THE EVOLUTION OF THE S100A4 ENSEMBLE

by

ANTONIA DE ANDRACA SERRANO

A THESIS

Presented to the Department of Chemistry & Biochemistry
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

May 2024

An Abstract of the Thesis of

Antonia de Andraca Serrano for the degree of Bachelor of Science
in the Department of Chemistry & Biochemistry to be awarded June 2024

Title: The Evolution of the S100A4 Ensemble

Approved: *Michael J. Harms, Ph.D.*
Primary Thesis Advisor

Understanding the evolution of proteins is essential for interpreting the complex mechanisms driving biological systems. The idea of intramolecular epistasis, which outlines a situation in which the effects of one mutation in a protein depend on the existence of other mutations in the same protein, is central to this investigation. Proteins' evolutionary paths are shaped by these complex interactions, which also affect the variety and adaptability of their functions. Proteins display an intriguing variety of epistatic patterns, ranging from long-range interactions between far-off sites to dynamic changes triggered by their environments. Evolutionary biochemistry has struggled for decades to provide thorough predictive models for these intricate patterns. Instead of treating proteins as static structures, this study views them as dynamic entities with a variety of conformations.

With a focus on human S100A4 and its ancestral counterparts, AncA4 and AncA2A4, this investigation seeks to shed light on the complex dynamics of ensemble epistasis and its evolutionary implications over time scales. Urea unfolding experiments for S100A4, AncA2A4, and AncA4 quantified free energy. Expected alignment of ΔG values across signals within each protein was not observed, indicating potential oligomerization taking place during unfolding. Observed variations in ΔG values during unfolding between these three proteins suggest potential influences of evolutionary mutations. Additionally, EDTA titrations for calcium binding demonstrated unexpected patterns, indicating a more intricate cooperativity in calcium binding dynamics within the protein ensemble. Nevertheless, a two-state system for binding calcium was observed, where low affinity binding took place first, followed by high affinity binding.

Acknowledgments

I would like to extend my deepest gratitude to Professor Mike Harms for his unwavering support throughout this past year. His guidance and encouragement has been invaluable in shaping my academic and research endeavors. I am also immensely grateful for Natalie Jaeger, a PhD student in the Harms Lab, whose assistance and commitment have significantly contributed to the progress of my work. Her expertise and dedication have been instrumental, and I am truly thankful for her exceptional support and guidance throughout this journey.

I am enormously thankful to the Clark Honors College for their support and the exceptional professors who have provided indispensable guidance during this challenging yet enriching journey. In particular, I would like to express my appreciation to Professor Dr. Carol Paty for her mentorship and wisdom, which have greatly enriched my academic experience.

My heartfelt appreciation goes out to my mom, whose unwavering love, support, and belief in me have been my constant source of strength and motivation. Her wise counsel and encouragement have been instrumental in navigating through the ups and downs of academic life. I am deeply grateful for all of her sacrifices and endless support.

I extend this appreciation to Dr. Andy Dey, whose mentorship and wisdom have made this journey immensely rewarding.

I also owe a big thank you to my siblings, Sofía and Tomás, for being the ultimate hype team during this journey. Their support, combined with their infectious positivity, has been a constant motivator, reminding me that I can achieve anything I set my mind to.

Finally, I also want to express a special thank you to Ella Gutierrez-Garner, Philip Jimenez, and Alina Scott who have been my pillars of strength during exams and busy weeks. I am truly blessed to have you both in my life.

Table of Contents

List of Figures

List of Tables

Introduction

Proteins are the fundamental building blocks of life; they are the molecular machines that organize vital processes in organisms. Proteins are big, intricate molecules made up of chains of amino acids that have been carefully folded into distinct three-dimensional shapes (Dobson, 2003). The functions of proteins are determined by these structures, which perform essential functions for the life and survival of cells (Alberts et al., 2002). Essentially, proteins are the instrumentalists in the symphony of life, each contributing uniquely to the vast orchestra of biological activities.

They are essential to almost every aspect of biology, serving as biological reaction catalysts, controlling cellular functions, and giving tissues and cells structural support. For example, enzymes, a type of protein, catalyze chemical reactions necessary for metabolism, allowing cells to extract energy from nutrients and synthesize essential molecules (Alberts et al., 2002). Proteins can also serve as transporters, moving molecules such as oxygen and nutrients across cell membranes (Alberts et al., 2002). Furthermore, by identifying and eliminating foreign invaders like bacteria and viruses, proteins are also essential for immunological function (Alberts et al., 2002).

Protein Functions Encoded by Ensembles of Structures

Understanding protein evolution is immensely beneficial as it provides a historical roadmap of how proteins have acquired their diverse functions. The story of protein function involves the dynamic interaction of structural ensembles. In the context of proteins, an ensemble is a dynamic collection of several conformations or structural states that a protein molecule can take on (Bonomi et al., 2017). The innate flexibility of proteins enables them to adjust and react to many environmental stimuli, contributing a subtle dimension to their multifaceted functional adaptability (Bonomi et al., 2017). Therefore, by understanding how proteins interact dynamically within these structures and by studying protein evolution, we are able to uncover the fundamental rules that govern how proteins work and adapt.

This knowledge is crucial in the design of experiments focused on probing protein dynamics, predicting functional changes based on evolutionary patterns and engineering proteins with tailored properties for specific uses in medicine, biotechnology, and beyond.

Proteins Evolve

Amazingly, protein functions also evolve over time. Proteins are linear chains of amino acids. The amino acid sequence encodes structure and ensemble. Mutations, defined as changes in the DNA sequence that result in modifications to the amino acid sequence of a protein (Strokach et al., 2019), can alter the function of a protein. If beneficial—or at least not deleterious—the copying error can get passed on. Understanding how proteins evolve is important for advancements in medicine, development of new technologies, and gaining deeper insights into the fundamental processes of life.

One poorly understood aspect of protein evolution is the interplay between the evolutionary process and protein thermodynamic ensembles. This theory, called ensemble evolution, captures the changes in the protein structural diversity due to changes in amino acid sequence. The dynamic variety of expressions an artist might present on stage is analogous to how an ensemble changes over time, reflecting the flexibility of the protein. A protein's functionality is determined by the changes within its structural ensemble over the course of its evolutionary history, much like an actor's performance is molded by the roles they play over their career.

Understanding ensemble evolution is essential to understanding how proteins evolve new functions (Harms, 2023. Grant, Unpublished). It enables us to understand how proteins' dynamic functions in cellular processes are influenced by their changing structural landscapes.

Exploring the importance of Ensemble Epistasis for evolution

The fact that proteins exist as ensembles of structures opens the opportunity for ensemble epistasis—in which several mutations cooperatively alter a protein's behavior and function. The main topic of this research will center around ensemble epistasis, which appears as a crucial aspect of the evolutionary script.

Figure 1 below illustrates the concept of epistasis, which plays a crucial role in understanding the complexities of protein evolution. Imagine a scenario where a protein undergoes a mutation, transitioning from an ab phenotype → to one labeled as Ab, undergoing an "A" mutation, or from an ab phenotype to → one labeled as aB with a "B" mutation. Individually, these mutations may not impact the protein's function significantly (as demonstrated by the protein retaining its purple color and maintaining a square shape) (Figure 1). However, if both mutations were to be present simultaneously, they could potentially affect the protein's functionality together. This demonstrates the phenomenon of epistasis: the combined effect of mutations differs from the sum of individual effects (this can be observed in the figure in where the AB mutation leads to an orange circle; i.e. change in phenotype and protein functionality) (Figure 1) (Ortlund et al., 2007). This interplay between mutations within a protein ensemble adds a layer of complexity to predicting mutation effects and underscores the importance of adopting an ensemble perspective in studying protein evolution (Harms, 2023 & Modi et. al. 2021).

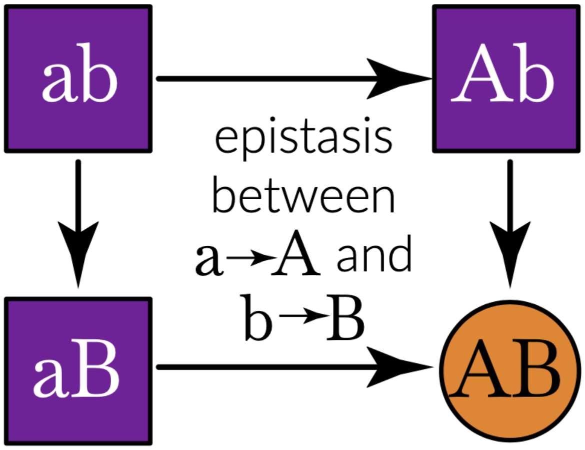


Figure 1. Epistasis. Squares and circles are different proteins. Different proteins can have different properties (purple square or orange circle). The sequence of each protein is shown as letters on each protein. The a→A and b→B mutations exhibit epistasis because they have no effect alone but together change the function of the protein.

Understanding the intricate nature of epistasis, characterized by complex interactions among mutations within proteins, is vital across various scientific domains. Notably, comprehending epistasis is crucial due to its significant impact on a protein's functionality, with mutations either enhancing or weakening individual effects (Morrison et al., 2021). This information is essential for developing focused treatments in the domains of molecular biology, medicine, and biotechnology. For example, by better understanding epistasis, new avenues may be made available for the development of drugs and therapies that target proteins associated with disease (Guerrero et. al, 2019 & Schenk et. al. 2013).

Ensemble epistasis refers to the phenomenon where mutations exhibit epistasis because they each have effects on the conformational ensemble populated by the protein. Epistasis of this sort has been observed previously (Morrison, 2021); however, its prevalence and importance for protein evolution remains poorly understood.

Studying Ensemble Evolution with S100A4

Understanding the intricate relationship between ensemble dynamics, epistasis, and evolutionary outcomes poses significant challenges in molecular biology and biophysics. The complexity of these subjects necessitates finding a protein with an ensemble complexity that strikes a balance—sufficiently intricate to be meaningfully simplified yet not overly complex as to be unmanageable. Therefore, for my investigation, I will delve into the role of human S100A4 and its ancestral relatives, AncA4 and AncA2A4, in shedding light on the intricate dynamics of ensemble epistasis over evolutionary time scales.

S100A4 proteins, also referred to as Mts1 or metastasis, are important modulators of cell motility. They accomplish this by attaching themselves to non-muscle myosin-IIA (NMIIA) and preventing actin filament production. S100A4 plays a critical role in the development of cancer and the spread of metastatic disease because this disturbance ultimately promotes cell invasion, motility, and metastasis. This protein's utility as a malignancy marker for a variety of cancer types and its participation in anti-cancer drug clinical studies highlight its clinical significance.

Additionally, S100A4 is especially well suited for this investigation due to its well-characterized conformational ensemble. The protein displays distinct conformations when calcium is present and changes from a closed to an open conformation when calcium binds as can be seen in Figure 2 below.

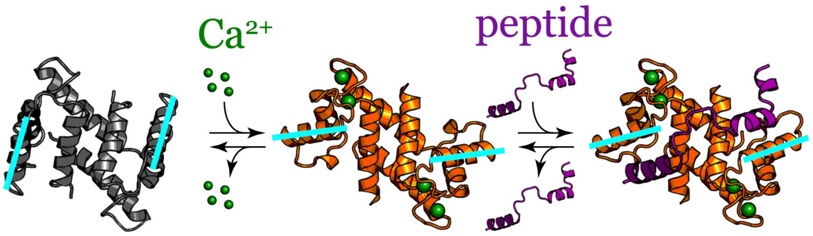


Figure 2. Simplified S100A4 Protein Ensemble. Structures depict S100A4 in three distinct conformations: unbound in the absence of calcium (grey structure), bound to calcium (middle structure), and bound both to calcium and peptide (furthest right structure). This is a visual representation of the structural change undergone by S100A4 in response to calcium binding and subsequent interactions with target peptide. Its unbound state in the absence of calcium shows a closed conformation, upon calcium binding, the protein transitions to an open state, exposing a hydrophobic patch. Here S100A4 adopts the specific conformation suitable for the protein-peptide interaction and binds peptide.

The S100A4 protein is also a great candidate for in-depth structural and dynamic investigations due to its small size, solubility, and ease of expression in *E. coli*. It is also amenable to biophysical characterization.

The evolutionary origin of S100A4 adds another level of complexity and intrigue to the research being conducted, on top of its functional ramifications. Previous work conducted in the Harms lab at the University of Oregon has allowed for the study of ancestor proteins of S100A4, which has illuminated its evolutionary history (see Figure 3 below for a phylogenetic tree of S100s). These investigations yielded intriguing discoveries, including temporal fluctuations in peptide binding specificity and the identification of unique protein-peptide interactions. In addition to improving our knowledge of S100A4's functional adaptations, an understanding of its evolutionary dynamics offers important insights into the protein's larger evolutionary context and its role in the biological system.

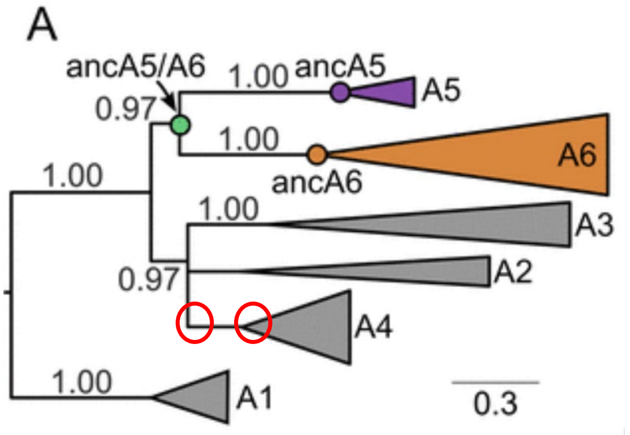


Figure 3. Phylogenetic tree of S100s, showing link between S100A4 to AncA4 & AncA2A4. AncA2A4 is represented by the circle furthest to the left in the diagram, while AncA4 is depicted by the circle closer to the right. The wedges within the tree represent multiple lineages. All A4 variants are grouped together in a single wedge. The numbers adjacent to each node represent statistical support values. A value approaching 1 signifies high confidence, indicating robust support for the depicted evolutionary relationship.

While a significant amount of research has been conducted showing that naturally occurring ensembles lead to various types of epistasis, the influence of changes in the conformational ensemble on the evolution of ancestral S100A4 proteins remains an unanswered question. Thus, the main questions driving my research are: To what extent did changes in the conformational ensemble influence the evolution of ancestral S100A4 proteins? Is Ensemble Epistasis observable in human S100A4, and if so, what does this tell us about the protein's evolutionary history?

Investigating these questions regarding the evolution of the S100A4 ensemble is highly valuable from a practical standpoint and provides important insights into the many functions of this important protein. Comprehending the evolutionary trajectory of the S100A4 protein, which contributes to cancer malignancy and metastasis and is important for human health, can offer new insights into structure-activity correlations and possible therapeutic approaches. Furthermore, identifying the impact of modifications in the S100A4 ensemble is essential for constructing epistasis prediction models and enhancing protein engineering approaches.

Methods/Experimental Approach

To attempt to answer my questions, the first steps involved conducting a meticulous characterization of the human S100A4 protein. Drawing from existing literature on the purification and structural ensemble of this protein, I set out to characterize its ensemble, which comprises five distinct conformations (see Figure 4). The exploration into the evolution of ensemble epistasis commences with a thorough examination of each structure within the protein's ensemble. This comprehensive analysis encompasses understanding the protein in its unfolded state (depicted as a squiggly line in Fig. 4), folded inactive state (grey circle), folded active state, shown as an orange circle with four white pockets (since the protein exists as a dimer, composed of two monomers with two sites per monomer), followed by the active folded state when bound to calcium, represented by green circles, and when bound to peptide, depicted as a purple bar. This exhaustive study is paramount for revealing the dynamic intricacies of the protein under diverse conditions, thus laying the foundation to evaluate ensemble epistasis for the S100A4 protein and its ancestral counterparts. Additionally, Fig. 4 includes an energy diagram illustrating how the protein exhibits a preference for binding to calcium and peptide, further enhancing our understanding of its structural dynamics and functional preferences.

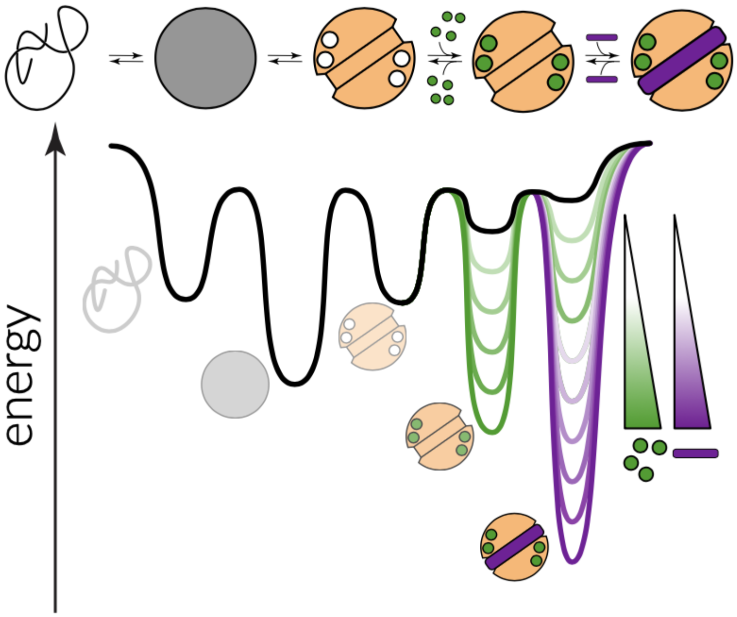


Figure 4. S100A4 Protein Ensemble and Energy Diagram. Illustration of the dynamic behavior of the S100A4 protein ensemble in response to calcium binding (depicted as small green circles) and peptide interaction (represented by purple bars). The protein's stability increases with higher calcium concentration and peptide binding. Protein ensembles starts with the protein being unfolded, then folded and inactive, then active, then binds calcium, and finally peptide. The energy diagram accompanying the ensemble showcases the energetic landscape of the protein's conformational shapes. Energy in this context, refers to the thermodynamic stability or the amount of energy held by a specific conformation. The squiggles and structures in the diagram represent different energy levels and conformational states for the protein. Lower energy levels correspond to more stable conformations, while higher energy levels indicate less stable or transitional states. The protein gains stability with more calcium (small green circles) and with increased peptide (purple bar).

Previous Literature

Previous literature has highlighted the profound impact of protein evolution on biological processes, from pathogenesis to drug resistance and cancer progression, emphasizing the necessity of a complex comprehension of these molecular changes (Bloom et al., 2010). This demonstrates the critical role that proteins play in addressing different challenges and opportunities in the biological world (Campbell et al., 2016).

Intramolecular epistasis is a phenomenon that is essential to the evolutionary story. It occurs when a mutation inside a protein modifies the action of another, affecting protein activities and forming evolutionary pathways (Bloom et al., 2010). The complexity of predicting intramolecular epistasis stems from its diverse and non-intuitive manifestation, such as mutations that are independent of the protein environment and those that are separated from one another in a protein structure (Otwinowski, 2018). Therefore, because of this, creating predictive models for intramolecular epistasis has proven difficult because of the variety of ways it can appear, including high-order epistasis involving three or more mutations and/or epistatic hotspots (Bershtein et al., 2006; Harms & Thornton, 2013). One major obstacle to comprehending the underlying dynamics of protein evolution is the absence of a unifying mechanism to explain and anticipate these inconsistent patterns (Duelli et al., 2014).

To address these challenges, I opted to investigate two ancestral forms of my “model” apo protein (S100A4), AncA2A4 and AncA4. My goal was to reconstruct their respective ensembles, a task for which I drew upon previous studies measuring calcium binding to the S100A4 protein, as well as peptide binding (Duelli, et al. 2014). I leveraged existing purification protocols to refine the method for purifying these ancestral proteins for the first time (Wheeler, et al., 2017). Moreover, I also focused some of my time purifying eight mutants of the S100A4 protein that show promise in participating in ensemble epistasis. These mutations are anticipated to have varying effects on different conformations within the ensemble. Specifically, six mutations are expected to impact the energy of the open form (F27A, M59A, F45A, E52A, K35A, and F55A), while two mutations are predicted to affect the energy of the closed form (L58A, M85A). We plan to introduce these mutations individually and thoroughly characterize their effects in the future once the human S100A4 and two ancestor ensembles are thoroughly defined. This will serve as a future path from which to continue the research on how ensemble epistasis influences the evolution and functional dynamics of the S100A4 protein.

Approach

The central focus of my research lies in characterizing the ensemble for S100A4, AncA2A4, and AncA4. To accomplish this, measuring the four free energies of the S100As ensembles is necessary (Harms, 2023 Grant, Unpublished). ΔG is a measure of the energy shift that occurs during processes such as protein folding or binding. It is useful in explaining the energy differences that occur before and after particular events. Essentially, ΔG can be thought of as the energy differential that exists before and after a specific occurrence, which helps determine how favorable or unfavorable a change is (Forsén & Linse, 1995). Hence, Figure 5 illustrates the precise energy needed to change a protein from its unfolded to its folded state, and from its inactive to its active form, before it binds to anything. Furthermore, it needs a certain amount of energy to bind to calcium (Ca^{2+}) and peptide. It will therefore be necessary to measure each of these ΔGs, as indicated by the red circles in Figure 5.

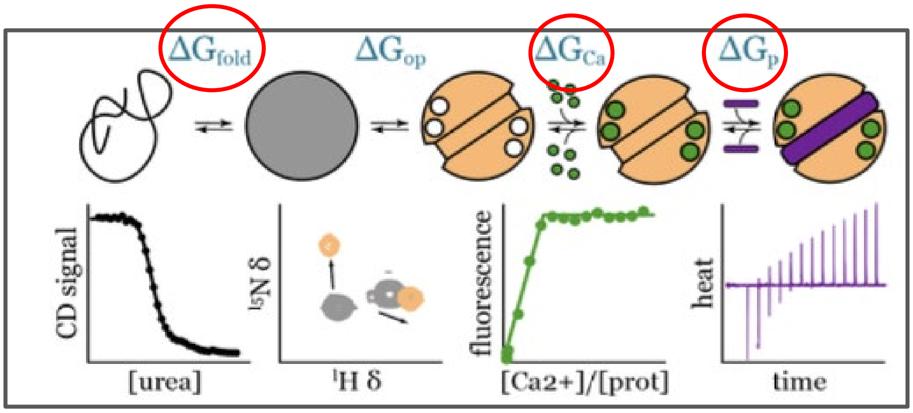


Figure 5. ΔG Measurements and Mathematical Modeling of S100A4 Ensemble. The figure provides a detailed analysis of the free energy changes (ΔG) associated with each conformation in the S100A4 protein ensemble.

The free energies that will be characterized are the ones circled in red: folding (ΔG_{fold}), calcium ion binding (ΔG_{Ca}), and peptide binding (ΔG_p). The measurements of these energies will be conducted using a combination of advanced biophysical techniques, including chemical denaturation for ΔG_{fold} using circular dichroism (CD), intrinsic fluorescence analysis with CD for ΔG_{Ca}, and isothermal titration calorimetry for ΔG_p (Shown but not explored in this paper, ΔG_{op}. Will be studied in the future with nuclear magnetic resonance (NMR) spectroscopy).

Utilizing a combination of biophysical instruments (such as Circular Dichroism and Isothermal Titration Calorimetry) allowed me to carry out these measurements. However, before diving into the measurement of free energies within each protein's ensemble, the initial hurdle involved purifying all the proteins (S100A4, AncA2A4 & AncA4).

Protein Purification

Protein purification is a complicated process essential for understanding protein structure, function, and interactions. My thesis required the perfecting of the purification protocol for the S100A4 protein, along with its ancestors. Once I was able to effectively achieve this, I extended this procedure to include the purification of both ancestor proteins (AncA2A4 and AncA4) as well as eight other S100A4 mutants.

Optimizing the Purification Protocol

I was able to successfully develop and optimize a procedure for purifying my S100A4 proteins by taking inspiration from the paper "Human S100A5 binds Ca^{2+} and Cu^{2+} independently," and its well-established purification protocol (Wheeler, et al., 2017).

Before initiating the purification process, the initial step involved obtaining plasmid sequences for our proteins, which were ordered from GenScript. All the protein sequences underwent a substitution of cysteine with serines and the deletion of the last 13 amino acids. These modifications were done to the proteins because these C-terminal residues have a positive charge and bind to the negatively charged residues of free Ca^{2+} and the binding weakens significantly when Ca^{2+} is present (Duelli et al., 2014). Consequently, the Δ13 mutant shows a higher affinity for Ca^{2+} and remains consistently bound at levels typical of cellular concentrations, which is extremely useful for our ensemble studies as they ensure that calcium can easily bind and remain bounded throughout experiments. These plasmids were transformed into XL10 Gold Competent Cells to amplify the plasmid and generate new DNA stocks. XL10 Gold cells have been engineered with specific mutations to enhance the uptake and maintenance of foreign plasmid DNA, ensuring high efficiency (Hte) (Transformation Efficiency Cells. Agilent). Once the plasmids were successfully transformed, they were further transformed into Rosetta (DE3) pLysS cells. This transformation into Rosetta cells allows for the induction of expression of our desired protein for purification using pET vectors (Tegel et al., 2010). Selection for transformants was performed on media containing chloramphenicol and the respective resistance marker of the transformed plasmid, in this case, kanamycin antibiotic resistance. Rosetta cells, derived from BL21 cells, possess mutations that prevent the degradation of heterologous proteins, enhance plasmid stability, and facilitate efficient protein expression upon induction with IPTG (Francis et al., 2010), which is why they served as great hosts for our protein purification studies.

Bacterial cultures were grown from these colonies once the plasmids had been introduced into Rosetta cells. This involved inoculating 5 mL of Luria Broth (LB), a nutritionally rich medium primarily used for bacterial growth, supplemented with 5 uL of Kanamycin (1000x concentration) and 5 uL of Chloramphenicol (1000x concentration) and allowing them to grow overnight at 37°C at 225-250 rpm. The following morning, the cultures were transferred into larger volumes of LB medium (1.5 L) containing 1.5 mL of Kanamycin and Chloramphenicol to maintain the same antibiotic concentration throughout. The cells were then induced with 1 mM IPTG and 2 mM $CaCl_2$ when they reached an optical density at 600 nm (OD600) of approximately 0.8, which took place approximately 3.5 hours post-inoculation. The cultures were subsequently incubated at 16°C at 225-250 rpm overnight. Finally, the cultures were spun down at 5000 rpm for 20 minutes. The cell pellets were subsequently collected, the supernatant was poured off, and the pellets were stored in the freezer.

The purification process began with lysing bacterial pellets, a crucial step to release proteins of interest. Lysis buffer composition and conditions were meticulously optimized for efficient cell disruption and protein solubility. The lysis buffer used had a pH of 7.4, a concentration of 25 mM Tris, and 300 mM NaCl. The first step in lysing cell pellets was adding one protease inhibitor pellet, 0.5 uL of lysozyme/mL, and 0.5 uL of DNaseI/mL to 3.5 mL of lysis solution per gram of cell pellet. After vortexing to resuspend the cells, they were sonicated in ice for 6 minutes with a 50% duty cycle (2 seconds on, 2 seconds off). The lysed mixture was then centrifuged at 15,000 rpm for 1 hour, yielding a supernatant containing the desired proteins. 2 mM $CaCl_2$ was added to the recovered supernatant.

After the cell lysis procedure, the purification process was carried out using two primary chromatography techniques on a Fast Protein Liquid Chromatography (FPLC) system: Hydrophobic Interaction Chromatography (HIC) and Ion Exchange Chromatography (IEC). HIC uses differences in hydrophobicity to separate proteins, whereas IEC uses charge differences. The supernatant underwent two separate FPLC runs to finish the purification process and achieve optimal protein purity. The initial step involved a day of HIC Gradient Elution, in which the supernatant was flushed through two 5 mL Phenyl FF HIC columns. It was first washed with 50 mL of HIC buffer A (25 mM Tris, 100 mM NaCl, 2 mM CaCl2, pH 7.4), followed by a step elution into 50 mL of HIC buffer B (25 mM Tris, 100 mM NaCl, 5 mM EDTA, pH 7.4), while collecting 4 mL fractions. Subsequently, an SDS-PAGE gel was run to evaluate the purity of these fractions. The protein fractions that showed the desired band at 10058.32 Da (MW of S100A4 protein) on the SDS gel, indicating the presence of our protein of interest, were combined. SDS gel denatures proteins and ensures they all become negatively charged, facilitating migration towards the positive electrode based on molecular weight for accurate identification and isolation (Thermo Fisher Scientific. Overview of Electrophoresis). After this, the combined fractions were dialyzed overnight in a 5 L container with IEC A buffer.

The following day, the dialyzed fractions were subjected to Ion Exchange Chromatography (IEC), passing through two 5 mL DEAE IEC columns. The process began with a wash using 50 mL of IEC buffer A (25 mM Tris, 0 mM NaCl, pH 7.8), followed by a step elution into 50 mL of IEC buffer B (25 mM Tris, 500 mM NaCl, pH 7.8), with fractions collected every 4 mL. Again, an SDS-PAGE gel was utilized to assess fraction purity post-purification. Specific buffer compositions tailored for HIC (HIC buffer A/B) and IEC (IEC buffer A/B) were crucial for preserving protein stability and enhancing binding affinity, contributing to the efficiency of the purification processes. Once, the IEC fractions were determined to contain our target protein, they were combined and flash-frozen using liquid nitrogen to maintain their integrity and subsequently stored at -80 °C.

Challenges Faced During Purification

The variation in elution fractions for each ancestral protein presented a significant obstacle to the purification process's optimization. In contrast to the human S100A4 purification, when purification procedures were run for AncA2A4 or AncA4 it resulted in some ancestors eluting earlier and others later, making it challenging to reliably identify the proper fractions to dialyze or subsequently freeze. To overcome this difficulty, careful documentation was necessary, including taking gel images and chromatographs, to confirm the efficacy of the protocol and track advancements over time. Despite encountering these challenges and undergoing numerous trial-and-error iterations that led to significant protein loss during purification, we successfully employed the refined protocol to purify one double mutant of S100A4 along with eight additional single mutants of S100A4. The journey of perfecting the protein purification protocol for S100A4 and its ancestors, highlights the meticulous attention to detail and scientific rigor essential in biochemistry research. This process of the research conducted demonstrates the great amount of labor needed to take place even before experiments are carried out to assess changes in free energy within a protein's ensemble.

Instruments Used to Measure ∆Gs: Isothermal Titration Calorimetry & Circular Dichroism

The investigation into the evolution of ensemble epistasis initiates with the comprehensive characterization of each structure within the protein's ensemble. We began by assessing S100A4's calcium-binding and peptide-binding affinity.

In order to understand the thermodynamics of these binding events we measured the heat changes associated with binding reactions, using Isothermal Titration Calorimetry (ITC). In the experiments we conducted, we had one solution containing the ligand (either peptide or calcium) and gradually injected it into another solution containing the binding partner (protein) while the temperature was kept constant (Lewis & Murphy, 2005). The changes in heat associated with the binding process were measured by recording the heat released or absorbed during each injection. This data calculated important characteristics such as the interaction's enthalpy, stoichiometry, binding constants, and entropy (Lewis & Murphy, 2005).

Unfortunately, our initial experiments, involving titration of peptide (specifically the = peptide with amino acid sequence: Acetylation - DAMNREVSSLKNKLRR - Amidation and a Molecular Weight: 1958.26 g/mol) into our protein or EDTA to assess calcium binding affinity (EDTA is a chelating agent that binds to calcium), yielded very poor results. We selected this peptide as our initial ligand due to its previously established binding partner sequence, derived from the C-terminal portion of the MIIA rod, which has been shown to interact with S100A4 (Malashkevich, et al, 2008). However, binding to this peptide did not result in high enough heat of binding to proceed with this line of experimentation. We spent several weeks adjusting protein and peptide concentrations in an attempt to obtain clearer and more interpretable data. Other ligands or buffers may make this experiment more tractable, but other avenues proved more fruitful for the time being. Therefore, considering the significant cost of peptide, we opted to transition to urea melts to study the folding and unfolding dynamics of the human S100A4 ensemble instead (the first ΔG of Fig. 5).

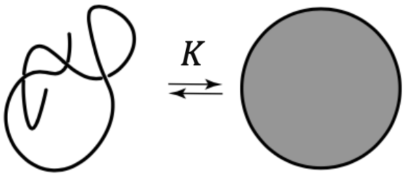
To measure the unfolded and folded states of our protein and determine the ΔG values, we used Circular Dichroism (CD). CD detects the differential absorption of circularly polarized light to characterize the structural features of a protein (Greenfield, 2006). Here I conducted two distinct experiments for the S100A4 protein. First, I titrated in 10 M Urea to measure its unfolding, as urea is a denaturant that disrupts protein structure by weakening hydrophobic interactions and hydrogen bonds. The second experiment involved titrating 2 mM EDTA into a sample of protein already bound to calcium. This procedure allowed me to observe the calcium-binding behavior of the protein and assess its affinity for calcium ions. I measured four signals during these experiments: a Far-CD signal at 222 nm, characteristic absorption of α-helical structures, and a Near-CD signal at 282 nm, characteristic absorption of β-sheet conformations. Additionally, I monitored two fluorescence signals. One at 235 nm (phenylalanine fluorescence) and another at 282 nm (tyrosine fluorescence). The changes in fluorescence observed in these amino acids indicate changes in the overall structure as its unfolding or binding to calcium.

The biggest hurdle with these CD experiments revolved around optimizing the standard conditions for running both the EDTA titrations and melts. Our primary challenge was figuring out how to balance the concentration of proteins such that it was high enough to observe unfolding and binding events but not so low that we couldn't get any useful signals or measurements. While it might seem intuitive to use a high concentration to ensure the collection of a strong, in this case, we opted for less protein. This decision was aimed at ensuring that the urea melt could effectively unfold the protein at each titration without encountering delays due to an overwhelming concentration of protein needing to unfold. We finally succeeded in obtaining dependable and consistent findings by conducting the EDTA titrations and melting at a protein concentration of 10 μM (monomer concentration) after much trial and error.

Characterization of S100A4

To assess the stability of S100A4, we denatured it with urea and evaluated changes in structure through the use of CD spectroscopy and fluorescence. This allowed us to gain a comprehensive understanding of the thermodynamic stability and structural dynamics of the protein. 10 M Urea with 10 μM denatured protein was titrated into a sample of 10 μM protein existing in 5 mM EDTA, 25 mM Tris, 100 mM NaCl, pH 7.4.

Key observations from these experiments after model fitting include identifying a two-state system with a singular K value for the folded to unfolded transition. This observation is supported by the absence of an intermediate state during the folding of each studied protein; instead, they transition directly from the folded to the unfolded state, which explains the distinct S-shaped curves observed in the data (Figures 7-9). In the context of studying protein unfolding, the equilibrium constant (K value) represents the ratio of the concentrations of the folded [F] and unfolded [U] states of the protein at equilibrium in the presence of urea. See Figure 6 below.



$$K = \frac{[F]}{[U]}$$

Figure 6. Equilibrium Transition of Protein Folding. Visual representation of the equilibrium transition of protein folding, showing the structural transformation from the unfolded state to the folded state and providing the breakdown of the K value.

We fit our denaturation data to a two-state unfolding model and extracted a ΔG value of 13.2 +/- 2.97 kcal/mol for the S100A4 protein. These findings, which can be seen in Figure 7 are supported by quantitative ΔG values displayed in Table 1. They provide insight into the folding and unfolding dynamics of our protein of study, and lay the groundwork necessary for comparison to its ancestors.

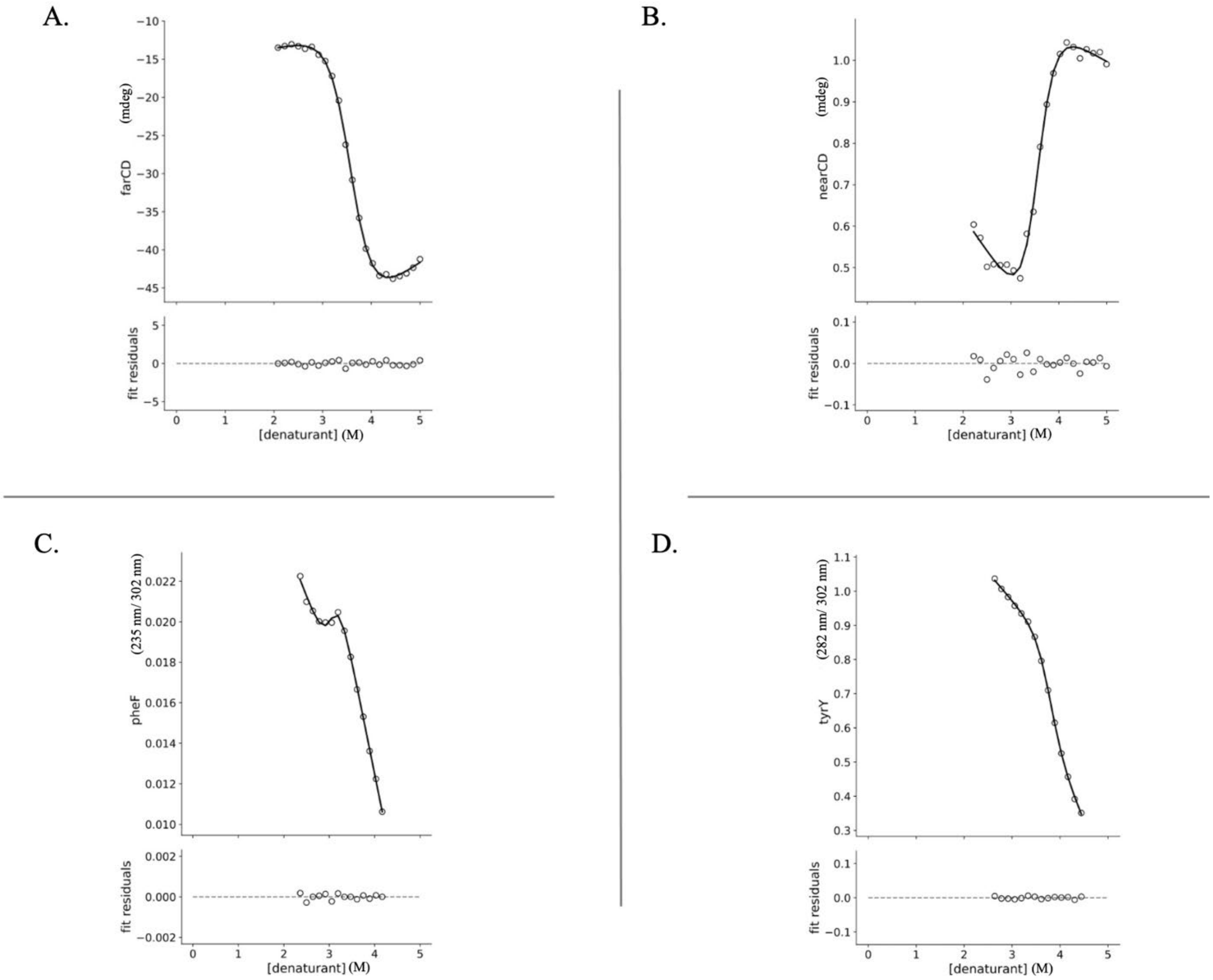


Figure 7. CD Signals for S100A4 Protein. (A) Far CD signal (collected at 222 nm) showing data points up to 2 M urea concentrations. (B) Truncated Near CD signal (Collected at 282 nm) showing data points up to 2 M urea concentrations. (C) Phenylalanine Fluorescence CD signal (Collected at 235 nm) showing data points between 2 M and 4 M urea concentrations. (D) Truncated Tyrosine CD Signal (Collected at 282 nm) showing data points between 2.5 M to 4.5 M urea concentrations.

(A) Far	**ml estimate**	**Lower 95**	**Upper 95**
ΔG (kcal/mol)	9.80 +/- 1.95	7.838268	11.74074
m	-2.73 +/- 0.54	-3.27656	-2.18709

(B) Near	**ml estimate**	**Lower 95**	**Upper 95**
ΔG (kcal/mol)	12.0 +/- 109	-97.0835	121.0963
m	-3.38 +/- 30.5	-33.9012	27.14304

(C) Phenylalanine	**ml estimate**	**Lower 95**	**Upper 95**
ΔG (kcal/mol)	16.66 +/- 3.0516 x10^4	-30499.3	30532.61
m	-5.41 +/- 9.68x10^3	-9686.18	9675.355

(D) Tyrosine	**ml estimate**	**Lower 95**	**Upper 95**
ΔG (kcal/mol)	14.4 +/- 554	-539.159	568.0064
m	-3.8 +/- 1.5 x10^2	-158.199	150.5239

Table 1. S100A4 Protein ΔGs from truncated CD Signals Model Fit

The data shown in Figure 7 and Table 1 above are limited to values within the range of approximately 2 M to 4 M Urea concentrations, aimed at improving the clarity of the raw urea melt data during the model fitting process. In all panels of Figure 7 (A-D), the bottom graph depicts the residuals, which represent the variance between the model and each data point. The aim is to ensure that these residuals are randomly distributed around the zero line, indicating a good fit of the model to the data. The fit result for each parameter is represented by the ml estimate column in each of the tables. The 95-percent confidence interval for the model fit to the data is indicated by the Lower 95 and Upper 95 columns. These values give conservative estimates of how tightly the data constrains the model parameter. The difference between the ml estimates and the Upper 95 or Lower 95 values is used to determine the uncertainty of the ml estimations. Standard deviations were then derived from the average ml estimates. Table 1 also includes the slope of a linear regression model as denoted by the "m" value. From this data, I successfully extrapolated an accurate ΔG from the far UV signal. However, the data from the other signals for S100A4 did not sufficiently constrain the model, preventing me from obtaining precise ΔGs, as evidenced by the significant uncertainties in Table 1 (shown by the difference between the ml estimate for the Upper 95 and the Lower 95 values).

To further characterize S100A4, EDTA titrations were carried out to measure the affinity of calcium-binding for the protein. These experiments entailed titrating in a 2 mM EDTA solution containing 10 μM denatured protein with 100 μM CaCl2, 25 mM Tris, and 100 mM NaCl concentrations into a sample of folded protein at a concentration of 10 μM also in 100 μM CaCl2, 25 mM Tris, 100 mM NaCl. In other words, these experiments allowed for the measurement of the third ΔG value in the protein's ensemble. See Figure 5. The data collection process mirrored that of the Urea Melts, with detailed logging of EDTA concentration during titration and recording of the same 4 CD signals for the sample. The next step was also followed by a model-fitting process.

To define the system and depict the calcium-binding process, a scheme was formulated. This scheme outlined four sites on an S100A4 dimer protein competing for calcium against EDTA. The scheme assumed a sequential binding of each calcium ion to each S100A4 binding site. Therefore prior to analyzing experimental data, an expression was derived that could calculate the concentration of every chemical species in the scheme using only equilibrium constants and the total concentrations of EDTA, S100A4, and calcium (Harms, 2023 EDTA-S100A4 Competition Scheme, Unpublished). This involved deriving expressions for the total concentrations of EDTA, S100A4, and calcium in terms of chemical species (Harms, 2023 EDTA-S100A4 Competition

Scheme, Unpublished). The equilibrium constants for the system were defined to reflect the affinity of each binding event, allowing for a nuanced understanding of the binding process. It was also assumed that there were two high-affinity sites (1 and 2) and two low-affinity sites (3 and 4) within the S100A4 dimer (Harms, 2023 EDTA-S100A4 Competition Scheme, Unpublished). The chemical reaction scheme is as follows:

$$E + A \cdot C_4 \rightleftarrows E + A \cdot C_3 + \mathrm{C} \rightleftarrows E + A \cdot C_2 + 2\mathrm{C} \rightleftarrows E + A \cdot C_1 + 3\mathrm{C} \rightleftarrows E + A + 4\mathrm{C} \rightleftarrows 4E \cdot C + A$$

This scheme describes the complex interplay between EDTA (E), calcium (C), and the S100A4 dimer (A), where each step represents a different binding event with its own equilibrium constant. From here, using this scheme, a model was fitted to the raw data for the S100A4 EDTA titration.

The S100A4 characterization serves as the foundation for comparing the two ancestral proteins, AncA2A4 and AncA4. Therefore, analyzing the ΔG values from identical experiments on these ancestors will offer relevant insights into the evolution of the S100A4 ensemble.

Results

Quantifying Unfolding and Folding Free Energy in Ancestor Proteins: A Comparative Analysis with S100A4

The same urea unfolding experiments were conducted with both ancestral proteins under the same conditions. CD measured protein changes during unfolding. For both AncA2A4 and AncA4, data truncation was also necessary. Data points were truncated ranging from 1M to 4 M urea.

The objective of this study was to detect a change in the behavior of the S100A4 protein across all four signals. Ideally, we aimed to witness a consistent alignment of ΔG values among all 4 of the signals for each of the three proteins: S100A4, AncA2A4, and AncA4. The ultimate goal was to overlay each signal type, such as the Far CD and/or Near CD signals, for all three proteins with the hopes to observe if the unfolding switch occurred uniformly at similar ΔG values. This analysis would have allowed us to determine if evolutionary changes had an impact in this part of the protein ensemble (Figure 5, first ΔG). However, our results did not meet these expectations.

While the characterization of S100A4 yielded signals similar in shape, and while some of the ancestor's signals matched the S-shaped shape to that of S100A4, none of the signals shared any quantitative similarity with the ΔG values collected. While a difference between proteins is expected, the observed variation within each protein's ΔG signals was unexpected (Figures 8 & 9). The data from each of the signals collected for the ancestors did not constrain the model well enough to obtain accurate ΔG values. This is evidenced by the massive uncertainties in tables 2 and 3, as shown by the large differences between the Upper and Lower 95 values. AncA24 had an average ΔG of 4.9 +/- 3.18 kcal/mol and AncA4 had an average ΔG of 7.8 +/- 8.29 kcal/mol. These results suggest that more replicates are needed to confirm the unfolding behavior of each protein. It is possible that each protein exhibits a different average ΔG for unfolding, indicating that evolutionary mutations have indeed influenced the protein's folding and unfolding processes. It is also possible that oligomerization could be taking place within each specific protein during unfolding and could account for the differences in ΔGs. Figures 8 & 9 below show the limited similarity AncA2A4 and AncA4 signals share to those of S100A4, further highlighting the variability in ΔG values observed and reported in tables 2 & 3. The data's variability highlights how difficult it is to make firm conclusions from it.

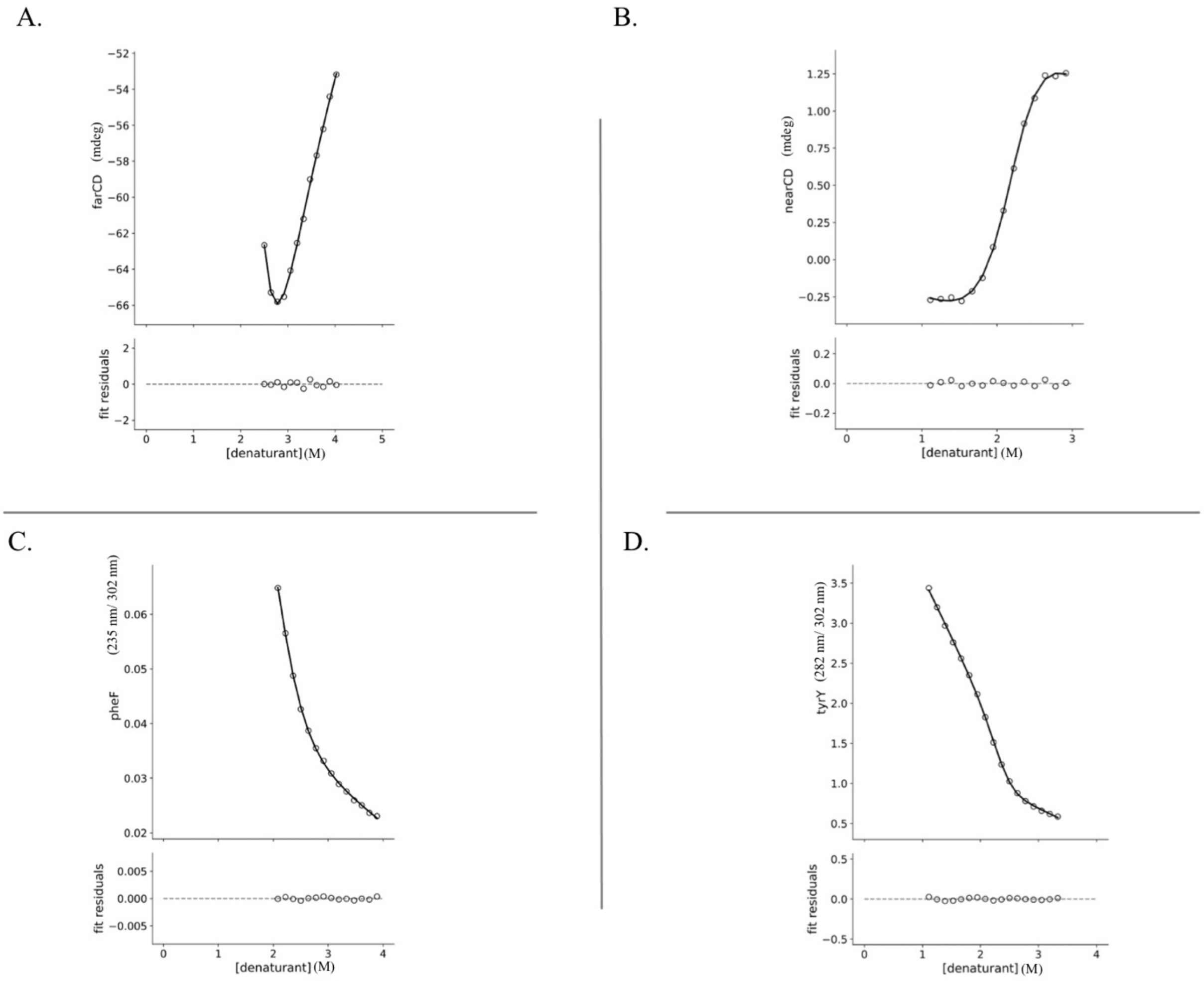


Figure 8. CD Signals for AncA2A4 Protein. (A) Far CD signal (collected at 222 nm) showing data points between 2 M and 4 M urea concentrations. (B) Near CD signal (Collected at 282 nm) showing data points between 1 M and 3 M urea concentrations. (C) Phenylalanine Fluorescence CD signal (Collected at 235 nm) showing data points between 2 M and 4 M urea concentrations (D) Truncated Tyrosine CD Signal (Collected at 282 nm) showing data points between 1 M and 3 M urea concentrations.

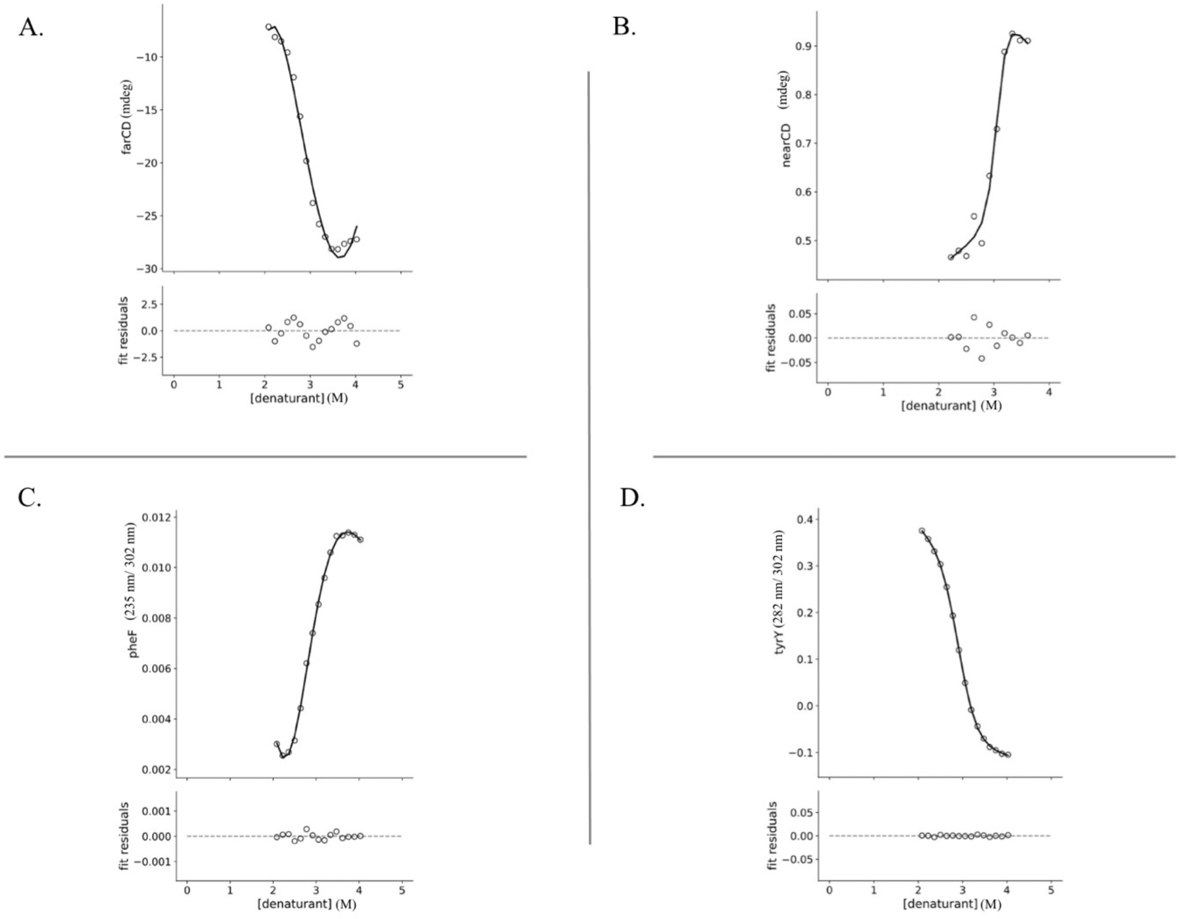


Figure 9. CD Signals for AncA4 Protein. (A) Truncated Far CD signal (collected at 222 nm) between 2 M and 4 M urea concentrations. (B) Truncated Near CD signal (Collected at 282 nm) showing data points between 2 M and 4 M urea concentrations. (C) Phenylalanine Fluorescence CD signal (Collected at 235 nm) showing data points between 2 M and 4 M urea concentrations. (D) Truncated Tyrosine CD Signal (Collected at 282 nm) showing data points between 2 M and 4 M urea concentrations.

(A) Far	ml estimate	Lower 95	Upper 95
ΔG (kcal/mol)	2.66 +/- 74.8	-72.1027	77.42575
m	-1.52 +/- 13.3	-14.8031	11.76217

(B) Near	ml estimate	Lower 95	Upper 95
ΔG (kcal/mol)	6.24 +/- 50.5	-44.2067	56.69233
m	-2.84 +/- 24.8	-27.6671	21.98391

(C) Phenylalanine	ml estimate	Lower 95	Upper 95
ΔG (kcal/mol)	1.94 +/- 4.65 x10^6	-4652704	4652708
m	-2.5 +/- 1.2 x10^4	-11508.6	11503.6

(D) Tyrosine	ml estimate	Lower 95	Upper 95
ΔG (kcal/mol)	8.74 +/- 97.3	-88.5146	106.0038
m	-3.68 +/- 40.6	-44.287	36.92366

Table 2. AncA2A4 Protein ΔGs from truncated CD Signals Model Fit

(A) Far	ml estimate	Lower 95	Upper 95
ΔG (kcal/mol)	1.46 x10^{-5} +/- 31.3	-31.3128	31.31282
m	0.48 +/- 3.00	-2.52664	3.481196

(B) Near	ml estimate	Lower 95	Upper 95
ΔG (kcal/mol)	19.16 +/- 540.5	-521.368	559.6784
m	-6.24 +/- 178.9	-185.19	172.7144

(C) Phenylalanine	**ml estimate**	**Lower 95**	**Upper 95**
ΔG (kcal/mol)	3.81 +/- 1.12 x10^4	-11216.1	11223.73
m	-1.54 +/- 3.29 x10^3	-3299.52	3296.444

(D) Tyrosine	**ml estimate**	**Lower 95**	**Upper 95**
ΔG (kcal/mol)	8.37 +/- 324	-316.449	333.1984
m	-2.86 +/- 105	-108.273	102.5432

Table 3. AncA4 Protein ΔGs from truncated CD Signals Model Fit

EDTA Titration Results

We conducted EDTA titrations for all three proteins, aiming to characterize the binding of calcium with each protein and by observing how EDTA displaces calcium (EDTA has a superior binding affinity for calcium compared to these proteins making it an excellent tool for observing the competition for calcium binding). However, upon implementing the binding scheme, we encountered an additional set of challenges. The expected calcium binding behavior, as depicted in Figure 10, did not align with our experimental observations. Our initial goal was to observe a sequential binding process, starting with our dimer protein fully bound to calcium and competing against EDTA to retain its calcium ions.

We expected the first transition to take place after enough EDTA had accumulated in the sample. We anticipated a flat curve initially due to the higher binding affinity of EDTA for free calcium. Next, after all the free calcium in the sample had been bound to EDTA, we expected to observe a change in the signal, as the first calcium ion would be removed from the protein. This signal would then plateau until the EDTA concentration reached a level where it was able to pull off the second calcium from the second low-affinity binding site, (green section of Figure 10A) followed by a plateau in the signal. So, after some additional accumulation of EDTA, we expected to see the dimer protein have its second calcium ion ripped off at its secondary low-affinity site, as indicated by the pink region in the figure. This transition would also require a waiting period for EDTA concentration to increase sufficiently for removal of the third calcium at the first high affinity site. From here, we expected to see a similar waiting and signal changes take place for each of the two high affinity sites of the protein (as depicted by the orange and blue sections of Figure 10A).

In an actual experiment I would have expected to observe a similar pattern to what is depicted in Figure 10B. This is because the results collected are an average representation across a vast number of molecules. This averaging effect causes the stepped curve in Fig. 10A to appear as a single smooth transition. The reason for this being that the molecules within the sample may have varying number of bindings – some might have one calcium bound, others two, and a few might have three. Consequently, instead of distinct steps in the curve, what emerges is a smooth S-shaped curve do to this mixture of binding states across the molecules in the sample.

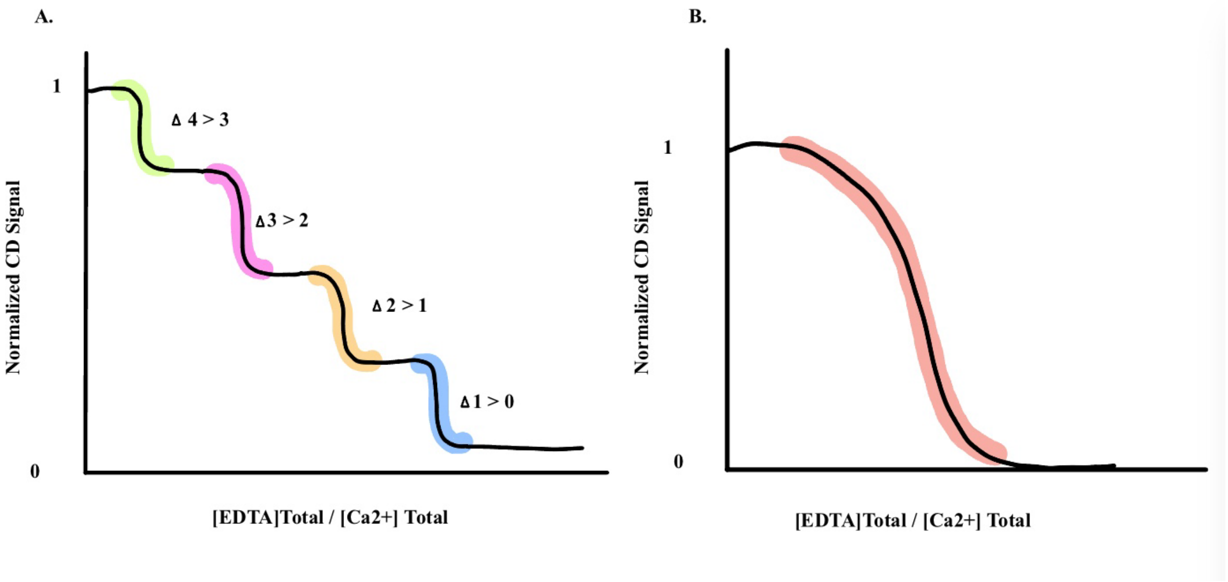


Figure 10. Prediction of Calcium Behavior in Protein Ensemble Using EDTA Titration (A) Conceptual diagram representing the behavior of a protein signal upon unbinding of each calcium. Signal changes when EDTA removes one calcium from the protein. Step curve demonstrates individual molecule behaviors.(B) Illustrates the concept of molecular averaging in experimental data. A smooth S-shaped curve represents the average effect, showcasing the blending of multiple molecules into a cohesive transition.

Our analysis of the data reveals a discrepancy between the expected calcium binding scheme and the actual patterns observed. Contrary to the anticipated binding scheme depicted in Figure 10, our observations align more closely with the dynamics illustrated in Figure 14. Raw data for each protein is presented in Figures 11-13. The graphs in these figures represent the Far CD signals of EDTA titrations carried out for all three proteins, in which the data is normalized to reflect a range from 0 (least structured) to 1 (most structured), with the x-axis indicating the ratio of total EDTA to total calcium. To add on, a molecular interpretation schematic for reciprocal calcium binding is shown in Figure 14. Here, the proteins are depicted as a circle, with smaller circles representing calcium binding sites and a star denoting high-affinity sites, while calcium ions are illustrated as green circles and EDTA as a blue square.

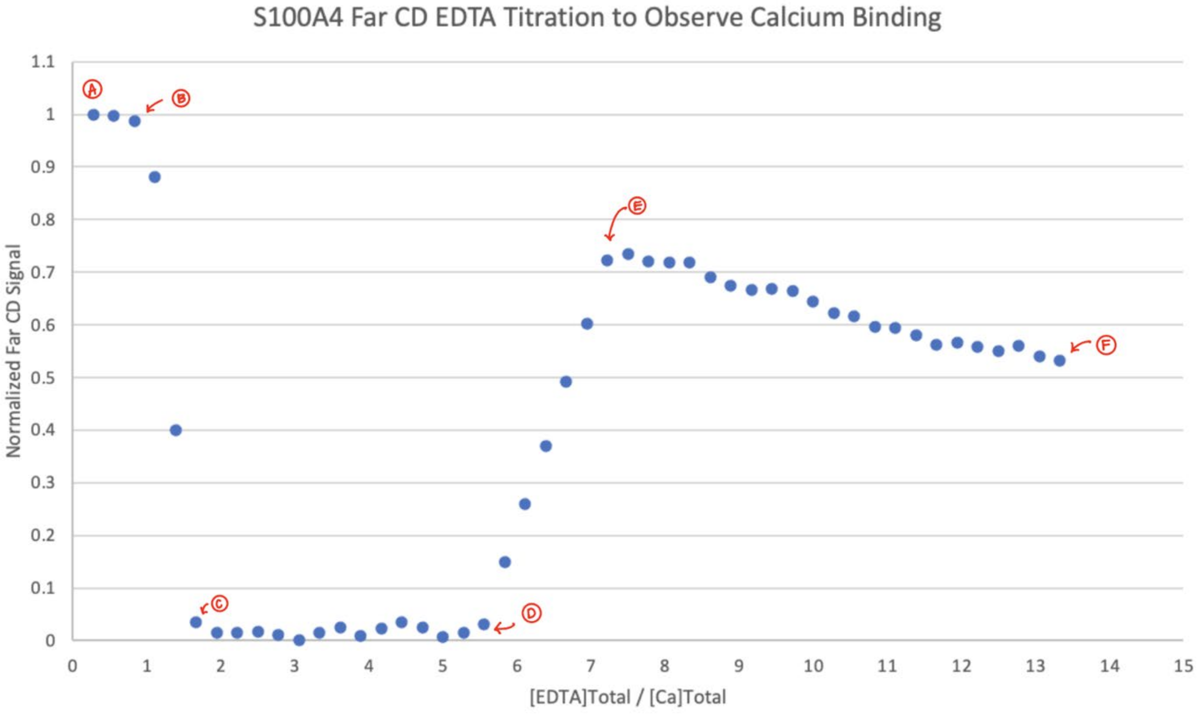


Figure 11. S100A4 Far CD EDTA Titration to Observe Calcium Binding

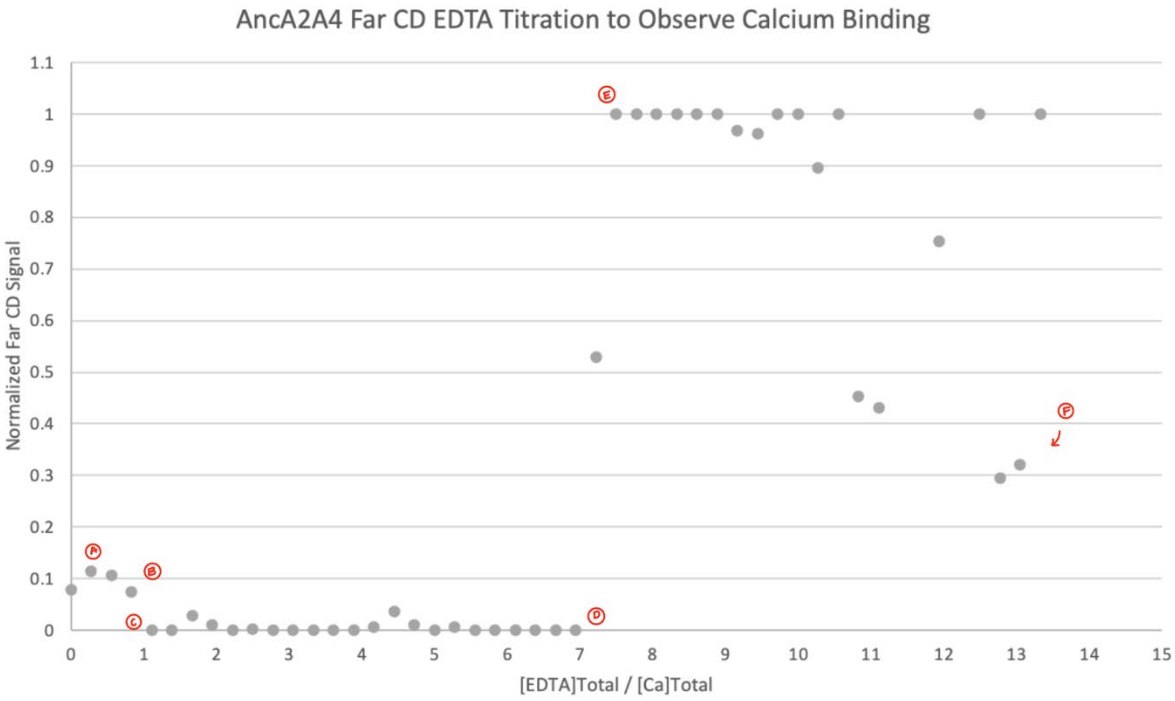


Figure 12. AncA2A4 Far CD EDTA Titration to Observe Calcium Binding

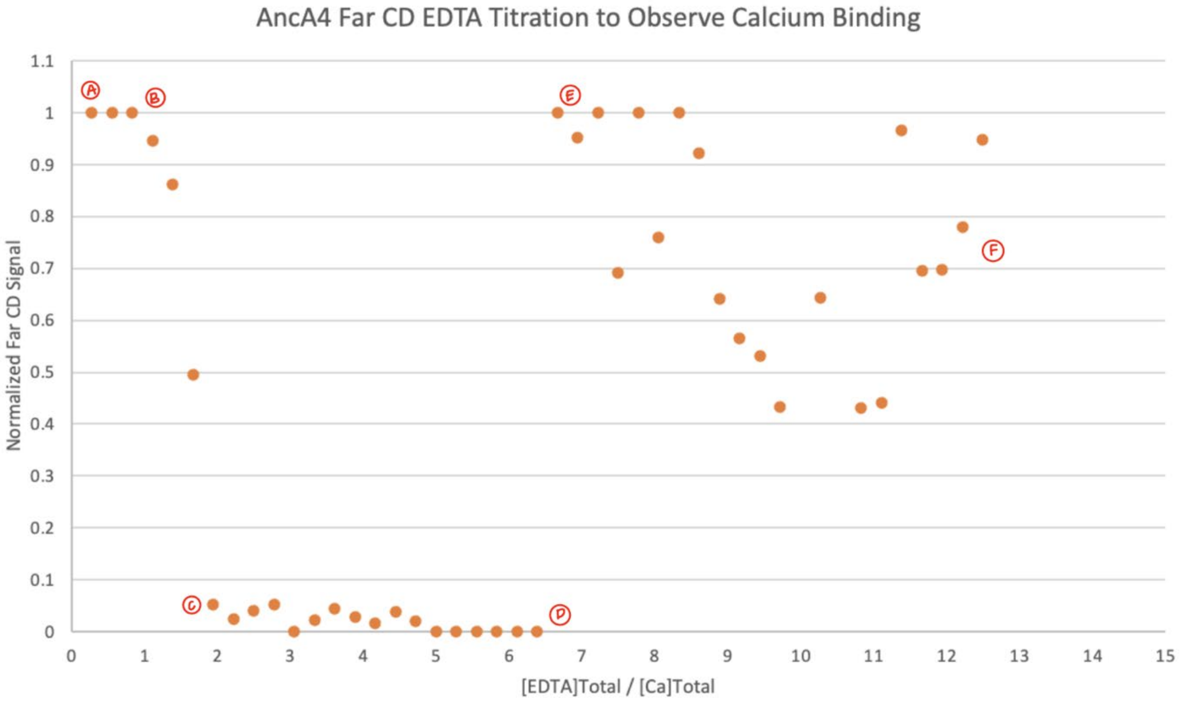


Figure 13. S100A4 Far CD EDTA Titration to Observe Calcium Binding

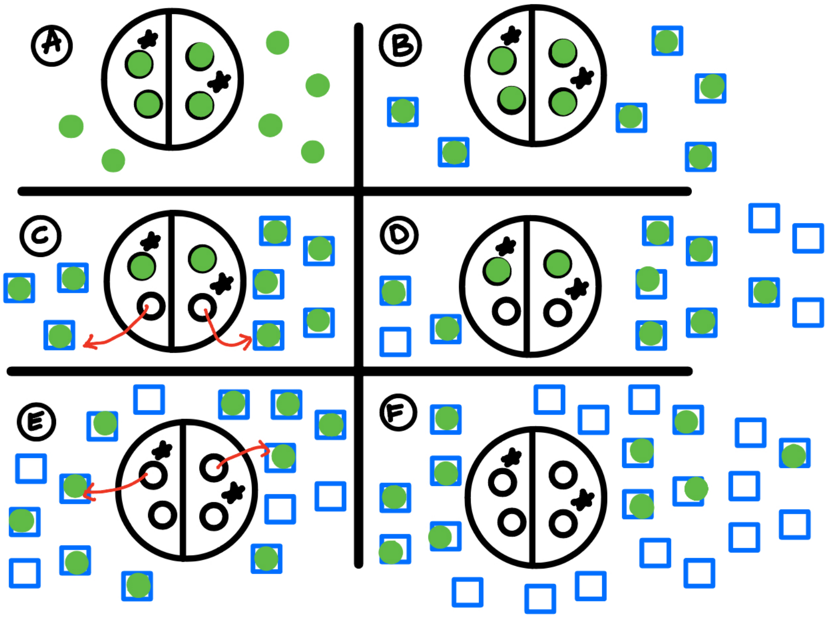


Figure 14. Molecular Mechanism of Reciprocal Calcium Binding

The binding process that we observed in Figures 11-13 appear to unfold in distinct steps: first, EDTA sequesters free calcium (A→B), followed by the removal of calcium from the protein's low-affinity binding sites (B→C). Subsequently, EDTA concentration increases but remains insufficient to extract calcium from high-affinity sites (C→D). But, as EDTA concentration further rises, it successfully extracts calcium from both of the high-affinity sites (D→E), ultimately resulting in the buildup of free EDTA in solution (E→F). These steps, clearly outlined in Figure 14, prompt the conclusion that our current understanding of calcium binding dynamics may not be entirely accurate. A more complicated cooperativity may be taking place within the protein when its binding calcium.

To resolve this discrepancy, additional mathematical modeling and experimental investigations will be necessary. While it's evident that calcium binding occurs in two distinct steps, supported by the raw data revealing a two-step transition (first, the removal of calcium from the low-affinity sites, followed by the removal from the high-affinity sites), the precise mechanism of calcium binding remains elusive based on the current data collected. This observation holds consistent across both the human S100A4 and the two ancestral proteins. Further exploration is warranted to gain a comprehensive understanding of how calcium binding unfolds in the context of S100A4 and its ancestral variants.

Discussion

The research conducted for this thesis sought to characterize the ensembles of S100A4, AncA2A4, and AncA4 proteins by measuring their free energies (ΔG) using biophysical techniques such as ITC and CD. However, the outcomes revealed challenges and limitations that impacted the ability to fully answer the research question with regards to an evaluation on whether the ensemble of protein S100A4 had evolved over time.

Nevertheless, the purification of S100A4 and its ancestors, AncA2A4 and AncA4, was successful. Protein purification protocols were developed and optimized, ensuring consistent outcomes. The biophysical measurements using ITC and CD were conducted to characterize the folding, unfolding, and calcium-binding behavior of the proteins. Initial experiments involving peptide titrations and EDTA to assess calcium binding affinity on the ITC yielded poor results, leading to a shift towards urea melts and EDTA titration on the CD to study unfolding and binding dynamics. These CD experiments provided insights into the stability and structural transitions of S100A4. However, attempts to replicate these experiments with AncA2A4 and AncA4 did not yield comparable results. The ΔG values across each of the 4 signals did not align for any of the proteins studied, indicating that potential oligomerization could be taking place as each of the proteins unfold. Furthermore, EDTA titrations aimed to characterize calcium binding exhibited discrepancies between expected and observed patterns. A two state system for binding calcium was observed, where low-affinity binding took place first, followed by high-affinity binding. The complexity of calcium binding dynamics within the protein ensemble highlighted the need for further mathematical modeling and experimental investigations.

Future Directions

Moving forward, addressing the limitations and struggles encountered in this study will be crucial for advancing the understanding of protein ensembles and their evolution. Future research will focus on conducting more replicates and experiments to confirm the observed discrepancies in ΔG values and unfolding behavior. Mathematical modeling will also play a significant role in refining the understanding of calcium binding dynamics and unraveling the complexities within the protein ensemble.

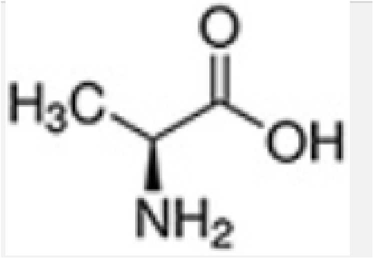
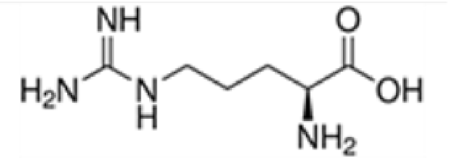
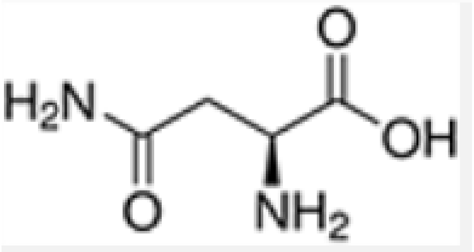
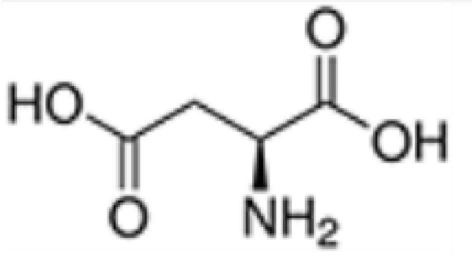
Additionally, exploring alternative biophysical techniques and going back to ITC experimental approaches will be key in providing complementary information about the thermodynamic stability and structural dynamics of S100A4 and its ancestral variants. Not to mention, with the purification process for eight of these S100A4 mutants successfully established, these mutants, when analyzed, will shed light on crucial aspects like folding, calcium-binding behavior, and evolutionary influences for different S100A4 ensembles. This will offer a glimpse into the complex world of intramolecular epistasis. Despite its diverse and challenging nature, understanding intramolecular epistasis is key for predicting and explaining various patterns seen in protein evolution, and considering conformational ensembles may hold the key to unraveling this complexity.

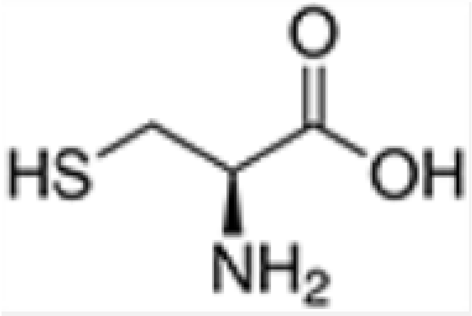
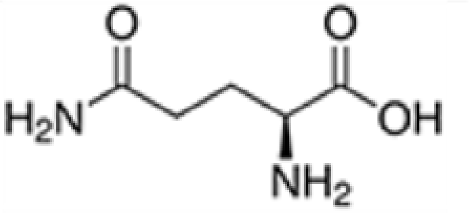
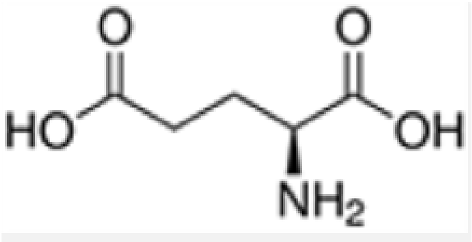
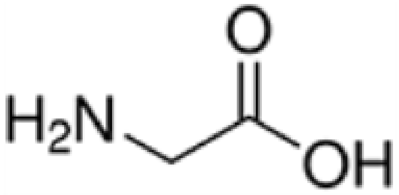
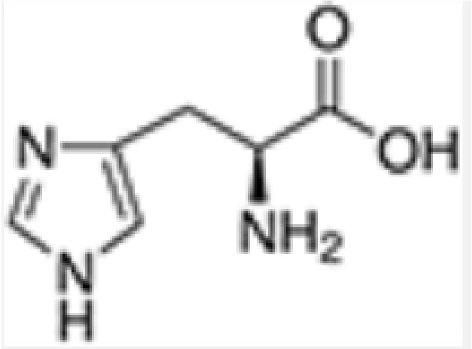
Overview

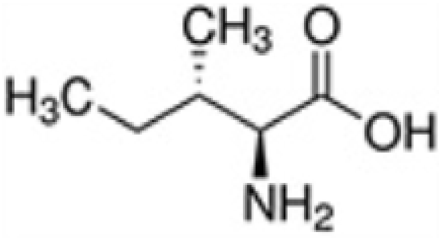
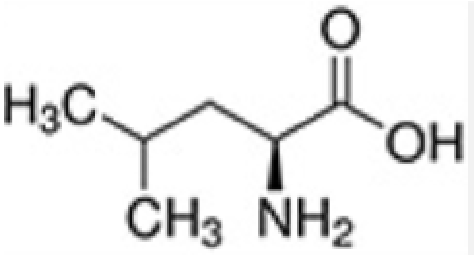
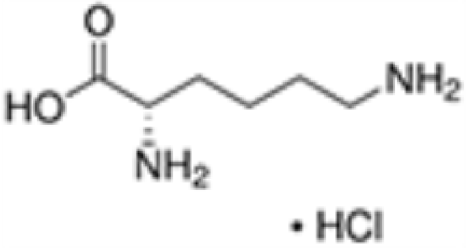
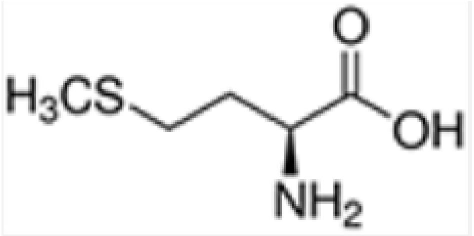
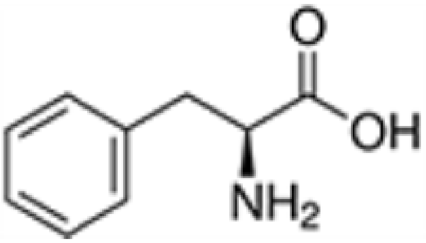
Despite encountering challenges and limitations, the research approach involving protein purification, biophysical measurements, and data analysis laid the groundwork for future investigations. The commitment to refining protocols and optimizing experimental conditions demonstrates the dedication to producing reliable and reproducible outcomes in biochemistry research. While the initial results did not fully answer my research question, they provide valuable information and directions regarding the evolution of the protein ensemble of S100A4.

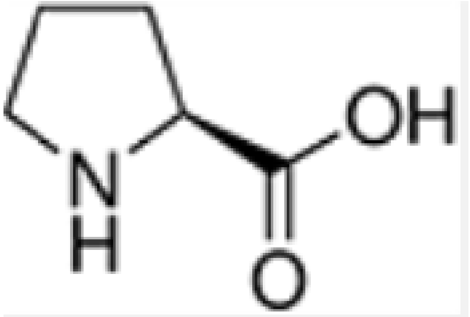
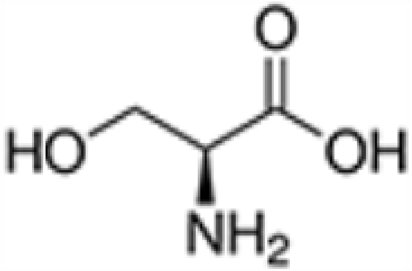
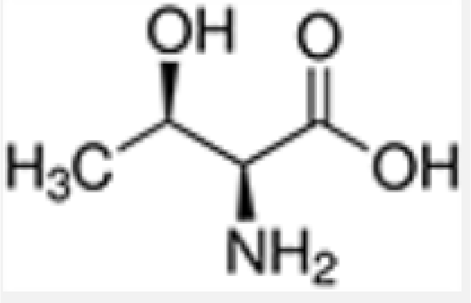
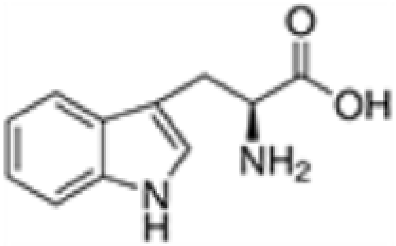
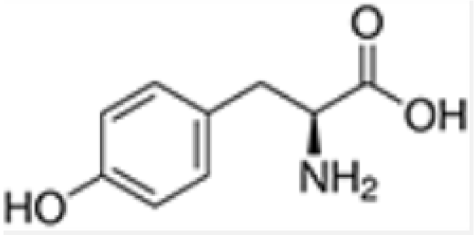
Glossary

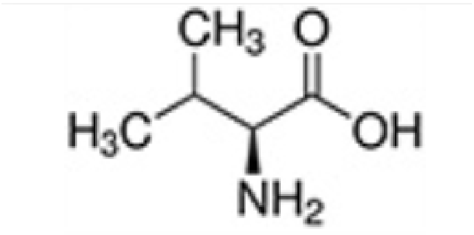
Amino Acid/residue: Amino acids are fundamental components of proteins, characterized by the presence of amine and carboxyl functional groups, as well as a distinct side chain unique to each amino acid. There are twenty different naturally occurring types of amino acids, each with its own name, properties, and chemical structures.

Full Name	1-letter code	Chemical Properties	Chemical Structure
Alanine	A	Hydrophobic, aliphatic	
Arginine	R	Charged, basic	
Asparagine	N	Polar, neutral	
Aspartic Acid	D	Charged, acidic	

Cysteine	C	Polar, neutral	
Glutamine	Q	Polar, neutral	
Glutamic Acid	E	Charged, acidic	
Glycine	G	Unique, no side chain	
Histidine	H	Charged, basic	

Isoleucine	I	Hydrophobic, aliphatic	
Leucine	L	Hydrophobic, aliphatic	
Lysine	K	Charged, basic	
Methionine	M	Hydrophobic, aliphatic	
Phenylalanine	F	Hydrophobic, aromatic	

Proline	P	Unique	
Serine	S	Polar, neutral	
Threonine	T	Polar, neutral	
Tryptophan	W	Hydrophobic, aromatic	
Tyrosine	Y	Hydrophobic, aromatic	

Valine	V	Hydrophobic, aliphatic	

Apo: a protein, that lacks its bound ligand or cofactor.

Biophysical Characterization: Analyzing the physical properties of biological molecules or systems, using techniques like spectroscopy or microscopy.

Buffer: A solution that resists changes in pH when acids or bases are added, crucial for maintaining stable conditions in biological reactions.

Bacterial Cultures: Cultivated populations of bacteria grown under controlled conditions

Conformations: Different spatial arrangements or shapes that molecules, especially proteins, can adopt, influencing their function.

Cell Membranes: Protective barriers surrounding cells, composed of lipids and proteins, controlling the movement of substances in and out of cells.

C-Terminal Residues: Amino acid residues at the end of a protein chain.

Circular Dichroism (CD): A spectroscopic technique used to study the secondary structure of proteins (such as α-helical structures, and β-sheets)

Dimer: A complex formed by two identical or similar molecules, often proteins, binding together.

Dialysis: A method for separating molecules in a solution based on their size and diffusion properties.

Elution: The process of extracting or separating a protein, from a column.

EDTA: Ethylenediaminetetraacetic acid, a chelating agent used to bind metal ions (such as calcium)

Enthalpy: a thermodynamic quantity that represents the total heat content of a system at constant pressure, including both the internal energy and the energy required to displace the system's surroundings.

Entropy: A measure of disorder or randomness in a system.

Ensemble: the dynamic collection of several conformations or structural states that a protein molecule can adopt

Epistasis: a phenomenon in which multiple mutations in a protein, while individually having minimal impact on its function, interact synergistically when present together, leading to significant alterations in the protein's behavior and functionality.

Epistatic Hotspots: Specific regions in a protein's sequence where epistatic interactions frequently occur.

Evolution: a process that results in changes in the genetic material (DNA) of a population over time.

Equilibrium Constant: A measure of the ratio of concentrations of reactants and products at equilibrium in a chemical reaction

Ensemble Epistasis: a phenomenon in which multiple mutations collaborate to collectively modify the behavior and function of a protein.

Intramolecular Epistasis: the phenomenon in which one mutation alters the effect of other mutations to the same protein

Isothermal Titration Calorimetry: A technique used to measure heat changes in biochemical reactions.

IPTG: Isopropyl β-D-1-thiogalactopyranoside, a chemical used to induce protein expression in recombinant DNA technology.

Fractions: individual portions obtained after running a sample through fast protein liquid chromatography (FPLC)

Hydrophobic: the property of a molecule that repels water

kDa: Kilodalton, a unit of molecular weight used to describe proteins and peptides.

Ligand: A molecule that binds to a receptor or protein

Lysis Buffer: A solution used to break open cells and release cellular components for analysis.

Monomer: the individual subunit of a protein molecule

Mutations: changes or alterations that occur in the DNA sequence of an organism, influencing traits and evolution.

Oligomerization: The process of forming complexes or structures composed of multiple subunits or molecules.

Pathogenesis: The mechanism by which a disease develops and progresses in an organism.

Peptide: A short chain of amino acids linked by peptide bonds, smaller than a protein.

Phenotype: The observable traits or characteristics of an organism, influenced by genetics and environment.

Plasmid: A small, circular DNA molecule used in genetic engineering and cloning.

Protein: A large biomolecule composed of amino acids, essential for the structure and function of cells.

S100A4: A human protein, member of the S100 protein family involved in cell signaling and cancer progression.

Stoichiometry: The quantitative relationship between reactants and products in a chemical reaction.

Supernatant: The liquid portion remaining after centrifugation of a mixture.

SDS-PAGE Gel: Sodium dodecyl sulfate polyacrylamide gel electrophoresis, a technique for separating proteins based on size.

Plasmid Transformation: Introducing foreign DNA into bacteria or other cells using plasmids.

Titrate: To slowly add more of something while measuring the response due to its addition.

Truncation: refers to the removal of a portion from a molecule, such as a protein, by cutting or deleting a specific segment. For instance, in the context of proteins, truncation involves eliminating a sequence of amino acids. For data, truncation can mean reducing the length or amount of information by removing part of it.

Two-State System: A system or process with two distinct states

Urea Melts: Experimental procedures using urea to denature or unfold proteins for analysis.

High Affinity Sites: Regions where molecules bind strongly

Low Affinity Sites: Regions where molecules bind weakly

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell* (4th ed.). Garland Science.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., & Tawfik, D. S. (2006). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature, 444(7121), 929–932.

Bloom, J. D., Gong, L. I., & Baltimore, D. (2010). Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. Science, 328(5983), 1272–1275. https://doi.org/10.1126/science.1187816

Bonomi, M., Heller, G. T., Camilloni, C., & Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Current opinion in structural biology*, *42*, 106–116. https://doi.org/10.1016/j.sbi.2016.12.004

Campbell, E., et al. (2016). The Role of Protein Dynamics in the Evolution of New Enzyme Function. Nature Chemical Biology, 12(12), 944–950.

Dobson C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), 884–890. https://doi.org/10.1038/nature02261

Duelli, A., Kiss, B., Lundholm, I., Bodor, A., Petoukhov, M. V., Svergun, D. I., Nyitray, L., & Katona, G. (2014). The C-Terminal Random Coil Region Tunes the Ca2+-Binding Affinity of S100A4 through Conformational Activation. PLOS ONE, 9(5), e97654–e97654. https://doi.org/10.1371/journal.pone.0097654

Forsén, S., & Linse, S. (1995). Cooperativity: over the Hill. Trends in biochemical sciences, 20(12), 495–497. https://doi.org/10.1016/s0968-0004(00)89115-x

Francis, D. M., & Page, R. (2010). Strategies to optimize protein expression in E. coli. *Current protocols in protein science*, *Chapter 5*(1), 5.24.1–5.24.29. https://doi.org/10.1002/0471140864.ps0524s61

Greenfield N. J. (2006). Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols*, *1*(6), 2876–2890. https://doi.org/10.1038/nprot.2006.202

Guerrero, R. F., Scarpino, S. V., Rodrigues, J. V., Hartl, D. L. & Ogbunugafor, C. B. Proteostasis Environment Shapes Higher-Order Epistasis Operating on Antibiotic Resistance. Genetics 212, 565–575 (2019).

Harms, M. J. (2023, November 6). Grant. Unpublished

Harms, M. (2023) EDTA-S100A4 Competition Scheme. Unpublished

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

Lewis, E. A., & Murphy, K. P. (2005). Isothermal titration calorimetry. *Methods in molecular biology (Clifton, N.J.)*, *305*, 1–16. https://doi.org/10.1385/1-59259-912-5:001

Malashkevich, V. N., Varney, K. M., Garrett, S. C., Wilder, P. T., Knight, D., Charpentier, T. H., Ramagopal, U. A., Almo, S. C., Weber, D. J., & Bresnick, A. R. (2008). Structure of Ca2+-bound S100A4 and its interaction with peptides derived from nonmuscle myosin-IIA. *Biochemistry*, *47*(18), 5111–5126. https://doi.org/10.1021/bi702537s

Modi, T., Campitelli, P., Kazan, I. C., & Ozkan, S. B. (2021). Protein folding stability and binding interactions through the lens of evolution: a dynamical perspective. Current Opinion in Structural Biology, 66, 207–215.

Morrison, A. J., Wonderlick, D. R. & Harms, M. J. Ensemble epistasis: thermodynamic origins of nonadditivity between mutations. Genetics (2021) doi:10.1093/genetics/iyab105.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., & Thornton, J. W. (2007). Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. Science, 317, 1544–1548.

Otwinowski, J. (2018). Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. Molecular Biology and Evolution, 35(10), 2345–2354. https://doi.org/10.1093/molbev/msy141

Schenk, M. F., Szendro, I. G., Salverda, M. L. M., Krug, J., & de Visser, J. A. G. M. (2013). Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene. Molecular Biology and Evolution, 30, 1779–1787.

Strokach, A., Corbi-Verge, C., Teyra, J., & Kim, P. M. (2019). Predicting the Effect of Mutations on Protein Folding and Protein-Protein Interactions. *Methods in molecular biology (Clifton, N.J.)*, *1851*, 1–17. https://doi.org/10.1007/978-1-4939-8736-8_1

Transformation Efficiency Cells, XL10-Gold Ultracompetent Cells | Agilent. (n.d.). https://www.agilent.com/en/product/mutagenesis-cloning/competent-cells-competent-cell-supplies/competent-cells-for-difficult-cloning/xl10-gold-ultracompetent-cells-233087#:~:text=The%20XL10%2DGold%20ultracompetent%20cells,faster%20growth%20and%20larger%20colonies.

Tegel, H., Tourle, S., Ottosson, J., & Persson, A. (2010). Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). *Protein expression and purification*, *69*(2), 159–167. https://doi.org/10.1016/j.pep.2009.08.017

Thermo Fisher Scientific. (n.d.). Overview of Electrophoresis. Retrieved from https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-electrophoresis.html#:~:text=SDS%2DPAGE%20separates%20proteins%20primarily,toward%20the%20positively%20charged%20electrode.

Wheeler, L., Harms, M. Human S100A5 binds Ca2+ and Cu2+ independently. BMC Biophys10, 8 (2017). https://doi.org/10.1186/s13628-017-0040-y