study sequence drift by introducing historical substitutions on different TetR backgrounds. This approach leverages the fact that each substitution must have been tolerated the time that they occurred, and epistasis must have changed the effects of residues that are deleterious. We generated libraries of AncBD and AncB variants, each containing one of 107 individual substitutions that occurred across two trajectories: AncBD to E.c. TetR(B) and AncBD to E.c TetR(D) (Figure 4.2A). We also leverage the previous dataset on E.c. TetR(B), where all 107 variants have already been quantified (165). Because each ancestor harbors have different amino acid sequences, the resulting libraries include unique subsets of variants, which consistent of substitutions that happened on a subsequence branch or reversions to ancestral amino acid states (Figure 4.2A).

We used fluorescence-activated cell sorting (FACS) coupled with deep sequencing to evaluate the effects of each variant on allostery by analyzing the inactive populations under tetracycline-bound (Tet+) and tetracycline-free (Tet–) conditions (Fig. 4.2B). Both libraries achieved full coverage, with highly reproducible results across two replicates ($R^2 > 75\%$ for both conditions; Fig. 4.2B & C). To quantify allosteric inactivation for each single mutation, we calculated an enrichment score by comparing the ratio of reads in the inactive bin of the Tet(+) sort to the inactive bin of the Tet(–) sort, normalized to the ancestral wild-type protein. A positive enrichment score indicates a substitution that disrupts allostery, and we defined allosterically inactivating residues as those with enrichment scores deviating by at least one standard deviation from the wild type, representing a significant shift toward allosteric inactivation (Fig. 4.2D). What fraction of the substitutions that occurred at some point break allosteric regulation when introduced into the backgrounds of AncBD, AncB, and E. coli TetR(B)? For AncBD, we find that 20 of the 107 total variants within the library are allosterically inactive (Fig 4.2E). For

AncB, only 11 of the 107 total variants within the library are allosterically inactive (Fig 4.2F). For the previously published E. coli TetR(B) dataset, we find that 22 of the 107 total variants are deleterious to allostery (Table A3.1). When we compare all deleterious variants across all backgrounds, we find that over 45% of amino acid states were allosterically inactivating at some point (Figure 4.2G). In contrast, only 1 mutant has been deleterious for DNA binding during this historical interval, or 0.009% of the entire library. Given our results, albeit a small sample size of all possible mutants, the likelihood of an epistatic interaction occurring on allosteric function is x5000 greater than epistatic interactions that occur on DNA binding during history.

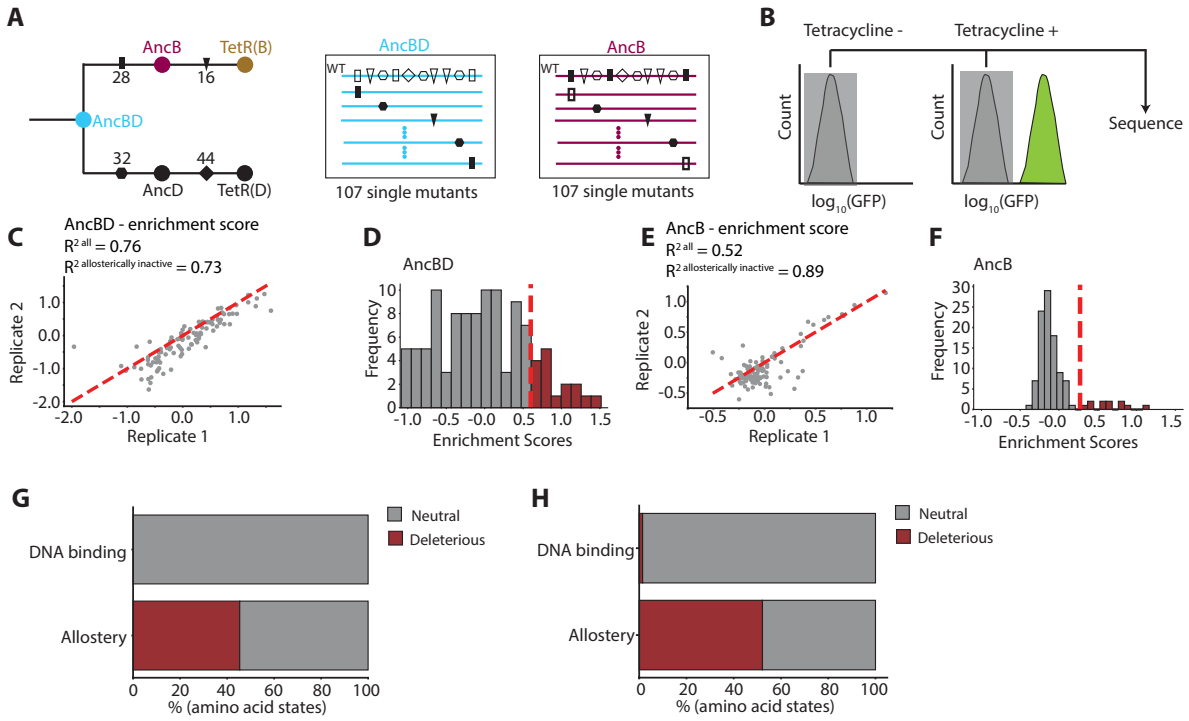

Figure 4.2. Single-mutant library of substitutions across two backgrounds. (A) *Left:* Simplified phylogeny of TetR phylogeny. Substitutions are numbers on branches that represent the number of amino acid residue changes that occur between ancestor and descendent proteins. Icons: square represent substitutions that occurred between AncBD and AncB, triangle were substitutions that occurred between AncB and E.c. TetR(B), hexagon were substitutions between AncBD and AncD, and diamond were substitutions between AncD and E.c. TetR(D). *Right:* For AncBD and AncB, a library containing all possible substitutions that occurred across this historical interval were created using *denovo* syntethsized genes. **(B)** Hypothetical distribution of TetR protein libraries in the presence and absence of tetraclycline. The low-GFP expressing bins

were sequence in both conditions to estimate enrichment relative to WT-protein. **(C)** Scatter plot showing the correlation between enrichment values across two biological replicates in AncBD. Grey points represent single mutant variants. Red dashed line is the x=y intersect. **(D)** Distribution of the average enrichment score for AncBD. Red dashed line is one standard deviation away from mean. Red bars are variants that are one standard deviation away from the mean in the positive direction, which are classified as allosterically inactivated protein variants. **(E)** Scatter plot showing correlation between enrichment values across two biological replicates in AncB. Grey points and red dashed line are same as mentioned above. **(F)** Distribution of average enrichment scores for AncB. Dashed line and red bars are same as mentioned above. **(G)** Stacked bar plot of percent amino acid states that are deleterious for DNA binding or allostery across the historical interval from AncBD to E.c. TetR(B). **(H)** Stacked bar plot of percent sites that are deleterious for DNA binding or allostery across the historical interval from AncBD to E.c. TetR(B).

4.6 Epistatic interactions shaped the history of TetR allostery

Does each library contain its own set of allosterically inactivating residues? Are residues whose functions have changed being biased to certain evolutionary outcomes or are they nearly equal in their direction? We compared shared allosteric inactivating variants in each library to see how these residues are conserved across multiple proteins. Across all three libraries, there are no shared allosterically inactivating mutations (Fig 4.3A). When comparing between pairs of libraries, there are some shared allosteric inactivating residues. Between AncBD and AncB, the number of shared were 2 and between AncBD and E. coli TetR(B) there was one variant. There were no shared vairants between AncB and E. coli TetR(B). Together, these results show that epistasis has changed most mutational effects of allosteric inactivating residues across each branch.

Epistatic interactions can alter a mutation's effect on allostery, making it either permitted or restricted. A permitted residue is one that was ancestrally deleterious but became functionally tolerated because of a substitution (Fig. 3B). Conversely, a restricted residue was ancestrally neutral for allosteric function but became deleterious following a substitution (Fig. 4.3C). Mapping the quantity of each category on each branch will reveal how many permitting or

restricting mutational pathways changed in the historical interval between AncBD and E.c. TetR(B).

We identified all permitted and restricted residues on the branches from AncBD to E.c. TetR(B). For the entire historical trajectory, we found 22 permitted residues and 29 restricted residues. On the branch between AncBD and AncB, there were 13 permitted residues and 8 restricted residues (Figure 4.3D). On the subsequent branch, between AncB and E.c. TetR(B), we found 9 permitted and 21 restricted residues. Together, these results show that quantity of permitted and restricted evolutionary paths is relatively equal across the entire trajectory between AncBD and E.c. TetR(B), but that they fluctuate in their direction on per-branch time scales.

What kinds of temporal relationships exist between epistatic changes in residue function and the timing of substitutions during evolution? These relationships can take several forms. In some cases, the relationship is immediate, where the epistatic change and the substitution occur on the same evolutionary branch. In other cases, the relationship is subsequent, where the epistatic change occurs on an ancestral branch prior to the substitution. Each of these scenarios provides different insights into how directly epistasis has influenced allosteric function in TetR proteins. We find evidence for both types of relationships. For immediate interactions between epistatic changes and substitutions affecting allosteric regulation, we identified six substitutions along the branch from AncBD to AncB—three of which were permitted and three restricted (Figure 4.3D). Along the subsequent branch from AncB to E. coli TetR(B), we observed one permitted substitution. Additionally, we found an example of a subsequent relationship in which a residue that was permissive on the AncBD-to-AncB branch later became a substitution on the following branch. These findings demonstrate that epistatic changes influencing allosteric function have directly shaped the trajectory of substitutions during TetR evolution.

In total, we identified 51 instances in which the effect of a residue on allosteric function changed across the evolutionary interval between AncBD and E. coli TetR(B). These changes include both permissive and restrictive epistatic interactions, with a slightly higher number of restricted paths. Notably, these interactions directly influenced the evolutionary trajectory of TetR allostery: for 8 substitutions, the allosteric effect of a residue changed either prior to or at the point of substitution, enabling the mutation or preventing reversion to its ancestral state. Together, these results demonstrate that epistatic interactions have shaped the mutational paths available to TetR allostery throughout its evolutionary history.

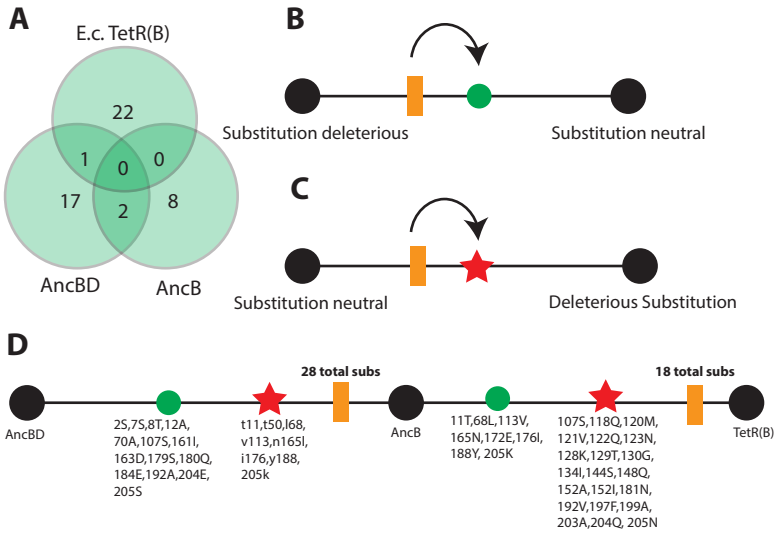

Figure 4.3. Permitted and restricted residues change across branches. (A) Venn diagram of allosterically inactivating residues across each TetR background studied. Numbers in overlapping portions of each circle represent shared residues that are deleterious for allosteric regulation. **(B)** Substitutions can permit residue changes by making ancestrally deleterious residues functionally neutral. Orange bar are substitutions. Green circle is a residue change whose functional effect were ancestrally deleterious prior to substitution. Black circles represent ancestral and derived proteins in a specified historical interval. **(C)** Substitutions can restrict residue reversions by making reversions to ancestral states allosterically deleterious. Red star represents a restricted reversion. Orange bar and black circle are same as mentioned previously. **(D)** All restricted reversions and permitted residue changes that occurred at each interval between AncBD and E.c. TetR(B). Orange bar, green circle, and red stars are same as mentioned previously.

4.7 Delocalized residue interaction underlying substitutions during TetR evolution

We were interested in understanding the types of epistasis that facilitate the turnover in sequence to allosteric regulation. To do so, we wanted to identify potential epistatic interactions between substitutions that occurred during history and then functionally these relationships.

How can we identify epistatic interactions between substitutions? We hypothesize that permissive substitutions and restricted reversions reflect underlying epistatic relationships. A substitution becomes restricted when its reversion is no longer tolerated, suggesting that other substitutions have become dependent on its presence to maintain allosteric function. Conversely, a substitution is permissive when its initially deleterious effect is tolerated due to compensatory changes elsewhere in the protein. Together, a restricted reversion may alter the functional effect of a deleterious mutation to permit a substitution.

We tested this hypothesis by examining the allosteric effects of three permitted substitutions and three restricted reversions that occurred on the evolutionary branch between AncBD and AncB (Figure 4.4A). To do this, we constructed a combinatorial library containing all possible combinations of these six mutations on the AncBD backgrounds. We picked isogenic variant and introduced them into E. coli, measuring allosteric response with a reporter strain that contains a GFP reporter gene under the control of a tetracycline-responsive promoter. This system allowed us to assess the functional consequences of each mutation by measuring GFP fluorescence in the presence and absence of tetracycline, providing a direct readout of each variant's ability to mediate transcriptional repression in response to effector binding.

Our mutational screen found an interaction between the residue substitution k205N, which permits the substitution e7S. The substitution e7S decreases allosteric activation significantly, near 0 for fold change difference, and k205N also decreases allosteric activation (Figure 4.4B).

The predicted fold change in allosteric regulation would be below 0, resulting in complete allosteric inactivation. However, the double mutant returns allosteric activation back to that of AncBD making the combined effects of these substitutions effectively neutral. The difference in precited and observed effects on allosteric regulation are 33-fold (Figure 4.4B).

We wondered where on the structures these mutations were occurring and if they are physically interacting with each other. Surprisingly, these substitutions span the entire protein: 7S in the DNA-binding domain and 205N in the ligand binding domain (Figure 4.4C). The measured distance between these two residue positions is 53.8 angstroms, which close to the largest possible position difference within TetR, 66 angstroms. Epistatic interactions within an allosteric protein can span nearly the whole structure.

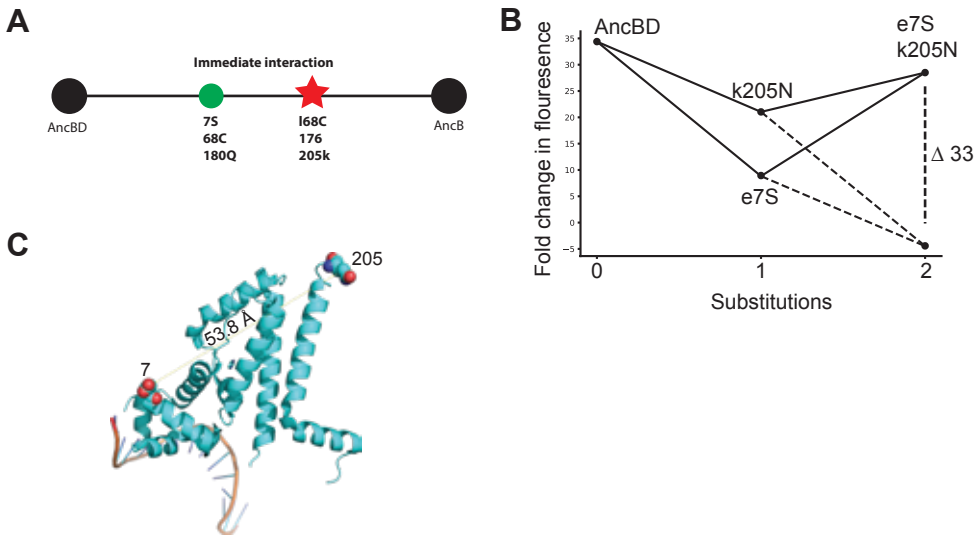

Figure 4.4 Epistatic interactions between historical substitutions in TetR history. (A) Restricted reversions and permitted changes of substitutions that occurred during the historical interval between AncBD and AncB. Residue positions and identity that are permitted substitutions are green circles. Red stars are restricted reversions. **(B)** Effect of no, single, and double substitutions for the substitutions k205N and e7S on the background of AncBD. Fold-change in fluorescence is the allosteric effect. **(C)** Structural position of 205N and 7S. Spheres are the sidechain for each residue, helices are TetR structure. Red atoms are oxygens.

4.8 Rampant epistasis leads to broader samples of allosteric residues across the entire proteobacter clade

To assess whether our findings hold across broader evolutionary timescales, we inferred a maximum likelihood phylogeny of the TetR family, incorporating members spanning the entire Proteobacteria clade, with Cyanobacteria as an outgroup (Fig 4.5A).

To examine how patterns of evolution have shifted across this bacterial phylogeny, we compiled a dataset of all amino acid substitutions that inactivate allostery—either from our single-mutant libraries or a previously published deep mutational scanning (DMS) dataset. We then mapped the occurrence of these residues throughout the phylogeny. Our analysis reveals that allosterically inactivating residues are widespread among TetR sequences. On average, a given TetR sequence contains more than 14 amino acids that inactivate allostery in at least one of our mutant libraries (Fig 4.5B). In contrast, these sequences contain, on average, only two amino acid substitutions that disrupt DNA binding or folding. In total, 200 out of 500 possible allosterically inactivating residues are sampled across this phylogeny, compared to only 10 out of 72 possible residues known to disrupt DNA binding (Fig 4.5C).

Together, these results demonstrate that our findings generalize across the TetR phylogeny. Allosterically inactivating residues are frequently sampled throughout evolutionary history and have likely altered their effects on allostery through epistatic interactions with other residues.

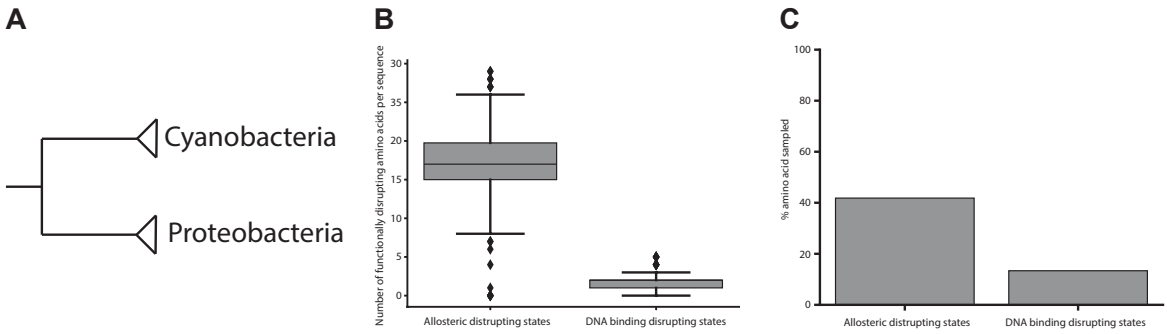

Figure 4.5. Epistatic is rampant across TetR family across proteobacteria. (A) Simplified phylogeny of the entire proteobacteria clade with cyanobacteria as the closest outgroup. **(B)** The average number of function disrupting states across all ancestors across all proteobacteria TetR genes. **(C)** The percent amino acid that appear across proteobacteria whose functions have been described previously.

Discussion

4.9 Genetic mechanisms in allosteric proteins altered via epistatic interactions

Our results show that the underlying genetic effect of residues within an allosteric protein change during history.

In this study, we utilized sequence drift in during evolutionary history to understand when and how particular residues are allowed in an allosteric protein (49). Our results show that even through the accumulate of one residue, the epistasis already states to arise in sites that are not linked structurally, leading to new paths for evolution within allosteric regulation. Our results suggest that underlying residues and positions within allosteric protein will change over time, incorporating different residues into the mechanisms by which allostery occurs.

If the underlying genetic mechanism within allostery changes, then why look at site specific changes during the allosteric transition of the protein? We believe that there is a lot that has been learned using these approaches, like understanding the very nature of the theoretical and physical mechanisms of allosteric transitions (62, 166). Understanding allosteric regulation within a particular protein using genetic changes to map out the allosteric pathways within proteins too has yield deep insights into the mechanisms (41, 43, 52, 71, 167). However, our results suggest that site-specific studies within single proteins may not capture the diversity of genetic mechanisms by which allostery occurs across protein families.

If the genetic and structural basis of allostery can change over evolutionary time, then understanding how this regulation arises and diversifies requires a broader comparative

framework. Rather than focusing solely on individual sites in a single protein, we propose that measuring allosteric regulation across all homologous proteins within a gene family, and correlating variation in function with patterns of sequence divergence, may reveal which residues or structural features are responsible for altering these regulatory mechanisms. This comparative approach can help identify generalizable principles of allostery and describe how constrained its evolutionary pathways are.

4.10 Sensitivity to epistasis arises because of complexity in allosteric systems

Our results show that allostery is sensitive to epistasis. We believe that this arises because of the complexity of allosteric systems. In addition to folding and performing their primary function, proteins must satisfy multiple conditions for allostery to occur: the protein must be able to exists in multiple conformational states that are functionally distinct, bind to an effector, and couple effector binding with shifts in functional conformations (126). Epistatic interaction between residues can thus arise within any of these conditions.

The complexity within allosteric regulation and the apparent changes in accessible paths during sequence drift would make the paths taken more idiosyncratic during history for particular lineages. Because epistatic interactions may arise from or between different interacting parts that make up an allosteric system, there will be a lot of epistasis as mutations accumulate. Between even closely related homologs, the set of mutational paths available for gaining, losing, or modifying allostery could differ substantially. This implies that each allosteric protein may have a distinct sequence space landscape, shaped by its unique history.

4.10 Residue interaction sensitivity can lead to drift through previously inaccessible areas of sequence space

If allostery is sensitive to epistasis, then how did TetR explore larger areas of previously inaccessible sequence space relative to DNA binding? We believe that allostery explored more areas of sequence space because of the larger number of residue positions that are involved in allostery. On average, about 20% of protein residues are involved in allosteric transitions (168). Within TetR, which seems to involve large spans of the protein structure, it maybe be larger because previous studies have shown that so much of the protein can inactivate allostery (165). When a mutation inactivates allosteric regulation mutations across the entire protein can compensate for its effect and rescue allostery (169). In contrast, DNA binding is defined by a smaller number of residue positions, which constrains the number of possible epistatic solution when a deleterious mutation appears at one of these sites (15).

This suggests that the sensitivity of allosteric regulation to epistatic interactions depends on the number of residues involved in the mechanism. As more residues contribute to allostery, the system becomes more likely to have epistatic interactions between residues, simply because there are more opportunities for mutations to disrupt or compensate function. Conversely, if allosteric regulation depends on only a few key residues, we expect it to be less sensitive to epistatic interactions, as fewer mutational combinations are required to maintain or alter function.

4.11 Future directions

Further work will have to be done to understand the quantity and mechanisms of these epistatic interactions within a protein. Here we describe that within the smaller evolutionary time scales, and amongst substitutions that are taken during history, epistatic interactions are rampant. However, it is unclear whether these epistatic shifts are caused by a small number of large effect mutations or if they are an accumulation of smaller effects that occur during history.

Understanding the effect of all mutations during history will let us understand the tempo of epistatic effects across time (18, 161).

Another way to gain insight into the nature of allosteric epistasis is by considering how epistatic relationships are structured by thermodynamic constraints. Allosteric proteins must satisfy multiple thermodynamic requirements: they must fold into stable structures, bind to effectors, and undergo functionally meaningful conformational changes. Epistasis may emerge from the interplay between mutations that affect one or more of these thermodynamic properties. For instance, some mutations might only become accessible after earlier substitutions shift the energetic balance of the system.

It's also important to emphasize that our study samples only a small portion of the total genotype-phenotype landscape. Whether the patterns we observe generalize to higher-order interactions, and how frequent or consequential these interactions are across a protein's evolutionary history, remain open and critical questions. Identifying the extent and structure of these epistatic interactions will be key to understanding how accessible allosteric traits are during evolution.

4.12 Methods

Single-mutant library construction

We built wild-type expression plasmids for AncB and AncBD backgrounds, which were used as template for library construction. We obtained the AncB and AncBD genes by commercial DNA synthesis (Twist Bioscience) and used Gibson assembly to clone each into a low-copy plasmid backbone (pSC101 origin) carrying resistance to spectinomycin. The ancestral TetR genes are driven by the strong constitutive promoter apFAB61. All primers used here and elsewhere were synthesized by IDT.

Library construction for both AncB and AncBD backgrounds was performed as follows. Oligonucleotides encoding each variant were synthesized by Twist Bioscience. Due to DNA synthesis length limitations, oligonucleotides were organized into two sub-libraries: Variants between residues 2-93, and those between residues 107-205. Within each synthesized oligonucleotide, the TetR variant sequence was flanked by 20-25bp constant regions homologous with the destination plasmid.

For each sub-library, a "backbone" fragment was amplified from the wild-type expression plasmid and digested with DpnI (NEB). Variant oligonucleotides were resuspended in dH2O, quantified, and pooled with equimolar ratios. Libraries were assembled using Gibson assembly at 50C for 1 hour using 200ng of total DNA with a 1:2 molar ratio of backbone to insert.

After assembly, the Gibson reaction was dialyzed for 1 hour with water on silica membranes (0.025um pores) before transforming 3uL dialyzed Gibson assembly into 25uL electrocompetent DH10B *E. coli*. After a 1-hour recovery in SOC medium at 37C, cells were spread on LB agar plates with 100ug/mL spectinomycin to assess transformation efficiency and diluted into fresh liquid culture for overnight growth. After confirming that transformation efficiency was sufficiently high, libraries were miniprepped and stored at -20C.

50ng of each library was transformed into 25uL electrocompetent *E. coli* harboring a reporter plasmid. The reporter plasmid was modified from pJ251-Gerc (Addgene.org plasmid #47441), expressing superfolder GFP under control of the TetR-inducible pLtetO promoter on a pColE1 backbone with kanamycin resistance. After a 1-hour recovery in SOC medium at 37C, cells were spread on LB agar plates with 100ug/mL spectinomycin and 50ug/mL kanamycin to assess transformation efficiency and diluted into fresh liquid culture for overnight growth. After confirming that transformation efficiency was sufficiently high, culture was miniprepped (Zymo

Research) and mixed 1:1 with 50% glycerol and stocked at -80C. Deep sequencing was used to confirm that the pre-selection library contained all expected variants and had minimal skew. All subsequent steps involving single-mutant libraries are performed for each sub-library in parallel.

Fluorescence measurements

Multiple colonies were picked from agar plates, inoculated into LB medium containing 100ug/mL spectinomycin and 50ug/mL kanamycin in 96-well plates, and grown overnight at 37C with shaking at 700rpm. These starter cultures were diluted 1:50 into fresh media and continued to grow with shaking at 37C. After 1 hour of growth, a final concentration of 1uM anhydrotetracycline or vehicle (dH2O) was added and allowed to incubate and induce for 3-4 hours before measuring OD-normalized GFP fluorescence in 96-well plates. Fluorescence was measured using a multi-well platereader (HTX Biotek) with excitation and emission wavelengths of 485nm and 528nm. Fluorescence intensity was normalized to OD600nm and background-subtracted (fluorescence of cells carrying no GFP).

Fluorescence-activated cell sorting

In duplicate, about 25uL of single-mutant libraries in DH10B *E. coli* with reporter plasmid was used to inoculate 5mL LB with 100ug/mL spectinomycin and 50ug/mL kanamycin and grown overnight. The following morning, starter cultures were diluted in duplicate 1:50 into fresh media with antibiotic selection and allowed to grow at 37C with shaking (250rpm) for 1 hour before addition of 1uM anhydrotetracycline or vehicle (dH2O). Growth continued until OD600 reached ~0.7 (about 4 hours) before placing on ice.

Chilled cultures were diluted 1:50 in ice-cold phosphate-buffered saline (PBS) and sorted with a Sony MA900 cell sorter (Sony Biotech) to enrich allosterically inactive variants (no fluorescence upon anhydrotetracycline induction). Briefly, uninduced cells were flowed first to ascertain the

level of fluorescence in the absence of induction. Then, the induced culture was flowed, and about 25k events from the bottom 2-3% of the GFP fluorescence distribution were sorted directly into fresh PBS (after gating for live cells and singlets). This sample was re-flowed to confirm sufficient purity, then at least 50k events from the bottom 2-3% of the fluorescence distribution were sorted directly into 1mL of SOC medium. This process was repeated for two independent replicates of each sub-library for AncB and AncBD single-mutant libraries.

SOC media containing sorted populations of libraries was added to fresh LB media with appropriate antibiotics and grown overnight at 37C with shaking. Sorted libraries were miniprepped and stored at -20C until ready for sequencing.

The AncBD single-mutant library contained a subpopulation of cells which allowed expression of GFP in the absence of induction. Therefore, as an initial preparatory step, we sorted the repression-competent variants from an uninduced AncBD single-mutant library with the same settings and procedure as above. This repression-competent version of the library was used as input for the induced sorts.

Deep sequencing

For both replicates of pre-selection and sorted libraries, the variable region of the TetR gene was amplified by PCR using primers that attach Illumina sequencing primer binding sites. This PCR used 1ng of template with a total of 12 PCR cycles. After purifying PCR product, a second PCR was performed to attach i5 and i7 indexes and P5 and P7 sequences. The second PCR used 10ng template with a total of 8 PCR cycles. The minimum number of cycles necessary to obtain enough DNA was used in both PCR steps to minimize PCR bias and template switching.

After confirming fragment size and purity by gel electrophoresis and measuring DNA concentration with a Qubit fluorimeter, libraries were pooled and prepared for sequencing on the

Illumina NextSeq 1000 using a 750pM library concentration and 5-15% PhiX spike-in to account for low base diversity.

Calculation of enrichment scores

Raw Fastq files were retrieved from the NextSeq 1000 and processed using compute resources at the UW Madison biochemistry department. First, paired-end reads were merged using PEAR with a minimum overlap of 10bp. Merged reads were filtered for quality, requiring all bases in each read have a Phred score of at least Q20 and at least 95% of bases have a Phred score of at least Q30.

Merged reads passing all quality filters were analyzed using a custom python script that computes functional scores. Functional scores are calculated using a wild-type normalized log ratio of pre- to post-selection counts, based on that used in Enrich2 software: $F = \ln((counts_{var,\ selected} + pseudocount)/counts_{WT,\ selected}) - \ln((counts_{var,\ unselected} + pseudocount)/counts_{WT,\ unselected})$. A pseudocount of 0.01 was used to allow calculation of scores for variants containing zero post-selection observations. Standard error for each score was calculated based on the method used in Enrich2: $SE = sqrt((1/(counts_{var,\ unselected} + pseudocount)) + (1/(counts_{WT,\ unselected} + pseudocount)) + (1/(counts_{var,\ selected} + pseudocount)) + (1/(counts_{WT,\ selected} + pseudocount)))$. Functional scores and standard errors were calculated for each replicate selection separately, then merged using a robust maximum likelihood estimator as in Enrich2 using 100 iterations.

Combinatorial library screen

To reduce uncertainty, we performed the combinatorial library screen using clonal fluorescence measurements. Each library was spread on LB agar plates containing 100ug/mL spectinomycin and 50ug/mL kanamycin. 288 colonies were picked from each library and used to inoculate 150uL LB with appropriate antibiotics in 96-well plates, then grown overnight at 37C with

shaking (700rpm). The next morning, cultures were diluted 1:50 into fresh media in duplicate. One replicate was grown for 1 hour, then induced with 1uM anhydrotetracycline. Fluorescence was measured for induced and uninduced plates after 4 hours of incubation and fold-induction was calculated after appropriate normalization (see "Fluorescence measurements" methods). Deep sequencing was used to identify the variant expressed in each well of the fluorescence experiment. In a plate format, 1uL of starter culture was used as template for PCRs to amplify the TetR gene and flanking barcodes and attach Illumina adapters. A second PCR appended i5 and i7 indexes and P5 and P7 sequences, using unique indexes for each plate row and column, such that sequencing reads from each plate well could be identified by their combination of i5 and i7 indexes. PCR product was pooled and purified, quality checked by gel electrophoresis, and fluorometrically quantified before sequencing on the NextSeq 1000.

After quality filtering as previously described, sequencing data was grouped by i5/i7 combination and barcodes were extracted to identify which variant was expressed in each well of the plate. This information was mapped back to the fluorescence induction screen to determine the inducibility of each variant.

Chapter 5

Conclusions

5.1 Simplicity in the origination of new molecular phenotypes

This thesis shows that the genetic mechanisms by which multiple ubiquitous proteins features arose during evolution was simple. By using hemoglobin as a model system, we have been able to pinpoint the exact residue changes that caused the emergence of three features: gain of new interface, specificity between paralogs, and allosteric regulation. These features arose through surprisingly few genetic changes, which became evident only when the substitutions were examined in their structural and biochemical context. These features are common in other proteins, suggesting that the evolution of new interfaces, specificity, or allostery may arise through similarly simple mechanisms.

5.2 Implications for novel molecular complexes

In acquiring a new interface, hemoglobin required only a single substitution to go from an ancestral dimer to a homotetramer. Acquiring this interface was aided by the preexisting dimeric structure that emerged prior in history. This dimer was symmetric, with each subunit oriented identically within the complex. Upon substitution of a bulky tryptophan, the residue appears multiple times, providing enough energy to allow for assembly into the tetramer. Symmetry facilitated the emergence of a new interface by multiplying the effect of a large bulky hydrophobic change.

These results have implications for the evolution of natural complexes. Many protein complexes display symmetry (75). This is thought to have occurred because of selection for these types of structures, which may allow for elaboration of new functions or types of regulation (106, 107). Instead, our work suggests that symmetry arise because of an intrinsic propensity for multiplying

the energetic effects of mutations. It would be impossible for selection to have built the hemoglobin structure over time, because a single substitution was sufficient for the emergence of tetramer. Other studies have found large structural transitions through single substitution which are facilited via symmetry, suggesting that these mechanisms may be a common route for the evolution of complexes (117).

5.2 Implications for specificity in complexes

We identified the mechanisms that lead to the hetero-specific assembly of the alpha and beta subunit within the heterotetramer in Hb. Within one of the two interfaces in the Hb tetramer, a single deletion in the alpha subunit led to the evolution of specificity after gene duplication. Here, we also find that symmetry facilitated the evolution of specificity between α and β subunits in hemoglobin. Because a single substitution in a homodimer appears twice, but only once in the heterodimer—and not at all in the alternate homodimer—its effect on binding affinity differs across subunit assemblies. This asymmetry in functional impact enabled the evolution of paralog-specific interactions.

These results have implications on the emergence of heterospecificity between sister proteins after gene duplication. Proteins after duplication often evolve specificity for heteromers and people have wondered whether this occurs neutrally (84, 93). Our work suggests that both genetic and thermodynamic factors may bias protein evolution toward the formation of heteromers. First, following the duplication of an ancestral homomer, heteromeric assemblies are statistically favored due to thermodynamic principles—there are more ways to form heteromers than homomers, resulting in 50% of all dimers being heteromers. Second, we find that a single amino acid substitution can be sufficient to confer specificity between subunits. This simplicity, combined with the high baseline occupancy of heteromeric assemblies, could create a production

bias that favors the evolution of new heteromers. As a result, heteromerization may represent the most accessible evolutionary pathway for recently duplicated paralogs.

The implications of this work open several promising avenues for future research. A key next step is to determine whether there is a mutational bias favoring heteromeric assembly, which could be assessed by measuring the distribution of mutational effects on specificity between recently duplicated paralogs. High throughput approaches such as yeast surface display can be used to quantify both heteromeric and homomeric binding affinities (170). These data can then be integrated with the thermodynamic models developed in our work to infer the occupancy of specific assembly across mutational landscapes. Another important direction is to investigate whether residue-level asymmetry exists within symmetric complexes and how it arises. One potential strategy is to use NMR to monitor the behavior of individual interface residues and compare the interaction profile of this residue with the identical residue in the other subunit, noting differences to provide insight into asymmetric residues within symmetric architectures (171).

5.3 implications for the evolution of allosteric regulation

Our work establishes the first historical description for the emergence of allosteric regulation. Our work agrees with much of the literature that describes latent allostery in extant proteins, where pre-existing molecular features can be coopted to produce regulation in the function (41-43, 128, 131). It extends these features by describing which features were latent and which were built through the interactions between residues. Together, we obtained a broad description of the underlying molecular forces that created allosteric regulation in hemoglobin.

This work has broad implications for understanding the evolution of allosteric regulation. First, it demonstrates that the emergence of novel regulatory mechanisms does not require extensive

sequence changes; instead, just a few substitutions at sites that influence ligand-sensitive tertiary movements can be sufficient. In globin folds, we find that this is a recurring evolutionary strategy: multimerization often occurs at or near helices that undergo conformational shifts upon oxygen binding. These structural dynamics create a natural scaffold for the evolution of cooperativity and allosteric communication between subunits (139, 154). Discreet and regulatable changes in protein conformations are a pre-requisite of allosteric regulation and nearly all proteins exhibit conformational dynamics to varying degrees (126, 155). Broadly, exploiting pre-existing movements through multimerization or ligand binding may be a more common strategy by which allosteric regulation occurs.

5.4 implications for epistasis and allostery

We have shown the underlying genetic basis for allosteric regulation is changes often during the natural history of TetR because of epistasis. This raises an important question: when are the genetic effects on homologous allosteric proteins no longer correlated? A key test will be to assess whether our ability to predict allosteric behavior declines significantly as sequence divergence increases, providing insights into how the genetic architecture of allostery evolves over time.

Understanding epistatic interactions within allosteric proteins may be key to determining how allosteric function is determine by their residues. One promising approach involves generating combinatorial libraries to systematically probe the effects of pairwise mutations across a protein, quantifying how these mutations, and their epistatic interactions, impact effector binding and overall protein function. While several computational models now exist to analyze large-scale datasets, current high-throughput assays typically focus on a single functional parameter (149, 172, 173). Future efforts should prioritize the development of platforms capable of measuring

multiple biophysical properties in parallel. This would enable a more comprehensive, quantitative understanding of how mutations and epistasis shape both individual molecular functions and their interdependencies within allosteric systems.

Appendix 1

Supplementary figures for Chapter 2: Symmetry facilitated the evolution of heterospecificity and high-order stoichiometry in vertebrate hemoglobin

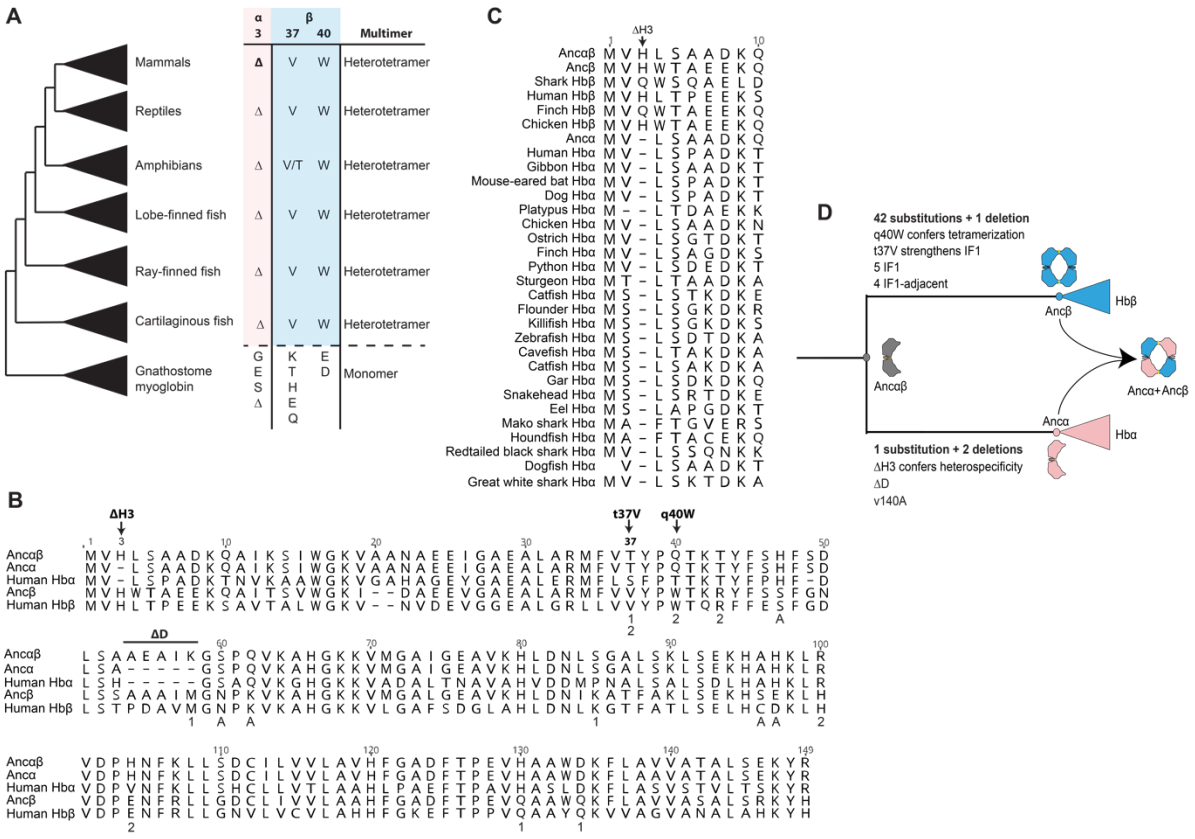

Figure A1.1. Key sequence changes are conserved in extant hemoglobin subunits. A) Multimerization state and residues at key sequence sites in extant Hb subunits from major jawed vertebrate taxa. Myoglobin is the closest paralogous protein in the globin family. For complete alignment, see https://tinyurl.com/yc6kvdb8. **B**) Alignment of extant human and reconstructed ancestral Hb subunits is shown. Historical sequence changes that confer heterospecificity (DH3 in Hba) and tetramerization (q40W in Hbb) are labeled and shown with arrows. Additional substitutions that occurred on the branch leading to Ancb and assayed in this paper are also labeled, including changes on the surface of IF1 (1), IF2 (2), and at sites adjacent to IF1 (A). **C**) Alignment of N-terminal portion of Hb subunits from species representative of major vertebrate taxa. The deletion of residue 3 in alpha subunits is marked. **D**) Summary of key sequence changes that occurred after the duplication of Ancab. Multimerization states of reconstructed ancestral proteins is shown.

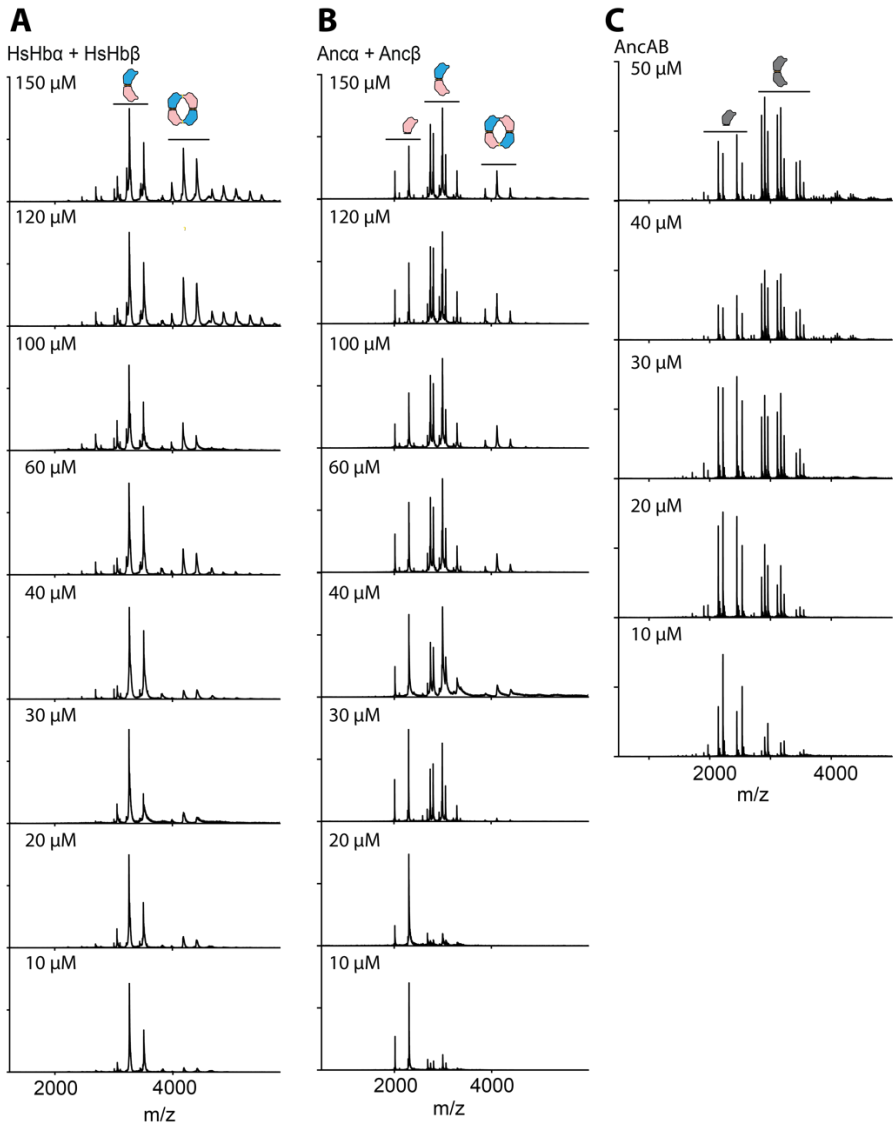

Figure A1.2. Native mass spectrometry spectra. nMS spectra across a concentration series is shown for A) human Hb, B) Ancα + Ancβ, and (C) Ancαβ. Peaks corresponding to monomers, dimers, and tetramer are labeled.

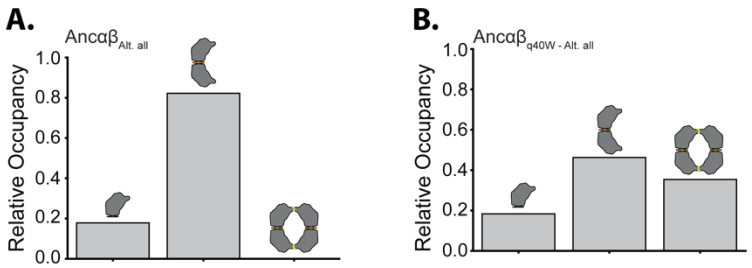

Figure A1.3. Effect of q40W tetramerization is robust to statistical uncertainty. (A) Relative occupancy of monomer, dimer, and tetramer of Ancαβ$_{Alt. all}$, an alternative reconstruction of Ancαβ that contains the second most likely state at all ambiguously reconstructed sites, measured at 20 μM total protein using native MS. (B) Relative occupancy Ancαβ$_{Alt. all}$ with substitution q40W.

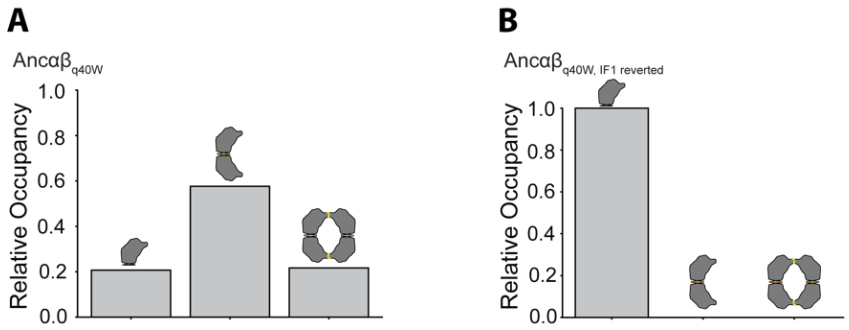

Figure A1.4. The effect of q40W on tetramerization depends on IF1. (A) Relative occupancy of Ancαβ$_{q40W}$, measured by native MS at 20 μM total protein. (B) Relative occupancy of Ancαβ$_{q40W_IF1\text{-}reverted}$, which contains mutation q40W, as well as reversions to the ancestral state found in AncMH of all residues that were substituted between AncMH and Ancαβ .

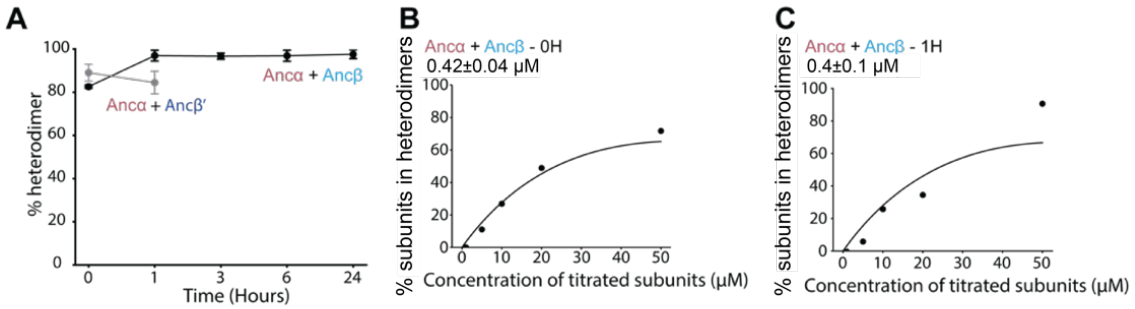

Figure A1.5. Heterodimer occupancy of Ancα and Ancβ is near equilibrium after mixing. (A) The percent of all dimers that are heterodimers, measured by nMS when proteins are mixed at 50 μM each and allowed to incubate for 0, 1, 3, 6, or 24 hours. Black line and points, Ancα + Ancβ (which only dimerize when expressed separately and then mixed). Grey line and points, Ancα + Ancβ' (Ancβ in which IF2 surface substitutions are reverted to their ancestral state in Ancαβ, thus preventing tetramerization). Each dot shows the mean of three replicates; error bars, SEM. (B) Affinity of monomer-to-heterodimer assembly measured by nMS immediately upon mixing of Ancα and Ancβ. Ancα was kept constant at 50 μM, while the concentration of Ancβ varied. Points, fraction of all subunits in the mixture that are incorporated into heterodimers. Line, best-fit binding curve. Estimated Kd and 95% confidence interval are shown. (C) Estimated heterodimerization affinity measured as in panel B, but 1 hour after mixing.

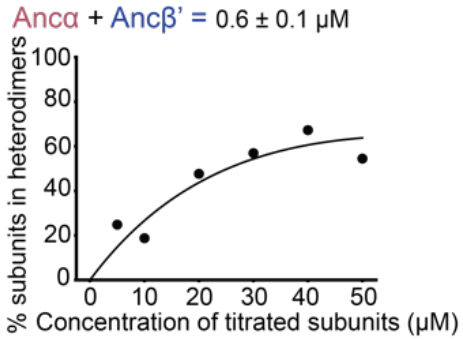

Figure A1.6. Heterodimerization by Ancα+Ancβ'. Monomer-to-heterodimer assembly measured by nMS. Ancα was kept constant at 50 μM while Ancβ' was at variable concentration. Points, fraction of all subunits in the mixture that are incorporated into heterodimers at each concentration. Line, best-fit binding curve. Estimated Kd and 95% confidence interval are shown.

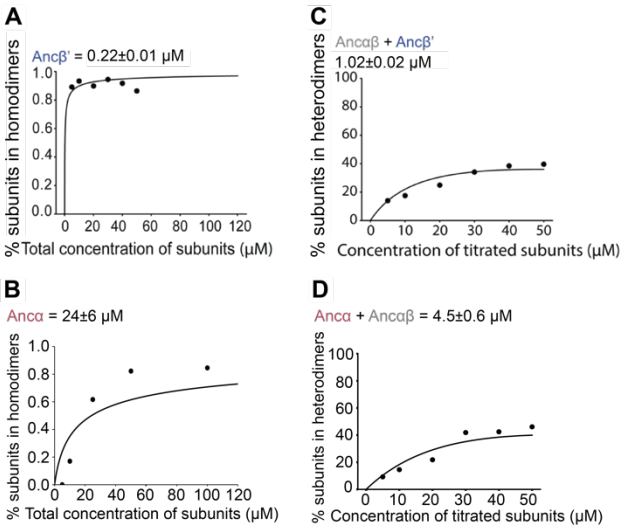

Figure A1.7. Dimerization by Ancα and Ancβ'. (A-B) Homodimerization by Ancβ' (panel A) and by Ancα (B). measured by nMS across a titration series. Each point shows the fraction of subunits incorporated into dimers as the concentration of protein varied. Best-fit binding curve, Kd, and 95% confidence interval are shown. (C-D) Heterodimerization by mixtures of Ancαβ+ Ancβ (C) and Ancαβ+Ancα and Ancα+Ancαβ (D). Each point shows the fraction of all subunits incorporated into heterodimers. In each case, one protein was held constant at 50 mM while the other was varied.

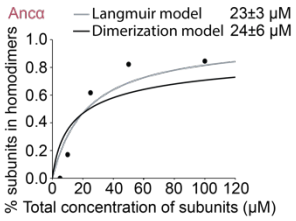

Figure A1.8. Estimated affinity of the monomer-dimer transition by the Ancα homodimer is robust to the binding model used. We fit two different binding equations (curves) to the stoichiometries of Ancα measured using nMS. The Langmuir-Hill model (black line) assumes that the concentration of free monomeric subunits is not depleted by dimerization. The dimerization model used throughout the rest of this paper accounts for this depletion (described in Materials and Methods). The estimated Kds (shown with their 95% confidence intervals) are statistically indistinguishable.

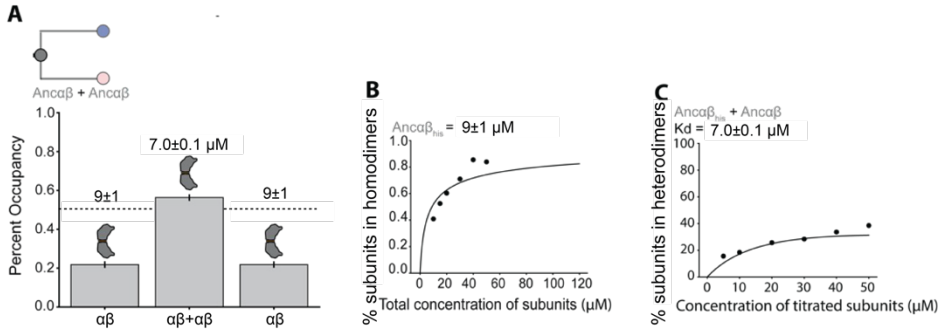

Figure A1.9. Dimerization affinities and occupancy of Ancαβ. (A) Expected fractional occupancies of homodimer and heterodimers when Ancαβ is mixed at equal concentrations with Ancαβ$_{his}$ (500 mM each), given the measured dimerization affinities (shown above each column, with 95% confidence interval). Ancαβ$_{his}$ is Ancαβ with an N-terminal polyhistidine tag, which allows the masses of the three kinds of dimer to be distinguished. (B-C) Homodimerization by Ancαβ$_{his}$ and heterodimerization by affinity of Ancαβ+ Ancαβ$_{his}$, measured and represented as in Fig. S5.

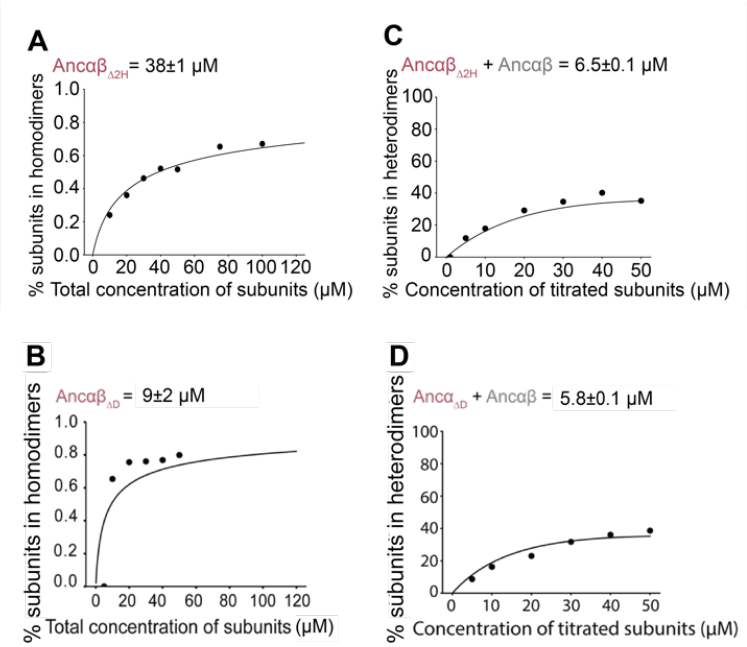

Figure A1.10. Effect of historical deletions on dimerization. (A-B) Homodimerization and (C-D) Heterodimerization by mixtures, measured and represented as in Fig. S5.

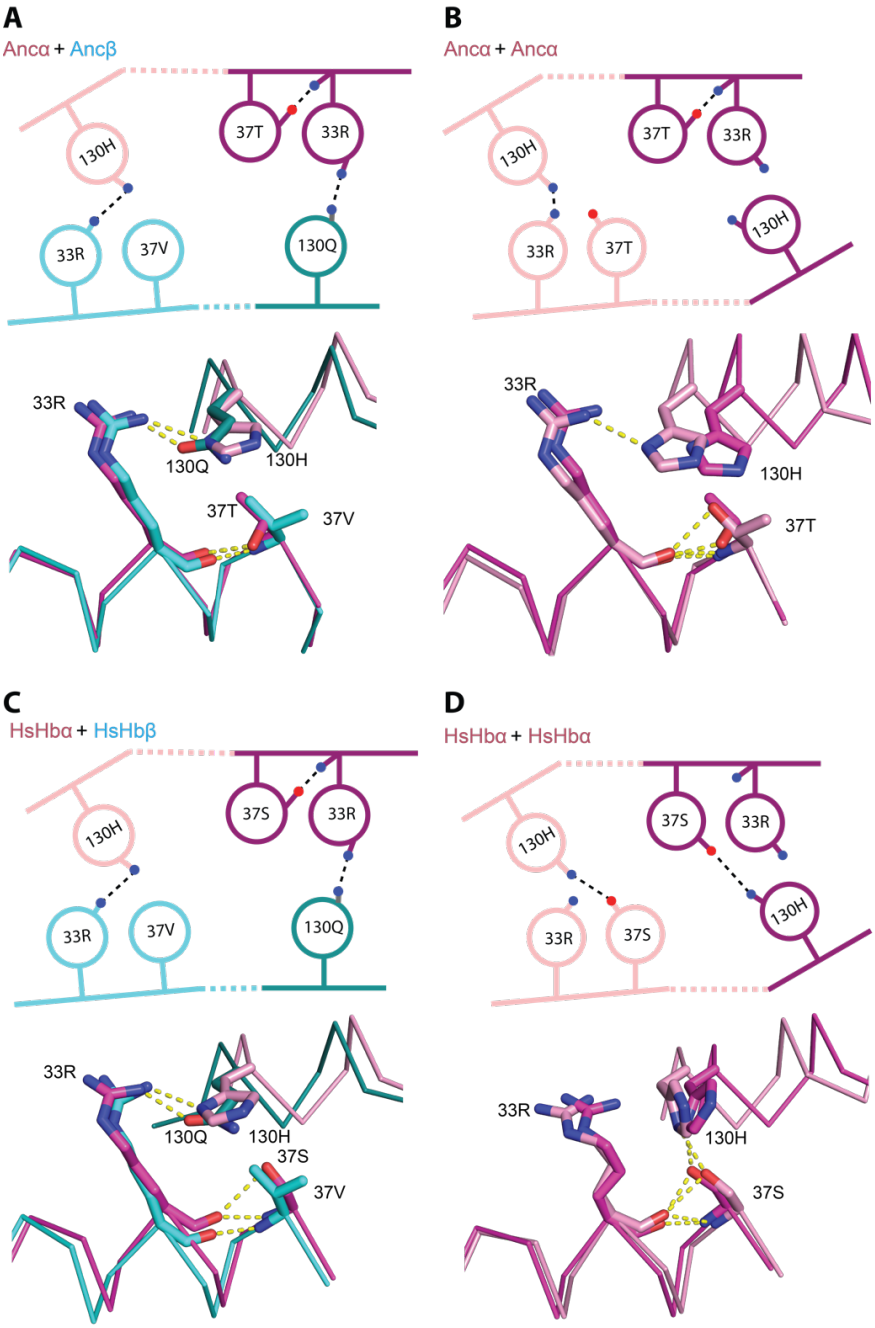

Figure A1.11. Nonadditive interactions that contribute to specificity are conserved in derived Hb complexes. In the modeled homodimers and heterodimers of Ancα+Ancβ (panels A, B) and X-ray crystal structure of human hemoglobin (PDB 4HHB and 3S48), the figure shows the key IF1 residues with nonadditive interactions in Ancαβ+Ancαβ$_{\Delta H3}$ (see Fig. 5G for comparison). Top, cartoon of key contacts. The two iterations of these interactions across the isologous interface are shown, one each in light or dark hue. Blue and red, nitrogen and oxygen atoms, respectively. Dotted lines, hydrogen bonds. Bottom, structural alignment of the two iterations of the isologous interface in each dimer. Each dimer structure was duplicated exactly

and then aligned to the original by targeting one subunit of the copy to align to the other subunit of the original. Hues correspond to the isologous iterations in the cartoon above

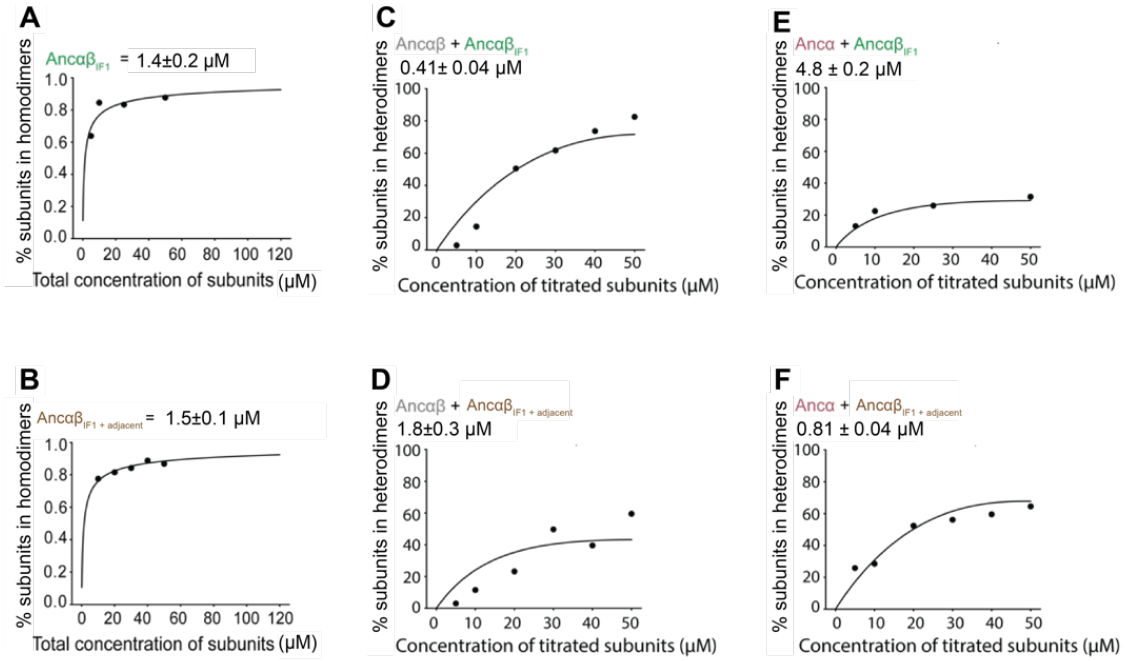

Figure A1.12. Homodimerization by Ancαβ$_{IF1}$ and Ancαβ$_{IF1 + Adjacent}$.(A,B) and heterodimerization by those proteins when mixed with Ancαβ (C,D) or Ancα (E,F). Measurements and representation as in Fig. S5.

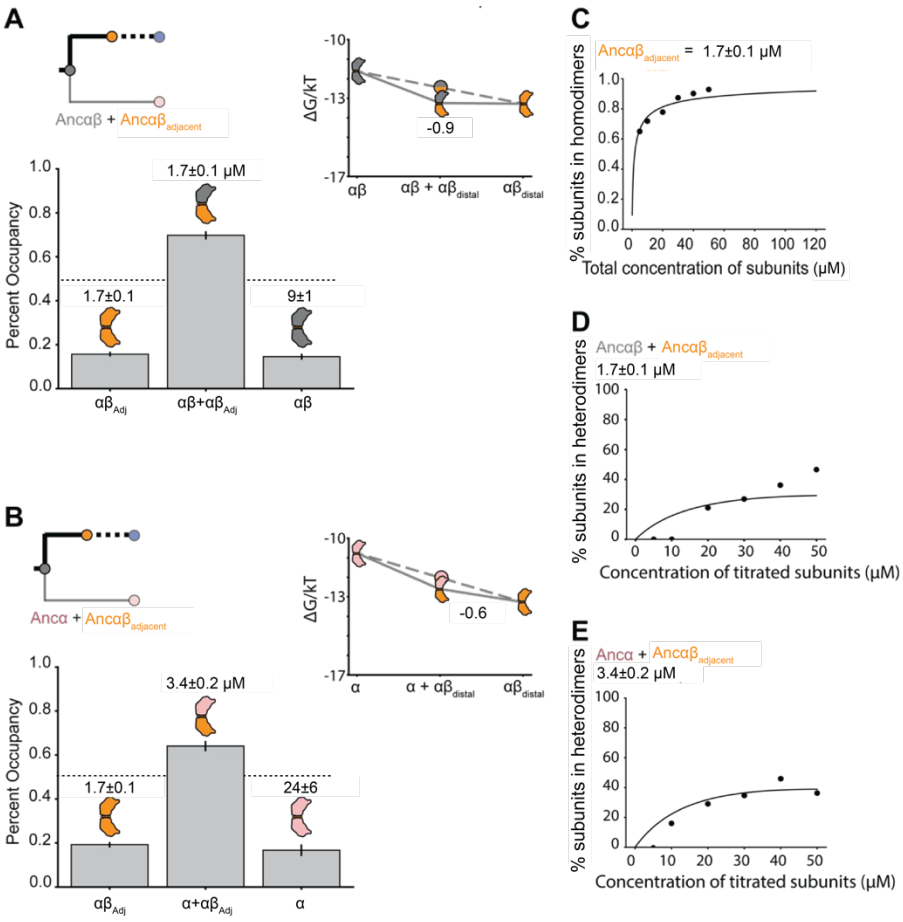

Figure A1.13. Dimerization affinity and occupancies for Ancαβ$_{Adjacent}$. Expected fractional occupancies of homodimer and heterodimers when Ancαβ$_{Adjacent}$ Is mixed with Ancαβ (A) or Ancα (B), each at (500 mM), given the measured dimerization affinities (shown above each column, with 95% confidence interval). Inset, ΔG of each dimerization (measured in units of kT), with ΔG_{spec} of the heterodimer shown. (C,D,E) Measurement of binding affinities, measured and represented as in Fig. S5.

Appendix 2

Supplementary Figures for Chapter 3: Evolutionary origins of allostery in vertebrate hemoglobin

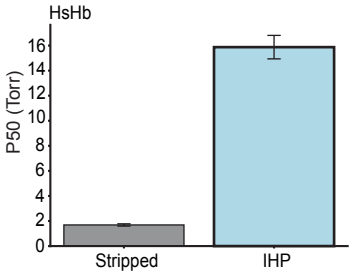

Figure A2.1. Oxygen affinity of Humam Hb. In grey, stripped condition, where no IHP is in solution. In blue, IHP condition, where 500 μM of IHP is added to solution. Error bars represent standard error of measurement, n = 5.

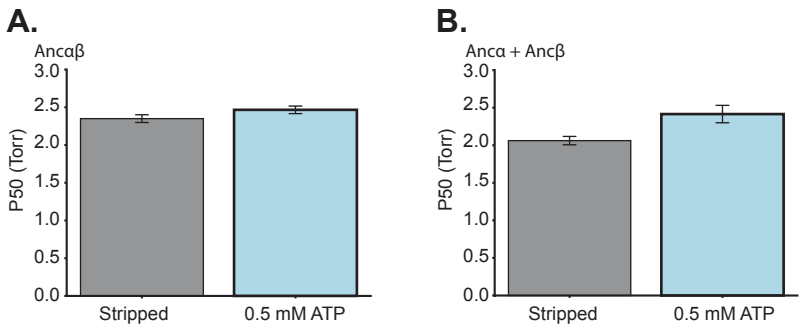

Figure A2.2. Oxygen affinity of Ancestral proteins with ATP. (A) Oxygen affinity of Ancαβ. In grey, stripped condition, where no ATP is in solution. In blue, ATP condition, where 500 μM of IHP is added to solution. Error bars represent standard error of measurement, n = 5. **(B)** Oxygen affinity of Ancα + Ancβ. In grey, stripped condition, where no ATP is in solution. In blue, ATP condition, where 500 μM of IHP is added to solution. Error bars represent standard error of measurement, n = 5.

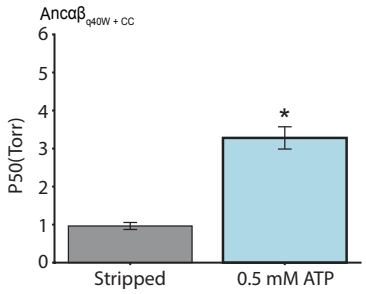

Figure A2.3. Oxygen affinity of Ancαβ$_{q40W + CC}$ with ATP. In grey, stripped condition, where no ATP is in solution. In blue, ATP condition, where 500 μM of IHP is added to solution. Error bars represent standard error of measurement, n = 5.

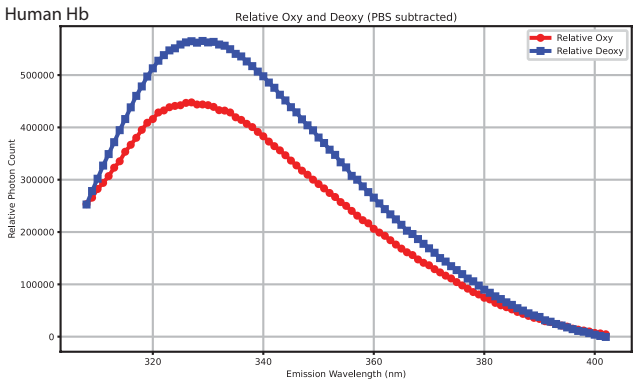

Figure A2.4. Fluorescence emission scan of Human Hb, excited at 280. In blue, Human Hb has been deoxygenated. In red, Human Hb is in oxygenated conditions. Error bars represent standard error of measurement, n = 10.

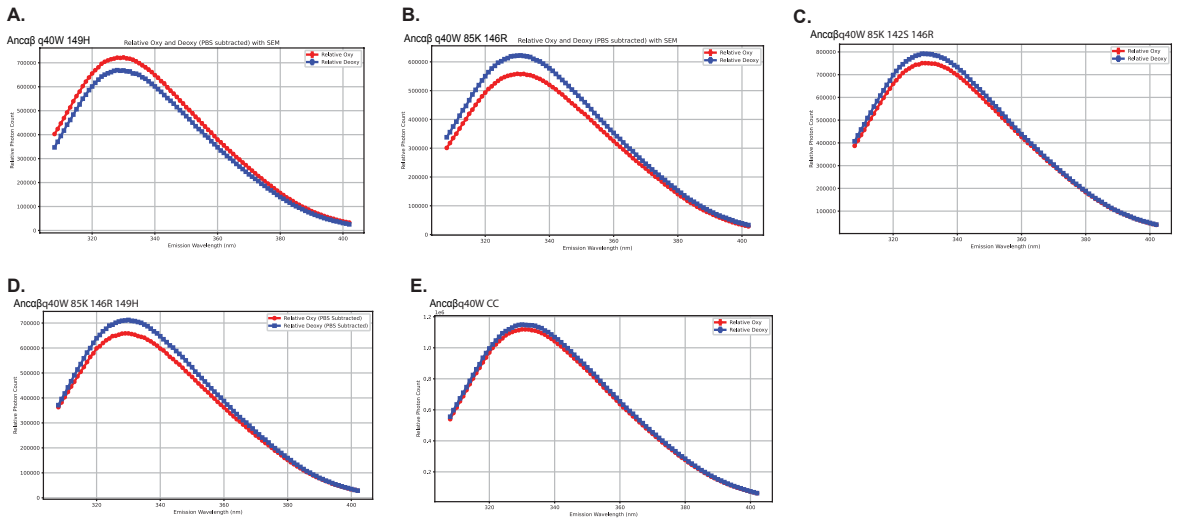

Figure A2.5. Flouresence emission scan of central cavity variants on the background of Ancaβ$_{q40W}$. (A) Emission scan of Ancaβ$_{q40W\ t149H}$ when excited at 280 nm. Blue points is deoxygenated condition. Red points is oxygenated condititon. Error bars represent standard error of measurement, n = 7. **(B)** Emission scan of Ancaβ$_{q40W\ 85K\ 146R}$ when excited at 280 nm. Blue points, red points, and error bars same as mentioned previously. **(C)** Emission scan of Ancaβ$_{q40W\ 85K\ 142S\ 146R}$ when excited at 280 nm. Blue points, red points, and error bars same as mentioned previously. **(D)** Emission scan of Ancaβ$_{q40W\ 85K\ 146R\ 149H}$ when excited at 280 nm. Blue points, red points, and error bars same as mentioned previously. **(E)** Emission scan of Ancaβ$_{q40W\ CC}$ when excited at 280 nm. Blue points, red points, and error bars same as mentioned previously.

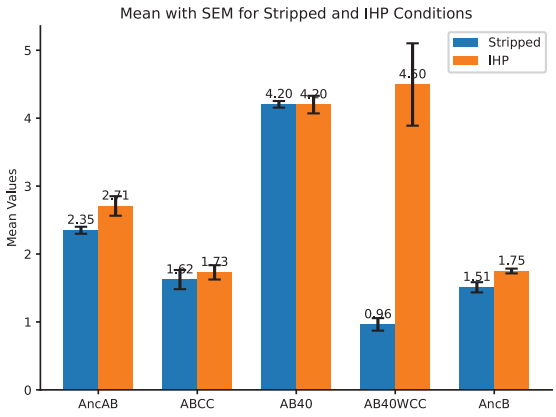

Figure A2.6. Oxygen affinity of multiple variants leading to Ancβ. Blue bars contain no effector. Orange bar contains 500 uM of IHP. Error bars are standard error of measurement, n = 3.

Appendix 3

Supplementary figure for Chapter 4: Rampant epistasis during the evolution of an allosteric transcription factor reveals shifting genetic basis.

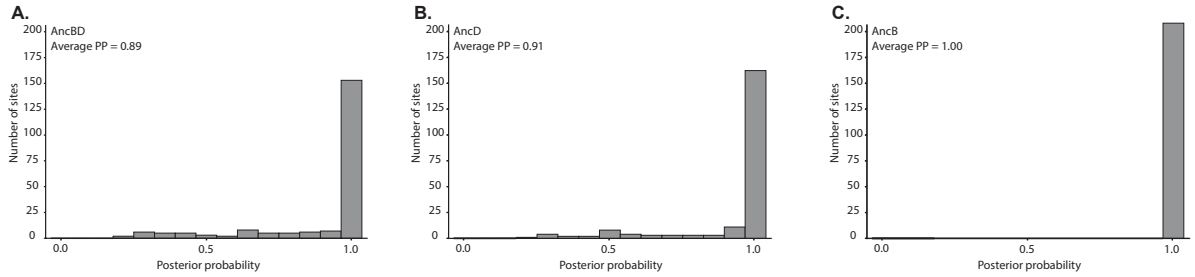

Figure A3.1. Posterior probably of states within ancestral proteins. (A) Statistical confidence of each maximun aposteriori (MAP) ancestral residue within AncBD. **(B)** Statistical confidence of MAP ancestral residues within AncD. **(C)** Statistical confidence of MAP ancestral residues within AncB.

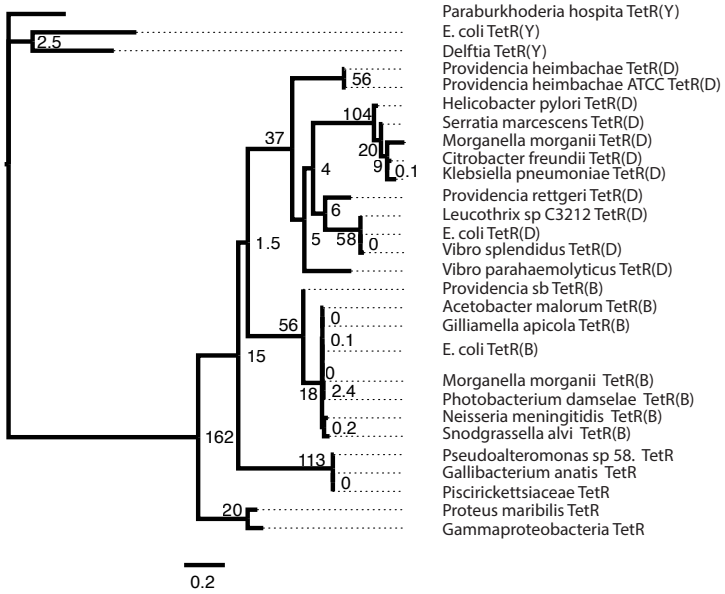

Figure A3.2. Maximum likelihood of TetR(B) and TetR(D) phylogeny. The aLRT scores of each node is shown.

AncBD Genotype		AncB		E.c TetR(B)	
A2S		A50T		2	A
E7S		C68L		107	S
K8T		D10T		120	M
S12A		S74T		121	V
L68C		F188Y		122	D
L70A		I113V		123	N
S74T		L176I		128	K
T107S		N205K		129	C
A161I		P184E		129	T
T163D		Q172E		129	D
Q172E		T165N		130	G
N179S				134	I
D180Q				144	S
Q184E				148	Q
V192A				152	I
Q204E				152	A
K205S				181	N
				192	V
				199	A
				203	A
				204	Q
				205	N

Table A3.1. Residue mutations that are allosterically inactivating. Each column represents the residue and its position that have an enrichment score at least one standard deviation away from the WT in the positive direction.

Bibliography

1. T. Sanvictores, F. Farci (October 31, 2022) Biochemistry, Primary Protein Structure. (StatPearls Publishing).

2. J. A. Alberts B, Lewis J, et al. (2002) Molecular Biology of the Cell. 4th edition. (Garland Science, New York).

3. G. K. A. Hochberg, J. W. Thornton, Reconstructing ancient proteins to understand the causes of structure and function. *Annual Review of Biophysics* **46**, 247-269 (2017).

4. M. J. Harms, J. W. Thornton, Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics* **14**, 559-571 (2013).

5. E. Zuckerkandl, L. Pauling, "Evolutionary Divergence and Convergence in Proteins" in Evolving Genes and Proteins*,* V. Bryson, H. J. Vogel, Eds. (Academic Press, 1965), pp. 97-166.

6. Z. Yang, B. Rannala, Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**, 303-314 (2012).

7. L. Pauling, E. Zuckerkandl, T. Henriksen, R. Lövstad, Chemical paleogenetics. *Acta chem. scand* **17**, S9-S16 (1963).

8. Z. Yang, S. Kumar, M. Nei, A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641-1650 (1995).

9. J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376 (1981).

10. J. Felsenstein, Phylogenies and the Comparative Method. *The American Naturalist* **125**, 1-15 (1985).

11. Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556 (1997).

12. M. A. Spence, J. A. Kaczmarski, J. W. Saunders, C. J. Jackson, Ancestral sequence reconstruction for protein engineers. *Current Opinion in Structural Biology* **69**, 131-141 (2021).

13. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Science* **25**, 1204-1218 (2016).

14. D. W. Anderson, A. N. McKeown, J. W. Thornton, Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* **4**, e07864 (2015).

15. B. P. H. Metzger, Y. Park, T. N. Starr, J. W. Thornton, Epistasis facilitates functional evolution in an ancient transcription factor. *eLife* **12**, RP88737 (2024).

16. C. Bank, Epistasis and adaptation on fitness landscapes. *Annual Review of Ecology, Evolution, and Systematics* **53**, 457-479 (2022).

17. K. Buda, C. M. Miton, N. Tokuriki, Pervasive epistasis exposes intramolecular networks in adaptive enzyme evolution. *Nat Commun* **14**, 8508 (2023).

18. L. Di Bari, M. Bisardi, S. Cotogno, M. Weigt, F. Zamponi, Emergent time scales of epistasis in protein evolution. *Proceedings of the National Academy of Sciences* **121**, e2406807121 (2024).

19. J. T. Bridgham, E. A. Ortlund, J. W. Thornton, An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515-519 (2009).

20. J. T. Bridgham, S. M. Carroll, J. W. Thornton, Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**, 97-101 (2006).

21. J. W. Thornton, E. Need, D. Crews, Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* **301**, 1714-1717 (2003).

22. Alesia N. McKeown *et al.*, Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell* **159**, 58-68 (2014).

23. G. K. A. Hochberg *et al.*, A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503-508 (2020).

24. C. The UniProt, UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53**, D609-D617 (2025).

25. T. Kortemme, De novo protein design—From new structures to programmable functions. *Cell* **187**, 526-544 (2024).

26. S. Jones, J. M. Thornton, Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20 (1996).

27. H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E. D. Levy, Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244-247 (2017).

28. A. Yang *et al.*, Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science* **381**, eadh1720.

29. O. Ashenberg, K. Rozen-Gagnon, M. T. Laub, A. E. Keating, Determinants of homodimerization specificity in histidine kinases. *J Mol Biol* **413**, 222-235 (2011).

30. S. E. Boyken *et al.*, De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science* **352**, 680-687 (2016).

31. J. Dauparas *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).

32. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).

33. J. L. Watson *et al.*, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023).

34. P. C. Després *et al.*, Compensatory mutations potentiate constructive neutral evolution by gene duplication. *Science* **385**, 770-775 (2024).

35. C. D. Aakre *et al.*, Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594-606 (2015).

36. E. J. Capra, B. S. Perchuk, J. M. Skerker, M. T. Laub, Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150**, 222-232 (2012).

37. D. Ding *et al.*, Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat Ecol Evol* **6**, 590-603 (2022).

38. C. J. McClune, A. Alvarez-Buylla, C. A. Voigt, M. T. Laub, Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature* **574**, 702-706 (2019).

39. S. Berger *et al.*, Computationally designed high specificity inhibitors delineate the roles of BCL2 family proteins in cancer. *eLife* **5**, e20352 (2016).

40. B. Liu *et al.*, Design of high specificity binders for peptide-MHC-I complexes. *bioRxiv*, 2024.2011.2028.625793 (2024).

41. J. W. McCormick, M. A. X. Russo, S. Thompson, A. Blevins, K. A. Reynolds, Structurally distributed surface sites tune allosteric regulation. *eLife* **10**, e68346 (2021).

42. D. Pincus *et al.*, Engineering allosteric regulation in protein kinases. *Sci. Signal.* **11**, eaar3250 (2018).

43. Kimberly A. Reynolds, Richard N. McLaughlin, R. Ranganathan, Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* **147**, 1564-1575 (2011).

44. G. Guntas, C. Purbeck, B. Kuhlman, Engineering a protein–protein interface using a computationally designed library. *Proceedings of the National Academy of Sciences* **107**, 19296-19301 (2010).

45. V. Mathieu, J. Fastrez, P. Soumillion, Engineering allosteric regulation into the hinge region of a circularly permuted TEM-1 β-lactamase. *Protein Engineering, Design and Selection* **23**, 699-709 (2010).

46. K. Deckert, S. J. Budiardjo, L. C. Brunner, S. Lovell, J. Karanicolas, Designing Allosteric Control into Enzymes by Chemical Rescue of Structure. *Journal of the American Chemical Society* **134**, 10055-10060 (2012).

47. Y. Xia *et al.*, The Designability of Protein Switches by Chemical Rescue of Structure: Mechanisms of Inactivation and Reactivation. *Journal of the American Chemical Society* **135**, 18840-18849 (2013).

48. A. Pillai *et al.*, De novo design of allosterically switchable protein assemblies. *Nature* **632**, 911-920 (2024).

49. S. Raman, Systems Approaches to Understanding and Designing Allosteric Proteins. *Biochemistry* **57**, 376-382 (2018).

50. K. E. Johnston *et al.*, Machine Learning for Protein Engineering. *ArXiv* (2023).

51. J. F. Storz, *Hemoglobin: insights into protein structure, function, and evolution* (Oxford University Press, 2018).

52. M. H. Ahmed, M. S. Ghatge, M. K. Safo, Hemoglobin: Structure, Function and Allostery. *Subcell Biochem* **94**, 345-382 (2020).

53. M. F. Perutz, H. Muirhead, J. M. Cox, L. C. G. Goaman, Three-dimensional Fourier Synthesis of Horse Oxyhaemoglobin at 2.8 Å Resolution: The Atomic Model. *Nature* **219**, 131-139 (1968).

54. G. K. Ackers, Energetics of subunit assembly and ligand binding in human hemoglobin. *Biophysical journal* **32**, 331-346 (1980).

55. W. P. Griffith, I. A. Kaltashov, Highly asymmetric interactions between globin chains during hemoglobin assembly revealed by electrospray ionization mass spectrometry. *Biochemistry* **42**, 10024-10033 (2003).

56. J. Liu, L. Konermann, Assembly of hemoglobin from denatured monomeric subunits: heme ligation effects and off-pathway intermediates studied by electrospray mass spectrometry. *Biochemistry* **52**, 1717-1724 (2013).

57. R. Benesch, R. E. Benesch, Homos and Heteros among the Hemos. *Science* **185**, 905-908 (1974).

58. S. J. Edelstein, Cooperative interactions of hemoglobin. *Annual review of biochemistry* **44**, 209-232 (1975).

59. M. F. Perutz, G. Fermi, B. Luisi, B. Shaanan, R. C. Liddington, Stereochemistry of cooperative mechanisms in hemoglobin. *Accounts of Chemical Research* **20**, 309-321 (1987).

60. J. Monod, J. Wyman, J.-P. Changeux, On the nature of allosteric transitions: a plausible model. *Journal of molecular biology* **12**, 88-118 (1965).

61. B. R. Gelin, M. Karplus, Mechanism of tertiary structural change in hemoglobin. *Proceedings of the National Academy of Sciences* **74**, 801-805 (1977).

62. Q. Cui, M. Karplus, Allostery and cooperativity revisited. *Protein science* **17**, 1295-1307 (2008).

63. K. Kanaori *et al.*, T-quaternary structure of oxy human adult hemoglobin in the presence of two allosteric effectors, L35 and IHP. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1807**, 1253-1261 (2011).

64. M. F. Perutz, Species adaptation in a protein molecule. *Molecular Biology and Evolution* **1**, 1-28 (1983).

65. C. R. Cortez-Romero, J. Lyu, A. S. Pillai, A. Laganowsky, J. W. Thornton, Symmetry facilitated the evolution of heterospecificity and high-order stoichiometry in vertebrate hemoglobin. *Proceedings of the National Academy of Sciences* **122**, e2414756122 (2025).

66. M. Berenbrink, P. Koldkjær, O. Kepp, A. R. Cossins, Evolution of Oxygen Secretion in Fishes and the Emergence of a Complex Physiological System. *Science* **307**, 1752-1757 (2005).

67. J. F. Storz, Gene Duplication and Evolutionary Innovations in Hemoglobin-Oxygen Transport. *Physiology (Bethesda)* **31**, 223-232 (2016).

68. M. T. Grispo *et al.*, Gene duplication and the evolution of hemoglobin isoform differentiation in birds. *Journal of Biological Chemistry* **287**, 37647-37658 (2012).

69. A. S. Pillai *et al.*, Origin of complexity in haemoglobin evolution. *Nature* **581**, 480-485 (2020).

70. J. L. Ramos *et al.*, The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* **69**, 326-356 (2005).

71. P. Orth, D. Schnappinger, W. Hillen, W. Saenger, W. Hinrichs, Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system. *Nature Structural Biology* **7**, 215-219 (2000).

72. J. Deng, Y. Yuan, Q. Cui, Modulation of Allostery with Multiple Mechanisms by Hotspot Mutations in TetR. *bioRxiv* (2023).

73. S. E. Reichheld, Z. Yu, A. R. Davidson, The induction of folding cooperativity by ligand binding drives the allosteric response of tetracycline repressor. *Proceedings of the National Academy of Sciences* **106**, 22263-22268 (2009).

74. Z. Liu, T. Gillis, S. Raman, Q. Cui (2024) A parametrized two-domain thermodynamic model explains diverse mutational effects on protein allostery. (eLife Sciences Publications, Ltd).

75. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure* **29**, 105-153 (2000).

76. J. A. Marsh, S. A. Teichmann, Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem* **84**, 551-575 (2015).

77. M. Lynch, The Evolution of Multimeric Protein Assemblages. *Molecular Biology and Evolution* **29**, 1353-1366 (2012).

78. M. Lynch, Evolutionary diversification of the multimeric states of proteins. *Proceedings of the National Academy of Sciences* **110**, E2821-E2828 (2013).

79. A. S. Pillai, G. K. A. Hochberg, J. W. Thornton, Simple mechanisms for the evolution of protein complexity. *Protein Sci* **31**, e4449 (2022).

80. G. K. A. Hochberg *et al.*, Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. *Science* **359**, 930-935 (2018).

81. S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).

82. G. Diss *et al.*, Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science* **355**, 630-634 (2017).

83. S. Mallik, D. S. Tawfik, Determining the interaction status and evolutionary fate of duplicated homomeric proteins. *PLoS Comput Biol* **16**, e1008145 (2020).

84. A. Marchant *et al.*, The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *Elife* **8** (2019).

85. J. B. Pereira-Leal, E. D. Levy, C. Kamp, S. A. Teichmann, Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* **8**, R51 (2007).

86. D. Grueninger *et al.*, Designed Protein-Protein Association. *Science* **319**, 206-209 (2008).

87. J. Monod, J. Wyman, J.-P. Changeux, On the nature of allosteric transitions: a plausible model. *J Mol Biol* **12**, 88-118 (1965).

88. D. A. Ghose, K. E. Przydzial, E. M. Mahoney, A. E. Keating, M. T. Laub, Marginal specificity in protein interactions constrains evolution of a paralogous family. *Proc Natl Acad Sci U S A* **120**, e2221163120 (2023).

89. T. V. Lite *et al.*, Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *Elife* **9** (2020).

90. I. Nocedal, M. T. Laub, Ancestral reconstruction of duplicated signaling proteins reveals the evolution of signaling specificity. *eLife* **11**, e77346 (2022).

91. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, Evolution of increased complexity in a molecular machine. *Nature* **481**, 360-364 (2012).

92. J. R. Emlaw *et al.*, A single historical substitution drives an increase in acetylcholine receptor complexity. *Proc Natl Acad Sci U S A* **118** (2021).

93. A. F. Cisneros, L. Nielly-Thibault, S. Mallik, E. D. Levy, C. R. Landry, Mutational biases favor complexity increases in protein interaction networks after gene duplication. *Mol Syst Biol* **20**, 549-572 (2024).

94. A. C. Leney, A. J. R. Heck, Native mass spectrometry: what is in the name? *Journal of the American Society for Mass Spectrometry* **28**, 5-13 (2016).

95. E. Stellwagen, "Chapter 23 Gel Filtration1" in Methods in Enzymology, R. R. Burgess, M. P. Deutscher, Eds. (Academic Press, 2009), vol. 463, pp. 373-385.

96. L. Kiger *et al.*, Thermodynamic studies on the equilibrium properties of a series of recombinant betaW37 hemoglobin mutants. *Biochemistry* **37**, 4336-4345 (1998).

97. B. L. Boys, L. Konermann, Folding and Assembly of Hemoglobin Monitored by Electrospray Mass Spectrometry Using an On-line Dialysis System. *Journal of the American Society for Mass Spectrometry* **18**, 8-16 (2007).

98. G. Snyder, B. Sheafor, Red Blood Cells: Centerpiece in the Evolution of the Vertebrate Circulatory System. *Am. Zool.* **39** (1999).

99. G. Fermi, M. F. Perutz, B. Shaanan, R. Fourme, The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *Journal of Molecular Biology* **175**, 159-174 (1984).

100. A. A. Bogan, K. S. Thorn, Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**, 1-9 (1998).

101. J. Jee, I.-J. L. Byeon, J. M. Louis, A. M. Gronenborn, The point mutation A34F causes dimerization of GB1. *Proteins: Structure, Function, and Bioinformatics* **71**, 1420-1431 (2008).

102. H. Garcia Seisdedos, T. Levin, G. Shapira, S. Freud, E. D. Levy, Mutant libraries reveal negative design shielding proteins from supramolecular self-assembly and relocalization in cells. *Proceedings of the National Academy of Sciences* **119**, e2101117119 (2022).

103. C.-S. Chen *et al.*, How to Change the Oligomeric State of a Circular Protein Assembly: Switch from 11-Subunit to 12-Subunit TRAP Suggests a General Mechanism. *PLOS ONE* **6**, e25296 (2011).

104. B. Kuhlman, J. W. O'Neill, D. E. Kim, K. Y. Zhang, D. Baker, Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc Natl Acad Sci U S A* **98**, 10687-10691 (2001).

105. H. Schweke *et al.*, An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999-1010.e1015 (2024).

106. J. A. Marsh *et al.*, Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461-470 (2013).

107. E. D. Levy, E. B. Erba, C. V. Robinson, S. A. Teichmann, Assembly reflects evolution of protein complexes. *Nature* **453**, 1262-1265 (2008).

108. R. P. Bahadur, F. Rodier, J. Janin, A dissection of the protein–protein interfaces in icosahedral virus capsids. *Journal of molecular biology* **367**, 574-590 (2007).

109. E. T. Powers, D. L. Powers, A perspective on mechanisms of protein tetramer formation. *Biophysical journal* **85**, 3587-3599 (2003).

110. J. T. Brennecke, B. L. de Groot, Quantifying Asymmetry of Multimeric Proteins. *The Journal of Physical Chemistry A* **122**, 7924-7930 (2018).

111. M. Bonjack-Shterengartz, D. Avnir, The near-symmetry of proteins. *Proteins: Structure, Function, and Bioinformatics* **83**, 722-734 (2015).

112. C. R. Baker, V. Hanson-Smith, A. D. Johnson, Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. *Science* **342**, 104-108 (2013).

113. J. T. Bridgham, J. E. Brown, A. Rodríguez-Marí, J. M. Catchen, J. W. Thornton, Evolution of a New Function by Degenerative Mutation in Cephalochordate Steroid Receptors. *PLOS Genetics* **4**, e1000191 (2008).

114. S. Liu *et al.*, Nonnatural protein–protein interaction-pair design by key residues grafting. *Proceedings of the National Academy of Sciences* **104**, 5330-5335 (2007).

115. D. P. Anderson *et al.*, Evolution of an ancient protein function involved in organized multicellularity in animals. *Elife* **5**, e10147 (2016).

116. L. Schulz *et al.*, Evolution of increased complexity and specificity at the dawn of form I Rubiscos. *Science* **378**, 155-160 (2022).

117. F. L. Sendker *et al.*, Emergence of fractal geometries in the evolution of a metabolic enzyme. *Nature* **628**, 894-900 (2024).

118. J. Yang, Z. Zhang, X. A. Zhang, Q. Luo, A ligation-independent cloning method using nicking DNA endonuclease. *Biotechniques* **49**, 817-821 (2010).

119. C. Natarajan *et al.*, Expression and purification of recombinant hemoglobin in Escherichia coli. *PLoS One* **6**, e20176 (2011).

120. M. T. Marty *et al.*, Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical chemistry* **87**, 4370-4376 (2015).

121. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261-272 (2020).

122. P. J. Carter, G. Winter, A. J. Wilkinson, A. R. Fersht, The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell* **38**, 835-840 (1984).

123. A. Horovitz, Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding and Design* **1**, R121-R126 (1996).

124. J. A. Wells, Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509-8517 (1990).

125. L. G. Nivón, R. Moretti, D. Baker, A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLOS ONE* **8**, e59004 (2013).

126. C.-J. Tsai, R. Nussinov, A Unified View of "How Allostery Works". *PLOS Computational Biology* **10**, e1003394 (2014).

127. J. Kuriyan, D. Eisenberg, The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983-990 (2007).

128. S. M. Coyle, J. Flores, W. A. Lim, Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell* **154**, 875-887 (2013).

129. A. S. Raman, K. I. White, R. Ranganathan, Origins of allostery and evolvability in proteins: a case study. *Cell* **166**, 468-480 (2016).

130. M. C. Nnamani *et al.*, A Derived Allosteric Switch Underlies the Evolution of Conditional Cooperativity between HOXA11 and FOXO1. *Cell Rep* **15**, 2097-2108 (2016).

131. R. N. McLaughlin Jr, F. J. Poelwijk, A. Raman, W. S. Gosal, R. Ranganathan, The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142 (2012).

132. M. Schupfner, K. Straub, F. Busch, R. Merkl, R. Sterner, Analysis of allosteric communication in a multienzyme complex by ancestral sequence reconstruction. *Proceedings of the National Academy of Sciences* **117**, 346-354 (2020).

133. A. J. Costello, W. E. Marshall, A. Omachi, T. O. Henderson, ATP binding to human hemoglobin in the presence and absence of magnesium ions investigated with 31P NMR spectroscopy and ultrafiltration. *Biochim Biophys Acta* **491**, 469-472 (1977).

134. D. S. Gottfried *et al.*, Probing the Hemoglobin Central Cavity by Direct Quantification of Effector Binding Using Fluorescence Lifetime Methods*. *Journal of Biological Chemistry* **272**, 1571-1578 (1997).

135. C. Natarajan *et al.*, Evolution and molecular basis of a novel allosteric property of crocodilian hemoglobin. *Current Biology* **33**, 98-108.e104 (2023).

136. R. E. Hirsch, R. S. Zukin, R. L. Nagel, Intrinsic fluorescence emission of intact oxy hemoglobins. *Biochem Biophys Res Commun* **93**, 432-439 (1980).

137. R. Hirsch, R. Nagel, Conformational studies of hemoglobins using intrinsic fluorescence measurements. *The Journal of biological chemistry* **256**, 1080-1083 (1981).

138. R. E. Hirsch, Hemoglobin fluorescence. *Methods Mol Med* **82**, 133-154 (2003).

139. T. R. M. Barends *et al.*, Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science* **350**, 445-450 (2015).

140. M. F. Perutz, Stereochemistry of cooperative effects in haemoglobin: haem–haem interaction and the problem of allostery. *Nature* **228**, 726-734 (1970).

141. B. Vallone, A. Nienhaus K Fau - Matthes, M. Matthes A Fau - Brunori, G. U. Brunori M Fau - Nienhaus, G. U. Nienhaus, The structure of carbonmonoxy neuroglobin reveals a heme-sliding mechanism for control of ligand affinity.

142. J. Chen, Y. L. Vishweshwaraiah, N. V. Dokholyan, Design and engineering of allosteric communications in proteins. *Curr Opin Struct Biol* **73**, 102334 (2022).

143. V. A. Feher, J. D. Durrant, A. T. Van Wart, R. E. Amaro, Computational approaches to mapping allosteric pathways. *Curr Opin Struct Biol* **25**, 98-103 (2014).

144. G. Schay *et al.*, Dissimilar flexibility of α and β subunits of human adult hemoglobin influences the protein dynamics and its alteration induced by allosteric effectors. *PLoS One* **13**, e0194994 (2018).

145. M. F. Perutz, Mechanisms of cooperativity and allosteric regulation in proteins. *Quarterly Reviews of Biophysics* **22**, 139-237 (1989).

146. J. S. Fraser *et al.*, Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669-673 (2009).

147. A. M. Damry, M. M. Mayer, A. Broom, N. K. Goto, R. A. Chica, Origin of conformational dynamics in a globular protein. *Communications Biology* **2**, 433 (2019).

148. K. A. Crowhurst, S. L. Mayo, NMR-detected conformational exchange observed in a computationally designed variant of protein Gβ1. *Protein Engineering, Design and Selection* **21**, 577-587 (2008).

149. A. J. Faure *et al.*, Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175-183 (2022).

150. A. Papkou, L. Garcia-Pastor, J. A. Escudero, A. Wagner, A rugged yet easily navigable fitness landscape. *Science* **382**, eadh3860.

151. H. A. Heaslet, W. E. Royer, 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**, 517-526 (1999).

152. S. Kundu, M. S. Hargrove, Distal heme pocket regulation of ligand binding and stability in soybean leghemoglobin. *Proteins* **50**, 239-248 (2003).

153. W. E. Royer, Jr., W. A. Hendrickson, E. Chiancone, Structural transitions upon ligand binding in a cooperative dimeric hemoglobin. *Science* **249**, 518-521 (1990).

154. H. Sugimoto *et al.*, Structural Basis of Human Cytoglobin for Ligand Binding. *Journal of Molecular Biology* **339**, 873-885 (2004).

155. K. Gunasekaran, B. Ma, R. Nussinov, Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics* **57**, 433-443 (2004).

156. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307-1320 (2008).

157. R. Elison Hirsch, "[12] Front-face fluorescence spectroscopy of hemoglobins" in Methods in Enzymology. (Academic Press, 1994), vol. 232, pp. 231-246.

158. G. M. Süel, S. W. Lockless, M. A. Wall, R. Ranganathan, Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology* **10**, 59-69 (2003).

159. H. N. Motlagh, J. O. Wrabl, J. Li, V. J. Hilser, The ensemble nature of allostery. *Nature* **508**, 331-339 (2014).

160. S. J. Wodak *et al.*, Allostery in its many disguises: from theory to applications. *Structure* (2019).

161. Y. Park, B. P. H. Metzger, J. W. Thornton, Epistatic drift causes gradual decay of predictability in protein evolution. *Science* **376**, 823-830 (2022).

162. G. Chure *et al.*, Predictive shifts in free energy couple mutations to their phenotypic consequences. *Proceedings of the National Academy of Sciences* **116**, 18275-18284 (2019).

163. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138**, 774-786 (2009).

164. S. W. Lockless, R. Ranganathan, Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **286**, 295-299 (1999).

165. M. Leander, Y. Yuan, A. Meger, Q. Cui, S. Raman, Functional plasticity and evolutionary adaptation of allosteric regulation. *Proceedings of the National Academy of Sciences* **117**, 25445-25454 (2020).

166. S. Ekambaram, G. Arakelov, N. V. Dokholyan, The Evolving Landscape of Protein Allostery: From Computational and Experimental Perspectives. *Journal of Molecular Biology*, 169060 (2025).

167. L. Swint-Kruse, K. S. Matthews, Allostery in the LacI/GalR family: variations on a theme. *Current Opinion in Microbiology* **12**, 129-137 (2009).

168. M. D. Daily, J. J. Gray, Local motions in a benchmark of allosteric proteins. *Proteins: Structure, Function, and Bioinformatics* **67**, 385-399 (2007).

169. A. J. Morrison, D. R. Wonderlick, M. J. Harms, Ensemble epistasis: thermodynamic origins of nonadditivity between mutations. *Genetics* **219**, iyab105 (2021).

170. B. D. Huisman, Z. Dai, D. K. Gifford, M. E. Birnbaum, A high-throughput yeast display approach to profile pathogen proteomes for MHC-II binding. *eLife* **11**, e78589 (2022).

171. J. A. Purslow, B. Khatiwada, M. J. Bayro, V. Venditti, NMR Methods for Structural Characterization of Protein-Protein Complexes. *Frontiers in Molecular Biosciences* **7** (2020).

172. A. J. Faure, B. Lehner, MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biology* **25**, 303 (2024).

173. C. Weng, A. J. Faure, A. Escobedo, B. Lehner, The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643-652 (2024).