# EVOLUTION OF METAL AND PEPTIDE BINDING IN THE S100 PROTEIN FAMILY

by

LUCAS CLAYTON WHEELER

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2017

DISSERTATION APPROVAL PAGE

Student: Lucas Clayton Wheeler

Title: Evolution of Metal and Peptide Binding in the S100 Protein Family

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Chemistry and Biochemistry by:

| | |
|---|---|
| James Prell | Chair |
| Michael Harms | Advisor |
| Bradley Nolen | Core Member |
| Patrick Phillips | Core Member |
| Alice Barkan | Institutional Representative |

and

| | |
|---|---|
| Sara D. Hodges | Interim Vice Provost and Dean of the Graduate School |

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2017

DISSERTATION ABSTRACT

Lucas Clayton Wheeler

Doctor of Philosophy

Department of Chemistry and Biochemistry

December 2017

Title: Evolution of Metal and Peptide Binding in the S100 Protein Family

Proteins perform an incredible array of functions facilitated by a diverse set of biochemical properties. Changing these properties is an essential molecular mechanism of evolutionary change, with major questions in protein evolution surrounding this topic. How do new functional biochemical features evolve? How do proteins change following gene duplication events? I used the S100 protein family as a model to probe these aspects of protein evolution. The S100s are signaling proteins that play a diverse range of biological roles binding Calcium ions, transition metal ions, and other proteins. Calcium drives a conformational change allowing S100s to bind to diverse peptide regions of target proteins. I used a phylogenetic approach to understand the evolution of these diverse biochemical features. Chapter I comprises an introduction to the disseration. Chapter II is a co-authored literature review assessing available evidence for global trends in protein evolution. Chapter III describes mapping of transition metal binding onto a maximum likelihood S100 phylogeny. Transition metal binding sites and metal-driven structural changes are a conserved, ancestral features of the S100s. However, they are highly labile at the amino acid level. Chapter IV further

characterizes the biophysics of metal binding in the S100A5 lineage, revealing that the oft–cited $Ca^{2+}/Cu^{2+}$ antagonism of S100A5 is likely due to an experimental artifact of previous studies. Chapter V uses the S100 family to investigate the evolution of binding specificity. Binding specificity for a small set of peptides in the duplicate S100A5 and S100A6 clades. Ancestral sequence reconstruction reveals a pattern of clade-level conservation and apparent subfunctionalization along both lineages. In chapter VI, peptide phage display, deep-sequencing, and machine-learning are combined to quantitatively reconstruct the evolution of specificity in S100A5 and S100A6. S100A5 has subfunctionalized from the ancestor, while S100A6 specificity has shifted. The importance of unbiased approaches to measure specificity are discussed. This work highlights the lability of conserved functions at the biochemical level, and measures changes in specificity following gene duplication. Chapter VII summarizes the results of the dissertation, considers the implications of these results, and discusses limitations and future directions.

This dissertation includes both previously published/unpublished and co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Lucas Clayton Wheeler

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Montana State University, Bozeman, MT

DEGREES AWARDED:

Doctor of Philosophy, Chemistry, 2017, University of Oregon
Bachelor of Science, Biochemistry, 2012, Montana State University

AREAS OF SPECIAL INTEREST:

Evolutionary biochemistry
Biophysics
Evolutionary biology

PROFESSIONAL EXPERIENCE:

PhD candidate & Graduate Research Fellow, University of Oregon, 2014-2017

PhD student & Graduate Teaching Fellow, University of Oregon, 2012-2013

Undergraduate Reseach Assistant, Montana State University, 2009-2012

PUBLICATIONS:

**Wheeler LC**, Harms MJ (2017). Increased peptide binding specificity for an S100 protein over evolutionary time (in prep)

**Wheeler LC**, Anderson JA, Morrison AJ, Wong CE, Harms MJ (2017). Conservation of specificity in two low specificity proteins (in review)

**Wheeler LC**, Harms MJ (2017). Human S100A5 binds $Ca^{2+}$ and $Cu^{2+}$ independently BMC Biophysics (in press)

**Wheeler LC**, Donor MT, Prell JS, Harms MJ (2016). Multiple Evolutionary Origins of Ubiquitous $Cu^{2+}$ and $Zn^{2+}$ Binding in the S100 protein Family. PLoS ONE 11(10): e0164740. doi:10.1371/journal.pone.0164740

**Wheeler LC**, An-Lim S, Marqusee S, Harms MJ (2016). The thermostability and specificity of ancient proteins. Curr Op Struct Biol. (LCW and SAL contributed equally to the work)

# ACKNOWLEDGEMENTS

I thank my advisor, Mike Harms, for his excellent mentorship and his patience. He has pushed me to wrestle with new concepts and given me the opportunity to learn a diverse set of skills. I also thank the members of the Harms group for their constructive feedback, helpful conversations, and friendship throughout my time in the lab. I would especially like to thank Zach Sailer for his friendship and for always being willing to help me wrestle with new ideas. Thanks are also in order for the members of the Beer and Theory Society, who have devoted their time weekly, over the course of several years, to creating an excellent environment for learning math and physics that we all wish we'd done a better job of learning in college. In the same vein, I thank the members of the Quantitative Problem Solving and Research Communication Consortium for their devotion to helping peers solve challenging problems and representing the ideas of open science and collaboration. Thanks are in order for my long-time roommates Adam and Forrest as well as for the members of my trivia team. We've had a lot of good times during my years in Eugene. My friend Stacey Wagner has been very helpful to me over the years and I appreciate her willingness to get together and trade advice. The research in this dissertation was supported in part by a grant, R01GM117140, from the National Institutes of Health, and I was personally supported by NIH training grant T32 GM007759 for three years of my PhD. I thank my committee for their assistance throughout my PhD work and with the preparation of this document. Special thanks go to Jim Prell, who has been an excellent committee chair as well a collaborator. I have been inspired by his knowledge, enthusiasm, and strong distaste for false dichotomies. Furthermore, I thank Jim Prell and Alice

Barkan for their assistance in applying for postdoctoral positions. I also thank Micah Donor, Shion An-Lim, and Susan Marqusee with whom I have collaborated. I thank Doug Turnbull and Maggie Weitzman in GC3F for their help with next-generation sequencing. I thank Carol Higginbotham at COCC and Ed LaChapelle at Bend Research, Inc. for encouraging me to pursue a scientific career very early in my life. I thank my undergraduate research advisor, Trevor Douglas, for helping me to develop a firm footing in biochemistry, molecular biology, and experimental design before starting graduate school. I would like to thank my good friend Ryan Russel Allen for his friendship, support, and steady stream of humorous correspondence throughout my undergraduate and graduate careers. Finally I would like to thank my family and my girlfriend for their constant support and love.

For my mother Terry, my father Ed, my sister April, my girlfriend Rutendo, and my consigliere Tucker, without all of whom I could never have maintained enough sanity to finish my PhD.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# CHAPTER I

# INTRODUCTION

## *Evolution is the Driving Force of Biological Diversity*

One of the most striking aspects of life is the vast diversity of forms and functions displayed by living things. Organisms are beautifully adapted to a broad range of environments and life styles; from bacteria that thrive in deep see thermal vents [1], to plants that live in high alpine meadows [2], to exquisitely colorful poisonous frogs that roam the rainforest [3]. With such diversity on display it is easy enough to forget that all organisms on Earth share a common ancestor in the distant past [4, 5, 6, 7]. Over unfathomable stretches of time, life has diversified from that common ancestor into the amazingly complex and dynamic biosphere that is familiar to us today. Perhaps the most incredible facet of this diversity is that it is produced via the stochastic process of evolution [4, 8, 9, 10, 11]. How this random process generates the rich biology observed on Earth is the primary driving question of evolutionary biology.

Evolution occurs via the change of heritable traits over the course of many generations of organisms. The process acts on traits that are displayed in some way at the macroscopic, organismal level [8, 12, 13, 14, 15]. However, at the heart of trait heritability are the genes that encode traits at the genetic level. The projection of underlying genetics into phenotypes is commonly referred to as the genotype-phenotype map [16, 17]. Evolutionary processes such as natural selection and genetic drift drive the fixation of mutant genes that lead to new traits in populations [18, 13, 12, 19]. Over time this fixation process can lead to substantial

changes in the genetic makeup of a population of organisms and result in the formation of new species with different organism-level traits [20, 21, 22, 23, 14].

Most functionally-important genes encode proteins, which are the workhorses of molecular processes in living organisms [24, 25]. They catalyze chemical reactions, form the basis of structural scaffolds, transport ions and small molecules, act as signals, and regulate the function and production of other molecules [24, 25]. The emergent outcome of all the intertwining protein roles is ultimately manifested in the macroscopic phenotype of an organism. The vast array of protein functions necessary to construct organisms requires a large diversity of proteins, all of which are encoded by genes and integrated into the broader system. This framework imposes an extremely complex set of constraints that govern the way in which new traits—that can be seen by evolution—can be achieved. Thus, a molecular-level understanding of evolution is critical to understanding the process on larger scales.

*Evolutionary Biochemistry is a Powerful Tool for Understanding Biology*

Most studies in traditional evolutionary biology have focused on understanding the genetics of evolution. Genetics provides a very useful tool to dissect the basis of evolutionary logic at the level of encoding architecture. A genetic framework has also been critical for developing a population-level understanding of evolution and creating useful mathematical models of evolutionary processes [14, 15, 26, 27, 18, 13, 28, 12, 19, 29]. These models have made it possible to make predictions that can be tested experimentally, thus furthering our ability to understand evolutionary dynamics and outcomes [30, 31, 32, 33, 34, 35]. However, the rules that ultimately govern the inner workings of biological organisms are those of physics and chemistry [36, 37, 38, 39, 40, 41].

Although the phenomenonological genetic "laws" that govern evolutionary processes are now largely understood, it is unclear how they are connected to the physical laws that govern the universe. This disconnect is one of the most prominent barriers to understanding molecular evolution. To truly understand how evolution works at the molecular level this relationship must be determined. The need to understand how physicochemical principles shape evolutionary outcomes has spawned the field of evolutionary biochemistry. This field seeks to understand the evolution of molecular phenotypes at the biochemical level and to relate the molecular phenotypes to implications for evolution at larger scales [42, 43, 44]. Many pressing questions remain unanswered. Are there general evolutionary trends in biochemical features over very long time scales? How robust are protein functions to alterations in amino acid coding sequence and how does this robustness affect the maintenance of important traits during evolution? How do protein copies evolve after they are generated by gene duplication events? How do correlations between mutations in protein sequences shape evolutionary possibilities? Can we understand large-scale evolutionary processes in terms of simpler molecular-level constraints and rules? These questions are unified by the broader inquiry: how do physical rules shape the genotype-phenotype map?

The field of evolutionary biochemistry has rapidly expanded since its inception and provided a great deal of insight into evolution at the molecular level. A critical workhorse of evolutionary biochemistry has been ancestral sequence reconstruction (ASR) [45, 43, 46, 47]. ASR is a statistical technique that utilizes a molecular phylogeny to infer the sequences of ancestral nodes [48, 49, 43, 50]. This technique has allowed many researches to directly assess ancestral protein activities using biochemical experiments, making it extremely

powerful for characterizing evolutionary history [44, 43, 46, 47]. In some cases, entire evolutionary trajectories—composed of historical substitutions—have been reconstructed [51, 52, 53]. Relationships between protein structure, function, and evolutionary history have been characterized for a wide variety of proteins [53, 54, 55, 52, 56, 46, 57, 58, 59, 60]. Much has been learned about how biochemistry and biophysics constrain and shape protein evolution. Furthermore, evolutionary approaches have been used—with great success—to winnow the substitutions observed in extant proteins down to those that are important for a given biochemical function. For example, ASR was used to identify residues that are important for binding selectivity of the drug Gleevec by Ab1 and Src kinases [59].

Detailed biochemical studies have also helped to clarify the importance of phenomena such as epistasis—the non-additivity of mutations [46, 32, 61, 62]—and pleiotropy—in which proteins have roles in multiple distinct biological processes [63, 64]—in determining evolutionary outcomes. These effects can reduce the evolutionary degrees of freedom allowed for a protein and result in effects such as historical contingency [65, 46]. For example, to evolve specificity for a new hormone ligand the glucocorticoid receptor required a permissive subtitution that alleviated the results of an otherwise deleterious functional substitution in the ligand binding site [46]. Studies that incorporate biochemical and functional work have futher demonstrated that the broader systemic architecture of the cellular environment can constrain the mechanisms by which biochemical changes underly organismal phenotypes [66, 67, 68, 69]. Certain systems have far greater constraints on the allowed biochemical changes. For example, the evolution of new flower colors in plants often requires both functional amino acid substitutions and regulatory

changes, but the genes that are subject to these different types of changes vary depending on pleiotropic consequences [70, 71, 72, 73].

Evolutionary biochemistry has provided great insight into the molecular mechanisms of evolution. However, there is a key limitation that is prevalent in most previous work. Evolutionary biochemical studies have focused almost exclusively on proteins that exhibit very rigidly defined biochemical features. For example, the evolution of binding specificity has largely been studied in proteins such as enzymes and transcription factors that exhibit exquisite binding specificity for targets [74, 75, 76, 77]. These studies have revealed key patterns in the evolution of specificity, such as consistent occurrence of subfunctionlization and neofunctionalization following gene duplications [53, 58, 52, 78, 79]. Similarly, studies of proteins binding to other biologically-relevant targets such as metal ions have traditionally considered very well-defined coordination systems, like those found in metalloproteases and Zinc finger proteins [80]. However, many proteins do not exhibit such exquisite such exquisite biochemical properties [81, 82, 83, 84, 85, 86, 87, 88]. A large number of proteins bind to targets with low specificity and limited binding-site conservation. The biological relevance of the biochemical properties of these proteins is less well understood. It is thus unclear how well evolutionary studies of typical protein model systems translate to the broad array of proteins with plastic biochemical properties.

### *The S100 Protein Family is a Useful Model System to Probe the Evolution of Low-specificity Proteins*

This dissertation focuses on case studies in evolutionary biochemistry that address unanswered questions in molecular evolution. Chapter II consists of a

literature review addressing the evidence for global trends in protein evolution over very long time scales. The remaining studies are unified by questions surrounding the evolution of protein-target interactions in proteins that have labile binding interfaces and/or highly-variable binding partners. Each case study dissects a specific aspect of evolution at the molecular level. The studies use a combination of experiments and computational analysis methods to address how biochemistry relates to broader questions in evolutionary biology.

Chapters III, IV, and V of this dissertation make extensive use of the S100 proteins as an experimental model system, which warrants an introduction to the protein family. The S100s are a large family of small, calcium-dependent signaling proteins [89, 90, 91, 92, 93]. The proteins are generally homodimeric and transduce signals via a calcium-ion driven conformational change [94, 89]. The family originated at the base of the Metazoan lineage and subsequently diversified over several hundred million years [95, 91, 96]. Mammals possess approximately thirty S100 genes including those encoding fusion proteins, in which the S100 acts as a single domain inside a larger domain architecture [91, 96, 97, 98, 99]. S100 proteins play a wide array of biological roles inside and outside of cells; including inflammatory signaling [100, 101], regulation of cell proliferation [102, 103, 104], antimicrobial activity [105, 106], and control of apoptosis in some cell types [107, 108]. The diversity of functions performed by the S100s is perhaps surprising considering the small size of the proteins, overall similarity of S100 amino acid sequences, and conservation of the folded form. However, the proteins have evolved an array of useful biochemical features that aid in carrying out biological functions. The proteins possess the ability to bind both calcium ions and other metal ions. Calcium-induced conformational changes result in the exposure of a

hydrophobic path on the S100 dimer surface, which facilitates binding of target proteins [94, 93]. The specificity of these hydrophobic binding sites varies among members of this family, although it has not been systematically studied prior to the work in this dissertation [93, 109]. It is sometimes presumed that this biochemical specificity contributes to the biological specialization of the S100s [93]. This notion is supported by the fact that only some S100s are capable of binding to certain target proteins. For example, many S100s bind to and activate the inflammatory RAGE protein, but this not a universal trait of the family [101, 104]. However, it has also been proposed that most functional specificity of S100 proteins is acheived by control of differential expression [89, 90, 100].

The biological importance of binding to metal ions other than calcium—which occurs at different ion binding sites—has not been well studied. This dissertation primarily uses the S100s as a model to address evolutionary questions, because they possess a rich evolutionary history, diverse biochemical features, and exhibit low specificity for interaction partners. However, the evolutionary biochemical work presented in chapters III, IV, and V also sheds light on biologically-relevant aspects of the S100 family. The work provides several opportunities and resources for more biologically-oriented future studies.

### *Chapter-by-chapter Breakdown of Dissertation*

Chapter II comprises a literature review—co-written with Shion An Lim (SAL), Susan Marqusee (SM), and my advisor Michael J. Harms (MJH). The review addresses the question of whether or not proteins display global evolutionary trends over very long time scales. Two case studies are used as key examples of hypothesized trends: the gradual reduction of protein thermostability due to

cooling of the Earth and the gradual increase in protein binding specificity due to continued specialization in ever-more-complex proteomes. Based on a thorough summary of evolutionary biochemistry literature, there does in fact appear to be some evidence for a gradual decline of thermostability on the billions-of-years time scale. However, there are still relatively few studies that probe this question. A more substantial body of evidence will need to be accumulated to make a strong argument for a global trend. The need for more experimental evidence is even more pronounced with regard to the question of broad trends in specificity. There are few studies that have addressed this question directly, and none to date that have done so using a truly unbiased experimental approach. This chapter provides an overview of the idea that there are global trends in protein evolution, makes a strong case that further experimental studies are needed to resolve the ongoing debates on this topic, and suggests strategies and experiments to maximize the current understanding in the field. The literature review presented in this chapter was published in the journal Current Opinions in Structural Biology [47].

Chapter III probes the evolutionary lability of a biologically-important biochemical feature. The S100 protein family is used as a model system to address this question. The phylogenetic history of the S100s is reconstructed to yield the highest-quality phylogeny of the S100 family to date. The history of transition metal binding in the S100s is then traced by mapping the results of detailed in vitro measurements of metal-ion binding onto this high-quality phylogeny. These results show that binding of transition metals is conserved across almost the entire S100 family, a more universal result than any previous study. By using mutagenesis studies it is further established that not all S100 proteins use the same amino acids or even the same site to bind metal ions. The binding of metal ions to a very early

branching S100 protein is measured for the first time, which demonstrates that binding of transition metals is an ancestral feature of the S100 protein family. The results of this chapter speak to the surprising level of lability—at the amino acid level—of S100 protein metal binding sites; highlighting the fact that an ancestral molecular phenotype can be maintained at the overall level of behavior even while the underlying biochemical basis fluctuates over evolutionary time. The work in this chapter has been published as a research article in PLoS One, co-authored with Micah T. Donor (MTD), James S. Prell (JSP), and Michael J. Harms (LC Wheeler is the first author) [96].

Chapter IV delves further into the biophysics of metal binding in one particular member of the S100 protein family, S100A5. Little is known about the biological roles of S100A5. A previous publication indicated that the protein exhibits antagonism between the binding of $Ca^{2+}$ and $Cu^{2+}$ ions. This feature is unique amongst S100 proteins and has been considered one of the key features of S100A5. Proposed biological roles for the protein typically involve $Ca^{2+}/Cu^{2+}$ antagonism. In chapter IV, it is demonstrated that antagonism between the binding of $Ca^{2+}$ and $Cu^{2+}$ is likely an artifact of the experiments done in the original study. Instead, it is shown that S100A5 can bind $Ca^{2+}$ and $Cu^{2+}$ independently, which changes the biological implications of metal binding to the protein. Furthermore, this chapter adds to the evolutionary story of metal binding by demonstrating another unique biochemical modification that has evolved in the S100 family. The work in this paper is currently in press as a research article in the journal BMC Biophysics, co-authored with Michael J. Harms.

Chapters V and VI address the evolution of binding specificity in two proteins following gene duplication from a common ancestor. Again, the S100 protein

family proves to be a useful model system to address this question. The proteins S100A5 and S100A6 arose from a duplication approximately 300 million years ago. They subsequently evolved to have different protein-binding specificity, distinct expression patterns, and perform different cellular roles. Despite having distinct specificity, both proteins can be described as sloppy or having very low biochemical specificity. Previous studies have addressing the question of evolving specificity have used highly-specific proteins and small sets of known binding partners that are biased by a priori knowledge. For these reasons, previous studies are limited in understanding the evolution of specificity in low-specificity proteins. The sloppiness of S100s makes them an excellent system to study how binding specificity evolves in an inherently noisy low-specificity system.

Chapter V comprises a biochemical study of the evolution of peptide binding specificity in the S100A5-S100A5 clade. The oldest ancestor of S100A5 and S100A6 is resurrected using ancestral sequence reconstruction (ASR). Detailed calorimetric measurements of binding to a small set of peptide targets are then used to compare specificity across a set of orthologous and paralagous S100A5 and S100A6 proteins. It is demonstrated that peptide binding is driven primarily by the hydrophobic effect and that specificity is readily changed by the addition of mutants into the peptide binding interface. Furthermore, this work reveals that the specificity of S100A5 and S100A6 have undergone an apparent pattern of subfunctionilization. This result is striking, because it demonstrates that proteins with very low biochemical specificity can undergo similar patterns of evolution to proteins with high specificity. The work in this chapter is in review as a research article in the journal Biochemistry, co-authored with Jeremy A. Anderson (JAA),

Anneliese J. Morrison (AJM), Caitlyn E. Wong (CEW), and Michael J. Harms. The submitted article has also been uploaded to the preprint server BioArxiv [110].

Chapter VI introduces new experimental and analysis pipelines for studying the evolution of specificity. An unbiased high-throughput approach, incorporating phage display and deep sequencing, is used to measure the binding of a large random peptide library to human S100A5, human S100A6, and the last common ancestor. Strikingly, the pipeline uncovers the lack of sequence-based rules that govern binding preferences of the S100 proteins. Instead, preferences appear to be defined by general physicochemical features of the peptide targets that can be used to generate a predictive model. The pipeline reveals overall patterns in the evolution of specificity along the S100A5 and S100A6 lineages. S100A5 exhibits a strong signal of subfunctionilization, while S100A6 appears to differ little from the ancestor. This chapter highlights the importance of using unbiased approaches to study the evolution of specificity and speaks to the necessity of understanding different classes of protein features when probing molecular evolution. The work in this paper is being prepared as a research article that will be submitted to the journal MBE, co-authored with Michael J. Harms.

### *Broader Impacts*

The studies described in this dissertation contribute the broader evolutionary biochemistry literature by addressing a set of topics that have remained ambiguous. There has been a lack of studies addressing the evolution of biochemical features in proteins that have highly diverse sets of binding partners. Much of the experimental basis for understanding evolution of protein binding specificity has instead been based on proteins with exquisite specificity profiles [74, 111]. For

example, enzymes, receptors, and transcription factors that have well-defined chemical binding preferences are workhorses of evolutionary biochemistry studies [46, 53, 52, 58]. The work presented in the following chapters probes key aspects of the evolution of binding specificity in proteins without such obvious rules. The S100 proteins act as an excellent model system to tackle these problems, because they have a variety of conserved biochemical behaviors that have nontheless been labile at the amino acid level during diversification of the family [96]. In particular, the ability of the S100s to bind to a variety of transition metals with similar affinities, and the ability to bind extremely diverse short peptide regions of target proteins are used as exemplary biochemical features.

Studies on both the binding of metal ions and peptides reveal several key evolutionary trends that speak to the evolution of biochemical features in sloppy proteins such as the S100s. 1) a biochemical output—such as binding of transition metals or peptides with moderate afffinity—can be acheived and conserved despite extensive variability in amino acid ligands that form binding sites. 2) Specificity can nontheless be achieved and conserved in proteins with highly diverse binding partners and labile binding sites. 3) Evolutionary patterns in proteins with low biochemical specificity nontheless resemble those observed in high-specificity proteins. 4) Evolutionary patterns can differ along duplicate lineages following gene duplication. 5) Unbiased high-throughput techniques are essential for inferring historical patterns of specificity in proteins with large diverse sets of binding partners. These observations contribute substantially to our understanding of what types of biochemical features are important during the evolution of proteins that do not meet the criterion of exquisite binding specificity. Despite relaxed binding rules, flexible binding sites, and highly-diverse binding partners these

proteins nonetheless exhibit evolutionary patterns that are reminiscent of those the field has come to expect from canonical examples. This key result suggests that proteins such as the S100s—despite the variability of there biochemical behaviors—are therefore operating under similar rules to other proteins. Therefore, proteins with highly variable binding partners and labile binding sites do not necessarily represent a fundamentally different class of proteins—subject to special evolutionary constraints—but rather are similarly constrained by evolutionary and biochemical forces in a way that can be understood by careful experimentation.

# CHAPTER II

# THERMOSTABILITY AND SPECIFICITY OF ANCIENT PROTEINS: ASSESSING THE EVIDENCE FOR GLOBAL TRENDS

## Author Contributions

Lucas Wheeler, Shion An-Lim, Michael Harms, and Susan Marqusee conceptualized the review and chose specific topics. Lucas Wheeler and Shion An-Lim conducted the literature review. MJH and SM administered the project. Michael Harms and Lucas Wheeler generated figures. Lucas Wheeler, Shion An-Lim, Michael Harms, and Susan Marqusee wrote and edited the manuscript.

## Abstract

Were ancient proteins systematically different than modern proteins? The answer to this question is profoundly important, shaping how we understand the origins of protein biochemical, biophysical, and functional properties. Ancestral sequence reconstruction (ASR), a phylogenetic approach to infer the sequences of ancestral proteins, may reveal such trends. We discuss two proposed trends: a transition from higher to lower thermostability and a tendency for proteins to acquire higher specificity over time. We review the evidence for elevated ancestral thermostability and discuss its possible origins in a changing environmental temperature and/or reconstruction bias. We also conclude that there is, as yet, insufficient data to support a trend from promiscuity to specificity. Finally, we propose future work to understand these proposed evolutionary trends.

## Introduction

Ancestral sequence reconstruction (ASR) has opened a window into the sequences and properties of ancient proteins [45, 44]. In ASR, a multiple sequence alignment of modern protein sequences is used to construct a phylogenetic tree and the sequences of ancient proteins are inferred for specific ancestors on this tree (Figure 1a). By synthesizing the genes encoding these sequences, these reconstructed ancient proteins can be experimentally characterized. This approach has yielded an explosion of results in recent years, revealing important mechanistic insights into the evolution of protein forms and functions [112, 113, 51, 114, 115, 116, 117, 118, 60, 56].

One intriguing possibility is to use ASR to investigate whether ancient proteins were systematically different in the past, leading to parallel, directional changes in properties over evolutionary time (Figure 1a). Such trends are inaccessible using comparisons between modern pro- teins. For example, studies of the modern proteins in Figure 1b would lead one to believe the last common ancestor had a 'blue' trait. By allowing direct measurement of ancestral properties, ASR can reveal properties ('red', in this case) not evident in the modern proteins.

If the evolution of protein properties were directional, it would provide a new level at which to explain and understand these properties. This is of deep interest to evolutionary biochemists seeking to identify the general principles that shape protein evolution. Further, a trend could mean that sampling evolutionary history would provide access to qualitatively different proteins [119] — a boon to engineers looking for proteins with desirable properties as templates for further engineering [120, 121].

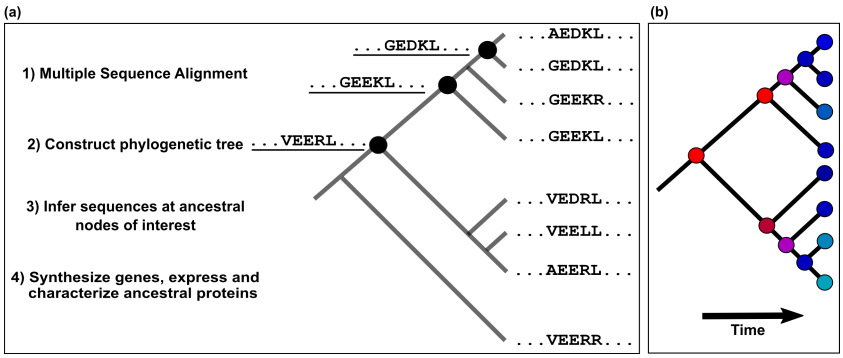


FIGURE 1 Ancestral Sequence Reconstruction (ASR) can be used to trace the history of evolving proteins. (a) The ASR pipeline. A multiple sequence alignment (MSA) of extant sequences of a protein family is generated using an alignment tool. The MSA is then used to estimate an appropriate model of sequence evolution and to estimate a phylogenetic tree. The sequences at ancestral nodes of interest (filled black circles) are then inferred (underlined) based on the tree and a phylogenetic evolutionary model. The maximum likelihood sequences are those with the highest likelihood of generating the known sequences of modern proteins given the tree and phylogenetic model. Genes encoding the inferred ancestral proteins can be synthesized, expressed, and purified using standard molecular biology tools. The properties of the ancestral proteins can then be experimentally characterized. (b) A phylogenetic tree showing the evolution of a protein that can vary between two properties—red and blue. The last common ancestor was red, but the modern proteins are blue because of parallel changes along the lineages. This red ancestor can only be accessed using an approach like ASR.

Recent work has suggested two trends over evolutionary time: decreasing protein stability [112] and increasing specificity [60]. Particularly for protein engineers, these trends could be extremely powerful, as high stability and broad substrate specificity are desirable traits that could be accessed using ASR. In this review, we review the evidence supporting and contradicting these trends, as well as the future work required to test and extend these conclusions.

## Reconstructed precambrian ancient proteins exhibit elevated thermostability

We begin by evaluating evidence from ASR studies that indicate the deepest ancestors of mesophilic proteins were highly thermostable. Over billion-year timescales, reconstructed ancestral proteins display systematically higher thermostability. Reconstructed EF-Tu [112], thioredoxin [118], DNA gyrase [116], nucleotide diphosphate kinase [115], and $\beta$-lactamase [60] all exhibit melting temperatures ($T_m$) far higher than their extant descendants. Some have argued that this is a universal trend [119] and have interpreted this as evidence for an ancient, hot environment [112]. The evidence, however, is not completely universal, as reconstructed RNase H along a mesophilic lineage gives a relatively flat trend in stability over similar time scales [56].

One difficulty in comparing these studies is that different proteins have different absolute requirements for stability. For example, the $T_m$'s of EF-Tu bacterial homologs are generally $\sim$2 °C above the environmental temperature ($T_{env}$), while the $T_m$s of RNase H are $\sim$30 °C above $T_{env}$. As a result, $T_m$s between protein families are not directly comparable. One way to overcome this challenge is to convert the measured $T_m$ of each protein to an estimate of $T_{env}$, as

$T_m$ often correlates with the growth temperature of the organism from which it was derived [122]. In most cases, this correlation arises to maintain stability above some critical threshold [123]. Empirically, $T_m$ generally rises by $\sim$1 °C per 1 °C of $T_{env}$, with an offset reflecting the required stability of the protein (e.g. 2 °C for EF-Tu and 30 °C for RNase H) [122]. This correlation has been directly established for three of the proteins above — EF-Tu, DNA gyrase and RNase H [112, 115, 56] — and holds generally for many other proteins [122].

When placed on the $T_{env}$ scale, reconstructed proteins report an elevated environmental temperature $\sim$3 billion years ago, though with significant scatter. Figure 2a shows the estimated $T_{env}$ over time for 17 ancestors of proteins found in the lineages leading to mesophilic *E. coli*. A total least-squares fit to the data reveals a highly significant negative slope that explains 75% of the variation in the data ($R^2 = 0.75$). In contrast, the estimated $T_{env}$ over time for ancestors leading to thermophile *T. thermophilus* exhibits a slope statistically indistinguishable from 0 ( Figure 2b). When taken in aggregate, these data support the hypothesis that the deepest ancestors had stabilities similar to proteins from modern thermophiles. While these data focus on the *E. coli* and *T. thermophilus* lineages, their deepest ancestors are shared both with each other and with most modern bacteria, thus suggesting a global transition away from ancient thermostability, at least along mesophilic lineages. It is not clear from these sparse, lineage-specific data whether mesophilicity evolved in parallel along many lineages or whether it evolved on a few key, early ancestral branches.

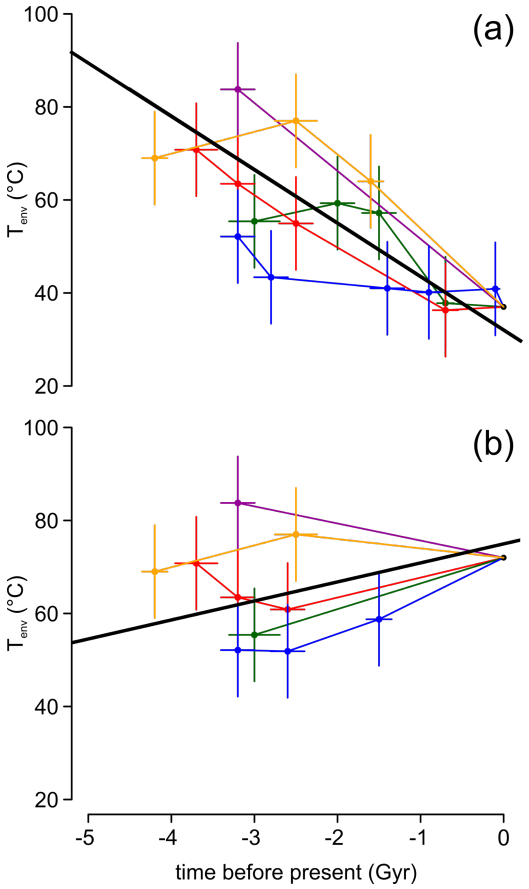


FIGURE 2 Ancient reconstructed ancestors exhibit elevated thermostability. Estimated environmental temperatures experienced by proteins on lineages leading to (panel a) *E. coli* or (panel b) *T. thermophilus*. Point/line series indicate individual protein families: EF-Tu (red), thioredoxin (orange), $\beta$–lactamase (green), RNase H (blue), and nucleotide diphosphate kinase (purple). Measured melting temperatures for ancestors that give rise to E. coli proteins were mined from published literature [112, 115, 118, 60, 56]. These were then converted to estimates of $T_{env}$ using measured relationships [112, 115, 56] or by adding an offset determined by the difference in $T_m$ and $T_{env}$ for the *E. coli* (panel a) or *T. thermophilus* homologs (panel B). Time estimates were drawn from original publications or estimated from Battistuzzi et al. [124]. Time errors are standard errors. $T_{env}$ standard errors were set to $\pm$ 10 °C to account for uncertainty in $T_m$ and the $T_m/T_{env}$ correlation. (This is a conservative estimate: when measured for NDK, RNase H, and EF-Tu [112, 115, 56] , the $T_m$ the standard error was <5 °C and the $T_m$ to $T_{env}$ variance was <5 °C.) Black line is a fit determined by total linear regression. To find the standard deviation of fit slopes, we generated 1000 pseudo datasets sampled from the time and $T_{env}$ uncertainties. For E. coli, the fits reject a slope = 0 ($p = 3 \times 10^{-8}$). For *T. thermophilus*, the fits fail to reject a zero slope ($p = 0.45$).

### Trends in Thermostability Are Complex

While ASR studies suggest that the most ancient proteins were highly thermostable, they do not support a smooth trend in thermostability over time. Ancestors exhibit extensive random scatter to the proposed trend. Such variation is expected as, over more recent timescales, protein stability fluctuates in response to neutral drift or adaptation in apparently random fashion [114, 117, 125, 126, 127]. The observed variation may also reflect uncertainty in the reconstruction, multiple heterogeneous environments experienced by ancient organisms, or uncertainty in the map between $T_m$ and $T_{env}$.

This scatter extends to the mechanism of stabilization. A recent study of the evolution of thermostability in RNase H revealed that the thermodynamic mechanism of stabilization for the ancestral proteins could fluctuate, even as the $T_m$s of the proteins varied smoothly [56]. This indicates that, even while under selection to maintain stability in a given environment, proteins are free to accumulate mutations to access alternate mechanisms of stabilization. Practically, studying multiple ancestors may reveal new sequence and thermodynamic determinants of stability. Although thermostability and the mechanism of stabilization appear to change independently for RNase H, the generality of this result for other proteins remains unknown.

Finally, these ASR studies generally used small, monomeric, and well-behaved proteins. Although such simple proteins may be representative of the first proteins to arise, studies on a greater diversity of protein families will reveal whether observed trends are applicable to the entire proteome.

### Can Reconstruction Errors Inflate Ancestral Thermostability?

While existing data are suggestive, further work must be done to test the hypothesis of ancient thermostability. The primary concern is that ancestral proteins are statistical reconstructions that cannot be directly verified. Even with good statistical support, it is unlikely that the reconstructed ancestor will have the exact sequence of the true ancestral state. Addressing and understanding this uncertainty will be critical for establishing or refuting the hypothesis that the earliest proteins were thermostable.

High stability is unlikely to arise from random errors in the reconstruction. To account for uncertainty, ASR studies have generated different versions of ancestral sequences to assess the robustness of the measured stability to phylogenetic errors. For example, Hart et al. measured ten alternate sequences of a ∼3 billion year-old ancestor and found a $T_m$ of 76.7 ± 2 °C (compared to 68.0 °C of RNase H from *E. coli*) [56]. Using such approaches, many sources of random error have been investigated: uncertain tree topology [112, 115, 128, 49], alternate evolutionary models [129], choice of reconstruction method [114, 49], different amino acid frequencies [112], and reconstruction ambiguity [112, 115, 119, 56, 130]. In all such studies, the properties of the ancestors have proven robust to uncertainty.

Of bigger concern are sources of systematic error in ASR — in particular, a bias towards elevated stability for deeper ancestors [131, 132, 133, 134]. Some have argued that ASR could be biased towards consensus sequences, which may lead to an increase in stability [132, 135, 136]. Simulations have also suggested that maximum likelihood (ML), the most popular form of ASR, may give rise to artificially elevated stability [131]. If different stabilizing mutations accumulate

along different lineages, ML may incorrectly incorporate all of the stabilizing mutations, creating an artificially stable ancestor. There is also concern that variable amino acid distributions and mutation rates can alter reconstructions [133, 134].

There have been some limited experimental tests of these computational predictions of bias. Comparisons between ancestral and consensus sequences have shown distinct statistical and functional properties [115, 116, 120, 137]. This suggests that any consensus bias that exists must be subtle. Other work has indirectly addressed this concern - the molecular basis of stability fluctuating over evolutionary time in the RNase H family is not consistent with bias arising from a single, convergent stabilization mechanism [56, 131].

Important experiments remain. One test would be a systematic comparison of ancestors reconstructed using both ML and an alternative, Bayesian, method. A Bayesian reconstruction averages over uncertainty; therefore, it is not expected to have the same stability bias as ML reconstructions [131]. Observing high thermostability in ancient Bayesian ancestors would be strong evidence that thermostability is not an artifact of the ML method. The experiment is not perfect, however, as Bayesian ancestors have more errors than ML ancestors as a result of incorporating uncertainty [49]. Because of this, they may not accurately reflect the ancestral state. For example, one study found that a Bayesian ancestor had fundamentally different folding properties than the ML ancestor or any modern protein in the family [114], consistent with a poor reconstruction.

Another test for bias would be to study the thermostability of reconstructed, recent ancestors of rapidly evolving proteins with known mesophilic ancestral environments. A rapidly evolving protein will accumulate similar amounts of

mutations relative to the deep ancestors studied to date, albeit on a much shorter timescale. If ML reconstructions lead to biased stability, we would predict that recent ancestors of rapidly evolving proteins would exhibit erroneously elevated stability.

### A Trend from Promiscuous to Specific is Not Yet Established

Another proposed trend is that proteins have, on average, changed from lower to higher specificity over deep evolutionary time [60, 119]. This stems from the idea that low specificity proteins — particularly enzymes — were important for the ability of primordial organisms to perform diverse chemical processes with a limited proteome [138] (Figure 3a). It is also well established that increased specificity often follows gene duplication via subfunctionalization from a multi-functional or promiscuous ancestral protein [139, 140] (Figure 3b). Given these considerations, proteins may, on average, increase in specificity over time.

To date, few attempts have been made to investigate the specificity of the deepest ancestors. One recent study found that an ancestral $\beta$-lactamase was both promiscuous and less efficient than its descendants [60]. Likewise, a study of RuBISCO found a promiscuous and inefficient ancestor, though this may be an artifact of poor reconstruction [141]. Other studies have determined the activities of ancient proteins, but not their specificity [114, 115]. On the basis of these data, it is difficult to make solid conclusions about specificity trends; more measurements of ancestral specificity are warranted.

The second model — gene duplication followed by subfunctionalization — could conceivably operate continuously through evolution, leading to progressively higher specificity proteins over all evolutionary timescales (Figure 3b). Studies of

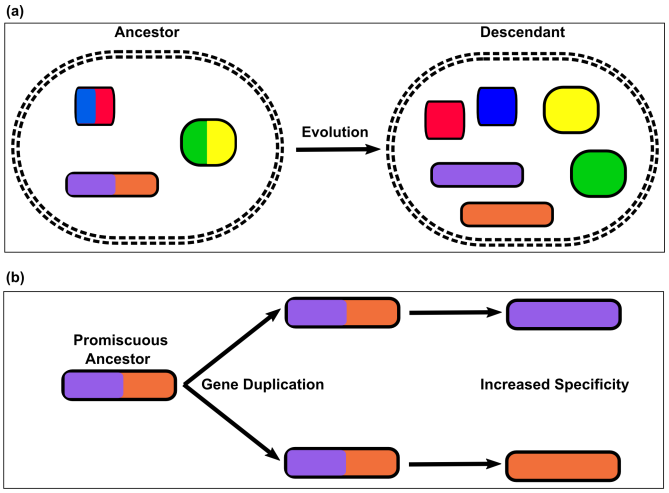


FIGURE 3 Models for increased specificity of proteins over time.(a) Large dotted ellipses denote cells. Small ellipses are proteins, colored by their specificity. Because early proteomes were presumably smaller than modern proteomes, it has been proposed that ancient proteins had to be promiscuous to achieve all the necessary chemistry. As organisms evolved, their proteomes expanded, allowing each protein to become more specific. (b) Higher specificity (subfunctionalization) is one of the possible outcomes of a gene duplication event. A gene encoding a low-specificity ancestral protein duplicates. Its descendants can then gain specificity and lose the promiscuous trait.

the evolution of specificity for ancestors from the last ∼500 million years suggest, however, that on average, proteins do not tend towards higher specificity over time. Some promiscuity-to-specificity transitions have been identified [60, 58, 53, 142, 59]. However, other studies have found switches between two high-specificity states [52, 55], evolution through a less-specific intermediate [143, 57, 79], and even decreased specificity over time [144].

This complexity likely arises because specificity is, at minimum, a bimolecular process that involves both the protein and its target. Further, constraints placed by the architecture of the larger system into which the proteins are embedded have been shown to shape specificity [79, 144, 145, 146, 147, 148, 149, 111]. For example, bioinformatic analyses have revealed that protein components of higher-complexity regulatory modules tend to possess lower specificity than those in simpler modules [150]. We therefore believe that it will be difficult to resolve a global evolutionary trend from lower to higher specificity.

## Conclusions

A number of ASR studies are starting to reveal a consistent pattern of elevated thermostability for the deepest ancestors. This trend of decreasing thermostability among mesophilic lineages is not smooth, involving fluctuations in both $T_m$ and mechanism of stabilization. Whether this reflects a real evolutionary signal or simply an artifact of the reconstruction method remains to be seen. From an engineering perspective, a ML reconstruction of an ancient ancestor appears to be a reasonable strategy for generating a thermostable, thermophilic-like protein that differs from a simple consensus sequence. This approach is not guaranteed — for example, reconstructed RNase H displays non-thermophilic-like thermostability

$\sim$3 billion years in the past — however, on average, deep ancestral proteins appear to be more stable than their modern counterparts. We should also note that these are deep trends, and thus we would not predict recent ancestors to exhibit any detectable trend in stability, consistent with recent studies [114, 117, 126].

Information about the specificity of deep ancestral proteins remains sparse and will thus require further investigation. Studies of more recent proteins indicate that multiple modes of specificity evolution can be at play, suggesting a lack of general trends.

Protein evolution is often viewed as a random, microscopically-reversible trajectory along a fitness landscape. A global trend would suggest that the fitness landscape changed in a systematic way, even while microscopic reversibility held. Such systematic changes in fitness landscape would, in turn, shape the pathways taken by proteins and provide another level at which to understand the emergence of new properties. ASR studies are hinting at a change in fitness landscape. This may help us, at a broad brush level, gain insight into the origins of protein features and properties.

## Bridge to Chapter III

In this chapter, the current evolutionary biochemistry literature was reviewed to assess the available evidence for broad evolutionary trends in protein properties. Two highly-referenced examples of trends were analyzed: the hypothesis that proteins have undergone a gradual, monotonic decrease in thermal stability over long time scales and the assertion that proteins generally become more specific over time as proteomes become increasingly complex. The conclusion was drawn that there is some evidence to support a long-term decreasing trend in thermostability.

However, there is still relatively sparse information available. More experiments will need to be targeted toward addressing this question to establish firm conclusions. With regard to the evolution of specificity, this chapter concluded that there is vastly insufficient evidence to draw strong conclusions. Furthermore, unlike decreases in thermostability there is no unifying theoretical reason to expect global, parallel increases in specificity over time. This chapter established the need for further experimentation and more complete theories to address the issue of global trends in protein evolution. Chapter III addresses trends in a specific biochemical feature during the evolution of an entire protein family. An interesting observation is made regarding the evolutionary lability of this feature and its underpinnings at the amino acid level.

# CHAPTER III

# MULTIPLE EVOLUTIONARY ORIGINS OF UBIQUITOUS CU2+ AND ZN2+ BINDING IN THE S100 PROTEIN FAMILY

## Author Contributions

Lucas Wheeler and Michael Harms conceptualized the study and designed experiments. Michael Harms acquired funding for the study. Lucas Wheeler and Micah Donor performed experiments and analyzed experimental data. Michael Harms and James Prell administered the project. Michael Harms conducted phylogenetic analyses. Lucas Wheeler and Michael Harms wrote and edited the manuscript.

## Abstract

The S100 proteins are a large family of signaling proteins that play critical roles in biology and disease. Many S100 proteins bind $Zn^{2+}$, $Cu^{2+}$, and/or $Mn^{2+}$ as part of their biological functions; however, the evolutionary origins of binding remain obscure. One key question is whether divalent transition metal binding is ancestral, or instead arose independently on multiple lineages. To tackle this question, we combined phylogenetics with biophysical characterization of modern S100 proteins. We demonstrate an earlier origin for established S100 subfamilies than previously believed, and reveal that transition metal binding is widely distributed across the tree. Using isothermal titration calorimetry, we found that $Cu^{2+}$ and $Zn^{2+}$ binding are common features of the family: the full breadth of human S100 paralogs—as well as two early-branching S100 proteins found in the

tunicate *Oikopleura dioica*—bind these metals with $\mu$M affinity and stoichiometries ranging from 1:1 to 3:1 (metal:protein). While binding is consistent across the tree, structural responses to binding are quite variable. Further, mutational analysis and structural modeling revealed that transition metal binding occurs at different sites in different S100 proteins. This is consistent with multiple origins of transition metal binding over the evolution of this protein family. Our work reveals an evolutionary pattern in which the overall phenotype of binding is a constant feature of S100 proteins, even while the site and mechanism of binding is evolutionarily labile.

## Introduction

The S100 protein family is an important group of calcium binding proteins found in vertebrates [89, 91]. Humans possess 27 family members that play diverse functional roles in inflammation [151, 101, 152], cell proliferation [153, 154, 155], and innate immunity [156, 105, 157]. S100 proteins are particularly prominent in inflammatory diseases and cancers, where they are used both as clinical markers and drug targets [100, 158, 159, 160, 161, 102, 162, 163, 164, 165]. S100 proteins are found only in chordates and are highly diverged from other calcium binding proteins [91, 100].

Most S100 proteins share a common homodimeric structure in which $\sim$10 kDa monomers come together to form a compact $\alpha$-helical fold (Fig 4A). Each monomer binds two $Ca^{2+}$ ions in conserved calcium binding motifs, inducing a conformational change that exposes a hydrophobic surface [166, 94, 167]. This surface can then interact with and modulate the activity of downstream target proteins [168, 169].

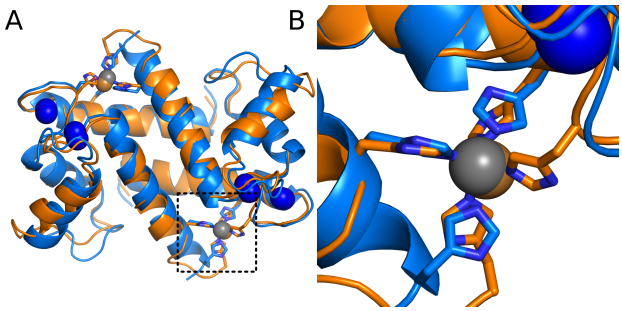


FIGURE 4 Transition metal binding occurs at a common site in diverse S100 proteins. Overlay of the crystal structures of S100B (orange, PDB 3CZT) and S100A12 (blue, PDB 1ODB) bound to Ca2+ and transition metals. Ions are shown as colored spheres: $Ca^{2+}$ (blue), $Zn^{2+}$ (gray) and $Cu^{2+}$ (copper). Residues ligating the transition metals are are shown as sticks. Boxed region is shown in detail in panel B.

In addition to $Ca^{2+}$, many S100 proteins interact with divalent transition metals such as $Zn^{2+}$, $Cu^{2+}$, or $Mn^{2+}$ as part of their biological functions [170, 171]. Such functions include metal transport [172], modulation of signaling [173], and antimicrobial activity [105]. Their transition metal binding constants tend to be $\sim\mu$M, consistent with their roles in metal transport and metal-dependent signaling [174, 175]. Despite the importance played by these metals, transition metal binding has not been studied systematically across the family [170, 171]. While one key transition metal site—at the dimer interface—has been studied extensively (Fig 4B), the transition metal binding capacity of many S100 proteins remains unknown. For many others, there are conflicting reports about the binding affinities, sites, and stoichiometries for binding to divalent transition metals [170, 171].

Evolutionary history provides a powerful lens through which to understand this metal binding diversity and its accompanying functional diversity. Understanding when a feature evolved in the family, and thus which homologs might share the feature, helps translate observations for one family member into predictions about other family members. One key question is whether

transition metal binding is a shared ancestral feature, or whether it has been acquired independently on multiple lineages. Although all five crystal structures of S100 proteins bound to transition metals have similar binding sites (Fig 4B), experimental evidence suggests that other S100s bind to divalent transition metals at a different site than the one identified crystallographically [176, 177], consistent with at least one more acquisition of transition metal binding.

A well-supported phylogeny of the S100 protein family would allow observations of transition metal binding to be mapped as evolutionary characters, thereby allowing inferences about the evolutionary history of the character. Several phylogenies have been published [91, 100, 178, 95, 179], however, these trees are not fully consonant with one another, making interpretation difficult. Previous analyses were limited by the number of S100 sequences available, particularly from early-branching vertebrate species. Further, all but one [95] relied on distance-based phylogenetic methods. Increased taxonomic sampling, combined with more advanced phylogenetic methods, will provide a much clearer picture of S100 evolution.

We therefore set out to understand the evolution of transition metal binding in this family through a combination of phylogenetic analysis and biochemical characterization of select human paralogs. Further, to establish the ancient features of the family, we performed the first-ever biochemical characterization of two early-diverging S100 proteins from the tunicate *Oikopleura dioica*. Our work sheds light on the evolutionary process that gave the diversity of modern S100 proteins, as well as revealing the broad-brush evolution of the transition-metal binding phenotype of this important protein family.

## Results

### *The S100 family arose in the ancestor of Olfactores*

Our first goal was to establish the taxonomic distribution of the S100 family. We began with an iterative BLAST approach. We used the full set of 27 human S100 family members (S1 Table in supplementary directory) as a starting point for PSI-BLAST against the NCBI non-redundant protein database. In addition to identifying thousands of S100 sequences, this protocol picked up non-S100 calcium binding proteins such as calmodulin and troponin, indicating that we had saturated S100 proteins in the database. We filtered our hits by reverse BLAST. All S100 hits were within vertebrates, with the exception of four hits from the tunicate *Oikopleura dioica*. To further support the taxonomic distribution of the S100s, we then used BLAST to search directly in the genomes and transcriptomes from representative tunicates, cephalochordates, hemichordates, and echinoderms. Only a transcriptome from the tunicate Molgula tectiform yielded a further S100 hit. We also queried the HMMER database, but found no new S100 family members. The presence of S100 proteins in tunicates and vertebrates (Olfactores), but not other chordates, suggests that the first S100 arose in the last common ancestor of tunicates and vertebrates, ∼700 million years ago [180]. These results are consistent with previous studies that noted the relative youth of the S100 family[91, 100, 95, 179].

### *Model-based phylogenetic approaches reveal well-supported clades*

We next constructed a phylogenetic tree, using sequences drawn from across Olfactores. Phylogenetic analyses of this family are challenging as it is large

and diverse. For example, the average sequence identity of the 27 human family members is 29.5%, with the most divergent pair (A3 and A14) only 13.2% identical. Further, the small size of these proteins (∼100 amino acids) means they have few evolutionary characters and, thus, relatively weak phylogenetic signal. Finally, many S100 paralogs exhibit highly specific tissue distributions, meaning that transcriptomes can provide very incomplete pictures of the S100 complement of a given organism.

To construct a tree despite these difficulties, we assembled a high-quality dataset of 564 sequences, from 52 species, through targeted searches of key genome/transcriptome/proteome databases (S2 Table, S1 Spreadsheet in supplementary directory). In an effort to bracket the class-level evolutionary origin of each S100 ortholog—despite incomplete sequence data and possible differential loss along each lineage—we included multiple species within each class: two Tunicata (one Ascidiacea, one Appendicularia), two Agnathan (jawless fishes), seven Chondrichthyans (cartilaginous fishes), eight Actinopterygii (ray-finned fishes), three Sarcopterygii (lobe-finned fishes), seven Amphibians, fourteen Sauropsids (birds and reptiles), and seven Mammals (two monotremes, two therians, and three eutherians). We generated a 133 character alignment from these sequences (Fig 24 in supplement and S2 Fig in supplementary directory, S1 Alignment) and used this for model-based phylogenetics.

We used both maximum likelihood (ML) and Bayesian approaches to construct phylogenetic trees for the family (Fig 5, S1 Tree and S2 Tree in supplementary directory, Fig 25 in supplement ). Both approaches resolved well-supported clades containing each of the human seed paralogs. This allowed us to assign the orthology, relative to the human proteins, for 500 of the 564 sequences

in our data set (S1 Spreadsheet in supplementary directory). In addition, the ML and Bayesian approaches revealed a set of consonant clades: A2/A3/A4; A5/A6; the calgranulins (A7/A8/A9/A12); A13/A14; and the so-called "fused" family (cornulin/ trichohyalin/repetin/hornerin/filaggrin) (Fig 5 and Fig 25 in supplement). In the Bayesian consensus tree, no further relationships could be resolved. Several other clades were resolved in the ML tree (Fig 5); A2/A3/A4 groups with A4/A5; A10 with A11; and A13/A14 groups with A16. In both trees, the sum of the branch lengths was extremely long, reflecting the high diversity of the family.

We were particularly interested in placing the tunicate S100 proteins on the tree. If we could assign the orthology of these proteins, we could potentially identify the most ancient S100 orthlog(s). Unfortunately, the placement of these sequences on the tree was neither evolutionarily reasonable nor stable between phylogenetic runs. For example, a single tunicate protein might end up on a long branch within a clade of mammalian proteins in one analysis, and then in an entirely different location in another. We thus excluded the tunicate proteins from the final phylogenetic analysis.

Uncertainty in the deepest branching pattern precluded rooting of the tree. We attempted to root the phylogeny by three methods; however, none proved successful. The first method was to include non-S100 calcium-binding proteins identified in our BLAST searches (sentan, calcineurins, troponins, and calmodulins) as an outgroup. With the exception of sentan, these non-S100 proteins grouped together; however, the branch leading to the clade was too long to allow robust placement relative to the S100 proteins—minor changes to the alignment and/or tree-building protocol would radically change their relationship to the rest of

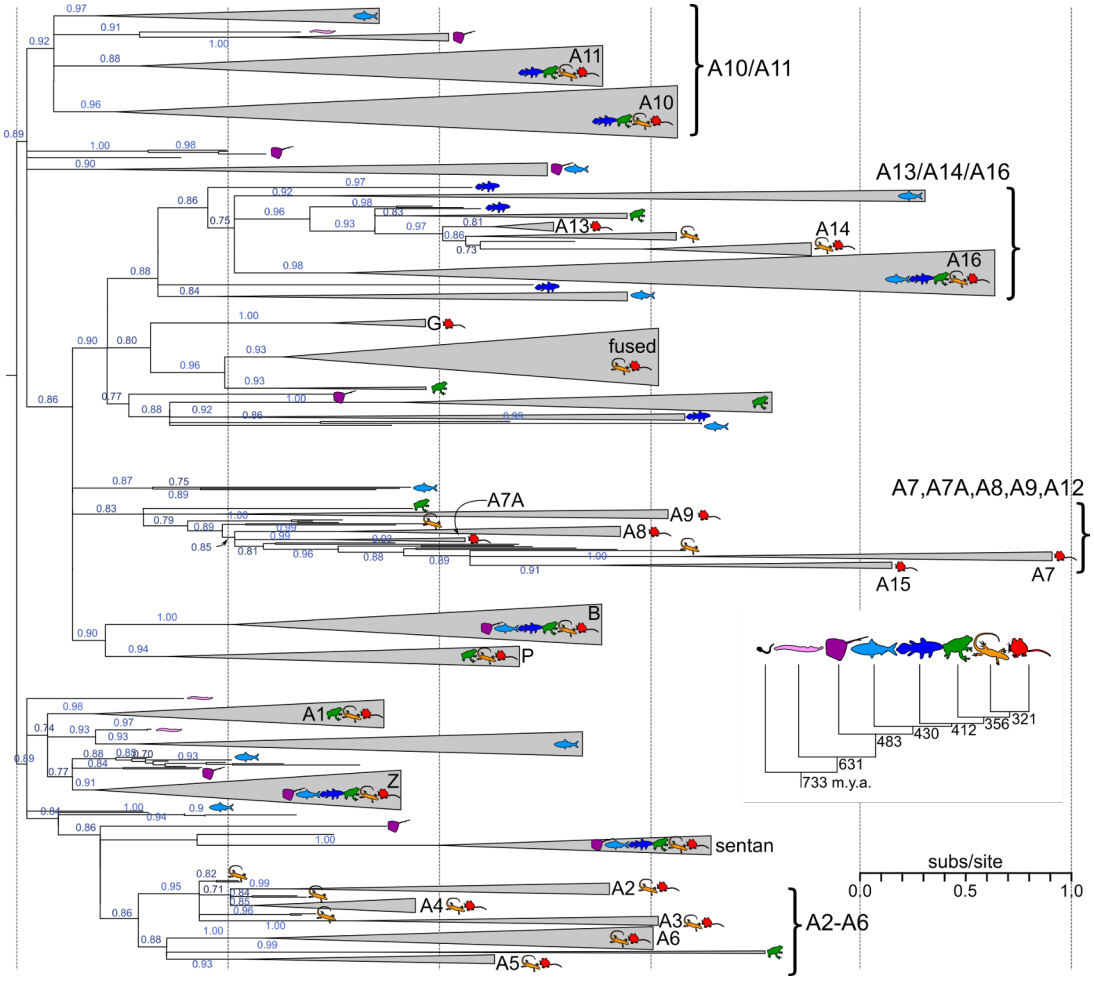


FIGURE 5 Model-based phylogenetics reveal several S100 subfamilies. Maximum likelihood phylogeny of 564 S100 proteins drawn from 52 Olfactores species. Wedges are collapsed clades of shared orthologs, with wedge height denoting number of included taxa and wedge length denoting longest branch length with the clade. Support values are SH-supports, derived from an approximate likelihood ratio test. Rooting is arbitrary, but roughly balances the distribution of jawless fishes across the ancestral node. Icons indicate taxonomic classes represented within each clade: tunicates (black), jawless fishes (pink), cartilaginous fishes (purple), ray-finned fishes (light blue), lobe-finned fishes (blue), amphibans (green), birds/reptiles (yellow), and mammals (red). Inset shows estimated divergence times for each taxonomic class in millions of years before present.

the tree. We also attempted to use the tunicate proteins, but as they could not be placed, this was ineffective. Finally, we attempted to minimize the number of duplications and losses across the tree; however, the lack of resolution of the deepest nodes also made identifying the precise origin (and thus gain/loss) of each paralog problematic.

### *Synteny and taxonomic distribution further support relationships among S100 proteins*

Because model-based phylogenetic methods provided relatively weak support for relationships within in the family, we used the taxonomic distribution of orthologs and synteny to further support the relationships we observed in the model-based approaches. Fig 6 shows distribution of observed orthologs to human genes across the species included in our analysis. (Species phylogenies taken from [181, 182, 183, 184, 185, 186, 187, 188, 189]). We mapped these orthologs onto the arrangement of these genes in the human genome (top). Four S100 genes (G, B, P, and Z) are scattered on different chromosomes, while twenty-two S100 genes (A1 through A10) form a contiguous block on a single chromosome. This tight linkage group has been noted previously [100, 179, 190], and arose at least as early as the bony vertebrates [179].

There is strong correlation between the S100 subfamilies identified in model-based phylogenetics and the distribution of the genes across human chromosome I. Proteins with shared evolutionary relationships form blocks across this region, suggesting local expansion by gene duplication. The ML relationship between orthologs are shown above the plot in Fig 6. The clades identified in our model-based phylogenetics form individually contiguous blocks: A13-A16, A2-A6, A7-A12,

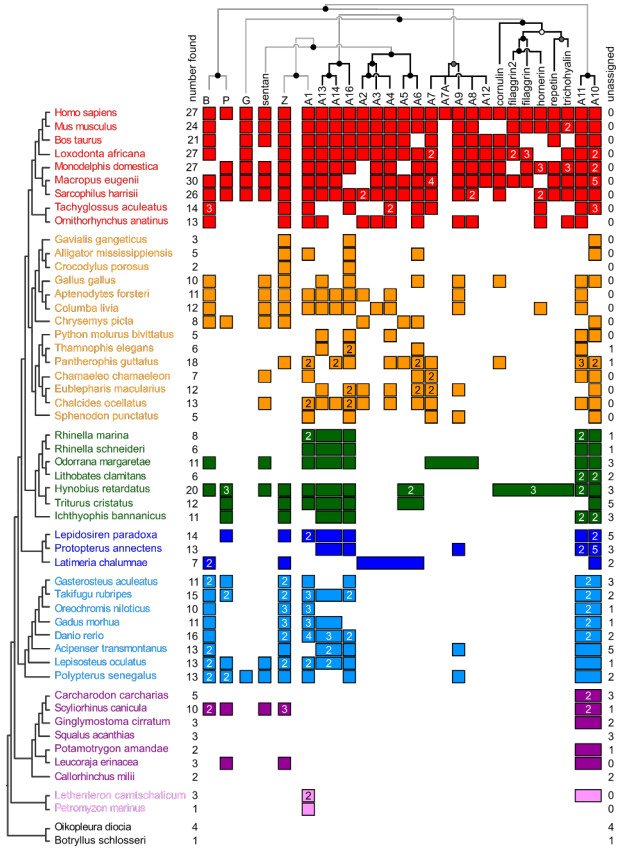


FIGURE 6 Model-based phylogeny, synteny, and taxonomic distribution provide a consonant picture of S100 evolution. The human S100 orthologs are shown across the top, in the order they occur in the human genome. B, P, G, and Z occur on different chromosomes; A1-A10 are in a contiguous region of chromosome I. Sentan, an evolutionary relative, is also on a different chromosome. Species are shown on the left, organized by taxonomy. Color indicates taxonomic class, as in Fig 5. Squares denote the presence of an ortholog to the human gene for each species; a number in the box indicates the number of co-orthologous genes found in that species (if more than one); squares fused into a rectangle indicate a gene found in an earlier branching lineage that subsequently duplicated somewhere along the lineage leading to Homo sapiens. Total number of genes found for each species are shown on the left. The number of genes that were not orthologous to human genes (or could not be classified) are shown on right. Top tree shows the maximum-likelihood phylogeny of the family mapped onto the S100 genes found in the human genome. Circles denote SH support $\geq 0.85$ (black); $\geq 0.75$ (gray), $< 0.75$ (white). Branches supported by both the ML phylogeny and synteny are shown in black; branches supported by only the ML tree are shown in gray.

S100-fused, and A11-A10. This consonance between the phylogenetic signal and genomic arrangement supports the shared ancestry of these subfamilies.

The species distribution of these orthologs then provides insight into the diversification of the family. For example, A10, A11, or their common ancestor (A10/A11) are found in all vertebrates, demonstrating that this protein arose no later than the last common ancestor of vertebrates. Because some genes may have been missed within each species—either through lineage-specific loss or incomplete genomic/transcriptomic coverage—this is a lower bound on the age of the gene. After its origin, A10/A11 then diversified in later lineages. In the bony fishes, A10 expanded, as reflected in the increased numbers of genes co-orthologous to A10/A11. A10/A11 gave rise to the tetrapod paralogs A10 and A11 via tandem gene duplication in the ancestor of the lobe-finned fishes.

Another ancient S100 by this analysis is A1, which, intriguingly, brackets the other end of the contiguous S100 genome region mammals and some fishes [179]. The simplest interpretation of this pattern would be that the A1 or A10/A11 gene was the earliest gene in this syntenic block, and that the remaining family arose by serial expansion from that starting point.

Other ancient S100 orthologs are B, P, and Z. Our tree provides some evidence that A1 and Z share a common ancestor, and that B and P share a common ancestor. Intriguingly, these four ancient proteins are scattered throughout vertebrate genomes, rather than being a part of the expanded gene region containing A1-A10. This suggests that the last common ancestor of jawed vertebrates had a collection of four to five S100 proteins, but that only the region containing A1-A10 then continued to expand with the radiation of the vertebrates. Sentan—a close evolutionary relative to the S100 family that does not possess

the diagnostic pseudo EF-hand of true family members—also arose in the early vertebrates. Given the ambiguity of the deepest branching of the tree, it is unclear whether it is an out group or, instead, a duplication of an established S100 paralog.

The gene block containing A1-A10 expanded by what appears to be a set of local gene duplication events. A13/A14 and A16 likely arose next, at least by the ancestor of bony vetebrates. Like A10/A11, these genes were duplicated through the whole genome duplications of teleost fishes, giving rise to multiple S100 genes that are co-orthogolous to the human genes in bony fishes. The tetrapod paralogs A13 and A14 did not arise until the amniotes, when they formed via duplication from A13/A14. The next phase of expansion was local duplication that led to the ancestors of A2-A6, A7-A12, and the S100-fused proteins in early tetrapods. These founding genes then expanded across the tetrapods, with several duplicates preserved in Sauropsids. The final mammalian complement was achieved by several more duplications. The A7-A12 and S100-fused clades—which are directly adjacent in mammalian chromosomes—continue to rapidly expand by duplication.

*Transition metal binding is nearly universal across the family*

With the phylogenetic tree in hand, we next set out to determine the distribution of transition metal binding across the tree. Previously reported transition metal binding is scattered across the tree (Fig 4, red proteins) [170, 171]. If this feature were ancestral, we predicted that transition metal binding would be present across the majority of the tree. To test this hypothesis we used isothermal titration calorimetry (ITC) to measure the ability of human S100 proteins to bind to $Zn^{2+}$ and $Cu^{2+}$—the two most prevalent transition metals encountered biologically—under approximately physiological conditions (125 mM ionic strength,

pH 7.4, 25°C). We chose proteins that would maximize the sampling across clades. Some of the proteins we selected have been reported to bind transition metals, albeit with variable stoichiometry [177, 191]. The other paralogs have, to our knowledge, yet to be characterized.

We found that $Zn^{2+}$ and $Cu^{2+}$ binding was universally distributed across the tree: every single S100 protein we characterized bound to $Zn^{2+}$ and/or $Cu^{2+}$ with low micromolar affinity (Fig 4 and Fig 26 in supplement, S3 Table in supplementary directory) [105, 177, 192, 193, 194, 195, 196]. With one exception, stoichiometry ranged from 1:1 to 3:1 (metal:monomer). These binding affinities and stoichiometries are similar to previously measured transition metal binding affinities for S100 proteins [171, 192, 195, 197]. Buffer-specific enthalpies ranged from -5.4 to 6.1 kcal/mol; the majority of the enthalpies were negative. All of the proteins tested bound to both $Zn^{2+}$ and $Cu^{2+}$, with the exception of A1 which did not bind $Cu^{2+}$ under our experimental conditions. The $Zn^{2+}$ binding isotherm for A6 and the $Cu^{2+}$ binding isotherms for A2 and A4 were not well fit by standard binding models (as is often observed for metal binding studies by ITC: [198]), however, from the curves we could gain insight into their stoichiometry. The A6/$Zn^{2+}$ and A4/$Cu^{2+}$ curves exhibited two phases, consistent with two binding sites. The A2/$Cu^{2+}$ curve was quite broad, consistent with >2 metals binding per monomer. Representative binding isotherms for $Zn^{2+}$ and $Cu^{2+}$ to a variety of S100 proteins—including the three problematic curves—are shown in Fig 26 in supplement. All measured thermodynamic parameters are reported in S3 Table in supplementary directory.

We next asked if the structural response to these metals, like the binding constant, was consistent across the tree. We measured $Zn^{2+}$-induced changes in

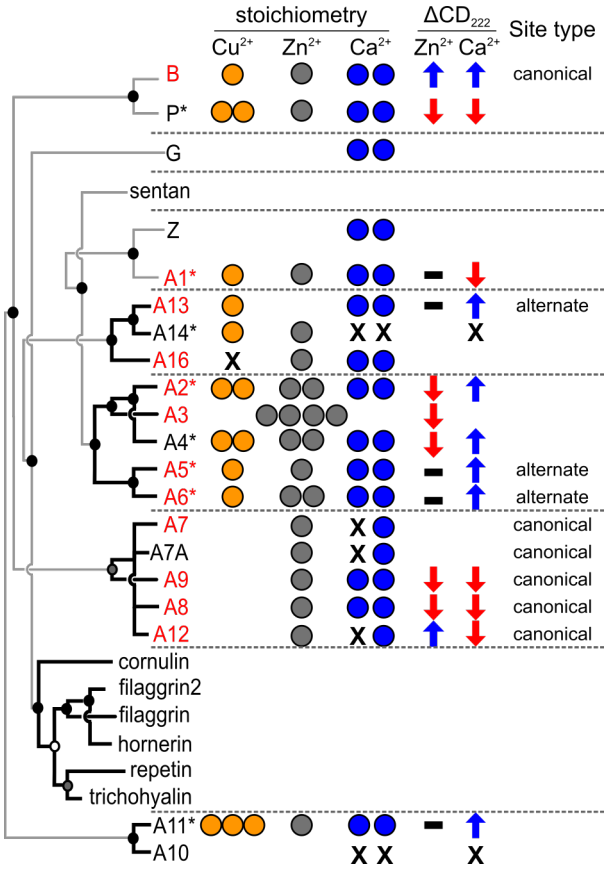


FIGURE 7 The human S100 paralogs are shown on the left, organized as on the top of Fig 6. Asterisks indicate S100 proteins investigated in the current study; red color indicates a protein for which transition metal binding has been noted in the literature previously. Biochemical properties of the human paralogs are shown as columns. Circles denote stoichiometry of binding for $Cu^{2+}$ (orange), $Zn^{2+}$ (gray), and $Ca^{2+}$ (blue). X indicates that the protein does not bind the metal; empty space is unmeasured. Arrows indicate the change in far-UV CD signal with the indicated metal: no change (black), increase (blue), and decrease (red). The transition metal binding site is indicated as canonical (B-like) or alternate (some other site).

secondary structure by comparing the far-UV circular dichroism (CD) spectra of these proteins with EDTA versus saturating $Zn^{2+}$ (Fig 27 in supplement). We found the response was variable across the family (Fig 4) [177, 192, 195, 199, 200, 201, 202, 203, 204, 205, 206]. For some proteins, $Zn^{2+}$ induced a decrease in CD signal (P, A2 and A4); in others, it had no effect (A1, A11, A5 and A6). We also observed $Zn^{2+}$-induced protein precipitation in the case of A14, which was rapidly reversible by the addition of excess EDTA. We also asked whether the structural response to $Zn^{2+}$ exhibited by these proteins correlated with the response to their canonical agonist $Ca^{2+}$. We found that they were largely uncorrelated (Fig 7 and Fig 27 in supplement). For example, P has decreased CD signal with both $Zn^{2+}$ and $Ca^{2+}$, while A2 shows decreased signal with $Zn^{2+}$ and increased signal with $Ca^{2+}$.

When placed onto the phylogenetic tree, a few patterns in these responses emerge (Fig 7). Phylogenetically close members of the family appear to display similar structural responses to $Zn^{2+}$ binding. For example, the closely related A2 and A4 proteins show qualitatively similar decreases in CD signal in the presence of $Zn^{2+}$ relative to the apo form. Likewise, the far-UV CD signal of direct sister proteins A5 and A6 is insensitive to $Zn^{2+}$. This said, such patterns are not universal. For example, B and P are directly sister but have opposite structural responses to $Zn^{2+}$. Further, family members exhibit all possible combinations of increased and decreased CD signal with the addition of $Ca^{2+}$ and $Zn^{2+}$, revealing the variability of this trait over evolutionary time.

*Early-diverging tunicate S100s bind transition metals*

Given that all human paralogs we characterized were capable of binding transition metals, we predicted that this was a conserved, early feature of the protein family. To test this prediction, we turned to two tunicate homologs, which represent some of the earliest-diverging S100 proteins. We selected two *Oikopleura dioica* proteins—tunA (tunicate A, CBY12809.1) and tunB (tunicate B, CBY30360.1)—for characterization. Although the orthology of these proteins is unclear, the proteins sample the breadth of tunicate S100 diversity, exhibiting only 26.2% identity. We expressed and purified these proteins, and then characterized their metal binding features.

Because these proteins have not been characterized previously, we first performed a baseline characterization to verify that they behave like other S100 proteins. We first measured $Ca^{2+}$ binding. Like many other S100 proteins, both tunA and tunB bound $Ca^{2+}$ with nanomolar to micromolar dissociation constants and 2:1 (per monomer) stoichiometry (Fig 8A and Fig 28 in supplement). Further, both proteins exhibited changes in secondary and/or tertiary structure—as measured by far-UV circular dichroism (CD) and intrinsic fluorescence—with the addition of saturating amounts of $Ca^{2+}$ (Fig 8B and 5C and Fig 28 in supplement). All of the observed changes were strictly metal dependent and reversible upon the addition of EDTA. Metal-dependent changes in conformation, as reflected in these changes in spectroscopic signals, are a hallmark of S100 proteins [177, 207, 208, 209].

We then assessed the ability of these proteins to form homodimers—a key feature of most S100 proteins—using native electrospray-ionization mass spectrometry (nanoESI) [210]. For tunB, we detected homodimers (Fig 8D). The

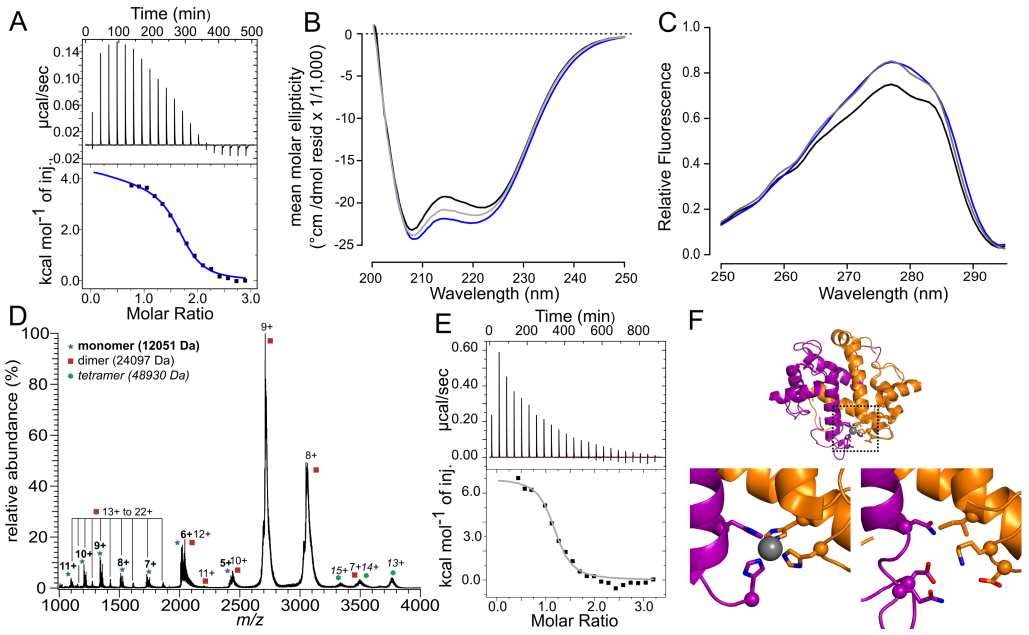


FIGURE 8 Early-branching tunicate S100 binds transition metals at a non-canonical site. Colors indicate the metal present during experiment: $Zn^{2+}$ is gray, $Ca^{2+}$ is blue. A) Ca2+ binding to tunB by ITC. Top panel shows power traces for injections; bottom curve shows integrated heats and model fit to extract thermodynamic parameters. B) Far-UV circular dichroism spectra of the apo protein (black), $Ca^{2+}$ bound protein (blue), or $Zn^{2+}$ bound protein (gray). C) Intrinsic fluorescence spectra, with samples colored as in panel B. D) Mass spectrum of tunB. Notes above each peak indicate molecular weight and corresponding oligomeric state. E) $Zn^{2+}$ binding to tunB by ITC, with subpanels as in A. E) Homology model of tunB overlaid on crystal structure of human S100B (PBD: 3ZCT). Ligating residues are shown as sticks, with C atoms shown as spheres. A and B chains of the dimer are shown in orange and purple, respectively. $Zn^{2+}$ ion is shown as gray sphere. Top panel shows overlay, with box highlighting the zoomed-up regions shown at right. Bottom left panel shows S100B structure with $Zn^{2+}$ chelation. Bottom left panel shows tunB homology model, highlighting residues that would have to chelate $Zn^{2+}$.

narrow distribution of relatively low charge states observed in the nanoESI mass spectra for both the monomer and dimer ions indicate that the proteins are not denatured under these conditions and undergo little unfolding during the ionization process. The broad mass spectral peaks observed are the result of adduction of residual sodium from solution that has survived buffer exchange. To see if the dimer peaks were the result of non-specific aggregation during the electrospray process, we measured dimerization at protein concentrations at which non-specific dimerization is not expected ($< 1$ $\mu$M, see methods). We found homodimers, even at 10 nM protein, consistent with a specific tunB dimer (Fig 29 in supplement). We also observed a small amount of homotetramer; however, the tetramer was not robust to dilution and is likely an artifact of the electrospray process (Fig 29 in supplement). For tunA, we detected homodimers; however, these were not robust to dilution, suggesting that dimerization is relatively weak for this protein (Fig 28 in supplement). We corroborated these observations for tunA and tunB using a sedimentation velocity experiment (Fig 30 in supplement). Under these conditions, we found that tunB was primarily a dimer. In contrast, tunA exhibited both monomer and dimer species, consistent with this protein forming a weaker dimer. Further work is required to determine the precise distribution of oligomeric species in solution for these proteins; however, these results are consistent with both proteins having the ability to form homodimers, like other S100 proteins [211].

We next turned our attention to $Zn^{2+}$ binding. By ITC, both tunA and tunB bound to $Zn^{2+}$ with nM to $\mu$M affinity and stoichiometries of 2:1 (Fig 8E and Fig 28 in supplement). We attempted to verify these stoichiometries by ESI-MS; however, we were unable to disentangle specific from non-specific metal adduction in these samples. We then measured the changes in secondary and

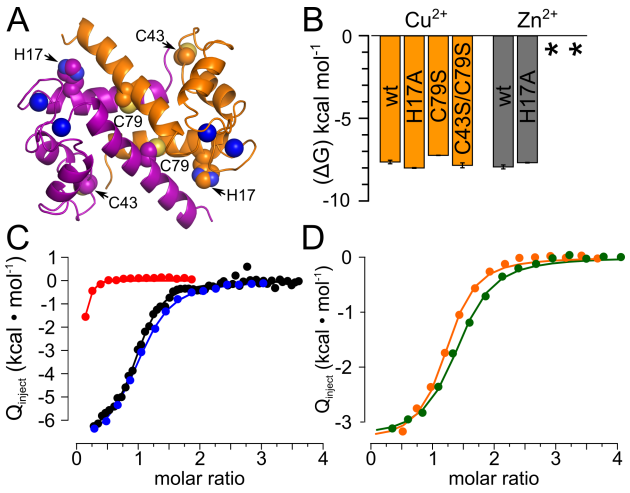


FIGURE 9 Human S100A5 does not bind transition metals at the same site as B and the calgranulins. A) Mutated residues mapped onto the NMR structure of $Ca^{2+}$–bound human A5 (PDB: 2KAY). Dimer chains are colored purple and orange. H17, C43, C79 and $Ca^{2+}$ ions are shown as spheres. The location of H17 corresponds to the transition metal site in calgranulins and B (Fig 4B); C43 and C79 are in different regions of S100A5. C) Binding free energies measured for $Cu^{2+}$ (copper) and $Zn^{2+}$ (gray) to human A5 and its mutants. $Zn^{2+}$ binding constants could not be extracted for the C43S and C43S/C79S proteins (*). C) Integrated heats for ITC titration of $Zn^{2+}$ onto A5 (black), A5/H17A (blue) and A5/C43S (red). D) Integrated heats for ITC titration of $Cu^{2+}$ onto A5 in the absence (orange) or presence (green) of saturating (500 $\mu$M) $Zn^{2+}$.

tertiary structure measured by far-UV CD and intrinsic tyrosine fluorescence. Although both proteins bound $Zn^{2+}$ tightly, only tunB displayed a pronounced structural response, similar to that induced by $Ca^{2+}$ binding (Figs 5B and 6C). The secondary structure of tunA was insensitive to $Zn^{2+}$ binding although the protein displayed a moderate increase in intrinsic tyrosine fluorescence (Fig 28 in supplement).

*Transition metal binding occurs at independently evolved binding sites*

The broad distribution of transition metal binding across human paralogs, along with the observed transition metal binding in the early-branching tunicate

proteins, suggests that transition metal binding is an essentially universal property of this family. We next sought to understand to what extent transition metal binding across the family reflects a common binding site, or rather convergent acquisition of metal binding on multiple lineages. Transition metal bind- ing to S100 proteins has been extensively characterized in B and the "calgranulin" clade (A7,A8,A9,A12,A15), where it occurs at the same site, using similar ligating residues (Fig 4B). B is an ancient protein, arising at least as early as the cartilaginous fishes (Fig 6). In contrast, the calgranulins arose ∼80 million years later in the ancestor of amniotes (Fig 6). If the common site reflects shared ancestry, we would expect to observe the same site across a wide variety of descendants—possibly explaining the ubiquity of transition-metal binding across the tree.

We first investigated the clade containing A2,A3,A4,A5, and A6. All members of this clade possess a conserved histidine that, in B and the calgranulins, coordinates transition metals (Fig 9A). We chose to investigate human A5, as it binds to both $Zn^{2+}$ and $Cu^{2+}$ with 1:1 stoichiometry, and thus simplifies identification of the binding site. We mutated His17 to Ala in human A5 and measured metal binding of the mutant. Surprisingly the H17A mutation had only a small effect on $Zn^{2+}$ binding (1.3 +/- 0.3 to 3.0 +/- 0.1 $\mu$M), suggesting it is not directly involved in the binding of $Zn^{2+}$ in human A5. Additionally, this mutation did not compromise $Cu^{2+}$ binding (Fig 9B). Previous reports suggested that Cys residues in the loop between helices 2 and 3, as well as those near the N and C-termini, could play a role in binding divalent transition metals in this clade [170, 177, 193]. We therefore mutated these residues to serine in A5 and measured binding of $Zn^{2+}$ and $Cu^{2+}$ to the mutants. Mutating the C-terminal Cys

(C79S) had no effect on $Cu^{2+}$ binding, but led to a drastic change in the $Zn^{2+}$ binding curve (Fig 9C). The apparent stoichiometry of binding was drastically reduced ($\sim$0.1), which is consistent with only a small fraction of the protein being competent to bind $Zn^{2+}$. Additionally, the enthalpy of binding is mostly ablated. These results clearly indicate that C79 is involved in $Zn^{2+}$, but not $Cu^{2+}$ binding. We attempted to ablate $Cu^{2+}$ binding by also mutating the loop Cys residue (C43S), but found that this double Cys mutant (C43S/C79S) still left $Cu^{2+}$ binding unaffected (Fig 9B). These results show that $Zn^{2+}$ and $Cu^{2+}$ not only bind outside the B/calgranulin site, but bind at different sites on the same protein. To confirm that these metals bind at different sites, we also measured binding of $Cu^{2+}$ to $Zn^{2+}$-saturated human A5 and found no evidence of competition between the two metals (Fig 6D). Finally, because mutating H17, C43, and C79 did not disrupt $Cu^{2+}$ binding, we hypothesized that the metal might bind at one of the $Ca^{2+}$ binding motifs. We therefore repeated the $Cu^{2+}$ binding curve in the presence of saturating (2 mM) $Ca^{2+}$. We observed extensive aggregation, however, which made interpretation of the ITC binding isotherm impossible. This suggests that previously-noted antagonism between $Ca^{2+}$ and $Cu^{2+}$ [195] may be an artifact of aggregation rather than true antagonism.

We next turned our attention to the tunicate protein tunB. This protein behaves like a conventional S100 protein, forming a homodimer, binding to $Ca^{2+}$ and changing its structure in response to metals (Fig 5A–5E). Further, it binds to transition metals with a 2:1 stoichiometry. To determine if it could bind metals at the canonical transition metal binding site, we constructed a homology model for the protein and then inspected the residues that would form the S100B/calgranulin binding pocket. These are Asp, Gln, Asn, and Lys (Fig 8F). The lack of a His or

Cys residue suggests this site is not capable of binding transition metals. Thus, transition metal binding in this early-branching ortholog almost certainly occurs at a different site.

## Discussion

Our work provides a high-level view of the evolution of the S100 protein family and the ability of its members to bind to divalent transition metals. Our work provides the best-resolved phylogeny yet determined for this family. All characterized human paralogs, as well as two early-branching tunicate S100 homologs, bind to transition metals with a physiologically relevant $\sim\mu$M binding constant. On the other hand, different S100 proteins bind at different transition metal binding sites. Thus, the apparently "conserved" feature of transition metal binding actually reflects independent acquisition of metal binding on multiple lineages. Further, the structural changes induced by transition metal binding are variable, suggesting quite different mechanisms of binding and possible functional consequences for different family members.

### *Transition metal binding occurs at independently evolved binding sites*

Our work, combined with previous publications, reveals at least four sites—and therefore four evolutionary origins—of transition metal binding in the S100 family: the B/calgranulin site (Fig 4B), A5's Cys-79 site (Fig 9), an N-terminal Cys in A2 [177], and a unique glutamate-rich site in human A13 [176]. The plasticity of this feature is likely because of the relative ease, biochemically, of creating transition metal binding sites [212, 213, 214]. A few amino acid substitutions can create a new site, while a few other substitutions ablate an

existing site. This is similar to the evolutionary behavior of phosphorylation sites, which can shift rapidly over evolutionary time [215]. Additionally, some of the proteins may bind to transition metals in one of the $Ca^{2+}$ binding motifs of an S100. For example, Gribenko et al. proposed that human S100P may bind $Zn^{2+}$ in one of the $Ca^{2+}$ binding motifs [216]. EF-hands often discriminate $Ca^{2+}$ from $Zn^{2+}$ and $Cu^{2+}$, however, so this likely does not explain all of the observed transition metal binding [176, 217, 218, 219].

Another feature of $Zn^{2+}$ and $Cu^{2+}$ binding in this family is that of variable structural responses to the same metal. Even closely related S100 proteins undergo different conformational changes when bound to a transition metal (Fig 7). This likely allows different orthologs to play different functional roles in response to transition metal binding. This can be seen for proteins that have been studied in detail. For example, human A13, which binds $Cu^{2+}$ at a unique site, has been proposed to be involved in chaperoning $Cu^{2+}$ as part of FGF release [172]. A9 provides another example of diverse responses to transition metals. When A9 is alone, $Zn^{2+}$ binding is strictly necessary for one function (TLR4-activation) [220], but strongly inhibits another function (arachidnoic acid binding) [221]. This site is modified in vivo through the formation of a heterodimer with A8, which changes the ligating residues for one half of the site [105, 222]. This creates an extremely high affinity site for $Mn^{2+}$ and $Zn^{2+}$ that inhibits bacterial growth by starving them of these metals [105].

Much of the transition metal binding we have observed plays no known role, but the observed binding constants ($\sim\mu$M) are consistent with biological concentrations of divalent transition metals. In particular, many S100 proteins are found in the extracellular space [90], where $Zn^{2+}$ concentrations can be high enough

to occupy these sites [223, 224]. We expect further roles of transition metal binding to be identified in this family as it is further characterized [170, 171].

*Expansion of the family*

In addition to providing insight into the evolution of transition metal binding, our phylogenetic analysis provides insight into the overall pattern of expansion of the S100 protein family. Previous phylogenies used highly incomplete taxonomic sampling and, with the exception of [95], distance-based phylogenetics [91, 100, 179]. We used many more sequences, from many more taxa, and applied a combined model-based/synteny analysis to better disentangle the history of the family. Our work provides support for evolutionary relationships between A13-A16, A2-A6, the calgranulins, the S100-fused proteins, and A10/A11 despite the relatively weak support for these clades taken from a purely model-based phylogenetic perspective. This also supports the previously proposed model of local gene duplication [91, 100, 179].

Our work provides evidence for earlier origins of many S100 family members than previously reported. For example, we found that the S100A2-A6 clade likely arose in ancestor of all tetrapods, and that it had the complete mammalian complement by the ancestor of amniotes. In contrast, Zimmer et al proposed this clade arose in the ancestor of mammals [91]. Some orthologs (A1, B, P, and Z) have likely been present since the last-common ancestor of vertebrates. Further, we expect that many S100 proteins actually arose even earlier than our analysis suggests. Despite having broader sampling than previous studies, our sampling of tunicates, jawless fishes, and cartilaginous fishes was still relatively sparse. Further, we relied heavily on transcriptomes, which likely underestimate the S100

complements for these organisms. As more genomic and transcriptomic datasets for these species become available, we expect to observe even earlier origins of many of the mammalian S100 orthologs.

Another difference between our tree and the published tree by Kraemer et al. [95] is that we do not see radical, parallel expansion of the S100s in bony fishes. Rather, most S100 proteins from the bony fishes are orthologous to mammalian S100s. For example, we identified 15 S100 proteins in Takifugu rubripes (pufferfish). All but two of them could be assigned as orthologs to human proteins (Fig 6). This said, many of these do represent lineage-specific duplications—likely via the whole genome duplications that have occurred in teleost fishes—that are co-orthologous to human proteins. The difference between our results and the previous phylogeny likely arises from our much broader sequence sampling, as the Kraemer et al. dataset was strongly biased towards sequences taken from teleosts [95].

Despite extensive taxonomic sampling, the phylogenetic tree we report is not fully resolved: the deepest branches remain obscure. This is because of the large amount of sequence divergence that has occurred between many S100 protein family members, their relatively short sequences, and the number of orthologs make full resolution of this family quite challenging. Resolution can likely be increased for individual subfamilies within the tree through even denser sampling. For example, adding further aminotes may help resolve the relationships between the amniote-specific clades identified in our analysis. We also believe increasing the sampling of amphibians would be particularly powerful, as we relied heavily on amphibian transcriptomes and likely missed S100 proteins. Better characterization of S100 proteins from amphibians may help disentangle the origins

and relationships of some of the tetrapod-specific S100s (such as the calgranulins) which are, as yet, difficult to resolve. Further, signal for these relatively recent proteins could be boosted by using codon rather than amino acid substitution models.

## Conclusion

Our work reveals that transition metal binding is both ubiquitous and evolutionarily labile within the S100 protein family. Many have noted that much of the diversity of S100 function is determined by altered expression of family members [100, 207, 225, 226, 227, 228, 229]; however, our work highlights that these regulatory changes have also been accompanied by changes in sequence and biochemistry. In particular, the ease of creating and destroying transition metal binding sites has allowed rapid changes in this feature of S100 proteins. As a result, new metal binding behavior can be exploited to achieve functional diversity in the family [170, 171], even while $Ca^{2+}$ binding and its induced structural changes remain relatively conserved (Fig 7).

This biochemical diversification occurred rapidly during the expansion of the S100 proteins, which are a relatively young protein family. The details of how this diversification occurred are likely to encompass a rich evolutionary story. As new S100s arose via gene duplication, were they required to maintain metal binding while continuing to evolve? Or, have there been multiple cycles of loss and subsequent regain over the course of S100 evolution? What was the exact nature of metal-binding in the last common ancestor of all S100 proteins? Our observations provide groundwork to begin to ask these questions.

## Materials and Methods

### *Sequence Set*

We generated a database of 564 S100 protein sequences, sampled from 52 chordate species, with an emphasis on even taxonomic sampling (S1 Spreadsheet). Previous publications and preliminary database searches revealed S100 proteins were restricted to the chordates, [91, 100, 95] so we selected specific chordate species and characterized their S100 protein complements through extensive BLAST searches [230]. We used human proteins as seed sequences (including sentan and the S100-fused proteins, S1 Table in supplementary directory). No published genome or transcriptome data were available for some species, so we generated de novo transcriptomes from RNAseq data in the short reads archive [231] using Trinity with default parameters [232]. The sources for our analysis are shown in S2 Table in supplementary directory.

We removed duplicate sequences (>95% identity) from within each species using cdhit [233], and removed sequences less than 45 amino acids long. We then reverse BLAST'd all remaining sequences against the human proteome to verify they encoded S100 proteins. We aligned the sequences using msaprobs [234] followed by manual refinement in aliview [235]. Refinement was minimal and consisted of truncating variable N-terminal and C-terminal extensions, as well as several ambiguous indels. (We truncated the fused S100 protein sequences to 150 amino acids covering the S100 domain prior to alignment). The final alignment was 132 columns and had robustly aligned key columns (Fig 24 in supplement and S2 Fig in supplementary directory, S1 Alignment in supplementary directory).

*Phylogenetic Trees*

We generated the ML tree using phyml [236] with SPR moves starting from the neighbor-joining tree. 10 random starting trees did not yield a higher likelihood tree. We found LG+8 was the highest likelihood model [237]. We calculated aLRT-SH supports for each node [238]. In pilot analyses, the tunicate sequences were placed in random and unpredictable places on the tree (for example, coming out with mammals or in other nonsensical places on the tree). We therefore excluded them from the final ML analysis (S1 Tree in supplementary directory).

We generated a Bayesian phylogenetic tree using Exabayes [239]. We ran two replicate MCMC runs starting from different random trees, each consisting of one main and three heated chains. We stopped the runs after 10 million generations, giving a final average split frequency of 3.97% and log likelihood ESS of 3,315. We sampled substitution models in addition to trees, giving a final 99.8% posterior probability for the JTT model [240]. We used uniform priors for all parameters. We discarded the first 15% of the trees as burn-in and generated a consensus tree by majority-rule, collapsing all nodes with posterior probabilities <50% (S2 Tree).

*Molecular cloning and Protein Expression/Purification*

S100 proteins were expressed from synthesized genes in a pET28/30 vector that had an N-terminal, TEV-cleavable His tag (Millipore). Proteins were expressed in Rosetta (DE3) pLysS E. coli cells (Millipore). A saturated overnight culture was used to inoculate 1.5 L cultures at 1:150 ratio. Bacteria were grown to log-phase (OD600 $\sim$ 0.8–1.0) shaking at 37°C, followed by induction of protein expression in 1 mM IPTG for $\sim$16 hr at 16°C. Cells were harvested by centrifugation. Pellets were frozen at -20°C, where they were stored for up

to 2 months. Cells were lysed by sonication in 25mM Tris, 100mM NaCl, 25mM imidazole, pH 7.4.

Primary purification was done with a 5 mL HiTrap Ni-affinity column (GE Health Science) on an *Ä*kta PrimePlus FPLC (GE Health Science), using a 25mL gradient between 25 and 500 mM imidazole. Pooled fractions were then incubated overnight at 4°C in the presence of ∼1:5 TEV protease. This cleaves the His-tag from the protein, leaving the amino acids Ser-Asn in front of the wildtype starting methionine. Proteins were further purified by hydrophobic interaction chromatography (HIC) using a 5 mL HiTrap phenyl-sepharose column (GE Health Science). This step takes advantage of the $Ca^{2+}$-dependent exposure of a hydrophobic binding surface on the S100 proteins. Proteins were equilibrated with 2 mM $CaCl_2$ and loaded onto the HIC column, followed by a 30mL gradient elution in 25mM Tris, 100mM NaCl, 5mM EDTA, pH 7.4. Proteins were then dialyzed into 4 L of 25 mM Tris, 100 mM NaCl, pH 7.4 buffer overnight at 4°C. To remove the small amount of uncleaved His-tagged protein present, proteins were then passed over another 5 mL HiTrap Ni-affinity column and the flow through collected. Finally, if any protein contaminants remained by SDS-PAGE, we performed a final anion chromatography step using a 5mL HiTrap DEAE column (GE), 25mM Tris, pH 7.0–8.5 (depending on protein) buffer with a 50mL gradient to 500 mM NaCl.

Purified proteins were dialyzed overnight against 2L of 25mM TES (or Tris), 100mM NaCl, pH 7.4, containing 2 g Chelex-100 resin (BioRad) to remove divalent metals. Purity of final protein products were >95% by SDS PAGE and MALDI-TOF mass spectrometry. Final protein products were flash frozen, dropwise, in liquid nitrogen and stored at -80°C. Typical protein yields were ∼20mg/L of culture.

*Protein characterization*

Prior to all biophysical measurements, we thawed and exchanged all proteins into an appropriate buffer by two serial NAP-25 desalting columns (GE Health Science). We then used A280 to determine protein concentration using an empirical extinction coefficient for each protein. To determine extinction coefficients, we first used ProtParam [241, 242] to calculate the extinction coefficient for each protein in 6 M GdmHCl ($\varepsilon$6MGdm). We then measured the difference in A280 for an identical concentration of protein in native buffer versus in 6 M GdmHCl. We could then estimate a native extinction coefficient using the relationship $\varepsilon_{native} = \varepsilon_{6MGdm} \cdot A_{280,native}/A_{280,6MGdm}$. For some proteins no correction from the predicted extinction coefficient was necessary. Extinction coefficients used for calculation of protein concentration are as follows: (hA5: 5540 $M^{-1}cm^{-1}$, hA6:5434 $M^{-1}cm^{-1}$, tunA:1490 $M^{-1}cm^{-1}$, tunB: 5699 $M^{-1}cm^{-1}$, hA2:3230 $M^{-1}cm^{-1}$, hA4:3230 $M^{-1}cm^{-1}$, hA14:7115 $M^{-1}cm^{-1}$, hA1:8480 $M^{-1}cm^{-1}$, hA11:4595 $M^{-1}cm^{-1}$, hP:2980 $M^{-1}cm^{-1}$). We also corrected for scatter in all A280 measurements [243].

We performed ITC experiments in 25 mM buffer, 100mM NaCl at pH 7.4 that had been chelex-treated and filtered at 0.22 $\mu$m. We selected Tris or TES as the buffering species on a case-by-case basis to ensure observable heats of binding. We equilibrated and simultaneously degassed, either by application of vacuum to the solution or by centrifugation at 18,000 $\times$ g at the experimental temperature for 60 minutes. We dissolved metals ($CaCl_2$, $ZnCl_2$, or $CuCl_2$) directly into the experimental buffer immediately prior to each experiment. We performed all experiments at 25°C using a MicroCal ITC-200 or a MicroCal VP-ITC (GE Health Sciences). Data were collected using low gain or no gain, with 750 rpm syringe stir speed. Shot spacing ranged from 120s-2400s depending on gain settings and

relaxation time of the binding process. These setting were optimized on a per protein basis. Data were fit to one or two site models using the Origin 7 software. For binding curves with obvious 1:1 stoichiometry the one-site model in Origin was used. For data with apparent 2:1 stoichiometry, evident from location of inflection points in the data, a fit of the included two-site model was attempted. If the two-site model could not be fit, we then used a single-site binding model with a floating stoichiometry to extract an apparent binding constant across sites.

We collected far-UV circular dichroism data between 200–250 nm using a J-815 CD spectrometer (Jasco) with a 1 mm quartz spectrophotometer cell (Starna Cells, Inc. Catalog No. 1-Q-1). We prepared 20–50 $\mu$M samples in a TES buffer identical to that used for ITC. We centrifuged at 18,000 x g in a temperature-controlled centrifuge at the experimental temperature prior to experiments. We collected 5 scans for each condition, and then averaged the spectra and subtracted a blank buffer spectrum using the Jasco spectra analysis software suite. We converted raw ellipticity into mean molar ellipticity using the concentration and number of residues in each protein. We collected intrinsic tyrosine and/or tryptophan fluorescence using a J-815 CD spectrometer (Jasco) with an attached model FDT-455 fluorescence detector (Jasco) using a 1 cm quartz cuvette (Starna Cells, Inc.). We prepared 5–20 $\mu$M samples exactly as we did for our CD experiments. We collected 3–5 replicate scans for each condition, and then averaged the spectra and subtracted a blank buffer spectrum (averaged from 10–15 buffer blank spectra) using the Jasco spectra analysis software suite. For all spectroscopic measurements, we verified the reversibility of metal-induced changes to the spectra by measuring the apo spectrum, adding the appropriate metal and

re-measuring the spectrum, and then adding excess EDTA and re-measuring the spectrum.

*Native electrospray ionization time-of-flight mass spectrometry (nano ESI-MS)*

To prepare samples for mass spectrometry experiments small ($\sim$200$\mu$L) samples of the proteins used in MS experiments were dialyzed for at least 24 hr against 2–4 L of either 10 or 100mM ammonium acetate, pH 7.4 to remove salt and exchange into a more optimal buffer for MS. Samples were then diluted to $\sim$10$\mu$M in the dialysis buffer prior to experiment. All mass spectra were acquired using a Waters Synapt G2-Si ion-mobility mass spectrometer equipped with a nanoelectrospray (nanoESI) source and operated in Sensitivity mode. NanoESI emitters were pulled from borosilicate capillaries (ID 0.78 mm) to a tip ID of approximately 1$\mu$m using a Sutter Flaming-Brown P-97 micropipette puller. 3–5$\mu$L of sample were loaded into an emitter, a platinum wire was placed in electrical contact with the solution, and a potential of +0.8–1.2 kV was applied to the wire to initiate electrospray. The source temperature was equilibrated to ambient temperature, trap and transfer collision voltages were set to 25 V and 5 V, respectively, and the trap gas used was argon at a flow rate of 5 mL/min. Reported spectra are the sum of $\sim$3 minutes of continuously-collected data. Mass calibration was achieved using the series of $Cs(CsI)_n{}^{1+}$ peaks produced from nanoESI of 0.1 M aqueous cesium iodide (Aldrich).

We carefully controlled for spurious dimers in our nanoESI-MS experiments. Non-specific dimers (and high-order oligomers) can arise if, by chance, more than one monomer ends up in an electrospray drop. These non-specific aggregates are expected to follow a roughly Poisson distribution of oligomeric states, governed

by the bulk concentration of monomers in solution. These non-specific species can be distinguished from specific oligomeric species by measuring the mass spectrum over a wide range of protein concentrations. Dimers observed at 10 $\mu$M could be the result of non-specific interactions; dimers observed at 10 nM are almost certainly not. This can be seen by considering the distribution of non-specific species across drops. Under our instrumental conditions, electrospray creates drops $\sim$100–200 nm in diameter, meaning that 10 nM protein solution will yield drops that contain, on average, $\sim$0.003–0.025 protein molecules. Taking the upper limit of 0.025 protein molecules per drop, one would expect only 0.2% of drops to have non-specific dimers. Increasing to 100 nM protein takes this to 2.4% of drops. If one goes to 1 $\mu$M, non-specific dimers become quite significant (25.6%), but this is accompanied by a large number of non-specific trimers (21.4%). Although many factors, including relative ionization efficiency and instrumental conditions, can affect the observed abundances of ions formed from electrospray, these effects should be largely independent of initial solution concentration under the instrumental conditions used here.

We interpreted the mass spectra shown in Fig 8D and S14 using this logic. Mass spectra of proteins at low concentrations (10–100 nM) exhibit unexpectedly abundant monomers and dimers, consistent with a specific dimer. Mass spectra at high concentrations (1–10$\mu$M) exhibit dimers but not trimers, again consistent with a specific dimer rather than non-specific, Poisson-governed aggregation in drops. The small population of tetramer for tunB at 10 $\mu$M could either reflect a true tetramer or a random partitioning of two dimers into an electrospray drop at this high concentration.

*Sedimentation velocity analytical ultracentrifugation*

Samples were concentrated to ∼50$\mu$M and then dialyzed against 2L of 25 mM TES, 100mM NaCl, 1mM TCEP, pH 7.4) overnight at 4°C using 6000–80000 MWCO dialysis tubing. Prior to sedimentation velocity experiments proteins were then centrifuged at >18000 × g for 30 min. in a temperature-controlled centrifuge. AUC experiments were performed at 50k x g in sector-shaped cells with sapphire windows (Beckman) on a Beckman ProteomeLab XL-1 analytical ultracentrifuge. Due to the low extinction coefficients of the proteins, sedimentation was monitored using interference mode rather than absorbance at 280nm. Sedimentation velocity data was fit numerically to the Lamm equation and the c(s) distribution determined using SedFit [244, 245]. Estimated sedimentation coefficients and molecular masses of species present in solution were calculated from the fits.

*Homology model*

The homology model of tunB was constructed using Modeller 9.17 [246] using 46 $Ca^{2+}$ bound crystal structures (without bound peptide targets) as combined template(PDB:1e8a, 1gqm, 1j55, 1k96, 1k9k, 1mho, 1mr8, 1odb, 1qlk, 1xk4, 1xyd, 1yut, 1yuu, 1zfs, 2egd, 2h2k, 2h61, 2k7o, 2kay, 2l51, 2psr, 2q91, 2wnd, 2wor, 2wos, 2y5i, 3c1v, 3cga, 3cr2, 3cr4, 3cr5, 3czt, 3d0y, 3d10, 3gk1, 3gk2, 3gk4, 3hcm, 3icb, 3iqo, 3lk0, 3lk1, 3lle, 3m0w, 3psr, 3rlz, and 4duq). Alignment was generated using the PAIRWISE alignment method with default parameters. Model was generated as a dimer, with the single tunicate sequence mapped to both the A and B chains. Automodel was used to generate models, using default parameters. 20 models were generated and the best selected by DOPE score. The final model had an RMSD of

0.65 $\mathring{A}^2$ relative to the crystal structure of S100B bound to $Ca^{2+}$ and $Zn^{2+}$ (PDB: 3czt).

## Bridge to Chapter IV

In this chapter, the phylogenetic tree of the S100 protein family was reconstructed. Subsequently, the binding of transition metal ions to the proteins was mapped onto the phylogeny by using a large set of human paralogs as representative clade members. Some data were already available in the literature and these were incorporated into the analysis. The results indicated that binding of transition metal ions is an almost universally-conserved feature of the S100 family. However, binding stoichiometry, metal-driven conformational changes, binding site ligands, and binding site location varied across the family. It was thus concluded that binding of transition metals is a conserved feature of S100s at the level of activity, but has diversified extensively at the biochemical level. Two early-branching S100 proteins from the tunicate *Oikopleura dioica* were also characterized for the first time. These proteins displayed all the hallmark biochemical features of the more well-studied S100 proteins from higher metazoans: homodimerization, calcium-binding, transition metal binding, and metal-ion driven conformational changes. These results indicate that these canonical biochemical features of the S100s are ancestral to the family. This chapter highlights the striking evolutionary lability of an overall conserved biochemical feature, which has broader implications for understanding the evolutionary meanderings of protein traits. Chapter IV delves deeper into the biochemistry of metal binding in a specific member of the S100 protein family. S100A5 is a calcium binding protein found in a small subset of human tissues. Little is known about the biological roles of S100A5.

Previous work indicated that S100A5 displays antagonism between binding of Ca2+ and Cu2+ ions, which is one of the most commonly cited features of the protein. The interplay between Ca2+ and Cu2+ binding by S100A5 is further characterized in this chapter. It is shown that S100A5 can actually bind to both Ca2+ and Cu2+ simultaneously without antagonism. Furthermore, it is demonstrated that the apparent antagonism observed in previous studies is likely due to aggregation of the protein induced by binding of metals. This chapter highlights further the diversity of biochemical modifications found in the S100 family and provides important data on S100A5 that will be useful for other researchers trying to understand it's biological functions.

# CHAPTER IV

# HUMAN S100A5 BINDS CA2+ AND CU2+ INDEPENDENTLY

## Author Contributions

Lucas Wheeler and Michael Harms conceived the study and designed the experiments. Lucas Wheeler performed all experiments and data analysis. Michael Harms secured funding for the work. Lucas Wheeler wrote the manuscript and generated the figures. Both authors have read and approved the manuscript.

## Abstract

S100A5 is a calcium binding protein found in a small subset of amniote tissues. Little is known about the biological roles of S100A5, but it may be involved in inflammation and olfactory signaling. Previous work indicated that S100A5 displays antagonism between binding of $Ca^{2+}$ and $Cu^{2+}$ ions—one of the most commonly cited features of the protein. We set out to characterize the interplay between $Ca^{2+}$ and $Cu^{2+}$ binding by S100A5 using isothermal titration calorimetry (ITC), circular dichroism spectroscopy (CD), and analytical ultracentrifugation (AUC).

We found that human S100A5 is capable of binding both $Cu^{2+}$ and $Ca^{2+}$ ions simultaneously. The wildtype protein was extremely aggregation-prone in the presence of $Cu^{2+}$ and $Ca^{2+}$. A Cys-free version of S100A5, however, was not prone to precipitation or oligomerization. Mutation of the cysteines does not disrupt the binding of either $Ca^{2+}$ or $Cu^{2+}$ to S100A5. In the Cys-free background, we measured $Ca^{2+}$ and $Cu^{2+}$ binding in the presence and absence of the other metal

using ITC. Saturating concentrations of $Ca^{2+}$ or $Cu^{2+}$ do not disrupt the binding of one another. $Ca^{2+}$ and $Cu^{2+}$ binding induce structural changes in S100A5, which are measurable using CD spectroscopy. We show via sedimentation velocity AUC that the wildtype protein is prone to the formation of soluble oligomers, which are not present in Cys-free samples.

S100A5 can bind $Ca^{2+}$ and $Cu^{2+}$ ions simultaneously and independently. This observation is in direct contrast to previously-reported antagonism between binding of $Cu^{2+}$ and $Ca^{2+}$ ions. The previous result is likely due to metal-dependent aggregation. Little is known about the biology of S100A5, so an accurate understanding of the biochemistry is necessary to make informed biological hypotheses. Our observations suggest the possibility of independent biological functions for $Cu^{2+}$ and $Ca^{2+}$ binding by S100A5.

## Background

S100A5 is a member of the calcium-binding S100 protein family. The protein is primarily homodimeric and is capable of binding one $Ca^{2+}$ ion each at it's EF-hand and pseudo-EF-hand sites [207, 247]. S100A5 undergoes a notable conformational change upon calcium-binding, resulting in the rotation and extension of a helix [207]. This $Ca^{2+}$-driven exposure of a hydrophobic surface is the primary mode of signal transduction in the S100 proteins [94]. Through interactions with metals and protein targets, S100s play a variety of biological roles including control of cell proliferation, inflammatory signalling, and antimicrobial activity [101, 92, 103, 89].

S100A5 is expressed primarily in the olfactory bulb and olfactory sensory neurons (OSNs). Its expression is dramatically upregulated by odor stimulation

[248, 195, 249]. It has been proposed that S100A5 is actively involved in olfactory signalling due to its expression profile [195]. Expression of the protein has also been observed in a small number of other tissues [249]. It is used as a bio-marker for several types of brain cancers and inflammatory disorders and appears to be involved in inflammation via activation of RAGE [104, 247, 250]. Genetic work on S100A5 has been minimal, which has limited our understanding of its biological roles.

The first biochemical study of human S100A5 identified it as a novel $Ca^{2+}$, $Cu^{2+}$, and $Zn^{2+}$ binding protein [195]. The authors used flow-dialysis to measure binding of the metal ions to the protein and concluded that S100A5 is capable of binding four $Ca^{2+}$ ions, four $Cu^{2+}$ ions, and two $Zn^{2+}$ ions per homodimer. One of the most striking observations of that study was the strong antagonism between the binding of $Cu^{2+}$ and $Ca^{2+}$ ions to the protein. This feature is one of the most highly cited aspects of S100A5. Because little is known about the protein, this fact is present in descriptions found across databases such as Uniprot, NCBI, Wikigenes, and Genecards [251, 252, 253]. While most S100s are capable of binding transition metal ions, antagonism with binding of $Ca^{2+}$ is not known outside the S100A5 lineage. Thus, this unique feature of S100A5 provoked speculation about its possible biological implications [195, 170]. It was suggested that S100A5 might act as a $Cu^{2+}$ and $Ca^{2+}$ regulated signal during olfaction or as a $Cu^{2+}$ sink to accommodate high $Cu^{2+}$ concentrations in the olfactory bulb [195].

We sought to characterize this presumably important feature of S100A5 in more detail. Previously, we characterized the binding of $Cu^{2+}$ and $Zn^{2+}$ to a large number of S100 proteins including S100A5 [96]. Via ITC competition experiments, we established that these two metals bind at different sites on the protein and

do not compete for binding [96]. We found that mutation of Cys43 and Cys79 lead to a loss of $Zn^{2+}$ binding. In contrast neither of these residues was necessary for binding of $Cu^{2+}$. Due to the original report of $Ca^{2+}$ /$Cu^{2+}$ antagonism we suspected that $Ca^{2+}$ and $Cu^{2+}$ may compete for the same sites on S100A5.

Here we report our study of the interplay between $Ca^{2+}$ and $Cu^{2+}$ binding by S100A5. Using a Cysteine-free variant (C43S/C79S) of the protein, we show that binding of $Ca^{2+}$ and $Cu^{2+}$ are not in fact antagonistic. The protein is capable of binding the two metals—which induce notable structural changes—simultaneously and independently. Furthermore, we establish that the Cysteine-containing (WT) protein is prone to the formation of high-ordered oligomers in solution, while the Cysteine-free variant is almost entirely dimeric. We suggest that this propensity for formation of large oligomeric species and precipitation under our experimental conditions may underlie the apparent antagonism observed in the original S100A5 report. Our results may suggest new biological roles for $Cu^{2+}$ binding by this protein.

## Results

### *$Ca^{2+}$ and $Cu^{2+}$ binding to S100A5 are not antagonistic*

Antagonism between $Cu^{2+}$ and $Ca^{2+}$ binding was previously identified as a distinct feature of S100A5 relative to other S100 proteins [195, 170, 171]. We hypothesized that $Cu^{2+}$ and $Ca^{2+}$ may bind using the same ligands, thus explaining the antagonism as direct competition. It was suggested in the original paper that $Cu^{2+}$ and $Ca^{2+}$ might share some ligands [195]. We performed ITC competition experiments to test whether $Cu^{2+}$ and $Ca^{2+}$ directly compete. We titrated $Cu^{2+}$ onto S100A5 in the presence of saturating $Ca^{2+}$. However, these experiments were

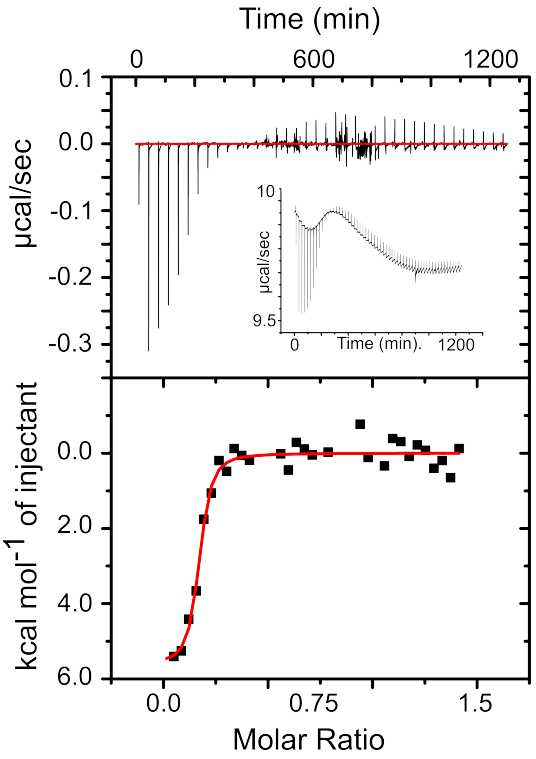


FIGURE 10 Measurements of $Cu^{2+}$ binding to wildtype S100A5 in the presence of $Ca^{2+}$ are difficult to interpret. Representative ITC trace showing $Cu^{2+}$ titrated onto $Ca^{2+}$-bound wildtype S100A5. Inset shows raw data trace. Data were characteristically noisy and the apparent fraction competent was systematically low.

difficult to interpret due to extensive precipitation in the samples containing both ions. ITC traces were very noisy and apparent stoichiometries were systematically low ($\approx$ 0.2), suggesting that a large portion of the protein sample was not competent to bind $Cu^{2+}$ (Figure 10). Together these observations suggested that a metal-driven aggregation process could be occurring in our samples.

We found previously that neither of the two native Cys residues in S100A5 were required for $Cu^{2+}$ binding [96]. We also noticed that—unlike the wildtype protein—the Cys-free mutant did not precipitate in the presence of saturating $Ca^{2+}$ and $Cu^{2+}$. We thus sought to use ITC to characterize the interaction between binding of the two metal ions using the Cys-Ser double mutant. Because some

of the metal-binding curves were complex and difficult to fit, we used a Bayesian Markov Chain Monte Carlo sampler—as implemented in pytc—to estimate thermodynamic parameters for all binding models [254]. We also included a floating "fraction competent" parameter to capture uncertainty in the relative protein and metal concentrations (following SEDPHAT [255]). This was necessary because a number of factors make it difficult to obtain accurate estimates of concentrations for components of this system. S100A5 has no tryptophan residues and, therefore, a low extinction coefficient that makes absorbance-based concentration estimates unreliable. Further, water absorption by dry metal salts, as well as interactions between metal ions and buffer, can also make estimates of metal concentration difficult. Because of these of uncertainties, ITC has been noted to provide poor estimates of stoichiometry for protein metal binding [256].

We first used ITC to remeasure binding of $Cu^{2+}$ ions to the apo form of the S100A5 double mutant. We found the $Cu^{2+}$ binding data was best described with a single-site binding model. In line with our previous observations, the protein bound $Cu^{2+}$ with a $K_d(\mu M)$ that had a 95% credibility region of $0.94 \leq 1.81 \leq 3.90$ (Figure 11A, Table 1). We next measured the binding of $Cu^{2+}$ in the presence of saturating $Ca^{2+}$. $Ca^{2+}$ had no detectable effect on the binding of $Cu^{2+}$ to the protein, giving a $K_d(\mu M)$ of $0.65 \leq 0.96 \leq 1.47$ (Figure 11B; Table 1).

We next performed the inverse set of experiments. We used ITC to measure the binding of $Ca^{2+}$ to the protein in the apo and $Cu^{2+}$ saturated forms. For each condition, we used four different titrant/stationary ratios to better resolve the complex $Ca^{2+}$ binding curve and then globally fit a binding model to all four datasets (Figure 11C). This binding curve had two distinct phases and could be fit with a two-site binding polynomial (Figure 11C). These $Ca^{2+}$ binding curves

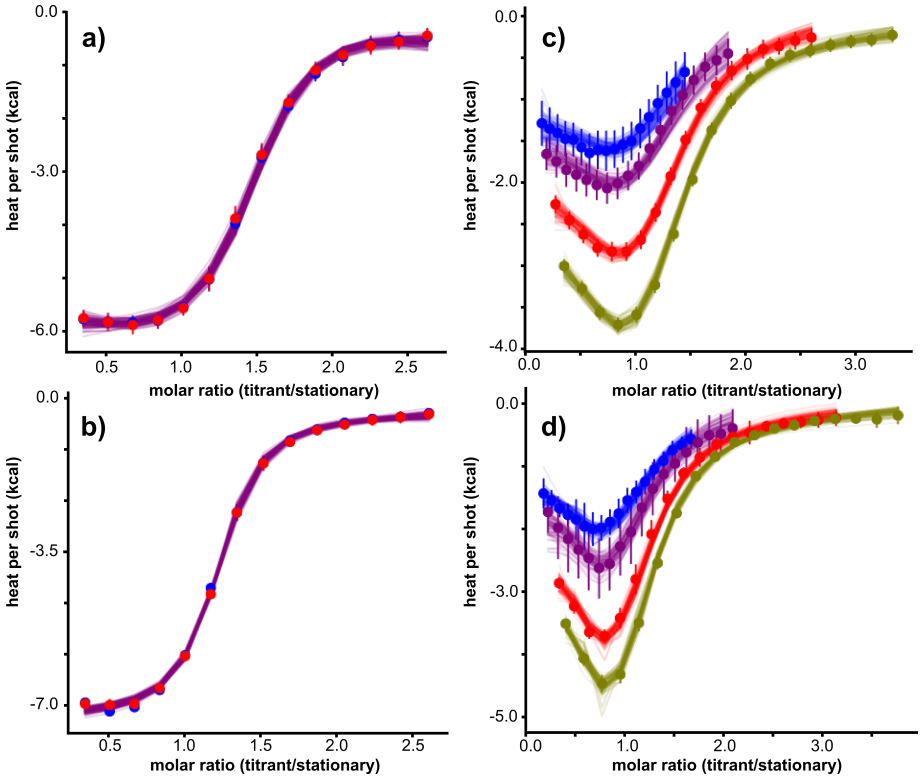


FIGURE 11 S100A5 can bind $Ca^{2+}$ and $Cu^{2+}$ simultaneously without antagonism. Plots show integrated data and global Bayesian fits from replicate isothermal titration calorimetry experiments: a) $Cu^{2+}$ binding to apo protein, b) $Cu^{2+}$ binding to $Ca^{2+}$-saturated protein, c) $Ca^{2+}$ binding to apo protein, and d) $Ca^{2+}$ binding to $Cu^{2+}$-saturated protein. Points are integrated titration shots. Lines are 100 curves drawn from the posterior distribution of the MCMC samples. For $Cu^{2+}$ binding experiments technical replicates are shown in blue and red. $Ca^{2+}$ binding experiments were performed with fixed protein concentration and four different titrant/titrate ratios: 8X (blue), 10X (purple), 15X (red), and 18X (green). For clarity Y-axes display total heat per shot, so that curves from different titrant concentrations fall on different areas of the graph. Raw data corresponding to these integrated heats are displayed in figure 31 in supplement.

presented a challenging model-fitting problem due to the complex shape of the curve. The individual enthalpies and binding constants may therefore be underdetermined in our analysis. To resolve realistic parameter values from the binding polynomial model, we constrained the dilution heat and dilution intercept in the Bayesian fit to reasonable values.

TABLE 1 Table contains values for key parameters determined via global fits of ITC data using the Bayesian MCMC fitter in pytc. 95% credibility regions from the posterior distributions are reported for parameter values. $\Delta H$ values are reported in $kcal \cdot mol^{-1}$, $K_d$ values in $\mu M$. Final parameter is fraction competent, a nuisance parameter that captures what fraction of the metal and protein in solution are competent for the measured reaction.

| ion | $Cu^{2+}$ | | $Ca^{2+}$ | |
|---|---|---|---|---|
| competitor | none | $Ca^{2+}$ | none | $Cu^{2+}$ |
| $\Delta H_1$ | $-5.7 \leq -3.4 \leq -2.8$ | $-4.1 \leq -3.6 \leq -3.2$ | $-1.8 \leq -1.4 \leq -0.7$ | $-1.5 \leq -1.2 \leq -0.7$ |
| $\Delta H_2$ | — | — | $-5.7 \leq -4.6 \leq -3.7$ | $-5.0 \leq -4.1 \leq -3.4$ |
| $K_{d,1}$ | $0.9 \leq 1.8 \leq 3.9$ | $0.7 \leq 1.0 \leq 1.5$ | $0.4 \leq 0.5 \leq 2.7$ | $0.03 \leq 0.2 \leq 2.2$ |
| $K_{d,2}$ | — | — | $1.9 \leq 6.3 \leq 34.9$ | $1.9 \leq 10.5 \leq 100$ |
| fx. comp. | $1.40 \leq 1.43 \leq 1.47$ | $1.15 \leq 1.17 \leq 1.19$ | $0.61 \leq 0.66 \leq 0.69$ | $0.54 \leq 0.58 \leq 0.61$ |

We observed one high-affinity site ($K_d(\mu M)$: $0.14 \leq 0.46 \leq 2.68$) and one lower-affinity site ($K_d(\mu M)$: $1.85 \leq 6.33 \leq 34.88$). The values were roughly consistent with those reported in the literature [247]. The presence of saturating $Cu^{2+}$ did not inhibit the binding of $Ca^{2+}$ ions (Figure 11D; Table 1). The $K_d$ value of the low affinity site ($K_d(\mu M)$: $0.03 \leq 0.18 \leq 2.16$) was not distinguishable within uncertainty from that of the apo protein. The $K_d$ of the high affinity site ($K_d(\mu M)$: $1.86 \leq 10.46 \leq 100.3$) is similarly indistinguishable from that for the apo protein (Table 1). Our results clearly demonstrate that $Ca^{2+}$ and $Cu^{2+}$ ions do not display strong antagonism when binding to S100A5.

### *S100A5 is prone to oligomerization and metal-driven aggregation*

We hypothesized that the metal-driven aggregation process observed in our ITC experiments with the wildtype protein contributed to the apparent antagonism

that was previously reported. To further examine this aggregation process we used sedimentation velocity AUC to test for the presence of oligomers in solution. We hypothesized that the oligomerization of the wildtype protein was driven by the presence of Cysteine residues. Due to the presence of $Cu^{2+}$ in some samples we were unable to use a reducing agent in either the ITC or AUC experiments.

We performed sedimentation velocity AUC experiments on both the wildtype and Cys-Ser double mutant proteins in both the apo form and the form loaded simultaneously with $Cu^{2+}$ and $Ca^{2+}$. We fit the Lamm equation to the sedimentation data using SedFit to calculate the c(s) distribution of the protein in each condition [244, 245]. We found that apo S100A5 formed high-ordered oligomers, ranging to at least dodecamers (Figure 12A). Addition of $Cu^{2+}$ and $Ca^{2+}$ caused a large amount of precipitation in wildtype S100A5 that we removed by extensive centrifugation prior to loading the cell. The remaining soluble protein was indistinguishable from the apo protein (Figure 12). In contrast, when we performed the same experiments with the Cys-Ser double mutant we found that the protein was primarily dimeric in solution (Figure 12C, 12D), even with the addition of $Cu^{2+}$ and $Ca^{2+}$. However, monomers were also detectable in the double mutant samples. The monomer peak appears to be more prominent in the apo-protein sample than in the sample saturated with $Cu^{2+}$ and $Ca^{2+}$, suggesting that binding of metals may stabilize the dimeric form (Figure 12C, 12D). Our AUC results clearly demonstrate that oligomerization of S100A5 is driven by the native cysteine residues, which also likely cause the visible aggregation we observed in the ITC experiments. This observation strongly suggests that the previously-reported apparent antagonism between $Ca^{2+}$ and $Cu^{2+}$ was due to oligomerization and/or aggregation.

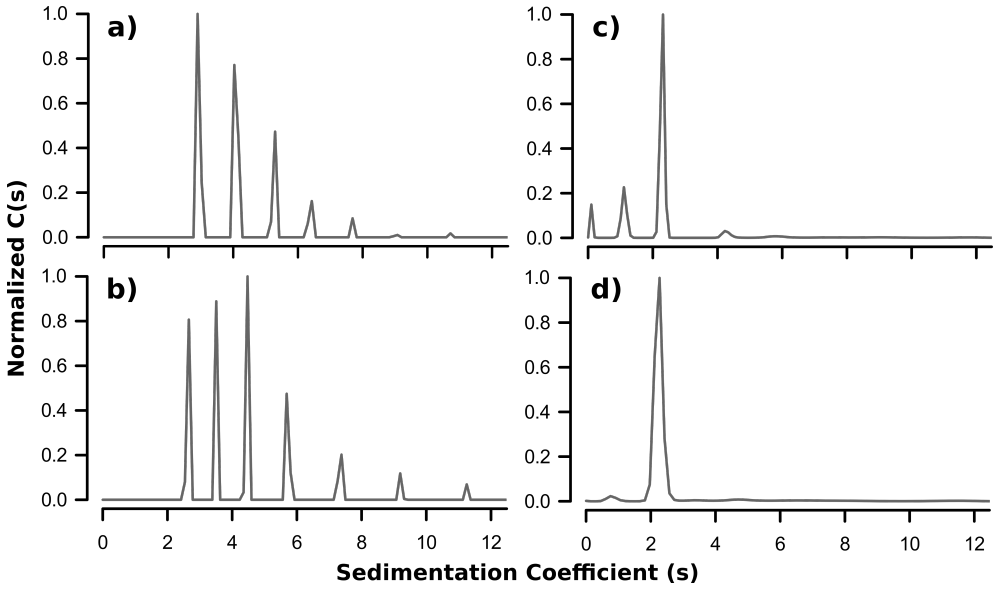


FIGURE 12 Wildtype S100A5 is prone to the formation of high-ordered oligomers. Sedimentation velocity AUC distribution plots showing a) apo wildtype S100A5, b) wildtype S100A5 saturated with $Cu^{2+}$ and $Ca^{2+}$, c) apo Cys-Ser double mutant, and d) Cys-Ser double mutant saturated with $Cu^{2+}$ and $Ca^{2+}$. Data are normalized to the same scale. Homodimers are the peaks near $s = 2$. The Cys-Ser double mutant plots show evidence of some monomer (peak near $s = 1$) in solution.

*Binding of $Ca^{2+}$ and $Cu^{2+}$ induce reversible changes in S100A5 secondary structure*

One hallmark feature of the S100 proteins is the change in secondary structure observed upon binding of metal ions [96, 166, 199]. Metal-induced conformational changes expose a binding interface that can bind downstream targets and regulate their activities [94, 247]. In the original publication on S100A5 biochemical characterization, the authors found that the secondary structure of the protein is insensitive to the binding of metal ions [195]. However, we previously found that binding of $Ca^{2+}$ ions to wildtype S100A5 induces a significant (≈25%) reversible increase in $\alpha$-helical secondary structure, which is consistent with the changes observed in published NMR data [207, 247]. Due to instantaneous sample precipitation, we were unable to reliably measure structural changes of wildtype S100A5 in the presence of $Cu^{2+}$. However, the Cys-Ser double mutant protein alleviates this issue. We collected far-UV circular dichroism spectra of the mutant protein in the apo form and bound to $Ca^{2+}$, $Cu^{2+}$, and $Ca^{2+}$ and $Cu^{2+}$ simultaneously. The Cys-Ser mutant displays a notable increase in alpha-helical signal (222nm) upon binding of $Ca^{2+}$, identical to the wildtype protein. Interestingly, addition of $Cu^{2+}$ also induces an increase in $\alpha$-helical signal that is approximately half of that induced by $Ca^{2+}$. The spectrum of S100A5 bound simultaneously to both metals is identical to that of the $Ca^{2+}$-bound form (Figure 13). This structural change is not due to oligomerization, as the protein remains a dimer under these conditions by AUC (Figure 13D). All the metal induced structural changes were instantly reversible by the addition of a molar excess of EDTA. These results may help to explain the minor differences—such as larger enthalpy—in $Ca^{2+}$ binding to the $Cu^{2+}$ bound form of the protein, which may

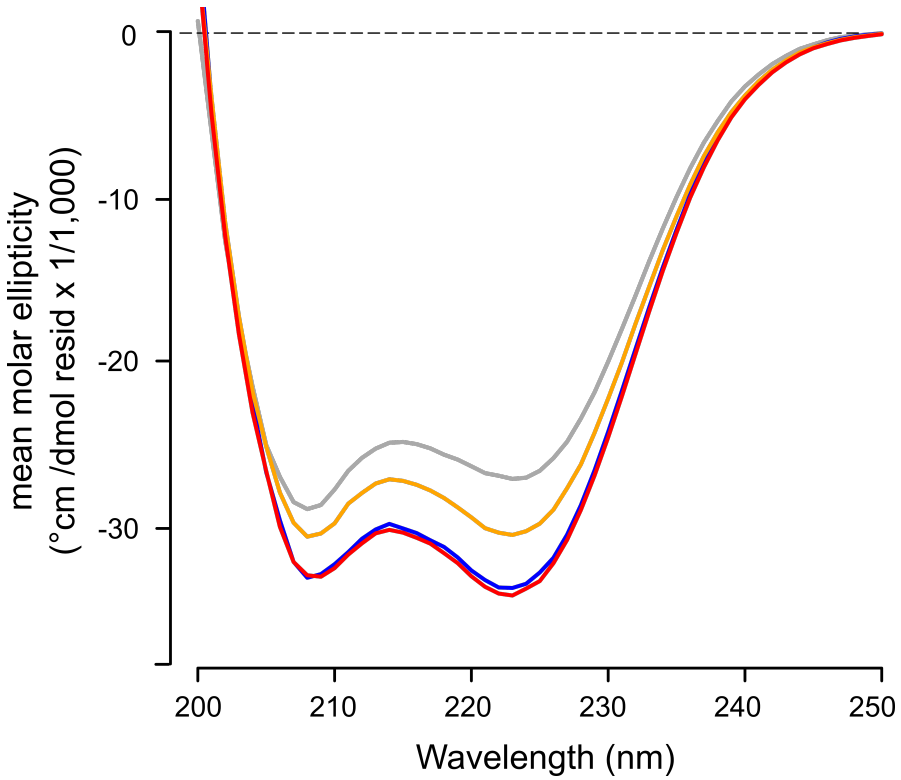


FIGURE 13 $Ca^{2+}$ and $Cu^{2+}$ induce increases in $\alpha$-helical secondary structure measured by far UV circular dichroism. Curves show mean molar ellipticity vs. wavelength for each experimental condition: Apo (gray), bound to $Cu^{2+}$ (orange), bound to $Ca^{2+}$ (blue), and bound to both $Cu^{2+}$ and $Ca^{2+}$ (red).

be due to moderate structural differences from the apo-protein. Despite the lack of antagonism between binding affinities for $Ca^{2+}$ and $Cu^{2+}$ ions, there is still indication of some structural interplay between the two metals.

## Discussion

S100A5 is one of the lesser-known members of the S100 protein family. Its expression pattern is very narrow and its biological functions are mostly uncharacterized. However, it has been the target of multiple biochemical studies that have sought to characterize the properties of the protein itself. Binding of metals and proteins to S100A5 have been studied using various techniques

[195, 96, 207, 104, 247]. X-ray crystallography and NMR have been used to solve structures of both apo and $Ca^{2+}$ bound forms of the protein [207, 257]. Despite the available biochemical data aspects of S100A5 have remained ambiguous. For example, the stoichiometry of transition metal binding and structural responses to metal binding have been variably reported [195, 96].

One of the most noted features of S100A5 is the strong antagonism between binding of $Ca^{2+}$ and $Cu^{2+}$ ions. This feature is reported in the gene descriptions found in many databases [252, 253, 251]. In this study we set out to characterize this unique feature of S100A5, hypothesizing that it was due to competition between the two metals for shared ligands. However, we found an absence of direct binding antagonism between $Ca^{2+}$ and $Cu^{2+}$. Neither metal ion affects the binding constant for the other. Instead, we observed a propensity of the protein for oligomerization and metal-induced aggregation. It is possible that the reduction of binding-competent protein caused by this aggregation process was interpreted in the original flow dialysis study of S100A5 as antagonism between $Ca^{2+}$ and $Cu^{2+}$. We also report notable changes in the secondary structure of S100A5 upon binding of both $Ca^{2+}$ and $Cu^{2+}$, which is contrary the original report that S100A5 structure is insensitive to the binding of metals.

One intriguing implication of our observations is that the $Cu^{2+}$ binding site of S100A5 must be quite distinct from that of other S100 proteins. $Ca^{2+}$ and $Cu^{2+}$ clearly do not share ligands, or there would be evidence of competition in our ITC experiments. Cysteine residues are thought to be involved in metal-binding in some other S100s [177, 170] and we previously showed that the Cys-free mutant of S100A5 displays compromised $Zn^{2+}$ binding [96]. However, neither native Cys residue of S100A5 is required for $Cu^{2+}$ binding. Furthermore, we showed that $Zn^{2+}$

and $Cu^{2+}$ do not share ligands, as they do not compete at all in ITC experiments [96]. In addition, mutation of His17—which is present in the canonical transition metal site of many S100s—also had no effect on $Cu^{2+}$ binding in S100A5 [96]. The results presented here with the Cys-free mutant also clearly rule out the possibility of oligomer-dependent $Cu^{2+}$ binding, such as could be achieved by the formation of a new site in a high-order oligomeric species. Thus, we still have no clues as to where $Cu^{2+}$ ions bind on S100A5. Further characterization—such as via scanning mutagenesis—will be necessary to determine the identity of $Cu^{2+}$ ligands.

Biological roles for the binding of transition metals have been established for some S100s and suggested for many others [170, 171, 105, 177, 172]. The binding constants that we measured for $Ca^{2+}$ and $Cu^{2+}$ suggest the possibility of physiologically relevant interactions in some tissues. Free $Ca^{2+}$ concentrations in rat olfactory neurons reach $\approx 2\ \mu M$ during nerve stimulation [258]. Likewise, pools of $Cu^{2+}$ are released in and around olfactory neurons during signaling, reaching concentrations as high as 10 $\mu M$ in the synapse [259, 260, 223, 261]. Further, despite high $Cu^{2+}$ concentrations, the olfactory bulb in rats does not have elevated expression of the typical copper chaperone metallothionein [262]. It has been suggested that S100A5 may play a role as a $Cu^{2+}$ buffer or chaperone in OSNs during olfactory signaling [195]. The fact that $Cu^{2+}$ is able to induce structural changes in S100A5 suggests it could play a more active role: S100A5 could actually respond to $Cu^{2+}$ and propagate a resulting signal by interacting with downstream targets.

Due to lack of antagonism, $Cu^{2+}$ dependent functions could be achieved even in the presence of saturating $Ca^{2+}$ levels. Furthermore, there could be synergistic functional roles for binding of $Ca^{2+}$ and $Cu^{2+}$. For example, if S100A5 is acting as

a $Cu^{2+}$ chaperone, binding of $Ca^{2+}$ could facilitate binding of protein targets—via exposure of the hydrophobic binding interface—to which $Cu^{2+}$ is being delivered. Furthermore, S100A5 is capable of binding $Zn^{2+}$ ions—which are also at high concentration in the olfactory bulb—with similar affinity to $Cu^{2+}$ [261]. $Zn^{2+}$ and $Cu^{2+}$ also bind noncompetitively and thus all three metals could potentially engage in synergistic activities [96].

One final possibility is that the oligomerization process we observed in this study may actually have a biological function. Wildtype S100A5 is prone to the formation of oligomers even in the apo form and is subject to extensive aggregation in solutions containing $Ca^{2+}$ and $Cu^{2+}$ even at relatively low protein concentrations. Roles for metal-driven oligomerization in S100s have been suggested previously [89, 211, 216, 263]. It is conceivable that $Ca^{2+}$ and $Cu^{2+}$ drive oligomerization of S100A5 in cells to facilitate a biological function, but further experiments would be required to determine if this process occurs in the reducing environment of the cell at physiologically-relevant concentrations of S100A5, $Ca^{2+}$ and $Cu^{2+}$.

Future experiments are needed to elucidate the biochemical features and biological functions of S100A5 that remain unknown. It will be important to identify the $Cu^{2+}$ ligands in S100A5 to fully understand the biochemical interplay between the binding of various biologically relevent metals. To understand how $Ca^{2+}$, $Cu^{2+}$, and $Zn^{2+}$ contribute to the biological activity of S100A5, experiments should be targeted at directly testing how these metals interact with the protein *in vivo*. The identification of more S100A5 biological targets and an increase in functional studies will be required to determine the chief roles of S100A5 in it's cellular environment.

## Conclusions

Antagonism between binding of $Ca^{2+}$ and $Cu^{2+}$ ions to S100A5 is one of the most oft-cited aspects of this protein. Several possible biological roles have been suggested. Using careful biophysical characterization, we discovered that binding of $Ca^{2+}$ and $Cu^{2+}$ ions is not antagonistic. A Cys-free mutant version of the protein makes measurements of metal binding using ITC possible and shows that the protein is capable of binding both metals simultaneously and independently. Rather than binding antagonism, it appears that the wildtype protein is prone to oligomerization and aggregation and that these behaviors may have contributed to the original interpretation. Furthermore, we also measured the effects of $Ca^{2+}$ and $Cu^{2+}$ binding on S100A5 secondary structure and found that both metals are capable of inducing increases in $\alpha$-helical secondary character. These results also contrast the original report on S100A5 [195], but are consistent with previously published NMR data [207]. The ability to bind $Ca^{2+}$ and $Cu^{2+}$ independently as well as the structural response to $Cu^{2+}$ may suggest new $Cu^{2+}$ dependent biological roles for S100A5.

## Methods

### *Protein expression and purification*

We previously generated the 6-histidine-tagged cysteine double-mutant construct in a pet28/30 vector [96]. In this study, the protein was expressed and purified using the same protocol detailed in the previous publication. Briefly, the protein was expressed in a 1.5L culture of Rosetta (DE3) pLysS cells (Millipore). Cells were lysed by sonication and treatment with DNase and

lysozyme. Subsequently, the tagged protein was purified using HisTrap $Ni^{2+}$ affinity columns (GE). The tag was then cleaved using TEV protease and the cleaved protein was further purified using $Ca^{2+}$-dependent hydrophobic interaction chromatography. Finally, the sample was run over a second HisTrap $Ni^{2+}$ affinity column to remove any uncleaved protein. The purified protein was dialyzed with 6000-8000 MWCO tubing (Fisher) against 2L 25 mM Tris, 100 mM NaCl, pH 7.4 with 2g chelex resin (BioRad). The dialyzed protein was filter-sterilized (0.22 $\mu m$), flash-frozen dropwise in liquid nitrogen, and stored at $-80°C$. We experimentally determined the extinction coefficient ($5002 M^{-1} cm^{-1}$) of the Cys-Ser double mutant. We measured the $A_{280}$ of the protein at the same concentration in both buffer and denaturing 6M GdHCl (Sigma). We used ProtParam [241] to predict an extinction coefficient for the protein based on sequence and then calculated the corrected coefficient using the equation $\varepsilon_{native} = \varepsilon_{6MGdm} \cdot A_{280,native}/A_{280,6MGdm}$. Concentration measurements were also corrected for scattering in samples [243]. Due to the low extinction coefficient of the protein, concentration is difficult to measure with high confidence, even with this careful protocol.

*Isothermal titration calorimetry*

Samples were prepared in 25 mM TES (Sigma), 100 mM NaCl (Thermo Scientific), buffer at pH 7.4. Protein was thawed from a frozen stock and exchanged into the experimental buffer using NAP-25 desalting columns (GE Healthcare). For competition experiments the experimental buffer also contained either 1 mM $CaCl_2$ (Sigma) or 0.25 mM $CuCl_2$ (Sigma). Titrant solutions were prepared in matching experimental buffer to ensure identical conditions to titrate. Anhydrous $CaCl_2$ or $CuCl_2$ was dissolved directly in the buffer and diluted to the appropriate

concentration immediately prior to experiments. Fresh stocks were made for each set of experiments. Experiments were performed with 50-80 $\mu M$ protein at 25°$C$. Two technical replicates of each $Cu^{2+}$ binding experiment were performed. To resolve the complex $Ca^{2+}$ binding curves, four $Ca^{2+}$ binding experiments were performed using four different concentrations of titrant. Raw data were integrated using the NITPIC software package—which allows uncertainty in the baseline—and the integrated heats were exported in standard SedPhat format [264]. We then used the Bayesian MCMC iterator included in pytc to estimate model parameters against all experiments simultaneously [254]. We used the maximum likelihood estimate as a starting point and then explored the likelihood surface with 100 walkers, each taking 20,000 steps. We discarded the first 10% of steps as burn in. We restricted parameters against all experiments simultaneously. We verified convergence by performing the sampling procedure several times. A single site binding model was used for $Cu^{2+}$ titration data and a two-site binding polynomial was used for $Ca^{2+}$ titration data [265, 266]. For $Ca^{2+}$ binding fits, we constrained the dilution heat and dilution intercept to between -3.0—0.0 kcal/mol and 0—10,000 kcal/mol/shot respectively. All other priors were uniform.

*Sedimentation velocity analytical ultracentrifugation*

Experiments were done in 25 mM TES (Sigma), 100 mM NaCl (Thermo Scientific), 100 $\mu M$ EDTA at pH 7.4 with the appropriate metal added directly to the buffer during preparation. Metals were added to a final concentration of 250 $\mu M$. Samples were prepared at 40 $\mu M$ in the appropriate experimental buffer by overnight dialysis (6000-8000 MWCO) against 2L at 4°$C$. Before ultracentrifugation samples were centrifuged at $18,000 \times g$ at 4°$C$ in a temperature-

controlled centrifuge for 30 minutes. Ultracentrifugation was done with sapphire windows at $50,000 \times g$ in sector-shaped cells (Beckman) on a Beckman ProteomeLab XL-1. Sedimentation was monitored using interference mode rather than absorbance at 280nm due to the low extinction coefficient of S100A5. The Lamm equation was fit to the sedimentation data—using SedFit—to calculate the continuous c(s) distribution [244, 245]. Estimated sedimentation coefficients of the species present in solution were calculated from the numerical fits.

*Circular dichroism spectroscopy*

Far-UV circular dichroism spectra (200250nm) were collected on a J-815 CD spectrometer (Jasco) with a 1 mm quartz cell (Starna Cells, Inc.). We prepared 50 $\mu M$ samples in a Chelex (Bio-Rad) treated, 25 mM TES (Sigma), 100 mM NaCl (Thermo Scientific), 100 $\mu M$ EDTA, buffer at pH 7.4. Samples were subsequently diluted to 25 $\mu M$ in buffers containing: no metal (apo), 1 mM $Ca^{2+}$, 1 mM $Cu^{2+}$, or both 1 mM $Ca^{2+}$ and 1 mM $Cu^{2+}$ —all prepared in the stock buffer above. Samples were centrifuged at $18,000 \times g$ at $25°C$ in a temperature-controlled centrifuge (Eppendorf) before experiments. Spectra were collected at $25°C$ in a Jasco peltier multi-cell sample unit. Reversibility of metal-induced structural changes was confirmed by adding a molar excess of EDTA to the metal-saturated samples and repeating spectra collection. In all cases, addition of EDTA returned the samples to the apo state. Five scans of each condition were collected. These scans were then averaged—using Jasco spectra analysis software—to minimize noise. Buffer blank spectra were generated for each condition. Applicable blanks were subtracted in the Jasco spectra analysis software. Blank-corrected data were exported as text files and raw signal was converted into mean molar ellipticity using

the concentration and the number of residues ($N_{res} = 95$) in our S100A5 construct using the equation: $MME = CD_{signal}/c(M) \cdot 10 \cdot L(cm) \cdot N_{res}$.

## Bridge to Chapter V

In this chapter, the metal binding behavior of S100A5 was characterized biophysically. A misunderstood aspect of S100A5 biochemistry was resolved. It has long been thought that S100A5 exhibits strong antagonism between the binding of Ca2+ and Cu2+ ions. However, it is demonstrated here that this antagonistic behavior was likely an artifact of techniques used in the original biochemical study of S100A5. Instead, it is shown that the protein is prone to the formation of high-ordered oligomeric species. By eliminating this oligomerization process with point mutations the metal binding behavior of S100A5 could be characterized. The protein is capable of binding Ca2+ and Cu2+ ions simultaneously. Additionally, the two metals were observed to induced distinct conformational changes in the protein. This chapter is an important addition to the S100 literature, because it overturns an erroneous paradigm and suggests new biological roles for Ca2+ and Cu2+ ion binding by S100A5. Chapter 6 turns to another basic biochemical feature of the S100 protein family; binding of small peptide regions of target proteins. Two S100 proteins, S100A5 and S100A6, that arose via gene duplication in the ancestor of amniotes are used as a model to study the diversification of binding specificity in duplicate lineages. Ancestral sequence reconstruction is used to resurrect the last common ancestor of all S100A5 and S100A6 proteins, allowing the evolutionary history of binding specificity to be directly characterized. The proteins are shown to display conserved specificity at the level of gene clades and both lineages appear to have subfunctionalized relative to the ancestor. The work in chapter

VI demonstrates that proteins with low biochemical specificity nonetheless display evolutionary patterns consistent with those observed in highly-specific proteins following gene duplication.

# CHAPTER V

# CONSERVATION OF PEPTIDE BINDING SPECIFICITY IN S100A5 AND S100A6

## Author Contributions

Lucas Wheeler and Michael Harms conceived the study and designed the experiments. Lucas Wheeler, Jeremy Anderson, Anneliese Morrison, and Caitlyn Wong performed the experiments. Lucas Wheeler and Michael Harms analyzed the experimental datasets. Michael Harms secured funding for the work. Michael Harms and Lucas Wheeler wrote the manuscript and generated figures. All authors have read and approved the manuscript.

## Abstract

S100 proteins bind linear peptide regions of target proteins and modulate their activity. The peptide binding interface, however, has remarkably low specificity and can interact with many target peptides. It is not clear if the interface discriminates targets in a biological context, or whether biological specificity is achieved exclusively through external factors such as subcellular localization. To discriminate these possibilities, we used an evolutionary biochemical approach to trace the evolution of paralogs S100A5 and S100A6. We first used isothermal titration calorimetry to study the binding of a collection of peptides with diverse sequence, hydrophobicity, and charge to human S100A5 and S100A6. These proteins bound distinct, but overlapping, sets of peptide targets. We then studied the peptide binding properties of S100A5 and S100A6 orthologs

sampled from across five representative amniote species. We found that the pattern of binding specificity was conserved along all lineages, for the last 320 million years, despite the low specificity of each protein. We next used Ancestral Sequence Reconstruction to determine the binding specificity of the last common ancestor of the paralogs. We found the ancestor bound the whole set of peptides bound by modern S100A5 and S100A6 proteins, suggesting that paralog specificity evolved by subfunctionalization. To rule out the possibility that specificity is conserved because it is difficult to modify, we identified a single historical mutation that, when reverted in human S100A5, gave it the ability to bind an S100A6–specific peptide. These results indicate that there are strong evolutionary constraints on peptide binding specificity, and that, despite being able to bind a large number of targets, the specificity of S100 peptide interfaces is indeed important for the biology of these proteins.

## Introduction

Many proteins have low specificity interfaces that can interact with a wide variety of targets [267, 84, 81, 87, 268, 269, 270, 85, 145, 79, 82]. Such interfaces are difficult to dissect. Crucially, it is not obvious that their specificity is biologically meaningful: maybe such proteins are essentially indiscriminate, and biological specificity is encoded by external factors such as subcellular localization or expression pattern [83, 81, 93].

An evolutionary perspective allows us to probe whether specificity is, indeed, an important aspect of these interfaces [43]. If there are functional and evolutionary constraints on binding partners, we would expect conservation of binding specificity similar to that observed for high–specificity protein families

[52, 53]. In contrast, if specificity is unimportant, we would expect it to fluctuate randomly over evolutionary time. Further, previous work on the evolution of specificity has revealed common patterns for the evolution of specificity [271, 76, 47], including partitioning of ancestral binding partners among descendant lineages [58, 272, 55, 273] and transitions through more promiscuous intermediates [57, 79, 143]. If low–specificity proteins exhibit similar patterns, it is strong evidence that the low specificity interface has conserved binding properties, and that the interface makes a meaningful contribution to biological specificity.

S100 proteins are an important group of low–specificity proteins [100, 89]. Members of the family act as metal sensors [172], pro–inflammatory signals [274, 156, 101, 104], and antimicrobial peptides [105]. Most S100s bind to linear peptide regions of target proteins via a short hydrophobic interface exposed on $Ca^{2+}$–binding (Fig 14A). S100s recognize extremely diverse protein targets [94, 89, 275]. No simple sequence motif for discriminating binders from non–binders has yet been defined. The breadth of targets is much more extreme than other low–specificity proteins such as kinases and some hub proteins, which recognize well–defined, but degenerate, sequence motifs [267, 81, 269, 79, 82].

We set out to determine whether there was conserved specificity for two S100 paralogs, S100A5 and S100A6. These proteins arose by gene duplication in the amniote ancestor $\approx$ 320 million years ago [276, 96]. S100A6 regulates the cell cycle and cellular motility in response to stress [277]. It binds to many targets including p53 [278, 279], RAGE [101], Annexin A1 [275], and Siah–interacting protein [280]. A crystal structure of human S100A6 bound to a fragment of Siah–interacting protein revealed that peptides bind via the canonical hydrophobic interface shared by most S100 proteins [280]. The biology of S100A5 is less well

understood. It binds both RAGE [101, 104] and a fragment of the protein NCX1 [281] at the canonical binding site. It is highly expressed in mammalian olfactory tissues [282, 283, 284], but its specific targets and their biological roles are not well understood.

Using a combination of *in vitro* biochemistry and molecular phylogenetics, we addressed three key questions regarding the evolution of specificity in S100A5 and S100A6. First: do the two human proteins exhibit specificity relative to one another? Second: is the set of binding partners recognized by each protein fixed over time, or does the set of partners fluctuate? And, third: do we see similar patterns of specificity change after gene duplication for these low–specificity proteins compared to high–specificity proteins? Unsurprisingly, we find that S100A5 and S100A6 both bind to a wide variety of diverse peptides. Surprisingly, we find that the set of partners, despite being diverse, has been conserved over hundreds of millions of years. Further, we observe a pattern of subfunctionalization for these low–specificity proteins that is identical to that observed in high–specificity proteins. This suggests that these low–specificity interfaces are indeed under selection to maintain a specific—if large—set of binding targets.

## Results

### *Human S100A5 and S100A6 interact with diverse peptides at the same binding site*

We first systematically compared the binding specificity of human S100A5 (hA5) relative to human S100A6 (hA6) for a collection of six peptides (Fig 14B). Peptide targets have been reported for both hA5 and hA6 [280, 101, 278, 275, 279, 281, 104], but only two targets have been directly compared between paralogs. Using Isothermal Titration Calorimetry (ITC), Streicher and colleagues found

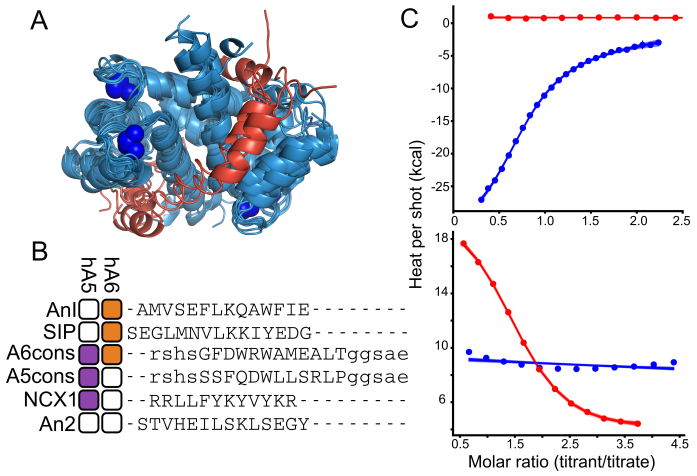


FIGURE 14 Human S100A5 and S100A6 exhibit peptide binding specificity. A) Published structures of S100 family members bound to both $Ca^{2+}$ and peptide targets at the canonical hydrophobic interface (PDB: 3IQQ, 1QLS, 3RM1, 2KRF, 4ETO, 2KBM, 1MWN, 3ZWH). Structures are aligned to the $Ca^{2+}$–bound structure of human S100A5 (2KAY). Peptides are shown in red. Blue spheres are $Ca^{2+}$ ions. B) Binding specificity of hA5 and hA6. Boxes indicate whether the peptide binds to hA5 (purple) and/or hA6 (orange). If peptide does not bind by ITC ($K_D$ >100 $\mu M$), the box is white. Peptide names are indicated on the left. Peptide sequences, aligned using MUSCLE [285], are shown on the right. Solubilizing flanks, which contribute minimally to binding (Table 3 in supplement), are shown in lowercase letters. Annexin 1 (An1) and Annexin 2 (An2) binding measurements are from a published study [275]. C) ITC heats for the titration of NCX1 (blue) and SIP (red) peptides onto hA5 (top) and hA6 (bottom). Points are integrated heats extracted from each shot. Lines are 100 different fit solutions drawn from the fit posterior probability distributions. For the hA5/NCX1 and hA6/SIP curves, we used a single–site binding model. For hA5/SIP and hA6/NCX1, we used a blank dilution model. Thermodynamic parameters for these fits are in Table 4–7 in supplement.

that a peptide fragment of Annexin 1 bound to hA6 but not hA5, and a peptide fragment of Annexin 2 bound to neither [275] (Fig 14B). To better quantify the relative specificity of these proteins, we used ITC to measure the binding of two additional peptides to recombinant hA5 and hA6. The first was a peptide from Siah–interacting protein (SIP) previously reported to bind to hA6 [280]. We found that this peptide bound to hA6 with a $K_D$ of 20 $\mu M$, but did not bind hA5 (Fig 14B, C). The second was a 12 amino acid fragment of the protein NCX1 that was reported to bind to hA5 [281]. We found that this peptide bound with to hA5 with a $K_D$ of 20 $\mu M$, but did not bind hA6 (Fig 14B, C).

To further characterize the specificity of the interface, we used phage display to identify two additional peptides that bound to each protein. We panned a commercial library of random 12–mer peptides fused to M13 phage with either hA5 or hA6. Phage enrichment was strictly dependent on $Ca^{2+}$ (Fig 32 in supplement). Three sequential rounds of binding and amplification with either hA5 or hA6 led to enrichment of the "A5cons" and "A6cons" peptides (Fig 15B, Fig 32 in supplement). We then used ITC to measure binding of these peptides to hA5 and hA6. To ensure solubility, we added polar N and C–terminal flanks before characterizing binding. A5cons bound to both hA5 and hA6 (Fig 14C). In contrast, A6cons, bound hA6 but not hA5 (Fig 14C). To verify that binding was driven by the central region, we re–measured binding in the presence and absence of different versions of the flanks (Table 3 in supplement).

The peptides that bind to hA5 and hA6 are diverse in sequence, hydrophobicity, and charge (Fig 14B). One explanation for this diversity could be that the peptides bind at different interfaces on the protein. To test for this possibility, we used NMR to identify residues whose chemical environment changed

on binding of peptide. We first verified the published assignments for hA5 using a 3D NOESY–TROSY experiment [207]. We then collected $^{1}H$ $-^{15}$ $N$ TROSY–HSQC NMR spectra of $Ca^{2+}$–bound protein in the presence of either the A5cons or A6cons peptide. By comparing the bound and unbound spectra, we could identify peaks whose location shifted dramatically or that broadened due to exchange. In addition to our own work, we also included previously reported experiments probing the hA5/NCX1 peptide interaction in the analysis [281]. For all three peptides, we observed a consistent pattern of perturbations in helices 3 and 4 and, to a lesser extent, helix 1 upon peptide binding (Fig 15A–C). These results suggest that all three peptides bind at the canonical interface. In addition to this spectroscopic evidence, binding of all of these peptides was strictly dependent on the presence of $Ca^{2+}$ (Fig 15D–F)—consistent with binding at the interface exposed on $Ca^{2+}$ binding [207].

*The S100A5 and S100A6 clades exhibit conserved binding specificity*

Although hA5 and hA6 bind to diverse peptide targets at the same interface, they exhibit distinct specificity relative to one another (Fig 14B). The particular peptides that bind or not could be random if specificity fluctuates over evolutionary time. In contrast, if specificity at the interface is strongly constrained, we would expect conserved specificity between paralogs. We therefore set out to study the evolution of the differences in peptide binding between the human proteins.

We first constructed a maximum–likelihood phylogeny of the clade containing S100A2, S100A3, S100A4, S100A5, and S100A6 (Fig 16A). We built the tree using the EX/EHO+$\Gamma_8$ evolutionary model [287], which uses different evolutionary models for sites in different structural classes. As expected from previous

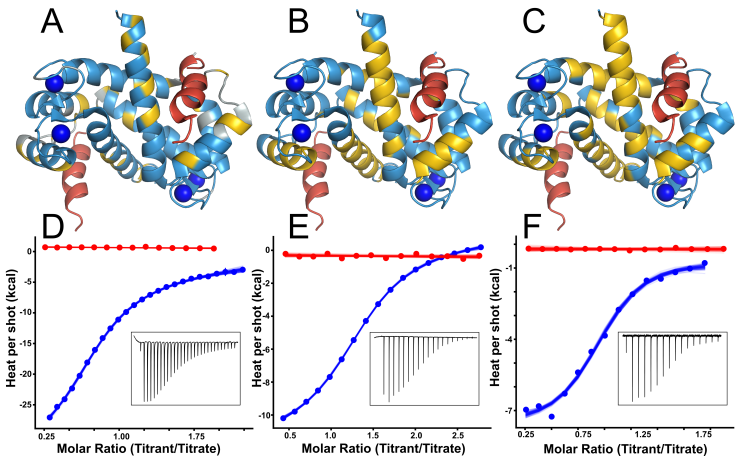


FIGURE 15 Diverse peptides bind at the human S100A5 peptide interface.Structures show NMR data mapped onto the structure of $Ca^{2+}$–bound hA5 (2KAY [207]). To indicate the expected peptide binding location, we aligned a structure of hA6 in complex with the SIP peptide (2JTT [280]) to the hA5 structure, and then displayed the SIP peptide in red. Panels A–C show binding for NCX1, A5cons, and A6cons respectively. In panel A, yellow residues are those noted as responsive to NCX1 binding in [281]. In panels B and C, yellow residues are those whose $^{1}H$–$^{15}N$ TROSY–HSQC peaks could not be identified in the peptide–bound spectrum because the peaks either shifted or broadened. Panels D–E show ITC data for binding of the peptides above in the presence of 2 mM $Ca^{2+}$ (blue) or 2 mM EDTA (red). Points are integrated heats extracted from each shot. Lines are 100 different fit solutions drawn from the fit posterior probability distributions. For the $Ca^{2+}$ curves, we used a single–site binding model. For the EDTA curves, we used a blank dilution model. Insets show raw ITC power traces for the $Ca^{2+}$ binding curves. Thermodynamic parameters for these fits are in Table 4–7 in supplement.

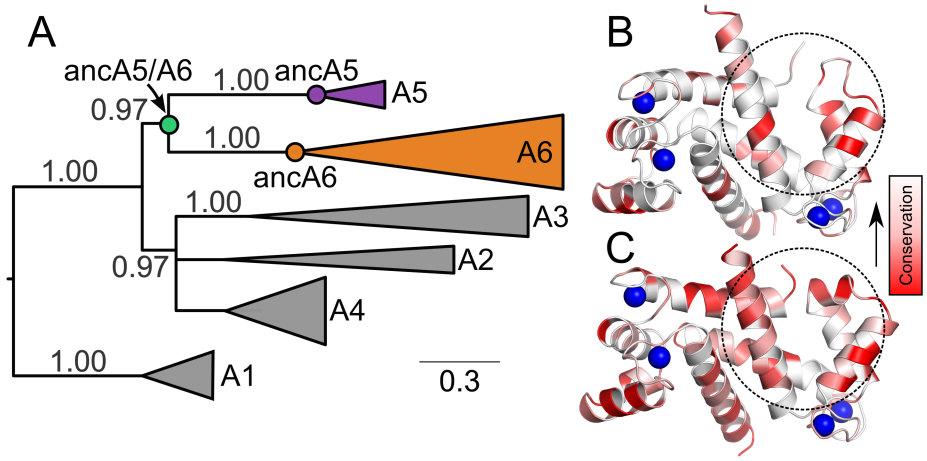


FIGURE 16 S100A5 and S100A6 arose by gene duplication in the amniote ancestor. A) Maximum likelihood phylogeny for S100A5, S100A6 and their close homologs. Wedges denote collections of paralogs (S100A1, S100A2, S100A3, S100A4, S100A5, or S100A6). Wedge height corresponds to the number of sequences and wedge length to the longest branch in that clade. SH supports, estimated using an approximate likelihood ratio test [236], are shown above the branches. Scale bar shows branch length in substitutions per site. Reconstructed ancestors are denoted with circles. All proteins, with the exception of those in the A1 clade, are taken from amniotes. A1 contains S100 proteins from bony vertebrates and was used as an out–group to root the tree. Panels B and C show relative conservation of residues across amniote paralogs mapped onto the structures of hA5 (2KAY, [207]) and hA6 (1K96, [286]). Colors denote conservation from $<$ 20 % (dark red) to 100 % white. Sequences were taken from the alignment used to generate the phylogeny in panel A. Dashed circles denote the peptide binding surface for one of the two chains. Blue spheres show the location of bound $Ca^{2+}$ in the structures.

phylogenetic and syntenic analyses [91, 96], S100A5 and S100A6 were paralogs that arose by gene duplication in the amniote ancestor, with S100A2, S100A3, and S100A4 forming a closely–related out group (Fig 16A). To set our expectation for conservation of specificity, we then calculated the conservation of residues at the binding site across S100A5 and S100A6 homologs. Fig 16B and C show the relative conservation of residues on hA5 (Fig 16B) and hA6 (Fig 16C). Taken as a whole, the peptide binding region does not exhibit higher conservation than other regions in the protein. We therefore predicted substantial variability in the peptide binding specificity across S100A5 and S100A6 orthologs.

To test the prediction that specificity has fluctuated over time, we expressed and purified S100A5 and S100A6 orthologs from human, mouse (*Mus musculus*), tasmanian devil (*Sarcophilus harrisii*), American alligator (*Alligator mississippiensis*), and chicken (*Gallus gallus*). We then characterized the peptide binding specificity of these S100A5 and S100A6 orthologs against four peptides: A5cons, A6cons, SIP, and NCX1 (Fig 17A). We selected these peptides because there is direct evidence that these peptides bind at the canonical binding interface (Fig 15, as well as [280, 281]). Surprisingly, we found that the S100A5 and S100A6 clades exhibited broadly similar, ortholog–specific binding specificity (Fig 17A). All S100A5 orthologs bound NCX1, A5cons, and A6cons, but not SIP. In contrast, all S100A6 orthologs bound SIP and A6cons, but not A5cons. The only labile character is NCX1 binding to S100A6. The sauropsid and marsupial S100A6 orthologs bound NCX1, but not the eutherian mammal representatives. We also characterized binding of these peptides to human S100A4 as an outgroup. Binding for this protein was intermediate between the S100A5 and S100A6 clades: it bound A5cons and A6cons, but not SIP or NCX1. Thermodynamic parameters for these

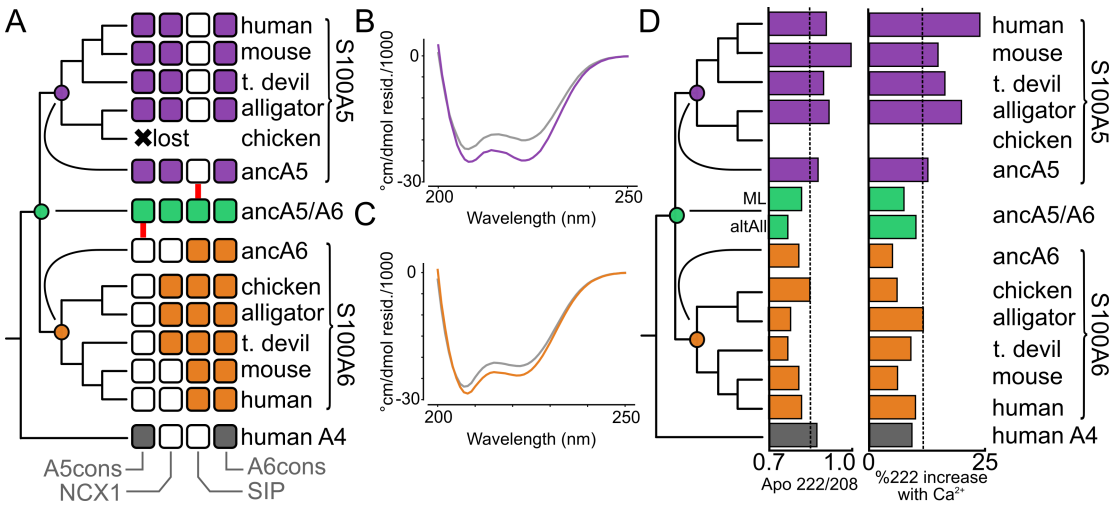


FIGURE 17 S100A5 and S100A6 paralogs exhibit conserved properties.A) Peptide binding specificity mapped onto the phylogenetic tree as a collection of binary characters. Each square denotes binding of a specific peptide to an ortholog sampled from the species indicated at right. Squares are filled if binding was observed by ITC. Ancestors are shown in the middle, with red arrows indicating changes that occurred after duplication that were then conserved across orthologs. The results for ancA5/A6 were identical for both the ML and "altAll" ancestors. Full thermodynamic parameters are in Table 4–7 in supplement. B) Far–UV spectra for apo (gray) and $Ca^{2+}$–bound (purple) hA5. C) Far–UV spectra for apo (gray) and $Ca^{2+}$–bound (orange) hA6. D) Spectroscopic properties mapped onto the phylogeny. The left column shows the ratio of absorbance at 222 nm/208 nm for the apo protein. The right column shows the percentage increase in signal at 222 nm upon addition of $Ca^{2+}$. Dashed lines show the mean values across all experiments. Raw spectra are given in Fig 34 in supplement.

binding experiments are given in Table 4–7 in supplement. Representative ITC traces for each protein are shown in Fig 33 in supplement.

The strong conservation of peptide binding suggested that other features—such as structural features—might be conserved between paralogs as well. To test for this, we characterized the secondary structure and response to $Ca^{2+}$ for all proteins using far–UV circular dichroism (CD) spectroscopy. A $Ca^{2+}$–driven change in $\alpha$–helical secondary structure is a conserved feature of S100 proteins [96, 100]. We asked whether this behavior was conserved across orthologs, which

would indicate similar structural properties. As with peptide binding, we found that the CD spectrum and response to $Ca^{2+}$ were diagnostic within each clade (Fig 17B–D, Fig 34 in supplement). S100A5 orthologs exhibited deep minima at 208 and 222 nm, corresponding to a largely $\alpha$–helical secondary structure (Fig 17B,D). This signal increased upon addition of saturating $Ca^{2+}$, consistent with the ordering of the C–terminus of the human protein reported by NMR [207]. In contrast, all S100A6 orthologs exhibited a deeper minimum at 208 nm, likely corresponding to a mixture of $\alpha$–helical and random coil secondary structure. The secondary structure of these proteins changed comparatively little on addition of $Ca^{2+}$ (Fig 17C,D).

*Specificity evolved from an apparently promiscuous ancestor*

Surprisingly, despite the diversity of peptides that bind to each paralog, peptide binding specificity is conserved across across paralogs. We next asked whether these proteins exhibited comparable evolutionary patterns to those observed in high–specificity proteins, such as the partitioning of ancestral binding partners along duplicate lineages [58, 272, 55]. Using our phylogeny, we used ancestral sequence reconstruction (ASR) to reconstruct the last common ancestors of S100A5 orthologs (ancA5) and S100A6 orthologs (ancA6) [288]. These proteins were well reconstructed, having mean posterior probabilities of 0.93 and 0.96, respectively. Their sequences are given in Table 8 in supplement. We expressed and purified both of these proteins. We found that they shared similar secondary structures and $Ca^{2+}$–binding responses with their descendants by far–UV CD (Fig 17C). We then measured binding to the suite of four peptides described above using ITC. These ancestors gave the pattern we would expect given the binding specificities of the derived proteins (Fig 17D). AncA5 is indistinguishable from a

modern S100A5 ortholog, binding A5cons, A6cons, and NCX1, but not SIP (Fig 4D). AncA6 also behaves as expected, binding A6cons and SIP, but not A5cons. It does not bind NCX1, consistent with this character being labile in the S100A6 lineage (Fig 17D).

We next characterized the last common ancestor of S100A5 and S100A6 (ancA5/A6). This reconstruction had a mean posterior probability of 0.83 (Table S6). AncA5/A6 has a secondary structure content identical to ancA6 and the S100A6 descendants. It also responds to $Ca^{2+}$ in a similar fashion (Fig 17C, Fig 33 in supplement). Unlike any modern protein, however, ancA5/A6 binds to all four peptides (Fig 18). To verify that this result was not an artifact of the reconstruction, we also made an "AltAll" ancestor of ancA5/A6 in which we swapped all ambiguous sites in the maximum–likelihood ancestor with their next most likely alternative [50] (Table 8 in supplement, methods). This protein is quite different than ancA5/A6—differing at 21 of 93 sites—but the binding profile for the four peptides was identical to the maximum–likelihood ancestor. Thermodynamic parameters for these binding experiments are given in Table 4–7 in supplement.

*Binding specificity can be changed with a single mutation*

Our work revealed that S100A5 and S100A6, despite having low overall specificity, display the same basic evolutionary patterns as high–specificity proteins [58, 55, 273]: they exhibit conserved partners across modern orthologs and display a pattern of subfunctionalization from a less specific ancestor. While suggestive, this does not establish that there is selection to maintain specificity. Another possibility is that switching specificity is intrinsically difficult, and that the pattern

we observe reflects this difficulty rather than selective pressure to maintain a particular specificity profile.

To distinguish these possibilities, we attempted to shift the binding specificity of hA5 by introducing mutations at the binding interface. We selected five historical substitutions that occurred along the branch between ancA5/A6 and ancA5: e2A, i44L, k54D, a78M, m83A (with the ancestral amino acid in lowercase and modern amino acid in uppercase). We chose these substitutions using three criteria: 1) the ancestral amino acid was conserved in S100A6 orthologs, 2) the derived amino acid was conserved in S100A5 orthologs, 3) and the mutations were located at the peptide binding interface. Fig 18A shows the positions of candidate substitutions mapped onto the structure of hA5 [207].

We reversed each of these sites individually to the ancestral state in hA5. We then measured binding of two clade–specific peptides, SIP and A5cons, to each mutant using ITC (Table 9 in supplement). We found that reverting a single substitution (A83m) to its ancestral state in hA5 enabled it to bind the SIP peptide (Fig 18B). This reversion does not compromise binding to A5cons, thus recapitulating the ancestral specificity (Table S3). Reversion to the ancestral methionine at residue 83 likely makes more favorable hydrophobic packing interactions with the SIP peptide than the extant alanine. This demonstrates that a single mutation at the peptide binding interface is capable of shifting specificity in S100A5. None of the remaining four ancestral reversions led to measurable changes in A5cons or SIP binding. Amino acids at these positions either do not interact with these peptides, or the ancestral and derived amino acids interact in roughly equivalent fashion.

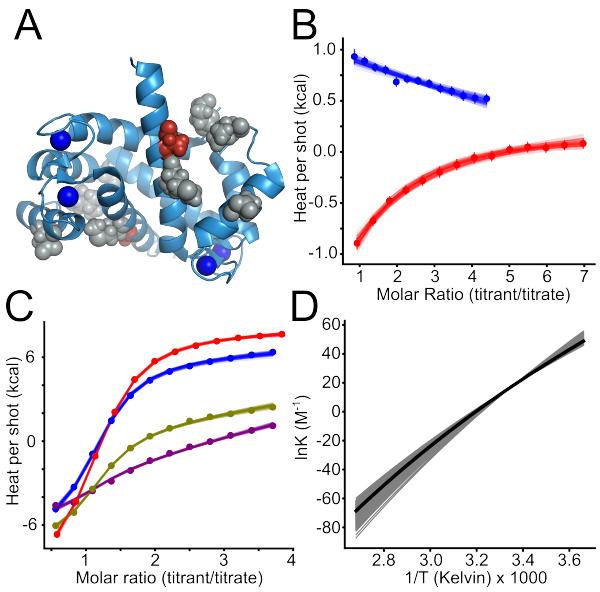


FIGURE 18 Small changes are sufficient to alter binding specificity at the interface. A) $Ca^{2+}$–bound structure of human S100A5 (2KAY) [207] with ancestral reversions marked in gray (no effect on SIP binding) and red (A83m—allows SIP binding). Blue spheres are $Ca^{2+}$ ions. B) ITC traces showing titration of SIP onto hA5 A83m (red) versus wildtype hA5 (blue). ITC experiments were performed at 25°$C$ in 25 mM TES, 100 mM NaCl, 2 mM $CaCl_2$, 1mM TCEP, pH 7.4. Points are integrated heats extracted from each shot. For each experiment, we sampled fit parameters using Bayesian Markov Chain Monte Carlo as implemented in pytc. For the A83m curve, we used a single–site binding model. For the wt curve, we used a blank dilution model, where the linear slope is indicative of peptide dilution without binding. Lines are 100 different solutions drawn from the Bayesian posterior probability distributions. C) ITC traces from experiments done at multiple temperatures: 10°$C$ (purple), 15°$C$ (green), 20°$C$ (blue), and 25°$C$ (red). Experiments were performed in 25mM TES, 100mM NaCl, 2mM $CaCl_2$, 1mM TCEP, pH 7.4. Dots are integrated heats with uncertainty calculate using NITPIC [264]. There is a clear temperature dependence of the binding enthalpy. A global Van't Hoff model was fit to the data using the Bayesian MCMC fitter in pytc. We were unable to fit the model without inclusion of a $\Delta C_p^\circ$ parameter ($-0.40 \leq -0.36 \leq -0.32 kcal \cdot mol^{-1} \cdot K^{-1}$), suggesting that there is a change in heat capacity as a function of temperature, which is indicative of a hydrophobically–driven interaction. Lines are 100 curves drawn from the posterior distribution of the fits. Table 9 in supplement. D) Van't Hoff plot of temperature dependence data. Thick black line shows Maximum Likelihood curve, gray lines are 500 curves drawn from the posterior distribution of the Bayesian fit. There is slight, but detectable curvature in the plot, consistent with the small $\Delta C_p^\circ$ parameter obtained from global model.

Another way to view specificity is in terms of binding mechanism. If binding affinity is mostly due to the hydrophobic effect, we would predict it would be relatively easy to alter binding by small changes to packing interactions. To test for relative contributions of the hydrophobic effect versus polar contacts to binding affinity, we did a van't Hoff analysis for the binding of A5cons to hA5. We performed ITC at temperatures ranging from 10 $^{\circ}C$ to 25 $^{\circ}C$ and then globally fit van't Hoff models to the binding isotherms (Fig 5C–D). We first attempted fits using a fixed enthalpy of binding ($\Delta C_p^{\circ} = 0.0$), but the fits did not converge. When we allowed $\Delta C_p^{\circ}$ to float, we found it was negative ($-0.40 \leq -0.36 \leq -0.32\ kcal \cdot mol^{-1} \cdot K^{-1}$), indicating that binding is driven by the hydrophobic effect [289]. This observation is consistent with binding at the hydrophobic surface exposed by the $Ca^{2+}$induced conformational change [207] and may help to explain why specificity can be readily altered via a single substitution in the interface.

## Discussion

Our work highlights the paradoxical nature of peptide binding specificity for these low–specificity S100 proteins. The binding interface has low specificity, interacting with very diverse peptides with no obvious binding motif (Fig 14B). Further, the specificity is fragile, and can be altered with a single point mutation (Fig 18). One might therefore conclude that this binding specificity is only weakly constrained. In contrast, binding specificity has been conserved over 320 million years along both lineages, exhibiting a pattern of subfunctionalization similar to what has been observed previously for the evolution of high–specificity proteins (Fig 17). This strongly points to the binding specificity being important, despite being very broad.

*Low specificity through a hydrophobic interface*

The binding specificity of these proteins is likely driven almost entirely by shape complementarity and packing. The protein interface exposed on $Ca^{2+}$ binding is hydrophobic and likely makes few protein–peptide polar contacts. This prediction is validated, at least for the hA5/A5cons interaction, by the negative $\Delta C_p^\circ$ on binding, pointing to an important contribution from the hydrophobic effect on binding (Fig 18C). The lack of polar contacts is the likely explanation for the low specificity of the interface. Peptides need only match hydrophobicity and packing, meaning that a large number of possible peptides bind with similar affinity.

The hydrophobic nature of the interface explains the low specificity, but makes the conservation of specificity over 320 million years quite surprising. There is likely no diagnostic set of polar contacts that can be conserved maintain specificity. It should therefore be straightforward to change specificity with minimal perturbation. Indeed, we found that a single mutation, from a small to a large hydrophobic amino acid, is able to switch the specificity of the interface (Fig 18A). Yet, over evolutionary time, binding specificity—at least for this set of targets—has been maintained (Fig 17). Amazingly, this is achieved without strict conservation of the binding site. The peptide binding region does not exhibit higher conservation than other residues in either S100A5 or S100A6 (Fig 16B–C).

Our work shows that protein binding specificity is likely an important feature of these proteins, but does not reveal the set of biological targets for S100A5 and S100A6. Identifying these targets will require further experiments. This could include coupling S100A5 and S100A6 knockouts to proteomics or transcriptomics, pull downs followed by proteomics, and/or large–scale screens of peptide targets

via a technique like phage display. We also anticipate that external factors—such as coexpression, large complex assembly, and subcellular localization—will add critical additional layers of specificity to the low–specificity binding interfaces of these proteins. Understanding the interplay between the biochemical specificity and these external factors will be important for dissecting the biology of these proteins.

*S100s may allow the evolution of new calcium regulation*

The existence of a conserved set of binding partners also has intriguing implications for the evolution of $Ca^{2+}$ signaling pathways in vertebrates. This can be seen by contrasting S100 proteins with calmodulin, a protein that also exposes a protein interaction surface and regulates the activity of target proteins in response to $Ca^{2+}$ [84]. It has been proposed that calmodulin provides a universal $Ca^{2+}$ response across tissues, while S100 proteins allow for fine–tuned, tissue–specific responses [100, 89]. Our results allow us to extend this idea along an evolutionary axis.

Our results suggest that S100 proteins may provide a minimally pleiotropic pathway for the evolution of new $Ca^{2+}$ regulation. Calmodulin is broadly expressed across tissues. As a result, a mutation that causes a protein to interact with calmodulin will have the same effect in all tissues where that protein is expressed. This could lead to unfavorable pleiotropic effects that prevent fixation of the mutation. In contrast, S100 proteins have highly differentiated tissue expression. S100A5, for example, is expressed almost exclusively in olfactory tissues. This means that a protein that acquires an interaction with S100A5 will do so only in olfactory tissue, with minimal pleiotropic effects in other tissues. The pattern of subfunctionalization we observed is consistent with this idea (Fig 17D), as

subfunctionalization is one way to escape adaptive conflict that arises due to pleiotropic effects of mutations [66, 290]. This is only possible because S100A5 evolved a distinct binding profile relative to S100A6 (and presumably other S100 proteins), meaning that acquisition of a new S100A5 interaction does not imply an interaction with a large number of other S100 proteins, which would itself lead to extensive pleiotropy.

Additionally, our results suggest that S100 proteins would provide a much simpler path for the evolution of new $Ca^{2+}$ regulation than calmodulin. The calmodulin sequence has been conserved for over a billion years and is basically unchanged across fungi and animals. As a result, evolution of a new calmodulin–regulated target requires that the target change its sequence to bind to calmodulin. This would likely mean that slowly evolving proteins would not be able to evolve $Ca^{2+}$ regulation, as neither the calmodulin nor possible new target would be able to acquire the necessary mutations to form the new interaction. In contrast, S100 proteins are evolving rapidly. For example, human S100A5 and S100A6 only exhibit 53% sequence identity, despite sharing an ancestor $\approx$ 320 million years ago. This means that, particularly after gene duplication, S100 proteins can acquire new interactions through mutations to the S100 itself. This would allow them to capture slowly evolving target proteins, opening a different avenue for the evolution of $Ca^{2+}$ regulation that would not be accessible by calmodulin alone.

### *Evolution of low–specificity proteins*

Our results also shed light on the evolution of low specificity proteins in general. Many proteins besides S100 proteins exhibit low specificity including other signaling proteins [84, 83], hub proteins [81, 269, 145, 82], and many others

[267, 79, 268, 85, 87]. Further experiments will be required to determine the generality of our observations for low–specificity proteins, but our work suggests that low–specificity proteins can evolve with similar dynamics to the high–specificity proteins that have been studied in detail. Partners for low–specificity proteins can be strongly conserved and evolve by subfunctionalization, just like a high–specificity protein.

One important question is whether S100A5 and S100A6 did, indeed, gain specificity over time. The current study, like many others [291, 292, 293, 271, 294, 58, 119], revealed an ancestral protein that appears less specific than its descendants. Some have proposed this is a general evolutionary trend [291, 271, 119]. Caution is warranted before interpreting these data as evidence for this hypothesis. We selected a small set of peptides to study; therefore, other patterns may be consistent with our observations. For example, it could be that the proteins both acquired more peptides that we did not sample in this experiment (actual neofunctionalization), while becoming more specific for the chosen set of targets (apparent subfunctionalization). Particularly given the large number of targets for these proteins, distinguishing these possibilities will require an unbiased, high–throughout approach to measuring specificity. Advances in high–throughput protein characterization have made such experiments tractable [295, 296, 149, 297, 298]. With the right method, we will be able to resolve whether the shifts in specificity we observed indeed reflect increased specificity over evolutionary time, or instead the small size of the binding set we investigated.

Whatever the precise evolutionary process, our results reveal that S100 proteins—despite binding diverse peptides at a low–specificity hydrophobic

interface—have maintained the same binding profile for the last 320 million years. Low–specificity does not imply no specificity, nor a lack of evolutionary constraint.

## Materials and Methods

### *Molecular cloning, expression and purification of proteins*

Synthetic genes encoding the S100 proteins and codon–optimized for expression in *E. coli* were ordered from Genscript. The accession numbers for the modern sequences are: *Homo sapiens* S100A5: P33763, S100A6: P06703; *Mus musculus* S100A5: P63084, S100A6: P14069; *Sarcophilus harrisii* S100A5: G3W581, S100A6: G3W4S8; *Alligator mississippiensis* S100A5: XP_006264408.1, S100A6: XP_006264409.1; *Gallus gallus* S100A6: Q98953. All accession numbers are for the uniprot database [299], with the exception of the *Alligator mississippiensis* accessions, which are for the NCBI database [300].

Genes were sub–cloned into a pET28/30 vector containing an N–terminal His tag with a TEV protease cleavage site (Millipore). Expression was carried out in Rosetta (DE3) pLysS *E. coli* cells. 1.5 L cultures were inoculated at a 1:100 ratio with saturated overnight culture. E.coli were grown to high log–phase ($OD_{600}\approx 0.8 - 1.0$) with 250rpm shaking at $37°C$. Cultures were induced by addition of 1 mM IPTG along with 0.2% glucose overnight at 16C. Cultures were centrifuged and the cell pellets were frozen at $-20°C$ and stored for up to 2 months. Lysis of the cells was carried out via sonication in 25mM Tris, 100mM NaCl, 25mM imidazole, pH 7.4.

Purification of all S100s used in this study was carried out as follows. The initial purification step was performed using a 5 mL HiTrap Ni–affinity column (GE Health Science) on an kta PrimePlus FPLC (GE Health Science). Proteins were

eluted using a 25mL gradient from 25–500mM imidazole in a background buffer of 25mM Tris, 100mM NaCl, pH 7.4. Peak fractions were pooled and incubated overnight at $4°C$ with ≈1:5 TEV protease (produced in the lab). TEV protease removes the N–terminal His–tag from the protein and leaves a small Ser–Asn sequence N–terminal to the wildtype starting methionine. Next hydrophobic interaction chromatography (HIC) was used to purify the S100s from remaining bacterial proteins and the added TEV protease. Proteins were passed over a 5 mL HiTrap phenyl–sepharose column (GE Health Science). Due to the $Ca^{2+}$–dependent exposure of a hydrophobic binding, the S100 proteins proteins adhere to the column only in the presence of $Ca^{2+}$. Proteins were pre–saturated with 2mM $Ca^{2+}$ before loading on the column and eluted with a 30mL gradient from 0mM to 5mM EDTA in 25mM Tris, 100mM NaCl, pH 7.4. Peak fractions were pooled and dialyzed against 4 L of 25 mM Tris, 100 mM NaCl, pH 7.4 buffer overnight at $4°C$ to remove excess EDTA. The proteins were then passed once more over the 5 mL HiTrap Ni–affinity column (GE Health Science) to removed any uncleaved His–tagged protein. The cleaved protein was collected in the flow–through. Finally, protein purity was examined by SDS–PAGE. If any trace contaminants appeared to be present we performed anion chromatography with a 5mL HiTrap DEAE column (GE). Proteins were eluted with a 50mL gradient from 0–500mM NaCl in 25mM Tris, pH 7.08.5 (dependent on protein isolectric point) buffer. Pure proteins were dialyzed overnight against 2L of 25mM TES (or Tris), 100mM NaCl, pH 7.4, containing 2 g Chelex–100 resin (BioRad) to remove divalent metals. After final purification step, the purity of proteins products was assessed by SDS PAGE and MALDI–TOF mass spectrometry to be >95%. Final protein products were flash

frozen, dropwise, in liquid nitrogen to form frozen spherical pellets and stored at $-80°C$. Protein yields were typically on the order of 25mg/1.5L of culture.

*Isothermal titration calorimetry*

ITC experiments were performed in 25 mM TES, 100mM NaCl, 2mM CaCl2, 1mM TCEP, pH 7.4. Although most experiments were performed at $25°C$ , some were done at cooler temperatures depending to ensure measurable binding heats and sufficient curvature for fitting. Samples were equilibrated and degassed by centrifugation at $18,000xg$ at the experimental temperature for 30 minutes. Peptides (GenScript, Inc.) were dissolved directly into the experimental buffer prior to each experiment. All experiments were performed at on a MicroCal ITC–200 or a MicroCal VP–ITC (Malvern). Gain settings were determined on a case–by–case basis to ensured quality data. A 750 rpm syringe stir speed was used for all ITC–200 experiments while 400rpm speed was used for experiments on the VP–ITC. Spacing between injections ranged from 300s–900s depending on gain settings and relaxation time of the binding process. These setting were optimized for each binding interaction that was measured. Titration data were fit to a single–site binding model using the Bayesian fitter in pytc. For each protein/peptide combination, one clean ITC trace was used to fit the binding model. Negative results were double–checked to ensure accuracy. Some were done at lower temperatures ($10°C$ or $15°C$) to confirm lack of binding, because peptide binding enthalpy should be dependent on temperature.

*2D HSQC NMR experiments*

We collected 2D $^{1}H$ $-^{15}$ $N$ TROSY-HSQC NMR spectra for 2 $mM$ hA5 in the presence of $Ca^{2+}$ alone and with the addition of the 2 $mM$ A5cons. We also collected the spectra of 0.5 $mM$ hA5 with the addition of 0.5 $mM$ A6cons peptide, which was done at lower concentration due to poorer solubility of A6cons in the aqueous buffer. We transfered published assignments to the $Ca^{2+}$-alone spectrum (BMRB: 16033, [207]), and then used 3D NOESY-TROSY spectra to verify the assignments. We were able to unambiguously assign 76 peaks of the 91 non-proline amino acids in the $Ca^{2+}$-bound form. We then added saturating A5cons or A6cons peptide to the sample and remeasured the TROSY-HSCQ spectrum. We then noted which peaks had either shifted or entered intermediate exchange upon addition of the peptide. Of the 76 unambiguously assigned non-proline amino acids 26 shifted or disappeared in the A5cons-bound form, and 35 shifted or disappeared in the A6cons bound form.

All NMR experiments were performed at 25 $^{\circ}C$ on an 800 MHz (18.8T) Bruker spectrometer at Oregon State University. TROSY spectra were collected with 32 transients, 1024 direct points with a signal width of 12820, and 256 indirect points with a signal width of 2837 Hz in $^{15}N$. NOESY-TROSYs were run with 8 transients, non-uniform sampling with 15% of data points used, and a 150 ms mixing time. All spectra were processed using NMRPipe [301]; data were visualized and assignments transferred using the CCPNMR analysis program [302].

*Far–UV CD spectroscopy*

Far–UV circular dichroism spectra (200250nm) were collected on a J–815 CD spectrometer (Jasco) with a 1 mm quartz cell (Starna Cells, Inc.). We

prepared 20–40 $\mu M$ samples in a Chelex (Bio–Rad) treated, 25mM TES (Sigma), 100mM NaCl (Thermo Scientific) buffer at pH 7.4. Samples were centrifuged at 18,000 x g at 25°$C$ in a temperature–controlled centrifuge (Eppendorf) before experiments. Spectra were measured in the absence and presence of saturating $Ca^{2+}$. Reversibility of $Ca^{2+}$–induced structural changes was confirmed by subsequently adding a molar excess of EDTA to the $Ca^{2+}$–saturated samples and repeating the measurements. Five scans were collected for each condition and averaged to minimize noise. A buffer blank spectrum was subtracted with the built–in subtraction feature in the Jasco spectra analysis software. Raw ellipticity was later converted into mean molar ellipticity based on the concentration and residue length of each protein. These calculations were performed on the buffer–blanked data.

*Preparation of biotinylated proteins for phage display*

A small amount of the purified proteins were biotinylated in the following way using the EZ–link BMCC–biotin system (ThermoFisher Scientific). This kit used a maleimide linker to attach biotin at a Cys residue on the protein. ≈1mg BMCC–biotin was dissolved directly in 100% DMSO to a concentration of 8mM for labeling. Proteins were exchanged into 25mM phosphate, 100mM NaCl, pH 7.4 using a Nap–25 desalting column (GE Health Science) and degassed for 30 minutes at 25°$C$ using a vacuum pump (Malvern Instruments). While stirring at room temperature, 8mM BMCC–biotin was added dropwise to a final 10X molar excess. Reaction tubes were sealed with PARAFILM (Bemis) and the maleimide–thiol reactions were allowed to proceed for 1 hour at room temperature with stirring. The reactions were then transferred to 4°$C$ and incubated with stirring overnight

to allow completion of the reaction. Excess BMCC–biotin was removed from the labeled proteins by exchanging again over a Nap–25 column (GE Health Science), and subsequently a series of 3 concentration–wash steps on a NanoSep 3K spin column (Pall corporation), into the Ca–TeBST loading loading buffer. Complete labeling was confirmed by MALDI–TOF mass spectrometry by observing the ≈540Da shift in the protein peak. Final stocks of labeled proteins were prepared at 10 $\mu M$ by dilution into the loading buffer.

*Phage display panning*

Phage display experiments were performed using the PhD–12 peptide phage display kit (NEB). All steps involving the pipetting of phage–containing samples was done using filter tips to prevent cross–contamination (Rainin). 100L samples containing phage ($2.5x10^{10}$ $PFU$) and biotin–protein 0.01 $\mu M$ (or 0.01 $\mu M$ biotin in the negative control) and 50 $\mu M$ peptide competitor (in competitor samples) were prepared at room temperature in a background of Ca–TeBST loading buffer (25mM TES, 100mM NaCl, 2mM $CaCl_2$, 0.01% Tween–20, pH 7.4) to ensure saturation of the S100s with $Ca^{2+}$. Samples were incubated at room temperature for 1hr. Each sample was then applied to one well of a 96–well high–capacity streptavidin plate (previously blocked using PhD–12 kit blocking buffer and washed 6X with 150 $\mu L$ loading buffer). Samples were incubated on the plate with gentle shaking for 20min. 1 $\mu L$ of 10mM biotin (NEB) was then added to each sample on the plate and incubated for an additional five minutes to compete away purely biotin–dependent interactions. Samples were then pulled from the plate carefully by pipetting and discarded. Each well was washed 5X with 200 $\mu L$ of loading buffer by applying the solution to the well and then immediately pulling off by

pipetting. Finally, 100 $\mu L$ of EDTA–TeBST (25mM TES, 100mM NaCl, 5mM EDTA, 0.01% Tween–20, pH 7.4) elution buffer was applied to each well and the plate was incubated with gentle shaking for 1hr at room temperature to elute. Two replicates of the experiment were performed with each protein.

Eluates were pulled from the plate carefully by pipetting and stored at $4°C$ Eluates were titered to quantify enrichment as follows. Serial dilutions of the eluates from $10^{-1}$–$10^{-6}$ were prepared in LB medium. These were used to inoculate 200 $\mu L$ aliquots of mid–log–phase ER2738 *E. coli* (NEB) by adding 10 $\mu L$ to each. Each 200 $\mu L$ aliquot was then mixed with 3mL of pre–melted top agar, applied to a LB/agar/XGAL/IPTG (Rx Biosciences) plate, and allowed to cool. The plates were incubated overnight at $37°C$ to allow formation of plaques. The next morning, blue plaques were counted and used to calculate PFU/mL phage concentration. Enrichment was calculated as a ratio of experimental samples to the biotin–only negative control.

For subsequent rounds of panning the eluates were amplified as follows. 20mL 1:100 dilutions of an ER2738 overnight culture were prepared. Each 20mL culture was inoculated with one entire sample of remaining phage eluate. The cultures were incubated at $37°C$ with shaking for 4.5 hours to allow phage growth. Bacteria were then removed by centrifugation and the top 80% of the culture was removed carefully with a filtered serological pipette and transferred to a fresh tube containing 1/6 volume of PEG/NaCl (20% w/v PEG–8000, 2.5M NaCl). Samples were incubated overnight at $4°C$ to precipitate phage. Precipitated phage were isolated by centrifugation and subsequently purified by an additional PEG/NaCl precipitation on ice for 1hr. Isolated phage were resuspended in 200 $\mu L$ each sterile loading buffer, titered to measure PFU/mL, and stored at $4°C$ for use in the next

panning round. This process was repeated for 3 total rounds of panning. Plaques were pulled from final reound eluate titer plates and amplified in 1mL ER2738 culture for 4.5 hours. ssDNA was isolated from the phage cultures using the Qiagen M13 spin kit. 10 plaques per replicate experiment were Sanger sequenced (GeneWiz, Inc.). These plaque sequences were used to construct the A5cons and A6cons consensus peptides.

*Phylogenetics and ancestral reconstruction*

We used targeted BLAST searches to build an database of 49 S100A2–S100A6 sequences sampled from across the amniotes, as well as six telost fish S100A1 sequences as an outgroup. We attempted to achieve even taxonomic sampling across amniotes. Database accession numbers are in Table 9 in supplement. We used MSAPROBS for the initial alignment [234], followed by manual refinement. Our final alignment is available as a supplemental stockholm file (File S1 in supplementary directory).

We constructed our phylogenetic tree using the EX/EHO+$\Gamma_8$ model, which incorporates information about secondary structure and solvent accessibility into the phylogenetic inference [287]. We assigned the secondary structure and solvent accessibility of each site using 115 crystallographic and NMR structures of S100A2, S100A3, S100A4, S100A5 and S100A6 paralogs: 1a03, 1a4p, 1b4c, 1bt6, 1cb1, 1cdn, 1cfp, 1clb, 1cnp, 1ig5, 1igv, 1irj, 1jwd, 1k2h, 1k8u, 1k9p, 1ksm, 1kso, 1m31, 1mq1, 1nsh, 1ozo, 1psb, 1psr, 1sym, 1uwo, 1yur, 1yus, 2bca, 2bcb, 2cnp, 2cxj, 2jpt, 2jtt, 2k8m, 2kax, 2ki4, 2ki6, 2kot, 2l0p, 2l50, 2l5x, 2le9, 2lhl, 2llt, 2llu, 2lnk, 2pru, 2rgi, 2wc8, 2wcb, 2wce, 2wcf, 3ko0, 3nsi, 3nsk, 3nsl, 3nso, 3nxa, 1b1g, 1e8a, 1gqm, 1j55, 1k96, 1k9k, 1mho, 1mr8, 1odb, 1qlk, 1xk4, 1xyd, 1yut, 1yuu, 1zfs, 2egd, 2h2k,

2h61, 2k7o, 2kay, 2l51, 2psr, 2q91, 2wnd, 2wor, 2wos, 2y5i, 3c1v, 3cga, 3cr2, 3cr4, 3cr5, 3czt, 3d0y, 3d10, 3gk1, 3gk2, 3gk4, 3hcm, 3icb, 3iqo, 3lk0, 3lk1, 3lle, 3m0w, 3psr, 3rlz, 4duq, 1mwn, 1qls, 2k2f, 2kbm, 3iqq, 3rm1, 3zwh, 4eto. We calculated the secondary structure for each site using DSSP and the solvent accessibility using NACCESS [303, 304]. To remove redundancy—whether from identical sequences solved under slightly different conditions or from the multiple models in the NMR models—we took the majority rule consensus secondary structure and the average solvent accessibility for all structures with identical sequences before doing averages across unique sequences. We then assigned the secondary structure for each column using a majority–rule across unique sequences. We assigned the solvent accessibility as the average across unique sequences at that site. Our structural annotation is available in our alignment stockholm file (File S1 in supplementary directory).

We then constructed our tree using the EX/EHO+$\Gamma_8$ model [287], enforcing correct species relationships within groups of orthologs [128]. We compared the final likelihood of this tree to trees generated using LG+$\Gamma_8$ and JTT+$\Gamma_8$ models [240, 237]. Although the EX/EHO model has seven more floating parameters than either LG or JTT, the final tree had a log–likelihood 61 units higher than the next–best model. An AIC test strongly supports the more complex model ($p = 3 \times 10^{-30}$) . One important output from an EX/EHO calculation is $\chi$, a term that measures the fraction of sites that use the structural models relative to a linear combination of all of them [287]. For our analysis, $\chi = 0.72$. We rooted the tree using the S100A1 sequences, which included S100s from several bony fishes.

To reconstruct ancestors using the EX/EHO+$\Gamma_8$ model, we used PAML to reconstruct ancestors using each of the six possible EX/EHO matrices [288, 305], as well as their linear combination. We then mixed the resulting ancestral posterior

probabilities using the secondary structure calls and apparent accessibility at each site, as well as $\chi$ (see Equation 3 in [287]). The code implementing this approach is posted on github: https://github.com/harmslab/exexho_phylo_mixer. We assigned gaps using parsimony. We generated the AltAll sequence as described in Eick et al [50]. This incorporates uncertainty in the reconstruction by taking the next–best reconstruction at each all ambiguous sites. We took each site at which the posterior probability of the next–best reconstruction was greater than 0.20 and the introduced that alternate reconstruction at the site of interest. Our AltAll sequence differed from the maximum likelihood sequence at 21 positions (24% of sites). File S2 in supplementary directory has the posterior probabilities of reconstructions at each site in the ancestor, as well as the final sequences characterized.

## Bridge to Chapter VI

In this chapter, the behavior of two low–specificity proteins were characterized following gene duplication from a shared ancestor. Two members of the S100 protein family, S100A5 and S100A6, were used a model system. The biochemical specificity of the two human proteins was characterized by measuring the binding of two known peptide targets and two target peptides identified via phage display. The human proteins displayed obvious patterns of binding specificity. The study was then expanded to characterize conservation of the specificity profiles across clades of S100A5 and S100A6 orthologs. Surprisingly, despite the highly variable nature of S100 binding partners, there was a clear signal of conservation in specificity profiles. Finally, ancestral sequence reconstruction was used to resurrect the last common ancestor of all S100A5 and S100A6 proteins. The binding specificity of this ancestor for the same set of peptides was measured,

revealing an apparent pattern of subfunctionalization along both duplicate lineages. Furthermore, careful biophysical experiments and a mutagenesis study were used to determine that peptide binding specificity is readily altered by a single amino acid substitution and binding is driven primarily by the hydrophobic effect. This chapter revealed that proteins with low biochemical specificity nonetheless undergo similar patterns of evolutionary change to high–specificity proteins following gene duplications. Chapter VI introduces a new method for directly measuring the biochemical specificity of proteins in an unbiased fashion. Random–peptide phage display and high–throughput sequencing are combined with ancestral sequence reconstruction are applied to directly trace the evolution of binding specificity in S100A5 and S100A6. This method improves upon the results presented in chapter V, by allowing an estimate to be made of the entire set of possible binding partners for each protein, including the oldest ancestor. This technique highlights the subtlety of evolutionary changes in specificity following a gene duplication. While the low–throughput methods shown in chapter V indicate that specificity may have subfunctionalized in both the S100A5 and S100A6 lineages, the ubiased high–throughput approach introduced in chapter VI demonstrates that human S100A5 has indeed undergone a constriction of specificity onto a subset ancestral binding partners. Meanwhile, human S100A6 appears to have shifted relative to the ancestor, indicative of neofunctionalization. This key result shows the importance of using unbiased methods to probe the evolution of specificity. A low–throughput method can suggest an incorrect picture of how specificity evolved, simply due to a lack of sufficient statistical sampling.

# CHAPTER VI

# EVOLUTION OF INCREASED BINDING SPECIFICITY IN S100A5

## Author Contributions

Lucas Wheeler and Michael Harms conceived the study and designed the experiments. Lucas Wheeler performed all experiments. Michael Harms and Lucas Wheeler analyzed experimental datasets. Michael Harms secured funding for the work. Lucas Wheeler and Michael Harms wrote the manuscript and generated the figures. Michael Harms and Lucas Wheeler edited the manuscript. All authors have read and approved the manuscript.

## Abstract

Some have hypothesized that ancestral proteins are, on average, less specific than their descendants. If true, this would provide directionality to evolution and suggest that reconstructed ancestral proteins would be practical starting points for engineering. In support of this idea, studies of reconstructed ancestral proteins have revealed ancestors that interact with more targets than their descendants. These experimental results, are, however, also compatible with divergence from a common set of ancestral partners: the set of partners shifts, rather than shrinks, along each lineage. We set out to distinguish these two possibilities for a historical evolutionary transition. Previously, we studied the acquisition of peptide binding specificity in the proteins S100A5 and S100A6. Using a handful of peptides, we found that the reconstructed last common ancestor of these proteins bound to

more peptides than its descendants. In the current study, we revisit this transition, estimating changes in the total set of peptides that bind to each protein using a quantitative phage display experiment coupled to supervised machine learning. We uncover a more nuanced picture of the historical transition. Human S100A5 exhibits increased specificity over time, binding a subset of the peptides recognized by the ancestor. In contrast, human S100A6 actually loses specificity, acquiring new targets and binding to a larger number of peptides than the ancestral protein. The S100A5 result is a direct demonstration that the total set of partners recognized by a protein can shrink over time. In contrast, our findings along the S100A6 lineage caution against interpreting changes in binding for a small number of targets as evidence that the ancestor is less specific than its descendants.

## Introduction

Changes in protein specificity are critical for evolutionary change [292, 271, 290, 306, 77, 307, 55, 273]. One intriguing suggestion is that, on average, proteins become more specific over evolutionary time [138, 76, 47]. If true, this would be a directional "arrow" for protein evolution[112, 308, 119, 47]. Such features are rare—and controversial—but could ultimately provide fundamental insight into the evolutionary process [131, 47]. For example, increasing specificity might indicate that proteins become less evolvable over time, as they have fewer promiscuous interactions that can be exploited to acquire new functions. From a practical standpoint, it has also been suggested that less-specific reconstructed ancestors would be powerful starting points for engineering new protein functions [60].

There are several reasons that proteins may evolve towards higher specificity. First, gene duplication followed by subfunctionalization could lead to a partitioning

of ancestral binding partners between descendants, and thus increase specificity along each lineage [309, 58, 55, 273]. Second, as metabolic and interaction networks become more complex, proteins must use more sophisticated rules to "parse" the environment: if an ancestral protein had to discriminate between fewer targets than modern proteins, it could be less specific and still achieve the same biological activity [58]. Finally, on the deepest evolutionary timescales, it has been pointed out that the proteome of the last universal common ancestor was small. As a result, each protein would have been required to perform multiple tasks and hence have lower specificity [138, 76].

The increasing-specificity hypothesis can be represented as a Venn diagram: the set of targets recognized by the ancestor is larger than the sets of targets recognized by its descendants (Fig 19A). The sets in this diagram consist of all possible interaction targets, not just those encountered biologically. From an evolutionary perspective, promiscuous interactions—targets that a protein does not encounter biologically, but would recognize if present—are critical for the evolution of new function and are thus a component of its specificity. Furthermore, if ancestors are to be used as good starting points for engineering applications they must possess a larger set of allowed binding partners.

Much of the empirical support for the increasing-specificity hypothesis comes from ancestral reconstruction studies. The results from one such study are shown in Fig 19B. We previously studied the evolution of peptide binding specificity in the amniote proteins S100A5 and S100A6. These proteins bind to $\approx$ 12 amino acid linear peptide regions of target proteins to modulate their activity (Fig 19B) [94, 280, 207, 101, 277, 275, 278, 281, 89]. We found that S100A5 and S100A6 orthologs bound to distinct peptides, but that the last

common ancestor bound to all of the peptides we tested (Fig 19B) [110]. Other studies, probing other classes of interaction partners, have found similar results: the ancestor interacts with a broader range of partners than extant descendants [292, 58, 60, 310, 119, 291, 55, 293, 141, 311, 273, 110].

Such results are not, however, sufficient to test the increasing-specificity hypothesis. This can be seen in Fig 19C, which illustrates two radically different Venn diagrams consistent with our experimental observations of peptide binding in Fig 19B. One possibility is increasing specificity (the descendant sets are smaller than the ancestral set). Another possibility is shifting specificity (the descendant sets remain the same size but diverge in their composition). Distinguishing these possibilities requires estimating the populations in each region of the Venn diagram, which can only be done with a much larger, unbiased sample of the set of binding partners (Fig 19C).

To perform a proper test for the evolution of increased specificity, we set out to estimate changes in the total set of peptides between ancA5/A6 and two of its descendants—human S100A5 (hA5) and human S100A6 (hA6). This evolutionary transition is an ideal model to probe this question. We already have a reconstructed ancestral protein that exhibits an apparent gain in specificity over time, at least for a small collection of peptides [110]. Further, because they bind to ≈12 amino acid peptides, the set of binders is discrete and enumerable ($20^{12} = 4 \times 10^{15}$ targets).

We estimated changes in the total sets of partners recognized by these proteins using a combination of high-throughput characterization, machine learning, and *in vitro* biochemistry. We start by measuring the protein-specific enrichment of a huge collection of peptides using phage display. This is a noisy measure of

binding that also suffers from sampling issues, as each experiment samples a different set of peptides. To solve these problems, we use supervised machine learning to train models linking amino acid sequence to peptide enrichment for each protein. We then calibrate these models against measured binding constants for individual peptides. Finally, we apply each calibrated model to a common set of one million peptides, allowing us to estimate the changes in the binding set for the proteins over time. This approach provides a quantitative estimate of changes in specificity over time—revealing that S100A5 and S100A6 did not evolve by a simple process of increasing specificity. This implies the evidence for the global increasing specificity hypothesis should be re-evaluated.

## Results

Our goal was to measure changes in the total binding sets between human S100A5 (hA5), human S100A6 (hA6), and their last common ancestor (ancA5/A6). We therefore performed high-throughput characterization of peptide binding to these three proteins. To account for uncertainty in the reconstructed ancestral sequence, we studied two different versions of the last common ancestor: "ancA5/A6" and "altAll." ancA5/A6 is the maximum likelihood reconstruction of the ancestral sequence; altAll has all ambiguous sites in the reconstruction flipped to their next most-likely state [50, 110]. Both proteins have the same low-resolution peptide specificity (Fig 19B) [110].

### *Estimating peptide interactions by phage display*

We first assayed the binding of tens of thousands of peptides to each protein using phage display. We panned a commercial library of randomized 12-mer

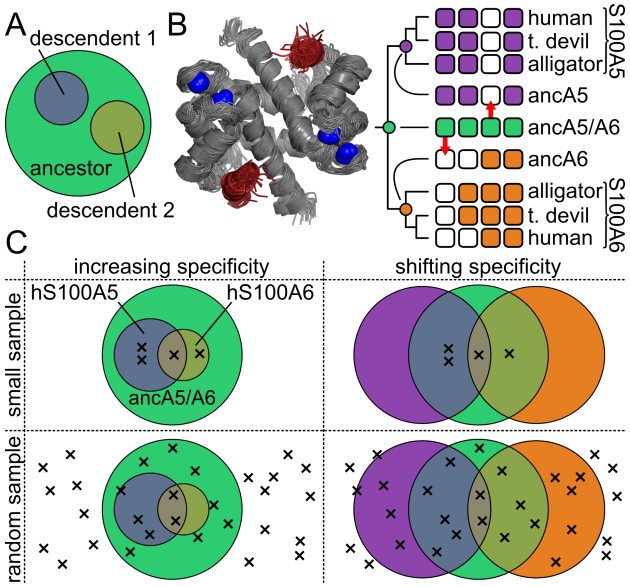


FIGURE 19 Testing the increased specificity hypothesis requires extensive sampling of targets. A) Venn diagram of the increasing-specificity hypothesis. The large circle is set of targets recognized by the ancestor; the smaller circles are sets of targets represented its descendants. There is no strict requirement that descendants be subsets of the ancestor. B) Experimentally measured changes in peptide binding specificity for S100A5 and S100A6 (taken from [110]). Structure: location of peptide (red) binding to a model of S100A5 (gray, PDB: 2KAY). Bound $Ca^{2+}$ are shown as blue spheres. Phylogeny: Boxes represent binding of four different peptides (arranged left to right) to nine different proteins (arranged top to bottom). A white box indicates the peptide does not bind that protein; a colored box indicates the peptide binds. Colors denote ancA5/A6 (green), S100A5 (purple), and S100A6 (orange). Red arrows highlight ancestral peptides lost in the modern proteins. C) Venn diagrams show overlap in peptide binding sets between ancA5/A6, S100A5, and S100A6. Crosses denote experimental observations. Columns show two evolutionary scenarios: increasing specificity (left) versus shifting specificity (right). Rows show to different sampling methods: small sample (top) versus random sampling (bottom). Colors are as in panel B.

peptides expressed as fusions with the M13 phage coat protein. The S100 peptide-binding interface is only exposed upon $Ca^{2+}$-binding (Fig 19B); therefore, we performed phage panning experiments in the presence of $Ca^{2+}$ and then eluted the bound phage using EDTA (Fig 20A). The population of enriched phage will be a mixture of phage that bind at the site of interest and phage that bind adventitiously (blue and purple phage, Fig 20A). Peptides in this latter category enrich in $Ca^{2+}$-dependent manner through avidity or binding at an alternate site [312, 313]. To separate these populations, we repeated the panning experiment in the presence of a saturating concentration of a competitor peptide known to bind at the site of interest (Fig 20B) [110]. This should lower enrichment of peptides that bind at the site of interest, while allowing any adventitious interactions to remain. By comparing the competitor and non-competitor pools, we can distinguish between actual and adventitious binders.

We performed this experiment with and without competitor, in biological duplicate, for hA5, hA6, ancA5/A6, and altAll. We found that phage enriched strongly for all proteins relative to a biotin-only control (Fig 35 in supplement). Further, the addition of competitor binding knocked down enrichment in all samples (Fig 35 in supplement). After panning, we sequenced the resulting phage pools, as well as the input library, by Illumina. We applied strict quality control (see methods), discarding any peptide that exhibited less than six counts (Fig 36 in supplement). After quality control, we had a total of 265 million reads spread over 17 samples (Table 10 in supplement).

We estimated changes in the frequencies of peptides between samples with and without competitor peptide. For each peptide $i$, we determined $E_i = -ln(\beta_i/\alpha_i)$, where $\beta_i$ and $\alpha_i$ are the frequencies of the peptide in the

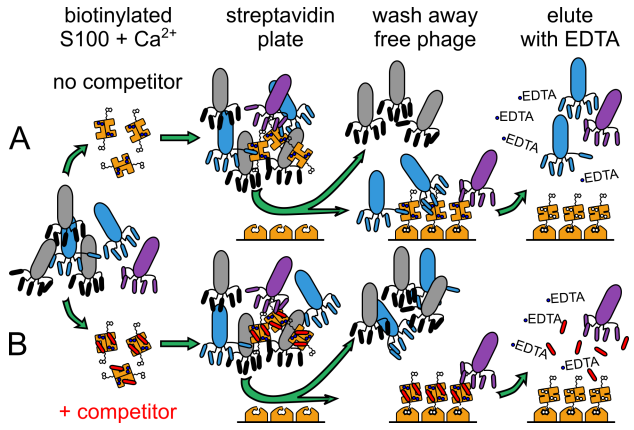


FIGURE 20 Set of binding peptides can be estimated using phage display. Rows show two different experiments, done in parallel, for each protein. Biotinylated, $Ca^{2+}$-loaded, S100 is added to a population of phage either alone (row A) or with saturating competitor peptide added in trans (row B). Phage that bind to the protein (blue or purple) are pulled down using a streptavidin plate. Bound phage are then eluted using EDTA, which disrupts the peptide binding interface. In the absence of competitor (row A), phage bind adventitiously (purple) as well as at the interface of interest (blue). In the presence of competitor (row B), only adventitious binders are present.

non-competitor and competitor samples, respectively. Defined this way, a more negative value of $E$ corresponds to a larger decrease in peptide frequency upon addition of competitor peptide. We used a clustering approach to estimate $E$ for $\approx 40,000$ different peptides for each protein. We found that $E$ exhibited a bimodal distribution for all four proteins, apparently reflecting two underlying processes (Fig 21A, Fig 37 in supplement). The dominant peak consists of "unresponsive" peptides whose frequencies change little in response to competitor peptide. A second, broader, distribution describes "responsive" peptides whose frequencies change dramatically with the addition of competitor. There was no systematic difference between estimates of $E$ between biological replicates (Fig 21B, Fig 38 in supplement). For hA5, the regression line between replicates has a slope 1.06 and an intercept of $-0.05$. This axis of variation explains $\approx 81\%$ of the total variation

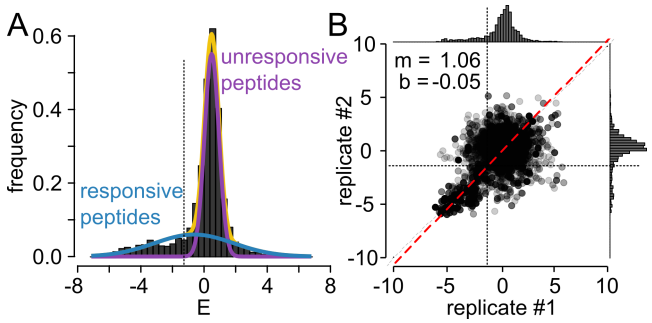


FIGURE 21 A subpopulation of the phage respond to the addition of competitor peptide. A) Distribution of enrichment values for peptides taken from pooled biological replicates of hA5. The measured distribution (gray) can be fit by the sum of two Gaussian distributions: responsive (blue) and unresponsive (purple), which sum to the total (yellow). B) Enrichment values from biological replicates are strongly correlated. Axes are enrichment for replicate #1 or replicate #2. Points are individual peptides. Distributions for each replicate are shown on the top and right, respectively. The red dashed line is the best fit line (orthoganol distance regression), explaining ≈81% of the variation in the data.

in the data. There are two distinct regions in the correlation plot, corresponding to the unresponsive and responsive peptide distributions. The unresponsive distribution forms a large cloud about zero. In contrast, the responsive peptide distribution extends along the 1:1 line in a correlated fashion.

*Supervised machine learning allows prediction of binding*

Our phage display experiment yielded a collection of peptides whose enrichment is disrupted by competitor, however, this information is not sufficient to construct the desired Venn diagram. First, a Venn diagram requires knowing the binding for a common set of peptides to all proteins. Because the total sets of partners are large for all proteins, we observed different peptides in each experiment (hA5 vs. hA6 vs. ancA5/A6 vs. altAll). Second, phage display is an imperfect proxy for binding. It has confounding non-biological factors: peptides are in the context of phage particles, the protein becomes immobilized by a biotin tag, there

is the possibility of avidity, and enrichment is determined by off-rate rather than equilibrium.

To solve these issues, we sought to relate our measured $E$ for these peptides back to binding of a common set of peptides. We used supervised machine learning to train models to predict binding from amino acid sequence for each protein. We then applied each model to an identical set of peptides, allowing us to directly compute a Venn diagram for peptide specificity.

We trained our models against 57 chemical features that we could could readily calculate from an amino acid sequence. These included measures of hydrophobicity, hydrogen bonding, geometry, secondary structure propensity, and electrostatics. In addition to these specific features, we also defined 20 "meta" features by taking the principle components of the entire aaindex database [314], which reports 590 quantitative values for each of the 20 amino acids. For most chemical features, we simply added the values for each amino acid in a sequence. For example, we would sum up the number of hydrogen bond donors across the sequence and treat that as a chemical feature. We also used CIDER to calculate a few non-additive electrostatic features for each sequence [315], such as the isoelectric point. A full list of the features we calculated is given in Table 11 (in supplement).

We calculated these 57 features for the entire sequence and for all sliding windows ranging from 1 to 11 amino acids (Fig 22A). This introduces neighbor-neighbor correlation between features that improves model power. Overall, we calculated the features for 78 sliding windows on each peptide, giving us a total of $57 \times 78 = 4,446$ features per sequence (Fig 22A). We then trained a random forest regression model to predict $E$ using the features of the $\approx 40,000$ we observed

for each protein. A random forest model finds weights for a collection of random decision trees based on a set of input features [316]. Prior to training, we withheld 10% of the peptides as a test set. We then optimized nuisance parameters such as the number of trees and choice of data weighting scheme using k-fold cross validation within the training set ($k = 10$). After training, the $R^2$ between our model and the training set was $\approx 97\%$ for all proteins (Table 2).

TABLE 2 Protein binding model statistics.

| protein | num. training observations | $R^2_{train}$ | $R^2_{test}$ | AUC | FPR | FNR |
|---|---|---|---|---|---|---|
| hA5 | 40,887 | 97.6 | 85.1 | 98.9 | 0.35 | 0.35 |
| hA6 | 42,156 | 97.4 | 82.9 | 96.1 | 0.41 | 0.41 |
| ancA5/A6 | 43,938 | 97.7 | 84.2 | 97.4 | 0.35 | 0.35 |
| altAll | 51,903 | 96.6 | 80.0 | 95.1 | 0.45 | 0.15 |

After our final optimization, we tested our models against their test sets. $R^2$ for test sets ranged from $80-87\%$ (Fig 22B, Table 1). For all models, the regression line reveals a slope slightly greater than one (e.g. 1.16 for hA5, Fig 22B). Further, the scatter is nonrandom, with the most negative values of $E$ being overestimated and the most positive values underestimated. This makes intuitive sense, as the best-of-the-best and the worst-of-the-worst enriching sequences likely depend strongly on details not captured by our rather crude amino acid model.

To calculate a Venn diagram, we need to classify peptides as binders or non-binders. We therefore tested how well our models would operate as classifiers. To facilitate this comparison, we normalized $E$ for each protein such that the competitor peptide had an enrichment value of -1. We did this by $E_{norm} = E/|E_{comp}|$, where $E_{comp}$ is the enrichment of the competitor peptide. We then asked whether our models could predict if peptides in the test set had measured $E_{norm} < -1$. We then attempted to classify peptides into the categories $E_{norm} < -1$ vs. $E_{norm} \geq -1$.

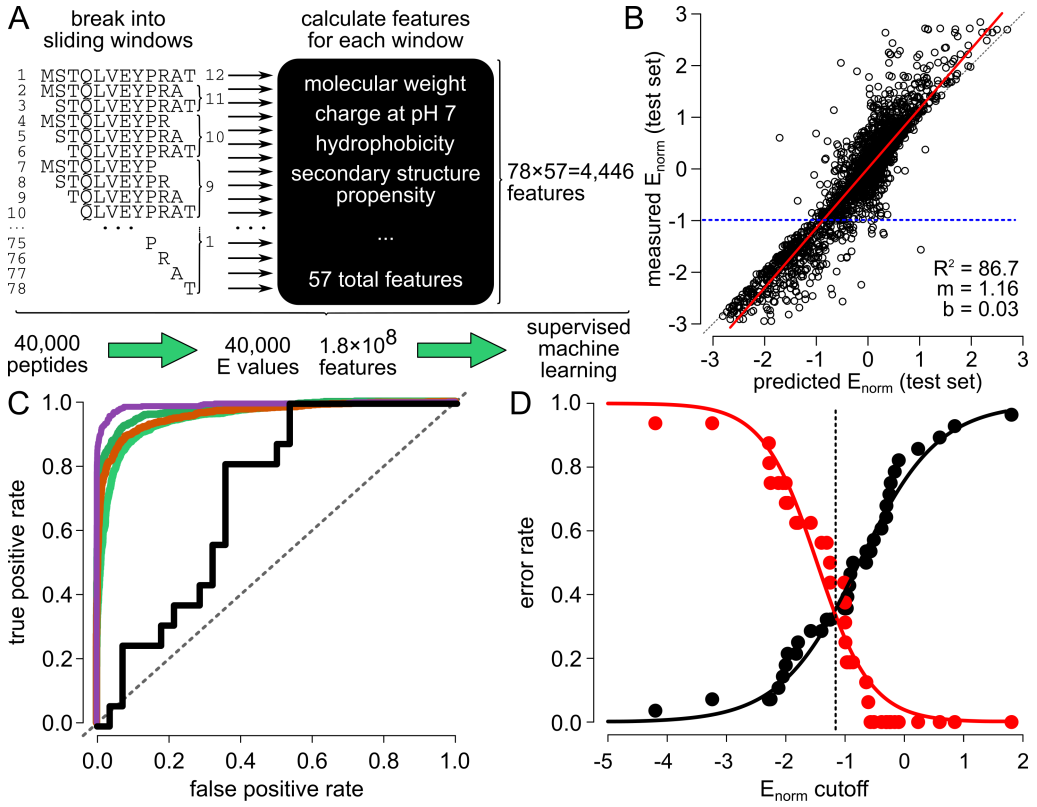


FIGURE 22 Peptide binding can be predicted from amino acid sequence. A) Schematic showing our strategy for training a binding model. We break the 12-mer peptide into 78 different sliding windows. For each peptide, we calculate 57 features (black box), giving a total of 4,446 features per peptide. We then use 40,000 peptides to train a model predicting $E$ (green arrows). B) Correlation between predicted $E_{norm}$ and measured $E_{norm}$ for ≈4,000 peptides in test set for hA5. Each point is a peptide. Red line is least squares regression line. Blue dashed line is our classification line (see panel C). C) Receiver Operator Characteristic (ROC) curves for binding models. Colored series show ability of models to classify measured $E_{norm}$ as $\leq -1$ (the blue dashed line form panel B). Curves are hA5 (purple), hA6 (orange), ancA5/A6 (dark green), and altAll (light green). Black line is the ROC curve for predicting the binding of 44 isolated peptides. D) Error rates for predicting isolated peptides that bind as function of $E_{norm}$ cutoff for the classifier. False negative rate (red) and false positive rate (red) cross at $E_{norm} = -1.19$ (dashed line) with a value of ≈0.35. Solid lines are fits of the modified Hill equation to the to error rates.

We swept along cutoffs in predicted values of $E_{norm}$ and calculated our false positive and false negative rate using the measured values of $E_{norm}$ for test-set peptides. As expected, increasing the cutoff increased the false positive rate and decreased the false negative rate for each model. We quantified this behavior with Receiver Operator Characteristic (ROC) curves. A ROC curve is a plot of the true positive rate against the false positive rate as one changes the classifier cutoff. A perfect predictor will have a cutoff value where the false positive rate is 0 and the true positive rate is 1. As a consequence, the Area Under the Curve (AUC) will be 1.0. In contrast, a random predictor will follow the 1:1 line and will have an AUC of 0.5. All of our models had steep ROC curves that gave AUC values from 0.95 to 0.99 (Fig 22C). Given the amino acid sequence of a 12-mer peptide, we can therefore predict with high confidence whether a peptide will respond to the addition of competitor peptide in a phage display experiment.

We next set out to calibrate our phage enrichment values against binding of isolated peptides. We did this by calculating $E_{norm}$ for 44 peptide/protein pairs and then measuring their binding using Isothermal Titration Calorimetry (Table 12 in supplement). We used 17 peptides, some with known binding properties [280, 275, 281], others that were in the freezer for other projects, and still others were extracted from the human proteome as possible S100 targets. We measured binding of 16 of these peptides to hA5, 13 to hA6, 8 to ancA5/A6, and 6 to altAll. We classified any peptide with a measurable binding constant ($K_D$ 100 $\mu M$) as "binding" and all others as "non-binding."

We then swept along $E_{norm}$ and attempted to classify the 44 measured binders. The ROC curve came off the diagonal, with an AUC of 0.71 (Fig 22C). While this is significantly worse than the predictions of $E_{norm} < -1$, it is not

unexpected given that we trained our model on phage display data and are now attempting to use it to predict isolated peptide binding. To verify that this low AUC curve indicated real binding signal, we simulated 44-observation ROC curves using a random predictor. We found that the probability of observing an AUC of 0.71 or greater by chance was 0.007—a strong indication that there is signal in our binding model.

To identify a cutoff for predicting binders, we plotted the false positive rate and false negative rate against $E_{norm}$ for all 44 peptides. We then identified the value of $E_{norm}$ that simultaneously minimized the false positive and false negative rates. To estimate the crossover point, we fit the modified Hill equation to each curve, which empirically captures the basic shape of these curves. We found that these curves crossed for $E_{norm} = -1.19$, with false positive and false negative rates of ≈0.35. These rates are high and therefore preclude confidently predicting whether a specific peptide binds. This is, however, sufficient to determine a Venn diagram for the binding specificity of these proteins.

*Venn diagrams can be estimated using MCMC*

We next used our trained and calibrated models to estimate the Venn diagram describing the binding sets for the modern and ancestral proteins (Fig 19). We applied our models for hA5, hA6, ancA5/A6, and altAll to a common collection of 1,000,000 random 12-mer peptides, classifying any peptide with $E_{norm} < -1.19$ as binding. We then calculated the overlap between these sets, placing the counts for each region of the Venn diagram into the vector $\vec{V}_{obs}$.

Because we have high (and uncertain) false positive and false negative rates, the counts in $\vec{V}_{obs}$ may not be identical to the real populations of the Venn diagram

($\vec{V}$). We therefore sampled over counts in $\vec{V}$, as well as possible false positive and false negative rates, using Bayesian Markov Chain Monte Carlo (MCMC). We wrote a transition matrix $\mathbf{T}$ that maps $\vec{V}$ into $\vec{V}_{obs}$ ($\vec{V}_{obs} = \vec{V} \cdot \mathbf{T}$). $\mathbf{T}$ defines the probability of each class of miss-call given all false positive and false negative rates. For example, one element in $\mathbf{T}$ encodes the probability that we mistakenly identify a hA5-specific peptide as a hA6-specific peptide (e.g. the false negative rate for hA5 times the false positive rate for hA6). The details of matrix construction are given in the supplemental text.

We allowed each protein to have its own false positive and false negative error rates. We set the prior probabilities for error rates by estimating the false positive and false negative rate for binding to each protein at the cutoff of $E_{norm} < -1.19$ (Fig 39 in supplement). We then used MCMC to sample values of $\vec{V}$ and the error rates, comparing the resulting vector to $\vec{V}_{obs}$. We ran two samplers in parallel until convergence (≈2 million steps each). This allowed us to estimate both the Venn diagram and our uncertainty in its composition.

*hA5 is more specific than hA6 or the ancestor*

We found that the total size of each binding set ranged from 1.3% [0.9,1.8] of peptides (for hA5) to 22.6% [21.8,22.9] of all peptides (for the altAll construct) (Fig 23). The values in the brackets denote the 95% credibility region from the posterior distribution. The large sizes of these sets likely reflects the low-specificity, hydrophobic nature of the S100 binding interface [110].

We found that hA5 exhibits increased specificity relative to the ancestral proteins, apparently evolving by a process of subfunctionalization. The hA5 peptide set is a subset of the ancestral binding set (Fig 23). While ≈85% of peptides are

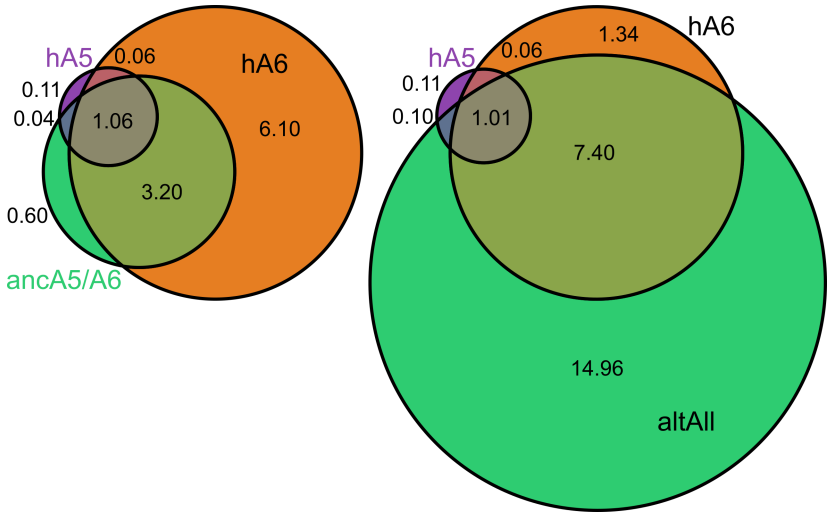


FIGURE 23 Changes in binding sets over time. Circles denote estimated binding sets for hA5 (purple), hA6 (orange), or ancestors (green). Areas and numbers in each region indicate the percent of random peptides in that region of the Venn diagram. The left panel shows the maximum likelihood ancestor (ancA5/A6); the right panel shows the altAll reconstruction.

shared with the ancestor, only ≈9% of peptides arose specifically for binding to hA5. This result is robust to phylogenetic uncertainty, as very similar overlaps are observed for ancA5/A6 (86.5% [83.9,95.2]) and the altAll construct (84.6% [81.1,91.4]). The hA5 peptide set was also largely a subset of the hA6 set: 80.6% [76.8,88.5] of hA5 peptides are also hA6 peptides. This demonstrates that, although the protein has subfunctionalized from the ancestor, it has constricted onto a set of peptides that mostly overlaps with the paralogous hA6 (Fig 23).

The results for hA6 depend on the reconstruction used for the ancestral state. The hA6 binding set was much larger than hA5—consisting of 10.1% [9.2,11.7] of peptides. This is is expanded relative to the ancA5/A6 set. While there is a extensive overlap (37.4% [36.4,38.4]), most hA6 binding targets were acquired after gene duplication (Fig 23). Fully 62.0% [61.1,63.1] of peptides are unique to hA6. In this scenario, hA6 kept its ancestral partners, and then added a large

collection of new partners. However, this pattern varies when inferred using the altAll construct. 82.9% [80.6,85.2] of hA6 peptides are shared with the altAll ancestor (Fig 23). Under this alternate scenario, hA6 binding specificity is the result of both subfunctionalization and the acquisition of new binding partners distinct from the ancestor (neofunctionalization). Subfunctionalization in this scenario appears to have occcurred to a far lesser extent than in the paralogous hA5 lineage, with the hA6 binding set representing only a minor constriction of the ancestral set. However, the maximum likelihood scenario is strongly supported over the alternative one. Furthermore, the overall distribution of enrichment values for the altAll construct was systematically lower than that for any other protein. This result suggests that the alternate construction may actually have compromised—rather than simply distinct—activity.

The alternate reconstruction is a very agressive attempt to incorporate phylogenetic uncertainty. In this case, the altAll protein differs at 21 sites from the maximum-likelihood ancA5/A6 and has a much lower likelihood. The results for hA5 are not changed between these estimates, but the interpretation of hA6 does differ. Thus it is difficult to conclude the exact nature of the transition that occurred along the hA6 lineage. Nonetheless, it is clear that the patterns observed along the hA5 and hA6 lineages are distinctly different. The transitions in specificity that occurred following gene duplication are more nuanced than a simple partitioning of ancestral binding partners across desecendant proteins.

## Discussion

Previously, we used a low-throughput approach to characterize the evolution of the biochemical specificity of S100A5 and S100A6 [110]. We observed a strong

signal of phylogenetic conservation in the pattern of peptide binding specificity. For a small set of peptides, the pattern of specificity is diagnostic for each clade. The small sample of binding targets suggested a pattern in which the ancestral binding partners were partitioned along descendant lineages to yield more specifiic derived proteins. This result was consistent with many previous low-throughput studies showing patterns of subfunctionilization following gene duplication [292, 58, 60, 310, 119, 291, 55, 293, 141, 311, 273, 110].

*Re-evaluating the empirical support for the increasing specificity hypothesis*

In this study we used an unbiased, quantitative phage display experiment to reveal a more subtle pattern. We observed increased specificity on one lineage: the set of peptides recognized by hA5 shrunk dramatically and consists almost entirely of peptides drawn from the ancestral set of peptides. In contrast, hA6 appears to have actually shifted and most likely expanded its set of peptides, although we were unable to resolve these two scenarios given the difference between the maximum likelihood and alternate ancestral reconstructions. The single pattern, when viewed with higher resolution, resolves into two distinct patterns. Our observations suggest that the empirical support for the increasing specificity hypothesis is rather weak. Patterns of specificity may evolve following gene duplication in much more complex ways than previously thought. Our results reveal how problematic inferring specificity from a small set of targets can be. Despite apparent subfunctionalization in low-throughput experiments, hA5 and hA6 exhibit opposite patterns of specificity when probed using a high-throughput approach. hA5 gained specificity, binding to a small subset of the ancestral sequences. hA6, in contrast, lost specificity, acquiring entirely new binding targets.

These findings do not refute the increasing specificity hypothesis, but rather help us to understand the nature of the inference. A small, biased set of targets is insufficient to infer "absolute" changes in specificity.

To date, there is still insufficient evidence to support or refute the hypothesis of increasing specificity over long times scales [47]. Our results suggest that unbiased high-throughput studies should be used to provide the necessary statistical power needed to address this question. The approach presented here will be broadly useful for addressing these questions systematically. Nonetheless, the key limitation of our results for inferring such global patterns in the evolution of specificity is the shallow time-scale of S100 evolution.S100A5 and S100A6 arose via gene duplication in the amniote ancestor ≈320 million years ago [91, 96, 110]. This time-scale is far shorter than that on which we might expect to observe the effects of global trends that permeate all of life. Thus, this study has instead tested what happens after a more recent gene duplication, presumably independent of global trends that have been proposed [138, 76, 47]. Currently, no high-throughput studies of very deep ancestral protein-protein interactions have been conducted. Several previous studies have targeted very old (>3 BYA) ancestral enzymes and observed apparent promiscuity-to-specificity transitions. However, these studies were on enzymes that are already exquisitely specificity compared to sloppy S100 protein-protein interactions. Furthermore, other evolutionary scenarios—such as neofunctionlization and transitions through promiscuous intermediates—are known to occur in some systems [53, 79, 143, 57, 317]. To directly address the hypothesis of increasing specificity, high-throughput studies should be performed that trace specificity in sequentially deeper ancestral proteins stretching backward

in evolutionary time. Such studies would provide a direct test of the hypothesis that proteins have undergone long time-scale promiscuity-to-specificity trends.

Our results display a very strong signal for subfunctionalization along the hA5 lineage. This observation suggests that some evolutionary patterns of specificity may be consistent with hypothetical expectations even in proteins with very low biochemical specificity. However, we have only characterized one gene duplication event, and even here have observed subtlety in the patterns of specificity. Ultimately, more proteins from a diversity of families should be studied using similar methods. Applying unbiased high-throughput screens to diverse proteins with diverse functions will further clarify the generality of our observations. Furthermore, empirical studies will help to build improved theoretical models of proteins with evolving specificity. The role of constraints such as pleiotropy, epistasis, and architectural properties of protein-protein interaction networks can be determined.

*Quantifying the evolution of specificity informs S100 biology and biochemistry*

The large sets for each protein likely reflect the hydrophobic nature of the hA5 and hA6 binding interfaces. Previously we showed that the binding of one A5-specific peptide was driven primarily by the hydrophobic effect [110]. The binding set of hA6 may be larger than that of hA5 due to its extended binding surface relative to other S100 proteins [280]. This larger extende surface may allow it to accommodate a larger number of register-shifted peptides. Rather than only using the canonical interface, peptides can wrap around the protein and bind into an extended groove. This may explain both its broader specificity and the acquisition of targets not observed in the ancestral protein.

Interestingly, the scope of these binding set sizes mirrors the tissue distributions of the two proteins. In mammals, S100A5 has an extremely narrow tissue distribution, being found primarily in the olfactory bulb and olfactory sensory neurons [282, 283, 284]. In contrast, S100A6 is expressed ubiquitously. This is counterintuitive if one starts with the "parsing environment" perspective, as S100A6 has broader specificity even while experiencing more diverse environments. This also suggests that S100A5 has become specialized for a subset of biological targets.

It remains unknown whether the hA5 and hA6 binding sets are shared among modern orthologs, or whether these sets have fluctuated relative to one another. We previously found strong evidence for conservation of specificity—for a small set of peptides—in orthologs across amniote species. This results suggested an overall conservation of biochemical specificity in the S100s. However, as noted above there is insufficient sampling in the low throughput experiments to distinguish differences in the gross specificity of the proteins. Thus, the high-throuhgput approach used in this study would need to be applied to sets of orthologs to quantitatively determine the degree to which patterns of specificity were conserved as the lineages diverged.

*Future directions*

The current analysis was specifically designed to probe targets that may not be realized biologically. Even very weak binders can act as starting points for future evolutionary optimization; therefore, our inclusive approach is the right one for addressing how biochemical specificity evolved in this system. However, this approach does leave an important questions unanswered; how do these results translate across a range of binding affinities? How does the inference of binding

sets change if we restrict our analysis to only the highest affinity binders? This cannot be answered given our data, as the correlation between enrichment in the phage display experiment and binding affinity is too noisy to allow this to be done rigorously. There are several ways the current work can be extended and developed.

An improved method with decreased noise in the esimate of affinities would allow these questions to be quantitatively answered. The scope of binding sets—as a function of stratified binding affinity—could be traced through time within and across lienages. Sampling multiple ancestors along a lineage, would allow us to detect patterns that occur as proteins evolved specificity for new binding partners. Would we observe continuous trends, akin to a gradually shrinking Venn diagram? Or would we instead instead observed random fluctuations, as the Venn diagram dilates between more and less specific nodes? Would these patterns be sensitive to the architectural constraints of the chose protein system?

Finally, similar methods could be utilized to study the evolution of other types of binding interactions. High-throuhgput "bind-and'seq" assays have previously been applied to study DNA and RNA binding specificity of extant proteins [295, 298, 318, 319]. High-throughput mass spectrometry has been used to characterize the enzymatic specificity of venom proteases [320]. One could envision a high-throughput enzyme assay to measure activity against a diverse, unbiased set of small molecule substrates. These approaches could be coupled to ASR—analagous to what we have done here—and used to probe a broad range of protein-target interactions.

*Implications for protein engineering*

Protein egineers seek to design proteins with specific functions; a goal that is often acheived by both rational design and directed evolution [321, 322, 323, 324, 325]. Protein engineers have proposed using ancestral proteins as starting points for engineering, as they may be less specific—and therefore be more generic starting points for an engineering protocol [60]. Thus characterizing global evolutionary trends in specificity—if they exist—would be potentially aid engineering efforts that seek to use ancestral starting points. The ability to build a quantitative picture of how specificity evolves in diverse protein systems would provide a framework for understanding and engineering protein binding properties. By applying unbiased high-throughput approaches such as ours can we understand patterns in biochemical specificity. The ability to control this feature rationally would be a boon to protein engineers.

## Conclusions

Our work provides direct evidence for a transition in which an ancestral set of binding partners was partitioned along a derived lineage. With respect to S100A5, the ancestor would indeed be a better engineering starting point—and presumably evolutionary starting point—given its lower overall speficity. The work also cautions against interpreting low-throughput data as evidence for such a change, as the pattern observed along the S100A6 lineage does not cleanly conform to the promiscuity-to-specificity concept. This protein appears to have expanded or shifted its binding set. Overall, the work presented here has allowed us to quantititavely characterize the subtleties of evolutionary transitions in specificity, reavealing a more nuanced picture than expected from simple hypotheses.

## Materials and Methods

*Molecular cloning, expression and purification in of S100 proteins*

Proteins were expressed in a pET28/30 vector containing an N-terminal His tag with a TEV protease cleavage site (Millipore). For each protein, expression was carried out in Rosetta *E.coli* (DE3) pLysS cells. 1.5 L cultures were inoculated at a 1:100 ratio with saturated overnight culture. *E.coli* were grown to high log-phase ($OD_{600}$ ≈0.8–1.0) with 250 rpm shaking at 37 $°C$. Cultures were induced by addition of 1 mM IPTG along with 0.2% glucose overnight at 16 $°C$. Cultures were centrifuged and the cell pellets were frozen at 20 $°C$ and stored for up to 2 months. Lysis of the cells was carried out via sonication on ice in 25 mM Tris, 100 mM NaCl, 25 mM imidazole, pH 7.4. The initial purification step was performed at 4 $°C$ using a 5 mL HiTrap Ni-affinity column (GE Health Science) on an kta PrimePlus FPLC (GE Health Science). Proteins were eluted using a 25 mL gradient from 25-500 mM imidazole in a background buffer of 25 mM Tris, 100mM NaCl, pH 7.4. Peak fractions were pooled and incubated overnight at 4 $°C$ with ≈1:5 TEV protease (produced in the lab). TEV protease removes the N-terminal His-tag from the protein and leaves a small Ser-Asn sequence N-terminal to the wildtype starting methionine. Next hydrophobic interaction chromatography (HIC) was used to purify the S100s from remaining bacterial proteins and the added TEV protease. Proteins were passed over a 5 mL HiTrap phenyl-sepharose column (GE Health Science). Due to the $Ca^{2+}$-dependent exposure of a hydrophobic binding, the S100 proteins proteins adhere to the column only in the presence of $Ca^{2+}$. Proteins were pre-saturated with 2mM $Ca^{2+}$ before loading on the column and

eluted with a 30mL gradient from 0 mM to 5 mM EDTA in 25 mM Tris, 100 mM NaCl, pH 7.4.

Peak fractions were pooled and dialyzed against 4 L of 25 mM Tris, 100 mM NaCl, pH 7.4 buffer overnight at 4 $°C$ to remove excess EDTA. The proteins were then passed once more over the 5 mL HiTrap Ni-affinity column (GE Health Science) to removed any uncleaved His-tagged protein. The cleaved protein was collected in the flow-through. Finally, protein purity was examined by SDS-PAGE. If any trace contaminants appeared to be present we performed anion chromatography with a 5 mL HiTrap DEAE column (GE). Proteins were eluted with a 50 mL gradient from 0-500 mM NaCl in 25 mM Tris, pH 7.4 buffer. Pure proteins were dialyzed overnight against 2L of 25 mM TES (or Tris), 100 mM NaCl, pH 7.4, containing 2 g Chelex-100 resin (BioRad) to remove divalent metals. After the final purification step, the purity of proteins products was assessed by SDS PAGE and MALDI-TOF mass spectrometry to be >95%. Final protein products were flash frozen, dropwise, in liquid nitrogen and stored at $-80$ $°C$. Protein yields were typically on the order of 25 mg/1.5 L of culture.

### *Isothermal Titration Calorimetry*

For all peptides, we attempted to measure binding at 25 $°C$. ITC experiments were performed in 25 mM TES, 100mM NaCl, 2 mM $CaCl_2$, 1mM TCEP, pH 7.4. Samples were equilibrated and degassed by centrifugation at $18,000xg$ at the experimental temperature for 35 minutes. Peptides were dissolved directly into the experimental buffer prior to each experiment. All experiments were performed at on a MicroCal ITC-200. Gain settings were determined on a case-by-case basis to ensured quality data. A 750 rpm syringe stir speed was used for

all experiments. Spacing between injections ranged from 300s-900s depending on gain settings and relaxation time of the binding process. These setting were optimized for each binding interaction that was measured. A single-site binding model was fit to the titration data using the Bayesian MCMC fitter in pytc (https://github.com/harmslab/pytc). For each protein/peptide combination, one clean ITC trace was used to fit the binding model. Negative results were double-checked to ensure accuracy.

*Preparation of biotinylated proteins for phage display*

A mutant version of hA5 with a single N-terminal Cys residues were generated via site-directed mutagenesis using the QuikChange lightning system (Agilent). The Cys was introduced in the Ser-Asn tag leftover from TEV protease cleavage as Ser-Asn-Cys. The proteins were expressed and purified as described in the previous section. A small amount of the purified proteins were biotinylated using the EZ-link BMCC-biotin system (ThermoFisher Scientific). ≈1 mg BMCC-biotin was dissolved directly in 100% DMSO to a concentration of 8 mM for labeling. Proteins were exchanged into 25mM phosphate, 100mM NaCl, pH 7.4 using a Nap-25 desalting column (GE Health Science) and degassed for 30 min at 25 $^{\circ}C$ using a vacuum pump (Malvern Instruments). While stirring at room temperature, 8mM BMCC-biotin was added dropwise to a final 10X molar excess. Reaction tubes were sealed with PARAFILM (Bemis) and the maleimide-thiol reactions were allowed to proceed for 1 hour at room temperature with stirring. The reactions were then transferred to $4^{\circ}C$ and incubated with stirring overnight to allow completion of the reaction. Excess BMCC-biotin was removed from the labeled proteins by exchanging again over a Nap-25 column (GE Health Science),

and subsequently a series of 3 concentration-wash steps on a NanoSep 3K spin column (Pall corporation), into the Ca-TeBST loading loading buffer. Complete labeling was confirmed by MALDI-TOF mass spectrometry by observing the ≈540Da shift in the protein peak. Final stocks of labeled proteins were prepared at 10 $\mu M$ by dilution into the loading buffer.

*Phage display*

Phage display experiments were performed using the PhD-12 peptide phage display kit (NEB). All steps involving the pipetting of phage-containing samples was done using filter tips (Rainin). We prepared 100 $\mu L$ samples containing phage ($5.5 \times 10^{11}$ PFU) and 0.01 $\mu M$ biotin-protein (or biotin alone in the negative control) and 20 $\mu M$ peptide competitor (in competitor samples) were prepared at room temperature in a background of $Ca^{2+}$-TeBST loading buffer (50mM TES, 100mM NaCl, 2mM $CaCl_2$, 0.01% Tween-20, pH 7.4) to ensure $Ca^{2+}$-saturation of the S100 proteinss. For the experiments including the use of a peptide competitor, the peptide was included at 20 $\mu M$ in the loading buffer. Samples were incubated at room temperature for 2hr. Each sample was then applied to one well of a 96-well high-capacity streptavidin plate (previously blocked using PhD-12 kit blocking buffer and washed 6X with 150 $\mu L$ loading buffer). Samples were incubated on the plate with gentle shaking for 20min. 1 $\mu L$ of 10 $mM$ biotin (NEB) was then added to each sample on the plate and incubated for an additional five minutes to compete away purely biotin-dependent interactions. Samples were then pulled from the plate carefully by pipetting and discarded. Each well was washed 5X with 200 $L$ of loading buffer by applying the solution to the well and then immediately pulling off by pipetting. Finally, 100 $L$ of EDTA-TeBST elution buffer (50mM

TES, 100mM NaCl, 5mM EDTA, 0.01% Tween-20, pH 7.4) was applied to each well and the plate was incubated with gentle shaking for 1hr at room temperature to elute. Eluates were pulled from the plate carefully by pipetting and stored at $4°C$. Eluates were titered to quantify eluted phage as follows. Serial dilutions of the eluates from $1:10$–$1:10^5$ were prepared in LB medium. These were used to inoculate 200 $L$ aliquots of mid-log-phase ER2738 *E. coli* (NEB) by adding 10 $L$ to each. Each 200 $L$ aliquot was then mixed with 3mL of pre-melted top agar, applied to a LB agar XGAL/IPTG (Rx Biosciences) plate, and allowed to cool. The plates were incubated overnight at $37°C$ to allow formation of plaques. The next morning, blue plaques were counted and used to calculate PFU/mL phage concentration. Enrichment was calculated as a ratio of experimental samples to the biotin-only negative control.

To generate the pre-conditioned phage library the nave library was first screened in duplicate against each of the four proteins as descrived above. Each of these lineages was subsequently amplified in ER2738 *E. coli* (NEB) as follows. 20mL 1:100 dilutions of an ER2738 overnight culture were prepared. Each 20mL culture was inoculated with one entire sample of remaining phage eluate. The cultures were incubated at $37°C$ with shaking for 4.5 hours to allow phage growth. Bacteria were then removed by centrifugation and the top 80% of the culture was removed carefully with a filtered serological pipette and transferred to a fresh tube containing 1/6 volume of PEG/NaCl (20% w/v PEG-8000, 2.5M NaCl). Samples were incubated overnight at $4°C$ to precipitate phage. Precipitated phage were isolated by centrifugation and subsequently purified by an additional PEG/NaCl precipitation on ice for 1hr. These individually amplified pools were then resuspended in 200 $L$ each of terile loading buffer and mixed together to

form a pre-conditioned library in order to minimize the impact of sampling on the subsequent panning experiment. The pool was diluted 1:1 with 100% glycerol and stored at $-20°C$ for use in the final panning experiments.

*Preparation of deep sequencing libraries*

Phage genomic ssDNA was isolated from leftover amplified eluates from each round of panning using the M13 spin kit (Qiagen). Products were stored in low TE buffer. These ssDNA were used as the template for 2 replicate PCRs with the Cs1 forward (5′-acactgacgacatggttctacagtggtacctttctattctcactct-3′) and PhD96seq-Cs2 reverse (5′-tacggtagcagagacttggtctccctcatagttagcgtaacg-3′) primers. Products were isolated from these PCR products using the GeneJet gel extraction kit (Thermo Scientific) and pooled. The pooled products were then used as templates for a secondary reaction with the barcoded primers. Products were isolated from these final PCRs using the GeneJet gel extraction kit. Concentration of barcoded samples was measured by $A_{260}/A_{280}$ using a 1mm cuvette on an Eppendorf biospectrometer. Multiplexing was done by mixing samples according to mass. The concentration of the multiplexed library was corrected using qPCR with the P5 and P7 Illumina flow-cell primers. The library was then diluted to a final concentration of 10nM and Illumina sequenced on two lanes of a HiSeq 4000 instrument, using the Cs1 F' as the R1 sequencing primer. The lanes were spiked with 20% PhiX control DNA due to the relatively low diversity of the library.

*Phage display analysis pipeline*

We performed quality control on three read features. First, we verified that the sequence had exactly the anticipated length from the start of the phage

sequence through the stop codon. Second, we only took sequences in which the invariant phage sequence differed by at most one base from the anticipated sequence. This allows for a single point mutation and or sequencing errors, but not wholesale changes in the sequence. Finally, we took only reads with an average phred score better than 15. The vast majority of the reads that failed our quality control did not have the variable region, representing reversion to phage with a wildtype-like coat protein. This analysis is encoded in the *hops_count.py* script, which takes a gzipped fastq file as input and returns the counts for every peptide in the file. Before our main analysis, we discarded any peptide that had fewer than 6 reads associated with it (see Table 10 in supplement). In total, 74.0% of reads passed our quality control and read cutoff.

We clustered peptides using our own implementation of the DBSCAN algorithm [326] using the Damerau-Levensthein distance [327]. The main parameter for DBSCAN clustering is $\varepsilon$—the neighborhood cutoff. Clusters are defined as sequences that can be reached through a series of $\varepsilon$-step moves. We found that $\varepsilon = 1$ gave the best results for our downstream machine learning analysis. Our whole enrichment pipeline—including clustering—can be run given a peptide count file for the non-competitor experiment and a peptide-count file for the competitor experiment using the *hops_enrich.py* script.

We implemented our machine learning model in Python 3 extended with numpy [328], scipy [329], and matplotlib [330]. We used sklearn for our random forest regression [331, 316, 332]. A full list of the calculated features is shown in Table 11 (in supplement). As noted, some features were calculated using CIDER [315]. Our full implementation, including all data files, is available at https://github.com/harmslab/hops.

*Identifying the read count cutoff*

One critical question is at what point the number of reads correlates with the frequency of a peptide. If we set the cutoff too low, we incorporate noise into downstream analyses. If we set the cutoff too high, we remove valuable observations from our dataset. To identify an appropriate cutoff, we studied the mapping between $c_i$ (the number of reads arising from peptide $i$) and $f_i$ (the actual frequency of peptide $i$ in the experiments). Our goal was to find $P(f_i|c_i, N)$: the probability peptide $i$ is at $f_i$ given we observe it $c_i$ times in $N$ counts. Using Bayes theorem, we can write

$$P(f_i|c_i, N) = \frac{P(c_i|f_i, N)P(f_i)}{P(c_i)},$$

where $N$ is the total number of reads. We calculated $P(c_i|f_i, N)$ assuming a binomial sampling process: what is the probability of observing exactly $c$ counts given $N$ independent samples when a population with a peptide frequency $f_i$? This gives the curve seen in Fig 36A (in supplement). We then estimated $P(\hat{f}_i)$ from the distribution of frequencies in the input library, constructing a histogram of apparent peptide frequencies (Fig 36B in supplement). Empirically, we found that frequencies followed an exponential distribution over the measurable range of frequencies. Finally, we assumed that all counts have equal prior probabilities, turning $P(c_i)$ into a scalar that normalizes the integral of $P(f_i|c_i, N)$ so it sums to 1.

Using the information from Fig 36A and B (in supplement), we could then calculate $P(f_i|c_i, N)$ for any number of reads in an experiment $N$. Fig 36C (in supplement) shows this calculation for $N = 2.0 \times 10^7$ reads—a typical number

of reads from our experimental replicates. This curve is linear above 6 reads. Below this, counts no longer correlates linearly with frequency, as it is possible to obtain 5 reads random sampling from low frequency library members. We therefore used a cutoff of 6 counts for all downstream analyses.

*Measuring enrichment values*

We next set out to measure changes in the frequency of peptides between the competitor and non-competitor samples. The simplest way to do this would be to identify peptides seen in both experiments, and then measure how their frequencies change between conditions. Unfortunately, these proteins all bind a wide swath of peptide targets and relatively few peptides were shared between conditions. This approach would thus exclude the majority of sequences. For example, only 8,672 of the 112,681 unique peptides observed for hA5 were present in both the competitor and non-competitor, even after pooling biological replicates. Worse, because we are interested in peptides that are lost when competitor peptide is added, ignoring peptides with no counts in the competitor sample means ignoring some of the most informative peptides.

To solve this problem, we clustered similar peptides and measured enrichment for peptide clusters rather than individual peptides. We extracted all peptides that were observed across the competitor and non-competitor samples for a given protein. We then used DBSCAN to cluster those peptides according to sequence similarity, as measured by their their Damerau-Levenshtein distance [327, 326]. This revealed extensive structure in our data. For example, hA5 yielded 8,645 clusters with more than one peptide, incorporating more than half of the unique peptides (Fig 21A, Fig 38A in supplement). We chose clustering parameters

that led to highly similar peptides within each cluster, as can be seen by the representative sequence logos for three clusters of hA5 (Fig 38B in supplement). Sequences that were not placed in clusters were treated as clusters with a size of one.

We then used the enrichment of each cluster to estimate the enrichment of individual peptides. We defined enrichment as:

$$E_{cluster} = -ln\left(\frac{\sum_{i=1}^{i\leq N}\beta_i}{\sum_{i=1}^{i\leq N}\alpha_i}\right), \tag{6.1}$$

where $N$ is the total number of peptides in the cluster, $\beta_i$ is the frequency of peptide $i$ in the competitor sample, and $\alpha_i$ is the frequency of peptide $i$ in the non-competitor sample. We then made the approximation that all members of the cluster have the same enrichment:

$$E_i \approx E_{cluster}, \tag{6.2}$$

allowing us to estimate the enrichment of all $i$ peptides in the cluster (Fig 38C in supplement). Peptides lost because of competition for the interface will add zeros to the numerator of Eq. 6.1, leading to an overall decrease in enrichment. Peptides missed because of finite sampling will add zeros evenly to the competitor and non-competitor samples, leading to no net enrichment.

We tested this cluster-based approximation using the 8,672 peptides of hA5 for which we could directly calculate enrichment (that is, those peptides seen in both the competitor and non-competitor experiments). We calculated the enrichment of each peptide individually and compared these values to those obtained by the cluster method. There is no systematic difference in the values

estimated using the two methods, and the linear model explains 98.4% of the variation between the two methods.

*Principle Component Analysis*

To generate the aaindex meta features, we performed a principle component analysis on all 590 features from the aaindex database. Any missing value was assigned the mean value of that feature. Prior to performing the PCA, we standardized all values to a mean of zero and a standard deviation of 1. This yielded 20 principle components.

*Incorporating uncertainty into an estimate of a Venn diagram*

We used a Bayesian approach to estimate the overlaps between the binding sets of proteins, despite high false positive and false negative rates. Consider a set of peptides binding to the proteins $A$ and $B$. The binding of these peptides can be described by a Venn diagram with four regions: $[A \cup B]^c$ (peptides that bind neither $A$ nor $B$), $A \backslash B$ (peptides that bind $A$ alone), $B \backslash A$ (peptides that bind $B$ alone), and $A \cap B$ (peptides that bind both $A$ and $B$). The number of peptides in each region is given by $\vec{V}$, while the number of peptides observed in each region is given by $\vec{V}_{obs}$. $\vec{V}$ and $\vec{V}_{obs}$ can differ as there may be both false positives (at rates $m_A$ and $m_B$) and false negatives (at rates $n_A$ and $n_B$). We can write a row-stochastic matrix that describes the probability of observing a peptide in a region given its actual region as:

$$\mathbf{T} = \begin{bmatrix} P([A\cup B]^c|[A\cup B]^c) & P(A\backslash B|[A\cup B]^c) & P(B\backslash A|[A\cup B]^c) & P(A\cap B|[A\cup B]^c) \\ P([A\cup B]^c|A\backslash B) & P(A\backslash B|A\backslash B) & P(B\backslash A|A\backslash B) & P(A\cap B|A\backslash B) \\ P([A\cup B]^c|B\backslash A) & P(A\backslash B|B\backslash A) & P(B\backslash A|B\backslash A) & P(A\cap B|B\backslash A) \\ P([A\cup B]^c|A\cap B) & P(A\backslash B|A\cap B) & P(B\backslash A|A\cap B) & P(A\cap B|A\cap B) \end{bmatrix}$$

where each conditional probability $P(X|Y)$ describes the probability of observing the peptide in region $X$ given it is actually in region $Y$. If we know this matrix and we know the real population in each region, we can calculate $\vec{V}_{obs}$ by:

$$\vec{V}_{obs} = \vec{V} \cdot \mathbf{T}.$$

We can construct $\mathbf{T}$ using the false positive and false negative rates for binding to protein $A$ or $B$. For example, the probability of seeing a peptide that binds to $A$ alone when it actually does not bind to either $A$ or $B$ would be

$$P(A\backslash B|[A \cup B]^c) = m_A - m_A m_B :$$

the probability of a false positive for $A$ less the probability of a false positive for both $A$ and $B$. Using appropriate combinations of false positive and false negative rates, we can calculate every value in $\mathbf{T}$:

$$\mathbf{T} = \begin{bmatrix} 1-(m_A+m_B-m_A m_B) & m_A-m_A m_B & m_B-m_A m_B & m_A m_B \\ n_A-n_A m_B & 1-(n_A+m_B-n_A m_B) & n_A m_B & m_B-n_A m_B \\ n_B-m_A n_B & m_A n_B & 1-(m_A+n_B-m_A n_B) & m_A-m_A n_B \\ n_A n_B & n_B-n_A n_B & n_A-n_A n_B & 1-(n_A+n_B-n_A n_B) \end{bmatrix}$$

This can be readily extended to any number of proteins with any number of possible overlaps.

We can then estimate $\vec{V}$ using Bayesian Markov Chain Monte Carlo (MCCE). We first write a likelihood function:

$$ln\left[P(\vec{V}_{obs}|\vec{V},\{m\},\{n\})\right] = -\frac{1}{2}\sum_i \left[(\vec{V}_{obs,i} - \vec{V}_i\mathbf{T})^2/\sigma_i^2 + ln(\sigma_i^2)\right]$$

where $i$ indexes regions in the Venn diagram, $\sigma_i^2$ is the uncertainty of the counts in region $i$, $\{m\}$ is the set of false positive rates and $\{n\}$ is the set of false negative rates. We can then sample values in $\vec{V}$, $\{m\}$ and $\{n\}$ by MCCE. For $\vec{V}$, we used the prior:

$$ln\left[P(\vec{V})\right] = \begin{cases} -\infty & \vec{V} < 0 \\ 0 & \vec{V} \geq 0 \end{cases},$$

thus requiring all regions to have positive counts. We also constrained the number of counts in $\vec{V}$ be within 5% of the number of counts in $\vec{V}_{obs}$ ($N$):

$$ln[P(\vec{V})] = \begin{cases} -\infty & \sum \vec{V} < 0.95N \\ 0 & 0.95N \leq \sum \vec{V} \leq 1.05N \\ -\infty & \sum \vec{V} > 1.05N \end{cases}$$

For every false positive or false negative rate (denoted as $r_j$), we used the prior:

$$ln\left[P(r_j)\right] = \begin{cases} -\infty & r_j < 0 \\ -\frac{(r_j-\hat{\mu}_j)^2}{2\sigma_j^2} + \sqrt{2\pi\sigma_j^2} & 0 \leq r_j \leq 1, \\ -\infty & r_j > 1 \end{cases}$$

where $\hat{\mu}_j$ is the estimate of the value of $r_j$ from our binding experiments and $\sigma_j$ was set to 0.2. For values outside of 0 and 1, the log prior is $-\infty$, enforcing bounds on these parameters.

## Bridge to Chapter VII

In this chapter a novel experimental pipeline was developed to quantify protein binding specificity in an unbiased fashion. Peptide phage display and high-throughput sequencing were coupled to trace the meanderings of peptide binding specificity in a clade formed by two S100s—S100A5 and S100A6—resultant from an ancient gene duplication. Vast datasets of peptides bound by the proteins were generated and the data were then analyzed using an advanced machine learning approach. The results of this pipeline demonstrate that the bulk of explanatory power that allows peptide binding partners to be predicted comes from overall biochemical features of the peptide targets. This is in stark contrast to studies done previously on more specific protein families, wherein sequence-properties are sufficient to predict targets. Estimating the total sets of binding partners recognized by S100A5, S100A6, and ancA5/A6 (the resurrected ancestor of the two clades) reveals how total sets of potential binding partners and the biochemical determinants of these sets evolved. An interesting historical evolutionary pattern is uncovered, wherein diversification appears to have happened on both protein lineages following gene duplication. However, the S100A5 lineage experiences subfunctionilzation of binding partners relative to the ancestral protein, while S100A6 appears to have shifted it's specificity laterally. This result demonstrates that diversification of biochemical phenotypes can be nuanced and complex when viewed from an unbiased, global perspective. It further highlights a disagreement between low-specificity and high-specificity methods for measuring changes in specificity and introduced a broadly-useful technique for addressing this issue. Furthermore, this chapter emphasizes that experiments with low-specificity proteins—such as the S100s—should focus on understanding how global biochemical features underly binding preferences, rather than attempting to focus on sequence motifs. In chapter VII, the results of the entire dissertation are summarized and the implications are discussed. The broader

contributions to the field of evolutionary biochemistry are highlighted. The limitations of the results are also discussed and future directions are outlined that expand on the work presented here.

# CHAPTER VII

# SUMMARY AND CONCLUDING REMARKS

## *Contributions to the field of evolutionary biochemistry*

The tools of evolutionary biochemistry have become an important approach to understanding molecular evolution. This field provides the basis to determine how the biochemical properties of organisms—at the level of proteins—shape the evolution of new phenotypes. A broad array of biologically-relevant molecular-level changes have been characterized, revealing the importance of biochemical constraints in contributing to evolutionary changes in signalling [52], metabolism [53], nutrient transport [55], and many other aspects of biology. In many cases these changes required biochemical modifications that altered the recognized binding partners of proteins [58, 53, 272, 273, 60, 290, 292]. Despite substantial progress in understaning the molecular basis of evolutionary alterations, there are still a vast number of unsanwered questions. One limitation of previous studies has been to focus on proteins with very specific sets of binding partners. Exquisite specificity is important for biology. However, many proteins exhibit the ability to interact with highly-diverse binding partners. These proteins are also critical for many biological processes, but previous work has shied away from using such proteins as models. The biological roles and relevance of biochemically-defined sets of binding partners are less obvious in these cases than in more canonical examples of proteins with tight specificity. This dissertation focused on helping to close the gap in understanding how proteins with highly-variable binding partners and binding sites evolve new biochemical specificity.

The work presented in this dissertation represents a contribution to the field of evolutionary biochemistry. It makes important contributions to understanding the evolution of proteins with diverse biochemical features, which differ from many other

model protein systems in the variability of their binding partners. The S100 proteins proved to be a useful model for probing the evolution of binding specificity. Tracing the evolution of of both metal ions and small peptide binding by the S100s revealed several key observations. 1) A biochemical output can be conserved over evolutionary time despite extensive amino acid turnover in binding sites. 2) Proteins with low biochemical specificity can nonetheless be subject to evolutionary constraints that maintain a given specificity profile. 3) The evolutionary patterns that follow gene duplications in low-specicificty proteins are similar to those observed in high-specificity proteins. 4) Following gene duplication the duplicate lineages can undergo differential changes in specificity, such as subfunctionalization on one lineage and neofunctionalizatin on the other.

*Limitations and future directions*

The work in this dissertation has contributed to understanding how proteins with highly-variable binding partners evolve new biochemical specificity. The studies presented here are the first to address this problem in a systematic way. It paves the way for future studies to improve upon methods, model systems, and connect what has been learned to more nuanced questions regarding the evolution of binding specificity.

Nonetheless, there are limitations to this work that should be considered. For example, the evolution of metal binding was traced across the entire S100 protein family. This work revealed two key observations: 1) the overall biochemical output of transition metal can be maintained despite extreme lability of metal-binding sites, and 2) there is substantial variation in the structural output of metal binding. These observations suggested that there has been some degree of biochemical specilization in the response to metal binding that could potentially have direct biological consequences. The known roles of transition metal binding in S100s range from antimicrobial sequestration metals [105] to metal-chaperoning as part of a signalling pathway [172]. It seems unlikely that a biochemical ouput conserved across the family has been maintained for hundreds of

millions of years in the absence of a biological role. However, the roles of transition metal binding remain unknown for most S100 proteins [170, 171, 96]. The work presented here remains merely suggestive of the biological relevance of this biochemical behavior. The downstream output of the S100 biochemistry was not characterized in its natural cellular environment. Future work should thus focus on understanding what role transition metal binding is playing in the relevant physiological context. Furthermore, although this work revealed the extreme variability of binding site ligands and locations, it nonetheless leaves open the question of exactly which ligands are used throughout the family. For example, the $Cu^{2+}$ binding ligands of S100A5 still remain unknown. Future studies should this also seek to map out the transition metal binding sites of S100s. Eliminating binding sites *in vivo* via site-directed mutagenesis will further allow their biology of transition metal binding to be probed.

With regard to the studies of peptide binding specificity presented in this dissertation, an unbiased approach was used to trace the evolutionary history of peptide binding specificity in S100A5 and S100A6 following gene duplication. This method allowed the total scope of binding partners to be estimate for each of the duplicate lineages. Combining the method with ASR further allowed the evolutionary dimension to be directly assessed, revealing the historical patterns of changes in specificity that occurred following duplication. The patterns, when compared to those observed via a more traditional low-throughput method, highlighted the importance of using such a global, unbiased approach. The low-throughput method provided valuable, gold-standard information that clearly demonstrates conservation of specificity profiles within paralogous lineages. However, it lacked the resolution to characterize how the size and divsersity of binding sets had changed during evolution. What appeared to be a symmetric pattern of subfunctionilzation from a less-specific ancestor on both the S100A5 and S100A6 lineages was revealed by the high-throughput approach to be far more nuanced. S100A5 did in fact undergo subfunctionlization, but S100A6 actually underwent

a shift in specificity. In fact, it is possible that the scope of binding partners for S100A6 actually increased over evolutionary time. This result clearly shows the superiority of using an unbiased high-throughput approach to chacterize the evolution of specificity, which is further accentuated by the very low specificity of proteins like the S100s.

The unbiased nature of the approach used to measure specificity is the key strength of the method. However, this is also one of the key limiations. The very fact that the approach utilized a random set of peptide targets divorces it from direct biological implications. The low-throughput study revealed strong conservation of binding specificity profiles in duplicate lineage. This results strongly argues that there is indeed biological relevance for the biochemical specificity of these proteins, because it seems very unlikely that these profiles would be maintained purely by chance over 320 million years of evolution without some sort of selection to maintain the set of binding patners. This argument is particularly strong considering that specificiy can be readily altered in these proteins by a single amino acid substitution. However, the biological implications of biochemical specificity in S100A5 and S100A6 were not directly assessed and remain firmly in the realm of speculation. This limitation opens up three key questions that future studies should seek to address. 1) How does the biologically-realized set of binding partners compare to the scope of all possible partners defined by biochemistry? 2) What are the biological forces that winnow the possible set of partners to the realized set? 3) What is the biological output of the biochemical specificity that has been conserved for so long? Answering these questions will provide a more complete picture of how specificity evolves in the S100s, the forces that shape it, and the biological implications for evolutionary changes in specificity.

# APPENDIX A

# SUPPLEMENTAL MATERIAL FOR CHAPTER III

## Supplemental Figures

This section includes the supplemental figures referenced in chapter III. Other supplemental files such as spreadsheets, newick trees, and multiple sequence alignments are included in the chapter 3 sub–directory of the zipped supplemental directory submitted with this dissertation.

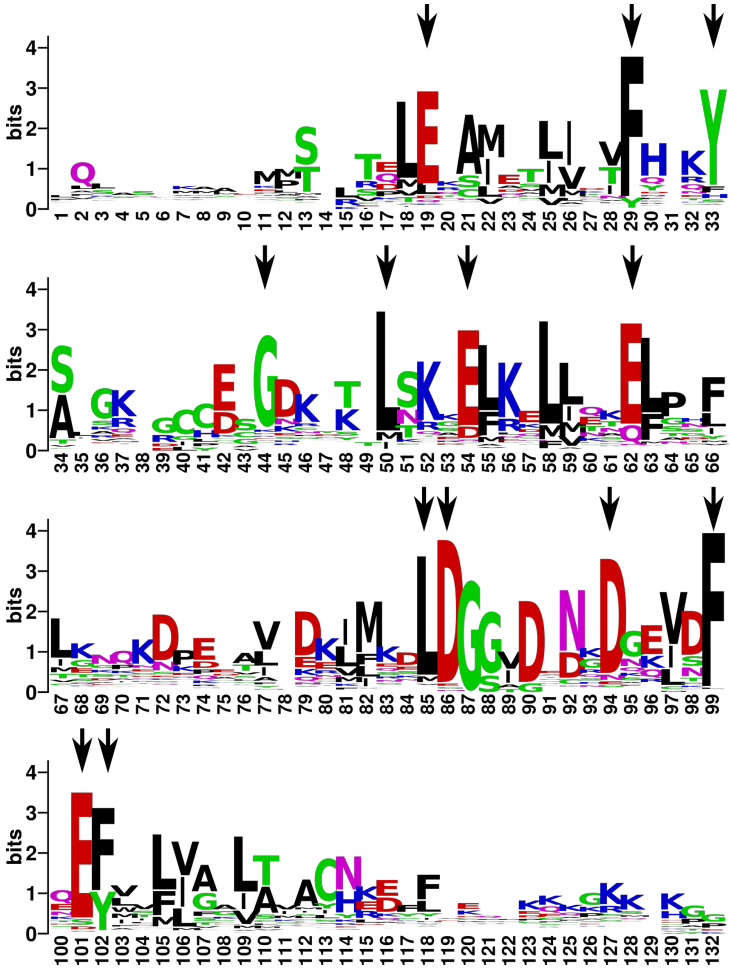


FIGURE 24 Sequence logo indicates relative frequency of amino acids at each position in the alignment. Taller letters indicate higher frequency at that position. Arrows indicate 13 key residues we used to verify/anchor the alignment.

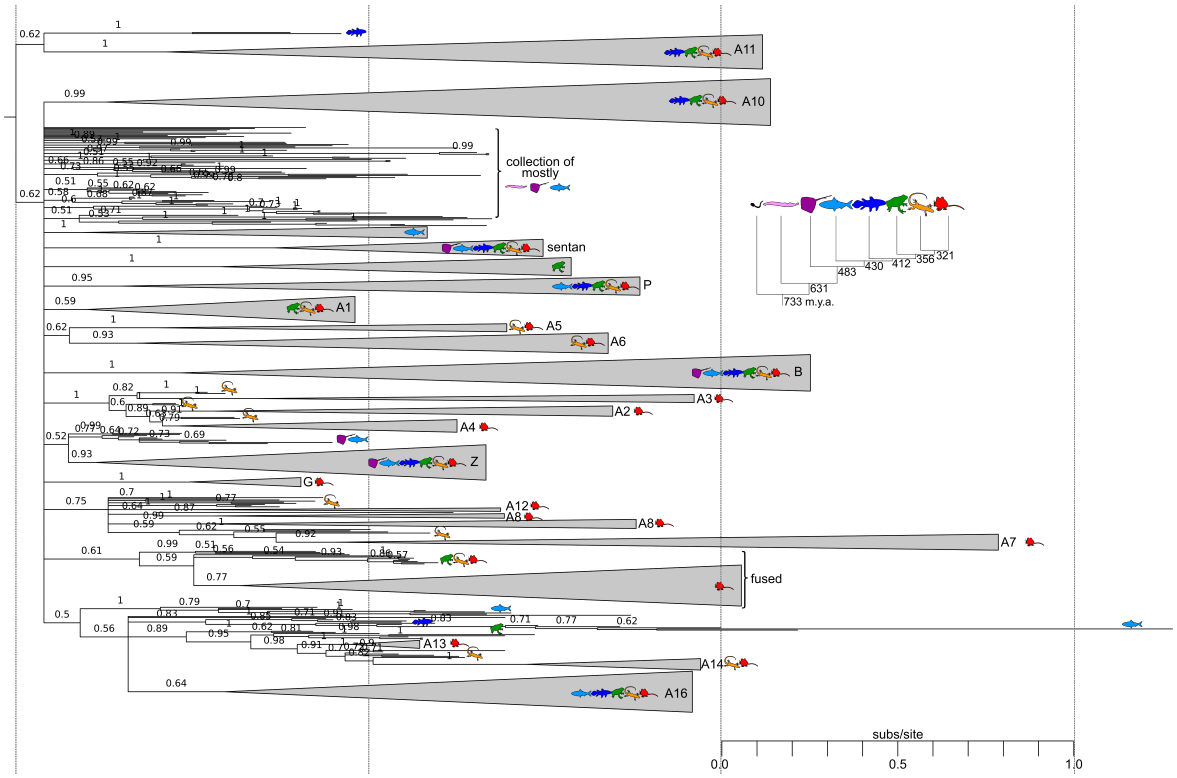


FIGURE 25 Tree is a majority rule consensus tree, with all nodes with posterior probabilities <50% collapsed into polytomies. Wedges are collapsed clades of shared orthologs, with wedge height denoting number of included taxa and wedge length denoting longest branch length with the clade. Support values are posterior probabilities. Rooting is arbitrary given the poor resolution at the base of the taxonomic tree. Icons indicate taxonomic classes represented within each clade: tunicates (black sea squirt), jawless fishes (pink lamprey), cartilaginous fishes (purple ray), ray-finned fishes (light blue fish), lobe-finned fishes (blue coelacanth), amphibans (green frog), birds/reptiles (yellow lizard), and mammals (red mouse). Inset shows estimated divergence times for each taxonomic class in millions of years before present.

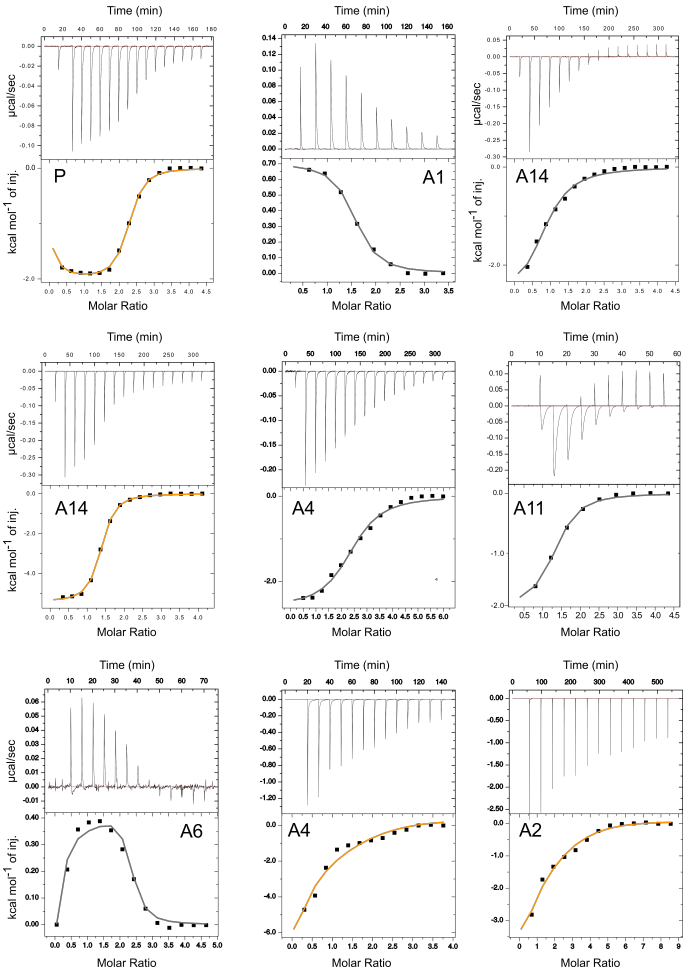


FIGURE 26 Each panel is a single human paralog, indicated by the name on the graph. Color of fit indicates metal used as titrant: $Zn^{2+}$ (gray) or $Cu^{2+}$ (copper). Top sub-panel for each panel is a raw power vs. time curve. Bottom sub-panel for each panel is integrated heat versus molar ratio. The model fit is denoted by the heavy line through the fit points.

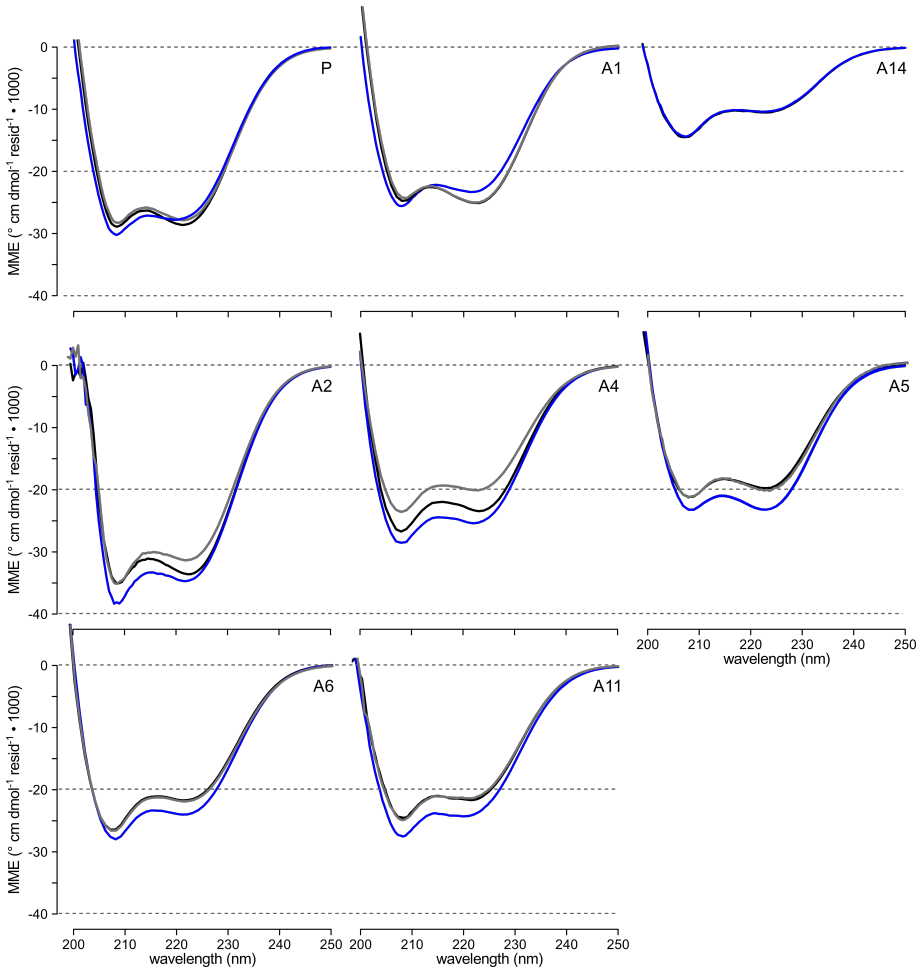


FIGURE 27 Curves are far-UV CD spectra (mean molar ellipticity vs. wavelength). Colors represent metal: apo (black), $Zn^{2+}$ (gray), and $Ca^{2+}$ (blue). Paralog is indicated to the right of each spectrum.

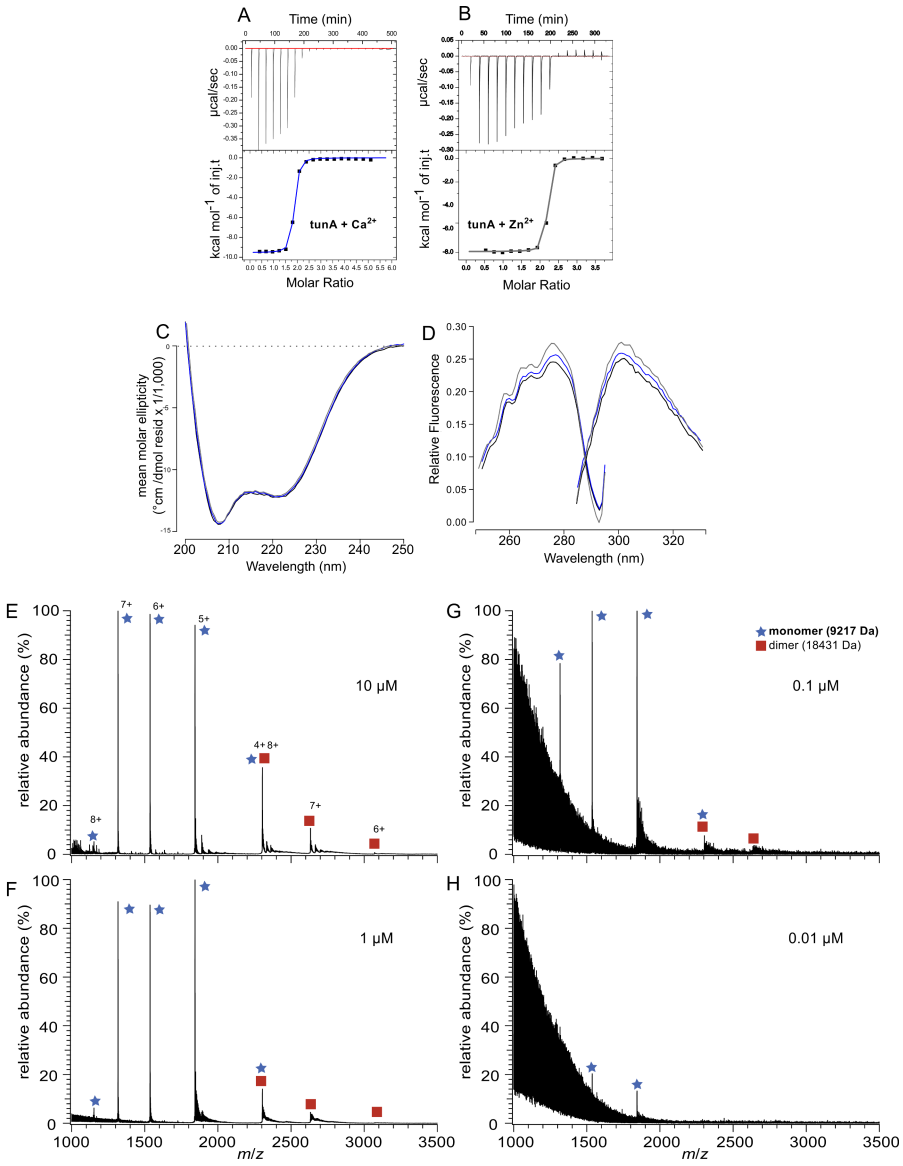


FIGURE 28 A) ITC trace for binding of $Ca^{2+}$. B) ITC trace for binding of $Zn^{2+}$. C) Far-UV CD spectra for tunA in apo form (black), presence of $Ca^{2+}$ (blue) and presence of $Zn^{2+}$ (gray). D) Intrinsic fluorescence spectra for tunA with conditions as in panel C. E-H) ESI-MS spectra for tunA, titrating from 10 $\mu$M to 0.01 $\mu$M protein. Icons indicate species (monomer or dimer). Numbers indicate charge state. Dimer is lost preferentially during dilution, suggesting it is an artifact of electrospray process.

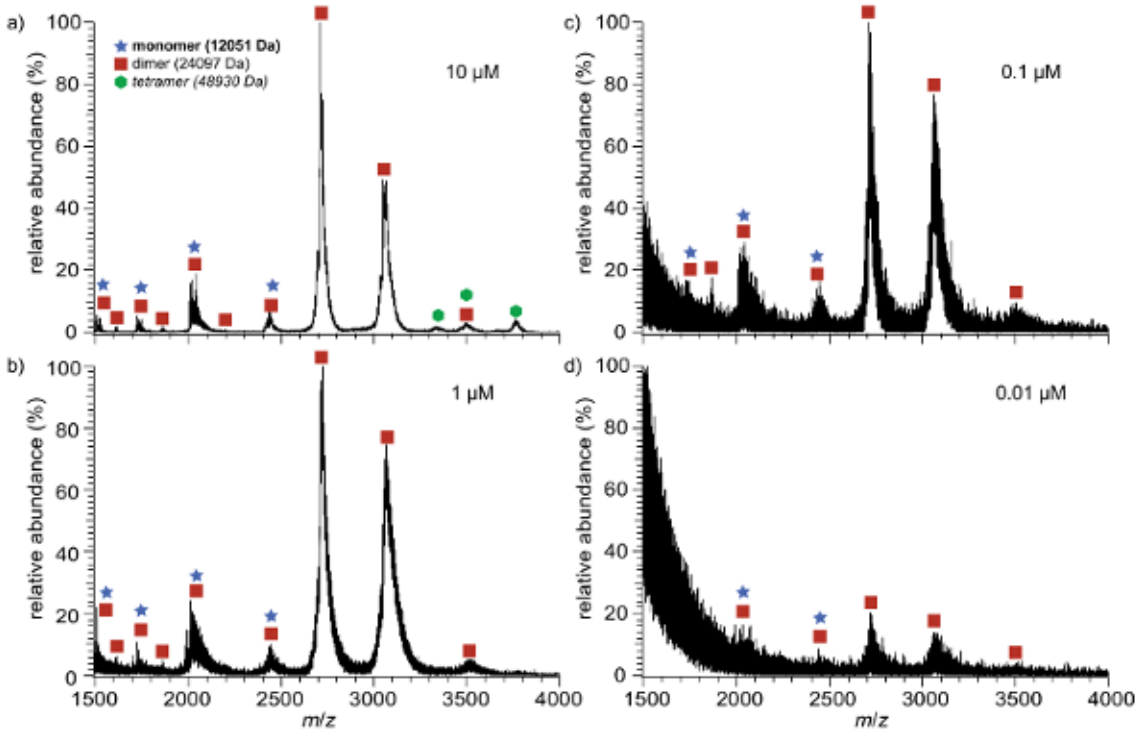


FIGURE 29 tunB mass spectra at concentrations of a) 10 $\mu$M, b) 1 $\mu$M, c) 0.1 $\mu$M, and d) 0.01 $\mu$M demonstrate that tunB homodimers are robust to dilution, indicating that this is a specific interaction. Homotetramer is observed only in the most concentrated sample, thus homotetramer signal likely arises from non-specific interactions during the electrospray process.

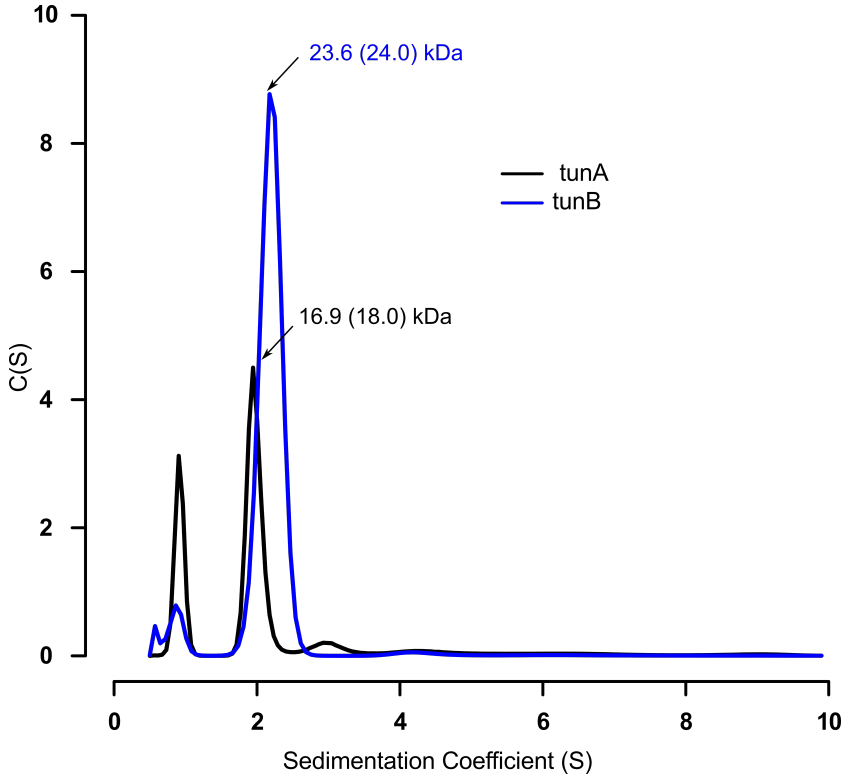


FIGURE 30 Graph shows the distribution of sedimentation coefficient determined for tunA (black) and tunB (blue). The apparent mass of the homodimer peaks are indicated above each peak, with the mass expected from the amino acid sequence of the protein in parentheses.

# APPENDIX B

# SUPPLEMENTAL MATERIAL FOR CHAPTER IV

## Supplemental Figures

This section includes the supplemental figures referenced in chapter IV.

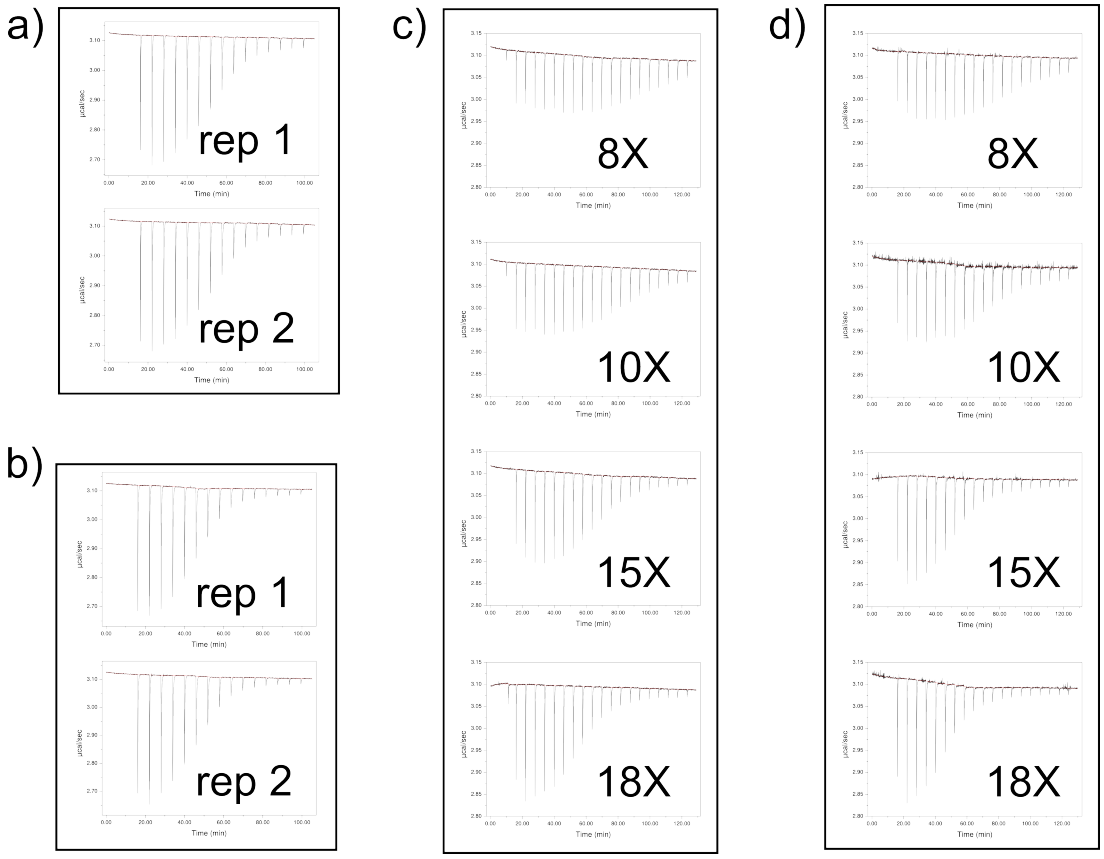


FIGURE 31 Raw data corresponding to integrated heats in figure 11. a) hA5 binding $Cu^{2+}$ , b) Ca 2+ loaded hA5 binding Cu 2+ , c) hA5 binding $Ca^{2+}$ , and d) $Cu^{2+}$—loaded hA5 binding $Ca^{2+}$ .

APPENDIX C

SUPPLEMENTAL MATERIAL FOR CHAPTER V

## Supplemental Figures

This section includes the supplemental figures referenced in chapter V. Other supplemental files such as spreadsheets, newick trees, and multiple sequence alignments are included in the chapter 5 sub–directory of the zipped supplemental directory submitted with this dissertation.

TABLE 3 Binding of 12-mer phage display peptides does not depend on solubilizing flanks. List of phage display consensus peptides used in the study. The sequences of flank variants of A5cons and A6cons are shown. Flanks are indicated by lower-case letters. The third column shows dissociation constants for peptides binding to hA5 with 95% credibility regions from Bayesian fits of one ITC dataset per variant. Flank variants bind with similar $K_D$.

| Peptide Name | Amino Acid Sequence | $K_D(\mu M)$ |
|---|---|---|
| A5cons (variant 1) | `rshsSSFQDWLLSRLPgggsae` | $4.9 \leq 6.1 \leq 7.8$ |
| A5cons (variant 2) | `----SSFQDWLLSRLP-ggsae` | $1.1 \leq 2.8 \leq 7.9$ |
| A5cons (variant 3) | `rshsSSFQDWLLSRLP------` | $7.2 \leq 9.6 \leq 13.1$ |
| A6cons (variant 1) | `rshsGFDWRWGMEALTgggsae` | $0.3 \leq 0.9 \leq 2.4$ |
| A6cons (variant 2) | `----GFDWRWGMEALT-ggsae` | $1.5 \leq 2.5 \leq 4.0$ |

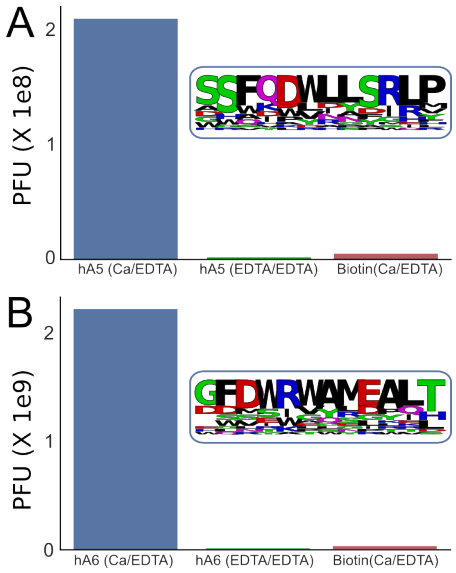


FIGURE 32 Randomer phage enrichment is dependent on $Ca^{2+}$ and protein. Bar graphs show the plaque forming units (PFU) for phage solutions after the third round of enrichment for screens using hA5 (A) or hA6 (B). For each round of panning, we incubated phage with biotinylated protein, pulled down bound phage via a streptavidin plate, and finally eluted the phage from the protein with an elution buffer. To verify that binding occurred in a $Ca^{2+}$-dependent manner, we compared $Ca^{+}$-loading/$EDTA$-elution to $EDTA$-loading/$EDTA$-elution. We also performed a $Ca^{2+}$-loading/$EDTA$-elution experiment using biotin alone. Insets show sequence logos (WebLogo) generated from 20 plaque sequences from each $Ca^{2+}/EDTA$ panning experiment. The most frequent residue at each position was used to generate the A5cons and A6cons peptides.

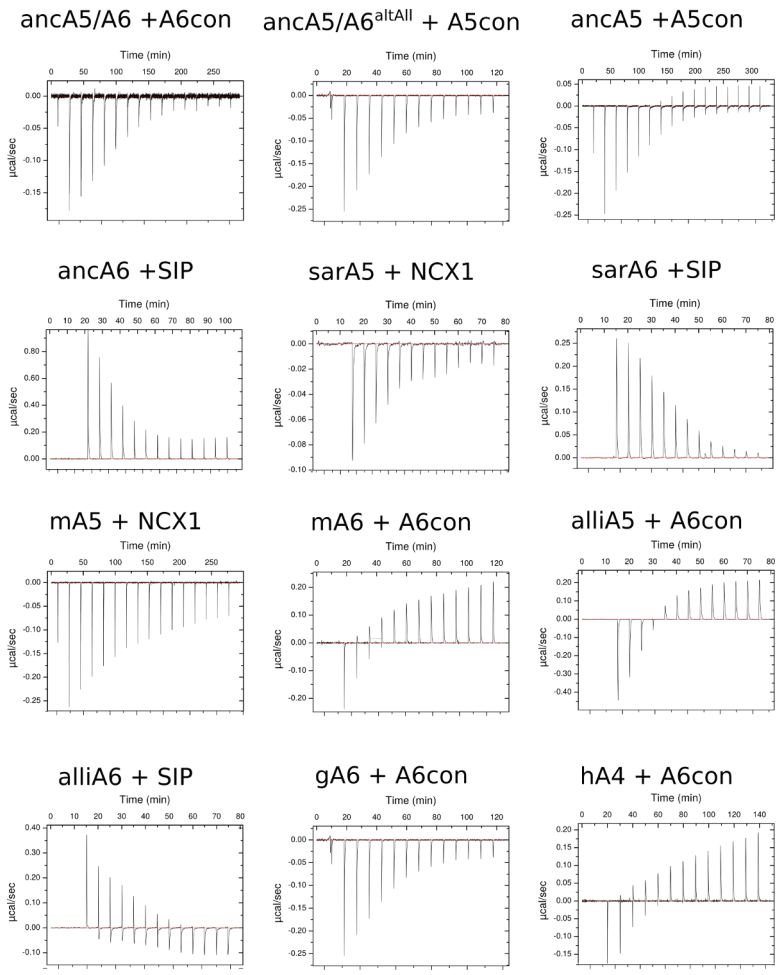


FIGURE 33 ITC traces show baseline-corrected titration of various peptides onto S100 proteins in the presence of 2 $mM$ $Ca^{2+}$. All experiments were done with $\approx$ 100 $\mu M$ protein in 25 $mM$ TES, 100 $mM$ NaCl, 1 $mM$ $TCEP$ at $pH$ 7.4, 25 $^{\circ}C$.CD spectra are mapped onto a diagram of the S100A5-S100A6 clade. Curves are spectra of apo (gray) and $Ca^{2+}$bound (orange/purple) proteins. The S100A5 proteins (purple) are characterized by a deep alpha-helical signal at 222nm that substantially increases in response to binding of $Ca^{2+}$. S100A6 proteins (orange) show comparatively minimal response and maintain a deeper peak at 208nm. These patterns hold for the ancestors at the base of each clade. The spectra of ancA5/A6 and the ancA5/A6 altAll version (both shown in green) resemble that of an extant S100A6, indicating that the large $Ca^{2+}$-driven conformational change seen in the extant S100A5s is a derived feature of this lineage.

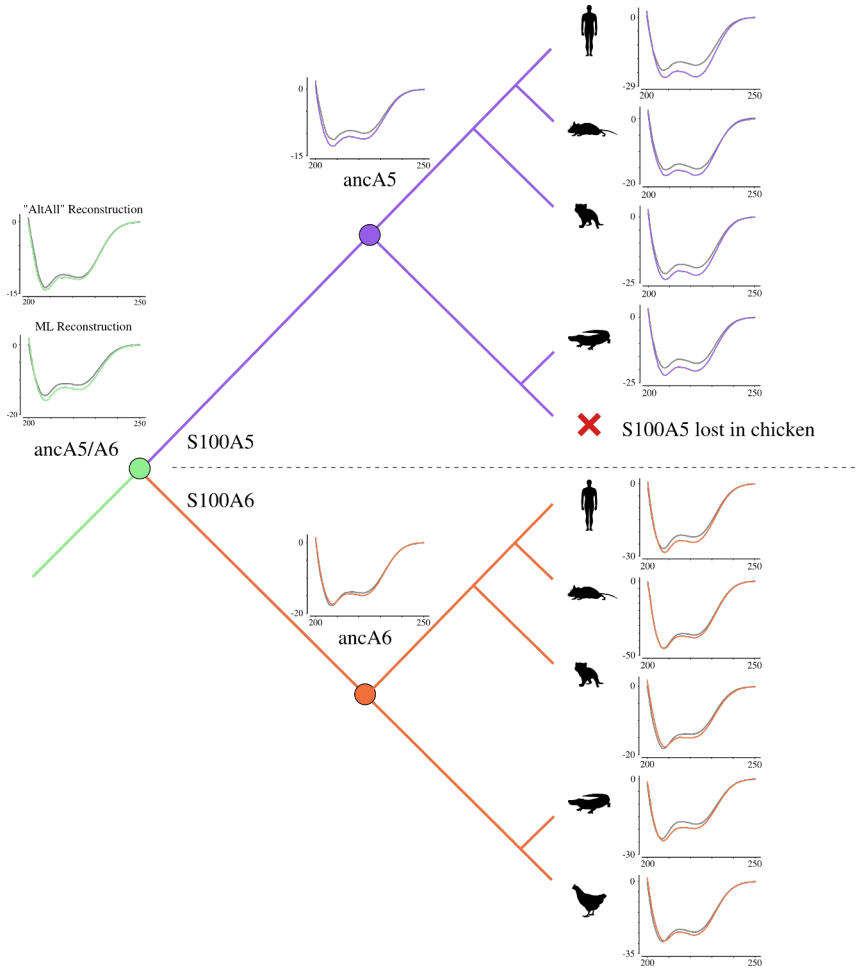


FIGURE 34 Far UV CD spectra are diagnostic for the S100A5 and S100A6 clades. CD spectra are mapped onto a diagram of the S100A5-S100A6 clade. Curves are spectra of apo (gray) and $Ca^{2+}$bound (orange/purple) proteins. The S100A5 proteins (purple) are characterized by a deep alpha-helical signal at 222nm that substantially increases in response to binding of $Ca^{2+}$. S100A6 proteins (orange) show comparatively minimal response and maintain a deeper peak at 208nm. These patterns hold for the ancestors at the base of each clade. The spectra of ancA5/A6 and the ancA5/A6 altAll version (both shown in green) resemble that of an extant S100A6, indicating that the large $Ca^{2+}$-driven conformational change seen in the extant S100A5s is a derived feature of this lineage.

TABLE 4 Thermodynamic parameters for binding of the peptide rshsGFDWRWAMEALTggsae (A6cons) to S100A5 and S100A6 proteins.Species abbreviations are "alli" (alligator), "gal" (chicken), "sar" (tasmanian devil), "m" (mouse), and "h" (human). Fit parameters, with standard deviation from fits, for to the data shown schematically in Fig 4A. Parameters are for a single-site binding model. "NA" indicates that there was no detectable binding. We floated the fraction competent parameter to capture uncertainty in peptide and protein concentration, particularly given the low extinction coefficients of S100A5 and S100A6. If an experiment was done at both 10 and 25 $°C$, the parameters correspond to the 10 $°C$ experiment.

| protein | $K_A$ $(M^{-1})$ | $\Delta H°$ $(kcal/mol)$ | $fx\ comp.$ | $num\ reps$ | $T$ $(°C)$ |
|---|---|---|---|---|---|
| ancA5/A6 | $8.30e5 \pm 1.9e5$ | $-12.20 \pm 1.2$ | $0.70 \pm 0.02$ | 2 | 25 |
| altAll | $1.10e5 \pm 5.2e4$ | $-8.70 \pm 0.6$ | $0.90 \pm 0.07$ | 2 | 25 |
| ancA5 | $7.70e5 \pm 2.1e5$ | $-3.30 \pm 0.8$ | $0.80 \pm 0.07$ | 2 | 25 |
| alliA5 | $4.40e5 \pm 6.8e4$ | $-10.60 \pm 0.7$ | $0.70 \pm 0.01$ | 2 | 25 |
| sarA5 | $2.50e5 \pm 1.6e5$ | $-5.90 \pm 2.5$ | $0.90 \pm 0.13$ | 2 | 25 |
| mA5 | $2.10e5 \pm 5.4e4$ | $-11.70 \pm 2.7$ | $1.10 \pm 0.05$ | 2 | 25 |
| hA5 | $4.10e5 \pm 9.8e4$ | $8.50 \pm 1.5$ | $1.00 \pm 0.02$ | 2 | 25 |
| ancA6 | $2.80e5 \pm 1.9e5$ | $-6.40 \pm 2.7$ | $1.10 \pm 0.14$ | 2 | 25 |
| alliA6 | $9.50e4 \pm 4.7e4$ | $-10.40 \pm 3.7$ | $0.60 \pm 0.09$ | 2 | 25 |
| gA6 | $4.20e5 \pm 2.0e5$ | $-8.10 \pm 2.0$ | $0.70 \pm 0.06$ | 2 | 25 |
| sarA6 | $1.40e5 \pm 6.7e4$ | $-6.20 \pm 1.9$ | $0.80 \pm 0.10$ | 2 | 25 |
| mA6 | $2.60e5 \pm 1.2e5$ | $-6.40 \pm 1.5$ | $0.60 \pm 0.05$ | 2 | 25 |
| hA6 | $2.00e5 \pm 4.8e4$ | $9.60 \pm 1.4$ | $0.80 \pm 0.02$ | 2 | 25 |
| hA4 | $2.80e6 \pm 6.5e6$ | $-1.80 \pm 0.5$ | $0.60 \pm 0.04$ | 2 | 25 |

TABLE 5 Thermodynamic parameters for binding of the peptide rshsSSFQDWLLSRLPgggsae (A5cons) to S100A5 and S100A6 proteins.Species abbreviations are "alli" (alligator), "gal" (chicken), "sar" (tasmanian devil), "m" (mouse), and "h" (human). Fit parameters, with standard deviation from fits, for to the data shown schematically in Fig 4A. Parameters are for a single-site binding model. "NA" indicates that there was no detectable binding. We floated the fraction competent parameter to capture uncertainty in peptide and protein concentration, particularly given the low extinction coefficients of S100A5 and S100A6. If an experiment was done at both 10 and 25 $°C$, the parameters correspond to the 10 $°C$ experiment.

| protein | $K_A$ $(M^{-1})$ | $\Delta H°$ $(kcal/mol)$ | $fx\ comp.$ | $num\ reps$ | $T$ $(°C)$ |
|---|---|---|---|---|---|
| ancA5/A6 | $9.30e4 \pm 3.0e4$ | $-5.20 \pm 1.6$ | $1.40 \pm 0.07$ | 2 | 25 |
| altAll | $4.70e4 \pm 2.2e4$ | $-3.90 \pm 1.3$ | $1.30 \pm 0.19$ | 2 | 25 |
| ancA5 | $1.30e5 \pm 3.6e4$ | $-6.90 \pm 1.3$ | $0.90 \pm 0.05$ | 2 | 10, 25 |
| alliA5 | $2.30e4 \pm 3.8e3$ | $13.80 \pm 2.4$ | $1.10 \pm 0.07$ | 2 | 10, 25 |
| sarA5 | $2.10e5 \pm 1.5e5$ | $-4.80 \pm 1.9$ | $0.70 \pm 0.1$ | 2 | 25 |
| mA5 | $4.70e4 \pm 1.9e4$ | $-6.90 \pm 2.1$ | $0.60 \pm 0.08$ | 2 | 25 |
| hA5 | $3.60e5 \pm 2.1e5$ | $-5.70 \pm 1.7$ | $0.80 \pm 0.06$ | 2 | 25 |
| ancA6 | NA | NA | NA | 2 | 10, 25 |
| alliA6 | NA | NA | NA | 2 | 25 |
| gA6 | NA | NA | NA | 2 | 25 |
| sarA6 | NA | NA | NA | 2 | 10, 25 |
| mA6 | NA | NA | NA | 2 | 25 |
| hA6 | NA | NA | NA | 2 | 25 |
| hA4 | $1.70e4 \pm 5.1e3$ | $-4.10 \pm 0.8$ | $0.90 \pm 0.3$ | 2 | 25 |

TABLE 6 Thermodynamic parameters for binding of the peptide RRLLFYKYVYKR (NCX1) to S100A5 and S100A6 proteins.Species abbreviations are "alli" (alligator), "gal" (chicken), "sar" (tasmanian devil), "m" (mouse), and "h" (human). Fit parameters, with standard deviation from fits, for to the data shown schematically in Fig 4A. Parameters are for a single-site binding model. "NA" indicates that there was no detectable binding. We floated the fraction competent parameter to capture uncertainty in peptide and protein concentration, particularly given the low extinction coefficients of S100A5 and S100A6. If an experiment was done at both 10 and 25 $°C$, the parameters correspond to the 10 $°C$ experiment. (*) Data from ancA5 binding to NCX1 were difficult to fit. The binding curves for this interaction had shallow curvature and did not appear to reach baseline saturation even with higher titrant/titrate molar ratio, leading to the high fraction competent.

| protein | $K_A$ $(M^{-1})$ | $\Delta H°$ $(kcal/mol)$ | $fx\ comp.$ | $num\ reps$ | $T$ $(°C)$ |
|---|---|---|---|---|---|
| ancA5/A6 | $3.3e4 \pm 7.6e3$ | $-1.70 \pm 0.3$ | $0.60 \pm 0.05$ | 2 | 25 |
| altAll | $2.3e4 \pm 8.3e3$ | $-3.80 \pm 1.2$ | $0.60 \pm 0.05$ | 2 | 25 |
| ancA5* | $1.98e5 \pm 1.7e5$ | $-0.68 \pm 0.4$ | $2.90 \pm 0.20$ | 2 | 10 |
| alliA5 | $5.80e3 \pm 1.6e3$ | $-7.20 \pm 2.0$ | $0.90 \pm 0.26$ | 2 | 10, 25 |
| sarA5 | $2.50e4 \pm 1.7e4$ | $-2.80 \pm 1.3$ | $0.70 \pm 0.20$ | 2 | 25 |
| mA5 | $1.20e5 \pm 1.7e5$ | $-1.30 \pm 0.5$ | $0.80 \pm 0.20$ | 2 | 25 |
| hA5 | $5.50e4 \pm 1.3e4$ | $-3.60 \pm 0.9$ | $1.40 \pm 0.10$ | 2 | 25 |
| ancA6 | NA | NA | NA | 2 | 10, 25 |
| alliA6 | $4.60e4 \pm 3.3e4$ | $-2.50 \pm 0.2$ | $0.70 \pm 0.15$ | 2 | 10, 25 |
| gA6 | $1.10e5 \pm 1.7e4$ | $3.40 \pm 0.6$ | $1.70 \pm 0.05$ | 2 | 25 |
| sarA6 | $1.30e4 \pm 5.8e3$ | $-4.30 \pm 1.8$ | $0.90 \pm 0.30$ | 2 | 25 |
| mA6 | NA | NA | NA | 2 | 25 |
| hA6 | NA | NA | NA | 2 | 25 |
| hA4 | NA | NA | NA | 2 | 25 |

TABLE 7 Thermodynamic parameters for binding of the peptide SEGLMNVLKKIYEDG (SIP) to S100A5 and S100A6 proteins.Species abbreviations are "alli" (alligator), "gal" (chicken), "sar" (tasmanian devil), "m" (mouse), and "h" (human). Fit parameters, with standard deviation from fits, for to the data shown schematically in Fig 4A. Parameters are for a single-site binding model. "NA" indicates that there was no detectable binding. We floated the fraction competent parameter to capture uncertainty in peptide and protein concentration, particularly given the low extinction coefficients of S100A5 and S100A6. If an experiment was done at both 10 and 25 $°C$, the parameters correspond to the 10 $°C$ experiment.

| protein | $K_A$ $(M^{-1})$ | $\Delta H°$ $(kcal/mol)$ | $fx\ comp.$ | $num\ reps$ | $T$ $(°C)$ |
|---|---|---|---|---|---|
| ancA5/A6 | $1.30e4 \pm 1.7e3$ | $-8.10 \pm 0.9$ | $1.50 \pm 0.01$ | 2 | 25 |
| altAll | $2.40e4 \pm 1.2e4$ | $-1.50 \pm 0.5$ | $1.30 \pm 0.20$ | 2 | 25 |
| ancA5 | NA | NA | NA | 2 | 25 |
| alliA5 | NA | NA | NA | 2 | 10, 25 |
| sarA5 | NA | NA | NA | 2 | 25 |
| mA5 | NA | NA | NA | 2 | 25 |
| hA5 | NA | NA | NA | 2 | 25 |
| ancA6 | $3.90e4 \pm 3.0e2$ | $3.80 \pm 0.3$ | $1.20 \pm 0.02$ | 2 | 10, 25 |
| alliA6 | $3.00e4 \pm 9.3e3$ | $3.20 \pm 0.7$ | $1.40 \pm 0.09$ | 2 | 25 |
| gA6 | $5.80e4 \pm 8.9e3$ | $4.00 \pm 0.4$ | $1.90 \pm 0.04$ | 2 | 25 |
| sarA6 | $3.30e5 \pm 1.5e5$ | $0.90 \pm 0.1$ | $2.00 \pm 0.02$ | 2 | 25 |
| mA6 | $3.90e5 \pm 2.8e5$ | $0.50 \pm 0.3$ | $1.80 \pm 0.02$ | 2 | 10, 25 |
| hA6 | $3.80e4 \pm 5.7e3$ | $2.90 \pm 0.3$ | $1.50 \pm 0.03$ | 2 | 15, 25 |
| hA4 | NA | NA | NA | 2 | 25 |

TABLE 8 Thermodynamic parameters for binding of the A5cons and SIP peptides to hA5 ancestral reversion mutants.Table entries show 95% credibility region from the posterior distribution of each parameter. Parameters are for a single-site binding model. "NA" parameters indicate that there was no detectable binding. All experiments were done at 25 $°C$.

| protein | peptide | $K_A$ $(\times 10^5\ M^{-1})$ | $\Delta H°$ $(kcal/mol)$ | $fx\ comp.$ |
|---|---|---|---|---|
| hA5 | A5cons | | | |
| hA5 | SIP | NA | NA | NA |
| hA5/E2a | A5cons | $1.2 \leq 1.3 \leq 1.4$ | $-5.2 \leq -5.1 \leq -4.88$ | $0.87 \leq 0.90 \leq 0.93$ |
| hA5/E2a | SIP | NA | NA | NA |
| L44i | A5cons | $1.6 \leq 1.7 \leq 1.9$ | $-5.2 \leq -5.1 \leq -4.88$ | $0.87 \leq 0.90 \leq 0.93$ |
| L44i | SIP | NA | NA | NA |
| D54k | A5cons | $3.2 \leq 3.5 \leq 3.7$ | $-5.1 \leq -5.0 \leq -4.9$ | $1.15 \leq 1.17 \leq 1.19$ |
| D54k | SIP | NA | NA | NA |
| M78a | A5cons | $1.0 \leq 1.1 \leq 1.2$ | $-3.5 \leq -3.3 \leq -3.1$ | $1.24 \leq 1.28 \leq 1.32$ |
| M78a | SIP | NA | NA | NA |
| A83m | A5cons | $2.6 \leq 2.8 \leq 3.0$ | $-5.5 \leq -5.5 \leq -5.3$ | $1.52 \leq 1.53 \leq 1.55$ |
| A83m | SIP | $0.4 \leq 0.6 \leq 1.0$ | $-0.8 \leq -0.6 \leq -0.4$ | $0.99 \leq 1.22 \leq 1.54$ |

TABLE 9 Accession numbers of S100 proteins used to build the multiple sequence alignment.

| paralog | accession | species |
|---|---|---|
| A1 | F1R758 | *Danio rerio* |
| A1 | A5WW32 | *Danio rerio* |
| A1 | H2TQM5 | *Takifugu rubripes* |
| A1 | H2ST19 | *Takifugu rubripes* |
| A1 | H2L492 | *Oryzias latipes* |
| A1 | H2M1B8 | *Oryzias latipes* |
| A1 | G3NKS0 | *Gasterosteus aculeatus* |
| A1 | G3PEI0 | *Gasterosteus aculeatus* |
| A2 | P29034 | *Homo sapiens* |
| A2 | F6Q7Q8 | *Ornithorhynchus anatinus* |
| A2 | P10462 | *Bos taurus* |
| A2 | G3W672 | *Sarcophilus harrisii* |
| A2 | JH205580.1 | *Pelodiscus sinensis* |
| A3 | P33764 | *Homo sapiens* |
| A3 | P62818 | *Mus musculus* |
| A3 | A4FUH7 | *Bos taurus* |
| A3 | G3W5T7 | *Sarcophilus harrisii* |
| A3 | F6SL13 | *Monodelphis domestica* |
| A3 | F6Q7S6 | *Ornithorhynchus anatinus* |
| A3 | JH205580.1 | *Pelodiscus sinensis* |
| A4 | P35466 | *Bos saurus* |
| A4 | predicted* | *Crocodylus porosus* |
| A4 | P26447 | *Homo sapiens* |
| A4 | H0Z1G5 | *Taeniopygia guttata* |
| A4 | P07091 | *Mus musculus* |
| A4 | F6SKU1 | *Monodelphis domestica* |
| A4 | F6Q7T6 | *Ornithorhynchus anatinus* |
| A4 | XP_015743713.1 | *Python bivittatus* |
| A4 | JH205580.1 | *Pelodiscus sinensis* |
| A4 | G3W5H2 | *Sarcophilus harrisii* |
| A4 | H9H0S2 | *Meleagris gallopavo* |
| A5 | P33763 | *Homo sapiens* |
| A5 | P63084 | *Mus musculus* |
| A5 | E1B8S0 | *Bos taurus* |
| A5 | G3W581 | *Sarcophilus harrisii* |
| A5 | XP_019412310.1 | *Crocodylus porosus* |
| A5 | JH205580.1 | *Pelodiscus sinensis* |
| A6 | P06703 | *Homo sapiens* |
| A6 | P14069 | *Mus musculus* |
| A6 | F6SKR4 | *Monodelphis domesitica* |
| A6 | F6R394 | *Ornithorhynchus anatinus* |
| A6 | G3W4S8 | *Sarcophilus harrisii* |
| A6 | H9H0S3 | *Meleagris gallopavo* |
| A6 | XP_019412316.1 | *Crocodylus porosus* |
| A6 | Q98953 | *Gallus gallus* |
| A6 | EOB07085.1 | *Anas platyrhynchos* |
| A6 | XP_015284753.1 | *Gekko japonicus* |
| A6 | XP_007429160.1 | *Python bivittatus* |
| A6 | JH205580.1 | *Pelodiscus sinensis* |

APPENDIX D

SUPPLEMENTAL MATERIAL FOR CHAPTER VI

## Supplemental Figures

This section includes the supplemental figures and tables referenced in chapter VI.

TABLE 10 Number of sequencing reads for each sample. Sample, and whether or not competitor was added, are indicated on the right. Columns show biological replicates 1 or 2. "total" columns indicate reads returned by the Illumina software pipeline. "good" columns indicate reads that passed our quality control and were used to calculate enrichment values.

| | | rep1 | | rep2 | |
|---|---|---|---|---|---|
| sample | competitor | total | good | total | good |
| hA5 | - | 24,794,016 | 19,695,958 | 29,085,203 | 16,773,567 |
| hA5 | + | 15,053,706 | 11,523,991 | 17,631,137 | 13,612,463 |
| hA6 | - | 22,728,393 | 17,722,779 | 7,769,003 | 5,972,295 |
| hA6 | + | 13,953,466 | 11,004,701 | 23,026,469 | 18,128,759 |
| ancA5/A6 | - | 23,690,810 | 18,387,038 | 14,534,333 | 11,034,524 |
| ancA5/A6 | + | 19,441,043 | 15,053,276 | 18,030,887 | 14,217,877 |
| altAll | - | 34,565,905 | 18,387,038 | 17,975,086 | 13,343,678 |
| altAll | + | 17,091,918 | 13,111,649 | 19,703,343 | 15,300,950 |
| raw library | | 39,700,991 | 32,190,368 | — | — |

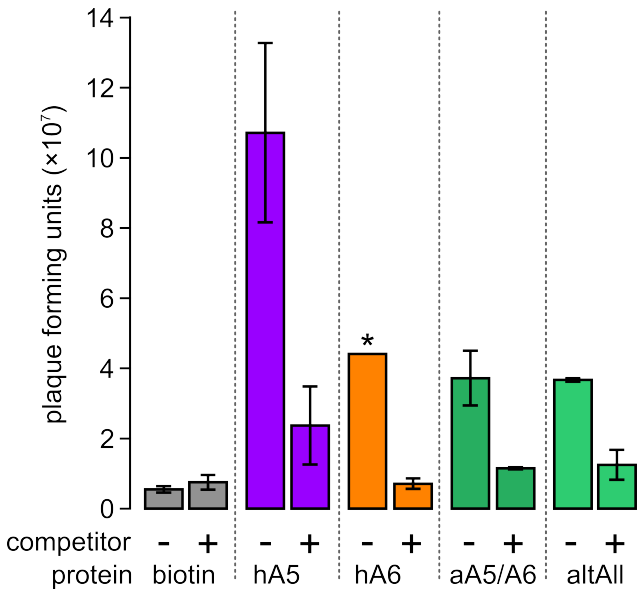


FIGURE 35 Phage enrichment is reduced in the presence of competitor peptide. Figure shows eluted plaque forming units (PFU) (estimated from phage titer) for two biological replicates of each condition. Enrichment is shown for biotin-only control (gray), hA5 (purple), hA6 (orange), ancA5/A6 (dark green), and ancA5/A6 atlAll (light green) with (+) and without (-) competitor peptide. Error bars show standard error for two biological replicates. (*) hA6 without competitor is shown for only one replicate due to failure of the titer for the other replicate.

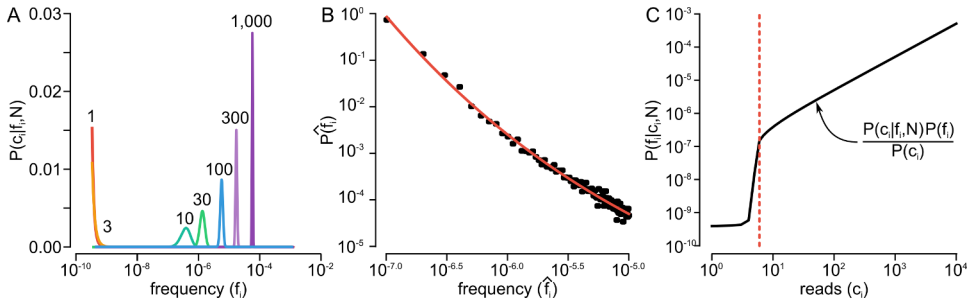


FIGURE 36 We can identify the number of counts that reliably reports on frequency in a sequenced phage pool. A) Using binomial sampling, we can calculate the probability of observing exactly $c_i$ counts in $N$ samples from that has a peptide of actual frequency $f_i$. Figure shows curves for counts ranging from 1 (red) to 1,000 (pink), all using $N = 2.0 \times 10^7$. B) Panel shows a histogram of frequencies estimated from $3.9 \times 10^7$ reads taken from the input library. The black points are experimental data. The red curve is an exponential distribution fit to that curve. C) Using the sampling from panel A and the fit curve from panel B, we can determine $P(f_i|c_i, N)$. The solid curve shows the relationship between the number of reads for peptide $i$ (x-axis) against the maximum-likelihood estimate of the frequency (y-axis). The red line highlights the cutoff we used in our experiments.

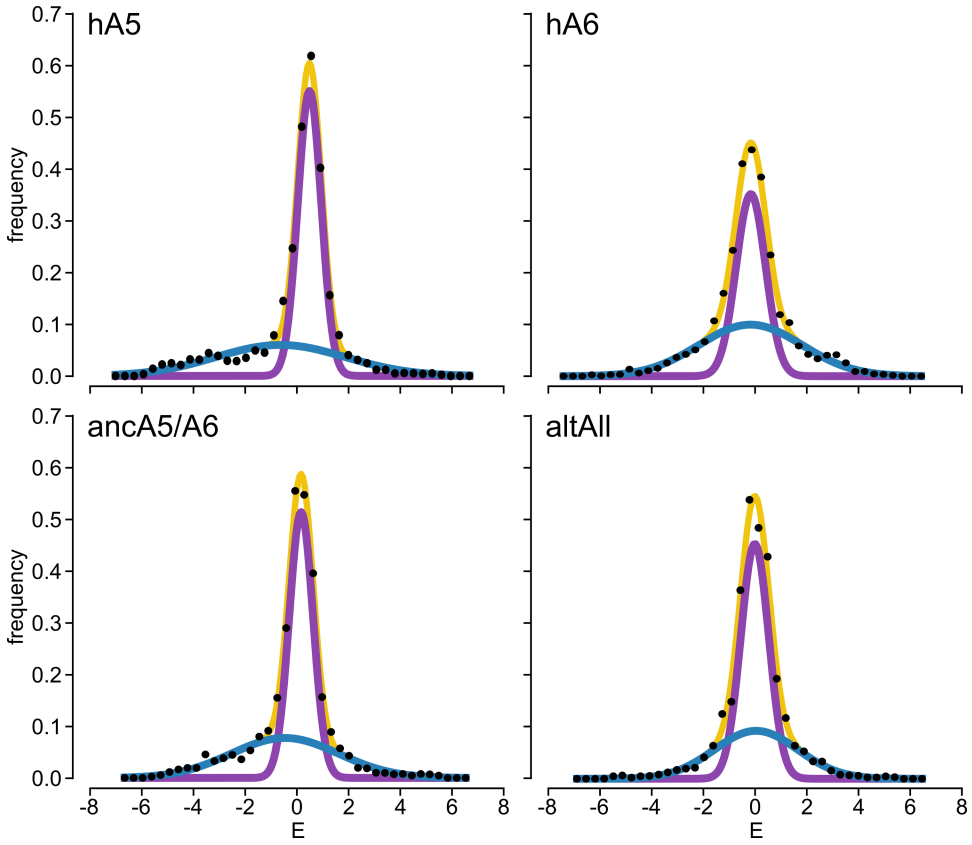


FIGURE 37 Enrichment distributions for all proteins. Panels show distribution of $E$ for each protein (pooled bio-replicates). Points are raw histograms. Curves are two Gaussian fit: blue (responsive), purple (unresponsive) and yellow (sum).

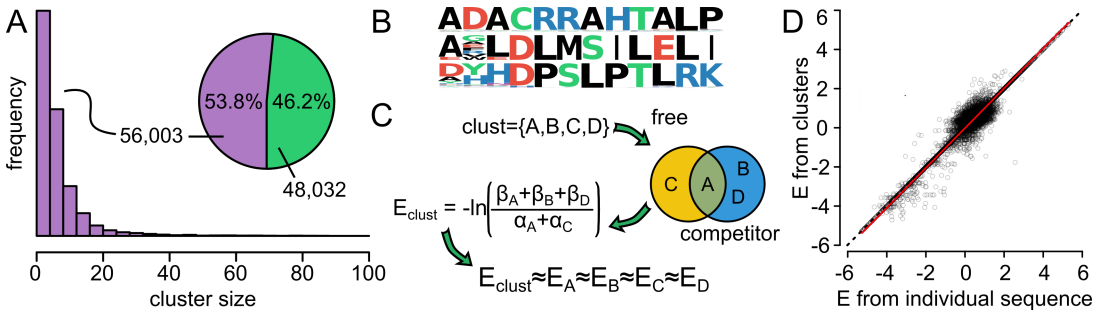


FIGURE 38 We can estimate how addition of competitor peptide alters the frequencies of peptides. A) Distribution of sizes of peptide clusters from hA5 experiment. Pie chart shows number of peptides placed in clusters (56,003; 53.8%) versus not (48,032; 46.2%). B) Three example clusters taken from the clusters in panel A. The letter height at each position indicates its frequency in the sequences within that cluster. C) Toy example showing how enrichment is calculated for a cluster containing peptides $\{A, B, C, D\}$. Peptides $A$ and $C$ were observed in the no competitor sample at frequencies $\alpha_A$ and $\alpha_C$. Peptides $A$, $B$, and $D$ were observed in the competitor sample at frequencies $\beta_A$, $\beta_B$ and $\beta_D$. The enrichment of the cluster is given by $E_{clust} = -ln[(\beta_A + \beta_B + \beta_D)/(\alpha_A + \alpha_C)]$. All members of the cluster are then assigned E $\approx E_{clust}$. D) Comparison of enrichment values for hA5 peptides determined using a direct comparison of frequencies with and without competitor (x-axis) versus the clustering method (y-axis). Each point is an individual peptide. Red line is a least-squares regression line fit to the data. The dashed line is the 1:1 line.

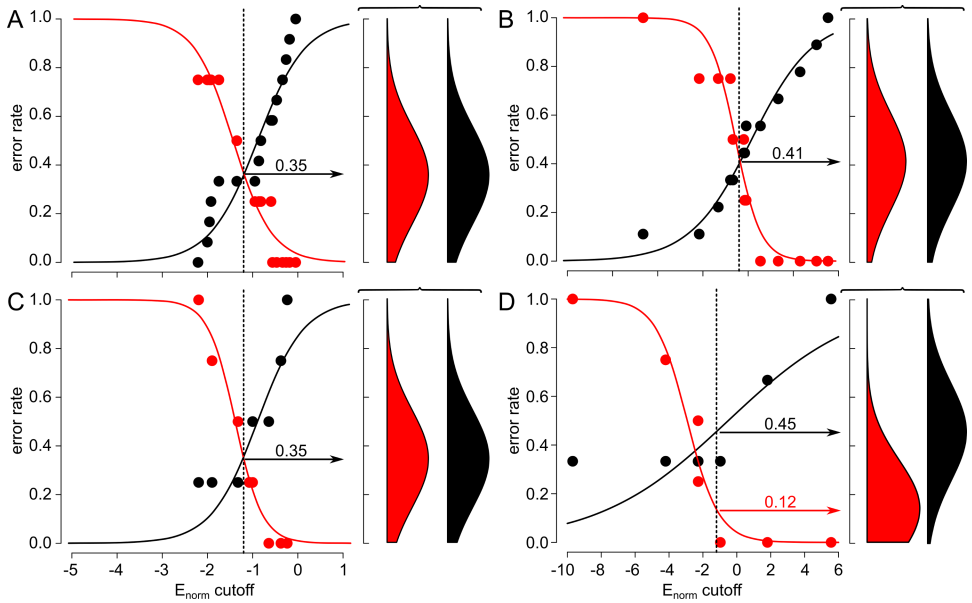


FIGURE 39 Estimating the error rates for individual models. Panels show individual models: A) hA5, B) hA6, C) ancA5/A6, and D) altAll. For each panel, the left graph shows the error rate for peptide binding as a function of the cutoff in $E_{norm}$ chosen for classification. Lines are fits of the modified Hill equation to the error rates. Colors indicate the false negative rate (red) and false positive rate (black). The dashed vertical line indicates the cutoff used for prediction of the Venn diagrams in Fig 6 ($E_{norm}$ = -1.19). The error rates associated with $E_{norm}$ = -1.19 are indicated with arrows pointing right. The distributions in each panel show the prior distributions used for the false negative (red) and false positive (black) error rates in the Bayesian estimator. These distributions are centered at the error rate estimate, with standard deviations of 0.2.

TABLE 11 Features used in for supervised machine learning. Features denoted (CIDER) were calculated using the CIDER software package [315]. Other features were calculated using our own software package (HOPS: https://github.com/harmslab/hops).

| feature | ref |
|---|---|
| num. hbond acceptors | — |
| num. hbond donors | — |
| $\kappa$ (CIDER) | [315] |
| $\Delta$ (CIDER) | [315] |
| $\Omega$ (CIDER) | [315] |
| FER (CIDER) | [315] |
| $\Sigma$ (CIDER) | [315] |
| dmax (CIDER) | [315] |
| $\Delta$max | [315] |
| NCPR | [315] |
| $F_+$(CIDER) | [315] |
| $F_-$(CIDER) | [315] |
| FCR (CIDER) | [315] |
| mean hydropathy (CIDER) | [315] |
| White Interface scale | [333] |
| Engleman scale | [334] |
| % buried in structures | [335] |
| Kyte/Doolittle scale | [336] |
| Octanol scale | [337] |
| Hopp-Woods scale | [338] |
| Uversky scale | [315] |
| cumulative mean hydropathy | [315] |
| side chain accessible area | [304] |
| main chain accessible area | [304] |
| Chou-Fasman, $\beta$ | [339] |
| Chou-Fasman, $\alpha$ | [339] |
| Chou-Fasman, turn | [339] |
| fraction poly-proline II | [315] |
| predicted charge at pH 4 | — |
| predicted charge at pH 5 | — |
| predicted charge at pH 6 | — |
| predicted charge at pH 7 | — |
| predicted charge at pH 8 | — |
| predicted charge at pH 9 | — |
| num. positive amino acids | — |
| num. neutral amino acids | — |
| num. negative amino acids | — |
| net charge | — |
| isoelectric point | — |
| knob main chain, b | [340] |
| socket main chain, x | [340] |
| socket main chain, y | [340] |
| socket main chain, h | [340] |
| knob side chain, b | [340] |
| socket side chain, x | [340] |
| socket side chain, y | [340] |
| socket side chain, h | [340] |
| side chain volume | [341] |
| molecular weight | — |
| aromatic | — |

TABLE 12 Predicted $E$ and measured binding constants for peptides. Columns indicate calculated $E$ and measured $K_D$ for the peptides indicated on the left. For $K_D$, an entry of "> 100" indicates that we performed an ITC experiment, but that no binding was detectable better than ≈100 $\mu M$. An entry of "—" indicates no experiment was performed.

| | | hA5 | | hA6 | | aA5A6 | | altAll | |
|---|---|---|---|---|---|---|---|---|---|
| name | sequence | $E$ | $K_D$ | $E$ | $K_D$ | $E$ | $K_D$ | $E$ | $K_D$ |
| Q86UW7 | AGSSQRAPPAPTREGRRD | -4.03 | >100 | 0.26 | — | -0.77 | — | 0.41 | — |
| An1 | AMVSEFLKQAWFIE | -1.71 | >100 | -1.68 | 13 | -1.73 | — | -0.37 | — |
| O75170 | DAPGAGAPPAPGKKEAPP | -3.94 | >100 | 0.50 | >100 | -0.74 | — | 0.78 | >100 |
| p3 | DWSSWVYRDTQTGGSAE | -1.26 | >100 | -1.10 | >100 | -1.28 | — | -0.11 | — |
| p1 | EPSPVSMNEGTFGGSAE | -0.27 | — | -0.54 | 10 | -0.34 | >100 | -0.45 | — |
| Q13424 | GAGGERWQRVLLSLAEDT | -4.45 | 3 | -1.11 | — | -1.98 | — | -0.10 | — |
| B2RNZ0 | KEIKTAMWRLFVKIYFLQK | -3.53 | >100 | -2.72 | >100 | -2.38 | >100 | -1.36 | >100 |
| p6 | QPELTQGRVGINGGGSAE | -1.02 | >100 | 0.30 | — | -0.79 | — | -0.08 | — |
| NCX1 | RRLLFYKYVYKR | -1.20 | 18 | -1.33 | >100 | -1.40 | 30 | -0.59 | 47 |
| A6cons | RSHSGFDWRWGMEALTGGGSAE | -1.97 | 3 | -0.84 | 5 | -1.12 | 1 | -0.14 | 9 |
| A5cons | RSHSSSFQDWLLSRLPGGGSAE | -2.75 | 3 | -0.81 | >100 | -2.05 | 11 | -0.32 | 24 |
| Q14147 | SEDDRAGPAPPGASDGVD | -3.88 | >100 | 0.71 | >100 | -1.20 | — | 0.47 | — |
| SIP | SEGLMNVLKKIYEDG | -0.60 | >100 | -1.05 | 26 | -0.64 | 77 | -0.32 | 42 |
| p4 | SIGASELHVYRSGGSAE | -0.76 | >100 | -0.21 | >100 | -0.18 | >100 | -0.40 | — |
| p7 | STTVRNGESPNCGGSAE | -0.45 | >100 | -0.84 | — | -0.24 | — | 0.16 | — |
| p5 | STVHEILSKLSEGY | -0.19 | >100 | -0.85 | >100 | -0.13 | — | 0.07 | — |
| p2 | TAKYLPMRPGPLGGGSAE | -1.79 | >100 | 0.20 | >100 | -1.04 | >100 | 0.25 | >100 |

REFERENCES CITED


[1] Robert A. Zierenberg, Michael W. W. Adams, and Alissa J. Arp. Life in extreme environments: Hydrothermal vents. *Proceedings of the National Academy of Sciences*, 97(24):12961–12962, November 2000. ISSN 0027-8424, 1091-6490.

[2] L. C. Bliss. Adaptations of Arctic and Alpine Plants to Environmental Conditions. *Arctic*, 15(2):117–144, 1962. ISSN 0004-0843.

[3] Kyle Summers and Mark E. Clough. The evolution of coloration and toxicity in the poison frog family (Dendrobatidae). *Proceedings of the National Academy of Sciences of the United States of America*, 98(11):6227–6232, May 2001. ISSN 0027-8424.

[4] Charles Darwin. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life /*, volume -1859. London :John Murray,, 1859.

[5] Madeline C. Weiss, Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi, and William F. Martin. The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, 1(9): nmicrobiol2016116, July 2016. ISSN 2058-5276.

[6] Nicolas Glansdorff, Ying Xu, and Bernard Labedan. The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biology Direct*, 3:29, July 2008. ISSN 1745-6150.

[7] Nick Lane, John F. Allen, and William Martin. How did LUCA make a living? Chemiosmosis in the origin of life. *BioEssays*, 32(4):271–280, April 2010. ISSN 1521-1878.

[8] Richard C. Lewontin. *The genetic basis of evolutionary change [by] R. C. Lewontin*. Columbia University Press New York, 1974. ISBN 0-231-03392-3 0-231-08318-1.

[9] F. Jacob. Evolution and tinkering. *Science*, 196(4295):1161–1166, June 1977. ISSN 0036-8075, 1095-9203.

[10] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, June 1990. ISSN 0027-8424, 1091-6490.

[11] Masatoshi Nei and Jianzhi Zhang. Molecular Origin of Species. *Science*, 282 (5393):1428–1429, November 1998. ISSN 0036-8075, 1095-9203.

[12] J.H. Gillespie. *Population Genetics: A Concise Guide*. A Johns Hopkins Paperback: Science. Johns Hopkins University Press, 1998. ISBN 978-0-8018-5754-6.

[13] John H. Gillespie. Molecular Evolution Over the Mutational Landscape. *Evolution*, 38(5):1116–1129, 1984. ISSN 0014-3820.

[14] Sir Fisher, Ronald Aylmer. *The genetical theory of natural selection.* OxfordClarendon Press.

[15] Sewall Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, March 1931. ISSN 0016-6731.

[16] Walter Fontana and Peter Schuster. Shaping Space: the Possible and the Attainable in RNA Genotypephenotype Mapping. *Journal of Theoretical Biology*, 194(4):491–515, October 1998. ISSN 0022-5193.

[17] Peter F. Stadler and Brbel M. R. Stadler. Genotype-Phenotype Maps. *Biological Theory*, 1(3):268–279, September 2006. ISSN 1555-5542, 1555-5550.

[18] John H. Gillespie. A simple stochastic gene substitution model. *Theoretical Population Biology*, 23(2):202–215, April 1983. ISSN 0040-5809.

[19] H. Allen Orr. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution; International Journal of Organic Evolution*, 56(7): 1317–1330, July 2002. ISSN 0014-3820.

[20] Vanda T. K. McNIVEN, Hlne LeVASSEUR-VIENS, Rachelle L. Kanippayoor, Meghan Laturney, and Amanda J. Moehring. The genetic basis of evolution, adaptation and speciation. *Molecular Ecology*, 20(24):5119–5122, December 2011. ISSN 1365-294X.

[21] R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Mst, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Vinl, J. B. W. Wolf, and D. Zinner. Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2):229–246, February 2013. ISSN 1420-9101.

[22] S. Blair Hedges, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar. Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, 32(4):835–845, April 2015. ISSN 0737-4038, 1537-1719.

[23] Scott P. Egan, Gregory J. Ragland, Lauren Assour, Thomas H.Q. Powell, Glen R. Hood, Scott Emrich, Patrik Nosil, and Jeffrey L. Feder. Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecology Letters*, 18(8):817–825, August 2015. ISSN 1461-0248.

[24] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Protein Function. 2002.

[25] D. Whitford. *Proteins: Structure and Function*. Wiley, 2005. ISBN 978-0-470-01241-3.

[26] Motoo Kimura and James F. Crow. The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics*, 49(4):725–738, April 1964. ISSN 0016-6731, 1943-2631.

[27] James F. Crow. The Mathematics of Heredity. Gustave Malcot. Translated from the French edition (Paris, 1948), revised, and edited by Demetrios M. Yermanos. Freeman, San Francisco, 1969. xx + 92 pp., illus. $4. *Science*, 168 (3932):721–721, May 1970. ISSN 0036-8075, 1095-9203.

[28] M. Nei. Relative Roles of Mutation and Selection in the Maintenance of Genetic Variability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 319(1196):615–629, 1988. ISSN 0080-4622.

[29] H. Allen Orr. The Probability of Parallel Evolution. *Evolution*, 59(1):216–220, January 2005. ISSN 1558-5646.

[30] Paul A. Hohenlohe, Patrick C. Phillips, and William A. Cresko. USING POPULATION GENOMICS TO DETECT SELECTION IN NATURAL POPULATIONS: KEY CONCEPTS AND METHODOLOGICAL CONSIDERATIONS. *International Journal of Plant Sciences*, 171(9): 1059–1071, November 2010. ISSN 1058-5893.

[31] J. Romiguier, P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, E. Loire, J. M. Lourenco, B. Nabholz, C. Roux, G. Tsagkogeorga, A. a.-T. Weber, L. A. Weinert, K. Belkhir, N. Bierne, S. Glmin, and N. Galtier. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263, November 2014. ISSN 0028-0836.

[32] Trudy F. C. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1):22–33, January 2014. ISSN 1471-0056.

[33] Peter Tiffin and Jeffrey Ross-Ibarra. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, 29 (12):673–680, December 2014. ISSN 0169-5347.

[34] Pu Huang, Maximilian Feldman, Stephan Schroder, Bochra A. Bahri, Xianmin Diao, Hui Zhi, Matt Estep, Ivan Baxter, Katrien M. Devos, and Elizabeth A. Kellogg. Population genetics of Setariaviridis, a new model system. *Molecular Ecology*, 23(20):4912–4925, October 2014. ISSN 1365-294X.

[35] Michael Lynch. Mutation and Human Exceptionalism: Our Future Genetic Load. *Genetics*, 202(3):869–875, March 2016. ISSN 0016-6731, 1943-2631.

[36] H.C. Berg. *Random Walks in Biology*. Princeton paperbacks. Princeton University Press, 1993. ISBN 978-0-691-00064-0.

[37] Guy Sella and Aaron E. Hirsh. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9541–9546, July 2005. ISSN 0027-8424, 1091-6490.

[38] K. Dill and S. Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. Taylor & Francis Group, 2010. ISBN 978-1-136-67299-6.

[39] Shigeru Kondo and Takashi Miura. Reaction-Diffusion Model as a Framework for Understanding Biological Pattern Formation. *Science*, 329(5999): 1616–1620, September 2010. ISSN 0036-8075, 1095-9203.

[40] Ken A. Dill, Kingshuk Ghosh, and Jeremy D. Schmit. Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences*, 108(44): 17876–17882, November 2011. ISSN 0027-8424, 1091-6490.

[41] Kingshuk Ghosh, Adam M. R. de Graff, Lucas Sawle, and Ken A. Dill. Role of Proteome Physical Chemistry in Cell Behavior. *The Journal of Physical Chemistry. B*, 120(36):9549, September 2016.

[42] R.E. Feeney and R.G. Allison. *Evolutionary biochemistry of proteins: homologous and analogous proteins from avian egg whites, blood sera, milk, and other substances*. Wiley-Interscience, 1969.

[43] Michael J. Harms and Joseph W. Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14(8):559–571, August 2013. ISSN 1471-0056.

[44] Michael J Harms and Joseph W Thornton. Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology*, 20(3):360–366, June 2010. ISSN 0959-440X.

[45] Linus Pauling and E. Zuckerkandl. Chemical Paleogenetics. *Acta chem. scand*, 17:S9–S16, 1963.

[46] Michael J. Harms and Joseph W. Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513): 203–207, August 2014. ISSN 0028-0836.

[47] Lucas C Wheeler, Shion A Lim, Susan Marqusee, and Michael J Harms. The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology*, 38:37–43, June 2016. ISSN 0959-440X.

[48] D.A. Liberles. *Ancestral Sequence Reconstruction*. Oxford biosciences. OUP Oxford, 2007. ISBN 978-0-19-929918-8.

[49] Victor Hanson-Smith, Bryan Kolaczkowski, and Joseph W. Thornton. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Molecular Biology and Evolution*, 27(9):1988–1999, September 2010. ISSN 0737-4038.

[50] Geeta N. Eick, Jamie T. Bridgham, Douglas P. Anderson, Michael J. Harms, and Joseph W. Thornton. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Molecular Biology and Evolution*, 34(2): 247–261, February 2017. ISSN 0737-4038.

[51] Jamie T. Bridgham, Eric A. Ortlund, and Joseph W. Thornton. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature*, 461(7263):515–519, September 2009. ISSN 0028-0836.

[52] AlesiaN. McKeown, JamieT. Bridgham, DaveW. Anderson, MichaelN. Murphy, EricA. Ortlund, and JosephW. Thornton. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell*, 159(1):58–68, September 2014. ISSN 0092-8674.

[53] Jeffrey I. Boucher, Joseph R. Jacobowitz, Brian C. Beckett, Scott Classen, and Douglas L. Theobald. An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *eLife*, 3:e02304, June 2014. ISSN 2050-084X.

[54] Dave W. Anderson, Alesia N. McKeown, and Joseph W. Thornton. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife*, 4:e07864, July 2015. ISSN 2050-084X.

[55] BenE. Clifton and ColinJ. Jackson. Ancestral Protein Reconstruction Yields Insights into Adaptive Evolution of Binding Specificity in Solute-Binding Proteins. *Cell Chemical Biology*, 23(2):236–245, February 2016. ISSN 2451-9456.

[56] Kathryn M. Hart, Michael J. Harms, Bryan H. Schmidt, Carolyn Elya, Joseph W. Thornton, and Susan Marqusee. Thermodynamic System Drift in Protein Evolution. *PLoS Biol*, 12(11):e1001994, November 2014.

[57] Christopher D. Aakre, Julien Herrou, Tuyen N. Phung, Barrett S. Perchuk, Sean Crosson, and Michael T. Laub. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell*, 163(3): 594–606, October 2015. ISSN 0092-8674.

[58] Geeta N. Eick, Jennifer K. Colucci, Michael J. Harms, Eric A. Ortlund, and Joseph W. Thornton. Evolution of Minimal Specificity and Promiscuity in Steroid Hormone Receptors. *PLoS Genetics*, 8(11), November 2012. ISSN 1553-7390.

[59] C. Wilson, R. V. Agafonov, M. Hoemberger, S. Kutter, A. Zorba, J. Halpin, V. Buosi, R. Otten, D. Waterman, D. L. Theobald, and D. Kern. Using ancient protein kinases to unravel a modern cancer drugs mechanism. *Science*, 347(6224):882–886, February 2015. ISSN 0036-8075, 1095-9203.

[60] Valeria A. Risso, Jose A. Gavira, Diego F. Mejia-Carmona, Eric A. Gaucher, and Jose M. Sanchez-Ruiz. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian -Lactamases. *Journal of the American Chemical Society*, 135(8):2899–2902, February 2013. ISSN 0002-7863.

[61] Tyler N. Starr and Joseph W. Thornton. Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218, July 2016. ISSN 1469-896X.

[62] Zachary R. Sailer and Michael J. Harms. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics*, 205(3):1079–1088, March 2017. ISSN 0016-6731, 1943-2631.

[63] Jason B. Wolf, Daniel Pomp, Eugene J. Eisen, James M. Cheverud, and Larry J. Leamy. The contribution of epistatic pleiotropy to the genetic architecture of covariation among polygenic traits in mice. *Evolution & Development*, 8(5):468–476, October 2006. ISSN 1520-541X.

[64] Gnter P. Wagner and Vincent J. Lynch. The gene regulatory logic of transcription factor evolution. *Trends in Ecology & Evolution*, 23(7):377–385, July 2008. ISSN 0169-5347.

[65] Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences*, 105(23):7899–7906, June 2008. ISSN 0027-8424, 1091-6490.

[66] David L. Des Marais and Mark D. Rausher. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 454(7205):762–765, August 2008. ISSN 0028-0836.

[67] Stacey D. Smith and Mark D. Rausher. Gene loss and parallel evolution contribute to species difference in flower color. *Molecular Biology and Evolution*, 28(10):2799–2810, October 2011. ISSN 1537-1719.

[68] Stacey D. Smith, Shunqi Wang, and Mark D. Rausher. Functional evolution of an anthocyanin pathway enzyme during a flower color transition. *Molecular Biology and Evolution*, 30(3):602–612, March 2013. ISSN 1537-1719.

[69] Trevor R. Sorrells, Lauren N. Booth, Brian B. Tuch, and Alexander D. Johnson. Intersecting transcription networks constrain gene regulatory evolution. *Nature*, 523(7560):361–365, July 2015. ISSN 0028-0836.

[70] Matthew A. Streisfeld and Mark D. Rausher. Altered trans-regulatory control of gene expression in multiple anthocyanin genes contributes to adaptive flower color evolution in Mimulus aurantiacus. *Molecular Biology and Evolution*, 26(2):433–444, February 2009. ISSN 1537-1719.

[71] Matthew A. Streisfeld and Mark D. Rausher. Genetic changes contributing to the parallel evolution of red floral pigmentation among Ipomoea species. *The New Phytologist*, 183(3):751–763, August 2009. ISSN 1469-8137.

[72] Carolyn A. Wessinger and Mark D. Rausher. Lessons from flower colour evolution on targets of selection. *Journal of Experimental Botany*, 63(16): 5741–5749, October 2012. ISSN 0022-0957.

[73] Carolyn A. Wessinger and Mark D. Rausher. Predictability and irreversibility of genetic changes associated with flower color evolution in Penstemon barbatus. *Evolution; International Journal of Organic Evolution*, 68(4):1058–1070, April 2014. ISSN 1558-5646.

[74] Ali Zarrinpar, Sang-Hyun Park, and Wendell A. Lim. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967):676–680, December 2003. ISSN 0028-0836.

[75] Daniel M. Weinreich, Nigel F. Delaney, Mark A. DePristo, and Daniel L. Hartl. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, 312(5770):111–114, April 2006. ISSN 0036-8075, 1095-9203.

[76] Shelley D. Copley. Toward a Systems Biology Perspective on Enzyme Evolution. *The Journal of Biological Chemistry*, 287(1):3–10, January 2012. ISSN 0021-9258.

[77] Aaron W. Reinke, Jiyeon Baek, Orr Ashenberg, and Amy E. Keating. Networks of bZIP Protein-Protein Interactions Diversified Over a Billion Years of Evolution. *Science*, 340(6133):730–734, May 2013. ISSN 0036-8075, 1095-9203.

[78] William H. Hudson and Eric A. Ortlund. The structure, function and evolution of proteins that bind DNA and RNA. *Nature Reviews Molecular Cell Biology*, 15(11):749–760, November 2014. ISSN 1471-0072.

[79] Conor J. Howard, Victor Hanson-Smith, Kristopher J. Kennedy, Chad J. Miller, Hua Jane Lou, Alexander D. Johnson, Benjamin E. Turk, and Liam J. Holt. Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. *eLife*, 3:e04126, November 2014. ISSN 2050-084X.

[80] Steven M Yannone, Sophia Hartung, Angeli L Menon, Michael WW Adams, and John A Tainer. Metals in biology: defining metalloproteomes. *Current Opinion in Biotechnology*, 23(1):89–95, February 2012. ISSN 0958-1669.

[81] Diana Ekman, Sara Light, sa K Bjrklund, and Arne Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biology*, 7(6):R45, 2006. ISSN 1465-6906.

[82] Nobuyuki Uchikoga, Yuri Matsuzaki, Masahito Ohue, and Yutaka Akiyama. Specificity of broad protein interaction surfaces for proteins with multiple binding partners. *Biophysics and Physicobiology*, 13:105–115, July 2016. ISSN 2189-4779.

[83] Shibani Bhattacharya, Christopher G. Bunick, and Walter J. Chazin. Target selectivity in EF-hand calcium binding proteins. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1742(13):69–79, December 2004. ISSN 0167-4889.

[84] David Chin and Anthony R Means. Calmodulin: a prototypical calcium sensor. *Trends in Cell Biology*, 10(8):322–328, August 2000. ISSN 0962-8924.

[85] Patrick S Mitchell, Michael Emerman, and Harmit S Malik. An evolutionary perspective on the broad antiviral specificity of MxA. *Current Opinion in Microbiology*, 16(4):493–499, August 2013. ISSN 1369-5274.

[86] D. Gfeller, F. Butty, M. Wierzbicka, E. Verschueren, P. Vanhee, H. Huang, A. Ernst, N. Dar, I. Stagljar, L. Serrano, S. S. Sidhu, G. D. Bader, and P. M. Kim. The multiple-specificity landscape of modular peptide recognition domains. *Molecular Systems Biology*, 7(1):484–484, April 2014. ISSN 1744-4292.

[87] Gideon Schreiber and Amy E Keating. Protein binding specificity versus promiscuity. *Current Opinion in Structural Biology*, 21(1):50–61, February 2011. ISSN 0959-440X.

[88] Shelley D. Copley. An evolutionary biochemist's perspective on promiscuity. *Trends in Biochemical Sciences*, 40(2):72–78, January 2015. ISSN 0968-0004.

[89] R. Donato, B.R. Cannon, G. Sorci, F. Riuzzi, K. Hsu, D.J. Weber, and C.L. Geczy. Functions of S100 Proteins. *Current molecular medicine*, 13(1):24–57, January 2013. ISSN 1566-5240.

[90] Rosario Donato. Intracellular and extracellular roles of S100 proteins. *Microscopy Research and Technique*, 60(6):540–551, April 2003. ISSN 1097-0029.

[91] Danna B. Zimmer, Jeannine O. Eubanks, Dhivya Ramakrishnan, and Michael F. Criscitiello. Evolution of the S100 family of calcium sensor proteins. *Cell Calcium*, 53(3):170–179, March 2013. ISSN 0143-4160.

[92] Claus W. Heizmann and Jos A. Cox. New perspectives on S100 proteins: a multi-functional Ca 2+ -, Zn 2+ - and Cu 2+ -binding protein family. *Biometals*, 11(4):383–397. ISSN 0966-0844, 1572-8773.

[93] Walter J. Chazin. Relating Form and Function of EF-hand Calcium Binding Proteins. *Accounts of chemical research*, 44(3):171–179, March 2011. ISSN 0001-4842.

[94] Liliana Santamaria-Kisiel, Anne C. Rintala-Dempsey, and Gary S. Shaw. Calcium-dependent and -independent interactions of the S100 protein family. *Biochemical Journal*, 396(2):201–214, June 2006. ISSN 0264-6021, 1470-8728.

[95] Andreas M. Kraemer, Luis R. Saraiva, and Sigrun I. Korsching. Structural and functional diversification in the teleost S100 family of calcium-binding proteins. *BMC Evolutionary Biology*, 8:48, 2008. ISSN 1471-2148.

[96] Lucas C. Wheeler, Micah T. Donor, James S. Prell, and Michael J. Harms. Multiple Evolutionary Origins of Ubiquitous Cu2+ and Zn2+ Binding in the S100 Protein Family. *PLOS ONE*, 11(10):e0164740, October 2016. ISSN 1932-6203.

[97] Kenji Kizawa, Hidenari Takahara, Masaki Unno, and Claus W. Heizmann. S100 and S100 fused-type protein families in epidermal maturation with special focus on S100a3 in mammalian hair cuticles. *Biochimie*, 93(12):2038–2047, December 2011. ISSN 1638-6183.

[98] Michael F. Gutknecht, Marc E. Seaman, Bo Ning, Daniel Auger Cornejo, Emily Mugler, Patrick F. Antkowiak, Christopher A. Moskaluk, Song Hu, Frederick H. Epstein, and Kimberly A. Kelly. Identification of the S100 fused-type protein hornerin as a regulator of tumor vascularity. *Nature Communications*, 8(1):552, September 2017. ISSN 2041-1723.

[99] Romuald Contzler, Bertrand Favre, Marcel Huber, and Daniel Hohl. Cornulin, a new member of the "fused gene" family, is expressed during epidermal differentiation. *The Journal of Investigative Dermatology*, 124(5):990–997, May 2005. ISSN 0022-202X.

[100] Ingo Marenholz, Claus W. Heizmann, and Gnter Fritz. S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochemical and Biophysical Research Communications*, 322(4):1111–1122, October 2004. ISSN 0006-291X.

[101] Estelle Leclerc, Gnter Fritz, Stefan W. Vetter, and Claus W. Heizmann. Binding of S100 proteins to RAGE: An update. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1793(6):993–1007, June 2009. ISSN 0167-4889.

[102] Francesca Riuzzi, Guglielmo Sorci, and Rosario Donato. S100b protein regulates myoblast proliferation and differentiation by activating FGFR1 in a bFGF-dependent manner. *J Cell Sci*, 124(14):2389–2400, July 2011. ISSN 0021-9533, 1477-9137.

[103] Weidong Zhu, Yi Xue, Chao Liang, Rihua Zhang, Zhihong Zhang, Hongyan Li, Dongming Su, Xiubin Liang, Yuanyuan Zhang, Qiong Huang, Menglan Liu, Lu Li, Dong Li, Allan Z. Zhao, and Yun Liu. S100a16 promotes cell proliferation and metastasis via AKT and ERK cell signaling pathways in human prostate cancer. *Tumor Biology*, 37(9):12241–12250, September 2016. ISSN 1010-4283, 1423-0380.

[104] Ching Chang Cho, Ruey Hwang Chou, and Chin Yu. Pentamidine blocks the interaction between mutant S100a5 and RAGE V domain and inhibits the RAGE signaling pathway. *Biochemical and Biophysical Research Communications*, 477(2):188–194, August 2016. ISSN 0006-291X.

[105] Steven M. Damo, Thomas E. Kehl-Fie, Norie Sugitani, Marilyn E. Holt, Subodh Rathi, Wesley J. Murphy, Yaofang Zhang, Christine Betz, Laura Hench, Gnter Fritz, Eric P. Skaar, and Walter J. Chazin. Molecular basis for manganese sequestration by calprotectin and roles in the innate immune response to invading bacterial pathogens. *Proceedings of the National Academy of Sciences*, 110(10):3841–3846, March 2013. ISSN 0027-8424, 1091-6490.

[106] Joshua A. Hayden, Megan Brunjes Brophy, Lisa S. Cunden, and Elizabeth M. Nolan. High-Affinity Manganese Coordination by Human Calprotectin Is Calcium-Dependent and Requires the Histidine-Rich Site Formed at the Dimer Interface. *Journal of the American Chemical Society*, 135(2):775–787, January 2013. ISSN 0002-7863.

[107] James N. Tsoporis, Alexander Marks, Abraham Haddad, Fayez Dawood, Peter P. Liu, and Thomas G. Parker. S100b Expression Modulates Left Ventricular Remodeling After Myocardial Infarction in Mice. *Circulation*, 111 (5):598–606, February 2005. ISSN 0009-7322, 1524-4539.

[108] James N. Tsoporis, Shehla Izhar, and Thomas G. Parker. Expression of S100a6 in Cardiac Myocytes Limits Apoptosis Induced by Tumor Necrosis Factor-. *Journal of Biological Chemistry*, 283(44):30174–30183, October 2008. ISSN 0021-9258, 1083-351X.

[109] Lucas N. Wafer, Franco O. Tzul, Pranav P. Pandharipande, and George I. Makhatadze. Novel Interactions of the TRTK12 Peptide with S100 Protein Family Members: Specificity and Thermodynamic Characterization. *Biochemistry*, 52(34):5844–5856, August 2013. ISSN 0006-2960.

[110] Lucas C. Wheeler, Jeremy A. Anderson, Annelise J. Morrison, Caitlyn E. Wong, and Michael J. Harms. Conservation of specificity in two low-specificity proteins. *bioRxiv*, page 207324, October 2017.

[111] Michael A. Stiffler, Jiunn R. Chen, Viara P. Grantcharova, Ying Lei, Daniel Fuchs, John E. Allen, Lioudmila A. Zaslavskaia, and Gavin MacBeath. PDZ Domain Binding Selectivity Is Optimized Across the Mouse Proteome. *Science*, 317(5836):364–369, July 2007. ISSN 0036-8075, 1095-9203.

[112] Eric A. Gaucher, Sridhar Govindarajan, and Omjoy K. Ganesh. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, 451(7179):704–707, February 2008. ISSN 0028-0836.

[113] Karin Voordeckers, Chris A. Brown, Kevin Vanneste, Elisa van der Zande, Arnout Voet, Steven Maere, and Kevin J. Verstrepen. Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biol*, 10(12): e1001446, December 2012.

[114] Joanne K. Hobbs, Charis Shepherd, David J. Saul, Nicholas J. Demetras, Svend Haaning, Colin R. Monk, Roy M. Daniel, and Vickery L. Arcus. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of Bacillus. *Molecular Biology and Evolution*, 29(2):825–835, February 2012. ISSN 0737-4038, 1537-1719.

[115] Satoshi Akanuma, Yoshiki Nakajima, Shin-ichi Yokobori, Mitsuo Kimura, Naoki Nemoto, Tomoko Mase, Ken-ichi Miyazono, Masaru Tanokura, and Akihiko Yamagishi. Experimental evidence for the thermophilicity of ancestral life. *Proceedings of the National Academy of Sciences*, 110(27): 11067–11072, July 2013. ISSN 0027-8424, 1091-6490.

[116] Satoshi Akanuma, Shoko Iwami, Tamaki Yokoi, Nana Nakamura, Hideaki Watanabe, Shin-ichi Yokobori, and Akihiko Yamagishi. Phylogeny-Based Design of a B-Subunit of DNA Gyrase and Its ATPase Domain Using a Small Set of Homologous Amino Acid Sequences. *Journal of Molecular Biology*, 412 (2):212–225, September 2011. ISSN 0022-2836.

[117] N. B. Loughran, M. J. O'Connell, B. O'Connor, and C. 'Fgin. Stability properties of an ancient plant peroxidase. *Biochimie*, 104:156–159, September 2014. ISSN 0300-9084.

[118] Raul Perez-Jimenez, Alvaro Ingls-Prieto, Zi-Ming Zhao, Inmaculada Sanchez-Romero, Jorge Alegre-Cebollada, Pallav Kosuri, Sergi Garcia-Manyes, T. Joseph Kappock, Masaru Tanokura, Arne Holmgren, Jose M. Sanchez-Ruiz, Eric A. Gaucher, and Julio M. Fernandez. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nature Structural & Molecular Biology*, 18(5):592–596, May 2011. ISSN 1545-9993.

[119] Valeria A. Risso, Jose A. Gavira, and Jose M. Sanchez-Ruiz. Thermostable and promiscuous Precambrian proteins. *Environmental Microbiology*, 16(6): 1485–1489, June 2014. ISSN 1462-2920.

[120] Megan F. Cole and Eric A. Gaucher. Utilizing natural diversity to evolve protein function: applications towards thermostability. *Current Opinion in Chemical Biology*, 15(3):399–406, June 2011. ISSN 1879-0402.

[121] Jason H. Whitfield, William H. Zhang, Michel K. Herde, Ben E. Clifton, Johanna Radziejewski, Harald Janovjak, Christian Henneberger, and Colin J. Jackson. Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Science*, 24(9):1412–1422, September 2015. ISSN 1469-896X.

[122] M. M. Gromiha, M. Oobatake, and A. Sarai. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemistry*, 82(1):51–67, November 1999. ISSN 0301-4622.

[123] Darin M. Taverna and Richard A. Goldstein. Why are proteins marginally stable? *Proteins*, 46(1):105–109, January 2002. ISSN 0887-3585.

[124] Fabia U. Battistuzzi, Andreia Feijao, and S. Blair Hedges. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*, 4:44, November 2004. ISSN 1471-2148.

[125] B. A. Malcolm, K. P. Wilson, B. W. Matthews, J. F. Kirsch, and A. C. Wilson. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, 345(6270):86–89, May 1990. ISSN 0028-0836.

[126] Pouria Dasmeh, Adrian W. R. Serohijos, Kasper P. Kepp, and Eugene I. Shakhnovich. Positively Selected Sites in Cetacean Myoglobins Contribute to Protein Stability. *PLOS Computational Biology*, 9(3):e1002929, March 2013. ISSN 1553-7358.

[127] Lizhi Ian Gong, Marc A. Suchard, and Jesse D. Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, May 2013. ISSN 2050-084X.

[128] Mathieu Groussin, Joanne K. Hobbs, Gergely J. Szllsi, Simonetta Gribaldo, Vickery L. Arcus, and Manolo Gouy. Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees. *Molecular Biology and Evolution*, 32(1):13–22, January 2015. ISSN 1537-1719.

[129] Satoshi Akanuma, Shin-ichi Yokobori, Yoshiki Nakajima, Mizumo Bessho, and Akihiko Yamagishi. Robustness of predictions of extremely thermally stable proteins in ancient organisms. *Evolution*, 69(11):2954–2962, November 2015. ISSN 1558-5646.

[130] Hagit Bar-Rogovsky, Adi Stern, Osnat Penn, Iris Kobl, Tal Pupko, and Dan S. Tawfik. Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein engineering, design & selection: PEDS*, 28(11): 507–518, November 2015. ISSN 1741-0134.

[131] Paul D. Williams, David D. Pollock, Benjamin P. Blackburne, and Richard A. Goldstein. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. *PLOS Computational Biology*, 2(6):e69, June 2006. ISSN 1553-7358.

[132] Shimon Bershtein, Korina Goldin, and Dan S. Tawfik. Intense neutral drifts yield robust and evolvable consensus proteins. *Journal of Molecular Biology*, 379(5):1029–1044, June 2008. ISSN 1089-8638.

[133] David D. Pollock, Grant Thiltgen, and Richard A. Goldstein. Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences*, 109(21):E1352–E1359, May 2012. ISSN 0027-8424, 1091-6490.

[134] Richard A. Goldstein, Stephen T. Pollard, Seena D. Shah, and David D. Pollock. Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Molecular Biology and Evolution*, 32(6):1373–1381, June 2015. ISSN 0737-4038.

[135] Brian Gaschen, Jesse Taylor, Karina Yusim, Brian Foley, Feng Gao, Dorothy Lang, Vladimir Novitsky, Barton Haynes, Beatrice H. Hahn, Tanmoy Bhattacharya, and Bette Korber. Diversity considerations in HIV-1 vaccine selection. *Science (New York, N.Y.)*, 296(5577):2354–2360, June 2002. ISSN 1095-9203.

[136] Denise L. Kothe, Yingying Li, Julie M. Decker, Frederic Bibollet-Ruche, Kenneth P. Zammit, Maria G. Salazar, Yalu Chen, Zhiping Weng, Eric A. Weaver, Feng Gao, Barton F. Haynes, George M. Shaw, Bette T. M. Korber, and Beatrice H. Hahn. Ancestral and consensus envelope immunogens for HIV-1 subtype C. *Virology*, 352(2):438–449, September 2006. ISSN 0042-6822.

[137] Valeria A. Risso, Jose A. Gavira, Eric A. Gaucher, and Jose M. Sanchez-Ruiz. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins*, 82(6):887–896, June 2014. ISSN 1097-0134.

[138] R. A. Jensen. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, 30(1):409–425, 1976.

[139] M Lynch and A Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, January 2000. ISSN 0016-6731.

[140] Gavin C. Conant and Kenneth H. Wolfe. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, December 2008. ISSN 1471-0056.

[141] Sheng Ma, Jacqueline Martin-Laffon, Morgane Mininno, Ocane Gigarel, Sabine Brugire, Olivier Bastien, Marianne Tardif, Stphane Ravanel, and Claude Alban. Molecular Evolution of the Substrate Specificity of Chloroplastic Aldolases/Rubisco Lysine Methyltransferases in Plants. *Molecular Plant*, 9(4): 569–581, April 2016. ISSN 1752-9867.

[142] Merridee A. Wouters, Ke Liu, Peter Riek, and Ahsan Husain. A despecialization step underlying evolution of a family of serine proteases. *Molecular Cell*, 12(2):343–354, August 2003. ISSN 1097-2765.

[143] Camille Sayou, Marie Monniaux, Max H. Nanao, Edwige Moyroud, Samuel F. Brockington, Emmanuel Thvenon, Hicham Chahtane, Norman Warthmann, Michael Melkonian, Yong Zhang, Gane Ka-Shu Wong, Detlef Weigel, Franois Parcy, and Renaud Dumas. A Promiscuous Intermediate Underlies the Evolution of LEAFY DNA Binding Specificity. *Science*, 343(6171):645–648, February 2014. ISSN 0036-8075, 1095-9203.

[144] A. Chinen, Y. Naito, N. Handa, and I. Kobayashi. Evolution of sequence recognition by restriction-modification enzymes: selective pressure for specificity decrease. *Molecular Biology and Evolution*, 17(11):1610–1619, November 2000. ISSN 0737-4038.

[145] Orit Peleg, Jeong-Mo Choi, and EugeneI. Shakhnovich. Evolution of Specificity in Protein-Protein Interactions. *Biophysical Journal*, 107(7): 1686–1696, October 2014. ISSN 0006-3495.

[146] Juhan Kim and Shelley D. Copley. Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42): E2856–E2864, October 2012. ISSN 0027-8424.

[147] Jungeui Hong and David Gresham. Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments. *PLoS Genet*, 10(1):e1004041, January 2014.

[148] Marjon G. J. de Vos, Alexandre Dawid, Vanda Sunderlikova, and Sander J. Tans. Breaking evolutionary constraint with a tradeoff ratchet. *Proceedings of the National Academy of Sciences*, 112(48):14906–14911, December 2015. ISSN 0027-8424, 1091-6490.

[149] Andreas Ernst, David Gfeller, Zhengyan Kan, Somasekar Seshagiri, Philip M. Kim, Gary D. Bader, and Sachdev S. Sidhu. Coevolution of PDZ domainligand interactions analyzed by high-throughput phage display and deep sequencing. *Molecular BioSystems*, 6(10):1782, 2010. ISSN 1742-206X, 1742-2051.

[150] Alexander J. Stewart and Joshua B. Plotkin. The evolution of complex gene regulation by low-specificity binding sites. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1768):20131313, October 2013. ISSN 0962-8452, 1471-2954.

[151] Ronald Wolf, O. M. Zack Howard, Hui-Fang Dong, Christopher Voscopoulos, Karen Boeshans, Jason Winston, Rao Divi, Michele Gunsior, Paul Goldsmith, Bijan Ahvazi, Triantafyllos Chavakis, Joost J. Oppenheim, and Stuart H. Yuspa. Chemotactic activity of S100a7 (Psoriasin) is mediated by the receptor for advanced glycation end products and potentiates inflammation with highly homologous but functionally distinct S100a15. *Journal of Immunology (Baltimore, Md.: 1950)*, 181(2):1499–1506, July 2008. ISSN 1550-6606.

[152] Guglielmo Sorci, Gloria Giovannini, Francesca Riuzzi, Pierluigi Bonifazi, Teresa Zelante, Silvia Zagarella, Francesco Bistoni, Rosario Donato, and Luigina Romani. The danger signal S100b integrates pathogen- and danger-sensing pathways to restrain inflammation. *PLoS pathogens*, 7(3): e1001315, March 2011. ISSN 1553-7374.

[153] Sean S. Shaw, Ann Marie Schmidt, Amy K. Banes, Xiaodan Wang, David M. Stern, and Mario B. Marrero. S100b-RAGE-mediated augmentation of angiotensin II-induced activation of JAK2 in vascular smooth muscle cells is dependent on PLD2. *Diabetes*, 52(9):2381–2388, September 2003. ISSN 0012-1797.

[154] Jrg Klingelhfer, Henrik D. Mller, Eren U. Sumer, Christian H. Berg, Maria Poulsen, Darya Kiryushko, Vladislav Soroka, Noona Ambartsumian, Mariam Grigorian, and Eugene M. Lukanidin. Epidermal growth factor receptor ligands as new extracellular targets for the metastasis-promoting S100a4 protein. *The FEBS journal*, 276(20):5936–5948, October 2009. ISSN 1742-4658.

[155] Xiangyu Wang, Jing Yang, Jingfeng Qian, Zhihua Liu, Hongyan Chen, and Zhumei Cui. S100a14, a mediator of epithelial-mesenchymal transition, regulates proliferation, migration and invasion of human cervical cancer cells. *American Journal of Cancer Research*, 5(4):1484–1495, March 2015. ISSN 2156-6976.

[156] Zheng Yang, Wei Xing Yan, Hong Cai, Nicodemus Tedla, Chris Armishaw, Nick Di Girolamo, Hong Wei Wang, Taline Hampartzoumian, Jodie L. Simpson, Peter G. Gibson, John Hunt, Prue Hart, J. Margaret Hughes, Michael A. Perry, Paul F. Alewood, and Carolyn L. Geczy. S100a12 provokes mast cell activation: a potential amplification pathway in asthma and innate immunity. *The Journal of Allergy and Clinical Immunology*, 119(1):106–114, January 2007. ISSN 0091-6749.

[157] Joseph P. Zackular, Walter J. Chazin, and Eric P. Skaar. Nutritional Immunity: S100 Proteins at the Host-Pathogen Interface. *Journal of Biological Chemistry*, 290(31):18991–18998, July 2015. ISSN 0021-9258, 1083-351X.

[158] F Sedaghat and A Notopoulos. S100 protein family and its application in clinical practice. *Hippokratia*, 12(4):198–204, 2008. ISSN 1108-4189.

[159] Rosario Donato. RAGE: A Single Receptor for Several Ligands and Different Cellular Responses: The Case of Certain S100 Proteins. *Current Molecular Medicine*, 7(8):711–724, December 2007.

[160] N. R. West and P. H. Watson. S100a7 (psoriasin) is induced by the proinflammatory cytokines oncostatin-M and interleukin-6 in human breast cancer. *Oncogene*, 29(14):2083–2092, April 2010. ISSN 0950-9232.

[161] Michelle M. Averill, Shelley Barnhart, Lev Becker, Xin Li, Jay W. Heinecke, Renee C. LeBoeuf, Jessica A. Hamerman, Clemens Sorg, Claus Kerkhoff, and Karin E. Bornfeldt. S100a9 Differentially Modifies Phenotypic States of Neutrophils, Macrophages, and Dendritic CellsClinical Perspective. *Circulation*, 123(11):1216–1226, March 2011. ISSN 0009-7322, 1524-4539.

[162] Kjetil Boye and Gunhild M. Mlandsmo. S100a4 and Metastasis: A Small Actor Playing Many Roles. *The American Journal of Pathology*, 176(2): 528–535, February 2010. ISSN 0002-9440.

[163] Masaya Yamaoka, Norikazu Maeda, Seiji Nakamura, Takuya Mori, Kana Inoue, Keisuke Matsuda, Ryohei Sekimoto, Susumu Kashine, Yasuhiko Nakagawa, Yu Tsushima, Yuya Fujishima, Noriyuki Komura, Ayumu Hirata, Hitoshi Nishizawa, Yuji Matsuzawa, Ken-ichi Matsubara, Tohru Funahashi, and Iichiro Shimomura. Gene expression levels of S100 protein family in blood cells are associated with insulin resistance and inflammation (Peripheral blood S100 mRNAs and metabolic syndrome). *Biochemical and Biophysical Research Communications*, 433(4):450–455, April 2013. ISSN 0006-291X.

[164] Stephane R. Gross, Connie Goh Then Sin, Roger Barraclough, and Philip S. Rudland. Joining S100 proteins and migration: for better or for worse, in sickness and in health. *Cellular and Molecular Life Sciences*, 71(9):1551–1579, June 2013. ISSN 1420-682X, 1420-9071.

[165] Anne R. Bresnick, David J. Weber, and Danna B. Zimmer. S100 proteins in cancer. *Nature Reviews Cancer*, 15(2):96–109, February 2015. ISSN 1474-175X.

[166] Ivano Bertini, Valentina Borsi, Linda Cerofolini, Soumyasri Das Gupta, Marco Fragai, and Claudio Luchinat. Solution structure and dynamics of human S100a14. *JBIC Journal of Biological Inorganic Chemistry*, 18(2):183–194, November 2012. ISSN 0949-8257, 1432-1327.

[167] Richard R. Rustandi, Alexander C. Drohat, Donna M. Baldisseri, Paul T. Wilder, and David J. Weber. The Ca2+-Dependent Interaction of S100b() with a Peptide Derived from p53. *Biochemistry*, 37(7):1951–1960, February 1998. ISSN 0006-2960.

[168] Danna B. Zimmer, Patti Wright Sadosky, and David J. Weber. Molecular mechanisms of S100-target protein interactions. *Microscopy Research and Technique*, 60(6):552–559, April 2003. ISSN 1059-910X.

[169] Danna B. Zimmer and David J. Weber. The Calcium-Dependent Interaction of S100b with Its Protein Targets. *Cardiovascular Psychiatry and Neurology*, 2010, 2010. ISSN 2090-0163.

[170] Olga V. Moroz, Keith S. Wilson, and Igor B. Bronstein. The role of zinc in the S100 proteins: insights from the X-ray structures. *Amino Acids*, 41(4): 761–772, March 2010. ISSN 0939-4451, 1438-2199.

[171] Benjamin A. Gilston, Eric P. Skaar, and Walter J. Chazin. Binding of transition metals to S100 proteins. *Science China Life Sciences*, pages 1–10, July 2016. ISSN 1674-7305, 1869-1889.

[172] Vaithiyalingam Sivaraja, Thallapuranam Krishnaswamy Suresh Kumar, Dakshinamurthy Rajalingam, Irene Graziani, Igor Prudovsky, and Chin Yu. Copper Binding Affinity of S100a13, a Key Component of the FGF-1 Nonclassical Copper-Dependent Release Complex. *Biophysical Journal*, 91(5): 1832–1843, September 2006. ISSN 0006-3495.

[173] Jrg Heierhorst, Richard J. Mann, and Bruce E. Kemp. Interaction of the Recombinant S100a1 Protein with Twitchin Kinase, and Comparison with Other Ca2+-Binding Proteins. *European Journal of Biochemistry*, 249(1): 127–133, October 1997. ISSN 1432-1033.

[174] Thomas V. O'Halloran and Valeria Cizewski Culotta. Metallochaperones, an Intracellular Shuttle Service for Metal Ions. *Journal of Biological Chemistry*, 275(33):25057–25060, August 2000. ISSN 0021-9258, 1083-351X.

[175] Wolfgang Maret. Zinc Biochemistry: From a Single Zinc Enzyme to a Key Element of Life. *Advances in Nutrition: An International Review Journal*, 4 (1):82–91, January 2013. ISSN , 2156-5376.

[176] Fabio Arnesano, Lucia Banci, Ivano Bertini, Adele Fantoni, Leonardo Tenori, and Maria Silvia Viezzoli. Structural Interplay between Calcium(II) and Copper(II) Binding to S100a13 Protein. *Angewandte Chemie International Edition*, 44(39):6341–6344, October 2005. ISSN 1521-3773.

[177] Michael Koch, Shibani Bhattacharya, Torsten Kehl, Mario Gimona, Milan Vak, Walter Chazin, Claus W. Heizmann, Peter M. H. Kroneck, and Gnter Fritz. Implications on zinc binding to S100a2. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1773(3):457–470, March 2007. ISSN 0167-4889.

[178] Timothy Ravasi, Kenneth Hsu, Jesse Goyette, Kate Schroder, Zheng Yang, Farid Rahimi, Les P. Miranda, Paul F. Alewood, David A. Hume, and Carolyn Geczy. Probing the S100 protein family through genomic and functional analysis. *Genomics*, 84(1):10–22, July 2004. ISSN 0888-7543.

[179] Xuan Shang, Hanhua Cheng, and Rongjia Zhou. Chromosomal mapping, differential origin and evolution of the S100 gene family. *Genetics Selection Evolution*, 40:449, 2008. ISSN 1297-9686.

[180] S. Blair Hedges, Fabia U. Battistuzzi, and Jaime E. Blair. Molecular Timescale of Evolution in the Proterozoic. In Shuhai Xiao and Alan J. Kaufman, editors, *Neoproterozoic Geobiology and Paleobiology*, number 27 in Topics in Geobiology, pages 199–229. Springer Netherlands, 2006. ISBN 978-1-4020-5201-9 978-1-4020-5202-6.

[181] R. Alexander Pyron and John J. Wiens. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, 61(2): 543–583, November 2011. ISSN 1055-7903.

[182] Ylenia Chiari, Vincent Cahais, Nicolas Galtier, and Frdric Delsuc. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biology*, 10:65, 2012. ISSN 1741-7007.

[183] Brant C. Faircloth, Laurie Sorenson, Francesco Santini, and Michael E. Alfaro. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). *PLOS ONE*, 8(6):e65923, June 2013. ISSN 1932-6203.

[184] Richard E. Green, Edward L. Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W. Vandewege, John A. St John, Salvador Capella-Gutirrez, Todd A. Castoe, Colin Kern, Matthew K. Fujita, Juan C. Opazo, Jerzy Jurka, Kenji K. Kojima, Juan Caballero, Robert M. Hubley, Arian F. Smit, Roy N. Platt, Christine A. Lavoie, Meganathan P. Ramakodi, John W. Finger, Alexander Suh, Sally R. Isberg, Lee Miles, Amanda Y. Chong, Weerachai Jaratlerdsiri, Jaime Gongora, Christopher Moran, Andrs Iriarte, John McCormack, Shane C. Burgess, Scott V. Edwards, Eric Lyons, Christina Williams, Matthew Breen, Jason T. Howard, Cathy R. Gresham, Daniel G. Peterson, Jrgen Schmitz, David D. Pollock, David Haussler, Eric W. Triplett, Guojie Zhang, Naoki Irie, Erich D. Jarvis, Christopher A. Brochu, Carl J. Schmidt, Fiona M. McCarthy, Brant C. Faircloth, Federico G. Hoffmann, Travis C. Glenn, Toni Gabaldn, Benedict Paten, and David A. Ray. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215):1254449, December 2014. ISSN 0036-8075, 1095-9203.

[185] N. Satoh, D. Rokhsar, and T. Nishikawa. Chordate evolution and the three-phylum system. *Proceedings of the Royal Society B: Biological Sciences*, 281(1794):20141729–20141729, September 2014. ISSN 0962-8452, 1471-2954.

[186] Susanne Gallus, Axel Janke, Vikas Kumar, and Maria A. Nilsson. Disentangling the relationship of the Australian marsupial orders using retrotransposon and evolutionary network analyses. *Genome Biology and Evolution*, 7(4):985–992, April 2015. ISSN 1759-6653.

[187] Richard O. Prum, Jacob S. Berv, Alex Dornburg, Daniel J. Field, Jeffrey P. Townsend, Emily Moriarty Lemmon, and Alan R. Lemmon. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, 526(7574):569–573, October 2015. ISSN 0028-0836.

[188] Pndaro Daz-Jaimes, Natalia J. Bayona-Vsquez, Douglas H. Adams, and Manuel Uribe-Alcocer. Complete mitochondrial DNA genome of bonnethead shark, Sphyrna tiburo, and phylogenetic relationships among main superorders of modern elasmobranchs. *Meta Gene*, 7:48–55, February 2016. ISSN 2214-5400.

[189] James E. Tarver, Mario dos Reis, Siavash Mirarab, Raymond J. Moran, Sean Parker, Joseph E. OReilly, Benjamin L. King, Mary J. OConnell, Robert J. Asher, Tandy Warnow, Kevin J. Peterson, Philip C. J. Donoghue, and Davide Pisani. The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biology and Evolution*, 8(2):330–344, February 2016. ISSN , 1759-6653.

[190] J. R. Dorin, E. Emslie, and V. van Heyningen. Related calcium-binding proteins map to the same subregion of chromosome 1q and to an extended region of synteny on mouse chromosome 3. *Genomics*, 8(3):420–426, November 1990. ISSN 0888-7543.

[191] P. O. Tsvetkov, F. Devred, and A. A. Makarov. Thermodynamics of zinc binding to human S100a2. *Molecular Biology*, 44(5):832–835, October 2010. ISSN 0026-8933, 1608-3245.

[192] H. Vorum, P. Madsen, H. H. Rasmussen, M. Etzerodt, I. Svendsen, J. E. Celis, and B. Honor. Expression and divalent cation binding properties of the novel chemotactic inflammatory protein psoriasin. *Electrophoresis*, 17(11): 1787–1796, November 1996. ISSN 0173-0835.

[193] Jolanta Kordowska, Walter F. Stafford, and C.-L. Albert Wang. Ca2+ and Zn2+ bind to different sites and induce different conformational changes in human calcyclin. *European Journal of Biochemistry*, 253(1):57–66, April 1998. ISSN 1432-1033.

[194] Gnter Fritz, Peer R. E. Mittl, Milan Vasak, Markus G. Grtter, and Claus W. Heizmann. The Crystal Structure of Metal-free Human EF-hand Protein S100a3 at 1.7- Resolution. *Journal of Biological Chemistry*, 277(36): 33092–33098, September 2002. ISSN 0021-9258, 1083-351X.

[195] Beat W. Schfer, Jean-Marc Fritschy, Petra Murmann, Heinz Troxler, Isabelle Durussel, Claus W. Heizmann, and Jos A. Cox. Brain S100a5 Is a Novel Calcium-, Zinc-, and Copper Ion-binding Protein of the EF-hand Superfamily. *Journal of Biological Chemistry*, 275(39):30623–30630, September 2000. ISSN 0021-9258, 1083-351X.

[196] J. Baudier, N. Glasser, and D. Gerard. Ions binding to S100 proteins. I. Calcium- and zinc-binding properties of bovine brain S100 alpha alpha, S100a (alpha beta), and S100b (beta beta) protein: Zn2+ regulates Ca2+ binding on S100b protein. *Journal of Biological Chemistry*, 261(18):8192–8203, June 1986. ISSN 0021-9258, 1083-351X.

[197] Olga V. Moroz, Will Burkitt, Helmut Wittkowski, Wei He, Anatoli Ianoul, Vera Novitskaya, Jingjing Xie, Oxana Polyakova, Igor K. Lednev, Alexander Shekhtman, Peter J. Derrick, Per Bjoerk, Dirk Foell, and Igor B. Bronstein. Both Ca2+ and Zn2+ are essential for S100a12 protein oligomerization and function. *BMC Biochemistry*, 10:11, 2009. ISSN 1471-2091.

[198] Dean E. Wilcox. Isothermal titration calorimetry of metal ions binding to proteins: An overview of recent studies. *Inorganica Chimica Acta*, 361(4): 857–867, March 2008. ISSN 0020-1693.

[199] Emmanuel Sturchler, Jos A. Cox, Isabelle Durussel, Mirjam Weibel, and Claus W. Heizmann. S100a16, a Novel Calcium-binding Protein of the EF-hand Superfamily. *Journal of Biological Chemistry*, 281(50):38905–38917, December 2006. ISSN 0021-9258, 1083-351X.

[200] T. Becker, V. Gerke, E. Kube, and K. Weber. S100p, a novel Ca(2+)-binding protein from human placenta. cDNA cloning, recombinant protein expression and Ca2+ binding properties. *European journal of biochemistry / FEBS*, 207 (2):541–547, July 1992. ISSN 0014-2956.

[201] S. Rty, J. Sopkova, M. Renouard, D. Osterloh, V. Gerke, S. Tabaries, F. Russo-Marie, and A. Lewit-Bentley. The crystal structure of a complex of p11 with the annexin II N-terminal peptide. *Nature Structural Biology*, 6(1): 89–95, January 1999. ISSN 1072-8368.

[202] Paul T. Wilder, Donna M. Baldisseri, Ryan Udan, Kristen M. Vallely, and David J. Weber. Location of the Zn2+-Binding Site on S100b As Determined by NMR Spectroscopy and Site-Directed Mutagenesis. *Biochemistry*, 42(46): 13410–13421, November 2003. ISSN 0006-2960.

[203] Nathan T. Wright, Kristen M. Varney, Karen C. Ellis, Joseph Markowitz, Rossitza K. Gitti, Danna B. Zimmer, and David J. Weber. The Three-dimensional Solution Structure of Ca2+-bound S100a1 as Determined by NMR Spectroscopy. *Journal of Molecular Biology*, 353(2):410–426, October 2005. ISSN 0022-2836.

[204] Sarah C. Garrett, Louis Hodgson, Andrew Rybin, Alexei Toutchkine, Klaus M. Hahn, David S. Lawrence, and Anne R. Bresnick. A biosensor of S100a4 metastasis factor activation: inhibitor screening and cellular activation dynamics. *Biochemistry*, 47(3):986–996, January 2008. ISSN 0006-2960.

[205] Jill I Murray, Michelle L Tonkin, Amanda L Whiting, Fangni Peng, Benjamin Farnell, Jay T Cullen, Fraser Hof, and Martin J Boulanger. Structural characterization of S100a15 reveals a novel zinc coordination site among S100 proteins and altered surface chemistry with functional implications for receptor binding. *BMC Structural Biology*, 12:16, July 2012. ISSN 1472-6807.

[206] Elena Babini, Ivano Bertini, Valentina Borsi, Vito Calderone, Xiaoyu Hu, Claudio Luchinat, and Giacomo Parigi. Structural characterization of human S100a16, a low-affinity calcium binder. *Journal of biological inorganic chemistry: JBIC: a publication of the Society of Biological Inorganic Chemistry*, 16(2):243–256, February 2011. ISSN 1432-1327.

[207] Ivano Bertini, Soumyasri Das Gupta, Xiaoyu Hu, Tilemachos Karavelas, Claudio Luchinat, Giacomo Parigi, and Jing Yuan. Solution structure and dynamics of S100a5 in the apo and Ca2+-bound states. *JBIC Journal of Biological Inorganic Chemistry*, 14(7):1097–1107, June 2009. ISSN 0949-8257, 1432-1327.

[208] Rajam S. Mani and Cyril M. Kay. Circular dichroism studies on the zinc-induced conformational changes in S-100a and S-100b proteins. *FEBS Letters*, 163(2):282–286, November 1983. ISSN 1873-3468.

[209] Beat W. Schfer and Claus W. Heizmann. The S100 family of EF-hand calcium-binding proteins: functions and pathology. *Trends in Biochemical Sciences*, 21(4):134–140, April 1996. ISSN 0968-0004.

[210] Helena Hernndez and Carol V. Robinson. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nature Protocols*, 2(3):715–726, March 2007. ISSN 1754-2189.

[211] Werner W. Streicher, Maria M. Lopez, and George I. Makhatadze. Modulation of Quaternary Structure of S100 Proteins by Calcium Ions. *Biophysical chemistry*, 151(3):181–186, October 2010. ISSN 0301-4622.

[212] M. M. Yamashita, L. Wesson, G. Eisenman, and D. Eisenberg. Where metal ions bind in proteins. *Proceedings of the National Academy of Sciences*, 87 (15):5648–5652, August 1990. ISSN 0027-8424, 1091-6490.

[213] Mariana Babor, Sergey Gerzon, Barak Raveh, Vladimir Sobolev, and Marvin Edelman. Prediction of transition metal-binding sites from apo protein structures. *Proteins: Structure, Function, and Bioinformatics*, 70(1):208–217, January 2008. ISSN 1097-0134.

[214] Jeffrey T. Rubino and Katherine J. Franz. Coordination chemistry of copper proteins: How nature handles a toxic cargo for essential function. *Journal of Inorganic Biochemistry*, 107(1):129–143, February 2012. ISSN 0162-0134.

[215] Liam J. Holt, Brian B. Tuch, Judit Villn, Alexander D. Johnson, Steven P. Gygi, and David O. Morgan. Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution. *Science*, 325(5948): 1682–1686, September 2009. ISSN 0036-8075, 1095-9203.

[216] Alexey V. Gribenko and George I. Makhatadze. Oligomerization and divalent ion binding properties of the S100p protein: a Ca2+/Mg2+-switch model1. *Journal of Molecular Biology*, 283(3):679–694, October 1998. ISSN 0022-2836.

[217] J. S. Mills and J. D. Johnson. Metal ions as allosteric regulators of calmodulin. *Journal of Biological Chemistry*, 260(28):15100–15105, December 1985. ISSN 0021-9258, 1083-351X.

[218] Zenon Grabarek. Insights into Modulation of Calcium Signaling by Magnesium in Calmodulin, Troponin C and Related EF-hand Proteins. *Biochimica et biophysica acta*, 1813(5):913–921, May 2011. ISSN 0006-3002.

[219] Hee Jung Chung, Du Young Ko, Hyo Jung Moon, and Byeongmoon Jeong. EF-Hand Mimicking Calcium Binding Polymer. *Biomacromolecules*, 17(3): 1075–1082, March 2016. ISSN 1526-4602.

[220] Per Bjrk, Anders Bjrk, Thomas Vogl, Martin Stenstrm, David Liberg, Anders Olsson, Johannes Roth, Fredrik Ivars, and Tomas Leanderson. Identification of Human S100a9 as a Novel Target for Treatment of Autoimmune Disease via Binding to Quinoline-3-Carboxamides. *PLOS Biol*, 7(4):e1000097, April 2009. ISSN 1545-7885.

[221] Claus Kerkhoff, Thomas Vogl, Wolfgang Nacken, Claudia Sopalla, and Clemens Sorg. Zinc binding reverses the calcium-induced arachidonic acid-binding capacity of the S100a8/A9 protein complex. *FEBS Letters*, 460 (1):134–138, October 1999. ISSN 1873-3468.

[222] Derek M. Gagnon, Megan Brunjes Brophy, Sarah E. J. Bowman, Troy A. Stich, Catherine L. Drennan, R. David Britt, and Elizabeth M. Nolan. Manganese binding properties of human calprotectin under conditions of high and low calcium: X-ray crystallographic and advanced electron paramagnetic resonance spectroscopic analysis. *Journal of the American Chemical Society*, 137(8):3004–3016, March 2015. ISSN 1520-5126.

[223] Alexander Hopt, Stefan Korte, Herbert Fink, Ulrich Panne, Reinhard Niessner, Reinhard Jahn, Hans Kretzschmar, and Jochen Herms. Methods for studying synaptosomal copper release. *Journal of Neuroscience Methods*, 128(1-2): 159–172, September 2003. ISSN 0165-0270.

[224] Taisun H. Hyun, Elizabeth Barrett-Connor, and David B. Milne. Zinc intakes and plasma concentrations in men with osteoporosis: the Rancho Bernardo Study. *The American Journal of Clinical Nutrition*, 80(3):715–721, September 2004. ISSN 0002-9165, 1938-3207.

[225] Haimoto H, Hosoda S, and Kato K. Differential distribution of immunoreactive S100-alpha and S100-beta proteins in normal nonnervous human tissues. *Laboratory investigation; a journal of technical methods and pathology*, 57(5): 489–498, 1987. ISSN 0023-6837.

[226] D. B. Zimmer and L. J. Van Eldik. Tissue distribution of rat S100 alpha and S100 beta and S100-binding proteins. *American Journal of Physiology - Cell Physiology*, 252(3):C285–C289, March 1987. ISSN 0363-6143, 1522-1563.

[227] Jacek Kunicki, Anna Filipek, Peter Heimann, Leszek Kaczmarek, and Boena Kamiska. Tissue specific distribution of calcyclin  10.5 kDa Ca2+ -binding protein. *FEBS Letters*, 254(1):141–144, August 1989. ISSN 0014-5793.

[228] Danna B. Zimmer, Emily H. Cornwall, Aimee Landar, and Wei Song. The S100 protein family: History, function, and expression. *Brain Research Bulletin*, 37(4):417–429, 1995. ISSN 0361-9230.

[229] Alexey V. Gribenko, James E. Hopper, and George I. Makhatadze. Molecular Characterization and Tissue Distribution of a Novel Member of the S100 Family of EF-Hand Proteins,. *Biochemistry*, 40(51):15538–15548, December 2001. ISSN 0006-2960.

[230] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. ISSN 0022-2836.

[231] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database issue):D19–D21, January 2011. ISSN 0305-1048.

[232] Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, 29(7):644–652, May 2011. ISSN 1087-0156.

[233] W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics (Oxford, England)*, 17(3):282–283, March 2001. ISSN 1367-4803.

[234] Yongchao Liu, Bertil Schmidt, and Douglas L. Maskell. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, 26(16):1958–1964, August 2010. ISSN 1367-4803, 1460-2059.

[235] Anders Larsson. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, November 2014. ISSN 1367-4803, 1460-2059.

[236] Stphane Guindon, Jean-Franois Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, May 2010. ISSN 1076-836X.

[237] Si Quang Le and Olivier Gascuel. An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25(7):1307–1320, July 2008. ISSN 0737-4038, 1537-1719.

[238] Maria Anisimova, Manuel Gil, Jean-Franois Dufayard, Christophe Dessimoz, and Olivier Gascuel. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*, page syr041, May 2011. ISSN 1063-5157, 1076-836X.

[239] Andre J. Aberer, Kassian Kobert, and Alexandros Stamatakis. ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Molecular Biology and Evolution*, page msu236, August 2014. ISSN 0737-4038, 1537-1719.

[240] David T. Jones, William R. Taylor, and Janet M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, 8(3):275–282, June 1992. ISSN 1367-4803, 1460-2059.

[241] Stanley C. Gill and Peter H. von Hippel. Calculation of protein extinction coefficients from amino acid sequence data. *Analytical Biochemistry*, 182(2): 319–326, November 1989. ISSN 0003-2697.

[242] John M. Walker, editor. *The Proteomics Protocols Handbook*. Humana Press, Totowa, NJ, 2005. ISBN 978-1-58829-343-5 978-1-59259-890-8.

[243] B. Birdsall, R. W. King, M. R. Wheeler, C. A. Lewis, S. R. Goode, R. B. Dunlap, and G. C. Roberts. Correction for light absorption in fluorescence studies of protein-ligand interactions. *Analytical Biochemistry*, 132(2): 353–361, July 1983. ISSN 0003-2697.

[244] P Schuck. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophysical Journal*, 78(3):1606–1619, March 2000. ISSN 0006-3495.

[245] Patrick H. Brown and Peter Schuck. Macromolecular Size-and-Shape Distributions by Sedimentation Velocity Analytical Ultracentrifugation. *Biophysical Journal*, 90(12):4651–4661, June 2006. ISSN 0006-3495.

[246] Benjamin Webb and Andrej Sali. Comparative Protein Structure Modeling Using MODELLER. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002. ISBN 978-0-471-25095-1.

[247] Iktae Kim, Ko On Lee, Young-Joo Yun, Jea Yeon Jeong, Eun-Hee Kim, Haekap Cheong, Kyoung-Seok Ryu, Nak-Kyoon Kim, and Jeong-Yong Suh. Biophysical characterization of Ca2+-binding of S100a5 and Ca2+-induced interaction with RAGE. *Biochemical and Biophysical Research Communications*, 483(1):332–338, January 2017. ISSN 0006-291X.

[248] Adrian M. Fischl, Paula M. Heron, Arnold J. Stromberg, and Timothy S. McClintock. Activity-Dependent Genes in Mouse Olfactory Sensory Neurons. *Chemical Senses*, 39(5):439–449, June 2014. ISSN 0379-864X.

[249] Takumi Teratani, Takumi Watanabe, Kaori Yamahara, Hiromichi Kumagai, Akira Ishikawa, Kazumori Arai, and Ryushi Nozawa. Restricted Expression of Calcium-Binding Protein S100a5 in Human Kidney. *Biochemical and Biophysical Research Communications*, 291(3):623–627, March 2002. ISSN 0006-291X.

[250] S. Hancq, I. Salmon, J. Brotchi, O. De Witte, H.-J. Gabius, C. W. Heizmann, R. Kiss, and C. Decaestecker. S100a5: a marker of recurrence in WHO grade I meningiomas. *Neuropathology and Applied Neurobiology*, 30(2):178–187, April 2004. ISSN 1365-2990.

[251] WikiGenes - Collaborative Publishing, .

[252] S100a5 Gene - GeneCards | S10a5 Protein | S10a5 Antibody, .

[253] S100a5 - Protein S100-A5 - Homo sapiens (Human) - S100a5 gene & protein, .

[254] pytc: python program for analyzing isothermal titration calorimetry data, May 2017.

[255] Huaying Zhao, Grzegorz Piszczek, and Peter Schuck. SEDPHAT  A platform for global ITC analysis and global multi-method analysis of molecular interactions. *Methods*, 76(Supplement C):137–148, April 2015. ISSN 1046-2023.

[256] Yi Zhang, Shreeram Akilesh, and Dean E. Wilcox. Isothermal Titration Calorimetry Measurements of Ni(II) and Cu(II) Binding to His, GlyGlyHis, HisGlyHis, and Bovine Serum Albumin: A Critical Evaluation. *Inorganic Chemistry*, 39(14):3057–3064, July 2000. ISSN 0020-1669.

[257] Melissa Liriano. Protein dynamics of calcium-S100a5 in the presence and absence of target peptide. *Grantome*.

[258] Johannes Reisert, Paul J. Bauer, King-Wai Yau, and Stephan Frings. The Ca-activated Cl Channel and its Control in Rat Olfactory Receptor Neurons. *The Journal of General Physiology*, 122(3):349–364, September 2003. ISSN 0022-1295, 1540-7748.

[259] Bronwen Gardner, Birger V. Dieriks, Steve Cameron, Lakshini H. S. Mendis, Clinton Turner, Richard L. M. Faull, and Maurice A. Curtis. Metal concentrations and distributions in the human olfactory bulb in Parkinsons disease. *Scientific Reports*, 7(1):10454, September 2017. ISSN 2045-2322.

[260] J. Herms, T. Tings, S. Gall, A. Madlung, A. Giese, H. Siebert, P. Schrmann, O. Windl, N. Brose, and H. Kretzschmar. Evidence of presynaptic location and function of the prion protein. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 19(20):8866–8875, October 1999. ISSN 1529-2401.

[261] M. S. Horning and P. Q. Trombley. Zinc and copper influence excitability of rat olfactory bulb neurons by multiple mechanisms. *Journal of Neurophysiology*, 86(4):1652–1660, October 2001. ISSN 0022-3077.

[262] Shin-Ichi Ono and M. George Cherian. Regional distribution of metallothionein, zinc, and copper in the brain of different strains of rats. *Biological Trace Element Research*, 69(2):151–159, August 1999. ISSN 0163-4984, 1559-0720.

[263] Olga V. Moroz, Elena V. Blagova, Anthony J. Wilkinson, Keith S. Wilson, and Igor B. Bronstein. The Crystal Structures of Human S100a12 in Apo Form and in Complex with Zinc: New Insights into S100a12 Oligomerisation. *Journal of Molecular Biology*, 391(3):536–551, August 2009. ISSN 0022-2836.

[264] Sandro Keller, Carolyn Vargas, Huaying Zhao, Grzegorz Piszczek, Chad A. Brautigam, and Peter Schuck. High-Precision Isothermal Titration Calorimetry with Automated Peak Shape Analysis. *Analytical Chemistry*, 84 (11):5066–5073, June 2012. ISSN 0003-2700.

[265] T. Wiseman, S. Williston, J. F. Brandts, and L. N. Lin. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Analytical Biochemistry*, 179(1):131–137, May 1989. ISSN 0003-2697.

[266] Ernesto Freire, Arne Schn, and Adrian VelazquezCampoy. Chapter 5 Isothermal Titration Calorimetry: General Formalism Using Binding Polynomials. In *Methods in Enzymology*, volume 455 of *Biothermodynamics, Part A*, pages 127–155. Academic Press, January 2009.

[267] Andres Kreegipuu, Nikolaj Blom, Sren Brunak, and Jaak Jrv. Statistical analysis of protein kinase specificity determinants. *FEBS Letters*, 430(1): 45–50, June 1998. ISSN 0014-5793.

[268] Kenji S. Nakahara, Chikara Masuta, Syouta Yamada, Hanako Shimura, Yukiko Kashihara, Tomoko S. Wada, Ayano Meguro, Kazunori Goto, Kazuki Tadamura, Kae Sueda, Toru Sekiguchi, Jun Shao, Noriko Itchoda, Takeshi Matsumura, Manabu Igarashi, Kimihito Ito, Richard W. Carthew, and Ichiro Uyeda. Tobacco calmodulin-like protein provides secondary defense by binding to and directing degradation of virus RNA silencing suppressors. *Proceedings of the National Academy of Sciences*, 109(25):10113–10118, June 2012. ISSN 0027-8424, 1091-6490.

[269] Paola Bertolazzi, Mary Ellen Bock, and Concettina Guerra. On the functional and structural characterization of hubs in protein-protein interaction networks. *Biotechnology Advances*, 31(2):274–286, April 2013. ISSN 1873-1899.

[270] So Nakagawa, Stephen S. Gisselbrecht, Julia M. Rogers, Daniel L. Hartl, and Martha L. Bulyk. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences*, 110(30):12349–12354, July 2013. ISSN 0027-8424, 1091-6490.

[271] Olga Khersonsky and Dan S. Tawfik. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79(1):471–505, 2010.

[272] William H. Hudson, Bradley R. Kossmann, Ian Mitchelle S. de Vera, Shih-Wei Chuo, Emily R. Weikum, Geeta N. Eick, Joseph W. Thornton, Ivaylo N. Ivanov, Douglas J. Kojetin, and Eric A. Ortlund. Distal substitutions drive divergent DNA specificity among paralogous transcription factors through subdivision of conformational space. *Proceedings of the National Academy of Sciences*, page 201518960, December 2015. ISSN 0027-8424, 1091-6490.

[273] T. Alhindi, Z. Zhang, P. Ruelens, H. Coenen, H. Degroote, N. Iraci, and K. Geuten. Protein interaction evolution from promiscuity to specificity with reduced flexibility in an increasingly complex network. *Scientific Reports*, 7, March 2017. ISSN 2045-2322.

[274] Carla Mouta Carreira, Theresa M. LaVallee, Francesca Tarantini, Anthony Jackson, Julia Tait Lathrop, Brian Hampton, Wilson H. Burgess, and Thomas Maciag. S100a13 Is Involved in the Regulation of Fibroblast Growth Factor-1 and p40 Synaptotagmin-1 Release in Vitro. *Journal of Biological Chemistry*, 273(35):22224–22231, August 1998. ISSN 0021-9258, 1083-351X.

[275] Werner W. Streicher, Maria M. Lopez, and George I. Makhatadze. Annexin I and Annexin II N-Terminal Peptides Binding to S100 Protein Family Members: Specificity and Thermodynamic Characterization. *Biochemistry*, 48 (12):2788–2798, March 2009. ISSN 0006-2960.

[276] S. Blair Hedges, Joel Dudley, and Sudhir Kumar. TimeTree: A Public Knowledge-Base of Divergence Times among Organisms. *Bioinformatics*, 22 (23):2971–2972, December 2006. ISSN 1367-4803.

[277] Wies\lawa Leniak, \Lukasz P. S\lomnicki, and Anna Filipek. S100a6 New Facts and Features. *Biochemical and Biophysical Research Communications*, 390(4):1087–1092, December 2009. ISSN 0006-291X.

[278] \Lukasz P. S\lomnicki, Barbara Nawrot, and Wies\lawa Leniak. S100a6 Binds P53 and Affects Its Activity. *The International Journal of Biochemistry & Cell Biology*, 41(4):784–790, April 2009. ISSN 1357-2725.

[279] Jan van Dieck, Maria R. Fernandez-Fernandez, Dmitry B. Veprintsev, and Alan R. Fersht. Modulation of the Oligomerization State of P53 by Differential Binding of Proteins of the S100 Family to P53 Monomers and Tetramers. *Journal of Biological Chemistry*, 284(20):13804–13811, May 2009. ISSN 0021-9258, 1083-351X.

[280] Young-Tae Lee, Yoana N. Dimitrova, Gabriela Schneider, Whitney B. Ridenour, Shibani Bhattacharya, Sarah E. Soss, Richard M. Caprioli, Anna Filipek, and Walter J. Chazin. Structure of the S100a6 Complex with a Fragment from the C-Terminal Domain of Siah-1 Interacting Protein: A Novel Mode for S100 Protein Target Recognition. *Biochemistry*, 47(41): 10921–10932, October 2008. ISSN 0006-2960.

[281] Melissa A. Liriano. *Structure, Dynamics and Function of S100B and S100A5 Complexes*. Ph.D., University of Maryland, Baltimore, United States Maryland, 2012.

[282] Thomas K. Knott, Pasil A. Madany, Ashley A. Faden, Mei Xu, Jrg Strotmann, Timothy R. Henion, and Gerald A. Schwarting. Olfactory Discrimination Largely Persists in Mice with Defects in Odorant Receptor Expression and Axon Guidance. *Neural development*, 7(1):17, 2012.

[283] Jeremy C. McIntyre, Erica E. Davis, Ariell Joiner, Corey L. Williams, I.-Chun Tsai, Paul M. Jenkins, Dyke P. McEwen, Lian Zhang, John Escobado, Sophie Thomas, and others. Gene Therapy Rescues Cilia Defects and Restores Olfactory Function in a Mammalian Ciliopathy Model. *Nature medicine*, 18 (9):1423–1428, 2012.

[284] Tsviya Olender, Ifat Keydar, Jayant M. Pinto, Pavlo Tatarskyy, Anna Alkelai, Ming-Shan Chien, Simon Fishilevich, Diego Restrepo, Hiroaki Matsunami, Yoav Gilad, and Doron Lancet. The Human Olfactory Transcriptome. *BMC genomics*, 17(1):619, August 2016. ISSN 1471-2164.

[285] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, March 2004. ISSN 0305-1048.

[286] Ludovic R. Otterbein, Jolanta Kordowska, Carlos Witte-Hoffmann, C. L. Albert Wang, and Roberto Dominguez. Crystal Structures of S100a6 in the Ca2+-Free and Ca2+-Bound States: The Calcium Sensor Mechanism of S100 Proteins Revealed at Atomic Resolution. *Structure*, 10(4):557–567, April 2002. ISSN 0969-2126.

[287] Si Quang Le and Olivier Gascuel. Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial. *Systematic Biology*, 59(3):277–287, May 2010. ISSN 1063-5157.

[288] Z. Yang, S. Kumar, and M. Nei. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. *Genetics*, 141(4):1641–1650, December 1995. ISSN 0016-6731, 1943-2631.

[289] P R Connelly and J A Thomson. Heat Capacity Changes and Hydrophobic Interactions in the Binding of FK506 and Rapamycin to the FK506 Binding Protein. *Proceedings of the National Academy of Sciences of the United States of America*, 89(11):4781–4785, June 1992. ISSN 0027-8424.

[290] Misha Soskine and Dan S. Tawfik. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics*, 11(8):572–582, August 2010. ISSN 1471-0056.

[291] Taisong Zou, Valeria A. Risso, Jose A. Gavira, Jose M. Sanchez-Ruiz, and S. Banu Ozkan. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Molecular Biology and Evolution*, 32(1):132–143, January 2015. ISSN 0737-4038, 1537-1719.

[292] Sean Michael Carroll, Jamie T. Bridgham, and Joseph W. Thornton. Evolution of Hormone Signaling in Elasmobranchs by Exploitation of Promiscuous Receptors. *Molecular Biology and Evolution*, 25(12):2643–2652, December 2008. ISSN 0737-4038.

[293] Titu Devamani, Alissa M. Rauwerdink, Mark Lunzer, Bryan J. Jones, Joanna L. Mooney, Maxilmilien Alaric O. Tan, Zhi-Jun Zhang, Jian-He Xu, Antony M. Dean, and Romas J. Kazlauskas. Catalytic Promiscuity of Ancestral Esterases and Hydroxynitrile Lyases. *Journal of the American Chemical Society*, 138(3):1046–1056, January 2016. ISSN 0002-7863.

[294] Karin Voordeckers, Ksenia Pougach, and Kevin J Verstrepen. How do regulatory networks evolve and expand throughout evolution? *Current Opinion in Biotechnology*, 34(Supplement C):180–188, August 2015. ISSN 0958-1669.

[295] Clayton D. Carlson, Christopher L. Warren, Karl E. Hauschild, Mary S. Ozers, Naveeda Qadir, Devesh Bhimsaria, Youngsook Lee, Franco Cerrina, and Aseem Z. Ansari. Specificity landscapes of DNA binding molecules elucidate biological function. *Proceedings of the National Academy of Sciences*, 107(10): 4544–4549, March 2010. ISSN 0027-8424, 1091-6490.

[296] Douglas M. Fowler, Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature Methods*, 7(9):741–746, September 2010. ISSN 1548-7091.

[297] Joan Teyra, Sachdev S. Sidhu, and Philip M. Kim. Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. *FEBS letters*, 586(17):2631–2637, August 2012. ISSN 1873-3468.

[298] Matthew Slattery, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen J. Bussemaker, and Richard S. Mann. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, 147(6):1270–1282, December 2011. ISSN 0092-8674.

[299] UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1): D204–D212, January 2015. ISSN 0305-1048.

[300] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33 (suppl_1):D54–D58, January 2005. ISSN 0305-1048.

[301] Frank Delaglio, Stephan Grzesiek, Geerten W. Vuister, Guang Zhu, John Pfeifer, and Ad Bax. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *Journal of Biomolecular NMR*, 6(3):277–293, November 1995. ISSN 0925-2738, 1573-5001.

[302] S. P. Skinner, B. T. Goult, R. H. Fogh, W. Boucher, T. J. Stevens, E. D. Laue, and G. W. Vuister. Structure Calculation, Refinement and Validation Using CcpNmr Analysis. *Acta Crystallographica Section D: Biological Crystallography*, 71(1):154–161, January 2015. ISSN 1399-0047.

[303] Dmitrij Frishman and Patrick Argos. Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Bioinformatics*, 23 (4):566–579, December 1995. ISSN 1097-0134.

[304] Simon J. Hubbard and Janet M. Thornton. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1), 1993.

[305] Ziheng Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, August 2007. ISSN 0737-4038.

[306] Hiroyuki Kanzaki, Kentaro Yoshida, Hiromasa Saitoh, Koki Fujisaki, Akiko Hirabuchi, Ludovic Alaux, Elisabeth Fournier, Didier Tharreau, and Ryohei Terauchi. Arms race co-evolution of Magnaporthe oryzae AVR-Pik and rice Pik genes driven by their physical interactions. *The Plant Journal*, 72(6): 894–907, December 2012. ISSN 1365-313X.

[307] Miriam Kaltenbach and Nobuhiko Tokuriki. Dynamics and constraints of enzyme evolution. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 322(7):468–487, November 2014. ISSN 1552-5015.

[308] Ranjan V. Mannige, Charles L. Brooks, and Eugene I. Shakhnovich. A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor. *PLoS Comput Biol*, 8(12):e1002839, December 2012.

[309] Chris Todd Hittinger and Sean B. Carroll. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163):677–681, October 2007. ISSN 0028-0836.

[310] Ksenia Pougach, Arnout Voet, Fyodor A. Kondrashov, Karin Voordeckers, Joaquin F. Christiaens, Bianka Baying, Vladimir Benes, Ryo Sakai, Jan Aerts, Bo Zhu, Patrick Van Dijck, and Kevin J. Verstrepen. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Nature Communications*, 5, September 2014. ISSN 2041-1723.

[311] Alissa Rauwerdink, Mark Lunzer, Titu Devamani, Bryan Jones, Joanna Mooney, Zhi-Jun Zhang, Jian-He Xu, Romas J. Kazlauskas, and Antony M. Dean. Evolution of a Catalytic Mechanism. *Molecular Biology and Evolution*, 33(4):971–979, April 2016. ISSN 1537-1719.

[312] Sachdev S. Sidhu, Henry B. Lowman, Brian C. Cunningham, and James A. Wells. Phage Display for Selection of Novel Binding Peptides. *Methods in Enzymology*, 328:333–IN5, January 2000. ISSN 0076-6879.

[313] William G. T. Willats. Phage Display: Practicalities and Prospects. *Plant Molecular Biology*, 50(6):837–854, December 2002. ISSN 0167-4412, 1573-5028.

[314] Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Research*, 36(Database issue): D202–205, 2008. ISSN 1362-4962.

[315] Alex S. Holehouse, James Ahad, Rahul K. Das, and Rohit V. Pappu. CIDER: Classification of Intrinsically Disordered Ensemble Regions. *Biophysical Journal*, 108(2):228a, January 2015. ISSN 0006-3495.

[316] Leo Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001.

[317] Phillip A. Steindel, Emily H. Chen, Jacob D. Wirth, and Douglas L. Theobald. Gradual neofunctionalization in the convergent evolution of trichomonad lactate and malate dehydrogenases. *Protein Science*, 25(7):1319–1331, July 2016. ISSN 1469-896X.

[318] Rafael G. Miranda, Margarita Rojas, Michael P. Montgomery, Kyle P. Gribbin, and Alice Barkan. RNA binding specificity landscape of the pentatricopeptide repeat protein PPR10. *RNA*, page rna.059568.116, January 2017. ISSN 1355-8382, 1469-9001.

[319] Tyler N. Starr, Lora K. Picton, and Joseph W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549 (7672):409–413, September 2017. ISSN 0028-0836.

[320] Andr Zelanis, Pitter F. Huesgen, Ana Karina Oliveira, Alexandre K. Tashima, Solange M. T. Serrano, and Christopher M. Overall. Snake venom serine proteinases specificity mapping by proteomic identification of cleavage sites. *Journal of Proteomics*, 113:260–267, January 2015. ISSN 1874-3919.

[321] Frances H. Arnold. Protein engineering for unusual environments. *Current Opinion in Biotechnology*, 4(4):450–455, August 1993. ISSN 0958-1669.

[322] Prachi Anand, Alison ONeil, Emily Lin, Trevor Douglas, and Mand Holford. Tailored delivery of analgesic ziconotide across a blood brain barrier model using viral nanocontainers. *Scientific Reports*, 5:srep12497, August 2015. ISSN 2045-2322.

[323] Benjamin Schwarz, Kaitlyn M. Morabito, Tracy J. Ruckwardt, Dustin P. Patterson, John Avera, Heini M. Miettinen, Barney S. Graham, and Trevor Douglas. Viruslike Particles Encapsidating Respiratory Syncytial Virus M and M2 Proteins Induce Robust T Cell Responses. *ACS Biomaterials Science & Engineering*, 2(12):2324–2332, December 2016.

[324] Frances H. Arnold. Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie International Edition*, pages n/a–n/a. ISSN 1521-3773.

[325] Stephan C. Hammer, Anders M. Knight, and Frances H. Arnold. Design and evolution of enzymes for non-natural chemistry. *Current Opinion in Green and Sustainable Chemistry*, 7(Supplement C):23–30, October 2017. ISSN 2452-2236.

[326] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. pages 226–231. AAAI Press, 1996.

[327] Fred J. Damerau. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM*, 7(3):171–176, March 1964. ISSN 0001-0782.

[328] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, 13(2):22–30, March 2011. ISSN 1521-9615.

[329] Eric Jones, Travis Oliphant, Pearu Peterson, and others. *SciPy: Open source scientific tools for Python*. 2001.

[330] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[331] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

[332] Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[333] Tara Hessa, Hyun Kim, Karl Bihlmaier, Carolina Lundin, and others. Recognition of Transmembrane Helices by the Endoplasmic Reticulum Translocon. *Nature*, 433(7024):377, 2005.

[334] D. M. Engelman, T. A. Steitz, Goldman, and A. Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 15(1):321–353, 1986.

[335] Catherine H. Schein. Solubility as a Function of Protein Structure and Solvent Components. *Nature Biotechnology*, 8(4):308–317, April 1990. ISSN 1087-0156.

[336] J. Kyte and R. F. Doolittle. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 157(1):105–132, May 1982. ISSN 0022-2836.

[337] William C. Wimley and Stephen H. White. Experimentally Determined Hydrophobicity Scale for Proteins at Membrane Interfaces. *Nature Structural & Molecular Biology*, 3(10):842–848, October 1996.

[338] T. P. Hopp and K. R. Woods. Prediction of Protein Antigenic Determinants from Amino Acid Sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 78(6):3824–3828, June 1981. ISSN 0027-8424.

[339] P. Y. Chou and G. D. Fasman. Empirical Predictions of Protein Conformation. *Annual Review of Biochemistry*, 47:251–276, 1978. ISSN 0066-4154.

[340] Hyun Joo and Jerry Tsai. An Amino Acid Code for -Sheet Packing Structure. *Proteins*, 82(9):2128–2140, September 2014. ISSN 0887-3585.

[341] F. M. Richards. Areas, Volumes, Packing and Protein Structure. *Annual Review of Biophysics and Bioengineering*, 6:151–176, 1977. ISSN 0084-6589.